IBM SPSS Modeler 17.1 Guida al mining nel database

Nota
Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni in “Note” a pagina 117.

Informazioni sul prodotto

Questa edizione si applica alla versione 17, release 1, modifica 0 di IBM(r) SPSS(r) Modeler e a tutte le modifiche e release successive se non diversamente indicato nelle nuove edizioni.

Indice

Prefazione vii

Capitolo 1. Informazioni su IBM SPSS Modeler 1

Prodotti IBM SPSS Modeler 1
IBM SPSS Modeler 1
IBM SPSS Modeler Server 1
IBM SPSS Modeler Administration Console . . . 2
IBM SPSS Modeler Batch 2
IBM SPSS Modeler Solution Publisher 2
Adattatori IBM SPSS Modeler Server per IBM SPSS Collaboration and Deployment Services . . 2
Edizioni di IBM SPSS Modeler 2
Documentazione di IBM SPSS Modeler 3
Documentazione di SPSS Modeler Professional . . 3
Documentazione di SPSS Modeler Premium . . . 4
Esempi di applicazioni 5
Cartella Demos 5

Capitolo 2. Mining nel database 7

Panoramica sulla modellazione di database 7
Requisiti necessari 7
Creazione del modello 8
Data Preparation 8
Calcolo del punteggio del modello 8
Esportazione e salvataggio di modelli di database 9
Uniformità dei modelli 9
Visualizzazione ed esportazione di codice SQL generato 9

Capitolo 3. Modellazione di database con Microsoft Analysis Services 11

IBM SPSS Modeler e Microsoft Analysis Services . . 11
Requisiti per l'integrazione con Microsoft Analysis Services 12
Attivazione dell'integrazione con Analysis Services 13
Creazioni di modelli con Analysis Services 15
Gestione di modelli di Analysis Services. 15
Impostazioni comuni a tutti i nodi degli algoritmi 17
Opzioni avanzate Struttura ad albero delle decisioni MS 18
Opzioni avanzate Raggruppamento cluster MS 18
Opzioni avanzate Naive Bayes MS. 18
Opzioni avanzate Regressione lineare MS . . . 18
Opzioni avanzate Rete neurale MS 18
Opzioni avanzate Regressione logistica MS . . . 18
Nodo Regole di associazione MS 18
Nodo Serie temporali MS. 19
Nodo Cluster di sequenze MS 20
Calcolo del punteggio per i modelli di Analysis Services 21
Impostazioni comuni a tutti i modelli di Analysis Services 22
Nugget del modello Serie temporali MS 23
Nugget del modello Cluster di sequenze MS . . 24
Esportazione di modelli e generazione di nodi. . 24
Esempi di mining con Analysis Services 24
Flusso di esempio: strutture ad albero delle decisioni 25

Capitolo 4. Modellazione di database con Oracle Data Mining 29

Informazioni su Oracle Data Mining 29
Requisiti per l'integrazione con Oracle 29
Attivazione dell'integrazione con Oracle 30
Creazione di modelli con Oracle Data Mining . . . 31
Opzioni della scheda Server dei modelli Oracle 32
Costi classificazione errata 32
Naive Bayes Oracle. 33
Opzioni del modello Naive Bayes 33
Opzioni avanzate di Naive Bayes 34
Bayes adattivi Oracle 34
Opzioni del modello Bayes adattivo 34
Opzioni avanzate di Bayes adattivo 35
Support Vector Machine Oracle (SVM) 35
Opzioni del modello SVM Oracle 36
Opzioni avanzate di SVM Oracle 36
Opzioni Pesi di SVM Oracle 37
Modelli lineari generalizzati Oracle (GLM) 37
Opzioni del modello GLM Oracle 38
Opzioni avanzate di GLM Oracle 38
Opzioni Pesi di GLM Oracle. 39
Struttura ad albero delle decisioni Oracle 39
Opzioni della scheda Modello per il nodo Struttura ad albero delle decisioni 40
Opzioni avanzate Struttura ad albero delle decisioni 40
O-Cluster Oracle. 41
Opzioni del modello O-Cluster 41
Opzioni avanzate di O-Cluster 41
Medie K Oracle 41
Opzioni del modello Medie K 42
Opzioni avanzate del nodo Medie K 42
NMF di Oracle (fattorizzazione a matrice non negativa) 42
Opzioni del modello NMF 43
Opzioni avanzate NMF 43
Apriori Oracle 43
Opzioni dei campi Apriori 44
Opzioni del modello Apriori. 45
Oracle MDL (Lunghezza descrizione minima) . . . 45
Opzioni del modello MDL 46
Importanza attributo Oracle (AI) 46
Opzioni modello AI 46
Opzioni di selezione AI 46
Scheda Modello del nugget del modello AI. . . 47
Gestione dei modelli Oracle 47
Scheda Server del nugget del modello Oracle . . 47

Scheda Riepilogo del nugget del modello Oracle 47
Scheda Impostazioni del nugget del modello Oracle 48
Elenco dei modelli Oracle 48
Oracle Data Miner 48
Preparazione dei dati 49
Esempi di Oracle Data Mining 50
Flusso di esempio: caricamento dati 50
Flusso di esempio: esplorazione dati 51
Flusso di esempio: creazione modello. 51
Flusso di esempio: valutazione modello 51
Flusso di esempio: Deployment del modello . . 51

Capitolo 5. Modellazione di database con IBM InfoSphere Warehouse 53

IBM InfoSphere Warehouse e IBM SPSS Modeler . . 53
Requisiti per l'integrazione con IBM InfoSphere Warehouse. 53
Attivazione dell'integrazione con IBM InfoSphere Warehouse. 53
Creazione di modelli con IBM InfoSphere Warehouse Data Mining 57
Calcolo del punteggio e deployment del modello 57
Gestione dei modelli DB2. 58
Elenco dei modelli in-database 59
Visualizzazione dei modelli 59
Esportazione di modelli e generazione di nodi. . 59
Impostazioni dei nodi comuni a tutti gli algoritmi 59
Struttura ad albero delle decisioni ISW 61
Opzioni della scheda Modello per il nodo Struttura ad albero delle decisioni ISW 62
Opzioni avanzate Struttura ad albero delle decisioni ISW. 62
Associazione ISW 62
Opzioni dei campi Associazione ISW 63
Opzioni della scheda Modello per il nodo Associazione ISW 64
Opzioni della scheda Opzioni avanzate per il nodo Associazione ISW 65
Opzioni della scheda Tassonomia per ISW . . . 65
Sequenza ISW 66
Opzioni della scheda Modello per il nodo Sequenza ISW 66
Opzioni della scheda Opzioni avanzate per il nodo Sequenza ISW 67
Regressione ISW. 67
Opzioni della scheda Modello per il nodo Regressione ISW. 68
Opzioni avanzate del nodo Regressione ISW . . 69
Raggruppamento cluster ISW 70
Opzioni della scheda Modello per il nodo Raggruppamento cluster ISW 71
Opzioni avanzate del nodo Raggruppamento cluster ISW 71
Naive Bayes ISW 73
Opzioni del modello Naive Bayes ISW 73
Regressione logistica ISW. 73
Opzioni del modello di Regressione logistica ISW 73
Serie temporali ISW 73
Opzioni Campi Serie temporali ISW 74
Opzioni del modello di serie temporali ISW . . 74
Opzioni avanzate per le serie temporali ISW . . 75
Visualizzazione dei modelli di serie temporali ISW 75
Nugget del modello di ISW Data Mining 75
Scheda Server del nugget del modello ISW . . . 75
Scheda Impostazioni del nugget del modello ISW 76
Scheda Riepilogo del nugget del modello ISW. . 76
Esempi di ISW Data Mining. 77
Flusso di esempio: caricamento dati 77
Flusso di esempio: Esplorazione dati 77
Flusso di esempio: creazione modello. 77
Flusso di esempio: valutazione modello 78
Flusso di esempio: Deployment del modello . . 78

Capitolo 6. Modellazione di database con IBM Netezza Analytics 79

IBM SPSS Modeler and IBM Netezza Analytics . . 79
Requisiti per l'integrazione con IBM Netezza Analytics 79
Abilitazione dell'integrazione con IBM Netezza Analytics 80
Configurazione di IBM Netezza Analytics . . . 80
Creazione di un'origine ODBC per IBM Netezza Analytics 80
Attivazione dell'integrazione IBM Netezza Analytics in IBM SPSS Modeler. 81
Attivazione di generazione e ottimizzazione SQL 82
Creazione dei modelli con IBM Netezza Analytics 82
Opzioni della scheda Campi dei modelli Netezza 83
Opzioni della scheda Server dei modelli Netezza 83
Modelli Netezza - Opzioni Modello 84
Gestione dei modelli Netezza 84
Elenco dei modelli in-database 84
Struttura ad albero di regressione Netezza 85
Opzioni di creazione della struttura ad albero di regressione Netezza - Espansione della struttura ad albero 85
Opzioni di creazione della struttura ad albero di regressione Netezza - Taglio della struttura ad albero 86
Raggruppamento cluster divisivo Netezza 86
Opzioni dei campi di raggruppamento cluster divisivo Netezza. 87
Opzioni di creazione del raggruppamento cluster divisivo Netezza. 87
Lineare generalizzato Netezza 88
Opzioni del campo del modello lineare generalizzato Netezza 88
Opzioni del modello lineare generalizzato Netezza - Generale 89
Opzioni Modello lineare generalizzato Nettezza - Interazione 90
Opzioni del modello lineare generalizzato Netezza - Opzioni di calcolo del punteggio. . . 91
Strutture ad albero delle decisioni di Netezza . . . 91
Pesi delle istanze e delle classi 91
Opzioni dei campi della struttura ad albero delle decisioni di Netezza 92
Opzioni di creazione della struttura ad albero delle decisioni di Netezza 93
Regressione lineare Netezza 94

Opzioni di creazione della regressione lineare Netezza 94
KNN Netezza 95
Opzioni del modello KNN Netezza - Generale . 95
Opzioni del modello KNN Netezza - Opzioni di calcolo del punteggio 96
Medie K Netezza 96
Opzioni dei campi Medie K di Netezza 97
Scheda Opzioni di creazione K-medie di Netezza 97
Naive Bayes Netezza 98
Rete di Bayes Netezza 98
Opzioni dei campi della rete di Bayes Netezza . 98
Opzioni di creazione della rete di Bayes Netezza 99
Serie temporali Netezza 99
Interpolazione dei valori nella serie temporale Netezza 100
Opzioni dei campi della serie temporale Netezza 101
Opzioni di creazione della serie temporale Netezza 102
Opzioni del modello di serie temporali Netezza 104
Netezza TwoStep 104
Opzioni del campo TwoStep di Netezza . . . 105
Opzioni di creazione TwoStep Netezza 105
PCA Netezza 106
Opzioni dei campi PCA Netezza 106
Opzioni di creazione PCA Netezza 106
Gestione di modelli di IBM Netezza Analytics . . 107
Calcolo del punteggio dei modelli IBM Netezza Analytics 107
Scheda Server del nugget del modello Netezza 107
Nugget del modello Struttura ad albero delle decisioni di Netezza 108
Nugget del modello Medie K di Netezza . . . 109
Nugget del modello di rete di Bayes Netezza 109
Nugget del modello Naive Bayes Netezza . . . 110
Nugget del modello KNN Netezza 111
Nugget del modello di raggruppamento cluster divisivo Netezza 112
Nugget del modello PCA Netezza 113
Nugget del modello della struttura ad albero di regressione Netezza 113
Nugget del modello di regressione lineare Netezza 114
Nugget del modello di serie temporali Netezza 115
Nugget del modello lineari generalizzati Netezza 115
Nugget del modello TwoStep di Netezza . . . 116

Note 117
Marchi 118

Indice analitico. 121

Prefazione

IBM® SPSS Modeler è l'efficace workbench di data mining aziendale di IBM Corp.. SPSS Modeler consente alle organizzazioni di migliorare le relazioni con i clienti e con il pubblico grazie a un'analisi approfondita dei dati. Le organizzazioni potranno utilizzare le informazioni ottenute tramite SPSS Modeler per mantenere i clienti di valore, cogliere opportunità di vendite incrociate, attrarre nuovi clienti, individuare frodi, diminuire i rischi e migliorare l'offerta di servizi a livello statale.

L'interfaccia visuale SPSS Modeler invita gli utenti ad applicare la propria competenza di business specifica che conduce a modelli predittivi più efficaci e alla riduzione dei tempi di risoluzione dei problemi. SPSS Modeler offre molte tecniche di modellazione come la previsione, la classificazione, la segmentazione e algoritmi di rilevamento delle associazioni. IBM SPSS Modeler Solution Publisher consente quindi di distribuire a livello aziendale i modelli creati in modo che vengano utilizzati dai responsabili dei processi decisionali oppure inseriti in un database.

Informazioni su IBM Business Analytics

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni di business. Un ampio portafoglio di applicazioni di business intelligence, analisi predittiva, gestione delle prestazioni e delle strategie finanziarie e analisi offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività di business. Clienti commerciali, istituzionali e accademici di tutto il mondo confidano sulla tecnologia IBM SPSS considerandola un vantaggio competitivo per attrarre, mantenere e accrescere il numero di clienti e al contempo ridurre le frodi e mitigare i rischi. Integrando il software IBM SPSS nei processi di tutti i giorni, le aziende diventano aziende predittive - in grado di assumere ed automatizzare decisioni per centrare gli obiettivi ed ottenere vantaggi competitivi misurabili. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito http://www.ibm.com/spss.

Supporto tecnico

Il supporto tecnico è a disposizione dei clienti che dispongono di un contratto di manutenzione. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo di IBM Corp. o per l'installazione di uno degli ambienti hardware supportati. Per contattare il supporto tecnico, consultare il sito Web IBM Corp. all'indirizzo http://www.ibm.com/support. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del accordo di manutenzione.

Capitolo 1. Informazioni su IBM SPSS Modeler

IBM SPSS Modeler è un insieme di strumenti di data mining che consente di sviluppare rapidamente modelli predittivi con l'ausilio di competenze di business e di eseguirne la distribuzione nelle operazioni di business per migliorare i processi decisionali. Progettato secondo il modello CRISP-DM conforme agli standard di settore, IBM SPSS Modeler supporta l'intero processo di data mining, dai dati a risultati di business migliori.

IBM SPSS Modeler offre numerosi metodi di modellazione ricavati dall'apprendimento automatico, dall'intelligenza artificiale e dalla statistica. I metodi disponibili nella palette Modelli consentono di ricavare nuove informazioni dai dati e di sviluppare modelli predittivi. Ogni metodo ha determinati punti di forza e si presta meglio per particolari tipi di problemi.

SPSS Modeler può essere acquistato come prodotto autonomo oppure utilizzato come client in combinazione con SPSS Modeler Server. È inoltre disponibile una serie di opzioni, come illustrato nelle sezioni seguenti. Per ulteriori informazioni, consultare http://www.ibm.com/software/analytics/spss/products/modeler/.

Prodotti IBM SPSS Modeler

La famiglia di prodotti IBM SPSS Modeler e del software associato comprende quanto segue.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Adattatori IBM SPSS Modeler Server per IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler è una versione del prodotto con funzionalità complete che viene installata ed eseguita sul proprio PC. È possibile eseguire SPSS Modeler in modalità locale come prodotto autonomo oppure in modalità distribuita assieme a IBM SPSS Modeler Server per ottenere una migliore performance su insiemi di dati di grandi dimensioni.

Grazie a SPSS Modeler si possono creare, in modo veloce e intuitivo, modelli predittivi accurati senza ricorrere alla programmazione. La sua avanzata interfaccia visiva permette di visualizzare con facilità il processo di data mining. Grazie alle funzionalità di analisi avanzate incorporate nel prodotto, l'utente potrà rilevare la presenza di pattern e tendenze, che altrimenti rimarrebbero occulti, all'interno dei dati. La modellazione dei risultati e la comprensione dei fattori che li influenzano consente di beneficiare di maggiori opportunità di business e, al contempo, di ridurre i rischi.

SPSS Modeler è disponibile in due edizioni: SPSS Modeler Professional e SPSS Modeler Premium. Per ulteriori informazioni, consultare l'argomento “Edizioni di IBM SPSS Modeler” a pagina 2.

IBM SPSS Modeler Server

SPSS Modeler utilizza un'architettura client/server per distribuire le richieste di operazioni che utilizzano molte risorse a potenti componenti software server, con un conseguente miglioramento della performance su insiemi di dati di grandi dimensioni.

SPSS Modeler Server è un prodotto con licenza separata che viene eseguito continuamente in modalità di analisi distribuita su un host server insieme a una o più installazioni IBM SPSS Modeler. Una configurazione di questo tipo consente a SPSS Modeler Server di ottenere prestazioni migliori quando si lavora su insiemi di dati di grandi dimensioni, in quanto le operazioni che richiedono un utilizzo consistente della memoria possono essere eseguite sul server senza scaricare i dati sul computer client. IBM SPSS Modeler Server offre inoltre il supporto delle funzionalità di ottimizzazione SQL e di modellazione nel database, garantendo ulteriori benefici dal punto di vista delle prestazioni e del livello di automazione.

IBM SPSS Modeler Administration Console

Modeler Administration Console è un'applicazione grafica per la gestione di molte delle opzioni di configurazione di SPSS Modeler Server, la cui configurazione può avvenire, inoltre, mediante un file delle opzioni. L'applicazione fornisce un'interfaccia utente di console per monitorare e configurare le installazioni di SPSS Modeler Server ed è disponibile gratuitamente per i clienti esistenti di SPSS Modeler Server. L'applicazione può essere installata solo sui computer Windows; tuttavia, può gestire un server installato su qualsiasi piattaforma supportata.

IBM SPSS Modeler Batch

Nonostante il data mining sia generalmente un processo di tipo interattivo, è possibile eseguire SPSS Modeler da una riga di comando senza il bisogno di ricorrere all'interfaccia utente grafica. Poniamo, ad esempio, che si debbano svolgere varie attività laboriose e ripetitive che non richiedono l'intervento di un utente. SPSS Modeler Batch è una versione speciale del prodotto che supporta l'intera gamma di funzionalità analitiche di SPSS Modeler senza richiedere l'accesso all'interfaccia utente normale. Per utilizzare SPSS Modeler Batch, è richiesto SPSS Modeler Server.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher è uno strumento che consente di creare una versione a pacchetto di un flusso SPSS Modeler che potrà essere eseguito da un motore di runtime esterno oppure incorporato in una applicazione esterna. Questo permette di pubblicare e sottoporre a deployment stream SPSS Modeler completi in ambienti in cui SPSS Modeler non è installato. SPSS Modeler Solution Publisher è distribuito come parte del servizioIBM SPSS Collaboration and Deployment Services - Scoring, per cui è necessario procurarsi una licenza separata. Insieme alla licenza, si riceve SPSS Modeler Solution Publisher Runtime, che consente di eseguire i flussi pubblicati.

Per ulteriori informazioni relative a SPSS Modeler Solution Publisher, consultare la documentazione di IBM SPSS Collaboration and Deployment Services. Il IBM SPSS Collaboration and Deployment Services Knowledge Center contiene sezioni denominate "IBM SPSS Modeler Solution Publisher" e "IBM SPSS Analytics Toolkit".

Adattatori IBM SPSS Modeler Server per IBM SPSS Collaboration and Deployment Services

È disponibile una serie di adattatori per IBM SPSS Collaboration and Deployment Services che abilitano l'interazione di SPSS Modeler e SPSS Modeler Server con un repository IBM SPSS Collaboration and Deployment Services. In questo modo, un flusso SPSS Modeler sottoposto a deployment sul repository potrà essere condiviso da più utenti oppure risulterà accessibile dall'applicazione thin client IBM SPSS Modeler Advantage. L'adattatore va installato sul sistema che ospita il repository.

Edizioni di IBM SPSS Modeler

SPSS Modeler è disponibile nelle edizioni seguenti.

SPSS Modeler Professional

SPSS Modeler Professional contiene tutti gli strumenti necessari per utilizzare la maggior parte dei tipi di dati strutturati, quali comportamenti e interazioni registrati in sistemi CRM, dati demografici, dati sulle vendite e sul comportamento d'acquisto.

SPSS Modeler Premium

SPSS Modeler Premium è un prodotto con licenza separata che amplia l'ambito di utilizzo di SPSS Modeler Professional aggiungendo il supporto di dati speciali, quali quelli usati per l'analisi delle entità o dei social network, e di dati di testo non strutturati. SPSS Modeler Premium comprende i seguenti componenti.

IBM SPSS Modeler Entity Analytics aggiunge una dimensione supplementare alle analisi predittive di IBM SPSS Modeler . Se l'analisi predittiva tenta di prevedere il comportamento futuro sulla base di dati precedenti, l'analisi dell'entità si concentra sul miglioramento della coerenza dei dati correnti risolvendo i conflitti tra gli stessi record. Un'identità può essere di un individuo, un'organizzazione, un oggetto o qualsiasi altra entità per cui possa esistere ambiguità. La risoluzione dell'identità può essere essenziale in diversi campi, tra cui la gestione delle relazioni con i clienti, il rilevamento di frodi, il riciclaggio di denaro e la sicurezza nazionale e internazionale.

IBM SPSS Modeler Social Network Analysis trasforma le informazioni sulle relazioni in campi che caratterizzano il comportamento sociale di individui e gruppi. Facendo leva sui dati che descrivono le relazioni esistenti nelle reti sociali, IBM SPSS Modeler Social Network Analysis riesce a individuare i leader in grado di influenzare il comportamento degli altri membri della rete. Consente inoltre di stabilire quali individui della rete sono maggiormente influenzati dagli altri membri. La combinazione di questi risultati ad altre misurazioni permette di delineare profili complessi degli individui su cui basare dei modelli predittivi. I modelli che contengono informazioni sociali generano risultati più accurati rispetto agli altri.

IBM SPSS Modeler Text Analytics utilizza tecnologie linguistiche avanzate e di NLP (Natural Language Processing) per elaborare rapidamente una grande varietà di dati di testo non strutturati, estrarre ed organizzare i concetti chiave e raggruppare tali concetti in categorie. È quindi possibile combinare i concetti e le categorie estratti con dati strutturati esistenti, per esempio dati demografici, e applicarli alla modellazione utilizzando la suite completa degli strumenti di data mining di IBM SPSS Modeler per prendere decisioni migliori e più mirate.

Documentazione di IBM SPSS Modeler

La documentazione nel formato guida in linea è disponibile nel menu Aiuto di SPSS Modeler. Sono incluse la documentazione per SPSS Modeler, SPSS Modeler Server, nonché la Guida alle applicazioni (indicato anche come Supporto didattico) e altro materiale di supporto.

La documentazione completa in formato PDF dei singoli prodotti, istruzioni di installazione comprese, è disponibile nella cartella *Documentation* del DVD di ciascun prodotto. I documenti relativi all'installazione possono anche essere scaricati dal Web, alla pagina http://www.ibm.com/support/docview.wss?uid=swg27043831.

La documentazione in entrambi i formati è disponibile anche da SPSS Modeler Knowledge Center all'indirizzo http://www-01.ibm.com/support/knowledgecenter/SS3RA7_17.0.0.0.

Documentazione di SPSS Modeler Professional

La documentazione completa di SPSS Modeler Professional, escluse le istruzioni di installazione, è la seguente.

- **IBM SPSS Modeler - Guida per l'utente.** Introduzione generale all'utilizzo di SPSS Modeler che illustra come creare flussi di dati, gestire valori mancanti, generare espressioni CLEM, utilizzare progetti e report e assemblare flussi per la distribuzione tramite IBM SPSS Collaboration and Deployment Services, le applicazioni predittive o IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler - Nodi origine, elaborazione e output.** Descrizioni di tutti i nodi utilizzati per leggere, elaborare e generare dati di output in vari formati, ovvero di nodi ad eccezione dei nodi Modelli.
- **IBM SPSS Modeler - Nodi di modellazione.** Descrizioni di tutti i nodi utilizzati per creare modelli data mining. IBM SPSS Modeler offre numerosi metodi di modellazione ricavati dall'apprendimento automatico, dall'intelligenza artificiale e dalla statistica.
- **IBM SPSS Modeler Algorithms Guide.** Descrizione dei fondamenti di matematica per i metodi di modellazione utilizzati in IBM SPSS Modeler. Questa guida è disponibile solo in formato PDF.
- **IBM SPSS Modeler Applications Guide.** Gli esempi inclusi in questa guida forniscono indicazioni mirate e sintetiche su specifici metodi e tecniche di modellazione. Una versione in linea di questa guida è inoltre disponibile dal menu Aiuto. Per ulteriori informazioni, consultare l'argomento "Esempi di applicazioni" a pagina 5.
- **IBM SPSS Modeler Python Scripting and Automation.** Informazioni sull'automazione del sistema mediante gli script Python. incluse le proprietà che è possibile utilizzare per manipolare nodi e flussi.
- **IBM SPSS Modeler - Guida alla distribuzione.** Informazioni sull'esecuzione di flussi e scenari IBM SPSS Modelercome fasi dell'elaborazione di lavori in IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler - Guida per lo sviluppatore CLEF.** CLEF consente di integrare programmi di terze parti come routine di elaborazione dei dati o algoritmi di modellazione come nodi in IBM SPSS Modeler.
- **IBM SPSS Modeler - Guida al mining nel database.** Informazioni sulle modalità per utilizzare al meglio la potenza del database in uso al fine di ottenere prestazioni migliori ed estendere la gamma di funzionalità analitiche tramite algoritmi di terze parti.
- **IBM SPSS Modeler Server - Guida della performance e amministrazione.** Informazioni su come configurare e amministrare IBM SPSS Modeler Server.
- **IBM SPSS Modeler Administration Console - Guida per l'utente.** Informazioni sull'installazione e l'utilizzo dell'interfaccia utente della console per il monitoraggio e la configurazione di IBM SPSS Modeler Server. La console viene implementata come plug-in dell'applicazione Deployment Manager.
- **IBM SPSS Modeler - Guida CRISP-DM.** Guida passo a passo al data mining tramite la metodologia CRISP-DM con SPSS Modeler.
- **IBM SPSS Modeler Batch - Guida per l'utente.** Guida completa all'utilizzo di IBM SPSS Modeler in modalità batch, contenente dettagli per l'esecuzione della modalità batch e gli argomenti della riga di comando. Questa guida è disponibile solo in formato PDF.

Documentazione di SPSS Modeler Premium

La documentazione completa di SPSS Modeler Premium, escluse le istruzioni di installazione, è la seguente.

- **IBM SPSS Modeler Entity Analytics User Guide.** Contiene informazioni per l'utilizzo dell'analisi delle entità con SPSS Modeler; descrive l'installazione e la configurazione di repository, i nodi Entity Analytics e le attività amministrative.
- **IBM SPSS Modeler Social Network Analysis User Guide.** Guida che spiega come eseguire l'analisi dei social network con SPSS Modeler; comprende l'analisi di gruppo e l'analisi di diffusione.
- **SPSS Modeler Text Analytics - Guida per l'utente.** Contiene informazioni per l'utilizzo di analisi di testo con SPSS Modeler; descrive i nodi di text mining, il workbench interattivo, i modelli e altre risorse.

Esempi di applicazioni

Mentre gli strumenti per il data mining di SPSS Modeler consentono di risolvere un'ampia gamma di problemi a livello di business e organizzativo, gli esempi di applicazioni forniscono indicazioni mirate e sintetiche su specifici metodi e tecniche di modellazione. Gli insiemi di dati utilizzati negli esempi hanno dimensioni molto più limitate rispetto agli enormi archivi di dati gestiti da alcuni data miner, ma i concetti e i metodi coinvolti sono rapportabili alle applicazioni del mondo reale.

È possibile accedere agli esempi facendo clic su **Esempi di applicazioni** nel menu Aiuto di SPSS Modeler. I file di dati e i flussi di esempio sono installati nella cartella *Demos* nella directory di installazione del prodotto. Per ulteriori informazioni, consultare l'argomento "Cartella Demos".

Esempi di modellazione del database. Vedere gli esempi nella *IBM SPSS Modeler Guida al mining nel database*.

Esempi di script. Vedere gli esempi nella *IBM SPSS Modeler Guida per script e automazione*.

Cartella Demos

I file di dati e i flussi di esempio utilizzati negli esempi di applicazioni sono installati nella cartella *Demos* nella directory di installazione del prodotto. È possibile accedere a questa cartella anche dal gruppo di programmi IBM SPSS Modeler nel menu Start di Windows oppure facendo clic su *Demos* nell'elenco delle directory recenti nella finestra di dialogo Apri file.

Capitolo 2. Mining nel database

Panoramica sulla modellazione di database

IBM SPSS Modeler Server supporta l'integrazione con gli strumenti di data mining e di modellazione offerti dai fornitori di database, quali IBM Netezza, IBM DB2 InfoSphere Warehouse, Oracle Data Miner e Microsoft Analysis Services. È possibile creare, calcolare il punteggio e archiviare modelli in tutti i database dall'interno dell'applicazione IBM SPSS Modeler. Ciò consente di combinare le funzionalità analitiche e la semplicità d'uso di IBM SPSS Modeler con la potenza e le performance di un database, sfruttando gli algoritmi nativi del database distribuiti da questi fornitori. I modelli vengono creati all'interno del database e possono essere successivamente selezionati per calcolarne il punteggio attraverso l'interfaccia di IBM SPSS Modeler secondo la procedura standard; se necessario, ne può essere eseguita la distribuzione attraverso IBM SPSS Modeler Solution Publisher. Gli algoritmi supportati sono elencati nella palette Modelli in-database di IBM SPSS Modeler.

L'utilizzo di IBM SPSS Modeler per accedere ad algoritmi nativi di database assicura numerosi vantaggi:

- Gli algoritmi in-database sono spesso strettamente integrati con il server di database e possono offrire performance migliorate.
- I modelli creati e archiviati "in database" possono essere facilmente distribuiti e condivisi con qualsiasi applicazione che può accedere al database.

Generazione SQL. La modellazione nel database è distinta dalla generazione SQL altrimenti noto come "push back SQL". Questa funzione consente di generare istruzioni SQL per le operazioni native di IBM SPSS Modeler che possono essere rinviate "pushed back" (cioè eseguite nel) database per migliorare le prestazioni. Per esempio, i nodi Unione, Aggregazione e Seleziona generano tutti codice SQL che può essere rinviato al database per l'esecuzione. L'utilizzo della generazione SQL in combinazione con la modellazione nel database può generare flussi eseguibili dall'inizio alla fine nel database, con significativi miglioramenti a livello di prestazioni rispetto ai flussi eseguiti in IBM SPSS Modeler.

Nota: le funzionalità di modellazione nel database e ottimizzazione SQL richiedono che la connettività IBM SPSS Modeler Server venga abilitata sul computer IBM SPSS Modeler. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da IBM SPSS Modeler, e accedere a IBM SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu IBM SPSS Modeler.

Guida > **Informazioni su** > **Dettagli aggiuntivi**

Se la connettività è abilitata, l'opzione **Abilitazione server** viene visualizzata nella scheda Stato della licenza.

Per informazioni sugli algoritmi supportati, fare riferimento alle sezioni relative ai fornitori specifici riportate di seguito.

Requisiti necessari

Per eseguire la modellazione di database, occorre disporre di quanto elencato di seguito:

- Una connessione ODBC a un database appropriato, in cui siano installati i componenti analitici richiesti (Microsoft Analysis Services, Oracle Data Miner o IBM DB2 InfoSphere Warehouse).
- In IBM SPSS Modeler la modellazione di database deve essere attivata nella finestra di dialogo Applicazioni di supporto (**Strumenti > Applicazioni di supporto**).
- In IBM SPSS Modeler e in IBM SPSS Modeler Server (se utilizzato) le impostazioni **Genera SQL** e **Ottimizzazione SQL** devono essere attivate nella finestra di dialogo Opzioni utente. Tener presente che

l'ottimizzazione SQL non è strettamente richiesta per la modellazione del database da utilizzare, ma è vivamente consigliato per motivi legati alle prestazioni.

Nota: le funzionalità di modellazione nel database e ottimizzazione SQL richiedono che la connettività IBM SPSS Modeler Server venga abilitata sul computer IBM SPSS Modeler. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da IBM SPSS Modeler, e accedere a IBM SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu IBM SPSS Modeler.

Guida > **Informazioni su** > **Dettagli aggiuntivi**

Se la connettività è abilitata, l'opzione **Abilitazione server** viene visualizzata nella scheda Stato della licenza.

Per informazioni dettagliate, vedere le sezioni relative ai fornitori specifici riportate di seguito.

Creazione del modello

Il processo di creazione di modelli e di calcolo del relativo punteggio mediante algoritmi di database presenta molte analogie con altri tipi di data mining all'interno di IBM SPSS Modeler. Il processo generale di utilizzo di nodi e "nugget" di modelli è analogo a qualsiasi altro flusso quando si lavora in IBM SPSS Modeler. L'unica differenza è rappresentata dal fatto che l'elaborazione e la creazione del modello effettive sono rinviate al database.

Un flusso di modellazione del database è concettualmente identico ad altri flussi in IBM SPSS Modeler; tuttavia, questo flusso esegue tutte le operazioni in un database compreso , ad esempio, la creazione del modello utilizzando il nodo della struttura ad albero delle decisioni Microsoft. Quando si esegue il flusso, IBM SPSS Modeler fornisce al database le istruzioni necessarie per creare e archiviare il modello risultante e i dettagli vengono scaricati all'interno di IBM SPSS Modeler. L'esecuzione del database è indicata nel flusso dall'utilizzo di nodi con ombreggiatura viola.+

Data Preparation

Indipendentemente dal fatto che siano utilizzati o meno algoritmi nativi di database, le preparazioni dei dati devono sempre essere rinviate al database quando possibile per migliorare le prestazioni.

- Se i dati originali sono archiviati nel database, l'obiettivo è quello di mantenerli nel database assicurandosi che tutte le operazioni upstream necessarie possano essere convertite in SQL. Ciò impedisce che i dati siano scaricati in IBM SPSS Modeler—evitando colli di bottiglia che potrebbero annullare i vantaggi e consentendo all'intero flusso di essere eseguito nel database.
- Se i dati originali *non* sono archiviati nel database, sarà comunque possibile utilizzare la modellazione di database. In questo caso, la preparazione dei dati viene effettuata all'interno di IBM SPSS Modeler e l'insieme dei dati preparato viene automaticamente caricato nel database per la creazione del modello.

Calcolo del punteggio del modello

I modelli generati da IBM SPSS Modeler utilizzando il mining nel database sono diversi dai normali modelli di di IBM SPSS Modeler. Sebbene essi siano visualizzati nel Model manager come un modello come “nugget,” del modello generato, sono effettivamente modelli remoti memorizzati sul data mining o sul server di database remoto. Quelli visibili in IBM SPSS Modeler sono semplicemente dei riferimenti a tali modelli remoti. In altre parole, il modello IBM SPSS Modeler che viene visualizzato è un modello vuoto che contiene le informazioni come ad esempio il nome host del server di database, il nome del database e il nome del modello. Si tratta di una distinzione importante da comprendere per la visualizzazione e il calcolo del punteggio dei modelli creati utilizzando gli algoritmi nativi di database.

Una volta creato un nuovo modello, è possibile aggiungerlo al flusso per il calcolo del punteggio seguendo la prassi utilizzata per qualsiasi altro modello generato in IBM SPSS Modeler. Tutti i calcoli di punteggio vengono eseguiti all'interno del database, anche se le operazioni upstream vengono eseguite

altrove. (Le operazioni upstream possono essere rimandate al database per migliorare le performance, ma questo non è necessario perché avvenga il calcolo del punteggio.) Nella maggior parte dei casi, è anche possibile sfogliare il modello generato utilizzando il browser standard offerto dal fornitore di database.

Per sfogliare e calcolare i punteggi, è necessario disporre di una connessione live al server su cui vengono eseguiti Oracle Data Miner, IBM DB2 InfoSphere Warehouse oppure Microsoft Analysis Services.

Visualizzazione dei risultati e specifica delle impostazioni

Per visualizzare i risultati e specificare le impostazioni inerenti al calcolo del punteggio, fare doppio clic sul modello nell'area del flusso. In alternativa, è possibile fare clic con il pulsante destro del mouse sul modello e scegliere **Visualizza** o **Modifica**. Le impostazioni specifiche dipendono dal tipo di modello.

Esportazione e salvataggio di modelli di database

I modelli e i riepiloghi del database possono essere esportati dal visualizzatore modelli con la stessa procedura impiegata per altri modelli creati in IBM SPSS Modeler, utilizzando le opzioni disponibili nel menu File.

1. Dal menu File del visualizzatore modelli scegliere una qualsiasi delle seguenti opzioni:

- **Esporta testo** esporta il riepilogo di modello in un file di testo
- **Esporta HTML** esporta il riepilogo di modello in un file HTML
- **Esporta PMML** (supportata solo per i modelli IBM DB2 IM) esporta il modello come PMML (Predictive Model Markup Language), che può essere utilizzato con altri software compatibili con PMML.

Nota: è possibile salvare un modello generato anche scegliendo **Salva nodo** dal menu File.

Uniformità dei modelli

Per ogni modello di database generato, IBM SPSS Modeler archivia una descrizione della relativa struttura insieme a un riferimento al modello con lo stesso nome memorizzato nel database. Nella scheda Server di un modello generato viene visualizzata una chiave univoca generata specificamente per il modello in questione che corrisponde al modello effettivo nel database.

IBM SPSS Modeler utilizza queste chiavi generate casualmente per controllare l'uniformità dei modelli. La chiave viene archiviata nella descrizione del modello al momento della creazione. È consigliabile verificare la corrispondenza delle chiavi prima di eseguire un flusso di distribuzione.

1. Per verificare l'uniformità del modello archiviato nel database confrontando la relativa descrizione con la chiave casuale memorizzata da IBM SPSS Modeler, fare clic sul pulsante **Controlla**. Se non è possibile trovare il modello di database o la chiave non corrisponde, verrà segnalato un errore.

Visualizzazione ed esportazione di codice SQL generato

Prima di procedere all'esecuzione, è possibile visualizzare un anteprima del codice SQL, il che può essere molto utile ai fini del debug.

Capitolo 3. Modellazione di database con Microsoft Analysis Services

IBM SPSS Modeler e Microsoft Analysis Services

IBM SPSS Modeler supporta l'integrazione con Microsoft SQL Server Analysis Services. Questa funzionalità viene implementata sotto forma di nodi Modelli in IBM SPSS Modeler ed è disponibile nella palette Modelli in-database. Se la palette non è visibile, è possibile attivarla abilitando l'integrazione con MS Analysis Services, disponibile nella scheda Microsoft della finestra di dialogo Applicazioni di supporto. Consultare l'argomento "Attivazione dell'integrazione con Analysis Services" a pagina 13 per ulteriori informazioni.

IBM SPSS Modeler supporta l'integrazione con i seguenti algoritmi di Analysis Services:

- Strutture ad albero delle decisioni
- Raggruppamento tramite cluster
- Regole di associazione
- Naive Bayes
- Regressione lineare
- Rete neurale
- Regressione logistica
- serie temporali
- Cluster di sequenze

Nel seguente diagramma è illustrato il flusso di dati dal client verso il server nei casi in cui il mining nel database è gestito da IBM SPSS Modeler Server. La creazione del modello viene eseguita mediante Analysis Services e il modello risultante è archiviato dallo stesso strumento. Un riferimento a tale modello viene conservato nei flussi di IBM SPSS Modeler. Il modello viene quindi scaricato da Analysis Services su Microsoft SQL Server o IBM SPSS Modeler per il calcolo del punteggio.

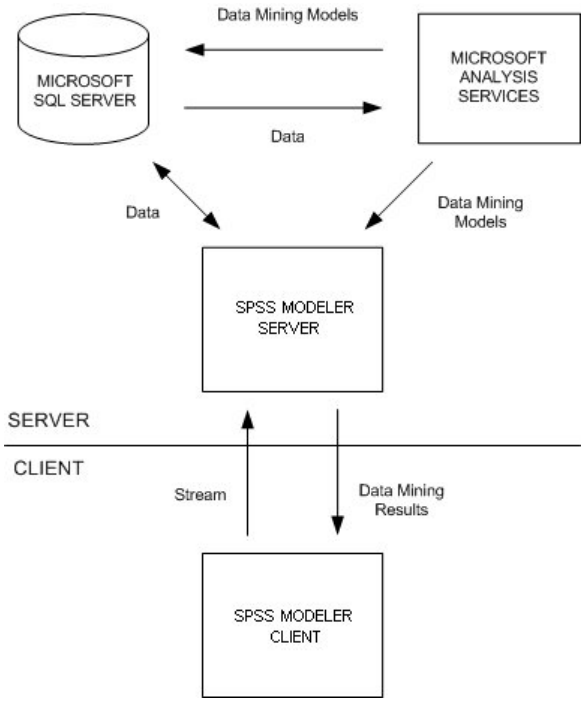

Figura 1. Flusso di dati tra IBM SPSS Modeler, Microsoft SQL Server e Microsoft Analysis Services durante la creazione del modello

Nota: Il IBM SPSS Modeler Server non è richiesto, anche se può essere utilizzato. Il client IBM SPSS Modeler è in grado di elaborare calcoli di mining nel database stesso.

Requisiti per l'integrazione con Microsoft Analysis Services

Di seguito sono riportati i prerequisiti richiesti per eseguire la modellazione nel database utilizzando gli algoritmi di Analysis Services con IBM SPSS Modeler. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di database.

- Esecuzione di IBM SPSS Modeler in un'installazione di IBM SPSS Modeler Server (modalità distribuita) su Windows. Le piattaforme UNIX non sono supportate in questa integrazione con Analysis Services.

  Importante: gli utenti IBM SPSS Modeler devono configurare la connessione ODBC utilizzando il driver SQL Native Client reso disponibile da Microsoft all'URL elencato in Requisiti *aggiuntivi di IBM SPSS Modeler Server*. *Il driver fornito con il IBM SPSS Data Access Pack (e generalmente consigliato per latri usi con IBM SPSS Modeler) non è consigliato per questo scopo.* È necessario configurare il driver per l'uso di SQL Server con l'opzione **Autenticazione integrata di Windows** attivata, poiché IBM SPSS Modeler non supporta l'autenticazione SQL Server. Per domande su come creare o impostare le autorizzazioni per le origini dati ODBC, contattare l'amministratore di database.
- È necessario aver installato sul computer SQL Server 2005 o 2008, sebbene non necessariamente sullo stesso host di IBM SPSS Modeler. Gli utenti di IBM SPSS Modeler devono disporre delle autorizzazioni richieste per leggere e scrivere dati nonché per creare ed eliminare tabelle e visualizzazioni.

  Nota: SQL Server Enterprise Edition è consigliato. La versione Enterprise Edition offre una flessibilità maggiore fornendo parametri avanzati che consentono di perfezionare i risultati degli algoritmi. La versione Standard Edition fornisce gli stessi parametri ma non consente agli utenti di modificare alcuni dei parametri avanzati.
- È necessario aver installato Microsoft SQL Server Analysis Services sullo stesso host di SQL Server.

Requisiti aggiuntivi di IBM SPSS Modeler Server

Per utilizzare gli algoritmi di Analysis Services con IBM SPSS Modeler Server, è necessario aver installato sull'host di IBM SPSS Modeler Server i seguenti componenti

Nota: se SQL Server è installato sullo stesso host di IBM SPSS Modeler Server, questi componenti saranno già disponibili.

- Microsoft .NET Framework Redistributable Package versione 2.0 (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Provider OLE DB Microsoft SQL Server 2008 Analysis Services 10.0 (avere cura di selezionare la variante corretta per il proprio sistema operativo)
- Microsoft SQL Server 2008 Native Client (avere cura di selezionare la variante corretta per il proprio sistema operativo)

Per scaricare questi componenti, accedere a *www.microsoft.com/downloads*, cercare **.NET Framework** o (per tutti gli altri componenti) **SQL Server Feature Pack** e selezionare il pacchetto più recente per la propria versione di SQL Server.

L'esecuzione di tali componenti potrebbe richiedere l'installazione di altri pacchetti, che dovrebbero essere disponibili anch'essi nell'area Download del sito Web di Microsoft.

Requisiti aggiuntivi di IBM SPSS Modeler

Per utilizzare gli algoritmi di Analysis Services con IBM SPSS Modeler, è necessario che siano installati gli stessi componenti riportati in precedenza, con l'aggiunta dei seguenti sul client:

- Microsoft SQL Server 2008 Datamining Viewer Controls (avere cura di selezionare la variante corretta per il proprio sistema operativo), che richiede inoltre:
- Microsoft ADOMD.NET

Per scaricare questi componenti, accedere a *www.microsoft.com/downloads*, cercare **SQL Server Feature Pack** e selezionare il pacchetto più recente per la propria versione di SQL Server.

Nota: le funzionalità di modellazione nel database e ottimizzazione SQL richiedono che la connettività IBM SPSS Modeler Server venga abilitata sul computer IBM SPSS Modeler. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da IBM SPSS Modeler, e accedere a IBM SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu IBM SPSS Modeler.

Guida > **Informazioni su** > **Dettagli aggiuntivi**

Se la connettività è abilitata, l'opzione **Abilitazione server** viene visualizzata nella scheda Stato della licenza.

Attivazione dell'integrazione con Analysis Services

Per attivare l'integrazione di IBM SPSS Modeler con Analysis Services, è necessario configurare SQL Server e Analysis Services e creare un'origine ODBC, quindi attivare l'integrazione nella finestra di dialogo di Applicazioni di supporto IBM SPSS Modeler e, infine, attivare la generazione e l'ottimizzazione SQL.

Nota: Microsoft SQL Server e Microsoft Analysis Services devono essere disponibili. Consultare l'argomento “Requisiti per l'integrazione con Microsoft Analysis Services” a pagina 12 per ulteriori informazioni.

Configurazione di SQL Server

Configurare SQL Server in modo da consentire che il calcolo del punteggio sia eseguito all'interno del database.

1. Creare la seguente chiave del Registro di sistema sul computer host SQL Server:
   `HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP`
2. Aggiungere quindi alla chiave il seguente valore `DWORD`:
   `AllowInProcess 1`
3. Riavviare SQL Server dopo aver apportato la modifica.

Configurazione di Analysis Services

Prima che IBM SPSS Modeler possa comunicare con Analysis Services, è necessario configurare manualmente due impostazioni nella finestra di dialogo Proprietà di Analysis Server:

1. Accedere ad Analysis Server tramite MS SQL Server Management Studio.
2. Accedere alla finestra di dialogo Proprietà facendo clic con il pulsante destro sul nome del server e scegliendo **Proprietà**.
3. Selezionare la casella di controllo **Mostra proprietà (tutte) avanzate**.
4. Modificare le seguenti proprietà:

- Modificare il valore di `DataMining\AllowAdHocOpenRowsetQueries` su `True` (il valore di default è `False`).
- Modificare il valore di `DataMining\AllowProvidersInOpenRowset` con `[all]` (non esiste un valore di default).

Creazione di un DSN ODBC per SQL Server

Per leggere o scrivere su un database, occorre che un'origine dati ODBC sia installata e configurata per il database in questione, con le relative autorizzazioni di lettura e scrittura. È necessario disporre del driver ODBC Microsoft SQL Native Client che viene installato automaticamente con SQL Server. *Il driver fornito con il IBM SPSS Data Access Pack (e generalmente consigliato per altri usi con IBM SPSS Modeler) non è consigliato per questo scopo.* Se IBM SPSS Modeler e SQL Server risiedono su host diversi, è possibile scaricare il driver ODBC Microsoft SQL Native Client. Consultare l'argomento "Requisiti per l'integrazione con Microsoft Analysis Services" a pagina 12 for more information.

Per domande su come creare o impostare le autorizzazioni per le origini dati ODBC, contattare l'amministratore di database.

1. Con il driver ODBC Microsoft SQL Native Client, creare un DSN ODBC che punta al database SQL Server utilizzato nel processo di data mining. Per le restanti impostazioni del driver, è necessario utilizzare le impostazioni di default.
2. Assicurarsi che per il DSN sia selezionata l'**autenticazione integrata di Windows**.

- Se IBM SPSS Modeler e IBM SPSS Modeler Server sono in esecuzione su differenti host, creare lo stesso DSN ODBC su ognuno degli host. Assicurarsi di utilizzare lo stesso nome DSN su ogni host.

Attivazione dell'integrazione di Analysis Services in IBM SPSS Modeler

Per consentire a IBM SPSS Modeler di utilizzare Analysis Services, è innanzitutto necessario specificare le informazioni sul server nella finestra di dialogo Applicazioni di supporto.

1. Dai menu di IBM SPSS Modeler scegliere:
   Strumenti > **Opzioni** > **Applicazione di supporto**
2. Fare clic sulla scheda **Microsoft**.

- **Abilita Microsoft Analysis Services Integration.** Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra IBM SPSS Modeler e aggiunge i nodi degli algoritmi di Analysis Services.
- **Host Analysis Server.** Specificare il nome del computer su cui è in esecuzione Analysis Services.

- **Database Analysis Server.** Selezionare il database desiderato facendo clic sul pulsante con i puntini di sospensione (...) che consente di aprire una sottofinestra di dialogo in cui è possibile scegliere tra i database disponibili. L'elenco contiene i database disponibili per il server Analysis specificato. Poiché Microsoft Analysis Services archivia i modelli di data mining all'interno di database denominati, è necessario selezionare il database appropriato in cui vengono archiviati i modelli Microsoft creati da IBM SPSS Modeler.
- **Connessione SQL Server.** Specificare le informazioni DSN utilizzate dal database SQL Server per archiviare i dati passati ad Analysis Server. Scegliere l'origine dati ODBC che verrà utilizzata per fornire i dati necessari per la creazione di modelli di data mining Analysis Services. Se si creano modelli Analysis Services a partire da dati forniti all'interno di file flat o origini dati ODBC, i dati verranno automaticamente caricati in una tabella temporanea creata nel database SQL Server al quale punta l'origine dati ODBC.
- **Avvisa quando si sovrascrive un modello di data mining.** Selezionare questa opzione per assicurarsi che i modelli archiviati nel database non vengano sovrascritti da IBM SPSS Modeler senza preavviso.

Nota: le impostazioni effettuate nella finestra di dialogo Applicazione di supporto possono essere sovrascritte all'interno dei vari nodi di Analysis Services.

Attivazione di generazione e ottimizzazione SQL

1. Dal menu IBM SPSS Modeler scegliere:

   Strumenti > **Proprietà flusso** > **Opzioni**
2. Fare clic sull'opzione **Ottimizzazione** nel riquadro di spostamento.
3. Confermare che l'opzione **Genera SQL** è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
4. Selezionare **Ottimizza generazione SQL** e **Ottimizza altre esecuzioni** (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Creazioni di modelli con Analysis Services

La creazione del modello di Analysis Services richiede che l'insieme di dati addestramento sia posizionato in una tabella o visualizzazione all'interno del database SQL Server. Se i dati non sono ubicati in SQL Server o devono essere elaborati in IBM SPSS Modeler come parte del processo di preparazione dei dati che non è possibile eseguire in SQL Server, tali dati vengono automaticamente caricati in una tabella temporanea di SQL Server prima della creazione del modello.

Gestione di modelli di Analysis Services

La creazione di un modello di Analysis Services tramite IBM SPSS Modeler comporta la creazione di un modello in IBM SPSS Modeler e la creazione o la sostituzione di un modello nel database SQL Server. Il modello di IBM SPSS Modeler fa riferimento al contenuto di un modello di database archiviato in un server di database. IBM SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello SQL Server che nel modello di IBM SPSS Modeler.

Il nodo Modelli **Struttura ad albero delle decisioni MS** è utilizzato nella modellazione predittiva di attributi sia categoriali che continui. Per gli attributi categoriali, il nodo esegue previsioni in base alle relazioni tra le colonne di input in un insieme di dati. Per esempio, in uno scenario per prevedere quali clienti è probabile che acquistino una bicicletta, se nove su dieci clienti più giovani acquistano una bicicletta, ma solo due su dieci clienti più anziani la acquistano, il nodo desume che l'età sia un buon predittore dell'acquisto di biciclette. La struttura ad albero delle decisioni esegue previsioni in base a questa tendenza verso un particolare risultato. Per gli attributi continui, l'algoritmo utilizza la regressione lineare per stabilire dove la struttura ad albero delle decisioni si suddivide. Se più di una colonna è impostata come prevedibile, o se i dati di input contengono una tabella nidificata che è impostata come prevedibile, il nodo genera una struttura ad albero delle decisioni separata per ogni colonna prevedibile.

Il nodo Modelli **Raggruppamento cluster MS** utilizza tecniche iterative per raggruppare i casi di un insieme di dati in cluster contenenti caratteristiche simili. Questi raggruppamenti sono utili per l'esplorazione dei dati, l'individuazione di anomalie nei dati e la creazione di previsioni. I modelli di raggruppamento tramite cluster individuano le relazioni di un insieme di dati che non potrebbero essere derivate logicamente dall'osservazione casuale. Per esempio, è possibile comprendere logicamente che le persone che si recano al lavoro in bicicletta in genere non abitano molto distante dal posto di lavoro. Tuttavia, l'algoritmo è in grado di trovare altre caratteristiche relative ai pendolari della bicicletta che non sono così ovvie. Il nodo di raggruppamento cluster differisce dagli altri nodi di data mining in quanto non è specificato alcun campo obiettivo. Il nodo di raggruppamento cluster addestra il modello partendo strettamente dalla relazione esistente nei dati e dai cluster identificati dal nodo.

Il nodo Modelli **Regole di associazione MS** è utile per i motori di raccomandazioni. Un motore di raccomandazioni consiglia i prodotti ai clienti in base agli elementi già acquistati o per i quali hanno mostrato un interesse. I modelli di associazione vengono costruiti sulla base di insiemi di dati che contengono identificatori sia per i singoli casi che per gli elementi contenuti nei casi. Un gruppo di elementi di un caso viene definito **insieme di elementi**. Un modello di associazione è costituito da una serie di insiemi di elementi e dalle regole che descrivono come questi elementi sono raggruppati all'interno dei casi. Le regole individuate dall'algoritmo possono essere utilizzate per prevedere i probabili acquisti futuri di un cliente, in base agli elementi già presenti nel suo carrello.

Il nodo Modelli **Naive Bayes MS** calcola la probabilità condizionale tra i campi obiettivo e predittore e presume che le colonne siano indipendenti. Il modello è definito naive perché considera le variabili di previsione proposte come indipendenti l'una dall'altra. Questo metodo è meno intenso dal punto di vista computazionale rispetto agli altri algoritmi Analysis Services e pertanto è utile per scoprire rapidamente le relazioni durante le fasi preliminari di modellazione. Questo nodo può essere utile per effettuare esplorazioni iniziali dei dati e successivamente applicare i risultati per creare modelli aggiuntivi con altri nodi che possono richiedere un tempo di calcolo più lungo ma fornire risultati più precisi.

Il nodo di modellazione **Regressione lineare MS** rappresenta una variazione del nodo Strutture ad albero delle decisioni in cui il parametro `MINIMUM_LEAF_CASES` è impostato per essere superiore o uguale al numero totale di casi nel dataset che il nodo utilizza per addestrare il modello di mining. Con il suddetto parametro impostato in questo modo, il nodo non creerà mai una suddivisione e verrà pertanto eseguita una regressione lineare.

Il nodo Modelli **Rete neurale MS** è simile al nodo Struttura ad albero delle decisioni MS, poiché calcola le probabilità per ogni possibile stato dell'attributo di input quando viene fornito ogni stato dell'attributo prevedibile. È quindi possibile utilizzare queste probabilità in un momento successivo per prevedere un risultato dell'attributo previsto, in base agli attributi di input.

Il nodo di modellazione **Regressione logistica MS** rappresenta una variazione del nodo rete neurale MS in cui il parametro HIDDEN_NODE_RATIO è impostato su 0. Questa impostazione crea un modello di rete neurale che non contiene un livello nascosto e quindi è equivalente alla regressione logistica.

Il nodo Modelli **Serie temporali MS** prevede degli algoritmi di regressione ottimizzati per la previsione di valori continui nel tempo, per esempio le vendite di un prodotto. A differenza di altri algoritmi Microsoft (quali le strutture ad albero delle decisioni), un modello di serie temporali non richiede colonne aggiuntive di nuove informazioni come input per prevedere una tendenza. I modelli di serie temporali possono infatti prevedere le tendenze solo in base all'insieme di dati originale utilizzato per creare il modello. È possibile anche aggiungere nuovi dati al modello quando si effettua una previsione e incorporare automaticamente i nuovi dati nell'analisi della tendenza. Consultare l'argomento "Nodo Serie temporali MS" a pagina 19 per ulteriori informazioni.

Il nodo Modelli **Cluster di sequenze MS** identifica le sequenze ordinate presenti nei dati e combina i risultati di questa analisi con le tecniche di raggruppamento tramite cluster per generare cluster basati sulle sequenze e su altri attributi. Consultare l'argomento "Nodo Cluster di sequenze MS" a pagina 20 per ulteriori informazioni.

È possibile accedere a ogni nodo dalla palette Modelli database nella parte inferiore della finestra di IBM SPSS Modeler.

Impostazioni comuni a tutti i nodi degli algoritmi

Le seguenti impostazioni sono valide per tutti gli algoritmi di Analysis Services.

Opzioni server

Nella scheda Server è possibile configurare l'host e il database di Analysis Server nonché l'origine dati SQL Server. Le opzioni specificate in questa posizione sovrascrivono quelle selezionate nella scheda Microsoft all'interno della finestra di dialogo Applicazioni di supporto. Consultare l'argomento "Attivazione dell'integrazione con Analysis Services" a pagina 13 per ulteriori informazioni.

Nota: è disponibile anche una variante di questa scheda durante il calcolo del punteggio dei modelli Analysis Services. Consultare l'argomento "Scheda Server del nugget del modello Analysis Services" a pagina 22 per ulteriori informazioni.

Opzioni modello

Per poter creare il modello di base, occorre specificare preliminarmente una serie di opzioni nella scheda Modello. Il metodo per il calcolo del punteggio e altre opzioni avanzate sono accessibili nella scheda Livello avanzato.

Di seguito sono riportate le principali opzioni di modellazione disponibili:

Nome modello. Specifica il nome assegnato al modello creato quando viene eseguito il nodo.

- **Automatico.** Genera il nome del modello automaticamente in base ai nomi dei campi ID e Obiettivo oppure il nome del tipo di modello nei casi in cui l'obiettivo non viene specificato (come i modelli cluster).
- **Personalizzato.** Consente di specificare un nome personalizzato per il modello creato.

Utilizza dati partizionati Suddivide i dati in sottoinsiemi separati, o campioni, per le fasi di addestramento, test e convalida in base al campo partizione corrente. Utilizzando un campione per creare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere una valida indicazione del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, più simili ai dati correnti. Se non viene specificato un campo partizione nel flusso, tale opzione viene ignorata.

Con drill-through. Se visualizzata, questa opzione consente di interrogare il modello per ottenere informazioni sui casi compresi nel modello.

Campo univoco. Dall'elenco a discesa,selezionare un campo che identifichi in modo univoco ogni caso. In genere, si tratta di un campo ID, per esempio **IDCliente**.

Opzioni avanzate Struttura ad albero delle decisioni MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Raggruppamento cluster MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Naive Bayes MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Regressione lineare MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Rete neurale MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Opzioni avanzate Regressione logistica MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Nodo Regole di associazione MS

Il nodo Modelli Regole di associazione Microsoft è utile per i motori di raccomandazioni. Un motore di raccomandazioni consiglia i prodotti ai clienti in base agli elementi già acquistati o per i quali hanno mostrato un interesse. I modelli di associazione vengono costruiti sulla base di insiemi di dati che contengono identificatori sia per i singoli casi che per gli elementi contenuti nei casi. Un gruppo di elementi di un caso viene definito **insieme di elementi**.

Un modello di associazione è costituito da una serie di insiemi di elementi e dalle regole che descrivono come questi elementi sono raggruppati all'interno dei casi. Le regole individuate dall'algoritmo possono essere utilizzate per prevedere i probabili acquisti futuri di un cliente, in base agli elementi già presenti nel suo carrello.

Per i dati in formato tabulare, l'algoritmo crea punteggi che rappresentano la probabilità ($MP-*campo*) per ogni raccomandazione generata ($M-*campo*). Per i dati in formato transazionale vengono creati punteggi per supporto ($MS-*campo*), probabilità ($MP-*campo*) e probabilità regolata ($MAP-*campo*) per ogni raccomandazione generata ($M-*campo*).

Requisiti

I requisiti di un modello di associazione transazionale sono i seguenti:

- **Campo univoco.** Un modello di regole di associazione richiede una chiave che identifichi i record in modo univoco.
- **Campo ID.** Quando si crea un modello Regole di associazione MS con dati in formato transazionale è necessario un campo ID che identifichi le singole transazioni. I campi ID si possono impostare sullo stesso valore del campo unico.
- **Almeno un campo di input.** L'algoritmo della regola di associazione richiede almeno un campo di input.
- **Campo obiettivo.** Quando si crea un modello Regole di associazione MS con dati transazionali, il campo obiettivo deve essere il campo della transazione, per esempio i prodotti acquistati da un utente.

Opzioni avanzate Regole di associazione MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Nodo Serie temporali MS

Il nodo Modelli Serie temporali MS supporta due tipi di previsioni:

- futura
- storica

Le **previsioni future** stimano i valori del campo obiettivo per un numero specificato di periodi di tempo successivi alla fine dei dati cronologici e vengono sempre eseguite. Le **previsioni storiche** sono valori stimati del campo obiettivo relativi a un numero specifico di periodi di tempo i cui valori sono effettivamente presenti nei dati cronologici. Le previsioni storiche si possono utilizzare per valutare la qualità del modello confrontando i dati storici effettivi con quelli previsti. Il valore del punto di partenza delle previsioni determina l'esecuzione o meno delle previsioni storiche.

A differenza del nodo Serie temporali IBM SPSS Modeler, il nodo Serie temporali MS non deve essere preceduto da un nodo Intervalli di tempo. Un'ulteriore differenza è il fatto che per default i punteggi sono calcolati solo per le righe previste e non per tutte le righe di dati storici della serie temporale.

Requisiti

I requisiti di un modello Serie temporali MS sono i seguenti:

- **Singolo campo tempo chiave.** Ogni modello deve contenere un campo numerico o data utilizzato come serie del caso, che definisce le sezioni temporali utilizzate dal modello. Il tipo di dati del campo tempo chiave può essere data/ora o numerico. Tuttavia, il campo deve contenere valori continui che devono inoltre essere univoci per ogni serie.
- **Un solo campo obiettivo.** È possibile specificare un solo campo obiettivo in ogni modello. Il tipo di dati del campo obiettivo deve avere valori continui. Per esempio, è possibile prevedere come variano nel tempo attributi numerici quali il reddito, le vendite o la temperatura. Non è invece consentito l'utilizzo di un campo contenente valori categoriali quali lo stato di acquisto o il livello di istruzione come campo obiettivo.
- **Almeno un campo di input.** L'algoritmo Serie temporali MS richiede almeno un campo di input. Il tipo di dati del campo di input deve avere valori continui. I campi di input non continui vengono ignorati al momento della creazione del modello.
- **Il dataset deve essere ordinato.** Il dataset di input deve essere ordinato (sul campo tempo chiave), altrimenti la creazione del modello si interromperà con un errore.

Opzioni del modello di serie temporali MS

Nome modello. Specifica il nome assegnato al modello creato quando viene eseguito il nodo.

- **Automatico.** Genera il nome del modello automaticamente in base ai nomi dei campi ID e Obiettivo oppure il nome del tipo di modello nei casi in cui l'obiettivo non viene specificato (come i modelli cluster).
- **Personalizzato.** Consente di specificare un nome personalizzato per il modello creato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Con drill-through. Se visualizzata, questa opzione consente di interrogare il modello per ottenere informazioni sui casi compresi nel modello.

Campo univoco. Dall'elenco a discesa, selezionare il campo tempo chiave utilizzato per creare il modello di serie temporali.

Opzioni avanzate Serie temporali MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

In caso di previsioni storiche, il numero di fasi storiche che è possibile includere nel risultato del calcolo del punteggio viene deciso dal valore di (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP). Per default, questo limite è 10, ovvero vengono effettuate solo 10 previsioni storiche. In questo caso, per esempio, si verifica un errore se si inserisce una valore inferiore a -10 per **Previsione storica** nella scheda Impostazioni del nugget del modello (vedere "Scheda Impostazioni del nugget del modello Serie temporali MS" a pagina 23). Per visualizzare un numero maggiore di previsioni storiche è possibile aumentare il valore di HISTORIC_MODEL_COUNT o HISTORIC_MODEL_GAP, ma questo aumenterà il tempo impiegato per la creazione del modello.

Opzioni di impostazione Serie temporali MS

Inizia stima. Specificare il periodo da cui si desidera iniziare le previsioni.

- **Inizia da: nuova previsione.** Il periodo da cui si desidera far iniziare le previsioni future, espresso come offset dall'ultimo periodo dei dati cronologici di cui si dispone. Per esempio, se i dati cronologici terminano il 12/99 e si desidera iniziare le previsioni il 01/00, utilizzare il valore 1; se invece si desidera iniziare le previsioni il 03/00, utilizzare il valore 3.
- **Inizia da: previsione storica** Il periodo da cui si desidera far iniziare le previsioni storiche, espresso come offset negativo dall'ultimo periodo dei dati cronologici di cui si dispone. Per esempio, se i dati cronologici terminano il 12/99 e si desidera fare previsioni storiche per gli ultimi cinque periodi di tempo dei propri dati, utilizzare il valore -5.

Termina stima. Specificare il periodo in cui si desidera terminare le previsioni.

- **Termina fase di previsione.** Il periodo in cui si desidera terminare le previsioni, espresso come offset dall'ultimo periodo dei dati cronologici di cui si dispone. Per esempio, se i dati cronologici terminano il 12/99 e si desidera che le previsioni terminino il 6/00, utilizzare il valore 6. Per le previsioni future, il valore deve sempre essere superiore o uguale al valore **Avvia da**.

Nodo Cluster di sequenze MS

Il nodo Cluster di sequenze MS utilizza un algoritmo di analisi delle sequenze che esplora i dati contenenti eventi collegabili seguendo dei percorsi, o *sequenze*. Alcuni esempi potrebbero essere i percorsi di navigazione creati dagli utenti che consultano un sito Web, o l'ordine con cui un cliente aggiunge gli articoli al carrello degli acquisti di un rivenditore online. L'algoritmo individua le sequenze più comuni raggruppando o *inserendo in cluster* le sequenze identiche.

Requisiti

I requisiti di un modello Cluster di sequenze Microsoft sono i seguenti:

- **Campo ID.** L'algoritmo di Cluster di sequenze Microsoft richiede che le informazioni delle sequenze siano archiviate in formato transazionale. A tale fine è necessario disporre di un campo ID che identifichi le singole transazioni.
- **Almeno un campo di input.** L'algoritmo richiede almeno un campo di input.
- **Campo Sequenza.** L'algoritmo richiede anche un campo di identificazione della sequenza che deve avere un livello di misurazione Continuo. Per esempio si può utilizzare un identificativo di pagina Web, un numero intero o una stringa di testo, purché il campo identifichi gli eventi di una sequenza. Per ogni sequenza è consentito un solo identificativo e per ogni modello è consentito un solo tipo di sequenza. Il campo Sequenza deve essere diverso dal campo ID e Unico.
- **Campo Obiettivo.** Quando si crea un modello di raggruppamento sequenze è necessario un campo obiettivo.
- **Campo univoco.** Un modello di cluster di sequenze richiede un campo chiave che identifichi i record in modo univoco. Il campo Unico si può impostare sullo stesso valore del campo ID.

Opzioni dei campi Cluster di sequenze MS

In tutti i nodi Modelli è disponibile una scheda Campi nella quale è possibile specificare i campi da utilizzare per la creazione del modello.

Per poter generare un modello di cluster di sequenze, è necessario prima specificare i campi da utilizzare come obiettivi e come input. Si noti che per il nodo Cluster di sequenze MS non è possibile utilizzare le informazioni dei campi di un nodo Tipo upstream: le impostazioni dei campi devono essere definite qui.

ID. Selezionare un campo ID dall'elenco. Come campo ID è possibile utilizzare campi numerici o simbolici. Ogni valore univoco di questo campo deve indicare una specifica unità di analisi. In un'applicazione per analisi market basket, per esempio, ciascun ID potrebbe rappresentare un singolo cliente. Per un'applicazione per analisi di registri Web, ciascun ID potrebbe rappresentare un computer (per indirizzo IP) o un utente (per dati di login).

Input. Selezionare il campo o i campi di input per il modello. Questi sono i campi che contengono gli eventi rilevanti nella creazione di modelli di sequenza.

Sequenza. Scegliere dall'elenco un campo da utilizzare come campo identificativo della sequenza. Per esempio si può utilizzare un identificativo di pagina Web, un numero intero o una stringa di testo, purché il campo identifichi gli eventi di una sequenza. Per ogni sequenza è consentito un solo identificativo e per ogni modello è consentito un solo tipo di sequenza. Il campo Sequenza deve essere diverso dal campo ID (specificato in questa scheda) e dal campo Unico (specificato nella scheda Modello).

Obiettivo. Scegliere un campo da utilizzare come campo obiettivo, cioè il campo di cui si sta cercando di prevedere il valore in base ai dati della sequenza.

Opzioni avanzate Cluster di sequenze MS

Le opzioni disponibili nella scheda Livello avanzato possono variare a seconda della struttura del flusso selezionato. Per ulteriori dettagli sulle opzioni di livello avanzato disponibili per il nodo del modello Analysis Services selezionato, fare riferimento alla guida a livello di campo dell'interfaccia utente.

Calcolo del punteggio per i modelli di Analysis Services

Il calcolo del punteggio del modello avviene in SQL Server ed è eseguito da Analysis Services. Potrebbe essere necessario caricare il dataset in una tabella temporanea se i dati vengono originati all'interno di IBM SPSS Modeler o è necessario prepararli all'interno di IBM SPSS Modeler. I modelli che l'utente crea da IBM SPSS Modeler mediante il mining nel database rappresentano in effetti modelli remoti

memorizzati nel server di database o di data mining. Si tratta di una distinzione importante da comprendere per la visualizzazione e il calcolo del punteggio dei modelli creati utilizzando gli algoritmi di Microsoft Analysis Services.

In IBM SPSS Modeler viene generalmente fornita una sola previsione con la probabilità o la confidenza associata.

Per gli di calcolo del punteggio del modello, consultare “Esempi di mining con Analysis Services” a pagina 24.

Impostazioni comuni a tutti i modelli di Analysis Services

Le seguenti impostazioni sono valide per tutti i modelli di Analysis Services.

Scheda Server del nugget del modello Analysis Services

La scheda Server consente di specificare le connessioni per il mining nel database. La scheda fornisce anche la chiave di modello univoca. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di IBM SPSS Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Nella scheda Server è possibile configurare l'host e il database di Analysis Server nonché l'origine dati SQL Server per le operazioni di calcolo del punteggio. Le opzioni specificate all'interno di questa scheda sovrascrivono quelle selezionate nelle finestre di dialogo Applicazioni di supporto o Creazione modello di IBM SPSS Modeler. Consultare l'argomento “Attivazione dell'integrazione con Analysis Services” a pagina 13 per ulteriori informazioni.

GUID modello. La chiave di modello viene visualizzata qui. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di IBM SPSS Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Controllo. Fare clic su questo pulsante per confrontare la chiave qui visualizzata con quella all'interno del modello archiviato nel database di Analysis Services. Ciò consente di verificare se il modello è ancora presente in Analysis Server e se la relativa struttura non è stata modificata.

Nota: Il pulsante Controllo è disponibile solo per i modelli aggiunti all'area del flusso nella preparazione per il calcolo del punteggio. Qualora il controllo abbia esito negativo, verificare se il modello è stato eliminato o sostituito da un modello diverso sul server.

Vista. Fare clic per ottenere una vista grafica del modello di struttura ad albero delle decisioni. La scheda Visualizzatore delle strutture ad albero delle decisioni è condivisa da altri algoritmi per strutture ad albero delle decisioni disponibili in IBM SPSS Modeler e la funzionalità è identica.

Scheda Riepilogo del nugget del modello Analysis Services

La scheda Riepilogo di un nugget del modello visualizza le informazioni sul modello stesso (*Analisi*), i campi utilizzati nel modello (*Campi*), le impostazioni utilizzate durante la creazione del modello (*Impostazioni di creazione*) e l'addestramento del modello (*Riepilogo addestramento*).

Quando si sfoglia per la prima volta il nodo, i risultati della scheda Riepilogo sono compressi. Per visualizzare i risultati, utilizzare il controllo dell'espansore a sinistra di un elemento se si desidera visualizzare solo i risultati di tale elemento oppure fare clic sul pulsante **Espandi tutto** se si desidera visualizzare tutti i risultati. Per nascondere i risultati, utilizzare il controllo dell'espansore di un elemento se si desidera nascondere solo i risultati di tale elemento oppure fare clic sul pulsante **Comprimi tutto** se si desidera nascondere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico. Se è stato eseguito un nodo Analisi collegato a questo nugget del modello, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi.

Campi. Elenca i campi utilizzati come obiettivi e come input nella creazione del modello.

Impostazioni di creazione. Contiene informazioni sulle impostazioni utilizzate nella creazione del modello.

Riepilogo addestramento. Mostra il tipo di modello, il flusso utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Nugget del modello Serie temporali MS

Il modello Serie temporali MS genera punteggi solo per i periodi di tempo previsti, non per i dati cronologici.

La seguente tabella mostra i campi che vengono aggiunti al modello.

Tabella 1. Campi aggiunti al modello

Nome campo	Descrizione
$M-*campo*	Valore previsto del *campo*.
$Var-*campo*	Varianza calcolata del *campo*
$Stdev-*campo*	Deviazione standard del *campo*

Scheda Server del nugget del modello Serie temporali MS

La scheda Server consente di specificare le connessioni per il mining nel database. La scheda fornisce anche la chiave di modello univoca. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di IBM SPSS Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Nella scheda Server è possibile configurare l'host e il database di Analysis Server nonché l'origine dati SQL Server per le operazioni di calcolo del punteggio. Le opzioni specificate all'interno di questa scheda sovrascrivono quelle selezionate nelle finestre di dialogo Applicazioni di supporto o Creazione modello di IBM SPSS Modeler. Consultare l'argomento “Attivazione dell'integrazione con Analysis Services” a pagina 13 per ulteriori informazioni.

GUID modello. La chiave di modello viene visualizzata qui. Tale chiave è generata in modo casuale quando il modello viene creato e archiviato sia all'interno del modello di IBM SPSS Modeler che nella descrizione dell'oggetto modello memorizzata nel database di Analysis Services.

Controllo. Fare clic su questo pulsante per confrontare la chiave qui visualizzata con quella all'interno del modello archiviato nel database di Analysis Services. Ciò consente di verificare se il modello è ancora presente in Analysis Server e se la relativa struttura non è stata modificata.

Nota: Il pulsante Controllo è disponibile solo per i modelli aggiunti all'area del flusso nella preparazione per il calcolo del punteggio. Qualora il controllo abbia esito negativo, verificare se il modello è stato eliminato o sostituito da un modello diverso sul server.

Vista. Fare clic per ottenere una vista grafica del modello di serie temporali. Analysis Services mostra il modello completo sotto forma di struttura ad albero. È possibile anche visualizzare un grafico che mostra il valore storico del campo obiettivo nel tempo, unitamente ai valori futuri previsti.

Per ulteriori informazioni, vedere la descrizione del visualizzatore di serie temporali nella libreria MSDN all'indirizzo http://msdn.microsoft.com/en-us/library/ms175331.aspx.

Scheda Impostazioni del nugget del modello Serie temporali MS

Inizia stima. Specificare il periodo da cui si desidera iniziare le previsioni.

- **Inizia da: nuova previsione.** Il periodo da cui si desidera far iniziare le previsioni future, espresso come offset dall'ultimo periodo dei dati cronologici di cui si dispone. Per esempio, se i dati cronologici terminano il 12/99 e si desidera iniziare le previsioni il 01/00, utilizzare il valore 1; se invece si desidera iniziare le previsioni il 03/00, utilizzare il valore 3.
- **Inizia da: previsione storica** Il periodo da cui si desidera far iniziare le previsioni storiche, espresso come offset negativo dall'ultimo periodo dei dati cronologici di cui si dispone. Per esempio, se i dati cronologici terminano il 12/99 e si desidera fare previsioni storiche per gli ultimi cinque periodi di tempo dei propri dati, utilizzare il valore -5.

Termina stima. Specificare il periodo in cui si desidera terminare le previsioni.

- **Termina fase di previsione.** Il periodo in cui si desidera terminare le previsioni, espresso come offset dall'ultimo periodo dei dati cronologici di cui si dispone. Per esempio, se i dati cronologici terminano il 12/99 e si desidera che le previsioni terminino il 6/00, utilizzare il valore 6. Per le previsioni future, il valore deve sempre essere superiore o uguale al valore **Avvia da**.

Nugget del modello Cluster di sequenze MS

La seguente tabella mostra i campi che vengono aggiunti al modello di Cluster di sequenze MS e dove *campo* è lo stesso nome del campo obiettivo.

Tabella 2. Campi aggiunti al modello

Nome campo	Descrizione
$MC-*campo*	Previsione del cluster a cui appartiene la sequenza.
$MCP-*campo*	Probabilità che la sequenza appartenga al cluster previsto.
$MS-*campo*	Valore previsto del *campo*.
$MSP-*campo*	Probabilità che il valore $MS-*campo* sia corretto.

Esportazione di modelli e generazione di nodi

È possibile esportare il riepilogo e la struttura di un modello in file formato testo e HTML, nonché generare i nodi Seleziona e Filtro appropriati laddove necessario.

Analogamente ad altri nugget del modello in IBM SPSS Modeler, i nugget del modello di Microsoft Analysis Services supportano la generazione diretta di nodi Operazioni su campi e record. Utilizzando le opzioni del menu Genera del nugget del modello, è possibile generare i seguenti nodi:

- Nodo Seleziona (solo se è stato selezionato un elemento nella scheda Modello)
- nodo Filtro

Esempi di mining con Analysis Services

È disponibile un'ampia gamma di flussi di esempio che illustrano l'utilizzo del data mining di MS Analysis Services con IBM SPSS Modeler. Tali flussi si trovano nella cartella di installazione di IBM SPSS Modeler in:

\Demos\Database_Modelling\Microsoft

Nota: è possibile accedere alla cartella Demos dal gruppo di programmi IBM SPSS Modeler nel nenu Start di Windows.

Flusso di esempio: strutture ad albero delle decisioni

I flussi riportati di seguito possono essere utilizzati insieme, in ordine sequenziale, come esempio del processo di mining nel database basato sull'algoritmo per strutture ad albero delle decisioni fornito da MS Analysis Services.

Tabella 3. Strutture ad albero delle decisioni - flussi di esempio

Flusso	Descrizione
1_upload_data.str	Utilizzato per la pulitura e il caricamento di dati da un file flat nel database.
2_explore_data.str	Utilizzato come esempio di esplorazione dati con IBM SPSS Modeler.
3_build_model.str	Genera il modello utilizzando l'algoritmo nativo del database.
4_evaluate_model.str	Utilizzato come esempio di valutazione di modelli con IBM SPSS Modeler.
5_deploy_model.str	Esegue la distribuzione del modello ai fini del calcolo del punteggio in-database.

Nota: per eseguire l'esempio, i flussi devono essere eseguiti in ordine. Inoltre, i nodi Origine e Modelli in ogni flusso devono essere aggiornati per far riferimento a un'origine dati valida per il database che si desidera utilizzare.

L'insieme di dati impiegato nei flussi di esempio concerne applicazioni relative alle carte di credito e presenta un problema di classificazione relativo a un insieme misto di predittori continui e categoriali. Per ulteriori informazioni su questo insieme di dati, vedere il file *crx.names* nella stessa cartella dei flussi di esempio.

Questo insieme di dati è disponibile in UCI Machine Learning Repository alla pagina *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*

Flusso di esempio: caricamento dati

Il primo flusso di esempio, *1_upload_data.str*, viene utilizzato per pulire e caricare dati da un file flat in SQL Server.

Poichè il data mining Analysis Services richiede un campo chiave, questo flusso iniziale utilizza un nodo Ricava per aggiungere un nuovo campo al dataset denominato *CHIAVE* con valori univoci *1,2,3* utilizzando la funzione IBM SPSS Modeler `@INDEX`.

Il successivo nodo Riempimento viene utilizzato per gestire l'assenza di valori e sostituisce i campi vuoti letti dal file di testo *crx.data* con valori *NULL*.

Flusso di esempio: esplorazione dati

Il secondo flusso di esempio, *2_explore_data.str*, viene utilizzato per illustrare l'uso di un nodo Esplora per acquisire una panoramica generale dei dati, comprese statistiche riassuntive e grafici.

Facendo doppio clic su un grafico nel report del nodo Esplora, si accede a una visualizzazione più dettagliata del grafico che consente un'esplorazione più approfondita di un dato campo.

Flusso di esempio: creazione modello

Il terzo flusso di esempio, *3_build_model.str*, illustra la creazione del modello in IBM SPSS Modeler. È possibile collegare il modello di database al flusso e fare doppio clic per specificare le impostazioni di creazione.

Nella scheda Modello della finestra di dialogo è possibile specificare quanto segue:

1. Selezionare il campo **Chiave** come campo ID univoco.

Nella scheda Livello avanzato è possibile regolare le impostazioni per la creazione del modello.

Prima di procedere all'esecuzione, assicurarsi di aver specificato il database corretto per la creazione del modello. Utilizzare la scheda Server per modificare le impostazioni.

Flusso di esempio: valutazione modello

Il quarto flusso di esempio, *4_evaluate_model.str*, illustra i vantaggi associati all'utilizzo di IBM SPSS Modeler per la modellazione nel database. Una volta eseguito il modello, è possibile aggiungerlo nuovamente al flusso di dati e valutare il modello utilizzando diversi strumenti offerti in IBM SPSS Modeler.

Visualizzazione dei risultati della modellazione

È possibile fare doppio clic sul nugget del modello per esplorare i risultati. La scheda Riepilogo fornisce una vista della struttura ad albero di regole dei risultati. È inoltre possibile fare clic sul pulsante **Visualizza** della scheda Server per visualizzare graficamente il modello Strutture ad albero delle decisioni.

Valutazione dei risultati della modellazione

Il nodo Analisi nel flusso di esempio crea una matrice di coincidenza che mostra lo schema di corrispondenze tra ogni campo previsto e il relativo campo obiettivo. Eseguire il nodo Analisi per visualizzare i risultati.

Il nodo Valutazione nel flusso di esempio può creare un grafico dei profitti, progettato per mostrare i miglioramenti in termini di precisione realizzati dal modello. Eseguire il nodo Valutazione per visualizzare i risultati.

Flusso di esempio: Deployment modello

Una volta raggiunto il livello di precisione del modello desiderato, è possibile eseguire la distribuzione del modello per consentirne l'utilizzo con applicazioni esterne o la ripubblicazione nel database. Nell'ultimo flusso di esempio, *5_deploy_model.str*, i dati vengono letti dalla tabella `CREDIT`, quindi viene eseguito il calcolo del punteggio e, infine, i dati vengono pubblicati nella tabella `CREDITSCORES` mediante il nodo di esportazione del database.

L'esecuzione del flusso genera il seguente codice SQL:

```
DROP TABLE CREDITSCORES

CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )

INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
 T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
 T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
 T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
 SELECT CONVERT(NVARCHAR,[TA].[field1])  AS C0, CONVERT(NVARCHAR,[TA].[field2])  AS C1,
  [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4])  AS C3,
  CONVERT(NVARCHAR,[TA].[field5])  AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
  CONVERT(NVARCHAR,[TA].[field7])  AS C6, [TA].[field8] AS C7,
  CONVERT(NVARCHAR,[TA].[field9])  AS C8, CONVERT(NVARCHAR,[TA].[field10])  AS C9,
  [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12])  AS C11,
  CONVERT(NVARCHAR,[TA].[field13])  AS C12, [TA].[field14] AS C13,
  [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16])  AS C15,
  [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16])  AS C17,
  [TA].[$MC-field16] AS C18
 FROM openrowset('MSOLAP',
```

```
  'Datasource=localhost;Initial catalog=FoodMart 2000',
  'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
   [T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
   [T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
   [T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
   [T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
   [T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
   PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
  FROM [CREDIT1] PREDICTION JOIN
   openrowset(''MSDASQL'',
   ''Dsn=LocalServer;Uid=;pwd='',''SELECT T0."field1" AS C0,T0."field2" AS C1,
   T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
   T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
   T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
   T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
   T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
  ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
   and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
   and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
   and [T].[C14] = [CREDIT1].[field15]') AS [TA]
)  T0
```

Capitolo 4. Modellazione di database con Oracle Data Mining

Informazioni su Oracle Data Mining

IBM SPSS Modeler supporta l'integrazione con Oracle Data Mining (ODM), che include una serie di algoritmi di data mining saldamente incorporati nel sistema Oracle RDBMS. È possibile accedere a queste funzioni mediante l'ambente di sviluppo basato sui flussi di lavoro e la GUI di IBM SPSS Modeler, consentendo all'utente di utilizzare gli algoritmi di data mining offerti da ODM.

IBM SPSS Modeler supporta l'integrazione dei seguenti algoritmi di Oracle Data Mining:

- Naive Bayes
- Bayes adattivo
- Support Vector Machine (SVM)
- Modelli lineari generalizzati (GLM)*
- Struttura ad albero delle decisioni
- O-Cluster
- Medie K
- NMF (fattorizzazione a matrice non negativa)
- Apriori
- MDL (Lunghezza descrizione minima)
- Importanza attributo (AI)

solo * 11*g* R1

Requisiti per l'integrazione con Oracle

Di seguito sono elencate le condizioni che costituiscono i prerequisiti indispensabili per l'esecuzione della modellazione nel database con Oracle Data Mining. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di database.

- Esecuzione di IBM SPSS Modeler in modalità locale o in un'installazione di IBM SPSS Modeler Server su Windows o UNIX.
- Oracle 10*g*R2 o 11*g*R1 (Database 10.2 o versione successiva) con l'opzione Oracle Data Mining.

Nota: 10*g*R2 fornisce supporto per tutti gli algoritmi di modellazione del database ad eccezione dei modelli lineari generalizzati (richiede 11*g*R1).

- Un'origine dati ODBC per la connessione a Oracle, come illustrato di seguito.

Nota: le funzionalità di modellazione nel database e ottimizzazione SQL richiedono che la connettività IBM SPSS Modeler Server venga abilitata sul computer IBM SPSS Modeler. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da IBM SPSS Modeler, e accedere a IBM SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu IBM SPSS Modeler.

Guida > **Informazioni su** > **Dettagli aggiuntivi**

Se la connettività è abilitata, l'opzione **Abilitazione server** viene visualizzata nella scheda Stato della licenza.

Attivazione dell'integrazione con Oracle

Per attivare l'integrazione di IBM SPSS Modeler con Oracle Data Mining, sarà necessario configurare Oracle e creare un'origine ODBC, attivare l'integrazione nella finestra di dialogo Applicazioni di supporto di IBM SPSS Modeler e abilitare la generazione e l'ottimizzazione SQL.

Configurazione di Oracle

Per installare e configurare Oracle Data Mining, consultare la documentazione Oracle in particolare *Oracle Administrator's Guide*—per ulteriori dettagli.

Creazione di un'origine ODBC per Oracle

Per attivare la connessione tra Oracle e IBM SPSS Modeler è necessario creare un nome di origine dati (DSN, Data Source Name) ODBC.

Per creare un DSN, è necessario avere una conoscenza di base delle origini dati e dei driver ODBC e disporre del supporto database in IBM SPSS Modeler.

Se l'applicazione è in esecuzione in modalità distribuita su IBM SPSS Modeler Server, creare il DSN sul computer server. Se invece è attiva la modalità locale (client), creare il DSN sul computer client.

1. Installare i driver ODBC. I driver sono disponibili sul disco di installazione di IBM SPSS Data Access Pack fornito con questa versione. Eseguire il file *setup.exe* per avviare il programma di installazione, e selezionare tutti i driver opportuni. Attenersi alle istruzioni visualizzate per installare i driver.
   a. Creare il DSN.

      Nota: La sequenza dei menu dipende dalla versione di Windows.

      - **Windows XP.** dal menu Start, scegliere **Pannello di controllo**. Fare doppio clic su **Strumenti di amministrazione** e doppio clic su **Origini dati (ODBC)**.
      - **Windows Vista.** Dal menu Start, scegliere **Pannello di controllo**, quindi **Strumenti di amministrazione**. Doppio clic su **Strumenti di amministrazione**, selezionare**Origini dati (ODBC)**, quindi fare clic su **Apri**.
      - **Windows 7.** dal menu Start, scegliere **pannello di controllo**, quindi**Sistema & Sicurezza**, quindi **Strumenti di amministrazione**. Selezionare**Origini dati (ODBC)**, then click **Open**.

   b. Fare clic sulla scheda **DSN di sistema**, quindi fare clic su **Aggiungi**.
2. Selezionare il driver **SPSS OEM 6.0 Oracle Wire Protocol**.
3. Fare clic su **Fine**.
4. Nella schermata di impostazione del driver ODBC Oracle Wire Protocol immettere il nome di una origine dati a scelta, il nome host del server Oracle, il numero di porta per la connessione e il SID dell'istanza Oracle in uso.

   Nome host, numero di porta e SID possono essere ottenuti dal file *tnsnames.ora*, presente sul computer server, se è stato implementato TNS con un file *tnsnames.ora*. Per ulteriori informazioni, contattare l'amministratore Oracle.
5. Fare clic sul pulsante **Test** per verificare la connessione.

Attivazione dell'integrazione di Oracle Data Mining in IBM SPSS Modeler

1. Dai menu di IBM SPSS Modeler scegliere:

   Strumenti > **Opzioni** > **Applicazione di supporto**
2. Fare clic sulla scheda **Oracle**.

Abilita integrazione Oracle Data Mining. Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra IBM SPSS Modeler e aggiunge i nodi degli algoritmi di Oracle Data Mining.

Connessione Oracle. Specificare l'origine dati ODBC Oracle di default, utilizzata per la creazione e l'archiviazione di modelli, insieme a un nome utente e una password validi. Questa impostazione può essere sovrascritta nei singoli nodi modelli e nugget del modello.

Nota: La connessione al database utilizzata per scopi di modellazione può o non può essere la stessa connessione utilizzata per accedere ai dati. Per esempio, è possibile utilizzare un flusso che accede ai dati di un database Oracle, li scarica in IBM SPSS Modeler per la pulitura o altre operazioni di modifica e, infine, li carica in un database Oracle differente per la modellazione. In alternativa, i dati originali possono risiedere in un file flat o in un'altra origine (non Oracle), nel qual caso sarà necessario caricarli in Oracle per il processo di modellazione. In tutti i casi, i dati verranno automaticamente caricati in una tabella temporanea creata nel database utilizzato per la modellazione.

Avvisa prima di sovrascrivere un modello Oracle Data Mining. Selezionare questa opzione per garantire che i modelli archiviati nel database non siano sovrascritti da IBM SPSS Modeler senza alcun preavviso.

Elenca i modelli Oracle Data Mining. Visualizza i modelli di data mining disponibili.

Attiva avvio di Oracle Data Miner. (facoltativo) Se attivata, consente a IBM SPSS Modeler di avviare l'applicazione Oracle Data Miner. Per ulteriori informazioni, fare riferimento a "Oracle Data Miner" a pagina 48.

Percorso file eseguibile di Oracle Data Miner. (facoltativo) Specifica la posizione fisica del file eseguibile di Oracle Data Miner per Windows (per esempio *C:\odm\bin\odminerw.exe*). Oracle Data Miner non viene installato insieme a IBM SPSS Modeler; è necessario scaricare la versione corretta dal sito Web di Oracle (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) e installarla sul client.

Attivazione di generazione e ottimizzazione SQL

1. Dal menu IBM SPSS Modeler scegliere:
   Strumenti > **Proprietà flusso** > **Opzioni**
2. Fare clic sull'opzione **Ottimizzazione** nel riquadro di spostamento.
3. Confermare che l'opzione **Genera SQL** è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
4. Selezionare **Ottimizza generazione SQL** e **Ottimizza altre esecuzioni** (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Creazione di modelli con Oracle Data Mining

I nodi di creazione modelli Oracle funzionano esattamente come gli altri nodi Modelli di IBM SPSS Modeler con alcune eccezioni. È possibile accedere a questi nodi dalla palette Modelli in-database, presente nella parte inferiore della finestra di IBM SPSS Modeler.

Considerazioni sui dati

Oracle richiede che i dati categoriali siano archiviati in formato stringa (CHAR o VARCHAR2). Di conseguenza, IBM SPSS Modeler non consentirà di specificare campi di archiviazione numerici con livello di misurazione *Flag* o *Nominale* (categoriali) come input per modelli ODM. Se necessario, i numeri possono essere convertiti in stringhe in IBM SPSS Modeler utilizzando il nodo Ricodifica.

Campo obiettivo. Nei modelli di classificazione ODM è possibile selezionare un solo campo come campo di output (obiettivo).

Nome modello. A partire da Oracle 11*g*R1, il nome `unique` p una parola chiave e non è possibile utilizzarla come un nome modello personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Commenti generali

- IBM SPSS Modeler non consente di eseguire operazioni di esportazione e importazione PMML per i modelli creati mediante Oracle Data Mining.
- In ODM viene sempre eseguito il calcolo del punteggio del modello. Può essere necessario caricare l'insieme di dati in una tabella temporanea, qualora i dati vengano originati in IBM SPSS Modeler o debbano essere preparati all'interno dell'applicazione.
- In IBM SPSS Modeler viene generalmente fornita una sola previsione con la probabilità o la confidenza associata.
- IBM SPSS Modeler restringe a 1.000 il numero di campi utilizzabili per la creazione del modello e il calcolo del punteggio.
- IBM SPSS Modeler è in grado di calcolare il punteggio dei modelli ODM dall'interno di flussi pubblicati per l'esecuzione utilizzando IBM SPSS Modeler Solution Publisher.

Opzioni della scheda Server dei modelli Oracle

Specificare la connessione Oracle utilizzata per caricare i dati per la modellazione. Se necessario, inoltre, è possibile selezionare nella scheda Server una connessione specifica per ogni nodo Modelli che sovrascriva la connessione Oracle di default indicata nella finestra di dialogo Applicazioni di supporto. Consultare l'argomento "Attivazione dell'integrazione con Oracle" a pagina 30 per ulteriori informazioni.

Commenti

- La connessione utilizzata per la modellazione può corrispondere o meno a quella impiegata nel nodo origine di un flusso. Per esempio, è possibile utilizzare un flusso che accede ai dati di un database Oracle, li scarica in IBM SPSS Modeler per la pulitura o altre operazioni di modifica e, infine, li carica in un database Oracle differente per la modellazione.
- Il nome dell'origine dati ODBC è efficacemente incorporato in ogni flusso di IBM SPSS Modeler. Se un flusso creato su un determinato host viene eseguito su un host differente, il nome dell'origine dati deve essere lo stesso su entrambi gli host. In alternativa, è possibile selezionare un'origine dati differente nella scheda Server all'interno di ogni nodo Modelli o di input.

Costi classificazione errata

In alcuni contesti, certi tipi di errori rappresentano un costo maggiore rispetto ad altri. Potrebbe essere più costoso, per esempio, classificare come a basso rischio chi richiede un fido ad alto rischio (un tipo di errore) di quanto non lo sia classificare come ad alto rischio chi chiede un fido a basso rischio (un tipo di errore diverso). I costi di errata classificazione permettono di specificare l'importanza relativa dei diversi tipi di errori di previsione.

Essenzialmente i costi di errata classificazione sono pesi applicati a risultati specifici. Tali pesi vengono fattorizzati nel modello e possono realmente modificare la previsione (come modo per proteggersi da errori che gravano sui costi).

Ad eccezione dei modelli C5.0, i costi di errata classificazione non sono applicati quando si calcola il punteggio di un modello e non vengono presi in considerazione quando si classificano o confrontano modelli con il nodo Classificatore automatico, un nodo Analisi o un grafico di valutazione. È possibile che un modello che include costi non generi un numero inferiore di errori rispetto a uno che non li

include e che non si classifichi meglio in termini di precisione complessiva, ma è probabile che fornisca prestazioni migliori in termini pratici poiché include una distorsione incorporata a favore degli errori *meno costosi*.

La matrice dei costi mostra il costo per ciascuna possibile combinazione di categoria prevista e categoria effettiva. Per default, tutti i costi di errata classificazione sono impostati su 1.0. Per immettere valori di costi personalizzati, selezionare **Utilizza costi di errata classificazione** e immettere i valori personalizzati nella matrice dei costi.

Per cambiare un costo di errata classificazione, selezionare la cella corrispondente alla combinazione desiderata di valori previsti e valori effettivi, eliminare il contenuto esistente della cella e immettere il costo desiderato. I costi non sono simmetrici automaticamente. Se si imposta, per esempio, il costo dell'errata classificazione di *A* come *B* su 2.0 e non viene esplicitamente cambiato il costo dell'errata classificazione di *B* come *A*, esso manterrà il valore di default 1.0.

Nota: solo il modello Strutture ad albero delle decisioni consente di specificare i costi al momento della creazione.

Naive Bayes Oracle

Naive Bayes è un algoritmo molto noto per problemi di classificazione. Il modello è denominato *naive* perché considera tutte le variabili di previsione proposte indipendenti l'una dall'altra. Naive Bayes è un algoritmo veloce e scalabile in grado di calcolare probabilità condizionali per combinazioni di attributi e per l'attributo obiettivo. Dai dati di addestramento viene stabilita una probabilità indipendente che serve a indicare la verosimiglianza di ciascuna classe obiettivo una volta specificata l'occorrenza di ogni categoria di valore da ogni variabile di input.

- La convalida incrociata viene utilizzata per verificare la precisione di un modello con gli stessi dati adoperati per creare il modello. Questa operazione risulta particolarmente utile quando il numero di casi disponibili per creare un modello è ridotto.
- L'output del modello può essere visualizzato in formato matrice. I numeri della matrice indicano probabilità condizionali che mettono in relazione le classi previste (colonne) e le combinazioni di valori-variabili dei predittori (righe).

Opzioni del modello Naive Bayes

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Opzioni avanzate di Naive Bayes

Al momento della creazione del modello, i valori o le coppie di valori dei singoli attributi dei predittori vengono ignorati, a meno che non ci siano sufficienti occorrenze di un determinato valore o coppia di valori nei dati di addestramento. Le soglie per ignorare i valori vengono specificate come frazioni basate sul numero di record presenti nei dati di addestramento. Adeguando queste soglie è possibile ridurre il rumore e migliorare la capacità del modello di essere generalizzato per altri insiemi di dati.

- **Soglia singleton.** Specifica la soglia per un determinato valore di attributo predittore. Il numero di occorrenze di un determinato valore deve essere uguale o maggiore della frazione specificata, altrimenti il valore verrà ignorato.
- **Soglia per coppia.** Specifica la soglia per una determinata coppia di valori di attributo e predittore. Il numero di occorrenze di una determinata coppia di valori deve essere uguale o maggiore della frazione specificata, altrimenti la coppia verrà ignorata.

Probabilità di previsione Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere **Seleziona**, fare clic sul pulsante **Specifica**, scegliere uno dei risultati possibili e fare clic su **Inserisci**.

Utilizza insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

Bayes adattivi Oracle

La rete di Bayes adattivi (ABN) crea classificatori di rete bayesiana utilizzando la lunghezza di descrizione minima (MDL) e la selezione automatica delle funzionalità. ABN funziona bene in alcune situazioni in cui il modello Naive Bayes non garantisce performance adeguate e in molti altri casi, sebbene con performance meno elevate. L'algoritmo ABN consente di creare tre tipi di modelli avanzati su base bayesiana, tra cui i modelli di struttura ad albero delle decisioni semplificata (funzione singola), Naive Bayes tagliato e multifunzione boosted.

Nota: L'algoritmo Oracle Adaptive Bayes è stato abbandonato in Oracle 12C e non è supportato in IBM SPSS Modeler quando si utilizza Oracle 12C. Consultare http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726.

Modelli generati

Nella modalità di creazione a funzione singola, ABN produce una struttura ad albero delle decisioni semplificata basata su un insieme di regole leggibili, che consente all'utente di business o all'analista di comprendere la base delle previsioni del modello, per agire di conseguenza o fornire spiegazioni ad altri. Consente di ottenere un vantaggio significativo rispetto ai modelli Naive Bayes e multifunzione. Queste regole possono essere visualizzate come un insieme di regole standard in IBM SPSS Modeler. Un semplice insieme di regole potrebbe avere il seguente aspetto:

```
IF MARITAL_STATUS = "Married"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "TRUE"
Confidence = .78, Support = 570 cases
```

I modelli Naive Bayes tagliato e multifunzione non possono essere visualizzati in IBM SPSS Modeler.

Opzioni del modello Bayes adattivo

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Tipo di modello

È possibile scegliere fra tre diverse modalità di creazione del modello.

- **Multifunzione.** Crea e confronta diversi modelli, inclusi un modello Naive Bayes e modelli di probabilità dei prodotti a funzione singola e multifunzione. Si tratta della modalità più completa e in genere comporta tempi di elaborazione maggiori. Le regole vengono prodotte solo se risulta che il modello a funzione singola è il migliore. Se si sceglie un modello multifunzione o Naive Bayes, non vengono prodotte regole.
- **Singola funzione.** Crea una struttura ad albero delle decisioni semplificata basata su un insieme di regole. Ogni regola contiene una condizione e le probabilità associate a ciascun risultato. Le regole si escludono a vicenda e vengono fornite in un formato leggibile, offrendo un significativo vantaggio rispetto ai modelli Naive Bayes e multifunzione.
- **Naive Bayes.** Crea un singolo modello Naive Bayes e lo confronta con l'a priori del campione globale (la distribuzione di valori obiettivo nel campione globale). Il modello Naive Bayes viene prodotto come output solo se risulta essere un predittore migliore dei valori obiettivo rispetto all'apriori globale. Altrimenti come output non viene prodotto alcun modello.

Opzioni avanzate di Bayes adattivo

Limita tempo di esecuzione. Selezionare questa opzione per specificare un tempo di creazione massimo in minuti. Ciò consente di produrre modelli in tempi più brevi, sebbene ne possa risultare una minore precisione. A ciascun passaggio importante del processo di modellazione, l'algoritmo controlla, prima di continuare, se sarà in grado di completare il passaggio successivo entro l'intervallo di tempo specificato e restituisce il miglior modello disponibile al raggiungimento del limite.

Numero massimo di predittori. Questa opzione consente di limitare la complessità del modello e migliorare le performance limitando il numero di predittori utilizzati. I predittori vengono classificati in base alla misura MDL della loro correlazione all'obiettivo, come misura della probabilità di essere inclusi nel modello.

Numero massimo di predittori Naive Bayes. Questa opzione specifica il numero massimo di predittori da utilizzare nel modello Naive Bayes.

Support Vector Machine Oracle (SVM)

SVM (Support Vector Machine) è un algoritmo di classificazione e regressione che utilizza la teoria di apprendimento automatico per ottimizzare la precisione predittiva senza sovradattare i dati. SVM utilizza una trasformazione non lineare opzionale dei dati di addestramento, seguita dalla ricerca di equazioni di regressione nei dati trasformati per la separazione delle classi (per gli obiettivi categoriali) o l'adattamento dell'obiettivo (per i target continui). L'implementazione Oracle di SVM consente di creare i modelli utilizzando uno dei due kernel disponibili, ossia lineare o gaussiano. Il kernel lineare omette completamente la trasformazione non lineare, in modo che il modello prodotto risulti essenzialmente un modello di regressione.

Per ulteriori informazioni, consultare *Oracle Data Mining Application Developer's Guide* e *Oracle Data Mining Concepts*.

Opzioni del modello SVM Oracle

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Apprendimento attivo. Fornisce un modo per gestire insiemi di creazione di grandi dimensioni. Mediante l'apprendimento attivo, l'algoritmo crea un modello iniziale basato su un piccolo esempio prima di applicarlo all'insieme di dati di addestramento completo e aggiorna in modo incrementale il campione e il modello in base ai risultati. Il ciclo viene ripetuto finché il modello converge sui dati di addestramento o finché non viene raggiunto il numero massimo di vettori di supporto consentiti.

Funzione Kernel. Selezionare **Lineare** o **Gaussiana**, oppure lasciare il valore di default **Determinato dal sistema** per consentire al sistema di scegliere il kernel più adatto. I kernel gaussiani sono in grado di apprendere relazioni più complesse, ma richiedono in genere tempi di elaborazione maggiori. Può essere opportuno iniziare con il kernel lineare, per poi passare al gaussiano solo se il kernel lineare non riesce a trovare un buon adattamento. Questa situazione si verifica in genere con i modelli di regressione, in cui la scelta del kernel ha un'importanza maggiore. Si noti, inoltre, che i modelli SVM creati con il kernel gaussiano non possono essere visualizzati in IBM SPSS Modeler. I modelli creati con il kernel lineare possono essere visualizzati in IBM SPSS Modeler esattamente come i modelli di regressione standard.

Metodo di normalizzazione Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere **Punteggio Z**, **Min-Max** o **Nessuno**. Oracle esegue automaticamente la normalizzazione se è selezionata la casella di controllo **Preparazione dati automatici**. Per impostare manualmente il metodo di normalizzazione, deselezionare la casella di controllo.

Opzioni avanzate di SVM Oracle

Dimensione cache Kernel. Specifica la dimensione in byte della cache da utilizzare per l'archiviazione dei kernel calcolati durante l'operazione di creazione. Com'è facilmente intuibile, cache di dimensioni maggiori consentono tempi di creazione più rapidi. Il valore predefinito è 50MB.

Tolleranza di convergenza. Specifica il valore di tolleranza consentito prima della terminazione per la creazione del modello. Questo valore deve essere compreso tra 0 e 1. L'impostazione di default è 0.001. Valori maggiori consentono tempi di creazione più rapidi ma producono modelli meno precisi.

Specifica la deviazione standard. Specifica il parametro di deviazione standard utilizzato dal kernel gaussiano. Questo parametro incide sul rapporto tra la complessità del modello e la possibilità di essere generalizzato ad altri insiemi di dati (sovradattando e sottoadattando i dati). Valori di deviazione standard maggiori favoriscono il sottoadattamento. Questo parametro viene calcolato di default a partire dai dati di addestramento.

Specifica epsilon. Nei modelli di regressione, specifica il valore dell'intervallo dell'errore consentito nella creazione di modelli senza rilevamento epsilon. In altre parole, distingue errori di piccola portata (che vengono ignorati) da errori più gravi (che non vengono ignorati). Il valore deve essere compreso tra 0 e 1 e viene calcolato di default dai dati di addestramento.

Specifica fattore di complessità Specifica il fattore di complessità, che bilancia il rapporto tra errore del modello (misurato a fronte dei dati di addestramento) e complessità del modello, per evitare il sovradattamento o il sottoadattamento dei dati. Valori maggiori comportano un livello di penalità più alto per gli errori, con un maggiore rischio di sovradattamento dei dati; valori minori, invece, comportano un livello di penalità più basso e possono portare al sottoadattamento.

Specifica tasso valore anomalo Specifica il tasso desiderato di valori anomali nei dati di addestramento. Valido solo per modelli SVM a una classe. Non può essere utilizzata insieme all'impostazione **Specifica fattore di complessità**.

Probabilità di previsione Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere **Seleziona**, fare clic sul pulsante **Specifica**, scegliere uno dei risultati possibili e fare clic su **Inserisci**.

Utilizza insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

Opzioni Pesi di SVM Oracle

In un modello di classificazione, i pesi consentono di specificare l'importanza relativa dei diversi valori di destinazione possibili. Questo può essere utile, per esempio, se i punti dati di addestramento non sono distribuiti in modo realistico tra le categorie. I pesi consentono di applicare una distorsione al modello in modo da compensare le categorie meno rappresentate nei dati. L'incremento del peso di un valore di destinazione dovrebbe aumentare la percentuale di previsioni corrette per quella categoria.

Esistono tre metodi per impostare i pesi:

- **Basato sui dati di addestramento.** Questa è l'opzione di default. I pesi si basano sulle frequenze relative delle categorie nei dati di addestramento.
- **Uguale per tutte le classi.** I pesi per tutte le categorie sono definiti come $1/k$, dove k è il numero di categorie obiettivo.
- **Personalizzato.** È possibile specificare pesi personalizzati. L'impostazione dei valori iniziali per i pesi è uguale per tutte le classi. È possibile impostare i pesi per le singole categorie su valori definiti dall'utente. Per regolare il peso di una categoria specifica, selezionare la cella Peso nella tabella corrispondente alla categoria desiderata, eliminare il contenuto della cella e immettere il valore desiderato.

La somma devi pesi di tutte le categorie deve essere uguale a 1,0. Se non assommano a 1.0 viene visualizzato un avviso e viene offerta la possibilità di normalizzare automaticamente i valori. Questa modifica automatica mantiene le proporzioni tra le varie categorie e al contempo applica il vincolo di peso. Tale modifica può essere eseguita in qualsiasi momento facendo clic sul pulsante **Normalizza**. Per riportare la tabella su valori uguali per tutte le categorie, fare clic sul pulsante **Equalizza**.

Modelli lineari generalizzati Oracle (GLM)

(11g soltanto) I modelli lineari generalizzati allentano le supposizioni restrittive effettuate dai modelli lineari. Tali supposizioni includono, per esempio, le supposizioni che la variabile obiettivo abbia una distribuzione normale e che l'effetto dei predittori su tale variabile sia lineare per natura. Un modello lineare generalizzato è adatto per le previsioni in cui è probabile che la distribuzione dell'obiettivo sia non normale, per esempio una distribuzione multinomiale o di Poisson. Analogamente, un modello lineare generalizzato è utile nei casi in cui è probabile che la relazione o il collegamento tra i predittori e l'obiettivo siano non-lineari.

Per ulteriori informazioni, consultare *Oracle Data Mining Application Developer's Guide* e *Oracle Data Mining Concepts*.

Opzioni del modello GLM Oracle

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Metodo di normalizzazione Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere **Punteggio Z**, **Min-Max** o **Nessuno**. Oracle esegue automaticamente la normalizzazione se è selezionata la casella di controllo **Preparazione dati automatici**. Per impostare manualmente il metodo di normalizzazione, deselezionare la casella di controllo.

Gestione valore mancante. Specifica come elaborare i valori mancanti nei dati di input:

- **Sostituisci con media o modalità** sostituisce i valori mancanti degli attributi numerici con il valore della media e sostituisce i valori mancanti degli attributi categoriali con la modalità.
- **Utilizza solo record completi** ignora i record con valori mancanti.

Opzioni avanzate di GLM Oracle

Utilizza pesi righe. Selezionare questa casella per attivare l'elenco a discesa adiacente da dove è possibile selezionare una colonna contenente un fattore di ponderazione per le righe.

Salva diagnostiche righe nella tabella. Selezionare questa casella di controllo per attivare il campo testo adiacente in cui è possibile specificare il nome di una tabella contenente diagnostiche a livello di riga.

Livello di confidenza del coefficiente Il grado di certezza, da 0,0 a 1,0, che il valore previsto per l'obiettivo rientrerà in un intervallo di confidenza calcolato dal modello. I limiti di confidenza vengono restituiti con le statistiche dei coefficienti.

Categoria di riferimento per l'obiettivo. Selezionare **Personalizzato** per scegliere un valore del campo obiettivo da utilizzare come categoria di riferimento oppure lasciare il valore di default **Auto** .

Regressione Ridge La regressione ridge è una tecnica che compensa nel caso in cui il grado di correlazione nelle variabili sia troppo elevato. È possibile utilizzare l'opzione **Automatico** per consentire all'algoritmo di controllare l'utilizzo di questa tecnica, oppure è possibile controllarlo manualmente mediante le opzioni **Disattiva** e **Attiva**. Se si sceglie di attivare manualmente la regressione ridge, è possibile ignorare il valore di default del sistema per il parametro ridge specificando un valore nel campo adiacente.

Produci VIF per la regressione ridge. Selezionare questa casella se si desidera produrre delle statistiche VIF (Variance Inflation Factor, fattore di inflazione della varianza) quando viene utilizzata la tecnica ridge per la regressione lineare.

Probabilità di previsione Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere **Seleziona**, fare clic sul pulsante **Specifica**, scegliere uno dei risultati possibili e fare clic su **Inserisci**.

Utilizza insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

Opzioni Pesi di GLM Oracle

In un modello di classificazione, i pesi consentono di specificare l'importanza relativa dei diversi valori di destinazione possibili. Questo può essere utile, per esempio, se i punti dati di addestramento non sono distribuiti in modo realistico tra le categorie. I pesi consentono di applicare una distorsione al modello in modo da compensare le categorie meno rappresentate nei dati. L'incremento del peso di un valore di destinazione dovrebbe aumentare la percentuale di previsioni corrette per quella categoria.

Esistono tre metodi per impostare i pesi:

- **Basato sui dati di addestramento.** Questa è l'opzione di default. I pesi si basano sulle frequenze relative delle categorie nei dati di addestramento.
- **Uguale per tutte le classi.** I pesi per tutte le categorie sono definiti come $1/k$, dove k è il numero di categorie obiettivo.
- **Personalizzato.** È possibile specificare pesi personalizzati. L'impostazione dei valori iniziali per i pesi è uguale per tutte le classi. È possibile impostare i pesi per le singole categorie su valori definiti dall'utente. Per regolare il peso di una categoria specifica, selezionare la cella Peso nella tabella corrispondente alla categoria desiderata, eliminare il contenuto della cella e immettere il valore desiderato.

La somma devi pesi di tutte le categorie deve essere uguale a 1,0. Se non assommano a 1.0 viene visualizzato un avviso e viene offerta la possibilità di normalizzare automaticamente i valori. Questa modifica automatica mantiene le proporzioni tra le varie categorie e al contempo applica il vincolo di peso. Tale modifica può essere eseguita in qualsiasi momento facendo clic sul pulsante **Normalizza**. Per riportare la tabella su valori uguali per tutte le categorie, fare clic sul pulsante **Equalizza**.

Struttura ad albero delle decisioni Oracle

Oracle Data Mining offre una classica funzionalità Struttura ad albero delle decisioni, basata sul diffuso algoritmo Strutture ad albero di regressione e di classificazione. Il modello Struttura ad albero delle decisioni ODM contiene informazioni complete sui singoli nodi, fra cui criterio di suddivisione, supporto e confidenza. È possibile visualizzare per intero la Regola per ciascun nodo; inoltre, per ogni nodo viene fornito un attributo surrogato, da utilizzare come sostituto quando si applica il modello a un caso in cui mancano dei valori.

Le strutture ad albero delle decisioni sono molto diffuse in quanto sono applicabili universalmente oltre a essere facili da utilizzare e capire. Le strutture ad albero delle decisioni vagliano ogni potenziale attributo di input alla ricerca della migliore "suddivisione," cioè il punto di divisione degli attributi (ad esempio `ETA' > 55`,) che suddivide i record dei dati a valle in popolazioni più omogenee. Dopo ogni decisione di suddivisione, ODM ripete il processo espandendo la struttura ad albero verso l'esterno e creando foglie terminali che rappresentano popolazioni di record, elementi o persone simili. Guardando verso il basso dal nodo root della struttura ad albero (per esempio la popolazione totale), le strutture ad albero delle decisioni forniscono regole leggibili di istruzioni `IF A, then B`. Queste regole della struttura ad albero delle decisioni forniscono anche il supporto e la confidenza per ciascun nodo della struttura ad albero.

Mentre le reti di Bayes adattivi possono anche fornire regole brevi e semplici, utili a offrire spiegazioni in merito alle singole previsioni, le strutture ad albero delle decisioni forniscono regole Oracle Data Mining complete per ciascuna decisione di divisione. Le strutture ad albero delle decisioni sono anche utili per sviluppare profili dettagliati dei clienti migliori, dei pazienti sani, dei fattori associati alla frode, e così via.

Opzioni della scheda Modello per il nodo Struttura ad albero delle decisioni

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Metrica di impurità. Specifica la metrica utilizzata per cercare la migliore domanda di test per la suddivisione dei dati nei singoli nodi. Il divisore e il valore di suddivisione migliori sono quelli che risultano nel maggior incremento dell'omogeneità del valore di destinazione per le entità presenti nel nodo. L'omogeneità viene misurata in base a un tipo di metrica. Sono supportate le metriche **gini** ed **entropia**.

Opzioni avanzate Struttura ad albero delle decisioni

Profondità massima. Imposta la profondità massima del modello di struttura ad albero da creare.

Percentuale minima di record in un nodo. Imposta la percentuale del numero minimo di record per nodo.

Percentuale minima di record per una suddivisione. Imposta il numero minimo di record in un nodo padre, espresso come valore percentuale del numero totale dei record utilizzati per l'addestramento del modello. Se il numero dei record è inferiore a questo valore percentuale, il sistema non esegue alcuna suddivisione.

Numero minimo di record in un nodo. Imposta il numero minimo di record restituiti.

Numero minimo di record per una suddivisione. Imposta il numero minimo di record in un nodo padre, espresso sotto forma di valore. Se il numero dei record è inferiore a questo valore, il sistema non esegue alcuna suddivisione.

Identificativo regola. Se selezionata, include nel modello una stringa per identificare il nodo della struttura ad albero in cui viene eseguita una determinata suddivisione.

Probabilità di previsione Consente al modello di includere la probabilità di una previsione corretta di un possibile risultato del campo obiettivo. Per abilitare questa funzione, scegliere **Seleziona**, fare clic sul pulsante **Specifica**, scegliere uno dei risultati possibili e fare clic su **Inserisci**.

Utilizza insieme di previsioni. Genera una tabella di tutti i risultati possibili relativi a tutti gli esiti possibili del campo obiettivo.

O-Cluster Oracle

L'algoritmo O-Cluster Oracle consente di identificare i raggruppamenti che si creano spontaneamente all'interno di una popolazione di dati. Il raggruppamento cluster a partizioni ortogonali, O-Cluster, è un algoritmo di raggruppamento proprietario di Oracle che consente di creare un modello di raggruppamento gerarchico basato su una griglia, vale a dire, crea partizioni ortogonali (parallele all'asse) nello spazio dell'attributo di input. Questo algoritmo funziona in modo ricorsivo. La struttura gerarchica risultante rappresenta una griglia irregolare che suddivide lo spazio dell'attributo in raggruppamenti a tasselli.

L'algoritmo O-Cluster gestisce sia gli attributi numerici sia gli attributi categoriali e ODM seleziona automaticamente le definizioni di raggruppamenti cluster migliori. ODM fornisce informazioni dettagliate sui cluster, le relative regole e i valori del baricentro e può esser utilizzato per calcolare il punteggio di una popolazione in base all'appartenenza al cluster.

Opzioni del modello O-Cluster

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Numero massimo di cluster. Imposta il numero massimo di cluster generati.

Opzioni avanzate di O-Cluster

Dimensione massima buffer. Imposta le dimensioni massime del buffer.

Sensibilità. Imposta una frazione che specifica la densità di picco necessaria per la separazione di un nuovo cluster. Il valore frazionale è relativo alla densità uniforme globale.

Medie K Oracle

L'algoritmo Medie K Oracle consente di identificare i raggruppamenti che si creano spontaneamente all'interno di una popolazione di dati. L'algoritmo Medie K è un algoritmo di raggruppamento cluster basato sulla distanza che suddivide i dati sotto forma di partizioni in un numero predeterminato di cluster (a condizione che ci sia un numero sufficiente di casi distinti). Gli algoritmi basati sulla distanza fanno affidamento su una metrica di distanza (funzione) per misurare la similarità tra i punti dati. I punti dati vengono assegnati al cluster più vicino in base alla metrica di distanza utilizzata. ODM fornisce una versione migliorata di Medie K.

L'algoritmo Medie K supporta i cluster gerarchici, gestisce gli attributi numerici e categoriali e suddivide la popolazione nel numero di cluster specificati dall'utente. ODM fornisce informazioni dettagliate sui cluster, le relative regole e i valori del baricentro e può esser utilizzato per calcolare il punteggio di una popolazione in base all'appartenenza al cluster.

Opzioni del modello Medie K

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Numero di cluster. Imposta il numero di cluster generati

Funzione Distanza. Specifica la funzione distanza utilizzata per il raggruppamento cluster Medie K.

Criterio di suddivisione. Specifica il criterio di suddivisione utilizzato per il raggruppamento cluster Medie K.

Metodo di normalizzazione. Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere **Punteggio Z**, **Min-Max** o **Nessuno**.

Opzioni avanzate del nodo Medie K

Iterazioni. Specifica il numero di iterazioni per l'algoritmo Medie K.

Tolleranza di convergenza. Imposta la tolleranza di convergenza per l'algoritmo Medie K.

Numero di bin. Specifica il numero di bin nell'istogramma dell'attributo prodotto da Medie K. I limiti del bin per i singoli attributi vengono calcolati globalmente sull'intero insieme di dati di addestramento. Il metodo di discretizzazione è ad ampiezza equivalente. Tutti gli attributi hanno lo stesso numero di bin ad eccezione degli attributi con un unico valore, che invece hanno un solo bin.

Espansione blocco. Imposta il fattore di espansione relativo alla memoria allocata per la memorizzazione dei dati dei cluster.

Supporto attributo percentuale minimo. Imposta la frazione dei valori dell'attributo che devono essere non null per far sì che l'attributo venga incluso nella descrizione della regola per il cluster. L'impostazione di un valore troppo alto per questo parametro nel caso di dati con valori mancanti può determinare la creazione di regole molto brevi o addirittura vuote.

NMF di Oracle (fattorizzazione a matrice non negativa)

L'algoritmo NMF è utile per ridurre un insieme di dati molto grosso in attributi rappresentativi. Concettualmente analogo all'algoritmo di Analisi dei componenti principali (PCA) ma in grado di gestire quantità di attributi maggiori in un modello di rappresentazione additivo, l'algoritmo NMF è un algoritmo di data mining potente e all'avanguardia, che può essere utilizzato in svariati casi.

L'algoritmo NMF può essere utilizzato per ridurre grandi quantità di dati (per esempio dati di testo) in rappresentazioni più piccole e sparse, che riducono la dimensionalità dei dati (le stesse informazioni possono essere conservate utilizzando un numero di variabili molto inferiore). L'output del modello NMF può essere analizzato mediante tecniche di apprendimento supervisionato, come le tecniche SVM, o non

supervisionato, come le tecniche di raggruppamento tramite cluster. Oracle Data Mining utilizza gli algoritmi NMF e SVM per eseguire il mining di dati di testo non strutturati.

Opzioni del modello NMF

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Metodo di normalizzazione Specifica il metodo di normalizzazione per i campi obiettivo e di input continuo. È possibile scegliere **Punteggio Z**, **Min-Max** o **Nessuno**. Oracle esegue automaticamente la normalizzazione se è selezionata la casella di controllo **Preparazione dati automatici**. Per impostare manualmente il metodo di normalizzazione, deselezionare la casella di controllo.

Opzioni avanzate NMF

Specifica il numero di funzioni. Specifica il numero delle funzionalità da estrarre.

Seme random. Imposta il seme random per l'algoritmo NMF.

Numero di iterazioni. Imposta il numero delle iterazioni per l'algoritmo NMF.

Tolleranza di convergenza. Imposta la tolleranza di convergenza per l'algoritmo NMF.

Visualizza tutte le funzioni. Consente di visualizzare ID e confidenza per tutte le funzionalità e non solo per la funzionalità migliore.

Apriori Oracle

L'algoritmo Apriori scopre le regole di associazione presenti nei dati. Ad esempio, "se un cliente acquista un rasoio e una lozione dopobarba, esiste una confidenza dell'80% che comprerà poi la schiuma da barba". Il problema di mining di associazione può essere scomposto in due sottoproblemi:

- Trovare tutte le combinazioni di elementi, denominate insiemi di elementi frequenti, il cui supporto sia superiore al valore di supporto minimo.
- Utilizzare gli insiemi di elementi frequenti per generare le regole desiderate. Il concetto è il seguente: se, per esempio, ABC e BC sono frequenti, la regola "A implica BC" è vera se il rapporto tra `support(ABC)` e `support(BC)` è grande almeno tanto quanto la confidenza minima. Si noti che la regola avrà un supporto minimo in quanto ABCD è frequente. L'associazione ODM supporta solo le regole conseguenti singole (ABC implica D).

Il numero degli insiemi di elementi frequenti è determinato dai parametri di supporto minimo. Il numero delle regole generate è determinato dal numero degli insiemi di elementi frequenti e dal parametro di confidenza. Se il parametro di confidenza è stato impostato su un valore troppo alto, è possibile che nel modello di associazione siano presenti degli insiemi di elementi frequenti ma nessuna regola.

ODM utilizza un'implementazione basata su SQL dell'algoritmo Apriori. La generazione di candidati e i passaggi di conteggio del supporto vengono implementati mediante query SQL. Non vengono invece utilizzate le strutture di dati in memoria specializzate. Le query SQL vengono perfezionate in modo da essere eseguite efficacemente nel server di database, utilizzando diversi suggerimenti.

Opzioni dei campi Apriori

In tutti i nodi Modelli è disponibile una scheda Campi nella quale è possibile specificare i campi da utilizzare per la creazione del modello.

Per poter generare un modello Apriori, è necessario prima specificare i campi da utilizzare come elementi rilevanti nella creazione di modelli di associazione.

Utilizzo delle impostazioni del nodo tipo. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo Tipo upstream. Questa è l'opzione di default.

Utilizzo delle impostazioni personalizzate. Questa opzione indica al nodo di utilizzare le informazioni sui campi specificate qui al posto di quelle date in un qualsiasi nodo Tipo upstream. Dopo avere selezionato questa opzione, specificare i campi rimanenti nella finestra di dialogo, che dipenderanno dall'utilizzo o meno del formato transazionale.

Se *non si utilizza* il formato transazionale, specificare:

- **Input.** Selezionare i campi input. Questa operazione è simile all'impostazione del ruolo di un campo su *Input* in un nodo Tipo.
- **Partizione.** Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello.

Se *si utilizza* il formato transazionale, specificare:

Utilizzo del formato transazionale. Utilizzare questa opzione se si desidera trasformare i dati da una riga per voce a una riga per caso.

La selezione di questa opzione modifica i controlli dei campi nella parte inferiore di questa finestra di dialogo:

Per il formato transazionale, specificare:

- **ID.** Selezionare un campo ID dall'elenco. Come campo ID è possibile utilizzare campi numerici o simbolici. Ogni valore univoco di questo campo deve indicare una specifica unità di analisi. In un'applicazione per analisi market basket, per esempio, ciascun ID potrebbe rappresentare un singolo cliente. Per un'applicazione per analisi di registri Web, ciascun ID potrebbe rappresentare un computer (per indirizzo IP) o un utente (per dati di login).
- **Contenuto.** Specificare il campo contenuto per il modello. Questo campo contiene l'elemento rilevante nella creazione di modelli di associazione.
- **Partizione.** Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello. Utilizzando un campione per creare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, nella scheda Campi di ogni nodo Modelli che utilizza il partizionamento è necessario selezionare un singolo campo partizione. Se è presente un'unica partizione, verrà utilizzata automaticamente quando si attiva il partizionamento. Si noti inoltre che per applicare la partizione selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.

Opzioni del modello Apriori

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Lunghezza massima regola. Imposta il numero massimo di precondizioni per le regole, un numero intero compreso tra 2 e 20. In questo modo è possibile ridurre la complessità delle regole. Se le regole sono troppo complesse o troppo specifiche, oppure se l'addestramento della propria regola sta richiedendo troppo tempo, provare a diminuire questo valore.

Confidenza minima. Imposta il livello di confidenza minimo, un valore tra 0 e 1. Le regole con una confidenza inferiore rispetto al criterio specificato vengono scartate.

Supporto minimo. Imposta la soglia di supporto minimo, un valore tra 0 e 1. Apriori scopre schemi con una frequenza superiore alla soglia di supporto minimo.

Oracle MDL (Lunghezza descrizione minima)

L'algoritmo di Oracle MDL (Lunghezza descrizione minima) facilita l'identificazione degli attributi che hanno il maggiore impatto su un attributo obiettivo. Spesso, sapere quali sono gli attributi più influenti aiuta a capire e gestire meglio il proprio business e consente di semplificare le attività di creazione dei modelli. Inoltre, questi attributi possono indicare i tipi di dati che è possibile aggiungere per espandere i modelli. L'algoritmo MDL può essere utilizzato, per esempio, per identificare gli attributi del processo più rilevanti per prevedere la qualità di un componente prodotto, i fattori associati al tasso di abbandono o i geni che potrebbero essere coinvolti nella cura di una malattia specifica.

Oracle MDL scarta i campi di input che considera privi di importanza nella previsione dell'obiettivo. Con i campi di input restanti crea quindi un nugget del modello grezzo associato a un modello Oracle, visibile in Oracle Data Miner. Quando si consulta il modello in Oracle Data Miner, viene visualizzato un grafico che mostra i campi di input rimanenti, in ordine di significatività ai fini della previsione dell'obiettivo.

Una classificazione negativa indica rumore. I campi di input classificati zero o meno di zero non contribuiscono alla previsione ed è consigliabile eliminarli dai dati.

Per visualizzare il grafico

1. Con il pulsante destro del mouse, fare clic sul nugget del modello grezzo nella palette Modelli e scegliere **Visualizza**.
2. Dalla finestra del modello, fare clic sul pulsante per lanciare Oracle Data Miner.
3. Connettersi a Oracle Data Miner. Consultare l'argomento “Oracle Data Miner” a pagina 48 per ulteriori informazioni.
4. Nel pannello di navigazione di Oracle Data Miner, espandere **Modelli** e quindi **Importanza attributo**.

5. Selezionare il modello Oracle desiderato (che avrà lo stesso nome del campo obiettivo specificato in IBM SPSS Modeler). Se non si conosce il modello corretto, selezionare la cartella Importanza attributo e cercare un modello in base alla data di creazione.

Opzioni del modello MDL

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Campo univoco. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. IBM SPSS Modeler impone il vincolo che questo campo chiave deve essere numerico.

Nota: questo campo è facoltativo per tutti i nodi Oracle tranne che Bayes adattivo Oracle, Oracle O-Cluster e Apriori Oracle.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Importanza attributo Oracle (AI)

Scopo dell'importanza attributo è individuare gli attributi dell'insieme di dati correlati al risultato, e in che misura influiscono sul risultato finale. Il nodo Importanza attributo Oracle analizza dati, individua schemi e prevede risultati con un livello di confidenza associato.

Opzioni modello AI

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Preparazione dati automatici. (solo 11*g*) Attiva (default) o disattiva la modalità di preparazione dei dati automatici di Oracle Data Mining. Se la casella è selezionata, ODM esegue automaticamente la trasformazione dati richiesti dall'algoritmo. Per ulteriori informazioni, vedere *Oracle Data Mining Concepts*.

Opzioni di selezione AI

La scheda Opzioni consente di specificare le impostazioni di default per la selezione o l'esclusione dei campi di input nel nugget del modello. In seguito è possibile aggiungere il modello a un flusso per selezionare un sottoinsieme di campi da utilizzare nelle successive operazioni di generazione dei modelli. In alternativa, è possibile ignorare queste impostazioni selezionando o deselezionando campi aggiuntivi nel browser dei modelli dopo aver generato il modello. Tuttavia, le impostazioni di default consentono di applicare il nugget del modello senza ulteriori modifiche, il che può rivelarsi particolarmente utile ai fini dello script.

Sono disponibili le seguenti opzioni:

Tutti i campi classificati. Seleziona i campi in base alla loro classificazione come *importante*, *marginale* o *non importante*. È possibile modificare l'etichetta di ogni classificazione nonché i valori di interruzione utilizzati per assegnare i record all'uno o all'altro rango.

Primi N campi. Seleziona i primi *N* campi in base all'importanza.

Importanza maggiore di. Seleziona tutti i campi con importanza maggiore del valore specificato.

Il campo obiettivo viene sempre mantenuto indipendentemente dalla selezione.

Scheda Modello del nugget del modello AI

La scheda Modello di un nugget del modello Oracle AI visualizza il rango e l'importanza di tutti gli input e consente di selezionare i campi da filtrare utilizzando le caselle di controllo nella colonna di sinistra. Quando si esegue il flusso, vengono mantenuti solo i campi selezionati insieme alla previsione dell'obiettivo. Gli altri campi di input vengono scartati. Le selezioni di default sono basate sulle opzioni specificate nel nodo Modelli, ma è possibile selezionare o deselezionare campi aggiuntivi, se necessario.

- Per ordinare l'elenco per rango, nome campo, importanza o in base a qualsiasi altra colonna visualizzata, fare clic sull'intestazione di colonna. In alternativa, selezionare l'elemento desiderato dall'elenco accanto al pulsante Ordina per e utilizzare le frecce su e giù per modificare la direzione dell'ordinamento.
- È possibile utilizzare la barra degli strumenti per selezionare o deselezionare tutti i campi e per accedere alla finestra di dialogo Seleziona campi, che consente di selezionare i campi per rango o per importanza. Per estendere la selezione è possibile anche premere il tasto Maiusc o Ctrl mentre si fa clic sui campi.
- I valori di soglia per la classificazione degli input come importante, marginale o non importante vengono visualizzati nella legenda sotto alla tabella. Questi valori sono specificati nel nodo Modelli.

Gestione dei modelli Oracle

I modelli Oracle vengono aggiunti alla palette Modelli esattamente come gli altri modelli IBM SPSS Modeler e possono essere utilizzati in modo sostanzialmente simile. Tuttavia, esistono alcune importanti differenze, dato che ogni modello Oracle creato in IBM SPSS Modeler fa riferimento a un modello archiviato in un server di database.

Scheda Server del nugget del modello Oracle

Se si crea un modello ODM tramite IBM SPSS Modeler, viene creato un modello in IBM SPSS Modeler e viene creato o sostituito un modello nel database Oracle. Un modello di IBM SPSS Modeler di questo tipo fa riferimento al contenuto di un modello di database archiviato in un server di database. IBM SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la **chiave del modello** generato sia nel modello di IBM SPSS Modeler che nel modello Oracle.

La stringa della chiave per ciascun modello Oracle viene visualizzata nella colonna *Informazioni sul modello* della finestra di dialogo Elenca modelli. La stringa della chiave di un modello di IBM SPSS Modeler viene invece visualizzata come **Chiave di modello** nella scheda Server di un modello IBM SPSS Modeler (se all'interno di un flusso).

È possibile utilizzare il pulsante Controllo della scheda Server di un nugget del modello per controllare che le chiavi di modello nei modelli IBM SPSS Modeler e Oracle corrispondano. Se in Oracle non è reperibile alcun modello con lo stesso nome o se le chiavi del modello non corrispondono, il modello Oracle è stato eliminato o ricreato dopo la creazione del modello di IBM SPSS Modeler.

Scheda Riepilogo del nugget del modello Oracle

La scheda Riepilogo di un nugget del modello visualizza le informazioni sul modello stesso (*Analisi*), i campi utilizzati nel modello (*Campi*), le impostazioni utilizzate durante la creazione del modello (*Impostazioni di creazione*) e l'addestramento del modello (*Riepilogo addestramento*).

Quando si sfoglia per la prima volta il nodo, i risultati della scheda Riepilogo sono compressi. Per visualizzare i risultati, utilizzare il controllo dell'espansore a sinistra di un elemento se si desidera visualizzare solo i risultati di tale elemento oppure fare clic sul pulsante **Espandi tutto** se si desidera

visualizzare tutti i risultati. Per nascondere i risultati, utilizzare il controllo dell'espansore di un elemento se si desidera nascondere solo i risultati di tale elemento oppure fare clic sul pulsante **Comprimi tutto** se si desidera nascondere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico. Se è stato eseguito un nodo Analisi collegato a questo nugget del modello, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi.

Campi. Elenca i campi utilizzati come obiettivi e come input nella creazione del modello.

Impostazioni di creazione. Contiene informazioni sulle impostazioni utilizzate nella creazione del modello.

Riepilogo addestramento. Mostra il tipo di modello, il flusso utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Scheda Impostazioni del nugget del modello Oracle

La scheda Impostazioni sul nugget del modello consente di ignorare l'impostazione di certe opzioni sul nodo Modelli a scopi di calcolo del punteggio.

Struttura ad albero delle decisioni Oracle

Utilizza costi di errata classificazione. Determina se utilizzare i costi di errata classificazione nel modello Struttura ad albero delle decisioni Oracle. Consultare l'argomento “Costi classificazione errata” a pagina 32 per ulteriori informazioni.

Identificativo regola. Se selezionata, aggiunge una colonna ID regola al modello Struttura ad albero delle decisioni Oracle. L'ID regola identifica il nodo nella struttura ad albero in cui viene eseguita una determinata suddivisione.

NMF Oracle

Visualizza tutte le funzioni. Se selezionata, consente di visualizzare ID e confidenza per tutte le funzionalità e non solo per la funzionalità migliore, nel modello NMF Oracle.

Elenco dei modelli Oracle

Il pulsante Elenca modelli Oracle Data Mining avvia una finestra di dialogo in cui sono elencati i modelli di database esistenti e da cui è possibile rimuovere i modelli. Questa finestra di dialogo può essere aperta dalla finestra di dialogo Applicazioni di supporto e dalle finestre di dialogo di creazione, visualizzazione e applicazione dei nodi correlati a ODM.

Di seguito sono riportate le informazioni visualizzate per ogni modello:

- **Nome modello.** Nome del modello, che viene utilizzato per ordinare l'elenco
- **Informazioni sul modello.** Informazioni chiave sul modello, inclusive di data e ora di creazione e nome della colonna obiettivo
- **Tipo di modello.** Nome dell'algoritmo di creazione del modello

Oracle Data Miner

Oracle Data Miner è l'interfaccia utente relativa a Oracle Data Mining (ODM) e sostituisce l'interfaccia utente IBM SPSS Modeler precedente. Oracle Data Miner è stato appositamente progettato per incrementare l'utilizzo corretto degli algoritmi ODM da parte degli analisti. Questi obiettivi vengono affrontati e risolti in diversi modi:

- Gli utenti necessitano di maggiore assistenza nell'applicazione di una metodologia che gestisce sia la preparazione dei dati sia la selezione degli algoritmi. Oracle Data Miner soddisfa questa esigenza fornendo delle specifiche attività di data mining che guidano gli utenti passo passo nell'utilizzo della metodologia corretta.
- Oracle Data Miner contiene un'euristica migliorata e ampliata nelle procedure guidate di creazione e trasformazione del modello, che consente di ridurre la possibilità di errori nella specifica delle impostazioni di trasformazione e di modello.

Definizione di una connessione Oracle Data Miner

1. Oracle Data Miner può essere avviato da tutti i nodi Applicazione e Creazione di Oracle e da qualsiasi finestra di dialogo di output mediante il pulsante **Avvia Oracle Data Miner**.

Figura 2. Pulsante Avvia Oracle Data Miner

2. La finestra di dialogo **Edit Connection** di Oracle Data Miner viene visualizzata all'utente prima che venga avviata l'applicazione esterna (a condizione che l'opzione Applicazione di supporto sia stata correttamente definita).

   Nota: questa finestra di dialogo viene visualizzata solo in assenza di un nome di connessione definito.

   - Indicare un nome per la connessione Data Miner e immettere le informazioni relative al server Oracle 10gR1 o 10gR2. Il server Oracle dovrebbe essere lo stesso server specificato in IBM SPSS Modeler.
3. La finestra di dialogo **Choose Connection** di Oracle Data Miner fornisce le opzioni per specificare il nome della connessione utilizzato, definito al punto precedente.

Per ulteriori informazioni sui requisiti, l'installazione e l'utilizzo di Oracle Data Miner, consultare la sezione Oracle Data Miner sul sito Web di Oracle.

Preparazione dei dati

Quando per la modellazione si utilizzano i modelli Naive Bayes, Bayes adattivo e SVM forniti con gli algoritmi Oracle Data Mining possono essere utili due tipi di preparazione dei dati:

- **Discretizzazione**, o conversione di campi di intervalli numerici continui in categorie per algoritmi che non possono accettare dati continui.
- **Normalizzazione**, o trasformazioni applicate a intervalli numerici in modo che abbiano medie e deviazioni standard simili.

Discretizzazione

Il nodo Discretizza IBM SPSS Modeleroffre un numero di tecniche per eseguire le operazioni di discretizzazione. Viene definita un'operazione di discretizzazione applicabile a uno o più campi. Se si esegue l'operazione di discretizzazione su un insieme di dati vengono create le soglie e consentita la creazione del nodo Ricava di IBM SPSS Modeler. L'operazione Ricava può essere convertita in SQL e applicata prima della creazione e del calcolo del punteggio del modello. Questo approccio crea una dipendenza tra il modello e il nodo Ricava che esegue la discretizzazione, ma consente di riutilizzare le specifiche di discretizzazione in diverse attività di modellazione.

Normalizzazione

I campi continui (intervallo numerico) utilizzati come input dei modelli SVM devono essere normalizzati prima della creazione del modello. Nel cado di modelli di regressione, la normalizzazione deve anche essere invertita per ricreare il punteggio dall'output del modello. Tra le impostazioni del modello SVM è

possibile scegliere **Punteggio Z**, **Min-Max** o **Nessuno**. I coefficienti di normalizzazione vengono creati da Oracle durante il processo di creazione del modello, per poi essere caricati in IBM SPSS Modeler e archiviati con il modello. In fase di applicazione, i coefficienti vengono convertiti in espressioni di derivazione IBM SPSS Modeler e utilizzati per preparare i dati per il calcolo del punteggio prima che siano passati al modello. In questo caso, la normalizzazione è strettamente associata all'attività di modellazione.

Esempi di Oracle Data Mining

È disponibile un'ampia gamma di flussi di esempio che illustrano l'utilizzo di ODM con IBM SPSS Modeler. Tali flussi sono disponibili nella cartella di installazione di IBM SPSS Modeler in *\Demos\Database_Modelling\Oracle Data Mining*.

Nota: è possibile accedere alla cartella Demos dal gruppo di programmi IBM SPSS Modeler nel menu Start di Windows.

I flussi nella seguente tabella possono essere utilizzati insieme in sequenza come un esempio del processo di mining del database, utilizzando l'algoritmo SVM (Support Vector Machine) fornito con Oracle Data Mining:

Tabella 4. Mining del database - flusso di esempio

Flusso	Descrizione
1_upload_data.str	Utilizzato per la pulitura e il caricamento di dati da un file flat nel database.
2_explore_data.str	Utilizzato come esempio di esplorazione dati con IBM SPSS Modeler.
3_build_model.str	Genera il modello utilizzando l'algoritmo nativo del database.
4_evaluate_model.str	Utilizzato come esempio di valutazione di modelli con IBM SPSS Modeler.
5_deploy_model.str	Esegue la distribuzione del modello ai fini del calcolo del punteggio in-database.

Nota: per eseguire l'esempio, i flussi deve essere eseguite in ordine. Inoltre, i nodi Origine e Modelli in ogni flusso devono essere aggiornati per far riferimento a un'origine dati valida per il database che si desidera utilizzare.

L'insieme di dati impiegato nei flussi di esempio concerne applicazioni relative alle carte di credito e presenta un problema di classificazione relativo a un insieme misto di predittori continui e categoriali. Per ulteriori informazioni su questo insieme di dati, vedere il file *crx.names* nella stessa cartella dei flussi di esempio.

Questo insieme di dati è disponibile in UCI Machine Learning Repository alla pagina *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*

Flusso di esempio: caricamento dati

Il primo flusso di esempio, *1_upload_data.str*, viene utilizzato per pulire e caricare dati da un file flat in Oracle.

Poiché Oracle Data Mining richiede un campo ID univoco, questo flusso iniziale utilizza un nodo Ricava per aggiungere un nuovo campo al dataset denominato *ID*, con i valori univoci 1,2,3, utilizzando la funzione IBM SPSS Modeler `@INDEX`.

Il nodo Riempimento viene utilizzato per la gestione dei valori mancanti e sostituisce i campi vuoti letti dal file di testo *crx.data* con valori *NULL*.

Flusso di esempio: esplorazione dati

Il secondo flusso di esempio, *2_explore_data.str*, viene utilizzato per illustrare l'uso di un nodo Esplora per acquisire una panoramica generale dei dati, comprese statistiche riassuntive e grafici.

Facendo doppio clic su un grafico nel report del nodo Esplora, si accede a una visualizzazione più dettagliata del grafico che consente un'esplorazione più approfondita di un dato campo.

Flusso di esempio: creazione modello

Il terzo flusso di esempio, *3_build_model.str*, illustra la creazione del modello in IBM SPSS Modeler. Fare doppio clic sul nodo origine del database (etichettato CREDIT) per specificare l'origine dati. Per specificare le impostazioni di creazione, fare doppio clic sul nodo di creazione (inizialmente etichettato CLASS e modificato in FIELD16 quando viene specificata l'origine dati).

Nella scheda Modello della finestra di dialogo:

1. Verificare che **ID** sia selezionato come campo Unico.
2. Verificare che **Lineare** sia selezionato come funzione Kernel e **Punteggio Z** come metodo di normalizzazione.

Flusso di esempio: valutazione modello

Il quarto flusso di esempio, *4_evaluate_model.str*, illustra i vantaggi associati all'utilizzo di IBM SPSS Modeler per la modellazione nel database. Una volta eseguito il modello, è possibile aggiungerlo nuovamente al flusso di dati e valutarlo con il supporto di un'ampia gamma di strumenti mirati disponibili in IBM SPSS Modeler.

Visualizzazione dei risultati della modellazione

Collegare un nodo Tabella al nugget del modello per esplorare i risultati. Il campo **$O-field16** mostra il valore previsto per *field16* in ciascun caso e il campo **$OC-field16** mostra il valore di confidenza per questa previsione.

Valutazione dei risultati della modellazione

È possibile utilizzare il nodo Analisi per creare una matrice di coincidenza che mostri lo schema di corrispondenze tra ogni campo previsto e il relativo campo obiettivo. Eseguire il nodo Analisi per visualizzare i risultati.

È possibile utilizzare il nodo Valutazione per creare un grafico dei profitti, progettato per mostrare i miglioramenti in termini di precisione realizzati dal modello. Eseguire il nodo Valutazione per visualizzare i risultati.

Flusso di esempio: Deployment del modello

Una volta raggiunto il livello di precisione del modello desiderato, è possibile eseguire la distribuzione del modello per consentirne l'utilizzo con applicazioni esterne o la ripubblicazione nel database. Nell'ultimo stream di esempio, *5_deploy_model.str*, i dati vengono letti dalla tabella `CREDITDATA`, quindi viene eseguito il calcolo del punteggio e, infine, i dati vengono pubblicati nella tabella `CREDITSCORES` mediante il nodo Publisher denominato *soluzione di distribuzione*.

Capitolo 5. Modellazione di database con IBM InfoSphere Warehouse

IBM InfoSphere Warehouse e IBM SPSS Modeler

IBM InfoSphere Warehouse (ISW) fornisce una famiglia di algoritmi di data mining integrati con DB2 RDBMS di IBM. IBM SPSS Modeler fornisce nodi che supportano l'integrazione dei seguenti algoritmi IBM:

- Strutture ad albero delle decisioni
- Regole di associazione
- Raggruppamento tramite cluster demografici
- Raggruppamento tramite cluster Kohonen
- Regole di sequenza
- Regressione trasformazione
- Regressione lineare
- Regressione polinomiale
- Naive Bayes
- Regressione logistica
- serie temporali

Per ulteriori informazioni su questi algoritmi, consultare la documentazione fornita con l'installazione di IBM InfoSphere Warehouse.

Requisiti per l'integrazione con IBM InfoSphere Warehouse

Di seguito sono elencate le condizioni che costituiscono i prerequisiti indispensabili per l'esecuzione della modellazione nel database con InfoSphere Warehouse Data Mining. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di database.

- Esecuzione di IBM SPSS Modeler in un'installazione di IBM SPSS Modeler Server su Windows o UNIX.
- IBM DB2 Data Warehouse Edition Versione 9.1

  o
- IBM InfoSphere Warehouse Versione 9.5 Enterprise Edition
- Un'origine dati ODBC per la connessione a DB2, come illustrato di seguito.

Nota: le funzionalità di modellazione nel database e ottimizzazione SQL richiedono che la connettività IBM SPSS Modeler Server venga abilitata sul computer IBM SPSS Modeler. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da IBM SPSS Modeler, e accedere a IBM SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu IBM SPSS Modeler.

Guida > **Informazioni su** > **Dettagli aggiuntivi**

Se la connettività è abilitata, l'opzione **Abilitazione server** viene visualizzata nella scheda Stato della licenza.

Attivazione dell'integrazione con IBM InfoSphere Warehouse

Per attivare l'integrazione di IBM SPSS Modeler con IBM InfoSphere Warehouse (ISW) Data Mining, sarà necessario configurare ISW e creare un'origine ODBC, attivare l'integrazione nella finestra di dialogo Applicazioni di supporto di IBM SPSS Modeler e abilitare la generazione e l'ottimizzazione SQL.

Configurazione di ISW

Per installare e configurare ISW seguire le istruzioni contenute nella guida all'installazione di *InfoSphere Warehouse*.

Creazione di un'origine ODBC per ISW

Per attivare la connessione tra ISW e IBM SPSS Modeler è necessario creare un nome di origine dati (DSN, Data Source Name) ODBC.

Per creare un DSN, è necessario avere una conoscenza di base delle origini dati e dei driver ODBC e disporre del supporto database in IBM SPSS Modeler.

Se IBM SPSS Modeler Server e IBM InfoSphere Warehouse Data Mining sono in esecuzione su host differenti, creare lo stesso DSN ODBC su ogni host. Assicurarsi di utilizzare lo stesso nome per il DSN su entrambi gli host.

1. Installare i driver ODBC. I driver sono disponibili sul disco di installazione di IBM SPSS Data Access Pack fornito con questa versione. Eseguire il file *setup.exe* per avviare il programma di installazione, e selezionare tutti i driver opportuni. Attenersi alle istruzioni visualizzate per installare i driver.
   a. Creare il DSN.

      Nota: La sequenza dei menu dipende dalla versione di Windows.

      - **Windows XP.** dal menu Start, scegliere **Pannello di controllo**. Fare doppio clic su **Strumenti di amministrazione** e doppio clic su **Origini dati (ODBC)**.
      - **Windows Vista.** Dal menu Start, scegliere **Pannello di controllo**, quindi **Strumenti di amministrazione**. Doppio clic su **Strumenti di amministrazione**, selezionare**Origini dati (ODBC)**, quindi fare clic su **Apri**.
      - **Windows 7.** dal menu Start, scegliere **pannello di controllo**, quindi**Sistema & Sicurezza**, quindi **Strumenti di amministrazione**. Selezionare**Origini dati (ODBC)**, then click **Open**.

   b. Fare clic sulla scheda **DSN di sistema**, quindi fare clic su **Aggiungi**.
2. Selezionare il driver **SPSS OEM 6.0 DB2 Wire Protocol**.
3. Fare clic su **Fine**.
4. Nella finestra di dialogo ODBC DB2 Wire Protocol Driver Setup:
   - Specificare un nome di origine dati.
   - Per l'indirizzo IP, indicare il nome host del server su cui è in esecuzione il sistema DB2 RDBMS.
   - Accettare il valore di default relativo alla porta TCP (50000).
   - Specificare il nome del database con il quale verrà stabilita la connessione.
5. Fare clic su **Verifica connessione**.
6. Nella finestra di dialogo Connessione a DB2 Wire Protocol immettere il nome utente e la password ricevuti dall'amministratore di database, quindi fare clic su **OK**.

   Verrà visualizzato il messaggio **Connessione effettuata**.

   Driver IBM DB2 ODBC. Se il driver ODBC in uso corrisponde al driver IBM DB2 ODBC, applicare la seguente procedura per creare un DSN ODBC:
7. In Amministratore origine dati ODBC fare clic sulla scheda **DSN di sistema**, quindi fare clic su **Aggiungi**.
8. Selezionare **IBM DB2 ODBC DRIVER**, quindi fare clic su **Fine**.
9. Nella finestra Aggiungi IBM DB2 ODBC DRIVER, immettere un nome origine dati e quindi per l'alias del database, fare clic su **Aggiungi**.
10. Nella finestra Impostazioni CLI/ODBC—<Nome origine dati> nella scheda Origine dati, immettere l'ID utente e la password fontiti all'utente dall'amministratore di database e quindi fare clic sulla scheda **TCP/IP** tab.

11. Nella scheda TCP/IP immettere:
    - Il nome del database al quale si desidera connettersi
    - Un nome alias di database (non più di otto caratteri)
    - Il nome host del server di database al quale si desidera connettersi
    - Il numero di porta per la connessione
12. Fare clic sulla scheda **Opzioni di sicurezza** e selezionare **Specifica le opzioni di sicurezza (facoltativo)** e quindi accettare l'opzione predefinita**Utilizza valore di autenticazione nella configurazione DBM del server**).
13. Fare clic sulla scheda **Origine dati**, quindi su **Connetti**.

Verrà visualizzato il messaggio **Connessione verificata**.

Configurazione di ODBC per il feedback (opzionale)

Per ricevere feedback da IBM InfoSphere Warehouse Data Mining durante la creazione del modello e attivare IBM SPSS Modeler per l'annullamento del processo di creazione, attenersi alla procedura riportata di seguito in modo da configurare l'origine dati ODBC creata nella sezione precedente. Si noti che questa procedura di configurazione consente a IBM SPSS Modeler di leggere i dati DB2 che non possono essere salvati nel database eseguendo contemporaneamente transazioni. Se si nutrono dubbi circa le implicazioni di questa modifica, si consiglia di consultare l'amministratore di database.

Driver SPSS OEM 6.0 DB2 Wire Protocol. Per il driver Connect ODBC, effettuare le seguenti operazioni:

1. Avviare Amministratore origine dati ODBC, selezionare l'origine dati creata nella sezione precedente e fare clic sul pulsante **Configura**.
2. Nella finestra di dialogo ODBC DB2 Wire Protocol Driver Setup fare clic sulla scheda **Avanzate**.
3. Impostare il livello di isolamento di default su **0-READ UNCOMMITED**, quindi fare clic su **OK**.

Driver IBM DB2 ODBC. Per il driver IBM DB2, effettuare le seguenti operazioni:

4. Avviare Amministratore origine dati ODBC, selezionare l'origine dati creata nella sezione precedente, quindi fare clic sul pulsante **Configura**.
5. Nella finestra di dialogo CLI/ODBC Settings fare clic sulla scheda **Impostazioni avanzate**, quindi sul pulsante **Aggiungi**.
6. Nella finestra di dialogo Add CLI/ODBC Parameter selezionare il parametro **TXNISOLATION**, quindi fare clic su **OK**.
7. Nella finestra di dialogo Isolation level selezionare **Read Uncommitted**, quindi fare clic su **OK**.
8. Nella finestra di dialogo CLI/ODBC Settings, fare clic su **OK** per concludere la configurazione.

Si noti che il feedback riportato da IBM InfoSphere Warehouse Data Mining viene visualizzato nel seguente formato:

```
<ITERATIONNO> / <PROGRESS> /
<KERNELPHASE>
```

dove:

- `<ITERATIONNO>` indica il numero del passaggio corrente sui dati, a partire da 1.
- `<PROGRESS>` indica lo stato di avanzamento dell'iterazione corrente sotto forma di un numero compreso tra 0.0 e 1.0.
- `<KERNELPHASE>` descrive la fase corrente dell'algoritmo di data mining.

Attivazione di IBM InfoSphere Warehouse Data Mining Integration in IBM SPSS Modeler

Per attivare IBM SPSS Modeler in modo da consentire l'utilizzo di DB2 con IBM InfoSphere Warehouse Data Mining, è necessario prima fornire alcune specifiche nella finestra di dialogo Applicazioni di supporto.

1. Dai menu di IBM SPSS Modeler scegliere:
   Strumenti > **Opzioni** > **Applicazione di supporto**
2. Fare clic sulla scheda **IBM InfoSphere Warehouse**.

Abilita integrazione InfoSphere Warehouse Data Mining. Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra IBM SPSS Modeler e aggiunge i nodi degli algoritmi di ISW Data Mining.

Connessione DB2. Specifica l'origine dati ODBC DB2 di default utilizzata per la creazione e l'archiviazione di modelli. Questa impostazione può essere sovrascritta nei singoli nodi di modelli generati e creazione del modello. Fare clic sul pulsante con i puntini di sospensione (**...**) per scegliere l'origine dati.

la connessione al database utilizzata a fini di modellazione può corrispondere o meno a quella impiegata per accedere ai dati. Per esempio, è possibile utilizzare un flusso che accede ai dati di un database DB2, li scarica in IBM SPSS Modeler per la pulitura o altre operazioni di modifica e, infine, li carica su un database DB2 differente per la modellazione. In alternativa, i dati originali possono risiedere in un file flat o in un'altra origine (non DB2) e occorrerà pertanto caricarli in DB2 per il processo di modellazione. In tutti i casi, se necessario, i dati verranno automaticamente caricati in una tabella temporanea creata nel database utilizzato per la modellazione.

Avvisa quando si sovrascrive un modello di integrazione InfoSphere Warehouse Data Mining. Selezionare questa opzione per assicurarsi che i modelli archiviati nel database non vengano sovrascritti da IBM SPSS Modeler senza preavviso.

Elenca i modelli InfoSphere Warehouse Data Mining. Consente di elencare ed eliminare i modelli archiviati in DB2. Consultare l'argomento “Elenco dei modelli in-database” a pagina 59 per ulteriori informazioni.

Abilita l'avvio InfoSphere Warehouse Data Mining Visualization. Se è stato installato il modulo Visualization, è necessario attivarlo qui per l'utilizzo di IBM SPSS Modeler.

Percorso per l'eseguibile Visualization La posizione del file eseguibile del modulo Visualization (se installato), per esempio *C:\Programmi\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

Directory plug-in di visualizzazione delle serie temporali Il percorso del plug-in Flash di visualizzazione delle serie temporali (se installato), per esempio *C:\Programmi\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash_2.2.1.v20091111_0915*.

Abilita opzioni avanzate InfoSphere Warehouse Data Mining. È possibile impostare un limite all'utilizzo di memoria su un algoritmo di mining nel database e specificare altre opzioni arbitrarie sotto forma di riga di comando per modelli specifici. La definizione del limite consente di controllare l'utilizzo di memoria e specificare un valore per l'opzione avanzata `-buf`. È possibile specificare in questa posizione anche altre opzioni avanzate sotto forma di riga di comando e passarle a IBM InfoSphere Warehouse Data Mining. Consultare l'argomento “Opzioni avanzate” a pagina 60 per ulteriori informazioni.

Verifica la versione Check InfoSphere Warehouse. Controlla la versione di IBM InfoSphere Warehouse in uso e visualizza un messaggio di errore se si tenta di utilizzare una funzione di data mining non supportata in quella versione.

Attivazione di generazione e ottimizzazione SQL

1. Dal menu IBM SPSS Modeler scegliere:
   Strumenti > **Proprietà flusso** > **Opzioni**
2. Fare clic sull'opzione **Ottimizzazione** nel riquadro di spostamento.

3. Confermare che l'opzione **Genera SQL** è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
4. Selezionare **Ottimizza generazione SQL** e **Ottimizza altre esecuzioni** (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Creazione di modelli con IBM InfoSphere Warehouse Data Mining

La creazione del modello di IBM InfoSphere Warehouse Data Mining richiede che l'insieme di dati addestramento sia posizionato in una tabella o visualizzazione all'interno del database DB2. Se i dati non sono ubicati in DB2 o devono essere elaborati in IBM SPSS Modeler come parte del processo di preparazione dei dati che non è possibile eseguire in DB2, tali dati vengono automaticamente caricati in una tabella temporanea di DB2 prima della creazione del modello.

Calcolo del punteggio e deployment del modello

Il calcolo del punteggio del modello avviene sempre all'interno di DB2 ed è sempre eseguito da IBM InfoSphere Warehouse Data Mining. Può essere necessario caricare l'insieme di dati in una tabella temporanea, qualora i dati vengano originati in IBM SPSS Modeler o debbano essere preparati all'interno dell'applicazione. In IBM SPSS Modeler, per i modelli di cluster, Struttura ad albero delle decisioni e Regressione viene generalmente fornita una sola previsione con la probabilità o la confidenza associata. È inoltre disponibile un'opzione utente per la visualizzazione delle confidenze per ogni possibile risultato (simile a quella della regressione logistica) che rappresenta un'opzione tempo punteggio ubicata all'interno della scheda Impostazioni del nugget del modello (la casella di controllo **Includi confidenze per tutte le classi**). Per i modelli di sequenza e associazione di IBM SPSS Modeler vengono forniti diversi valori. IBM SPSS Modeler può calcolare i modelli IBM InfoSphere Warehouse Data Mining da stream pubblicati per l'esecuzione mediante IBM SPSS Modeler Solution Publisher.

La seguente tabella descrive i campi che vengono generati dai modelli del calcolo del punteggio.

Tabella 5. Campi di calcolo del punteggio dei modelli

Tipo di modello	Colonne punteggio	Significato
Strutture ad albero delle decisioni	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.
	$IC-*valore1*, ..., $IC-*valoreN*	(facoltativo) Confidenza di ciascuno dei possibili valori *N* per il *campo*.
Regressione	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.
Raggruppamento tramite cluster	$I-*nome_modello*	Migliore assegnazione cluster per il record di input.
	$IC-*nome_modello*	Confidenza della migliore assegnazione cluster per il record di input.
Associazione	$I-*nome_modello*	Identificatore della regola corrispondente.
	$IH-*nome_modello*	Elemento dell'intestazione.
	$IHN-*nome_modello*	Nome dell'elemento dell'intestazione.
	$IS-*nome_modello*	Valore di supporto della regola corrispondente.
	$IC-*nome_modello*	Valore di confidenza della regola corrispondente.

Tabella 5. Campi di calcolo del punteggio dei modelli (Continua)

Tipo di modello	Colonne punteggio	Significato
	$IL-*nome_modello*	Valore di guadagno cumulativo della regola corrispondente.
	$IMB-*nome_modello*	Numero di elementi del corpo o di serie di elementi del corpo corrispondenti (dal momento che tutti gli elementi o le serie di elementi del corpo devono corrispondere a questo numero, esso è uguale al numero di elementi o di serie di elementi del corpo).
Sequenza	$I-*nome_modello*	Identificatore della regola corrispondente
	$IH-*nome_modello*	Serie di elementi dell'intestazione della regola corrispondente
	$IHN-*nome_modello*	Nomi degli elementi della serie di elementi dell'intestazione della regola corrispondente
	$IS-*nome_modello*	Valore di supporto della regola corrispondente
	$IC-*nome_modello*	Valore di confidenza della regola corrispondente
	$IL-*nome_modello*	Valore di guadagno cumulativo della regola corrispondente
	$IMB-*nome_modello*	Numero di elementi del corpo o di serie di elementi del corpo corrispondenti (dal momento che tutti gli elementi o le serie di elementi del corpo devono corrispondere a questo numero, esso è uguale al numero di elementi o di serie di elementi del corpo)
Naive Bayes	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.
Regressione logistica	$I-*campo*	Previsione migliore per il *campo*.
	$IC-*campo*	Confidenza della previsione migliore per il *campo*.

Gestione dei modelli DB2

La creazione di un modello di IBM InfoSphere Warehouse Data Mining tramite IBM SPSS Modeler comporta la creazione di un modello in IBM SPSS Modeler e la creazione o la sostituzione di un modello nel database DB2. Il modello IBM SPSS Modeler di questo tipo fa riferimento al contenuto di un modello di database archiviato su un server di database. IBM SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello DB2 che nel modello di IBM SPSS Modeler.

La stringa della chiave di ogni modello DB2 viene visualizzata nella colonna *Informazioni sul modello* all'interno della finestra di dialogo Elenco dei modelli in-database. La stringa della chiave di un modello di IBM SPSS Modeler viene visualizzata come Chiave di modello nella scheda Server di un modello di IBM SPSS Modeler (se all'interno di un flusso).

È possibile utilizzare il pulsante Controllo per verificare la corrispondenza delle chiavi nel modello DB2 e in quello di IBM SPSS Modeler. Se in DB2 non è reperibile alcun modello con lo stesso nome o se le chiavi del modello non corrispondono, il modello DB2 è stato eliminato o ricreato dopo la creazione del modello di IBM SPSS Modeler. Consultare l'argomento “Scheda Server del nugget del modello ISW” a pagina 75 per ulteriori informazioni.

Elenco dei modelli in-database

IBM SPSS Modeler fornisce una finestra di dialogo per elencare i modelli che sono archiviati in IBM InfoSphere Warehouse Data Mining e consente l'eliminazione di modelli. Si può accedere a questa finestra di dialogo dalle applicazioni di supporto IBM e dalle finestre di dialogo Crea, Visualizza e Applica per i nodi correlati a IBM InfoSphere Warehouse Data Mining. Di seguito sono riportate le informazioni visualizzate per ogni modello:

- Nome del modello (utilizzato per ordinare l'elenco).
- Informazioni sul modello (informazioni sulla chiave di modello, da una chiave casuale che viene generata quando IBM SPSS Modeler crea il modello).
- Tipo di modello (la tabella DB2 in cui IBM InfoSphere Warehouse Data Mining ha archiviato il modello).

Visualizzazione dei modelli

Lo strumento visualizzatore è l'unico modo per sfogliare i modelli di InfoSphere Warehouse Data Mining. Lo strumento può essere installato facoltativamente con InfoSphere Warehouse Data Mining. Consultare l'argomento “Attivazione dell'integrazione con IBM InfoSphere Warehouse” a pagina 53 per ulteriori informazioni.

- Fare clic su **Visualizza** per avviare lo strumento visualizzatore. Ciò che viene visualizzato dallo strumento dipende dal tipo di nodo generato. Ad esempio, lo strumento visualizzatore restituirà una vista Classi previste quando viene avviato da un nugget del modello Struttura ad albero delle decisioni ISW
- Fare clic su **Risultati del test** (solo Strutture ad albero delle decisioni e Sequenza) per avviare lo strumento visualizzatore e visualizzare la qualità globale del modello generato.

Esportazione di modelli e generazione di nodi

È possibile eseguire importazioni ed esportazioni del PMML nei modelli di IBM InfoSphere Warehouse Data Mining. Il PMML che viene esportato è quello originale generato da IBM InfoSphere Warehouse Data Mining. La funzione di esportazione restituisce il modello in formato PMML.

È possibile esportare il riepilogo e la struttura di un modello in file formato testo e HTML, nonché generare i nodi Filtro, Seleziona e Nuovo Campo appropriati laddove necessario. Per ulteriori informazioni, consultare "Esportazione di modelli" *IBM SPSS Modeler Manuale dell'utente*.

Impostazioni dei nodi comuni a tutti gli algoritmi

Le seguenti impostazioni sono valide per molti degli algoritmi di IBM InfoSphere Warehouse Data Mining:

Obiettivo e predittori. È possibile specificare obiettivo e predittori utilizzando il nodo Tipo oppure manualmente mediante la scheda Campi del nodo builder del modello, come è prassi standard in IBM SPSS Modeler.

Origine dati ODBC. Consente all'utente di sovrascrivere l'origine dati ODBC predefinita per il modello corrente. (L'impostazione di default è specificata nella finestra di dialogo Applicazioni di supporto. Consultare l'argomento “Attivazione dell'integrazione con IBM InfoSphere Warehouse” a pagina 53 per ulteriori informazioni.)

Opzioni della scheda Server ISW

È possibile specificare la connessione DB2 utilizzata per il caricamento dei dati per la modellazione. Se necessario, inoltre, nella scheda Server è possibile selezionare una connessione specifica per ogni nodo Modelli che sovrascriva la connessione DB2 di default indicata nella finestra di dialogo Applicazioni di supporto. Consultare l'argomento "Attivazione dell'integrazione con IBM InfoSphere Warehouse" a pagina 53 per ulteriori informazioni.

La connessione utilizzata per la modellazione può corrispondere o meno a quella impiegata nel nodo origine di un flusso. Per esempio, è possibile utilizzare un flusso che accede ai dati di un database DB2, li scarica in IBM SPSS Modeler per la pulitura o altre operazioni di modifica e, infine, li carica su un database DB2 differente per la modellazione.

Il nome dell'origine dati ODBC è incorporato in ogni flusso di IBM SPSS Modeler. Se un flusso creato su un determinato host viene eseguito su un host differente, il nome dell'origine dati deve essere lo stesso su entrambi gli host. In alternativa, è possibile selezionare un'origine dati differente nella scheda Server all'interno di ogni nodo Modelli o di input.

Per ricevere feedback durante la creazione di un modello, utilizzare le seguenti opzioni:

- **Abilita feedback.** Selezionare questa opzione per ricevere feedback durante la creazione di un modello (per default, l'opzione è disattivata).
- **Intervallo di feedback (in secondi).** Specificare con quale frequenza IBM SPSS Modeler recupera feedback sullo stato di avanzamento della creazione dei modelli.

Abilita opzioni avanzate InfoSphere Warehouse Data Mining. Selezionare questa opzione per abilitare il pulsante **Opzioni avanzate**, che consente di specificare un numero di opzioni avanzate, tra cui un limite di memoria e un SQL personalizzato. Consultare l'argomento "Opzioni avanzate" per ulteriori informazioni.

Nella scheda Server di un nodo generato è disponibile un'opzione che consente di eseguire controlli di uniformità grazie all'archiviazione di una stringa identica contenente la chiave del modello generato sia nel modello DB2 che in quello di IBM SPSS Modeler. Consultare l'argomento "Scheda Server del nugget del modello ISW" a pagina 75 per ulteriori informazioni.

Opzioni avanzate

La scheda Server di tutti gli algoritmi comprende una casella di controllo per l'attivazione delle opzioni avanzate per ISW Modeling. Quando si fa clic sul pulsante **Opzioni avanzate**, viene visualizzata la finestra di dialogo Opzioni avanzate di ISW, che contiene una serie di opzioni per:

- Limite memoria.
- Altre opzioni avanzate.
- SQL personalizzato data mining.
- SQL personalizzato dati logici.
- SQL personalizzato impostazioni di mining.

Limite di memoria. Limita l'utilizzo di memoria di un algoritmo per la creazione di modelli. Si noti che l'opzione avanzata standard imposta un limite sul numero di valori discreti nei dati categoriali.

Altre opzioni avanzate. Consente di definire opzioni avanzate arbitrarie sotto forma di riga di comando per soluzioni o modelli specifici. Le specifiche possono variare in base al tipo di implementazione o soluzione. È possibile estendere manualmente l'SQL generato da IBM SPSS Modeler in modo da definire un'attività di creazione modelli.

SQL personalizzato data mining. È possibile aggiungere chiamate ai metodi per modificare l'oggetto `DM_MiningData`. Per esempio, se si immette il seguente codice SQL, ai dati utilizzati nella creazione del modello verrà aggiunto un filtro basato su un campo denominato *Partizione*:

```
..DM_setWhereClause('"Partition" = 1')
```

SQL personalizzato dati logici. È possibile aggiungere chiamate ai metodi per modificare l'oggetto `DM_LogicalDataSpec`. Per esempio, il seguente codice SQL rimuove un campo dall'insieme di campi utilizzato per la creazione del modello:

```
..DM_remDataSpecFld('field6')
```

SQL personalizzato impostazioni di mining. È possibile aggiungere chiamate ai metodi per modificare l'oggetto `DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings`. Per esempio, se si immette il seguente codice SQL, IBM InfoSphere Warehouse Data Mining imposterà il campo *Partizione* su attivo (il che significa che verrà sempre incluso nel modello risultante):

```
..DM_setFldUsageType('Partition',1)
```

Opzioni dei costi di ISW

Nella scheda Costi è possibile modificare i costi di errata classificazione, in maniera da specificare l'importanza relativa dei diversi tipi di errori di previsione.

In alcuni contesti, certi tipi di errori rappresentano un costo maggiore rispetto ad altri. Potrebbe essere più costoso, per esempio, classificare come a basso rischio chi richiede un fido ad alto rischio (un tipo di errore) di quanto non lo sia classificare come ad alto rischio chi chiede un fido a basso rischio (un tipo di errore diverso). I costi di errata classificazione permettono di specificare l'importanza relativa dei diversi tipi di errori di previsione.

Essenzialmente i costi di errata classificazione sono pesi applicati a risultati specifici. Tali pesi vengono fattorizzati nel modello e possono realmente modificare la previsione (come modo per proteggersi da errori che gravano sui costi).

Ad eccezione dei modelli C5.0, i costi di errata classificazione non sono applicati quando si calcola il punteggio di un modello e non vengono presi in considerazione quando si classificano o confrontano modelli con il nodo Classificatore automatico, un nodo Analisi o un grafico di valutazione. È possibile che un modello che include costi non generi un numero inferiore di errori rispetto a uno che non li include e che non si classifichi meglio in termini di precisione complessiva, ma è probabile che fornisca prestazioni migliori in termini pratici poiché include una distorsione incorporata a favore degli errori *meno costosi*.

La matrice dei costi mostra il costo per ciascuna possibile combinazione di categoria prevista e categoria effettiva. Per default, tutti i costi di errata classificazione sono impostati su 1.0. Per immettere valori di costi personalizzati, selezionare **Utilizza costi di errata classificazione** e immettere i valori personalizzati nella matrice dei costi.

Per cambiare un costo di errata classificazione, selezionare la cella corrispondente alla combinazione desiderata di valori previsti e valori effettivi, eliminare il contenuto esistente della cella e immettere il costo desiderato. I costi non sono simmetrici automaticamente. Se si imposta, per esempio, il costo dell'errata classificazione di *A* come *B* su 2.0 e non viene esplicitamente cambiato il costo dell'errata classificazione di *B* come *A*, esso manterrà il valore di default 1.0.

Struttura ad albero delle decisioni ISW

I modelli Struttura ad albero delle decisioni consentono di sviluppare sistemi di classificazione in grado di prevedere o classificare le osservazioni future in base a un insieme di regole decisionali. Se i dati sono divisi in classi di interesse (per esempio, prestiti a basso vs. alto rischio, sottoscrittore vs. non sottoscrittore, votanti vs. non votanti, oppure tipi di batteri), è possibile utilizzare i dati per generare regole da utilizzare per classificare casi precedenti e nuovi con la massima precisione. Per esempio, è possibile creare una struttura ad albero che classifica il rischio sul credito o l'intento di acquisto in base all'età e ad altri fattori.

L'algoritmo per strutture ad albero delle decisioni ISW crea strutture ad albero di classificazione su dati di input categoriali. La struttura ad albero delle decisioni risultante è binaria. Per la creazione del modello, è possibile applicare diverse impostazioni, compresi i costi di errata classificazione.

Lo strumento ISW Visualizer è l'unico modo per sfogliare i modelli di IBM InfoSphere Warehouse Data Mining.

Opzioni della scheda Modello per il nodo Struttura ad albero delle decisioni ISW

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se si definisce un campo partizione, selezionare l'opzione **Utilizza dati partizionati**.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test IBM InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici guadagno cumulativo e così via.

Profondità massima della struttura ad albero. È possibile specificare la profondità massima della struttura ad albero. In tal modo, si stabilisce che la profondità della struttura ad albero non potrà superare il numero specificato di livelli. Se questa opzione non viene selezionata, non verrà applicato alcun limite. Per evitare modelli eccessivamente complessi, si sconsiglia, se non in caso di stretta necessità, di definire un valore superiore a 5.

Opzioni avanzate Struttura ad albero delle decisioni ISW

Massima purezza Questa opzione imposta la purezza massima per i nodi interni. Qualora, in seguito alla suddivisione di un nodo, uno dei figli superi la misura di purezza definita (se, per esempio, oltre il 90% dei casi rientra in una categoria specificata), il nodo non verrà suddiviso.

Numero di casi minimi per il nodo interno. Se il processo di suddivisione comporta la creazione di un nodo con un numero di casi inferiore al minimo specificato, il nodo non verrà suddiviso.

Associazione ISW

Il nodo Associazione ISW può essere utilizzato per trovare delle regole di associazione tra gli elementi presenti in un insieme di gruppi. Le regole di associazione consentono di associare una conclusione specifica, per esempio l'acquisto di un particolare prodotto, a un insieme di condizioni, come l'acquisto di numerosi altri prodotti.

È possibile decidere di includere o escludere le regole di associazione dal modello specificando dei **vincoli**. Se si decide di includere un determinato campo di input, vengono incluse nel modello le regole di associazione contenenti almeno uno degli elementi specificati. Se si esclude un campo di input, le regole di associazione che contengono uno qualsiasi degli elementi specificati vengono eliminate dai risultati.

Gli algoritmi di associazione e sequenza ISW possono utilizzare **tassomanie**. Queste ultime mappano singoli valori a concetti di livello superiore. Per esempio, penne e matite possono essere mappate a una categoria cancelleria.

Le regole di associazione hanno un solo conseguente (la conclusione) e più antecedenti (l'insieme di condizioni). Di seguito è riportato un esempio:

```
[Pane, Marmellata] ⌂ [Burro]
[Pane, Marmellata]
⌂ [Margarina]
```

Qui, `Pane` e `Marmellata` sono gli antecedenti (detti anche **corpo della regola**) e `Burro` o `Margarina` sono altrettanti esempi di conseguenti (detti anche **intestazione della regola**). La prima regola indica che una persona che ha acquistato pane e marmellata ha acquistato contemporaneamente anche del burro. La seconda regola identifica un cliente che al momento dell'acquisto della stessa combinazione (pane e marmellata) ha acquistato anche margarina nello stesso negozio.

Lo strumento visualizzatore è l'unico modo per sfogliare i modelli di IBM InfoSphere Warehouse Data Mining.

Opzioni dei campi Associazione ISW

Nella scheda Campi si indicano i campi da utilizzare nella creazione del modello.

Per poter generare un modello, è necessario prima specificare i campi da utilizzare come obiettivi e come input. Con alcune eccezioni, tutti i campi Modelli utilizzano le informazioni sui campi di un nodo Tipo upstream. Oltre all'impostazione di default relativa all'utilizzo del nodo Tipo per la selezione dei campi obiettivo e di input, in questa scheda è possibile modificare soltanto l'opzione relativa al layout di tabella per i dati non transazionali.

Utilizzare le impostazioni del nodo tipologia. Questa opzione specifica l'utilizzo delle informazioni sui campi da un nodo Tipo upstream. Questa è l'opzione di default.

Utilizzare le impostazioni personalizzate. Questa opzione specifica l'utilizzo delle informazioni sui campi immessi qui al posto di quelle date in un qualsiasi nodo Tipo upstream. Dopo avere selezionato questa opzione, specificare i campi nell'area sottostante come richiesto.

Utilizzo del formato transazionale. Selezionare la casella di controllo se i dati di origine sono in **formato transazionale**. I record in questo formato hanno due campi, uno per l'ID e uno per il contenuto. Ogni record rappresenta una singola transazione o elemento, e gli elementi associati sono collegati poiché hanno lo stesso ID. Deselezionare questa casella se i dati sono in **formato tabulare**, in cui gli elementi sono rappresentati da flag separati e ogni campo flag rappresenta la presenza o l'assenza di un elemento specifico, mentre ogni record rappresenta un insieme completo di elementi associati.

- **ID.** Per i dati transazionali, selezionare un campo ID dall'elenco. Come campo ID è possibile utilizzare campi numerici o simbolici. Ogni valore univoco di questo campo deve indicare una specifica unità di analisi. In un'applicazione per analisi market basket, per esempio, ciascun ID potrebbe rappresentare un singolo cliente. Per un'applicazione per analisi di registri Web, ciascun ID potrebbe rappresentare un computer (per indirizzo IP) o un utente (per dati di login).
- **Contenuto.** Specificare il campo o i campi contenuto per il modello. Questi campi contengono gli elementi rilevanti nella creazione di modelli di associazione. È possibile specificare un unico campo nominale quando i dati sono in formato transazionale.

Utilizzo del formato tabulare. Deselezionare la casella di controllo **Utilizza formato transazionale** se i dati di origine sono in formato tabulare.

- **Input.** Selezionare il campo o i campi di input. Questa operazione è simile all'impostazione del ruolo di un campo su *Input* in un nodo Tipo.
- **Partizione.** Questo campo consente di specificare un campo utilizzato per partizionare i dati in campioni distinti per le fasi di addestramento, di test e di convalida della creazione del modello. Utilizzando un campione per generare il modello e un altro campione per sottoporlo a verifica, è possibile ottenere un'indicazione valida del modo in cui il modello potrà essere esteso a insiemi di dati di dimensioni maggiori, simili ai dati correnti. Se sono stati definiti più campi partizione utilizzando nodi Tipo o Partizione, nella scheda Campi di ogni nodo Modelli che utilizza il partizionamento è necessario selezionare un singolo campo partizione. Se è presente un'unica partizione, verrà utilizzata

automaticamente quando si attiva il partizionamento. Si noti inoltre che per applicare la partizione selezionata nell'analisi, è necessario che il partizionamento sia attivato anche nella scheda Opzioni modello relativa al nodo. La disattivazione di questa opzione consente di disattivare il partizionamento senza dover modificare le impostazioni dei campi.

Layout di tabella per dati non transazionali. Per i dati in formato tabellare, è possibile scegliere un layout di tabella standard (opzione di default) o layout a lunghezza degli elementi limitata.

Nel layout di default, il numero di colonne è determinato dal numero totale di elementi associati.

Tabella 6. Layout di tabella di default.

ID gruppo	Conto corrente	Conto di risparmio	Carta di credito	Prestito	Deposito titoli
Smith	Y	Y	Y	-	-
Jackson	Y	-	Y	Y	Y
Douglas	Y	-	-	-	Y

Nel layout a lunghezza degli elementi limitata, il numero di colonne è determinato dal numero più alto di elementi associati in una qualsiasi delle righe.

Tabella 7. Layout di tabella a lunghezza degli elementi limitata.

ID gruppo	Elemento1	Elemento2	Elemento3	Elemento4
Smith	conto corrente	conto di risparmio	carta di credito	-
Jackson	conto corrente	carta di credito	prestito	deposito titoli
Douglas	conto corrente	deposito titoli	-	-

Opzioni della scheda Modello per il nodo Associazione ISW

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Supporto minimo per la regola (%). Livello di supporto minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di supporto. Il valore è calcolato come A/B*100, dove A è il numero dei gruppi che contengono tutti gli elementi presenti nella regola e B è il numero totale dei gruppi considerati. Per prendere in esame le associazioni o le sequenze più comuni, aumentare il valore di questa impostazione.

Confidenza minima per la regola Livello di confidenza minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di confidenza. Il valore è calcolato come m/n*100, dove m è il numero dei gruppi che contengono l'intestazione della regola (conseguente) unita al corpo della regola (antecedente), e n è il numero dei gruppi che contengono il corpo della regola. Se si ottengono associazioni o sequenze in numero eccessivo o non interessanti, provare ad aumentare il valore di questa impostazione. Se le associazioni o le sequenze ottenute sono troppo poche, provare a diminuire il valore di questa impostazione.

Dimensione massima della regola. Numero massimo di elementi consentito in una regola, compreso l'elemento conseguente. Se le associazioni o le sequenze interessanti sono relativamente brevi, è possibile diminuire il valore di questa impostazione per accelerare la creazione dell'insieme.

Nota: viene calcolato il punteggio solo per i nodi con il formato di input transazionale; le tabelle di verità (dati in formato tabellare) restano grezze.

Opzioni della scheda Opzioni avanzate per il nodo Associazione ISW

Sulla scheda Livello avanzato del nodo Associazione è possibile indicare le regole di associazione da includere o da escludere dai risultati. Se si decide di includere elementi specificati, vengono incluse nel modello le regole contenenti almeno uno degli elementi specificati. Se si decide di escludere gli elementi specificati, le regole che contengono uno qualsiasi degli elementi specificati vengono eliminate dai risultati.

Quando è selezionata l'opzione **Utilizza vincoli elemento** qualsiasi elemento aggiunto all'elenco dei vincoli verrà incluso o escluso dai risultati a seconda dell'impostazione per **Tipo di vincolo**

Tipo di vincolo. Decidere se includere o escludere dai risultati le regole di associazione che contengono gli elementi specificati.

Modifica vincolo. Per aggiungere un elemento all'elenco degli elementi vincolati, selezionarlo nell'elenco Elementi e fare clic sul pulsante freccia destra.

Opzioni della scheda Tassonomia per ISW

Gli algoritmi di associazione e sequenza ISW possono utilizzare **tassonomie**. Queste ultime mappano singoli valori a concetti di livello superiore. Per esempio, penne e matite possono essere mappate a una categoria cancelleria.

Nella scheda Tassonomia è possibile definire mappe di categorie per esprimere tassonomie sui dati. Per esempio, una tassonomia può creare due categorie (`Prodotti alimentari di base` e `Prodotti di lusso`) e quindi assegnare gli elementi fondamentali a ognuna di esse. Per esempio, `vino` è assegnato a `Prodotti di lusso` e `pane` è assegnato a `Prodotti alimentari di base`. La tassonomia presenta una struttura padre-figlio come illustrato nella seguente tabella.

Tabella 8. Esempio di struttura di tassonomia

Figlio	Padre
vino	Prodotti di lusso
pane	Prodotti alimentari di base

Utilizzando questa tassonomia, è possibile creare un modello di associazione o di sequenza comprendente regole che coinvolgono sia le categorie che gli elementi fondamentali.

Nota: per attivare le opzioni in questa scheda, i dati di origine devono essere nel formato transazionale ed è necessario selezionare **Utilizza formato transazionale** nella scheda **Campi**, quindi selezionare **Utilizza tassonomia** in questa scheda.

Nome tabella. Consente di specificare il nome della tabella DB2 in cui archiviare i dettagli relativi alla tassonomia.

Colonna Figlio. Consente di specificare il nome della colonna figlio nella tabella di tassonomia. Tale colonna contiene i nomi di elemento o i nomi di categoria.

Colonna Padre. Consente di specificare il nome della colonna padre nella tabella di tassonomia. Tale colonna contiene i nomi di categoria.

Carica dettagli nella tabella. Selezionare questa opzione se le informazioni sulla tassonomia archiviate in IBM SPSS Modeler devono essere caricate nella tabella di tassonomia alla creazione del modello. Si noti

che se già esiste una tabella di tassonomia, tale tabella verrà eliminata. Le informazioni di tassonomia vengono memorizzate con il nodo di creazione del modello e possono essere modificate utilizzando i pulsanti Modifica categorie e Modifica tassonomia.

Editor di categorie

La finestra di dialogo Modifica categorie consente di aggiungere o di eliminare categorie da un elenco ordinato.

Per aggiungere una categoria, digitarne il nome nel campo **Nuova categoria** e fare clic sul pulsante freccia per spostarla nell'elenco **Categorie**.

Per rimuovere una categoria, selezionarla nell'elenco **Categorie** e fare clic sul pulsante Elimina adiacente.

Editor di tassonomia

La finestra di dialogo Modifica tassonomia consente di combinare l'insieme di elementi fondamentali definiti nei dati e l'insieme di categorie per creare una tassonomia. Per aggiungere nuove voci alla tassonomia, selezionare uno o più elementi o una o più categorie dall'elenco a sinistra e una o più categorie dall'elenco a destra, quindi fare clic sul pulsante con la freccia. Si noti che, qualora eventuali aggiunte alla tassonomia producano un conflitto (per esempio, se si specifica sia `cat1 -> cat2` che l'opposto, `cat2 -> cat1`), tali aggiunte non verranno effettuate.

Sequenza ISW

Il nodo Sequenza consente di individuare gli schemi nei dati sequenziali o basati su valori temporali, nel formato `pane -> formaggio`. Gli elementi di una sequenza sono **serie di elementi** che costituiscono una singola transazione. Per esempio, se una persona si reca in negozio e compra pane e latte e dopo alcuni giorni torna per comprare del formaggio, la sua attività di acquisto può essere rappresentata come due serie di elementi. La prima serie di elementi contiene il pane e il latte e il secondo contiene il formaggio. Per **sequenza** si intende un elenco di serie di elementi che tendono a ricorrere secondo un ordine prevedibile. Mediante il nodo Sequenza è possibile rilevare sequenze frequenti e creare un nodo di modello generato utilizzabile per elaborare previsioni.

È possibile usare la funzione mining Regole di sequenza in diverse aree di business. Per esempio, nel settore delle vendite al dettaglio è possibile trovare insiemi di acquisti tipici. Questi insiemi mostrano le diverse combinazioni di clienti, prodotti e ora dell'acquisto. Mediante queste informazioni, è possibile identificare i clienti potenziali di un prodotto che non hanno ancora acquistato il prodotto. Inoltre, è possibile offrire prodotti ai clienti potenziali nei tempi previsti.

Una sequenza rappresenta una serie ordinata di serie di elementi. Le sequenze contengono i seguenti livelli di raggruppamento:

- Gli eventi che accadono simultaneamente formano un'unica transazione o una serie di elementi.
- Ogni elemento o serie di elementi appartiene a un gruppo di transazioni. Per esempio, un articolo acquistato appartiene a un cliente, un clic su una pagina specifica appartiene a un utente Web, un componente appartiene a un'automobile prodotta. Diverse serie di elementi che avvengono in momenti diversi e appartengono allo stesso gruppo di transazioni formano una sequenza.

Opzioni della scheda Modello per il nodo Sequenza ISW

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Minimo supporto per la regola (%). Livello di supporto minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di supporto. Il valore è calcolato come A/B*100, dove A è il numero dei gruppi che contengono tutti gli elementi presenti nella regola e B è il numero totale dei gruppi considerati. Per prendere in esame le associazioni o le sequenze più comuni, aumentare il valore di questa impostazione.

Minima confidenza per la regola (%). Livello di confidenza minimo per le regole di associazione o di sequenza. Vengono incluse nel modello solo le regole che raggiungono almeno questo livello di confidenza. Il valore è calcolato come *m*/*n**100, dove *m* è il numero dei gruppi che contengono l'intestazione della regola (conseguente) unita al corpo della regola (antecedente), e *n* è il numero dei gruppi che contengono il corpo della regola. Se si ottengono associazioni o sequenze in numero eccessivo o non interessanti, provare ad aumentare il valore di questa impostazione. Se le associazioni o le sequenze ottenute sono troppo poche, provare a diminuire il valore di questa impostazione.

Dimensione massima della regola. Numero massimo di elementi consentito in una regola, compreso l'elemento conseguente. Se le associazioni o le sequenze interessanti sono relativamente brevi, è possibile diminuire il valore di questa impostazione per accelerare la creazione dell'insieme.

Nota: viene calcolato il punteggio solo per i nodi con il formato di input transazionale; le tabelle di verità (dati in formato tabellare) restano grezze.

Opzioni della scheda Opzioni avanzate per il nodo Sequenza ISW

È possibile specificare le regole di sequenza da includere o da escludere dai risultati. Se si decide di includere elementi specificati, vengono incluse nel modello le regole contenenti almeno uno degli elementi specificati. Se si decide di escludere gli elementi specificati, le regole che contengono uno qualsiasi degli elementi specificati vengono eliminate dai risultati.

Quando è selezionata l'opzione **Utilizza vincoli elemento** , qualsiasi elemento aggiunto all'elenco dei vincoli verrà incluso o escluso dai risultati a seconda dell'impostazione per il **Tipo di vincolo**

tipo di vincolo. Decidere se includere o escludere dai risultati le regole di associazione che contengono gli elementi specificati.

Modifica vincoli. Per aggiungere un elemento all'elenco degli elementi vincolati, selezionarlo nell'elenco Elementi e fare clic sul pulsante freccia destra.

Regressione ISW

Il nodo Regressione ISW supporta i seguenti algoritmi di regressione:

- Trasformazioni (default).
- Lineare
- Polynomial
- RBF

Regressione trasformazione

L'algoritmo di regressione trasformazione ISW consente di creare modelli che rappresentano strutture ad albero delle decisioni con equazioni di regressione in corrispondenza delle tre foglie. Notare che il Visualizer IBM non visualizzerà la struttura di questi modelli.

Il browser di IBM SPSS Modeler visualizza le impostazioni e le annotazioni, ma non la struttura dei modelli. L'algoritmo prevede un numero relativamente esiguo di impostazioni configurabili dall'utente.

Regressione lineare

L'algoritmo di regressione lineare ISW presuppone che esista una relazione lineare tra i campi esplicativi e il campo obiettivo e genera modelli che rappresentano delle equazioni. È normale che il valore previsto sia diverso dal valore osservato, in quanto un'equazione di regressione rappresenta un'approssimazione del campo obiettivo. La differenza viene chiamata residuo.

Il modello di IBM InfoSphere Warehouse Data Mining riconosce i campi che non presentano un valore esplicativo. Per stabilire se un campo dispone di un valore esplicativo, l'algoritmo di regressione lineare esegue dei test statistici oltre alla selezione di variabili autonomiche. Se si conoscono già i campi che non presentano un valore esplicativo, è possibile selezionare automaticamente un sottoinsieme dei campi esplicativi in modo da ridurre i tempi di esecuzione.

L'algoritmo di regressione lineare offre i seguenti metodi per selezionare automaticamente i sottoinsiemi di campi esplicativi:

Regressione stepwise. Per utilizzare il metodo di regressione stepwise, è necessario specificare un livello di significatività minima. Solo i campi che presentano un livello di significatività superiore al valore specificato vengono utilizzati dall'algoritmo di regressione lineare.

Regressione R-quadrato. Il metodo di regressione di R-quadrato identifica il modello ideale mediante l'ottimizzazione di una misura di qualità del modello. Viene utilizzata una delle seguenti misure di qualità:

- Coefficiente di correlazione di Pearson quadrato
- Coefficiente di correlazione di Pearson quadrato corretto.

Per default, l'algoritmo di regressione lineare seleziona automaticamente i sottoinsiemi dei campi esplicativi, utilizzando il coefficiente di correlazione di Pearson quadrato corretto per ottimizzare la qualità del modello.

Regressione polinomiale

L'algoritmo di regressione polinomiale ISW presuppone una relazione polinomiale. Un modello di regressione polinomiale è un'equazione formata dalle seguenti parti:

- Il massimo grado di regressione polinomiale
- Un'approssimazione del campo obiettivo
- I campi esplicativi.

Regressione RBF

L'algoritmo di regressione RBF ISW presuppone che esista una relazione tra i campi esplicativi e il campo obiettivo. Questa relazione si può esprimere come una combinazione lineare di funzioni gaussiane. Le funzioni gaussiane sono funzioni RBF specifiche.

Opzioni della scheda Modello per il nodo Regressione ISW

Nella scheda Modello del nodo Regressione ISW è possibile specificare il tipo di algoritmo di regressione da utilizzare, oltre a:

- Se utilizzare o meno dati partizionati
- Se eseguire o meno un test
- Un limite per il valore R^2
- Un limite per il tempo di esecuzione

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Metodo di regressione. Scegliere il tipo di regressione da eseguire. Consultare l'argomento “Regressione ISW” a pagina 67 per ulteriori informazioni.

esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici guadagno cumulativo e così via.

Limita R quadrato. Questa opzione specifica l'errore sistematico massimo tollerato (il coefficiente di correlazione di Pearson quadrato, R^2). Questo coefficiente misura la correlazione tra l'errore di previsione sui dati di verifica e i valori obiettivo effettivi. Ha un valore compreso tra 0 (nessuna correlazione) e 1 (correlazione positiva o negativa perfetta). Il valore definito qui imposta il limite massimo per l'errore sistematico tollerato del modello.

Limita tempo di esecuzione Specificare in minuti il tempo di esecuzione massimo desiderato.

Opzioni avanzate del nodo Regressione ISW

Nella scheda Livello avanzato del nodo Regressione ISW è possibile specificare una serie di opzioni avanzate per la regressione lineare, polinomiale o RBF.

Opzioni avanzate per la regressione lineare o polinomiale

Limita grado di polinomiale. Imposta il grado massimo di regressione polinomiale. Se si imposta il grado di regressione polinomiale massimo su **1**, l'algoritmo di regressione polinomiale è identico a quello della regressione lineare. Se si specifica un valore alto per il grado massimo di regressione polinomiale, l'algoritmo di regressione polinomiale tenderà a sovradattare i dati. Questo significa che il modello risultante approssimerà in modo accurato i dati di addestramento, ma non risulterà valido se applicato a dati non utilizzati per l'addestramento.

Utilizza intercettazione. Quando è attivata questa opzione, la curva di regressione viene forzata a passare attraverso l'origine. Questo significa che il modello non conterrà un termine costante.

Utilizza selezione della funzione automatica Quando è attivata questa opzione, l'algoritmo tenta di determinare un sottoinsieme ottimale di possibili predittori se non si specifica un livello minimo di significatività.

Utilizza livello minimo di significatività Quando si specifica un livello minimo di significatività, viene utilizzata la regressione stepwise per determinare un sottoinsieme di possibili predittori. Solo i campi indipendenti la cui significatività è superiore al valore specificato contribuiscono al calcolo del modello di regressione.

Impostazioni campo. Per specificare le opzioni per i singoli campi di input, fare clic sulla riga corrispondente nella colonna Impostazioni della tabella Impostazioni campo e scegliere **<Specifica impostazioni>**. Consultare l'argomento “Specifica delle impostazioni dei campi per la regressione” a pagina 70 per ulteriori informazioni.

Opzioni avanzate per la regressione RBF

Utilizza dimensione campione output. Definisce un campione 1-ogni-N per la verifica e il test del modello.

Utilizza dimensione campione input. Definisce un campione 1-ogni-N per l'addestramento.

Utilizza numero massimo di centri. Il numero massimo di centri creati a ogni passaggio. Dal momento che il numero dei centri può aumentare fino al doppio del numero iniziale durante un passaggio, il numero effettivo di centri può essere superiore al numero specificato.

Utilizza la dimensione minima di area. Il numero minimo di record assegnati a un'area.

Utilizza numero massimo di passaggi di dati. Il numero massimo di passaggi effettuato dall'algoritmo nei dati di input. Se specificato, questo valore deve essere maggiore o uguale al numero minimo di passaggi.

Utilizza il numero minimo di passaggi di dati. Il numero minimo di passaggi effettuato dall'algoritmo nei dati di input. Specificare un valore elevato solo se si dispone di dati di addestramento sufficienti e se si è certi dell'esistenza di un buon modello.

Specifica delle impostazioni dei campi per la regressione

Nella finestra di dialogo Modifica impostazioni di regressione è possibile specificare l'intervallo di valori per un singolo campo di input per la regressione polinomiale o lineare.

Valore MIN. Valore valido minimo del campo di input.

Valore MAX. Valore valido massimo del campo di input.

Raggruppamento cluster ISW

La funzione di mining di raggruppamento tramite cluster cerca nei dati di input le caratteristiche comuni che ricorrono più frequentemente e raggruppa i dati di input in cluster. I membri di ciascun cluster hanno proprietà simili. Non vengono applicate nozioni preconcette su quali schemi esistano all'interno dei dati. Il raggruppamento tramite cluster è un processo di rilevamento.

Il nodo Raggruppamento cluster ISW offre i seguenti metodi di raggruppamento tramite cluster:

- Demografico
- Kohonen
- BIRCH ottimizzato (Balanced Iterative Reducing and Clustering using Hierarchies)

La tecnica di algoritmo di **raggruppamento tramite cluster demografici** è basata sulla distribuzione. Il raggruppamento tramite cluster basato sulla distribuzione consente un rapido e semplice raggruppamento dei database molto estesi. Il numero di cluster viene scelto automaticamente (l'utente può specificare il numero massimo). È disponibile un'ampia gamma di parametri configurabili dall'utente.

La tecnica di algoritmo **raggruppamento tramite cluster Kohonen** è basata sul centro. La mappa della funzione Kohonen cerca di posizionare i centri di cluster in modo da ridurre al minimo la distanza totale tra i record e il centro di cluster. La possibilità di separare i cluster non viene presa in considerazione. I vettori centrali vengono sistemati in una mappa con un determinato numero di colonne e righe. Questi vettori sono interconnessi in modo che oltre al vettore vincente, che si trova più vicino a un record di addestramento, vengano regolati anche tutti i vettori che si trovano nelle vicinanze. Tuttavia, la quantità di regolazione apportata diminuisce con l'aumentare della distanza dei centri.

La tecnica di algoritmo ottimizzato **raggruppamento Birch** è basata sulla distribuzione e tenta di rendere minima la distanza totale tra i record e i relativi cluster. La distanza di log-verosimiglianza viene utilizzata per default per determinare la distanza tra un record e un cluster; in alternativa, è possibile selezionare la distanza euclidea se tutti i campi attivi sono numerici. L'algoritmo BIRCH esegue due passaggi distinti: innanzitutto dispone i record di input in una struttura ad albero CF (Clustering Feature) in modo che i record simili appartengano agli stessi nodi della struttura ad albero, quindi raggruppa le foglie della struttura ad albero in memoria per generare il risultato di raggruppamento finale.

Opzioni della scheda Modello per il nodo Raggruppamento cluster ISW

Nella scheda Modello del nodo Raggruppamento cluster è possibile specificare il metodo da utilizzare per creare i cluster e alcune opzioni correlate.

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Metodo cluster Scegliere il metodo che si desidera utilizzare per creare i cluster: **Demografico**, **Kohonen** o **BIRCH ottimizzato**. Consultare l'argomento "Raggruppamento cluster ISW" a pagina 70 per ulteriori informazioni.

Limita numero di cluster. La limitazione del numero di cluster consente di risparmiare tempo in fase di esecuzione evitando la produzione di molti cluster di piccole dimensioni.

Numero di righe/Numero di colonne (Solo per il metodo Kohonen) Specifica il numero di righe e di colonne della mappa delle caratteristiche Kohonen. (Disponibile solo se è selezionato **Limita numero di passaggi di Kohonen** ed è deselezionato **Limita numero di cluster**.)

Limita numero di passaggi Kohonen. (Solo per il metodo Kohonen) Specifica il numero di passaggi compiuti dall'algoritmo di raggruppamento tramite cluster sui dati durante le esecuzioni dell'addestramento. A ogni passaggio, i vettori del centro vengono regolati in modo da ridurre al minimo la distanza tra i centri di cluster e i record. Inoltre, a ogni passaggio, diminuisce la quantità di regolazione apportata ai vettori. Al primo passaggio, le regolazioni sono approssimative. All'ultimo passaggio, l'entità della regolazione apportata ai centri è piuttosto ridotta. Vengono eseguite solo regolazioni minime.

Misura della distanza. (Solo per il metodo BIRCH ottimizzato) Selezionare la misura della distanza dal record al cluster utilizzata dall'algoritmo BIRCH. È possibile selezionare la distanza di verosimiglianza, che rappresenta l'importazione di default, oppure la distanza euclidea. *Nota*: è possibile scegliere solo distanza euclidea quando tutti i campi attivi sono numerici.

Numero massimo di nodi foglia. (Solo per il metodo BIRCH ottimizzato) Numero massimo di nodi foglia che si desidera includere nella struttura ad albero CF (Clustering Feature). La struttura ad albero CF (Clustering Feature) è il risultato del primo passaggio dell'algoritmo BIRCH ottimizzato, nel quale i record di dati sono disposti in una struttura ad albero in modo che i record simili appartengano allo stesso nodo foglia. Il tempo di esecuzione dell'algoritmo aumenta con il numero dei nodi foglia. Il valore predefinito è 1000.

Passaggi Birch. (Solo per il metodo BIRCH ottimizzato) Numero di passaggi eseguiti dall'algoritmo sui dati per perfezionare il risultato del raggruppamento in cluster. Il numero di passaggi ha effetto sul tempo di elaborazione delle esecuzioni di addestramento (poiché ciascun passaggio richiede una scansione completa dei dati) e sulla qualità del modello. Valori bassi limitano il tempo di elaborazione, ma possono ridurre la qualità dei modelli. Valori alti aumentano il tempo di elaborazione ma generano solitamente modelli migliori. In media, 3 o più passaggi producono buoni risultati. Il valore predefinito è 3.

Opzioni avanzate del nodo Raggruppamento cluster ISW

Nella scheda Livello avanzato del nodo Raggruppamento cluster è possibile specificare opzioni avanzate come soglie di similarità, limiti per il tempo di esecuzione e pesi dei campi.

Limita tempo di esecuzione. Selezionare questa casella per abilitare le opzioni che consentono di controllare il tempo impiegato per creare il modello. È possibile specificare un intervallo di tempo in

minuti, una percentuale minima di dati di addestramento da elaborare o entrambi i valori. Per il metodo Birch, è anche possibile specificare il numero massimo di nodi foglia da creare nella struttura ad albero CF.

Specifica soglia di similarità. (Solo per il raggruppamento tramite cluster demografici) Il limite inferiore della similarità di due record di dati appartenenti allo stesso cluster. Per esempio, un valore di 0,25 significa che i record con valori simili per il 25% saranno probabilmente assegnati allo stesso cluster. Un valore di 1,0 significa che i record devono essere identici per essere assegnati allo stesso cluster.

Impostazioni campo. Per specificare le opzioni per i singoli campi di input, fare clic sulla riga corrispondente nella colonna Impostazioni della tabella Impostazioni campo e scegliere **<Specifica impostazioni>**.

Specifica delle impostazioni dei campi per il raggruppamento tramite cluster

Nella finestra di dialogo Modifica impostazioni cluster è possibile specificare le opzioni per singoli campi di input.

Peso campo. Assegna più o meno peso al campo durante il processo di creazione del modello. Per esempio, se si ritiene che questo campo sia relativamente meno importante per il modello rispetto agli altri campi, diminuirne il peso rispetto agli altri campi.

Peso valore. Assegna più o meno peso a determinati valori del campo. Alcuni valori del campo potrebbero essere più comuni di altri. La coincidenza di valori rari in un campo potrebbe essere più significativa per un cluster della coincidenza di valori frequenti. Per ponderare i valori del campo è possibile scegliere uno dei seguenti metodi (in entrambi i casi, i valori rari hanno un peso notevole, mentre quelli frequenti hanno un peso scarso):

- **Logaritmico.** Assegna un peso a ciascun valore a seconda del logaritmo della sua probabilità nei dati di input.
- **Probabilistico.** Assegna un peso a ciascun valore a seconda della sua probabilità nei dati di input.

Per entrambi i metodi è possibile scegliere anche un'opzione **con compensazione** per compensare la ponderazione del valore applicata a ogni campo. Se si compensa la ponderazione del valore, l'importanza globale del campo ponderato è uguale a quella di un campo non ponderato, indipendentemente dal numero di valori possibili. La ponderazione compensata influisce solo sull'importanza relativa delle coincidenze nell'insieme dei valori possibili.

Utilizza scala di similarità. Selezionare questa casella per utilizzare la scala di similarità per controllare il calcolo della misurazione di similarità di un campo. La scala di similarità può essere indicata come numero assoluto. La specifica viene considerata solo per i campi numerici attivi. Se non si specifica una scala di similarità, viene utilizzato il valore di default (la metà della deviazione standard). Per ottenere un gran numero di cluster, ridurre la similarità media fra coppie di cluster adottando scale di similarità più piccole per i campi numerici.

Trattamento valori anomali. I valori anomali sono valori al di fuori dell'intervallo dei valori specificati per un campo, definiti da **Valore MIN** e **Valore MAX**. È possibile decidere come gestire i valori anomali per questo campo.

- L'impostazione di default, **nessuno**, significa che per i valori anomali non è previsto alcun trattamento particolare.
- Se si sceglie **sostituire con MIN o MAX**, i valori di campo inferiori a **Valore MIN** o superiori a **Valore MAX** vengono sostituiti rispettivamente con i valori MIN o MAX. In questo caso è possibile impostare i valori di MIN e MAX.
- Se si sceglie **trattare come mancante**, i valori anomali vengono trattati come valori mancanti e ignorati. In questo caso è possibile impostare i valori di MIN e MAX.

Naive Bayes ISW

Naive Bayes è un algoritmo molto noto per problemi di classificazione. Il modello è denominato *naive* perché considera tutte le variabili di previsione proposte indipendenti l'una dall'altra. Naive Bayes è un algoritmo veloce e scalabile in grado di calcolare probabilità condizionali per combinazioni di attributi e per l'attributo obiettivo. Dai dati di addestramento viene stabilita una probabilità indipendente che serve a indicare la verosimiglianza di ciascuna classe obiettivo una volta specificata l'occorrenza di ogni categoria di valore da ogni variabile di input.

L'algoritmo di classificazione Naive Bayes ISW è un classificatore probabilistico basato su modelli di probabilità che integrano forti presupposizioni di indipendenza.

Opzioni del modello Naive Bayes ISW

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test IBM InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici guadagno cumulativo e così via.

Soglia probabilità. Definisce una probabilità per qualsiasi combinazione di valori predittore e obiettivo non evidenti nei dati di addestramento. Questa probabilità deve essere compresa tra 0 e 1. L'impostazione di default è 0,001.

Regressione logistica ISW

La regressione logistica, nota anche come regressione nominale, è una tecnica statistica per classificare i record in base ai valori dei campi di input. È simile alla regressione lineare, ma l'algoritmo di regressione logistica ISW richiede un campo obiettivo flag (binario) anziché un campo numerico.

Opzioni del modello di Regressione logistica ISW

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Esegui test. Se si seleziona questa opzione, dopo la creazione del modello nella partizione di addestramento verrà eseguito un test IBM InfoSphere Warehouse Data Mining. Ciò comporterà un passaggio sulla partizione di test per stabilire informazioni su qualità del modello, grafici guadagno cumulativo e così via.

Serie temporali ISW

Gli algoritmi serie temporali ISW consentono di prevedere eventi futuri in base a eventi noti verificatisi in passato.

Analogamente ai comuni metodi di regressione, gli algoritmi di serie temporali prevedono un valore numerico. A differenza di quanto accade per i comuni metodi di regressione, invece, le previsioni di serie temporali si concentrano sui valori futuri di una serie ordinata, definiti in genere previsioni.

Gli algoritmi di serie temporali sono univariati, nel senso che la variabile indipendente è una colonna tempo o ordine. Le previsioni si basano su valori passati e non su altre colonne indipendenti.

Gli algoritmi di serie temporali sono diversi dai comuni algoritmi di regressione perché non si limitano a prevedere valori futuri ma incorporano nella previsione anche dei cicli stagionali.

La funzione mining Serie temporali dispone dei seguenti algoritmi per la previsione di tendenze future:

- Modello Autoregressivo Integrato a Media Mobile (ARIMA)
- Livellamento esponenziale
- Scomposizione tendenza stagionale

L'algoritmo che crea la migliore previsione in base ai dati disponibili parte da ipotesi di modello diverse. È possibile calcolare tutte le previsioni contemporaneamente. Gli algoritmi calcolano una previsione dettagliata che comprende il comportamento stagionale della serie temporale originale. Se è installato il client IBM InfoSphere Warehouse è possibile utilizzare il Visualizer di serie temporali per valutare e confrontare le curve risultanti.

Opzioni Campi Serie temporali ISW

Ora. Selezionare il campo di input contenente la serie temporale. Deve trattarsi di un campo con tipo di archiviazione Data, Ora, Timestamp, Numero reale o Numero intero.

Utilizza le impostazioni del nodo Tipo. Questa opzione indica al nodo di utilizzare le informazioni sui campi da un nodo Tipo upstream. Questa è l'opzione di default.

Utilizza impostazioni personalizzate. Questa opzione indica al nodo di utilizzare le informazioni sui campi specificate qui al posto di quelle date in un qualsiasi nodo Tipo upstream. Dopo avere selezionato questa opzione, specificare i campi nell'area sottostante come richiesto.

Obiettivi. Selezionare uno o più campi obiettivo. Questa operazione è simile all'impostazione del ruolo di un campo su *Obiettivo* in un nodo Tipo.

Opzioni del modello di serie temporali ISW

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Algoritmi di previsione. Selezionare gli algoritmi da utilizzare per la modellazione. È possibile scegliere una o più delle opzioni seguenti:

- ARIMA
- Livellamento esponenziale
- Scomposizione tendenza stagionale.

Ora di fine della previsione. Specifica se l'ora di fine della previsione deve essere calcolata automaticamente o indicata manualmente.

Valore del campo Ora. Quando l'opzione **Ora di fine previsione** è impostata su manuale, immettere l'ora di fine della previsione. Il valore che è possibile inserire dipende dal tipo del campo Ora; per esempio, se

il valore è un integer che rappresenta le ore è possibile inserire 48 per interrompere la previsione dopo l'elaborazione di 48 ore di dati. In alternativa, questo campo può richiedere di immettere una data o un'ora come valore di fine.

Opzioni avanzate per le serie temporali ISW

Utilizza tutti i record per creare il modello. Questa è l'impostazione di default; quando viene creato il modello, vengono analizzati tutti i record.

Utilizza un sottoinsieme di record per creare il modello. Se si desidera creare il modello a partire da solo alcuni dei dati disponibili, selezionare questa opzione. Per esempio, questo potrebbe essere necessario in presenza di un volume eccessivo di dati ripetitivi.

Immettere il **Valore ora di inizio** e il **Valore ora di fine** per identificare i dati da utilizzare. Si noti che i valori che è possibile digitare in questi campi dipendono dal tipo del campo Ora, che può essere per esempio un numero di ore o di giorni, oppure una data e un'ora specifica.

Metodo di interpolazione per i valori obiettivo mancanti. In caso di elaborazione di dati con uno o più valori mancanti, selezionare il metodo da utilizzare per calcolarli. È possibile scegliere tra le opzioni seguenti:

- Lineare
- Spline esponenziali
- Spline cubiche

Visualizzazione dei modelli di serie temporali ISW

I modelli di serie temporali ISW vengono creati sotto forma di modelli grezzi, contenenti informazioni estratte dai dati ma non destinati direttamente alla generazione di previsioni.

Figura 3. Icona di modello grezzo

Se è installato il client IBM InfoSphere Warehouse, è possibile utilizzare lo strumento Visualizer di serie temporali per visualizzare una riproduzione grafica dei dati di serie temporali.

Per utilizzare lo strumento Visualizer di serie temporali:

1. Verificare di avere eseguito tutte le attività per l'integrazione di IBM SPSS Modeler con IBM InfoSphere Warehouse. Consultare l'argomento “Attivazione dell'integrazione con IBM InfoSphere Warehouse” a pagina 53 per ulteriori informazioni.
2. Fare doppio clic sull'icona del modello grezzo nella palette Modelli.
3. Nella scheda Server della finestra di dialogo, fare clic sul pulsante Visualizza per richiamare il Visualizer nel browser di default.

Nugget del modello di ISW Data Mining

È possibile creare modelli dai nodi Struttura ad albero delle decisioni, Associazione, Sequenza, Regressione e Raggruppamento tramite cluster ISW inclusi in IBM SPSS Modeler.

Scheda Server del nugget del modello ISW

La scheda Server offre le opzioni che consentono di eseguire controlli di uniformità e avviare lo strumento IBM Visualizer.

IBM SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello ISW che nel modello di IBM SPSS Modeler. Per eseguire questo tipo di verifica, fare clic sul pulsante **Controllo** nella scheda Server. Consultare l'argomento “Gestione dei modelli DB2” a pagina 58 per ulteriori informazioni.

Lo strumento visualizzatore è l'unico modo per sfogliare i modelli di InfoSphere Warehouse Data Mining. Lo strumento può essere installato facoltativamente con InfoSphere Warehouse Data Mining. Consultare l'argomento “Attivazione dell'integrazione con IBM InfoSphere Warehouse” a pagina 53 per ulteriori informazioni.

- Fare clic su **Visualizza** per avviare lo strumento visualizzatore. Ciò che viene visualizzato dallo strumento dipende dal tipo di nodo generato. Ad esempio, lo strumento visualizzatore restituirà una visualizzazione Classi previste quando viene avviato dal nugget del modello della struttura ad albero delle decisioni ISW.
- Fare clic su **Risultati del test** (solo Strutture ad albero delle decisioni e Sequenza) per avviare lo strumento visualizzatore e visualizzare la qualità globale del modello generato.

Scheda Impostazioni del nugget del modello ISW

In IBM SPSS Modeler viene generalmente fornita una sola previsione con la probabilità o la confidenza associata. È inoltre disponibile un'opzione utente per la visualizzazione delle probabilità per ogni risultato (simile a quella della regressione logistica) che rappresenta un'opzione tempo punteggio ubicata all'interno della scheda Impostazioni del nugget del modello.

Includi confidenze per tutte le classi. Aggiunge una colonna con il livello di confidenza per ciascuno dei possibili risultati del campo obiettivo.

Scheda Riepilogo del nugget del modello ISW

La scheda Riepilogo di un nugget del modello visualizza le informazioni sul modello stesso (*Analisi*), i campi utilizzati nel modello (*Campi*), le impostazioni utilizzate durante la creazione del modello (*Impostazioni di creazione*) e l'addestramento del modello (*Riepilogo addestramento*).

Quando si sfoglia per la prima volta il nodo, i risultati della scheda Riepilogo sono compressi. Per visualizzare i risultati, utilizzare il controllo dell'espansore a sinistra di un elemento se si desidera visualizzare solo i risultati di tale elemento oppure fare clic sul pulsante **Espandi tutto** se si desidera visualizzare tutti i risultati. Per nascondere i risultati, utilizzare il controllo dell'espansore di un elemento se si desidera nascondere solo i risultati di tale elemento oppure fare clic sul pulsante **Comprimi tutto** se si desidera nascondere tutti i risultati.

Analisi. Visualizza informazioni sul modello specifico. Se è stato eseguito un nodo Analisi collegato a questo nugget del modello, in questa sezione verranno visualizzate anche le informazioni sui risultati di tale analisi.

Campi. Elenca i campi utilizzati come obiettivi e come input nella creazione del modello.

Impostazioni di creazione. Contiene informazioni sulle impostazioni utilizzate nella creazione del modello.

Riepilogo addestramento. Mostra il tipo di modello, il flusso utilizzato per creare tale modello, l'utente che lo ha creato, la data di creazione e il tempo trascorso per la creazione del modello.

Esempi di ISW Data Mining

IBM SPSS Modeler per Windows viene fornito con un'ampia gamma di flussi di esempio che illustrano il processo di mining nel database. Tali flussi si trovano nella cartella di installazione di IBM SPSS Modeler in:

\Demos\Database_Modeling\IBM DB2 ISW

Nota: è possibile accedere alla cartella Demos dal gruppo di programmi IBM SPSS Modeler nel menu Start e Windows.

I flussi riportati di seguito possono essere utilizzati insieme, in ordine sequenziale, come esempio del processo di mining nel database:

- *1_upload_data.str*—Utilizzato per ripulire e caricare i dati da un file flat in DB2.
- *2_explore_data.str*—Utilizzato come esempio di un'esplorazione dati con IBM SPSS Modeler.
- *3_build_model.str*—Utilizzato per creare un modello Struttura ad albero delle decisioni.
- *4_evaluate_model.str*—Utilizzato come esempio di valutazione del modello con IBM SPSS Modeler.
- *5_deploy_model.str*—Utilizzato per distribuire il modello per il calcolo del punteggio nel database.

L'insieme di dati impiegato nei flussi di esempio concerne applicazioni relative alle carte di credito e presenta un problema di classificazione relativo a un insieme misto di predittori continui e categoriali. Per ulteriori informazioni su questo insieme di dati, fare riferimento al seguente file della cartella di installazione di IBM SPSS Modeler in:

\Demos\Database_Modeling\IBM DB2 ISW\crx.names

Questo dataset è disponibile dal repository per l'apprendimento automatico UCI all'indirizzo http://archive.ics.uci.edu/ml/.

Flusso di esempio: caricamento dati

Il primo flusso di esempio, *1_upload_data.str*, viene utilizzato per pulire e caricare dati da un file flat in DB2.

Il nodo Riempimento viene utilizzato per gestire l'assenza di valori e sostituisce i campi vuoti letti dal file di testo *crx.data* con i valori *NULL*.

Flusso di esempio: Esplorazione dati

Il secondo flusso di esempio, *2_explore_data.str* viene utilizzato per dimostrare l'esplorazione dati in IBM SPSS Modeler.

Un passo tipico utilizzato durante l'esplorazione dati consiste nell'allegare un nodo Esplora ai dati. Il nodo Esplora è disponibile nella palette di nodi Output.

È possibile utilizzare l'output di un nodo Esplora per acquisire una panoramica generale sui campi e la distribuzione dei dati. Facendo doppio clic su un grafico nella finestra Esplora, si accede a una visualizzazione più dettagliata del grafico che consente un'esplorazione più approfondita di un dato campo.

Flusso di esempio: creazione modello

Il terzo flusso di esempio, *3_build_model.str*, illustra la creazione del modello in IBM SPSS Modeler. È possibile collegare il nodo Modelli database al flusso e fare doppio clic per specificare le impostazioni di creazione.

Utilizzando le schede Modello e Livello avanzato del nodo Modelli, è possibile modificare la profondità massima della struttura ad albero e bloccare l'ulteriore suddivisione di un nodo dal momento in cui viene creato la struttura ad albero delle decisioni iniziale impostando la massima purezza e il numero minimo di casi per nodo interno. Consultare l'argomento "Struttura ad albero delle decisioni ISW" a pagina 61 per ulteriori informazioni.

Flusso di esempio: valutazione modello

Il quarto flusso di esempio, *4_evaluate_model.str*, illustra i vantaggi associati all'utilizzo di IBM SPSS Modeler per la modellazione nel database. Una volta eseguito il modello, è possibile aggiungerlo nuovamente al flusso di dati e valutarlo con il supporto di un'ampia gamma di strumenti mirati disponibili in IBM SPSS Modeler.

Quando si apre il flusso per la prima volta, il nugget del modello (*campo16*) non è compreso nel flusso. Aprire il nodo origine CREDIT e verificare di avere specificato un'origine dati. Quindi, a condizione di avere eseguito il flusso *3_build_model.str* per creare un nugget *campo16* nella palette Modelli, è possibile eseguire i nodi disconnessi facendo clic sul pulsante **Esegui** nella barra degli strumenti (il pulsante contrassegnato da un triangolo verde). In tal modo viene eseguito uno script che copia il nugget *campo16* nel flusso, quindi lo connette ai nodi esistenti e, infine, esegue i nodi terminali nel flusso.

È possibile collegare un nodo Analisi (disponibile nella palette Output) per creare una matrice di coincidenza che mostri lo schema di corrispondenze tra ogni campo generato (previsto) e il relativo campo obiettivo. Eseguire il nodo Analisi per visualizzare i risultati.

È anche possibile creare un grafico dei profitti in modo da mostrare i miglioramenti in termini di precisione realizzati dal modello. Collegare un nodo Valutazione al modello generato, quindi eseguire il flusso per visualizzare i risultati.

Flusso di esempio: Deployment del modello

Una volta raggiunto il livello di precisione del modello desiderato, è possibile eseguire la distribuzione del modello per consentirne l'utilizzo con applicazioni esterne o la riscrittura dei punteggi nel database. Nel flusso di esempio *5_deploy_model.str* i dati vengono letti dalla tabella `CREDIT`. Quando si esegue il nodo di esportazione del database *soluzione di distribuzione*, il punteggio dei dati non viene calcolato. Il flusso crea invece il file di immagine pubblicato *credit_scorer.pim* e il file di parametri pubblicato *credit_scorer.par*.

Come nell'esempio precedente, il flusso esegue uno script che copia il nugget *campo16* nel flusso dalla palette Modelli, lo connette ai nodi esistenti e, infine, esegue i nodi terminali nel flusso. In questo caso occorre prima specificare un'origine dati sia nei nodi origine Database che nei nodi di esportazione.

Capitolo 6. Modellazione di database con IBM Netezza Analytics

IBM SPSS Modeler and IBM Netezza Analytics

IBM SPSS Modeler supporta l'integrazione con IBM Netezza Analytics, che consente di eseguire analisi avanzate sui server IBM Netezza. Queste funzionalità sono accessibili tramite l'interfaccia utente grafica e l'ambiente di sviluppo basato sui flussi di lavoro di IBM SPSS Modeler e consentono di eseguire gli algoritmi di data mining direttamente nell'ambiente IBM Netezza.

IBM SPSS Modeler supporta l'integrazione con i seguenti algoritmi di IBM Netezza Analytics.

- Strutture ad albero delle decisioni
- Medie K
- Rete di Bayes
- Naive Bayes
- KNN
- Raggruppamento cluster divisivo
- PCA
- Struttura ad albero di regressione
- Regressione lineare
- Serie temporali
- Lineare generalizzato

Per ulteriori informazioni relative agli algoritmi, consultare *IBM Netezza Analytics Developer's Guide* e *IBM Netezza Analytics Reference Guide*.

Requisiti per l'integrazione con IBM Netezza Analytics

Di seguito sono elencate le condizioni che costituiscono i prerequisiti indispensabili per l'esecuzione della modellazione nel database con IBM Netezza Analytics. Per garantire che queste condizioni vengano soddisfatte, può essere necessario consultare l'amministratore di database.

- IBM SPSS Modeler in esecuzione su un'installazione di IBM SPSS Modeler Server in ambiente Windows o UNIX (ad eccezione di zLinux, per cui i driver ODBC di IBM Netezza non sono disponibili).
- IBM Netezza Performance Server, con il pacchetto IBM Netezza Analytics.

  Nota: la versione minima di NPS (Netezza Performance Server) richiesta dipende dalla versione di INZA richiesta, come riportato di seguito:

  – Qualsiasi versione di NPS successiva alla versione 6.0.0 P8 supporterà le versioni di INZA precedenti alla versione 2.0.

  – Per utilizzare INZA 2.0 o versione successiva è necessario NPS 6.0.5 P5 o versione successiva.

  Lineare generalizzato Netezza e Serie temporali Netezza richiedono INZA 2.0 e versioni successive. Tutti gli altri nodi In-Database Netezza richiedono INZA 1.1 o versioni successive.
- Un'origine dati ODBC per la connessione a un database IBM Netezza. Consultare l'argomento "Abilitazione dell'integrazione con IBM Netezza Analytics" a pagina 80 per ulteriori informazioni.
- Generazione e ottimizzazione SQL abilitate in IBM SPSS Modeler. Consultare l'argomento "Abilitazione dell'integrazione con IBM Netezza Analytics" a pagina 80 per ulteriori informazioni.

Nota: le funzionalità di modellazione nel database e ottimizzazione SQL richiedono che la connettività IBM SPSS Modeler Server venga abilitata sul computer IBM SPSS Modeler. Con questa impostazione attivata, è possibile accedere agli algoritmi di database, restituire codice SQL direttamente da IBM SPSS Modeler, e accedere a IBM SPSS Modeler Server. Per verificare lo stato attuale della licenza, scegliere le seguenti opzioni dal menu IBM SPSS Modeler.

Guida > **Informazioni su** > **Dettagli aggiuntivi**

Se la connettività è abilitata, l'opzione **Abilitazione server** viene visualizzata nella scheda Stato della licenza.

Abilitazione dell'integrazione con IBM Netezza Analytics

L'attivazione dell'integrazione con IBM Netezza Analytics prevede i seguenti passaggi.

- Configurazione IBM Netezza Analytics
- Creazione di un'origine ODBC
- Attivazione dell'integrazione in IBM SPSS Modeler
- Attivazione della generazione e dell'ottimizzazione SQL in IBM SPSS Modeler

I passaggi sono descritti nelle sezioni che seguono.

Configurazione di IBM Netezza Analytics

Per installare e configurare IBM Netezza Analytics, consultare la documentazione IBM Netezza Analytics in particolare *IBM Netezza Analytics Installation Guide*—per ulteriori dettagli. La sezione *Setting Database Permissions* di quel manuale contiene informazioni sugli script che è necessario eseguire per consentire ai flussi di IBM SPSS Modeler di scrivere nel database.

Nota: Se si utilizzeranno i nodi che si basano sui calcoli della matrice (PCA Netezza e Regressione lineare Netezza) il motore della matrice Netezza deve essere inizializzato mediante l'esecuzione di `CALL NZM..INITIALIZE();` altrimenti l'esecuzione delle procedure archiviate avrà esito negativo. L'inizializzazione è un passaggio della configurazione da eseguire una sola volta per ogni database.

Creazione di un'origine ODBC per IBM Netezza Analytics

Per attivare la connessione tra il database IBM Netezza e IBM SPSS Modeler è necessario creare un nome di origine dati (DSN, Data Source Name) ODBC.

Per creare un DSN, è necessario avere una conoscenza di base delle origini dati e dei driver ODBC e disporre del supporto database in IBM SPSS Modeler.

Se l'applicazione è in esecuzione in modalità distribuita su IBM SPSS Modeler Server, creare il DSN sul computer server. Se invece è attiva la modalità locale (client), creare il DSN sul computer client.

Client Windows

1. Dal CD del *client Netezza* eseguire il file *nzodbcsetup.exe* per avviare il programma di installazione. Attenersi alle istruzioni visualizzate per installare il driver. Per le istruzioni complete, vedere IBM Netezza ODBC, JDBC, and OLE DB Installation and Configuration Guide.
   a. Creare il DSN.

      Nota: La sequenza dei menu dipende dalla versione di Windows.

      - **Windows XP.** dal menu Start, scegliere **Pannello di controllo**. Fare doppio clic su **Strumenti di amministrazione** e doppio clic su **Origini dati (ODBC)**.
      - **Windows Vista.** Dal menu Start, scegliere **Pannello di controllo**, quindi **Strumenti di amministrazione**. Doppio clic su **Strumenti di amministrazione**, selezionare**Origini dati (ODBC)**, quindi fare clic su **Apri**.

- **Windows 7.** dal menu Start, scegliere **pannello di controllo**, quindi**Sistema & Sicurezza**, quindi **Strumenti di amministrazione**. Selezionare**Origini dati (ODBC)**, then click **Open**.

b. Fare clic sulla scheda **DSN di sistema**, quindi fare clic su **Aggiungi**.

2. Selezionare **NetezzaSQL** dall'elenco e fare clic su **Fine**.
3. Nella scheda delle **opzioni DSN** della schermata IBM Netezza ODBC Driver Setup, digitare un nome di origine dati, il nome host o indirizzo IP del server IBM Netezza, il numero di porta per la connessione, il database dell'istanza Netezza in uso e il nome utente e la password utilizzati per la connessione al database. Fare clic sul pulsante **Guida** per visualizzare una spiegazione dei campi.
4. Fare clic sul pulsante **Verifica connessione** e assicurarsi di poter eseguire la connessione al database.
5. Stabilita correttamente la connessione, fare clic su **OK** più volte per uscire dalla schermata Amministratore origine dati ODBC.

Server Windows

La procedura per Windows Server è uguale alla procedura per il client in Windows XP.

Server UNIX o Linux

La procedura che segue è valida per i server UNIX o Linux (tranne zLinux, per il quale non sono disponibili driver ODBC di IBM Netezza).

1. Dal proprio CD/DVD del client Netezza, copiare il file `<platform>cli.package.tar.gz` pertinente in una posizione temporanea sul server.
2. Estrarre i contenuti dell'archivio utilizzando i comandi **`gunzip`** e **`untar`**.
3. Aggiungere le autorizzazioni per l'esecuzione allo script *unpack* estratto.
4. Eseguire lo script, rispondendo ai prompt visualizzati.
5. Modificare il file `modelersrv.sh` in modo che includa le righe riportate di seguito.

   ```
   . <SDAP Install Path>/odbc.sh
   LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
   NZ_ODBC_INI_PATH=<SDAP Install Path>; export NZ_ODBC_INI_PATH
   ```

   Ad esempio:

   ```
   . /usr/IBM/SPSS/SDAP/odbc.sh
   LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
   NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP; export NZ_ODBC_INI_PATH
   ```

6. Individuare il file `/usr/local/nz/lib64/odbc.ini` e copiarne il contenuto nel file `odbc.ini` installato con SDAP (quello definito dalla variabile di ambiente $ODBCINI).

   Nota: per i sistemi Linux a 64-bit, il parametro **`Driver`** fa riferimento per errore ad un driver a 32-bit. Quando si copia il contenuto di `odbc.ini` nel passaggio precedente, modificare il percorso in questo parametro come nell'esempio che segue:

   ```
   /usr/local/nz/lib64/libnzodbc.so
   ```

7. Modificare i parametri nella definizione DSN Netezza in modo che riflettano il database da utilizzare.
8. Riavviare IBM SPSS Modeler Server e provare a utilizzare i nodi di mining nel database di Netezza sul client.

Attivazione dell'integrazione IBM Netezza Analytics in IBM SPSS Modeler

1. Dal menu principale IBM SPSS Modeler, scegliere

   Strumenti > **Opzioni** > **Applicazioni di supporto**
2. Fare clic sulla scheda **IBM Netezza**.

Abilita l'integrazione Netezza Data Mining. Attiva la palette Modelli in-database (se non è già visualizzata) nella parte inferiore della finestra IBM SPSS Modeler e aggiunge i nodi degli algoritmi di Netezza Data Mining.

Connessione Netezza. Fare clic sul pulsante **Modifica** e scegliere la stringa di connessione Netezza specificata al momento della creazione dell'origine ODBC. Consultare l'argomento "Creazione di un'origine ODBC per IBM Netezza Analytics" a pagina 80 per ulteriori informazioni.

Attivazione di generazione e ottimizzazione SQL

Poiché è probabile che ci si trovi a lavorare con insiemi di dati di dimensioni molto grandi, per motivi di prestazioni è bene attivare le opzioni di generazione e ottimizzazione SQL in IBM SPSS Modeler.

1. Dal menu IBM SPSS Modeler scegliere:

   Strumenti > **Proprietà flusso** > **Opzioni**
2. Fare clic sull'opzione **Ottimizzazione** nel riquadro di spostamento.
3. Confermare che l'opzione **Genera SQL** è attivata. Questa impostazione è necessaria per il corretto funzionamento della modellazione di database.
4. Selezionare **Ottimizza generazione SQL** e **Ottimizza altre esecuzioni** (queste due opzioni non sono strettamente necessarie, tuttavia se ne consiglia la selezione per ottenere performance ottimizzate).

Creazione dei modelli con IBM Netezza Analytics

Per ognuno degli algoritmi supportati esiste un nodo Modelli corrispondente. Ai nodi Modelli di IBM Netezza è possibile accedere dalla scheda Modelli database nella palette dei nodi.

Considerazioni sui dati

I campi dell'origine dati possono contenere variabili di diversi tipi di dati, in base al nodo Modelli. In IBM SPSS Modeler, i tipi di dati sono noti come **livelli di misurazione**. La scheda Campi del nodo Modelli utilizza delle icone per indicare i tipi di livello di misurazione consentiti per i campi di input e obiettivo.

Campo obiettivo. Il campo obiettivo è il campo il cui valore si tenta di prevedere. Dove può essere specificato un obiettivo, è possibile selezionare come campo obiettivo un solo campo dati di origine.

Campo ID record. Specifica il campo utilizzato per identificare ciascun caso in modo univoco. Tale campo può coincidere, per esempio, con un campo ID quale *IDCliente*. Se i dati di origine non includono un campo ID, è possibile crearlo mediante un nodo Ricava, come indica la procedura che segue.

1. Selezionare il nodo origine.
2. Nella scheda Oper su campi della palette dei nodi, fare doppio clic sul nodo Ricava.
3. Aprire il nodo Ricava facendo doppio clic sulla relativa icona nell'area.
4. Nel campo **Ricava** digitare, per esempio, `ID`.
5. Nel campo **Formula** digitare `@INDEX` e fare clic su **OK**.
6. Collegare il nodo Ricava al resto del flusso.

Nota: Se si richiamano dati numerici estesi da un database Netezza utilizzando il tipo dati `NUMERIC(18,0)`, SPSS Modeler può a volte arrotondare i dati durante l'importazione. Per evitare questo problema, memorizzare i propri dati utilizzando il tipo dati `BIGINT` o `NUMERIC(36,0)`.

Gestione dei valori null

Se i dati di input contengono valori null, l'utilizzo di alcuni nodi Netezza potrebbe causare messaggi di errore o flussi lunghi da eseguire, quindi è consigliabile rimuovere i record che contengono valori null. Utilizzare il metodo seguente.

1. Collegare un nodo Seleziona al nodo origine.
2. Impostare l'opzione **Modalità** del nodo Seleziona su **Scarta**.
3. Immettere quanto segue nel campo **Condizione**:
   `@NULL(campo1) [or @NULL(campo2)[... or @NULL(campoN]])`

   Assicurarsi di includere tutti i campi di input.
4. Collegare il nodo Seleziona al resto del flusso.

Output del modello

È possibile che un flusso contenente un nodo di modellazione Netezza produca risultati leggermente diversi a ogni esecuzione. Questo si verifica perché l'ordine con il quale il nodo legge i dati di input non è sempre lo stesso poiché i dati vengono letti in tabelle temporanee prima della creazione del modello. Le differenze prodotte da questo effetto sono tuttavia trascurabili.

Commenti generali

- In IBM SPSS Collaboration and Deployment Services non è possibile creare configurazioni per il calcolo del punteggio utilizzando flussi che contengono nodi Modelli database IBM Netezza.
- Per i modelli creati dai nodi Netezza non è possibile eseguire l'esportazione o l'importazione PMML.

Opzioni della scheda Campi dei modelli Netezza

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi upstream oppure creare manualmente le assegnazioni dei campi.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo upstream o dalla scheda Tipi di un nodo origine upstream.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione per assegnare manualmente obiettivi, predittori e altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

Obiettivo. Scegliere un campo come obiettivo per la previsione. Per i modelli lineari generalizzati, vedere anche il campo **Prove** di questo schermo.

ID record. Campo da utilizzare come identificatore univoco del record.

Predittori (Input). Scegliere uno o più campi come input per la previsione.

Opzioni della scheda Server dei modelli Netezza

Nella scheda Server, è possibile specificare il database IBM in cui deve essere creato il modello.

Dettagli Server DB Netezza. Qui si specificano i dettagli della connessione per il database da utilizzare per il modello.

- **Utilizza connessione upstream**. (default) Utilizza i dettagli di connessione specificati in un nodo upstream, per esempio il nodo origine del database. *Nota*: questa opzione funziona solo se tutti i nodi upstream sono in grado di utilizzare Push back SQL. In questo caso non è necessario spostare i dati fuori dal database perché SQL supporta pienamente tutti i nodi upstream.

- **Sposta dati alla connessione**. Sposta i dati nel database indicato qui. In questo modo si consente al modello di lavorare se i dati si trovano su un altro database IBM Netezza, o su un database di un altro fornitore, o anche se i dati si trovano in un file flat. Inoltre i dati vengono riportati in questo database se sono stati in precedenza estratti perché un nodo non ha effettuato il push back SQL. Fare clic sul pulsante **Modifica** per reperire e selezionare una connessione. *Attenzione*: IBM Netezza Analytics viene generalmente utilizzato con data set molto grandi. Il trasferimento di grandi quantità di dati tra database, o anche dentro e fuori lo stesso database, può richiedere molto tempo ed è quindi da evitare se possibile.

Nota: Il nome dell'origine dati ODBC è efficacemente incorporato in ogni flusso di IBM SPSS Modeler. Se un flusso creato su un determinato host viene eseguito su un host differente, il nome dell'origine dati deve essere lo stesso su entrambi gli host. In alternativa, è possibile selezionare un'origine dati differente nella scheda Server all'interno di ogni nodo Modelli o di input.

Modelli Netezza - Opzioni Modello

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile impostare valori di default per le opzioni di calcolo del punteggio.

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Sostituisci esistente se il nome è stato utilizzato. Se questa casella di controllo viene selezionata, tutti i modelli esistenti con lo stesso nome verranno sovrascritti.

Rendi disponibile per il calcolo del punteggio. È possibile impostare qui i valori di default per le opzioni di calcolo del punteggio che appaiono nella finestra di dialogo del nugget del modello. Per maggiori dettagli sulle opzioni, vedere l'argomento della Guida relativo alla scheda Impostazioni del nugget del modello specifico.

Gestione dei modelli Netezza

Se si crea un modello IBM Netezza tramite IBM SPSS Modeler, viene creato un modello in IBM SPSS Modeler e viene creato o sostituito un modello nel database Netezza. Il modello IBM SPSS Modeler di questo tipo fa riferimento al contenuto di un modello di database archiviato su un server di database. IBM SPSS Modeler consente di eseguire un controllo dell'uniformità archiviando una stringa identica con la chiave del modello generato sia nel modello ISW che nel modello Netezza di IBM SPSS Modeler.

Il nome di ogni modello Netezza viene visualizzato nella colonna *Informazioni sul modello* all'interno della finestra di dialogo Elenco dei modelli in-database. Il nome di un modello di IBM SPSS Modeler viene visualizzato come Chiave di modello nella scheda Server di un modello di IBM SPSS Modeler (se all'interno di uno stream).

È possibile utilizzare il pulsante Controllo per verificare la corrispondenza delle chiavi nel modello Netezza e in quello di IBM SPSS Modeler. Se in Netezza non è reperibile alcun modello con lo stesso nome o se le chiavi del modello non corrispondono, il modello Netezza è stato eliminato o ricreato dopo la creazione del modello di IBM SPSS Modeler.

Elenco dei modelli in-database

IBM SPSS Modeler fornisce una finestra di dialogo per elencare i modelli che sono archiviati in IBM Netezza e consente l'eliminazione di modelli. Tale finestra di dialogo è accessibile dalla finestra Applicazioni di supporto IBM, nonché dalle finestre relative alle operazioni di creazione, visualizzazione e applicazione dei nodi correlati al data mining di IBM Netezza. Di seguito sono riportate le informazioni visualizzate per ogni modello:

- Nome del modello (utilizzato per ordinare l'elenco).
- Nome proprietario.
- L'algoritmo utilizzato nel modello.
- Lo stato attuale del modello, ad esempio, Completo.
- La data di creazione del modello.

Struttura ad albero di regressione Netezza

La struttura ad albero di regressione è un algoritmo basato su strutture ad albero che suddivide più volte un campione di casi per derivare sottoinsiemi dello stesso tipo, in base ai valori di un campo obiettivo numerico. Come le strutture ad albero delle decisioni, le strutture ad albero di regressione decompongono i dati in sottoinsiemi in cui le foglie della struttura ad albero corrispondono a sottoinsiemi sufficientemente piccoli o sufficientemente uniformi. Le suddivisioni vengono selezionate in modo da ridurre la dispersione dei valori dell'attributo obiettivo e quindi consentire una previsione soddisfacente da parte dei valori medi in corrispondenza delle foglie.

Opzioni di creazione della struttura ad albero di regressione Netezza - Espansione della struttura ad albero

È possibile impostare le opzioni per l'accrescimento e la riduzione della struttura ad albero.

Sono disponibili le seguenti opzioni di creazione per l'accrescimento della struttura ad albero:

Profondità massima della struttura ad albero. Numero massimo di foglie fino al quale la struttura ad albero può crescere sotto il nodo root, ovvero il numero di volte che il campione può essere suddiviso in modo ricorsivo. Il valore di default è 62, che è la massima profondità della struttura ad albero ai fini della modellazione.

Nota: Se il visualizzatore nel nugget del modello mostra la rappresentazione del modello, vengono visualizzati un massimo di 12 livelli della struttura ad albero.

Criteri di suddivisione. Queste opzioni controllano il momento in cui interrompere la suddivisione della struttura ad albero. Se non si desidera utilizzare i valori di default, fare clic su **Personalizza** e apportare le modifiche.

- **Misura di valutazione della suddivisione.** Questa misura di valutazione di classe individua il punto migliore in cui suddividere la struttura ad albero.

  Nota: attualmente Varianza è l'unica opzione possibile.
- **Miglioramento minimo per suddivisioni.** La quantità minima in base alla quale l'impurità deve essere ridotta prima di creare una nuova suddivisione nella struttura ad albero. La creazione della struttura ad albero è finalizzata alla creazione di sottogruppi con valori di output simili, per ridurre l'impurità all'interno di ogni nodo. Se la suddivisione migliore di un ramo riduce l'impurità di un valore inferiore a quello specificato dal criterio di suddivisione, il ramo non viene suddiviso.
- **Numero minimo di istanze per suddivisione.** Numero minimo di record che possono essere suddivisi. Quando la quantità di record ancora da suddividere è inferiore a questo numero, non verranno eseguite altre suddivisioni. È possibile utilizzare questo campo per impedire che vengano creati sottogruppi piccoli nella struttura ad albero.

Statistiche. Questo parametro definisce quante statistiche vengono incluse nel modello. Selezionare una delle seguenti opzioni:

- **Tutti.** Vengono incluse tutte le statistiche correlate alle colonne ed ai valori.

Nota: Questo parametro include il numero massimo di statistiche è può quindi influenzare le prestazioni del sistema. Se non si desidera visualizzare il modello in formato grafico, specificare **Nessuno**.

- **Colonne.** Vengono incluse le statistiche correlate alle colonne.
- **Nessuno.** Vengono incluse solo le statistiche richieste per assegnare un punteggio al modello.

Opzioni di creazione della struttura ad albero di regressione Netezza - Taglio della struttura ad albero

Le opzioni di taglio consentono di specificare i criteri con cui la struttura ad albero di regressione viene tagliata. Lo scopo del taglio è ridurre il rischio di sovradattamento rimuovendo i sottogruppi cresciuti troppo che non migliorano la precisione attesa nei nuovi dati.

Misura di taglio. La misura del taglio garantisce che la precisione stimata del modello rimanga entro limiti accettabili dopo la rimozione di una foglia dalla struttura ad albero. È possibile scegliere tra le misure seguenti:

- **mse.** Errore quadratico medio (default): misura la vicinanza di una retta interpolante ai punti dati.
- **r2.** R-quadrato: misura la proporzione di variabilità della variabile dipendente spiegata dal modello di regressione.
- **Pearson.** Coefficiente di correlazione di Pearson: misura la forza della relazione tra le variabili linearmente dipendenti che sono normalmente distribuite.
- **Spearman.** Il coefficiente di correlazione Spearman - rileva relazioni non lineari che sono deboli rispetto alla correlazione Pearson, ma che possono essere in realtà forti.

Dati per il taglio. È possibile utilizzare alcuni o tutti i dati di addestramento per stimare la precisione attesa dei nuovi dati. In alternativa, è possibile utilizzare un insieme di dati di taglio separato estratti da una tabella specifica.

- **Utilizza tutti i dati di addestramento**. Questa opzione (default) utilizza tutti i dati di addestramento per stimare la precisione del modello.
- **Utilizza % dei dati di addestramento per il taglio**. Utilizzare questa opzione per dividere i dati in due insiemi, uno per l'addestramento e uno per il taglio, usando la percentuale qui specificata per i dati del taglio.

  Selezionare **Replica risultati** se si desidera specificare un seme random per assicurarsi che i dati vengano partizionati nello stesso modo ogni volta che si esegue il flusso. È possibile specificare un valore intero nel campo **Seme utilizzato per il taglio** oppure fare clic su **Genera** per creare un intero pseudocasuale.
- **Utilizza dati da una tabella esistente**. Specificare il nome della tabella di un insieme di dati di taglio separato per la stima della precisione del modello. Questa operazione è più affidabile rispetto all'utilizzo dei dati di addestramento. Tuttavia, questa opzione può causare la rimozione di un grande sottoinsieme di dati dal set di addestramento, riducendo la qualità della struttura ad albero delle decisioni.

Raggruppamento cluster divisivo Netezza

Il raggruppamento cluster divisivo è un metodo di analisi in cluster in cui l'algoritmo viene eseguito ripetutamente in modo da suddividere i cluster in cluster secondari finché non si raggiunge un punto di arresto specifico.

La formazione del cluster inizia con un solo cluster contenente tutte le istanze di addestramento (record). La prima iterazione dell'algoritmo divide l'insieme di dati in due cluster secondari, che le successive iterazioni dividono in ulteriori cluster secondari. Il criterio di arresto viene specificato come numero massimo di iterazioni, numero massimo di livelli in cui l'insieme di dati viene suddiviso e numero minimo di istanze necessarie per l'ulteriore partizionamento.

Viene generato un raggruppamento in cluster con struttura gerarchica che consente di classificare le istanze propagandole a partire dal cluster radice, come nell'esempio che segue.

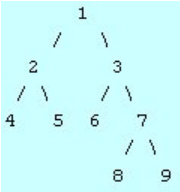

Figura 4. Esempio di struttura ad albero di raggruppamento cluster divisivo

A ogni livello viene scelto il cluster secondario con la migliore corrispondenza in base alla distanza dell'istanza dai centri di cluster secondari.

Quando viene calcolato il punteggio per le istanze con livello -1 della gerarchia (default), il calcolo del punteggio restituisce solo un cluster foglia, poiché le foglie sono identificate da un numero negativo. Nell'esempio, può trattarsi di uno dei cluster 4, 5, 6, 8 o 9. Tuttavia, se il livello della gerarchia è impostato su 2, per esempio, il calcolo del punteggio restituirà uno dei cluster al secondo livello sotto il cluster radice, ovvero 4, 5, 6 o 7.

Opzioni dei campi di raggruppamento cluster divisivo Netezza

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi upstream oppure creare manualmente le assegnazioni dei campi.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo upstream o dalla scheda Tipi di un nodo origine upstream.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione per assegnare manualmente obiettivi, predittori e altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

ID record. Campo da utilizzare come identificatore univoco del record.

Predittori (Input). Scegliere uno o più campi come input per la previsione.

Opzioni di creazione del raggruppamento cluster divisivo Netezza

La scheda Opzioni di creazione contiene tutte le opzioni che consentono di creare il modello. È possibile, ovviamente, fare clic sul pulsante **Esegui** per creare un modello con tutte le opzioni di default, ma normalmente si rende necessario personalizzare il modello in base alle proprie esigenze.

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dati più vicini all'origine.

- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero massimo di iterazioni. L'algoritmo funziona eseguendo una serie di iterazioni dello stesso processo. Questa opzione permette di interrompere l'addestramento del modello dopo il numero di iterazioni specificato.

Massima profondità delle strutture ad albero cluster. Numero massimo di livelli in cui è possibile suddividere l'insieme di dati.

Replica risultati. Selezionare questa casella per impostare un seme random che consentirà di replicare le analisi. È possibile specificare un valore intero o fare clic su **Genera** per creare un intero pseudocasuale.

Numero minimo di istanze per suddivisione. Numero minimo di record che possono essere suddivisi. Quando la quantità di record ancora da suddividere è inferiore a questo numero, non verranno eseguite altre suddivisioni. È possibile utilizzare questo campo per impedire che vengano creati sottogruppi molto piccoli nella struttura ad albero dei cluster.

Lineare generalizzato Netezza

La regressione lineare è una tecnica statistica impiegata da molto tempo che consente di classificare i record in base ai valori dei campi di input numerici. La regressione lineare rappresenta una linea retta o un piano che riduce al minimo le differenze tra i valori di output previsti e quelli effettivi. I modelli lineari sono utili per modellare un'ampia gamma di fenomeni del mondo reale in virtù della loro semplicità di addestramento e di applicazione ai modelli. Tuttavia, i modelli lineari presuppongono una distribuzione normale nella variabile dipendente (obiettivo) e un impatto lineare delle variabili indipendenti (predittori) sulla variabile dipendente.

Esistono molte situazioni in cui una regressione lineare risulta utile ma in cui i presupposti esposti sopra non sono applicabili. Ad esempio, quando si esegue la modellazione delle scelte dei consumatori tra un numero di prodotti discreto, è probabile che la variabile dipendente abbia una distribuzione multinomiale. Analogamente, quando la modellazione del reddito avviene rispetto all'età, generalmente il reddito cresce al crescere dell'età, ma è improbabile che il collegamento tra i due fattori sia una semplice linea retta.

Per questi scenari è possibile utilizzare un modello lineare generalizzato. I modelli lineari generalizzati ampliano il modello di regressione lineare in modo che la variabile dipendente sia correlata alle variabili predittore per mezzo di una funzione di collegamento specifica, per cui esiste una scelta di funzioni adatte. Inoltre, il modello consente alla variabile dipendente di avere una distribuzione non normale ad esempio una distribuzione Poisson.

L'algoritmo esegue una ricerca iterativa del modello più adatto arrivando al numero massimo di iterazioni specificato. Per il calcolo del modello più adatto, l'errore è rappresentato dalla somma dei quadrati delle differenze tra il valore previsto e il valore attuale della variabile dipendente.

Opzioni del campo del modello lineare generalizzato Netezza

Nella scheda Campi, è possibile indicare se si desidera utilizzare le impostazioni del ruolo del campo già definite nei nodi upstream oppure se si desidera eseguire le assegnazioni dei campi manualmente.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli, come obiettivi o predittori da un nodo Tipo upstream oppure dalla scheda Tipi di un nodo di origine upstream.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione se si desidera assegnare manualmente obiettivi, predittori ed altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

Obiettivo. Scegliere un campo come obiettivo per la previsione.

Id record. Campo da utilizzare come identificatore univoco del record. I valori di questo campo devono essere univoci per ciascun record, come, ad esempio, i numeri di ID cliente.

Peso istanza. Specificare un campo per utilizzare i pesi delle istanze. Un peso istanza è un peso per riga di dati di input. Per default, si suppone che tutti i record di input abbiano la stessa importanza relativa. È possibile modificare l'importanza assegnando pesi individuali ai record di input. Il campo specificato deve contenere un peso numerico per ciascuna riga di dati di input.

Predittori (Input). Selezionare il campo o i campi di input. Questa azione è simile all'impostazione del ruolo del campo su *Input* in un nodo Tipo.

Opzioni del modello lineare generalizzato Netezza - Generale

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. Si possono definire anche varie impostazioni relative al modello, alla funzione di collegamento, alle interazioni tra i campi di input (se presenti) e impostare i valori di default per le opzioni di calcolo del punteggio.

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Opzioni campo. È possibile specificare i ruoli dei campi di input per la creazione del modello.

Impostazioni generali. Queste impostazioni si riferiscono ai criteri di arresto dell'algoritmo.

- **Numero massimo di iterazioni.** Il numero massimo di iterazioni che l'algoritmo eseguirà; il numero minimo è 1, il valore predefinito è 20.
- **Numero massimo di errori (1e).** Il valore di errore massimo (in notazione scientifica) raggiunto il quale l'algoritmo deve interrompere la ricerca del modello più adatto. Il numero minimo è 0, il valore predefinito è -3, che indica 1E-3 o 0.001.
- **Soglia dei valori di errore insignificanti (1e).** Il valore (in notazione scientifica) sotto il quale gli errori vengono trattati come se avessero valore zero. Il valore minimo è -1, il valore predefinito è -7, il che indica che i valori di errore al di sotto di 1E-7 (o 0.0000001) sono considerati insignificanti.

Impostazioni distribuzione. Queste impostazioni sono relative alla distribuzione della variabile dipendente (obiettivo).

- **Distribuzione della variabile di risposta** Il tipo di distribuzione; uno di **Bernoulli** (predefinito), **Gaussian**, **Poisson**, **Binomiale**, **Binomiale negativa**, **Wald** (Gaussiana inversa) e **Gamma**.
- **Parametri.** (Solo Poisson o distribuzione binomiale) È necessario specificare una delle seguenti opzioni nel campo **Specifica parametro**:
  - Affinché il parametro venga stimato automaticamente dai dati, selezionare **Default**.
  - Per consentire l'ottimizzazione della quasi verosimiglianza di distribuzione, selezionare **Quasi**.
  - Per specificare in modo esplicito il valore del parametro, selezionare **Esplicito**.

(Solo distribuzione binomiale) È necessario specificare la colonna della tabella di input che deve essere utilizzata come campo delle prove come richiesto dalla distribuzione binomiale. Questa colonna contiene il numero di prove per la distribuzione binomiale.

(Solo distribuzione binomiale negativa) È possibile utilizzare il valore predefinito -1 o specificare un diverso valore di parametro.

Impostazioni funzione di collegamento. Queste impostazioni si riferiscono alla funzione di collegamento, che pone in correlazione la variabile dipendente con le variabili predittore.

- **Funzione di collegamento.** La funzione da utilizzare; una di **Identità**, **Inversa**, **Invnegative**, **Invsquare**, **Sqrt**, **Power**, **Oddspower**, **Log**, **Clog**, **Loglog**, **Cloglog**, **Logit** (predefinito), **Probit**, **Gaussit**, **Cauchit**, **Canbinom**, **Cangeom**, **Cannegbinom**.
- **Parametri.** (solo funzioni di collegamento Power o Oddspower) È possibile specificare un valore di parametro se la funzione di collegamento è **Power** o **Oddspower**. Scegliere se specificare un valore oppure utilizzare il valore predefinito 1.

Opzioni Modello lineare generalizzato Nettezza - Interazione

Il pannello Interazione contiene le opzioni per specificare le interazioni (cioè gli effetti moltiplicativi tra i campi di input).

Interazione colonna. Selezionare questa casella di controllo per specificare le interazioni tra i campi di input. Lasciare la casella vuota se non sono presenti interazioni.

Immettere le interazioni nel modello selezionando uno o più campi nell'elenco di origini e trascinandoli nell'elenco delle interazioni. Il tipo di interazione creato dipende dall'area sensibile nella quale si rilascia l'interazione.

- **Principale.** I campi rilasciati vengono visualizzati come interazioni principali separate in fondo all'elenco delle interazioni.
- **a 2 vie.** Tutte le possibili coppie dei campi rilasciati vengono visualizzate come interazioni a 2 vie in fondo all'elenco delle interazioni.
- **a 3 vie.** Tutti i possibili gruppi di tre dei campi rilasciati vengono visualizzati come interazioni a 3 vie in fondo all'elenco delle interazioni.
- ***.** L'insieme di tutti i campi eliminati viene visualizzato come una singola interazione alla fine dell'elenco delle interazioni.

Includi intercettazione. L'intercettazione viene generalmente inclusa nel modello. Se è possibile presumere che i dati passino attraverso l'origine, l'intercettazione può essere esclusa.

Pulsanti della finestra di dialogo

I pulsanti alla destra della visualizzazione consentono di effettuare modifiche ai termini utilizzati nel modello.

Figura 5. Pulsante Elimina

Eliminare i termini dal modello selezionando i termini che si desidera eliminare e facendo clic sul pulsante Elimina.

Figura 6. Pulsanti Riordina

Riordinare i termini all'interno del modello selezionando i termini che si desidera riordinare e facendo clic sulla freccia Su e Giù.

Figura 7. Pulsante personalizza interazione

Aggiungi termine personalizzato

È possibile specificare interazioni personalizzate nel formato *n1*x1*x1*x1...* Selezionare un campo dall'elenco **Campi**, fare clic sul pulsante freccia destra per aggiungere il campo **Termine personalizzato**, fare clic su **Per***, selezionare il successivo campo, fare clic sul pulsante freccia destra e così via. Quando si è creata l'interazione personalizzata, fare clic su **Aggiungi termine** per ritornare al pannello Interazione.

Opzioni del modello lineare generalizzato Netezza - Opzioni di calcolo del punteggio

Rendi disponibile per il calcolo del punteggio. È possibile impostare qui i valori di default per le opzioni di calcolo del punteggio che appaiono nella finestra di dialogo del nugget del modello. Consultare l'argomento “Nugget del modello lineare generalizzato Netezza - Scheda Impostazioni” a pagina 116 per ulteriori informazioni.

- **Includi campi di input.** Selezionare questa casella di controllo per visualizzare i campi di input nell'output del modello oltre alle previsioni.

Strutture ad albero delle decisioni di Netezza

La struttura ad albero delle decisioni è una struttura gerarchica che rappresenta un modello di classificazione. Con un modello di struttura ad albero delle decisioni, è possibile sviluppare un sistema di classificazione per prevedere o classificare future osservazioni provenienti da un insieme di dati di addestramento. La classificazione assume l'aspetto di una struttura ad albero in cui i rami rappresentano i punti di suddivisione nella classificazione. In tali punti i dati vengono suddivisi in sottogruppi in modo ricorsivo finché non viene raggiunto un punto di arresto. I nodi della struttura ad albero sono punti di arresto noti come **foglie**. Ogni foglia assegna un'etichetta, detta **etichetta di classe**, ai membri del relativo sottogruppo (classe).

Pesi delle istanze e delle classi

Per default, si presume che tutti i record di input e tutte le classi abbiano uguale importanza relativa. Questa impostazione si può modificare assegnando pesi individuali ai membri di uno di questi elementi o di entrambi. Questo può essere utile, per esempio, se i punti dati di addestramento non sono distribuiti in modo realistico tra le categorie. I pesi consentono di applicare una distorsione al modello in modo da compensare le categorie meno rappresentate nei dati. L'incremento del peso di un valore di destinazione dovrebbe aumentare la percentuale di previsioni corrette per quella categoria.

Nel nodo Modelli Struttura ad albero delle decisioni è possibile specificare due tipi di pesi. I **pesi delle istanze** assegnano un peso a ogni riga dei dati di input. I pesi sono generalmente specificati come 1.0 nella maggior parte dei casi con i valori più alti o più bassi assegnati solo a quei casi che sono più o meno importanti rispetto alla maggioranza dei casi come mostrato nella seguente tabella.

Tabella 9. Esempio peso istanza

ID record	Obiettivo	Peso istanza.
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

Pesi delle classi assegnano un peso a ciascuna categoria del campo obiettivo come mostrato nella seguente tabella

Tabella 10. Esempio peso classi

Classe	Peso della classe
drugA	1.0
drugB	1.5

È possibile utilizzare contemporaneamente entrambi i tipi di pesi, nel qual caso essi vengono moltiplicati insieme e utilizzati come peso delle istanze. Quindi, se i due esempi precedenti vengono utilizzati insieme, l'algoritmo dovrebbe utilizzare i pesi delle istanze come mostrato nella seguente tabella.

Tabella 11. Esempio calcolo del peso dell'istanza

ID record	Calcolo	Peso istanza.
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

Opzioni dei campi della struttura ad albero delle decisioni di Netezza

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi upstream oppure creare manualmente le assegnazioni dei campi.

Utilizza ruoli predefiniti Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo a monte o dalla scheda Tipi di un nodo origine a monte.

Utilizza assegnazioni di campo personalizzate Per assegnare manualmente gli obiettivi, predittori e altri ruoli, selezionare questa opzione.

Campi Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

Obiettivo Selezionare un campo come obiettivo per la previsione.

Id record. Campo da utilizzare come identificatore univoco del record. I valori di questo campo devono essere univoci per ogni record (per esempio, i numeri di ID dei clienti).

Peso istanza. Se si specifica un campo, è possibile utilizzare i pesi delle istanze (un peso per ogni riga di dati di input) in aggiunta ai pesi delle classi di default (un peso per ogni categoria per il campo obiettivo) o al posto degli stessi. Il campo da specificare qui deve contenere un peso numerico per ogni riga dei dati di input. Consultare l'argomento “Pesi delle istanze e delle classi” a pagina 91 per ulteriori informazioni.

Predittori (Input). Selezionare il campo o i campi di input. Questa operazione è simile all'impostazione del ruolo di un campo su *Input* in un nodo Tipo.

Opzioni di creazione della struttura ad albero delle decisioni di Netezza

Sono disponibili le seguenti opzioni di creazione per l'accrescimento della struttura ad albero:

Misura dell'accrescimento. Queste opzioni controllano il modo in cui viene misurato l'accrescimento della struttura ad albero.

- **Misura di impurità.** Questa misura individua il punto migliore in cui suddividere la struttura ad albero. Rappresenta la misura della variabilità in un sottogruppo o in un segmento di dati. Valori bassi d'impurità indicano un gruppo in cui molti membri hanno valori simili nei campi criterio o obiettivo.

  Le misure supportate sono **Entropia** e **Gini**. Queste misurazioni sono basate sulle probabilità di appartenenza alle categorie del ramo.
- **Profondità massima della struttura ad albero.** Numero massimo di foglie fino al quale la struttura ad albero può crescere sotto il nodo root, ovvero il numero di volte che il campione può essere suddiviso in modo ricorsivo. Il valore predefinito di questa proprietà è 10, e il valore massimo che è possibile impostare per questa proprietà è 62.

  Nota: Se il visualizzatore nel nugget del modello mostra la rappresentazione del modello, vengono visualizzati un massimo di 12 livelli della struttura ad albero.

Criteri di suddivisione. Queste opzioni controllano il momento in cui interrompere la suddivisione della struttura ad albero.

- **Miglioramento minimo per suddivisioni.** La quantità minima in base alla quale l'impurità deve essere ridotta prima di creare una nuova suddivisione nella struttura ad albero. La creazione della struttura ad albero è finalizzata alla creazione di sottogruppi con valori di output simili, per ridurre l'impurità all'interno di ogni nodo. Se la suddivisione migliore di un ramo riduce l'impurità di un valore inferiore a quello specificato dal criterio di suddivisione, il ramo non viene suddiviso.
- **Numero minimo di istanze per suddivisione.** Numero minimo di record che possono essere suddivisi. Quando la quantità di record ancora da suddividere è inferiore a questo numero, non verranno eseguite altre suddivisioni. È possibile utilizzare questo campo per impedire che vengano creati sottogruppi piccoli nella struttura ad albero.

Statistiche. Questo parametro definisce quante statistiche vengono incluse nel modello. Selezionare una delle seguenti opzioni:

- **Tutti.** Vengono incluse tutte le statistiche correlate alle colonne ed ai valori.

  Nota: Questo parametro include il numero massimo di statistiche è può quindi influenzare le prestazioni del sistema. Se non si desidera visualizzare il modello in formato grafico, specificare **Nessuno**.
- **Colonne.** Vengono incluse le statistiche correlate alle colonne.
- **Nessuno.** Vengono incluse solo le statistiche richieste per assegnare un punteggio al modello.

Nodo della struttura ad albero delle decisioni di Netezza - Pesi delle classi

È possibile assegnare pesi alle singole classi. L'impostazione di default è assegnare il valore 1 a tutte le classi in modo che abbiano lo stesso peso. Specificando valori numerici diversi per i pesi delle singole etichette di classe, si indica all'algoritmo di pesare i set di addestramento delle singole classi.

Per modificare un peso, farvi doppio clic sopra nella colonna **Peso** e apportare le modifiche desiderate.

Valore. L'insieme delle etichette di classe ricavato dai valori possibili del campo obiettivo.

Peso. Il peso da assegnare a una determinata classe. L'assegnazione di un peso superiore a una classe rende il modello più sensibile a quella classe rispetto alle altre.

I pesi delle classi si possono utilizzare insieme ai pesi delle istanze. Consultare l'argomento “Pesi delle istanze e delle classi” a pagina 91 per ulteriori informazioni.

Nodo della struttura ad albero delle decisioni di Netezza - Taglio della struttura ad albero

Le opzioni di taglio consentono di specificare i criteri con cui la struttura ad albero delle decisioni viene tagliato. Lo scopo del taglio è ridurre il rischio di sovradattamento rimuovendo i sottogruppi cresciuti troppo che non migliorano la precisione attesa nei nuovi dati.

Misura di taglio. La misura di default del taglio, **Precisione**, garantisce che la precisione stimata del modello rimanga entro limiti accettabili dopo la rimozione di una foglia dalla struttura ad albero. Utilizzare invece **Precisione ponderata**, se si desidera tenere in considerazione i pesi delle classi quando si applica il taglio.

Dati per il taglio. È possibile utilizzare alcuni o tutti i dati di addestramento per stimare la precisione attesa dei nuovi dati. In alternativa, è possibile utilizzare un insieme di dati di taglio separato estratti da una tabella specifica.

- **Utilizza tutti i dati di addestramento**. Questa opzione (default) utilizza tutti i dati di addestramento per stimare la precisione del modello.
- **Utilizza % dei dati di addestramento per il taglio**. Utilizzare questa opzione per dividere i dati in due insiemi, uno per l'addestramento e uno per il taglio, usando la percentuale qui specificata per i dati del taglio.

  Selezionare **Replica risultati** se si desidera specificare un seme random per assicurarsi che i dati vengano partizionati nello stesso modo ogni volta che si esegue il flusso. È possibile specificare un valore intero nel campo **Seme utilizzato per il taglio** oppure fare clic su **Genera** per creare un intero pseudocasuale.
- **Utilizza dati da una tabella esistente**. Specificare il nome della tabella di un insieme di dati di taglio separato per la stima della precisione del modello. Questa operazione è più affidabile rispetto all'utilizzo dei dati di addestramento. Tuttavia, questa opzione può causare la rimozione di un grande sottoinsieme di dati dal set di addestramento, riducendo la qualità della struttura ad albero delle decisioni.

Regressione lineare Netezza

I modelli lineari prevedono un target continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori. Benché limitati unicamente alla modellazione diretta delle relazioni lineari, i modelli di regressione lineare sono relativamente semplici e forniscono una formula matematica di facile interpretazione per il calcolo del punteggio. I modelli lineari sono veloci, efficaci e facili da utilizzare, anche se meno flessibili rispetto a quelli generati da algoritmi di regressione più sofisticati.

Opzioni di creazione della regressione lineare Netezza

La scheda Opzioni di creazione contiene tutte le opzioni che consentono di creare il modello. È possibile, ovviamente, fare clic sul pulsante **Esegui** per creare un modello con tutte le opzioni di default, ma normalmente si rende necessario personalizzare il modello in base alle proprie esigenze.

Utilizza Decomposizione del valore singolare per risolvere le equazioni. L'utilizzo della matrice di Decomposizione ai valori singolari, anziché della matrice originale, ha il vantaggio di essere robusta rispetto agli errori numerici e allo stesso tempo velocizza il calcolo.

Includi intercettazione nel modello. Includere l'intercettazione aumenta la precisione generale della soluzione.

Calcola la diagnostica del modello. Questa opzione consente di generare calcoli diagnostici per il modello. I risultati vengono archiviati in matrici o tabelle. per l'ultima revisione. La diagnostica include r-quadrato, somma dei quadrati residua, stima della varianza, deviazione standard, valore *p* e valore *t*.

La diagnostica è correlata alla validità e all'utilità del modello. È necessario eseguire la diagnostica separatamente sui dati sottostanti per garantire che soddisfi i presupposti di linearità.

KNN Netezza

L'analisi della approssimità è un metodo che consente la classificazione dei casi in base alla loro somiglianza con altri casi. Questa analisi è stata sviluppata per l'apprendimento automatico, come metodo per riconoscere gli schemi di dati senza che sia necessaria una corrispondenza esatta con gli schemi, o i casi, archiviati. I casi simili sono vicini gli uni agli altri, mentre i casi dissimili sono distanti gli uni dagli altri. Pertanto, la distanza tra due casi è una misura della loro dissimilarità.

I casi che sono vicini gli uni agli altri sono denominati “elementi adiacenti.” Quando viene presentato un nuovo caso (holdout), viene calcolata la sua distanza da ognuno dei casi nel modello. Le classificazioni dei casi più simili - gli elementi adiacenti più simili - vengono contate e il nuovo caso viene posizionato nella categoria che contiene il maggior numero di elementi adiacenti più simili.

È possibile specificare il numero di elementi adiacenti più simili da esaminare; questo valore viene chiamato *k*. Le figure mostrano come possono essere classificati utilizzando due differenti valori di *k*. Quando *k* = 5, il nuovo caso viene posizionato nella categoria *1* poiché la maggior parte di elementi adiacenti più simili appartiene alla categoria *1*. Tuttavia, quando *k* = 9, il nuovo caso viene posizionato nella categoria *0* perché la maggior parte di elementi adiacenti più simili appartiene alla categoria *0*.

L'analisi della approssimità può anche essere usata per calcolare i valori per un target continuo. In questa situazione, per ottenere il valore previsto per il nuovo caso, viene utilizzato il valore obiettivo medio o mediano degli elementi adiacenti più vicini.

Opzioni del modello KNN Netezza - Generale

Nella scheda Opzioni modello - Generale è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile impostare opzioni che controllano il modo in cui il numero di elementi adiacenti più vicini viene calcolato e impostare opzioni per ottimizzare la performance e la precisione del modello.

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Elementi adiacenti

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero di elementi adiacenti più vicini (k). Il numero di elementi adiacenti più vicini relativi ad un caso specifico. L'utilizzo di un numero maggiore di elementi adiacenti non garantisce necessariamente un modello più preciso.

La scelta di *k* controlla la proporzione tra la prevenzione del sovradattamento (può essere importante, soprattutto per i dati "rumorosi") e la risoluzione (con previsioni diverse per istanze simili). Normalmente è necessario adattare il valore di *k* per ogni insieme di dati; i valori tipici variano da 1 a diverse decine.

Ottimizza performance e precisione

Standardizza le misure prima di calcolare la distanza. Se selezionata, questa opzione standardizza le misure per i campi di input continui prima di calcolare i valori della distanza.

Utilizza insiemi centrali per incrementare le prestazioni per dataset di grandi dimensioni Se selezionata, questa opzione utilizza il campionamento degli insiemi centrali per accelerare il calcolo quando si lavora con insiemi di dati di grandi dimensioni.

Opzioni del modello KNN Netezza - Opzioni di calcolo del punteggio

Nella scheda Opzioni modello - Opzioni di calcolo del punteggio è possibile impostare il valore di default per un'opzione di calcolo del punteggio e assegnare pesi relativi alle singole classi.

Rendi disponibile per il calcolo del punteggio

Includi campi di input. Specifica se i campi di input vengono inclusi nel calcolo del punteggio per default.

Pesi delle classi

Utilizzare questa opzione se si desidera modificare l'importanza relativa delle singole classi durante la creazione del modello.

Nota: questa opzione è abilitata solo se si utilizza KNN per la classificazione. Se si esegue una regressione, ovvero il tipo di campo obiettivo è Continuo, l'opzione è disabilitata.

L'impostazione di default è assegnare il valore 1 a tutte le classi in modo che abbiano lo stesso peso. Specificando valori numerici diversi per i pesi delle singole etichette di classe, si indica all'algoritmo di pesare i set di addestramento delle singole classi.

Per modificare un peso, farvi doppio clic sopra nella colonna **Peso** e apportare le modifiche desiderate.

Valore. L'insieme delle etichette di classe ricavato dai valori possibili del campo obiettivo.

Peso. Il peso da assegnare a una determinata classe. L'assegnazione di un peso superiore a una classe rende il modello più sensibile a quella classe rispetto alle altre.

Medie K Netezza

Il nodo Medie K implementa l'algoritmo *k*-medie, che fornisce un metodo di analisi dei cluster. Questo nodo si può utilizzare per raggruppare un insieme di dati in gruppi distinti.

Si tratta di un algoritmo di cluster basato sulla distanza che utilizza una metrica di distanza (funzione) per misurare la similarità fra i punti dati. I punti dati vengono assegnati al cluster più vicino in base alla metrica di distanza utilizzata.

L'algoritmo funziona eseguendo una serie di iterazioni del medesimo processo di base in cui ogni istanza di addestramento viene assegnata al cluster più vicino (rispetto alla funzione di distanza specificata, applicata all'istanza e al centro di cluster). Tutti i centri di cluster vengono in seguito ricalcolati come vettori del valore medio degli attributi delle istanze assegnate a determinati cluster.

Opzioni dei campi Medie K di Netezza

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi upstream oppure creare manualmente le assegnazioni dei campi.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo upstream o dalla scheda Tipi di un nodo origine upstream.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione per assegnare manualmente obiettivi, predittori e altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

ID record. Campo da utilizzare come identificatore univoco del record.

Predittori (Input). Scegliere uno o più campi come input per la previsione.

Scheda Opzioni di creazione K-medie di Netezza

Impostando le opzioni di creazione, è possibile personalizzare la creazione del modello adattandolo alle proprie esigenze.

Se si desidera creare un modello con le opzioni di default, fare clic su **Esegui**.

Misura della distanza. Questo parametro definisce il metodo di misura della distanza tra i punti di dati. Distanze maggiori distanze indicano maggiori differenze. Selezionare una delle seguenti opzioni:

- **Euclidea** La misura euclidea è la distanza in "linea retta" tra due punti di dati.
- **Euclidea normalizzata.** La misura Euclidea normalizzata è simile alla misura Euclidea ma è normalizzata al quadrato della deviazione standard. A differenza della misura Euclidea, quella Euclidea normalizzata varia al variare della scala.
- **Mahalanobis.** La misura Mahalanobis è una misura Euclidea generalizzata che considera le correlazioni dei dati di input. Come la misura Euclidea, la misura Mahalanobis non varia al variare della scala.
- **Manhattan.** La misura Manhattan è la distanza fra due punti di dati calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra.** La misura Canberra e simile a quella Manhattan, ma è più sensibile ai punti di dati più vicini all'origine.
- **Massima.** La misura Massima è la distanza fra due punti di dati calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero di cluster. Questo parametro specifica il numero di cluster da creare.

Numero massimo di iterazioni. L'algoritmo esegue una serie di iterazioni dello stesso processo. Questo parametro specifica il numero di iterazioni dopo cui l'addestramento del modello viene interrotto.

Statistiche. Questo parametro definisce quante statistiche vengono incluse nel modello. Selezionare una delle seguenti opzioni:

- **Tutti.** Vengono incluse tutte le statistiche correlate alle colonne ed ai valori.

  Nota: Questo parametro include il numero massimo di statistiche è può quindi influenzare le prestazioni del sistema. Se non si desidera visualizzare il modello in formato grafico, specificare **Nessuno**.
- **Colonne.** Vengono incluse le statistiche correlate alle colonne.
- **Nessuno.** Vengono incluse solo le statistiche richieste per assegnare un punteggio al modello.

Replica risultati. Selezionare questa casella per impostare un seme random per la replica delle analisi. È possibile specificare un numero intero, o creare un numero intero pseudocasuale facendo clic su **Genera**.

Naive Bayes Netezza

Naive Bayes è un algoritmo molto noto per problemi di classificazione. Il modello è denominato *naive* perché considera tutte le variabili di previsione proposte indipendenti l'una dall'altra. Naive Bayes è un algoritmo veloce e scalabile in grado di calcolare probabilità condizionali per combinazioni di attributi e per l'attributo obiettivo. Dai dati di addestramento viene stabilita una probabilità indipendente che serve a indicare la verosimiglianza di ciascuna classe obiettivo una volta specificata l'occorrenza di ogni categoria di valore da ogni variabile di input.

Rete di Bayes Netezza

Una rete Bayesian è un modello che visualizza le variabili in un dataset e le indipendenze probabilistiche o condizionali tra di esse. Il nodo Rete di Bayes Netezza consente di generare un modello di probabilità combinando elementi osservati e registrati con conoscenze del mondo reale basate sul "buon senso" per stabilire la probabilità delle occorrenze utilizzando attributi apparentemente non collegati fra loro.

Opzioni dei campi della rete di Bayes Netezza

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi upstream oppure creare manualmente le assegnazioni dei campi.

Per questo nodo, il campo obiettivo è necessario solo per il calcolo del punteggio, quindi non viene visualizzato in questa scheda. È possibile impostare o modificare l'obiettivo in un nodo Tipo, nella scheda Opzioni modello di tale nodo o nella scheda Impostazioni del nugget del modello. Consultare l'argomento “Nugget di rete di Bayes Netezza - Scheda Impostazioni” a pagina 110 per ulteriori in formazioni.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo upstream o dalla scheda Tipi di un nodo origine upstream.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione per assegnare manualmente obiettivi, predittori e altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

Predittori (Input). Scegliere uno o più campi come input per la previsione.

Opzioni di creazione della rete di Bayes Netezza

La scheda Opzioni di creazione contiene tutte le opzioni che consentono di creare il modello. È possibile, ovviamente, fare clic sul pulsante **Esegui** per creare un modello con tutte le opzioni di default, ma normalmente si rende necessario personalizzare il modello in base alle proprie esigenze.

Indice di base. Identificatore numerico da assegnare al primo attributo (campo input) per semplificare la gestione interna.

Dimensione campione. Dimensione del campione da utilizzare se il numero degli attributi è talmente grande da allungare il tempo di elaborazione in modo inaccettabile.

Visualizza ulteriori informazioni durante l'esecuzione Se questa casella è selezionata (impostazione predefinita), ulteriori informazioni sull'avanzamento vengono visualizzate in una finestra di dialogo del messaggio.

Serie temporali Netezza

Una **serie temporale** è una sequenza di valori numerici misurati in momenti temporali successivi (anche se non necessariamente a intervalli regolari), ad esempio i prezzi giornalieri delle azioni o i dati di vendita settimanali. L'analisi di questo tipo di dati può essere utile, ad esempio, per evidenziare comportamenti che rivelano tendenze o stagionalità (pattern ripetuti), e per predire comportamenti futuri basandosi su eventi passati.

Serie temporali Netezza supporta i seguenti algoritmi per serie temporali.

- analisi spettrale
- livellamento esponenziale
- Modello Autoregressivo Integrato a Media Mobile (ARIMA)
- scomposizione tendenza stagionale

Lo scopo di questi algoritmi è di estrapolare dalla serie temporale una tendenza o un componente stagionale. I componenti verranno poi analizzati nell'ottica di creare un modello predittivo.

L'**analisi spettrale** individua i comportamenti periodici nelle serie temporali. Per le serie temporali composte di varie periodicità implicite o quando i dati contengono una quantità notevole di rumore casuale, l'analisi spettrale rappresenta il modo più chiaro per individuare i componenti periodici. Per rilevare la frequenza dei comportamenti periodici, questo metodo trasforma la serie temporale dall'ambito temporale all'ambito della frequenza.

Il **livellamento esponenziale** è un metodo di previsione che utilizza i valori ponderati delle osservazioni di serie precedenti per prevedere i valori futuri. Con il livellamento esponenziale, l'influenza delle osservazioni diminuisce nel tempo in modo esponenziale. Questo metodo esegue la previsione di un punto di tempo per volta, rettificando la previsione non appena riceve nuovi dati quali aggiunte, tendenza e stagionalità.

I modelli **ARIMA** forniscono metodi più sofisticati per la modellazione dei componenti di tendenza e stagionali rispetto ai modelli di livellamento esponenziale. Questo metodo comporta l'indicazione esplicita di ordini autoregressivi e di media mobile, nonché del grado di differenziazione.

Nota: in termini pratici, i modelli ARIMA sono utili soprattutto se si desidera includere dei predittori che possono contribuire a spiegare il comportamento della serie oggetto della previsione, quale il numero di cataloghi inviati per posta o il numero di risultati di ricerca ottenuti per la pagina Web di una società. I modelli di livellamento esponenziale descrivono il comportamento della serie temporale senza cercare di spiegare le ragioni di tale comportamento.

La **scomposizione tendenza stagionale** elimina il comportamento periodico dalla serie temporale per eseguire l'analisi della tendenza e quindi seleziona una forma semplice per la tendenza, come una funzione quadratica. Le forme semplici presentano un numero di parametri i cui valori sono determinati in modo da ridurre al minimo l'errore quadratico medio dei residui (vale a dire, le differenze tra i valori previsti e i valori osservati della serie temporale).

Interpolazione dei valori nella serie temporale Netezza

L'**interpolazione** è il processo di stima e inserimento dei valori mancanti nei dati di una serie temporale.

Se gli intervalli di una serie temporale sono regolari ma alcuni valori sono assenti, i valori mancanti potranno essere stimati mediante l'interpolazione lineare. Consideriamo la seguente serie che contiene gli arrivi mensili di passeggeri al terminal di un aeroporto.

Tabella 12. Arrivi mensili presso un terminal passeggeri

Mese	Passeggeri
3	3,500,000
4	3,900,000
5	-
6	3,400,000
7	4,500,000
8	3,900,000
9	5,800,000
10	6,000,000

In questo caso, l'interpolazione lineare stimerebbe il valore mancante per il mese 5 come 3.650.000 (il punto intermedio tra i mesi 4 e 6).

Gli intervalli irregolari vengono gestiti in modo diverso. Consideriamo la seguente serie di letture della temperatura.

Tabella 13. Letture temperatura

Data	Ora	Temperatura
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

In questo caso abbiamo letture rilevate in tre momenti nel corso di tre giorni, ma in orari diversi che si ripetono solo in alcuni giorni. Inoltre, solo due giorni su tre sono consecutivi.

Questa situazione può essere gestita in uno dei due modi: calcolando gli aggregati o determinando la dimensione della fase.

Gli aggregati potrebbero essere aggregati quotidiani calcolati con una formula basata sulla conoscenza semantica dei dati. Questa soluzione restituirebbe il seguente insieme di dati.

Tabella 14. Letture temperatura (aggregate)

Data	Ora	Temperatura
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	null
2011-07-27	24:00	72

In alternativa, l'algoritmo può trattare la serie come una serie distinta e determinare un'adeguata dimensione di passo. In questo caso, la dimensione di passo determinata dall'algoritmo potrebbe essere 8 ore, il che restituirebbe quanto segue.

Tabella 15. Letture temperatura con dimensione di passo

Data	Ora	Temperatura
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

In questo caso, solo quattro letture corrispondono alle letture originali, ma aiutandosi con altri valori conosciuti della serie originale, i valori mancanti possono essere nuovamente calcolati mediante l'interpolazione.

Opzioni dei campi della serie temporale Netezza

Nella scheda Campi si specificano i ruoli per i campi di input nell'origine dati.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Obiettivo. Scegliere un campo come obiettivo per la previsione. Questo campo deve avere un livello di misurazione Continuo.

(Predittore) - Punti di tempo. (obbligatorio) Il campo di input contenente i valori di data o ora per la serie temporale. Questo campo deve avere un livello di misurazione Continuo o Categoriale e un tipo di archiviazione dati Data, Ora, Timestamp o Numerico. Il tipo di archiviazione dati del campo specificato qui definisce anche il tipo di input di alcuni campi di altre schede di questo stesso nodo Modelli.

(Predittore) - ID serie temporali (Da). Campo contenente ID di serie temporali; utilizzarlo se l'input contiene più di una serie temporale.

Opzioni di creazione della serie temporale Netezza

Esistono due livelli di opzioni di creazione:

- Di base - impostazioni per la scelta dell'algoritmo, dell'interpolazione e dell'intervallo di tempo da usare.
- Opzioni avanzate - impostazione per la previsione

Questa sezione descrive le opzioni di base.

La scheda Opzioni di creazione contiene tutte le opzioni che consentono di creare il modello. È possibile, ovviamente, fare clic sul pulsante **Esegui** per creare un modello con tutte le opzioni di default, ma normalmente si rende necessario personalizzare il modello in base alle proprie esigenze.

Algoritmo

Sono le impostazioni relative all'algoritmo di serie temporali da utilizzare.

Nome algoritmo. Scegliere l'algoritmo di serie temporali da utilizzare. Gli algoritmi disponibili sono **Analisi spettrale**, **Livellamento esponenziale** (default), **ARIMA** e **Scomposizione tendenza stagionale**. Consultare l'argomento "Serie temporali Netezza" a pagina 99 per ulteriori informazioni.

Tendenza. (solo per Livellamento esponenziale) Il livellamento esponenziale semplice non restituisce buoni risultati se la serie temporale presenta una tendenza. Utilizzare questo campo per specificare la tendenza, se presente, in modo che l'algoritmo possa tenerne conto.

- **Determinato dal sistema.** (default) Il sistema tenta di trovare il valore ottimale per questo parametro.
- **Nessuno (N).** La serie temporale non presenta alcuna tendenza.
- **Additivo(A).** Tendenza che cresce nel tempo in maniera costante.
- **Additivo smorzato(DA).** Tendenza additiva che finisce per esaurirsi.
- **Moltiplicativo(M).** Tendenza che cresce nel tempo, generalmente in maniera più rapida rispetto a una tendenza additiva costante.
- **Moltiplicativo smorzato(DM).** Tendenza moltiplicativa che finisce per esaurirsi.

Stagionalità. (solo per Livellamento esponenziale) Utilizzare questo campo per specificare se i dati della serie temporale presentano pattern stagionali.

- **Determinato dal sistema.** (default) Il sistema tenta di trovare il valore ottimale per questo parametro.
- **Nessuno(N).** Le serie temporali non presentano pattern stagionali.
- **Additivo(A).** Il pattern delle fluttuazioni stagionali presenta una tendenza costante verso l'alto nel tempo.
- **Moltiplicativo(M).** Come la stagionalità additiva, con l'aggiunta che l'ampiezza (distanza tra i punti alto e basso) delle fluttuazioni stagionali cresce rispetto alla tendenza verso l'alto complessivo delle fluttuazioni.

Utilizza le impostazioni determinate dal sistema per ARIMA. (solo per ARIMA) Scegliere questa opzione per lasciare che il sistema determini le impostazioni per l'algoritmo ARIMA.

Specifica. (solo ARIMA) Scegliere questa opzione e fare clic sul pulsante per specificare manualmente le impostazioni per ARIMA.

Interpolazione

Se i dati di origine della serie temporale presentano dei valori mancanti, scegliere un metodo che consenta di riempire valori stimati al posto dei valori mancanti. Consultare l'argomento "Interpolazione dei valori nella serie temporale Netezza" a pagina 100 per ulteriori informazioni.

- **Lineare.** Scegliere questo metodo se gli intervalli della serie temporale sono regolari e alcuni valori semplicemente non sono dati.
- **Spline esponenziali.** Applica una curva dove i punti dati conosciuti aumentano o diminuiscono con una frequenza elevata.
- **Spline cubiche.** Applica una curva ai punti dati conosciuti per stimare i valori mancanti.

Intervallo temporale

Qui è possibile scegliere se, per creare il modello, si vuole utilizzare la gamma completa dei dati della serie temporale o un sottoinsieme contiguo di tali dati. Gli input validi per questi campi sono definiti dal tipo di archiviazione dei dati specificato per i Punti di tempo della scheda Campi. Consultare l'argomento "Opzioni dei campi della serie temporale Netezza" a pagina 101 per ulteriori informazioni.

- **Utilizza le date più lontane e più vicine disponibili nei dati.** Scegliere questa opzione per utilizzare la gamma completa dei dati della serie temporale.
- **Specifica la finestra temporale.** Scegliere questa opzione per utilizzare solo una porzione della serie temporale. Utilizzare i campi **Data più lontana (da)** e **Data più vicina (a)** per specificare i limiti.

Struttura ARIMA

Specificare i valori dei vari componenti stagionali e non stagionali del modello ARIMA. In ogni caso, impostare l'operatore a = (uguale a o <= (minore o uguale a), quindi specificare il valore nel campo adiacente. I valori devono essere numeri interi non negativi che specificano i gradi

Non stagionale. I valori dei vari componenti non stagionali del modello.

- **Gradi di autocorrelazione (p).** Il numero di ordini autoregressivi nel modello. Gli ordini autoregressivi specificano quali valori precedenti della serie vengono utilizzati per prevedere i valori correnti. Per esempio, un ordine autoregressivo 2 specifica di utilizzare il valore dei due periodi precedenti della serie per prevedere il valore corrente.
- **Derivazione (d).** Specifica l'ordine di differenziazione applicato alla serie prima di eseguire la stima dei modelli. La differenziazione è necessaria quando sono presenti delle tendenze (di norma, le serie che presentano delle tendenze sono non stazionarie e nei modelli ARIMA si presume che vi sia stazionarietà) e viene utilizzata per rimuoverne l'effetto. L'ordine di differenziazione corrisponde al grado di tendenza della serie, la differenziazione di primo grado tiene conto delle tendenze lineari, la differenziazione di secondo grado tiene conto delle tendenze quadratiche e così via.
- **Media mobile (q).** Il numero di ordini di media mobile nel modello. Gli ordini di media mobile specificano il modo in cui vengono utilizzate le deviazioni provenienti dalla media della serie per prevedere i valori correnti. Per esempio, gli ordini di media mobile 1 e 2 specificano di considerare le deviazioni dalla media della serie degli ultimi due periodi precedenti per prevedere i valori correnti della serie.

Stagionale. I componenti Autocorrelazione stagionale (SP), Derivazione (SD) e Media mobile (SQ) svolgono le stesse funzioni delle rispettive controparti non stagionali. Per gli ordini stagionali tuttavia, i valori di serie correnti vengono influenzati dai valori di serie precedenti separati da uno o più periodi stagionali. Ad esempio, per i dati mensili (periodo stagionale di 12), un ordine stagionale 1 è il valore della serie corrente è influenzato dal valore della serie che precede di 12 periodi quello corrente. Specificare un ordine stagionale 1, per i dati mensili, è quindi come specificare un ordine non stagionale 12.

Le impostazioni stagionali sono prese in considerazione solo se la stagionalità è rilevata nei dati oppure se si specificano impostazioni di Periodo nella scheda Avanzate.

Opzioni di creazione della serie temporale Netezza - Avanzate

Le impostazioni avanzate consentono di specificare opzioni per la previsione.

Utilizza le impostazioni determinate dal sistema per le opzioni di creazione del modello. Selezionare questa opzione per lasciare che il sistema determini le impostazioni avanzate.

Specifica. Selezionare questa opzione per specificare manualmente le opzioni avanzate. (L'opzione non è disponibile se l'algoritmo è analisi spettrale.)

- **Periodo/Unità per periodo.** Il periodo di tempo trascorso il quale un dato comportamento caratteristico della serie temporale si ripete. Ad esempio, per una serie temporale di dati di vendita settimanali si definirebbe 1 per il periodo e `Settimane` per le unità. **Periodo** deve essere un intero non negativo; **Unità di periodo** può essere **Millisecondi**, **Secondi**, **Minuti**, **Ore**, **Giorni**, **Settimane**, **Trimestri** o **Anni**. Non impostare **Unità di periodo** se **Periodo** non è impostato o se il tipo tempo non è numerico. Tuttavia, se si specifica **Periodo**, è necessario anche specificare **Unità di periodo**.

Impostazioni per la previsione. Si può scegliere di effettuare previsioni fino a uno specifico punto di tempo o in precisi punti di tempo. Gli input validi per questi campi sono definiti dal tipo di archiviazione dei dati specificato per i Punti di tempo della scheda Campi. Consultare l'argomento "Opzioni dei campi della serie temporale Netezza" a pagina 101 per ulteriori informazioni.

- **Orizzonte di previsione.** Selezionare questa opzione per specificare solo un punto finale in cui interrompere la previsione. Le previsioni verranno effettuate fino a questo punto di tempo.
- **Tempi di previsione.** Selezionare questa opzione per specificare uno o più punti di tempo per cui effettuare delle previsioni. Fare clic su **Aggiungi** per aggiungere una nuova riga alla tabella dei punti di tempo. Per eliminare una riga, selezionarla e fare clic su **Elimina**.

Opzioni del modello di serie temporali Netezza

Nella scheda Opzioni modello è possibile scegliere se specificare un nome per il modello o generare un nome automaticamente. È inoltre possibile impostare valori di default per le opzioni di output del modello.

Nome modello È possibile generare il nome del modello automaticamente in base al campo ID o obiettivo (oppure il tipo di modello nei casi non sia specificato tale campo) oppure indicare un nome personalizzato.

Rendi disponibile per il calcolo del punteggio. È possibile impostare qui i valori di default per le opzioni di calcolo del punteggio che appaiono nella finestra di dialogo del nugget del modello.

- **Includi valori storici nel risultato.** Di default, l'output del modello non include i valori cronologici (quelli utilizzati per effettuare la previsione). Selezionare questa casella di controllo per includere questi valori.
- **Includi valori interpolati nel risultato.** Se si opta per includere valori storici nell'output, selezionare questa casella per includere anche i valori interpolati, se presenti. Considerare che l'interpolazione funziona solo sui dati storici quindi questa casella non è disponibile se non è selezionata l'opzione **Includi valori storici nel risultato**. Consultare l'argomento "Interpolazione dei valori nella serie temporale Netezza" a pagina 100 per ulteriori informazioni.

Netezza TwoStep

Il nodo TwoStep implementa l'algoritmo TwoStep che fornisce un metodo per raggruppare i dati distribuiti in dataset di grandi dimensioni.

È possibile utilizzare questo nodo per raggruppare i dati mentre vengono prese in considerazione le risorse disponibili, ad esempio vincoli di tempo e memoria.

L'algoritmo TwoStep è un algoritmo di mining del database che raggruppa i dati nel modo seguente:

1. Viene creata una struttura ad albero CF (clustering feature). Questa struttura ad albero altamente bilanciata memorizza le CF (clustering features) per il raggruppamento cluster gerarchico in cui record di input simili diventano parte degli stessi nodi della struttura ad albero.
2. Le foglie della struttura ad albero CF sono raggruppate in memoria in modo gerarchico, per generare il risultato di raggruppamento finale. Il numero ottimale di cluster viene determinato automaticamente. Se si specifica un numero massimo di cluster, viene determinato il numero ottimale di cluster nel limite specificato.
3. Il risultato di raggruppamento viene rifinito in una seconda fase in cui ai dati viene applicato un algoritmo simile all'algoritmo Medie K.

Opzioni del campo TwoStep di Netezza

Impostando le opzioni del campo, è possibile specificare di utilizzare le impostazioni del ruolo del campo definite nei nodi upstream. È anche possibile effettuare le assegnazioni del campo manualmente.

Seleziona un elemento. Scegliere questa opzione per utilizzare le impostazioni del ruolo da un nodo Tipo upstream oppure dalla scheda Tipi di un nodo di origine upstream. Le impostazioni del ruolo sono, ad esempio, obiettivi e predittori.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione se si desidera assegnare obiettivi, predittori ed altri ruoli manualmente.

Campi. Utilizzare le frecce per assegnare manualmente le voci da questo elenco ai campi di ruolo a destra. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Id record. Campo da utilizzare come identificatore univoco del record.

Predittori (Input). Scegliere uno o più campi come input per la previsione.

Opzioni di creazione TwoStep Netezza

Impostando le opzioni di creazione, è possibile personalizzare la creazione del modello adattandolo alle proprie esigenze.

Se si desidera creare un modello con le opzioni di default, fare clic su **Esegui**.

Misura della distanza. Questo parametro definisce il metodo di misura della distanza tra i punti di dati. Distanze maggiori distanze indicano maggiori differenze. Le opzioni disponibili sono:

- **Verosimiglianza logaritmica.** La misura di verosimiglianza applica una distribuzione delle probabilità alle variabili. Si suppone che le variabili continue vengano distribuite normalmente e che le variabili categoriali siano multinomiale. Si suppone che tutte le variabili siano indipendenti.
- **Euclidea** La misura euclidea è la distanza in "linea retta" tra due punti di dati.
- **Euclidea normalizzata.** La misura Euclidea normalizzata è simile alla misura Euclidea ma è normalizzata al quadrato della deviazione standard. A differenza della misura Euclidea, quella Euclidea normalizzata varia al variare della scala.

Numero cluster. Questo parametro specifica il numero di cluster da creare. Le opzioni disponibili sono:

- **Calcola automaticamente il numero di cluster.** Il numero di cluster viene calcolato automaticamente. È possibile specificare il numero massimo di cluster nel campo **Massimo**.
- **Specifica numero di cluster.** Specificare il numero di cluster da creare.

Statistiche. Questo parametro definisce quante statistiche vengono incluse nel modello. Le opzioni disponibili sono:

- **Tutti.** Vengono incluse tutte le statistiche correlate alle colonne ed ai valori.

Nota: Questo parametro include il numero massimo di statistiche è può quindi influenzare le prestazioni del sistema. Se non si desidera visualizzare il modello in formato grafico, specificare **Nessuno**.

- **Colonne.** Vengono incluse le statistiche correlate alle colonne.
- **Nessuno.** Vengono incluse solo le statistiche richieste per assegnare un punteggio al modello.

Replica risultati. Selezionare questa casella per impostare un seme random per la replica delle analisi. È possibile specificare un numero intero, o creare un numero intero pseudocasuale facendo clic su **Genera**.

PCA Netezza

Principal Component Analysis (PCA), o analisi delle componenti principali, è una tecnica efficace progettata appositamente per ridurre la complessità dei dati. PCA trova le combinazioni lineari dei campi di input che catturano meglio la varianza nell'intero insieme di campi, dove le componenti sono ortogonali (e non correlate) le une rispetto alle altre. Lo scopo è individuare un numero limitato di campi derivati (le componenti principali) contenenti un riepilogo efficace delle informazioni presenti nell'insieme originale di campi di input.

Opzioni dei campi PCA Netezza

Nella scheda Campi è possibile indicare se si desidera utilizzare le impostazioni dei ruoli dei campi già definite nei nodi upstream oppure creare manualmente le assegnazioni dei campi.

Utilizza ruoli predefiniti. Questa opzione utilizza le impostazioni dei ruoli (obiettivi, predittori e così via) ottenute da un nodo Tipo upstream o dalla scheda Tipi di un nodo origine upstream.

Utilizza assegnazioni campi personalizzate. Scegliere questa opzione per assegnare manualmente obiettivi, predittori e altri ruoli in questa schermata.

Campi. Utilizzare i pulsanti con le frecce per assegnare manualmente le voci dell'elenco ai vari campi dei ruoli situati a destra sullo schermo. Le icone indicano i livelli di misurazione validi per ogni campo di ruolo.

Fare clic sul pulsante **Tutto** per selezionare tutti i campi dell'elenco o fare clic su un singolo pulsante di livello di misurazione per selezionare tutti i campi con tale livello.

ID record. Campo da utilizzare come identificatore univoco del record.

Predittori (Input). Scegliere uno o più campi come input per la previsione.

Opzioni di creazione PCA Netezza

La scheda Opzioni di creazione contiene tutte le opzioni che consentono di creare il modello. È possibile, ovviamente, fare clic sul pulsante **Esegui** per creare un modello con tutte le opzioni di default, ma normalmente si rende necessario personalizzare il modello in base alle proprie esigenze.

Centra i dati prima di calcolare PCA. Se selezionata (default), questa opzione esegue la centratura dei dati (nota anche come "sottrazione delle medie") prima dell'analisi. La centratura dei dati è necessaria per assicurarsi che la prima componente principale descriva la direzione della varianza massima, altrimenti la componente potrebbe corrispondere maggiormente alla media dei dati. Deselezionare questa opzione per migliorare la performance solo se i dati sono già stati preparati in questo modo.

Esegui scala dati prima di calcolare PCA. Questa opzione esegue la scala dei dati prima dell'analisi. Questa operazione può ridurre l'arbitrarietà dell'analisi quando vengono misurate diverse variabili in diverse unità. La forma più semplice di scala dei dati consiste nel dividere ogni variabile per la sua variazione standard.

Utilizza metodo più rapido, ma meno preciso per calcolare PCA. Questa opzione indica all'algoritmo di utilizzare un metodo meno preciso ma più rapido (forceEigensolve) per trovare le componenti principali.

Gestione di modelli di IBM Netezza Analytics

I modelli IBM Netezza Analytics vengono aggiunti all'area e alla palette Modelli secondo modalità analoghe agli altri modelli di IBM SPSS Modeler e si possono utilizzare praticamente nello stesso modo. Tuttavia, esistono alcune importanti differenze, dato che ogni modello IBM Netezza Analytics creato in IBM SPSS Modeler fa in realtà riferimento a un modello archiviato in un server di database. Quindi affinché un flusso funzioni correttamente deve connettersi al database su cui è stato creato il modello, e la tabella del modello non deve essere stata cambiata da un processo esterno.

Calcolo del punteggio dei modelli IBM Netezza Analytics

I modelli sono rappresentati nell'area da un'icona di nugget del modello dorata. Scopo principale di un nugget è calcolare il punteggio dei dati per generare previsioni o consentire ulteriori analisi delle proprietà del modello. I punteggi vengono aggiunti sotto forma di uno o più campi di dati aggiuntivi che possono essere visualizzati allegando un nodo Tabella al nugget ed eseguendo quel ramo del flusso, come descritto nella sezione che segue. Alcune finestre di dialogo dei nugget, ad esempio quelle relative alla struttura ad albero delle decisioni o alla struttura ad albero di regressione, contengono anche una scheda Modello che fornisce una rappresentazione grafica del modello.

I campi aggiuntivi sono individuati tramite il prefisso `$<id>-` aggiunto al nome del campo obiettivo in cui `<id>` dipende dal modello e identifica il tipo di informazioni che si sta aggiungendo. I diversi identificatori vengono descritti negli argomenti per ogni nugget del modello.

Per visualizzare i punteggi, completare la seguente procedura:

1. Collegare un nodo Tabella al nugget del modello.
2. Aprire il nodo Tabella.
3. Fare clic su **Esegui**.
4. Scorrere verso destra nella finestra di output della tabella per visualizzare i campi aggiuntivi e i relativi punteggi.

Scheda Server del nugget del modello Netezza

Nella scheda Server è possibile impostare le opzioni del server per il calcolo del punteggio del modello. È possibile continuare a utilizzare una connessione al server specificata come upstream oppure spostare i dati in un altro database specificato qui.

Dettagli Server DB Netezza. Qui si specificano i dettagli della connessione per il database da utilizzare per il modello.

- **Utilizza connessione upstream**. (default) Utilizza i dettagli di connessione specificati in un nodo upstream, per esempio il nodo origine del database. *Nota*: questa opzione funziona solo se tutti i nodi upstream sono in grado di utilizzare Push back SQL. In questo caso non è necessario spostare i dati fuori dal database perché SQL supporta pienamente tutti i nodi upstream.
- **Sposta dati alla connessione**. Sposta i dati nel database indicato qui. In questo modo si consente al modello di lavorare se i dati si trovano su un altro database IBM Netezza, o su un database di un altro fornitore, o anche se i dati si trovano in un file flat. Inoltre i dati vengono riportati in questo database se sono stati in precedenza estratti perché un nodo non ha effettuato il push back SQL. Fare clic sul pulsante **Modifica** per reperire e selezionare una connessione. *Attenzione*: IBM Netezza Analytics viene generalmente utilizzato con data set molto grandi. Il trasferimento di grandi quantità di dati tra database, o anche dentro e fuori lo stesso database, può richiedere molto tempo ed è quindi da evitare se possibile.

Nome modello. Il nome del modello. Il nome viene visualizzato solo a scopo informativo; non è possibile modificarlo qui.

Nugget del modello Struttura ad albero delle decisioni di Netezza

Il nugget del modello Struttura ad albero delle decisioni visualizza l'output prodotto dall'operazione di modellazione e consente anche di impostare alcune opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello della struttura ad albero delle decisioni, per default il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 16. Campo di calcolo del punteggio dei modelli per la struttura ad albero delle decisioni.

Nome del campo aggiunto	Significato
$I-*nome_obiettivo*	Valore previsto per il record corrente.

Se si seleziona l'opzione **Calcola probabilità di classi assegnate per il punteggio dei record** nel nodo Modelli o nel nugget del modello e si esegue il flusso, viene aggiunto un ulteriore campo.

Tabella 17. Campo di calcolo del punteggio dei modelli per la struttura ad albero delle decisioni - aggiuntivo.

Nome del campo aggiunto	Significato
$IP-*nome_obiettivo*	Valore di confidenza (da 0,0 a 1,0) per la previsione.

Nugget della struttura ad albero delle decisioni di Netezza - Scheda Modello

La scheda **Modello** mostra l'importanza predittore del modello di struttura ad albero di decisione in formato grafico. La lunghezza della barra rappresenta l'importanza del predittore.

Nota: Quando si utilizza IBM Netezza Analytics Versione 2.x o precedente, il contenuto del modello di struttura ad albero di decisione viene mostrato solo in formato testo.

Per queste versioni, vengono visualizzate le seguenti informazioni:

- Ogni riga di testo corrisponde a un nodo o una foglia.
- Il rientro riflette il livello della struttura ad albero.
- Per un nodo, viene visualizzata la condizione di suddivisione.
- Per una foglia appare l'etichetta di classe assegnata.

Nugget della struttura ad albero delle decisioni di Netezza - Scheda Impostazioni

La scheda Impostazioni consente di impostare alcune opzioni di calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Calcola probabilità di classi assegnate per il calcolo del punteggio dei record. (solo Struttura ad albero delle decisioni e Naive Bayes) Se selezionata, questa opzione indica che i campi di modellazione in più contengono un campo della confidenza (ovvero una probabilità) oltre al campo della previsione. Se si deseleziona questa casella di controllo viene generato solo il campo della previsione.

Utilizza dati di input deterministici. Se selezionata, questa opzione garantisce che qualsiasi algoritmo Netezza che esegue più passaggi della stessa vista utilizzerà lo stesso insieme di dati per ciascun passaggio. Se questa casella di controllo viene selezionata per indicare che vengono utilizzati dati non deterministici, viene creata una tabella temporanea che contiene i dati di output per l'elaborazione, come quelli prodotti da un nodo di partizione; questa tabella viene eliminata dopo la creazione del modello.

Nugget della struttura ad albero delle decisioni di Netezza - Scheda Visualizzatore

La scheda **Visualizzatore** mostra una presentazione di struttura ad albero del modello della struttura ad albero nello stesso modo in cui SPSS Modeler visualizza il proprio modello della struttura ad albero delle decisioni.

Nota: Se il modello è creato con IBM Netezza Analytics Versione 2.x o versione precedente, la scheda **Visualizzatore** è vuota.

Nugget del modello Medie K di Netezza

I nugget del modello Medie K contengono tutte le informazioni intercettate dal modello di cluster, nonché le informazioni sui dati di addestramento e sull'elaborazione della stima.

Quando si esegue un flusso che contiene un nugget del modello Medie K, il nodo aggiunge due nuovi campi che contengono l'appartenenza al cluster e la distanza dal centro del cluster assegnato per tale record. Il nuovo campo denominato $KM-K-Means è relativo all'appartenenza al cluster ed il nuovo campo denominato $KMD-K-Means è relativo alla distanza dal centro del cluster.

Netezza del modello K-medie di Netezza - Scheda Modello

La scheda **Modello** contiene diverse viste grafiche che mostrano distribuzioni e statistiche di riepilogo per i campi dei cluster. È possibile esportare i dati dal modello o le viste come grafici.

Quando si lavora con IBM Netezza Analytics Version 2.x o precedente o quando si crea il modello con Mahalanobis come misura della distanza, il contenuto dei modelli K-Means viene visualizzato solo in formato testuale.

Per queste versioni, vengono visualizzate le seguenti informazioni:

- **Statistiche di riepilogo.** Le statistiche di riepilogo mostrano il numero di record sia per i cluster piccoli che per quelli grandi. Viene inoltre visualizzata la percentuale di dataset rilevata da tali cluster. L'elenco mostra inoltre il rapporto fra le dimensioni del cluster più grande e quelle del più piccolo.
- **Riepilogo cluster**. Il riepilogo cluster elenca i cluster creati dall'algoritmo. Per ogni cluster, la tabella mostra il numero di record e la loro distanza media dal centro di cluster.

Nugget del modello Medie K di Netezza - Scheda Impostazioni

La scheda Impostazioni consente di impostare alcune opzioni di calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Nugget del modello di rete di Bayes Netezza

Il nugget del modello di rete di Bayes consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello di rete di Bayes, il nodo aggiunge un nuovo campo, il cui nome è derivato dal nome obiettivo.

Tabella 18. Campo di calcolo del punteggio dei modelli per rete di Bayes.

Nome del campo aggiunto	Significato
$BN-*nome_obiettivo*	Valore previsto per il record corrente.

Per visualizzare il campo aggiuntivo, collegare un nodo Tabella al nugget del modello ed eseguire il nodo Tabella.

Nugget di rete di Bayes Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Obiettivo. Se si desidera calcolare il punteggio di un campo obiettivo diverso dall'obiettivo corrente, scegliere qui il nuovo obiettivo.

ID record. Se non è specificato un campo ID record, scegliere qui il campo da utilizzare.

Tipo di previsione. La variazione dell'algoritmo di previsione da utilizzare:

- **Migliore (l'elemento adiacente con maggiore correlazione).** (default) Utilizza il nodo elemento adiacente con la maggiore correlazione.
- **Elementi adiacenti (previsione ponderata degli elementi adiacenti).** Utilizza una previsione ponderata di tutti i nodi elementi adiacenti.
- **Elementi adiacenti-NN (elementi adiacenti non null).** Equivale all'opzione precedente tranne per il fatto che ignora i nodi con valori null, ovvero i nodi corrispondenti agli attributi con valori mancanti per l'istanza per cui viene calcolata la previsione.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Nugget del modello Naive Bayes Netezza

Il nugget del modello Naive Bayes consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello Naive Bayes, per default il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 19. Campo di calcolo del punteggio dei modelli per Naive Bayes - default.

Nome del campo aggiunto	Significato
$I-*nome_obiettivo*	Valore previsto per il record corrente.

Se si seleziona l'opzione **Calcola probabilità di classi assegnate per il punteggio dei record** nel nodo Modelli o nel nugget del modello e si esegue il flusso, vengono aggiunti altri due campi.

Tabella 20. Campi di calcolo del punteggio dei modelli per Naive Bayes - aggiuntivi.

Nome del campo aggiunto	Significato
$IP-*nome_obiettivo*	Numeratore bayesiano della classe per l'istanza, ovvero il prodotto della probabilità della classe a priori e delle probabilità del valore dell'attributo dell'istanza condizionale.

Tabella 20. Campi di calcolo del punteggio dei modelli per Naive Bayes - aggiuntivi (Continua).

Nome del campo aggiunto	Significato
$ILP-*nome_obiettivo*	Algoritmo naturale del secondo elemento.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella al nugget del modello ed eseguire il nodo Tabella.

Nugget Naive Bayes Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Calcola probabilità di classi assegnate per il calcolo del punteggio dei record. (solo Struttura ad albero delle decisioni e Naive Bayes) Se selezionata, questa opzione indica che i campi di modellazione in più contengono un campo della confidenza (ovvero una probabilità) oltre al campo della previsione. Se si deseleziona questa casella di controllo viene generato solo il campo della previsione.

Migliora la precisione di probabilità per dataset piccoli o notevolmente non bilanciati. Quando si calcolano le probabilità, questa opzione richiama la tecnica di stima *m* per evitare le probabilità zero durante la stima. Questo tipo di stima delle probabilità può essere più lento ma fornisce risultati migliori per gli insiemi di dati di piccole dimensioni o notevolmente sbilanciati.

Nugget del modello KNN Netezza

Il nugget del modello KNN consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello KNN, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 21. Campo di calcolo del punteggio dei modelli per KNN.

Nome del campo aggiunto	Significato
$KNN-*nome_obiettivo*	Valore previsto per il record corrente.

Per visualizzare il campo aggiuntivo, collegare un nodo Tabella al nugget del modello ed eseguire il nodo Tabella.

Nugget KNN Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Numero di elementi adiacenti più vicini (k). Il numero di elementi adiacenti più vicini relativi ad un caso specifico. L'utilizzo di un numero maggiore di elementi adiacenti non garantisce necessariamente un modello più preciso.

La scelta di *k* controlla la proporzione tra la prevenzione del sovradattamento (può essere importante, soprattutto per i dati "rumorosi") e la risoluzione (con previsioni diverse per istanze simili). Normalmente è necessario adattare il valore di *k* per ogni insieme di dati; i valori tipici variano da 1 a diverse decine.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Standardizza le misure prima di calcolare la distanza. Se selezionata, questa opzione standardizza le misure per i campi di input continui prima di calcolare i valori della distanza.

Utilizza insiemi centrali per incrementare le prestazioni per dataset di grandi dimensioni Se selezionata, questa opzione utilizza il campionamento degli insiemi centrali per accelerare il calcolo quando si lavora con insiemi di dati di grandi dimensioni.

Nugget del modello di raggruppamento cluster divisivo Netezza

Il nugget del modello di raggruppamento cluster divisivo consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello di raggruppamento cluster divisivo, il nodo aggiunge due nuovi campi, i cui nomi vengono derivati dal nome obiettivo.

Tabella 22. Campi di calcolo del punteggio dei modelli per raggruppamento cluster divisivo.

Nome del campo aggiunto	Significato
$DC-*nome_obiettivo*	Identificatore del cluster secondario a cui viene assegnato il record corrente.
$DCD-*nome_obiettivo*	Distanza dal centro di cluster secondari per il record corrente.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella al nugget del modello ed eseguire il nodo Tabella.

Nugget di raggruppamento cluster divisivo Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Misura della distanza. Metodo utilizzato per misurare la distanza tra i punti dati; una maggiore distanza indica una maggiore dissimilarità. Le opzioni disponibili sono:

- **Euclidea**. (default) La distanza fra due punti calcolata unendoli con una linea retta.
- **Manhattan**. La distanza fra due punti calcolata come somma delle differenze assolute fra le loro coordinate.
- **Canberra**. Simile alla distanza di Manhattan, ma più sensibile ai punti dati più vicini all'origine.
- **Massimo**. La distanza fra due punti calcolata come differenza massima sulla dimensione di qualsiasi coordinata.

Livello di gerarchia applicato. Livello della gerarchia da applicare ai dati.

Nugget del modello PCA Netezza

Il nugget del modello PCA consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello PCA, per default il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 23. Campo di calcolo del punteggio dei modelli per PCA.

Nome del campo aggiunto	Significato
$F-*nome_obiettivo*	Valore previsto per il record corrente.

Se si specifica un valore maggiore di 1 nel campo **Numero di componenti principali ...** nel nodo Modelli o nel nugget del modello e si esegue il flusso, il nodo aggiunge un nuovo campo per ciascuna componente. In questo caso i nomi dei campi hanno il suffisso *-n*, dove *n* è il numero della componente. Per esempio, se il modello è denominato *pca* e contiene tre componenti, i nuovi campi saranno denominati *$F-pca-1*, *$F-pca-2* e *$F-pca-3*.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella al nugget del modello ed eseguire il nodo Tabella.

Nugget PCA Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Numero di componenti principali da utilizzare nella proiezione. Numero delle componenti principali a cui si desidera ridurre l'insieme di dati. Questo valore non deve superare il numero di attributi (campi di input).

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Nugget del modello della struttura ad albero di regressione Netezza

Il nugget del modello della struttura ad albero di regressione consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello della struttura ad albero di regressione, per default il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 24. Campo di calcolo del punteggio dei modelli per la struttura ad albero di regressione.

Nome del campo aggiunto	Significato
$I-*nome_obiettivo*	Valore previsto per il record corrente.

Se si seleziona l'opzione **Calcola varianza stimata** nel nodo Modelli o nel nugget del modello e si esegue il flusso, viene aggiunto un ulteriore campo.

Tabella 25. Campo di calcolo del punteggio dei modelli per la struttura ad albero di regressione - aggiuntivo.

Nome del campo aggiunto	Significato
$IV-*nome_obiettivo*	Varianza stimata del valore previsto.

Per visualizzare i campi aggiuntivi, collegare un nodo Tabella al nugget del modello ed eseguire il nodo Tabella.

Nugget della struttura ad albero di regressione Netezza - Scheda Modello

La scheda **Modello** mostra l'importanza predittore del modello di struttura ad albero di regressione in formato grafico. La lunghezza della barra rappresenta l'importanza del predittore.

Nota: Quando si utilizza IBM Netezza Analytics Versione 2.x o precedente, il contenuto del modello di struttura ad albero di regressione viene mostrato solo in formato testo.

Per queste versioni, vengono visualizzate le seguenti informazioni:

- Ogni riga di testo corrisponde a un nodo o una foglia.
- Il rientro riflette il livello della struttura ad albero.
- Per un nodo, viene visualizzata la condizione di suddivisione.
- Per una foglia appare l'etichetta di classe assegnata.

Nugget della struttura ad albero di regressione Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Calcola varianza stimata. Indica se le varianze delle classi assegnate devono essere incluse nell'output.

Nugget della struttura ad albero di regressione di Netezza - Scheda Visualizzatore

La scheda **Visualizzatore** mostra una presentazione di struttura ad albero del modello sella struttura ad albero nello stesso modo in cui SPSS Modeler visualizza il proprio modello della struttura ad albero di regressione.

Nota: Se il modello è creato con IBM Netezza Analytics Versione 2.x o versione precedente, la scheda **Visualizzatore** è vuota.

Nugget del modello di regressione lineare Netezza

Il nugget del modello di regressione lineare consente di impostare le opzioni per il calcolo del punteggio del modello.

Quando si esegue un flusso che contiene un nugget del modello di regressione lineare, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 26. Campo di calcolo del punteggio dei modelli per la regressione lineare.

Nome del campo aggiunto	Significato
$LR-*nome_obiettivo*	Valore previsto per il record corrente.

Nugget di regressione lineare Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile impostare le opzioni per il calcolo del punteggio del modello.

Includi campi di input. Se selezionata, questa opzione passa a valle tutti i campi di input originali, accodando il campo o i campi di modellazione in più a ogni riga di dati. Se si deseleziona questa casella di controllo vengono passati solo il campo ID record e i campi di modellazione in più, consentendo una più rapida esecuzione del flusso.

Nugget del modello di serie temporali Netezza

Il nugget del modello consente di accedere all'output dell'operazione di creazione di modelli di serie temporali. L'output è composto dai seguenti campi:

Tabella 27. Campi di output del modello di serie temporali

Campo	Descrizione
TSID	L'identificatore della serie temporale; il contenuto del campo ID serie temporali nella scheda Campi del nodo Modelli. Consultare l'argomento "Opzioni dei campi della serie temporale Netezza" a pagina 101 per ulteriori informazioni.
TIME	Il periodo di tempo coperto dalla serie temporale corrente.
HISTORY	I valori dei dati cronologici (utilizzati per effettuare la previsione). Il campo viene incluso solo se l'opzione **Includi valori storici nel risultato** è selezionata nella scheda Impostazioni del nugget del modello.
$TS-INTERPOLATED	I valori interpolati, se presenti. Il campo viene incluso solo se l'opzione **Includi valori interpolati nel risultato** è selezionata nella scheda Impostazioni del nugget del modello. Interpolazione è un'opzione della scheda Opzioni di creazione del nodo Modelli.
$TS-FORECAST	I valori di previsione per la serie temporale.

Per visualizzare l'output del modello, allegare un nodo Tabella (dalla scheda Output della palette dei nodi) al nugget del modello ed eseguire il nodo Tabella.

Nugget serie temporali Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile specificare opzioni per personalizzare l'output del modello.

Nome modello. Il nome del modello specificato nella scheda Opzioni modello del nodo Modelli.

Le altre opzioni sono le stesse della scheda Opzioni di modellazione del nodo Modelli.

Nugget del modello lineari generalizzati Netezza

Il nugget del modello consente di accedere all'output dell'operazione di creazione di modelli.

Quando si esegue un flusso che contiene un nugget del modello lineare generalizzato, il nodo aggiunge un nuovo campo, il cui nome viene derivato dal nome obiettivo.

Tabella 28. Campo del calcolo del punteggio per lineare generalizzato.

Nome del campo aggiunto	Significato
$GLM-*nome_obiettivo*	Valore previsto per il record corrente.

La scheda Modello visualizza varie statistiche relative al modello.

L'output è composto dai seguenti campi:

Tabella 29. Campi di output dal modello lineare generalizzato.

Campo di output	Descrizione
Parametro	I parametri (cioè le variabili predittore) utilizzati dal modello. Questi sono colonne numeriche e nominali nonché l'intercettazione (il termine costante nel modello di regressione).
Beta	Il coefficiente di correlazione (cioè il componente lineare del modello).
Errore Std	La deviazione standard per beta.

Tabella 29. Campi di output dal modello lineare generalizzato (Continua).

Campo di output	Descrizione
Test	Le statistiche di test utilizzate per valutare la validità del parametro.
Valore P	La probabilità di un errore quando si presuppone che il parametro è significativo.
Riepilogo residui	
Tipo residuo	Il tipo di residui della previsione per cui sono mostrati i valori di riepilogo.
RSS	Il valore del residuo.
df	I gradi di libertà del residuo.
Valore P	La probabilità di un errore. Un valore elevato indica un modello poco adattabile; un valore basso indica un buon adattamento.

Nugget del modello lineare generalizzato Netezza - Scheda Impostazioni

Nella scheda Impostazioni è possibile personalizzare l'output del modello.

L'opzione corrisponde a quella descritta per Opzioni di calcolo del punteggio del nodo Modelli. Consultare l'argomento "Opzioni del modello lineare generalizzato Netezza - Opzioni di calcolo del punteggio" a pagina 91 per ulteriori informazioni.

Nugget del modello TwoStep di Netezza

Quando viene eseguito un flusso che contiene un nugget del modello TwoStep, il nodo aggiunge due nuovi campi che contengono l'appartenenza al cluster e la distanza dal centro del cluster assegnato per tale record. Il nuovo campo con il nome $TS-Twostep è relativo all'appartenenza al cluster ed il nuovo campo con il nome $TSP-Twostep è relativo alla distanza dal centro del cluster.

Nugget TwoStep di Netezza - Scheda Modello

La scheda **Modello** contiene diverse viste grafiche che mostrano distribuzioni e statistiche di riepilogo per i campi dei cluster. È possibile esportare i dati dal modello o le viste come grafici.

Note

Queste informazioni sono state preparate per prodotti e servizi offerti in tutto il mondo.

IBM può non offrire i prodotti, i servizi o le funzioni presentati in questo documento in altri paesi. Consultare il rappresentante locale IBM per le informazioni sui prodotti e servizi attualmente disponibili nella propria zona. Qualsiasi riferimento ad un prodotto, programma o servizio IBM non implica o intende dichiarare che solo quel prodotto, programma o servizio IBM può essere utilizzato. In sostituzione a quelli forniti da IBM, è possibile utilizzare prodotti, programmi o servizi funzionalmente equivalenti che non comportino violazione dei diritti di proprietà intellettuale o di altri diritti IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può avere applicazioni di brevetti o brevetti in corso relativi all'argomento descritto in questo documento. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Chi desiderasse ricevere informazioni relative a licenze può rivolgersi per iscritto a:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Per richieste di licenze relative ad informazioni double-byte (DBCS) contattare il Dipartimento di Proprietà Intellettuale IBM nel proprio paese o inviare richieste per iscritto a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Il seguente paragrafo non è valido per il Regno Unito o per tutte le nazioni le cui leggi nazionali siano in contrasto con le disposizioni in esso contenute: L'INTERNATIONAL BUSINESS MACHINES CORPORATION FORNISCE QUESTA PUBBLICAZIONE "NELLO STATO IN CUI SI TROVA", SENZA ALCUNA GARANZIA, ESPLICITA O IMPLICITA, IVI INCLUSE EVENTUALI GARANZIE DI COMMERCIABILITÀ ED IDONEITÀ AD UNO SCOPO PARTICOLARE. Alcuni stati non consentono la rinuncia ad alcune garanzie espresse o implicite in determinate transazioni; pertanto, la presente dichiarazione potrebbe non essere sempre applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM si riserva il diritto di apportare miglioramenti e/o modifiche al prodotto o al programma descritto nel manuale in qualsiasi momento e senza preavviso.

Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. I materiali disponibili sui siti Web non fanno parte di questo prodotto IBM e l'utilizzo di questi è a discrezione dell'utente.

IBM può utilizzare o distribuire le informazioni fornite in qualsiasi modo ritenga appropriato senza incorrere in alcun obbligo verso l'utente.

Coloro che detengono la licenza su questo programma e desiderano avere informazioni su di esso allo scopo di consentire (i) uno scambio di informazioni tra programmi indipendenti ed altri (compreso questo) e (ii) l'uso reciproco di tali informazioni, dovrebbero rivolgersi a:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Queste informazioni possono essere rese disponibili secondo condizioni contrattuali appropriate, compreso, in alcuni casi, l'addebito di un canone.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale concesso in licenza disponibile sono forniti da IBM in base ai termini dell'IBM Customer Agreement, dell'IBM International Program License Agreement o di qualsiasi altro accordo equivalente tra le parti.

Tutti i dati relativi alle prestazioni contenuti in questa pubblicazione sono stati determinati in un ambiente controllato. Di conseguenza, i risultati ottenuti con sistemi operativi diversi possono variare in modo significativo. Alcune misurazioni potrebbero essere state effettuate su sistemi in corso di sviluppo e non c'è garanzia che tali misurazioni coincidano con quelle effettuate sui sistemi comunemente disponibili. Inoltre, alcune misurazioni potrebbero essere stime elaborate tramite l'estrapolazione. I risultati effettivi potrebbero variare. Gli utenti di questo documento devono verificare i dati applicabili al proprio ambiente specifico.

Le informazioni relative a prodotti non IBM sono ottenute dai fornitori di quei prodotti, dagli annunci pubblicati e da altre fonti disponibili al pubblico. IBM non ha testato tali prodotti e non può confermarne l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Tutte le dichiarazioni relative all'orientamento o alle intenzioni future di IBM sono soggette a modifica o a ritiro senza preavviso e rappresentano unicamente mete ed obiettivi.

Questa pubblicazione contiene esempi di dati e prospetti utilizzati quotidianamente nelle operazioni aziendali. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e qualsiasi somiglianza con nomi ed indirizzi utilizzati da gruppi aziendali realmente esistenti è puramente casuale.

Se si visualizza una copia elettronica di queste informazioni, è possibile che le illustrazioni a colori e le fotografie non vengano visualizzate.

Marchi

IBM, il logo IBM e ibm.com sono marchi o marchi registrati di International Business Machines Corp., registrati in numerose giurisdizioni del mondo. I nomi di altri prodotti e servizi potrebbero essere marchi di IBM o di altre società. Un elenco aggiornato di marchi IBM è disponibile sul Web nella sezione "Copyright and trademark information" all'indirizzo www.ibm.com/legal/copytrade.shtml.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o relative controllate negli Stati Uniti e altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o in altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o in altri paesi.

UNIX è un marchio registrato di Open Group negli Stati Uniti e/o in altri paesi.

Java e tutti i marchi e i logo relativi a Java sono marchi commerciali o marchi registrati di Oracle e/o delle sue affiliate.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Indice analitico

A

analisi spettrale, IBM Netezza Analytics 99
Analysis Services
  esempi 25
  gestione di modelli 15
  Strutture ad albero delle decisioni 25
Apriori
  Microsoft 18
  Oracle Data Mining 43, 44, 45

C

calcolo del punteggio 8, 107
calcolo del punteggio del modello
  InfoSphere Warehouse Data Mining 57
campi partizione
  selezione 44
campo unico
  Apriori Oracle 40, 45
  MDL Oracle 46
  Medie K Oracle 42
  Naive Bayes Oracle 33
  NMF Oracle 43
  O-Cluster Oracle 41
  Oracle Data Mining 31
  Rete di Bayes adattivi Oracle 34
  SVM Oracle 36
chiave
  chiavi dei modelli 9
cluster di sequenze
  opzioni modello 17
cluster di sequenze (Microsoft) 20
  Opzioni avanzate 21
  opzioni dei campi 21
convalida incrociata
  Naive Bayes Oracle 33
costi
  Oracle 32
costi di errata classificazione
  Oracle 32
creazione di modelli di associazione
  InfoSphere Warehouse Data Mining 62
criterio di suddivisione
  Medie K Oracle 42

D

database
  modellazione nel database 8, 11, 13, 15, 21
  modellazione nel database per ISW 53
dati in formato tabellare
  Nodo Associazione ISW 63
dati transazionali
  Nodo Associazione ISW 63
DB2
  gestione di modelli 58
decomposizione tendenza stagionale, IBM Netezza Analytics 99
deviazione standard
  SVM Oracle 36
discretizzazione dei dati
  modelli Oracle 49
distribuzione 26, 51, 78
documentazione 3
DSN
  configurazione 13

E

editor di categorie
  Nodo Associazione ISW 66
epsilon
  SVM Oracle 36
esempi
  cenni generali 5
  Guida alle applicazioni 3
  mining del database 24, 25, 26, 51, 77, 78
esempi di applicazioni 3
esplorazione 25, 51, 77
esportazione
  modelli Analysis Services 24
  modelli DB2 59
etichetta classe, in modelli di strutture ad albero Netezza 91

F

fattore di complessità
  SVM Oracle 36
file tnsnames.ora 30
flussi
  Esempi di InfoSphere Warehouse Data Mining 77
foglia, in modelli di strutture ad albero Netezza 91
funzione distanza
  Medie K Oracle 42

G

generazione di nodi 24
Generazione SQL 8

I

IBM
  creazione di modelli di associazione 53
  creazione di modelli di raggruppamento tramite cluster demografici 53
IBM *(Continua)*
  creazione di modelli di raggruppamento tramite cluster Kohonen 53
  creazione di modelli di regressione 53
  creazione di modelli di regressione lineare 53
  creazione di modelli di regressione logistica 53
  creazione di modelli di regressione polinomiale 53
  creazione di modelli di sequenza 53
  Creazione di modelli di serie temporali 53
  creazione di modelli di struttura ad albero delle decisioni 53
  gestione di modelli 58, 84
  modelli Naive Bayes 53
IBM Netezza Analytics 79
  configurazione con IBM SPSS Modeler 79, 80, 82, 83
  Elementi adiacenti più vicini (KNN) 95
  gestione di modelli 107
  Lineare generalizzato 88
  Medie K 96
  Naive Bayes 98
  nugget del modello della struttura ad albero di regressione 113, 114
  nugget del modello di regressione lineare 114
  Nugget del modello di serie temporali 115
  Nugget del modello KNN 111
  Nugget del modello lineare generalizzato 115
  Nugget del modello Medie K 109
  nugget del modello Naive Bayes 110, 111
  Nugget del modello PCA 113
  Nugget del modello raggruppamento cluster divisivo 112
  nugget del modello rete di Bayes 109, 110
  nugget del modello struttura ad albero delle decisioni 108, 109, 114
  nugget del modello TwoStep 116
  Nugget modello lineare generalizzato 88, 116
  opzioni dei campi 83
  Opzioni dei campi della serie temporale 101
  Opzioni dei campi Medie K 97
  opzioni del campo TwoStep 105
  Opzioni della scheda Campi della struttura ad albero delle decisioni 92
  Opzioni di creazione della regressione lineare 94

IBM Netezza Analytics *(Continua)*
opzioni di creazione della struttura ad albero delle decisioni 93, 94
opzioni di creazione della struttura ad albero di regressione 85, 86
Opzioni di creazione delle serie temporali 102, 104
opzioni di creazione Medie K 97
opzioni di creazione TwoStep 105
opzioni modello 84
Opzioni modello di serie storiche 104
Opzioni modello KNN 95, 96
Opzioni modello lineare generalizzato 89, 90
PCA 106
PCA, opzioni campi 106
PCA, opzioni di creazione 106
Raggruppamento cluster divisivo 86
Raggruppamento cluster divisivo, opzioni campi 87
Raggruppamento cluster divisivo, opzioni creazione 87
Regressione lineare 94
Rete di Bayes 98
Rete di Bayes, opzioni campi 98
Rete di Bayes, opzioni creazione 99
Serie temporali 99
Struttura ad albero di regressione 85
Strutture ad albero delle decisioni 91
TwoStep 104
IBM SPSS Modeler 1
documentazione 3
mining del database 7
IBM SPSS Modeler Server 1
IBM SPSS Modeler Solution Publisher
modelli Oracle Data Mining 31
Importanza attributo (AI)
Oracle Data Mining 46, 47
InfoSphere Warehouse (IBM), vedere ISW 53
InfoSphere Warehouse Data Mining
creazione di modelli di associazione 62
flusso di esempio 77
nodo Regressione 67
nodo Sequenza 66
nugget del modello 75
strutture ad albero delle decisioni 61
tassonomia 65
interpolazione di valori, Serie temporali
IBM Netezza Analytics 100
ISW
Connessione ODBC 53
integrazione con IBM SPSS Modeler 53
scheda Server 60

K

kernel gaussiano
SVM Oracle 35
kernel lineare
SVM Oracle 35

L

livellamento esponenziale
IBM Netezza Analytics 99
Lunghezza di descrizione minima 34

M

MDL 34
MDL (Lunghezza descrizione minima)
Oracle Data Mining 45, 46
Medie K
IBM Netezza Analytics 96, 97, 109
Oracle Data Mining 41, 42
metodo di normalizzazione
Medie K Oracle 42
NMF Oracle 43
SVM Oracle 36
metrica di impurità
Apriori Oracle 40
Microsoft
Analysis Services 11, 13, 21
Cluster di sequenze 11
creazione di modelli di regressione lineare 13, 21
creazione di modelli di regressione logistica 13, 21
creazione di modelli di reti neurali 13, 21
creazione di modelli di struttura ad albero delle decisioni 11, 13, 21
gestione di modelli 15
modelli di cluster 11, 13, 21
modelli di regole di associazione 11, 13, 21
modelli Naive Bayes 11, 13, 21
Regressione lineare 11
Regressione logistica 11
Rete neurale 11
Microsoft Analysis Services 23, 24
min-max
normalizzazione dei dati 36, 49
mining del database
configurazione 13
creazione di modelli 8
Data Preparation 8
opzioni di ottimizzazione 8
utilizzo di IBM SPSS Modeler 7
Mining del database
esempio 24, 77
misura di impurità entropia 93
misura di impurità Gini 93
misure di impurità
Struttura ad albero delle decisioni di Netezza 93
modellazione di database
IBM Netezza Analytics 79, 80, 82, 83
Oracle 29, 30, 31, 32
modellazione nel database 22
modelli
creazione di modelli in-database 8
elenco DB2 59
elenco Netezza 84
esportazione 9
gestione DB2 58
gestione di Analysis Services 15
gestione di Netezza 84
modelli *(Continua)*
modelli di calcolo del punteggio in-database 8
problemi di uniformità 9
salvataggio 9
valutazione 26, 51, 78
visualizzazione di DB2 59
visualizzazione di Oracle 34
modelli a funzione singola
Rete di Bayes adattivi Oracle 34
modelli ARIMA
IBM Netezza Analytics 99, 103
modelli dell'elemento adiacente più vicino
IBM Netezza Analytics 95, 96, 111
modelli di regole di associazione
Microsoft 18
modelli di struttura ad albero delle decisioni
InfoSphere Warehouse Data Mining 61
modelli KNN
IBM Netezza Analytics 111
modelli lineari generalizzati
IBM Netezza Analytics 88, 89, 90, 91, 115, 116
Modelli lineari generalizzati (GLM)
Oracle Data Mining 37, 38, 39
modelli multifunzione
Rete di Bayes adattivi Oracle 34
modelli Naive Bayes
IBM Netezza Analytics 111
Rete di Bayes adattivi Oracle 34
modelli Naive Bayes tagliato
Rete di Bayes adattivi Oracle 34
Modelli PCA
IBM Netezza Analytics 106, 113

N

Naive Bayes
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
IBM Netezza Analytics 98, 110
InfoSphere Warehouse Data Mining 73
Opzioni avanzate 18
opzioni modello 17
opzioni server 17
Oracle Data Mining 33, 34
Netezza
gestione di modelli 84
NMF
Oracle Data Mining 42, 43
nodi
generazione 24
nodi Modelli
Microsoft Sequence Clustering 15
modellazione nel database 8, 11, 13, 15, 21
modellazione nel database per ISW 53
Naive Bayes Microsoft 15
raggruppamento tramite cluster Microsoft 15

nodi Modelli *(Continua)*
regole di associazione Microsoft 15
Regressione lineare Microsoft 15
Regressione logistica Microsoft 15
Rete neurale Microsoft 15
Serie temporali Microsoft 15
Strutture ad albero delle decisioni Microsoft 15
nodo Esplora 25, 51, 77
nodo Publisher
modelli Oracle Data Mining 31
Nodo Raggruppamento cluster
InfoSphere Warehouse Data Mining 70
nodo Regressione
InfoSphere Warehouse Data Mining 67
nodo Regressione logistica
InfoSphere Warehouse Data Mining 73
nodo Sequenza
InfoSphere Warehouse Data Mining 66
nome host
connessione Oracle 30
normalizzazione dei dati
modelli Oracle 49
nugget del modello
IBM Netezza Analytics 88, 108, 109, 110, 111, 112, 113, 114, 115, 116
InfoSphere Warehouse Data Mining 75
numero di cluster
Medie K Oracle 42
O-Cluster Oracle 41

O

O-Cluster
Oracle Data Mining 41
ODBC
configurazione 13
configurazione con Oracle 29, 30, 31, 32
configurazione di ISW 53
configurazione di SQL Server 13
configurazione per IBM Netezza Analytics 79, 80, 82, 83
ODM. Vedere Oracle Data Mining 29
opzioni avanzate
ISW Data Mining 60
opzioni dei campi
IBM Netezza Analytics 83, 87, 92, 97, 98, 101, 105, 106
nodi Modelli 63
opzioni di creazione
IBM Netezza Analytics 85, 86, 87, 93, 94, 97, 99, 102, 104, 105
opzioni modello
IBM Netezza Analytics 84, 89, 90, 95, 96, 104
Oracle Data Miner 48
Oracle Data Mining 29
Apriori 43, 44, 45
configurazione con IBM SPSS Modeler 29, 30, 31, 32
costi di errata classificazione 48
Oracle Data Mining *(Continua)*
esempi 50, 51
gestione di modelli 47, 48
Importanza attributo (AI) 46, 47
MDL (Lunghezza descrizione minima) 45, 46
Medie K 41, 42
Modelli lineari generalizzati (GLM) 37, 38, 39
Naive Bayes 33, 34
NMF 42, 43
O-Cluster 41
preparazione dei dati 49
Rete di Bayes adattivi 34, 35
Struttura ad albero delle decisioni 39, 40
Support Vector Machine 35, 36
verifica dell'uniformità 47

P

partizionamento dei dati 44
penalità complessità 18, 19, 20
peso classe, in modelli di strutture ad albero Netezza 91
peso istanza, in modelli di strutture ad albero Netezza 91
port
connessione Oracle 30
probabilità a priori
Oracle Data Mining 37
punteggi z
normalizzazione dei dati 36, 49

R

raggruppamento cluster divisivo
IBM Netezza Analytics 86, 87
Raggruppamento cluster divisivo
IBM Netezza Analytics 112
raggruppamento tramite cluster
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
IBM Netezza Analytics 112
InfoSphere Warehouse Data Mining 70
Opzioni avanzate 18
opzioni modello 17
opzioni server 17
regole di associazione
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
Opzioni avanzate 19
opzioni modello 17
opzioni server 17
regressione lineare
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
IBM Netezza Analytics 85, 94, 114
Opzioni avanzate 18
regressione lineare *(Continua)*
opzioni modello 17
opzioni server 17
regressione logistica
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
Opzioni avanzate 18
opzioni modello 17
opzioni server 17
Rete bayesiana, modelli
IBM Netezza Analytics 98, 99, 109, 110
Rete di Bayes adattivi
Oracle Data Mining 34, 35
rete neurale
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
Opzioni avanzate 18
opzioni modello 17
opzioni server 17

S

scheda Server
ISW 60
serie temporali
IBM Netezza Analytics 101, 102, 104
InfoSphere Warehouse Data Mining 73, 74, 75
Serie temporali
IBM Netezza Analytics 104
serie temporali (IBM Netezza Analytics) 115
Serie temporali (IBM Netezza Analytics) 99
serie temporali (Microsoft) 19
Opzioni avanzate 20
opzioni di impostazione 20
opzioni modello 20
server
esecuzione di Analysis Services 17, 22
SID
connessione Oracle 30
soglia pairwise
Naive Bayes Oracle 34
soglia Singleton
Naive Bayes Oracle 34
Solution Publisher
modelli Oracle Data Mining 31
SQL Server 17, 22
configurazione 13
Connessione ODBC 13
Struttura ad albero delle decisioni
IBM Netezza Analytics 91, 92, 93, 94, 108, 109, 114
Oracle Data Mining 39, 40
strutture ad albero delle decisioni
calcolo del punteggio - opzioni riepilogo 22
calcolo del punteggio - opzioni server 22
Microsoft Analysis Services 11, 13, 21

strutture ad albero delle decisioni
(Continua)
Opzioni avanzate 18
opzioni modello 17
opzioni server 17
strutture ad albero di regressione
IBM Netezza Analytics 85, 86, 113, 114
Support Vector Machine
Oracle Data Mining 35, 36
SVM. Consultare Support Vector Machine 35

T

tassonomia
InfoSphere Warehouse Data Mining 65
tolleranza convergenza
SVM Oracle 36
twostep
IBM Netezza Analytics 104
TwoStep
IBM Netezza Analytics 105, 116

V

valutazione 26, 51, 78

Stampato in Italia