Modulo IV.1 - Tecniche di osservazione e valutazione delle risorse intelligenti : introduzione al Data Mining

1. Data Mining
2. Tipi di apprendimento nel Data Mining
3. Algoritmi di classificazione
4. Algoritmi di clustering
5. Algoritmi di regressione
6. KNIME
7. Materiali supplementari: utilizzo di KNIME

Bibliografia

Co-funded by the European Union
e-EarlyCare-T
gestionet Play Think Learn
ROMA TRE UNIVERSITÀ DEGLI STUDI
Hveloce I+D+i
MEDRI
UNIVERSIDAD DE BURGOS

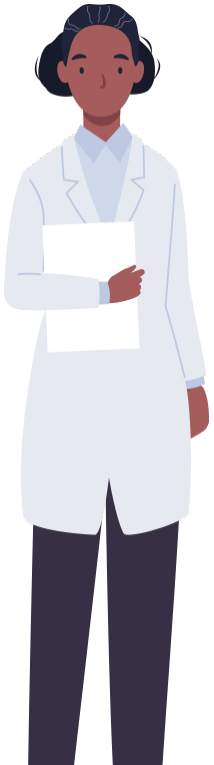
1. Data Mining

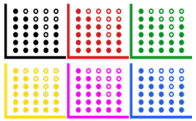
Il Data Mining è il processo di ricerca e analisi di grandi database per trovare informazioni utili al processo decisionale.

Esistono numerose tecniche di DM che impiegano l'analisi matematica per dedurre i modelli e le tendenze esistenti nei dati. In genere, questi schemi non possono essere rilevati con l'esplorazione tradizionale dei dati perché le relazioni sono troppo complesse o perché il volume di dati da analizzare è troppo grande.

Attualmente il Data Mining viene utilizzato continuamente per l'analisi di grandi quantità di dati in vari campi della conoscenza, come l'istruzione, l'economia, gli affari, l'ambiente, ...

1.1. Data Mining. Nozioni di base





1.2. Data Mining. Processo di applicazione di tecniche di data mining

Definizione del problema
Questa è la prima fase in cui un problema specifico viene tradotto in un problema di data mining, in cui vengono sollevati gli obiettivi dell'analisi e le domande di ricerca.

Preparazione e raccolta dei dati
È la fase più estesa del processo, poiché la qualità dei dati è una delle sfide più importanti del data mining. I dati grezzi devono essere identificati, puliti e archiviati in un formato predefinito.

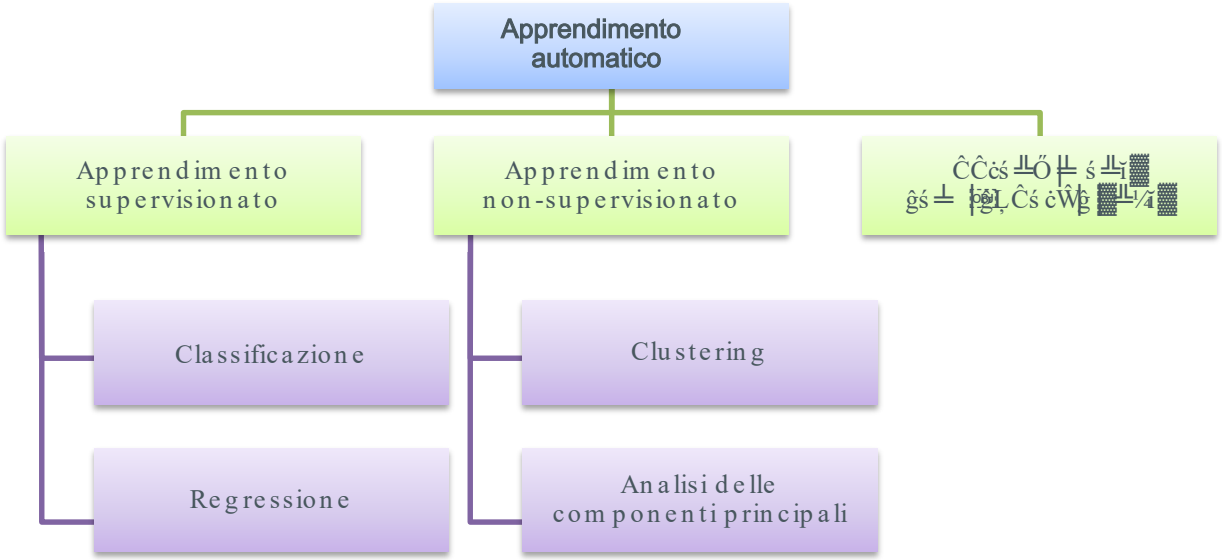
Modellazione e valutazione
In questa fase vengono selezionate e applicate diverse tecniche di modellazione dei dati (algoritmi), per poi stabilire i parametri e i valori ottimali di queste tecniche .

Distribuzione
È l'ultima fase in cui i risultati del data mining vengono organizzati e presentati tramite grafici e report.

1.2. Data Mining. Processo di applicazione di tecniche di data mining

È importante notare che ogni processo di data mining è un processo iterativo, il che significa che il processo non si ferma quando una particolare soluzione viene impiegata. Può essere solo una nuova voce per un altro processo di data mining (Rodríguez-Arribas, 2021). In altre parole, in molte occasioni l'applicazione delle tecniche di DM richiede diverse iterazioni e l'uso di diversi algoritmi per poter estrarre i risultati finali della ricerca che stiamo conducendo.

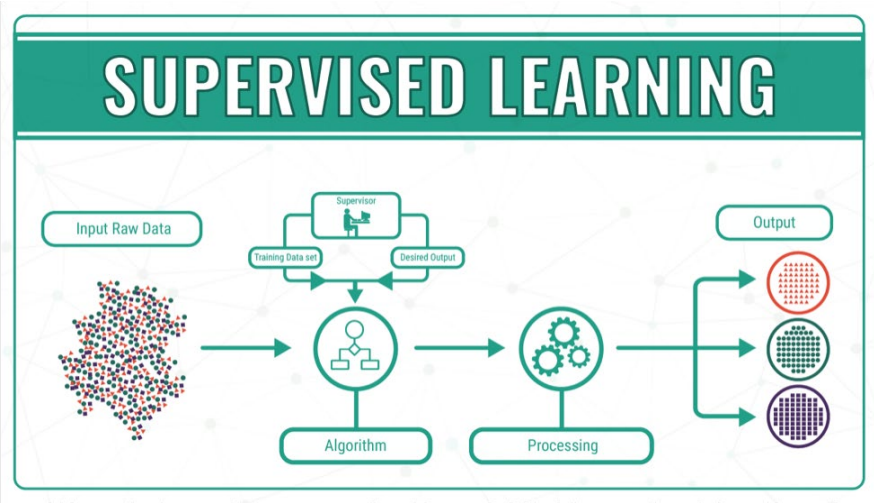
2. Tipi di apprendimento nel Data Mining



2.1. Tipi di apprendimento nel Data Mining

APPRENDIMENTO SUPERVISIONATO

L'obiettivo fondamentale dell'apprendimento supervisionato è la creazione di un modello in grado di prevedere i valori corrispondenti agli oggetti in ingresso dopo aver familiarizzato con una serie di esempi, i dati di addestramento .



Questa tecnica si compone di due fasi fondamentali :

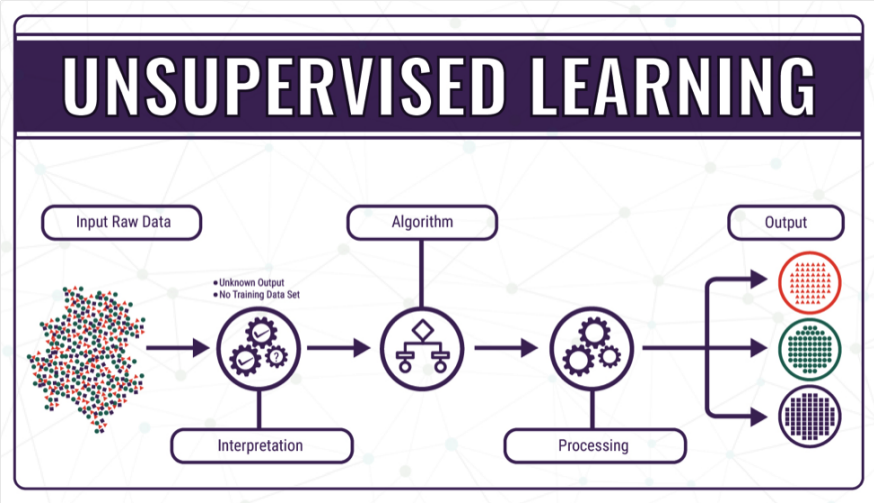
1. una fase di addestramento, in cui si utilizza un insieme di dati etichettati, che contengono i dati di input e i risultati desiderati per quei dati di addestramento con un algoritmo che permette di dedurre una funzione dai dati grezzi (raw data) che stiamo fornendo all'algoritmo ;
2. la fase di test, in cui la funzione ottenuta nella fase precedente viene utilizzata per generare nuove previsioni con nuovi set di dati .

Il processo è noto come apprendimento supervisionato, poiché conoscendo le risposte di ciascun esempio dell'insieme di formazione, è possibile correggere la funzione generata dall'algoritmo . La preparazione dell'algoritmo viene supervisionata correggendo i suoi parametri, in base ai risultati ottenuti iterativamente .

APPRENDIMENTO NON-SUPERVISIONATO

Questo tipo di apprendimento è l'altro approccio di base al Machine Learning (ML). L'apprendimento non supervisionato prevede dati non etichettati che l'algoritmo deve cercare di comprendere da solo.

L'obiettivo di questo tipo di apprendimento è far sì che la macchina impari senza l'aiuto o le indicazioni dei data scientist, cioè senza supervisione e senza un set di dati di addestramento. Inoltre, la macchina stessa aggiusterà i risultati e i raggruppamenti quando ci saranno risultati più adatti, permettendo alla macchina di comprendere i dati e di elaborarli nel modo migliore.



L'apprendimento non-supervisionato viene utilizzato per esplorare dati sconosciuti e non etichettati. Può rivelare modelli che potrebbero essere stati trascurati o esaminare grandi insiemi di dati che sarebbero troppo impegnativi per una singola persona.

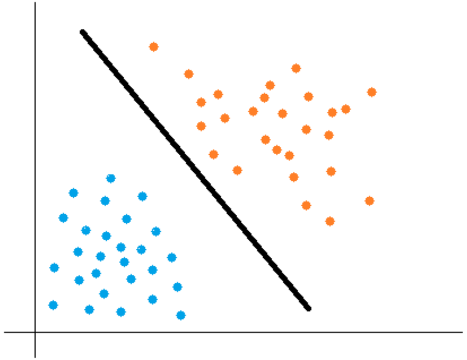
APPRENDIMENTO SEMI-SUPERVISIONATO

Attualmente sono in corso numerose ricerche sui metodi di apprendimento semi-supervisionato. Queste tecniche di apprendimento automatico utilizzano dati di addestramento sia etichettati che non etichettati: in genere, una piccola quantità di dati etichettati insieme a una grande quantità di dati non etichettati (Zhu e Goldberg, 2009).

In altre parole, cercano di migliorare i modelli di previsione ottenuti utilizzando esclusivamente dati etichettati, esplorando le informazioni strutturali contenute nei dati non etichettati. Possiamo dire che l'apprendimento semi-supervisionato cerca di combinare i due approcci tradizionali del data mining (apprendimento supervisionato e apprendimento non supervisionato) per mantenere il meglio di ciascuno di essi.

3. Algoritmi di classificazione

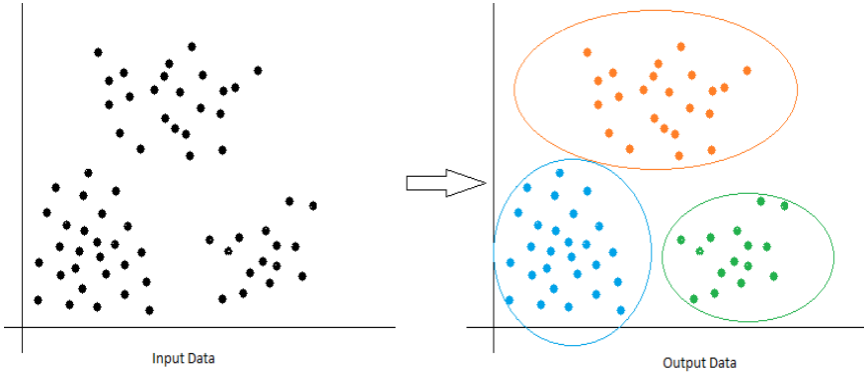
[illegible]



Questi algoritmi lavorano generalmente sulle informazioni fornite da un insieme di campioni, modelli, esempi o prototipi di addestramento che vengono presi come rappresentanti delle classi e conservano un'etichetta di classe corretta. Questo insieme di prototipi correttamente etichettati è chiamato insieme di addestramento e costituisce la conoscenza disponibile per la classificazione di nuovi campioni.
L'obiettivo della classificazione supervisionata è quello di determinare, in base a ciò che è noto, a quale classe dovrebbe appartenere un nuovo campione, considerando le informazioni che possono essere estratte.

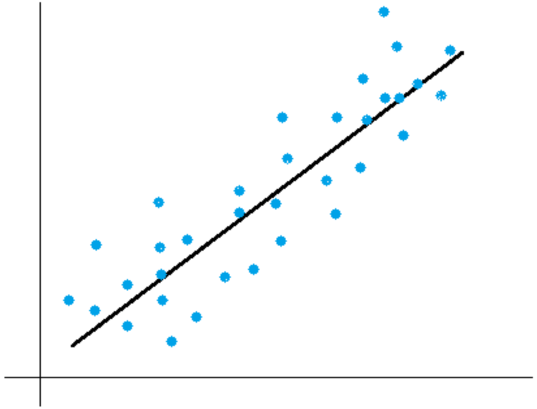
4 . Algoritmi di clustering

Gli algoritmi di clustering sono responsabili del raggruppamento degli oggetti in un set di dati in base alle loro somiglianze. In questo modo, gli oggetti che si trovano all'interno di un cluster o di un gruppo hanno più somiglianze tra loro che differenze. Questi algoritmi lavorano con dati non etichettati, quindi è l'algoritmo stesso che analizza i dati per trovare il numero ottimale di raggruppamenti per l'insieme di dati in ingresso, poiché non abbiamo una conoscenza preliminare delle caratteristiche dei dati e delle loro classi.

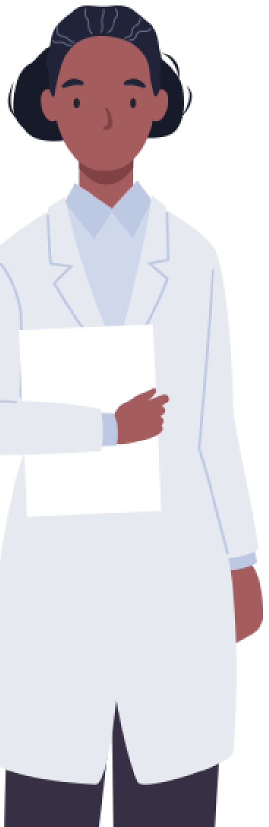


I raggruppamenti effettuati dagli algoritmi possono essere di due tipi:
hard cluster : ogni dato appartiene esclusivamente a un gruppo;
soft (diffuse) cluster : i dati possono appartenere a più gruppi in misura diversa, cioè gli stessi dati possono avere un grado di appartenenza del 60% al gruppo 1e del 40% al gruppo 2.

Gli algoritmi di regressione sono un sottocampo dell'apprendimento supervisionato il cui obiettivo è stabilire un metodo per la relazione tra un certo numero di caratteristiche e una variabile obiettivo continua.
Si tratta di algoritmi che stabiliscono una traiettoria per fornire la tendenza di un insieme di dati, cioè: lo scopo di questi algoritmi è di mettere in relazione un certo numero di caratteristiche e una variabile-obiettivo continua.



Questa tecnica è utile per prevedere esiti che sono valori continui, il che significa che la risposta alla domanda di ricerca è presentata da una quantità che può essere determinata in modo flessibile in base agli input del modello, piuttosto che essere limitata a un insieme finito di etichette come nel caso della classificazione.



6. KNIME

KNIME è un'applicazione open-source che ci permette di applicare ai nostri set di dati o a set di dati campione:

1. metodi statistici,
2. algoritmi di data mining o Machine Learning (apprendimento automatico),
3. tecniche di visualizzazione.

È costruito sulla piattaforma Eclipse ed è programmato in Java. Essere un software open-source ha molti vantaggi: il suo codice appartiene alla comunità di utenti e sviluppatori, il che garantisce che sarà sempre uno strumento libero che può essere scaricato e utilizzato gratuitamente secondo i termini della licenza GPLv3. Permette inoltre di incorporare codice sviluppato in R o Python.

È uno strumento di "programmazione visiva". L'analisi dei dati può essere effettuata in modo intuitivo, impostando il processo con un semplice clic del mouse. I "nodi" di cui abbiamo bisogno vengono posizionati, senza aver bisogno di conoscerne il nome o la configurazione, poiché è sempre possibile ricevere aiuto.

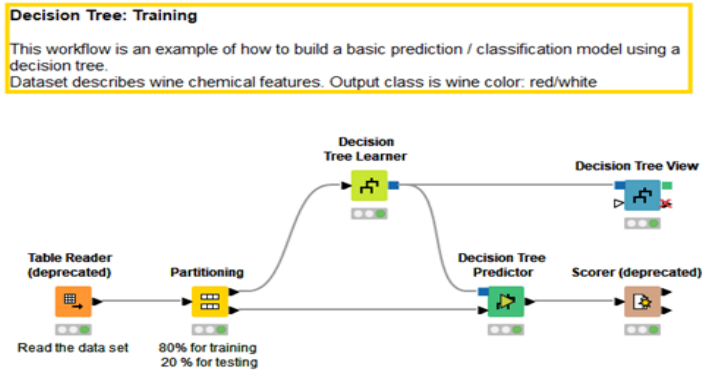
È uno strumento progettato per essere semplice da usare. Il concetto più importante nell'uso dello strumento è quello di flusso di lavoro (workflow). Un flusso di lavoro è una sequenza di passi configurati dall'utente. Formalmente è un insieme di nodi uniti da frecce che rappresentano il flusso di dati da un nodo all'altro. Un nodo racchiude diversi lavori che possono essere eseguiti con i dati; ci sono nodi per molti compiti.

Sono presenti nodi per:

- caricare i dati da file o database.
- Creare, modificare o eliminare righe o colonne dal set di dati con cui stiamo lavorando.
- Calcolare statistiche medie, percentili, correlazioni, ecc.
- Combinare dati provenienti da fonti diverse.
- Costruire e valutare modelli di Machine Learning come: classificazione, regressione o clustering.
- Visualizzare i dati utilizzando grafici a barre, a torta, a dispersione e altri tipi di grafici più avanzati.
- Generare report.

Un flusso di lavoro potrebbe avere un nodo per caricare un set di dati da un file Excel, poi un nodo per selezionare gli attributi (colonne) da quel set di dati e infine un altro nodo per visualizzare le statistiche degli attributi selezionati.

Qui di seguito, un esempio di «albero decisionale»



Bibliografia

d ½ċl’ ┤⁄₄ Œ‴ à Ļ ś ╨↑ ▓ É‴ Æl ś ċċś ċ⁄₄ c ‴ □⁄ ∧ ⁹ □‴ B ½ī ¼ ¬ ċś Ĉċ▓’ś ĝĝ ╠↑ ╠ B ½ī ¼ í ╠ ╠↑ □Ľ ÆŒ ₄ ╪Ŵ⁄₄Ő▓ċ d ½ċl’ ┤⁄₄ ÉĻ, ╠⁄╨ à Ļ ś ╨↑ ▓ c ċ⁄₄ľ’ ĝl’▓ l ś ċċś ċ⁄₄ ŒĈċ ╠↑ ś ċ Œ₄ ╪ẅœ ½ ╨ẅ⁄₄╨½ćś ĝ‴ í ‴; ‴ í ½ćĭ ┫’▓ćś ╨¼ ≈‴ Æ ċ╨¼ ╪ẅœ ▓╨ẅ⁄₄ ŵ‴ ‴ □⁄ ∧ ⅞ ⅞ □‴ ı╨ Ĉċ▓Ŵ╨ ś ╨ĭĝ ←ċ ĭ≡ś ċ⁄₄Ĉś Ļ ĭ ┫’ ╠ĭś ċŴ╨ĭ ▓╨ ←ė▓╨ ĭ≡ś Ļĝś ▓←ŵś Ľ ½ĈĈ ╪’ ½ī ▓╨ĝ ½╪Ő ╨ ¼’≡ ╠ś ╪½ċ▓╨↑ ĭś ľ≡╨Ċ Ļ ś ĝ ╠ Ő ┣←ś-ċś ╨ĭ ¼←ś-ľ ĭ ½ī ▓╨ĝ ╠ ľ≡ ┣Ďċś ╨ ¼┤ś Ő ∧ Áś ½ċĝ‴ ı╨ĭ‴ É‴ J ╨Ŵ┣▓╨‴ ≈ś ĝ‴ ¬ Ļ Ľ ╪’ l ś ¼╪≡‴ ⁹⁴‴ □□□□‴ ≡ĭĭ Ĉĝ‴ □Ő▓ ▓ċ↑ □⁹∧‴, , ⁴ ∧ □╪ċĈ≡ ⁹⁴ ⁹⁹ □□□□

Œ₄ ╪ẅœ ½ ╨ẅ⁄₄╨½ćś ĝ‴ í ‴; ‴ í ½ćĭ ┫’▓ćś ╨¼ ≈‴ ▪ ċ╨¼ ╪ẅ‴ ‴ □⁄ ∧ ⅞ ∧ □ J Ŵ₄ ╪ ½ī ▓╨ ▓←e Ļ ╨ľ’ĭ ▓╨¼ ┣ Ľ ┫╪ś ĝ ╠ ∧ □ ▓ś ½ċœ Ďĝ‴ ╨ ╨¼ ╪ĝ ĝ ŵ ┣≡ ĭ≡ś ś J ½ċ ┣É ½ćś ; ▓╨ ĈĻĭś ċ ĈĈ ┫’ ½ī ▓╨‴ □⁄ ∧ ⅞ ∧ □ ı╨ĭ‴ É‴ J ╨Ŵ┣▓╨‴ ≈ś ĝ‴ ¬ Ļ Ľ ┫’ l ś ¼╪≡‴ ⁹⁶ ⁴ □, , ⁹□ ▓ ≡ĭĭ Ĉĝ‴ □Ő▓ ▓ċ↑ □⁹∧‴, , ⁴ ∧ □╪ċĈ≡ ⁹⁶ ∧ ⁴, , ⁹□

Œ₄ ╪ẅœ ½ ╨ẅ⁄₄╨½ćś ĝ‴ í ‴; ‴ í ½ćĭ ┫’▓ćś ╨¼ ≈‴ ċ╨¼ ╪ẅœ ▓╨ẅ⁄₄ ŵ‴ ‴ B ┤ ŵ▓½ĝi▓ċ‴ É‴ c ‴ ▪ ≈▓Óċ↑ Ļ ś ŵ▓ċċ ┣Ľ½ĝ‴ Œ‴ □⁄ ∧ ⁹⁴‴ í ½ćl’≡□; ▓╨ ĈĻĭś ċ ½ĈĈ ┫’ ½ī ▓╨ ←ċ ĭ≡ś ċś↑ ĝĭ ċ⁄₄ ▓╨ ½╪Ő ¼Ļ ĭ ▓╨ ½ī ▓╨ ▓←ĭ≡ś l’▓ċċś l’ĭ ▓╨ ▓←¼←Ļ╨ľ’ĭ ▓╨¼ ┣ĝ ▓ Őś ĭś l’ĭ ▓╨ ĝl’¼ ┣ ╠ J ½ċ ┣É; ½ćś ‴ ⁹, ĭ≡ ı╨ĭś ċ╨¼ ▓╨¼ ┣ś ľ≡╨▓┫↑ ÆJ ŐĻ ľ ½ī ▓╨ ½╪Ő B ś Ŵ ▓Ĉ╨ ś╨ĭ ; ▓╨←ś-ċś ╨ľ’ś ¬ċ▓’ś ś Ő ╠↑ ĝ ▓←ŕî ‹ J B ⅞ ∧ ⁹⁴ ; ▓╨←ś-ċś ╨ľ’ś ⁹⁹ĭ≡▓ ĭ≡ □▓ ⅞ ⅞▓ ⅞ □□‴ ı ‹ J B ‴ ¸ ¼┣╨ľ’ ⅛ Ő▓┤ ⁹∧‴⅞ ⁹⁹⅞ □□┫╨ĭś Ő‴⅞ ∧ ⁹⁴‴⁹, ⅞ ∧

Modulo IV.1. Materiali supplementari : utilizzo di KNIME

1. KNIME, installazione

2. KNIME, flussi di lavoro

2. 1. Nodi

2.2. L'area di lavoro

3. Esempio generico: classificazione delle specie floreali

4. Esempio con i dati di un intervento terapeutico intelligente (EarlyCare)

Riferimenti dal web

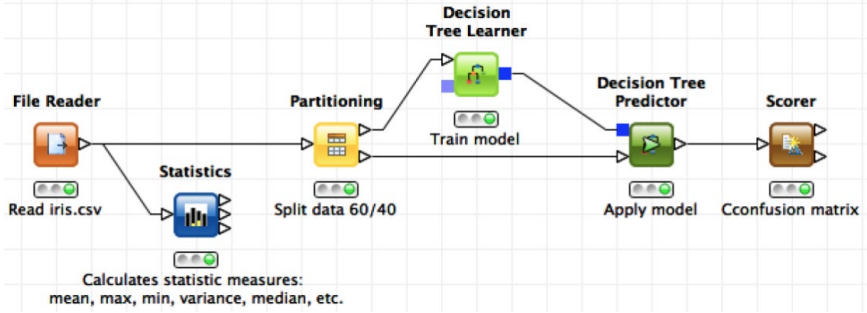
Co-funded by the European Union

1. KNIME, installazione

Öî rí J Φ Ļ ╩ ▒ĈĈ ╠ ¼ẅ▓╩ś É¼Ŵ¼ ╠═ ≡ś ĝ ╞ ╩ ╞ ¼ ≡ś Φ ╩ś ś ĝĝ½ć ▓ ¼Ŵ ćś ╠ĝĩ¼╠═▓ É¼Ŵ¼ Ĉć ╠ ¼Ő ▏ Ĉ▓ś ć ╠ĝĩ¼╠═ćś ś Ő ś ĝś ↑ Ļ ╞ś ╠Ĉć▓ ć¼╩ ═ ¼

¬ś ć ╠ĝĩ¼╠═ćś ╠ĝ▓ŧẇ½ćś ˝ Ő▓Ľ ▏¼╩ ▓ ¼╩Ő½ćś ĝĻ ╞ĝ ▏▓ ≡ĭĭĈĝ˝ □ẅ ŵ ŵ ˝ ╙╙ ╠ ś ˝ ľ ▓ ╩ □Ő▓ ŵ ╩ ▓ ⁄Őĝ˝ Ļ ╩¼ Ŵ▓ ╞¼ ╠ ĝľ ½ć ▏ ≡ś ćś ═ ▓ □Öî rí J ╩¼ ╞ĭ ▏ ĝ ¬ ╞ŧ ◄▓ć ═ □ĝľ ś ↑ ╠ ╩Ő▓ ╞₄ Ŵ ćĝ ▓╩ś ľ ▓ćc ╞ĝĈ▓╩Ős ╩ĭś ¼╞ ╩▓ẹĭ ć▓ Ĉś ćĝ▓╩¼ ╠ ▓╩ ĈĻĭś ć˝ í ¼ľ ˝ ″ ╠Ő▓ŵ ĝ , ⁒ Ľ ▏ □Ŵ ľ ľ ≡ ▏ľ ▓╩ ĈĻĭś ć□ ″ ╠Ő▓ŵ ĝ □* □ľ ▓╩ ĈĻĭś ć ═ ▓Ős ć╩ ▏▓ à ╠Ļ å˝

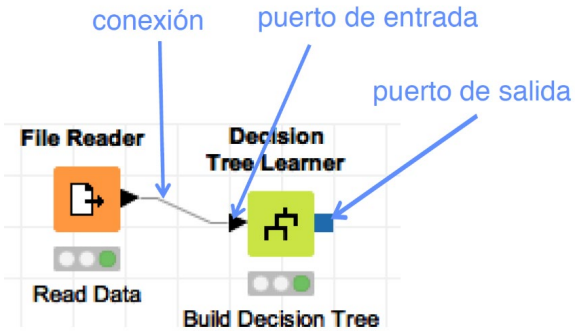
2. KNIME, flussi di lavoro



2.1. Nodi

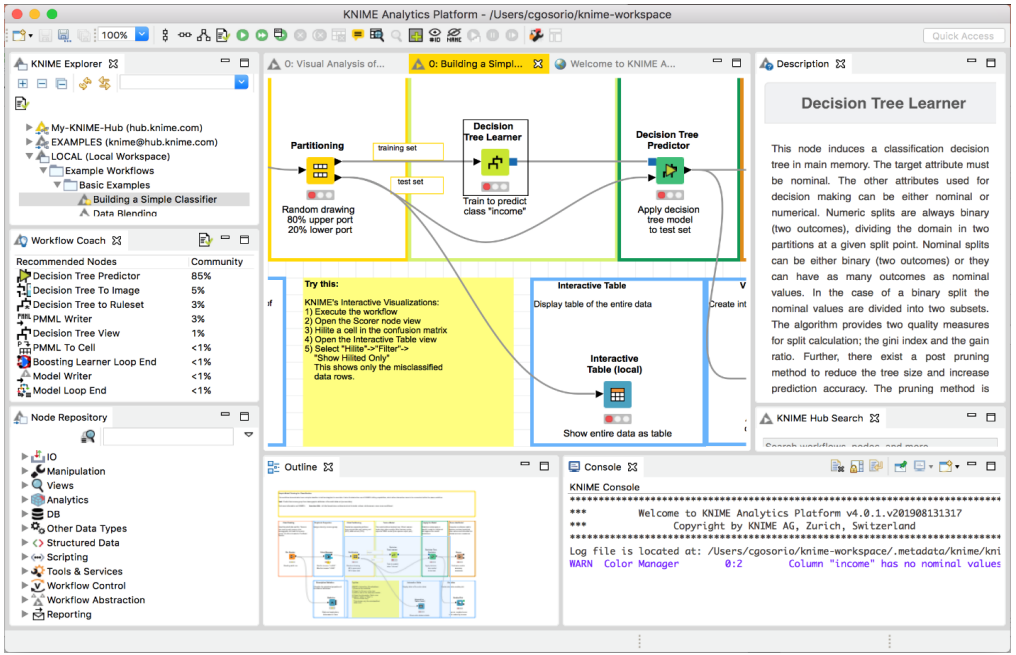
I nodi (nell'immagine accanto, i nodi di entrata e di uscita, oltre alla connessione) incapsulano gli algoritmi che implementano le azioni che possono essere eseguite sui dati:

- manipolazione di righe, colonne, ecc.
- Creazione di modelli di data mining.
- Valutazione dei modelli.
- Applicazione dei modelli a nuovi dati.
- Processi ETL (Extract, Load, Transform).
- Creazione di report personalizzati.



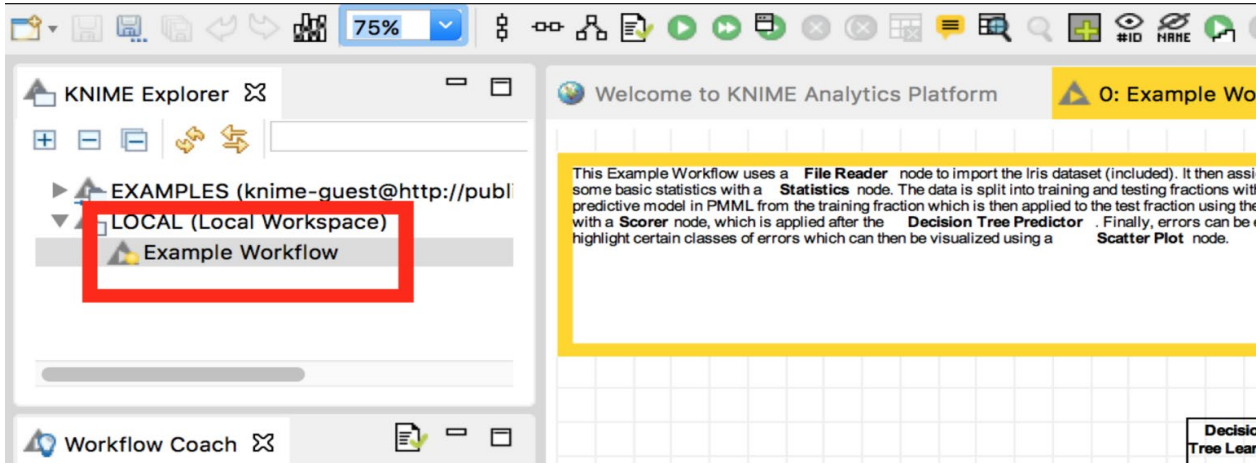
2.2. L’area di lavoro

L'area di lavoro è la cartella o directory del nostro computer in cui sono memorizzati tutti i progetti realizzati con KNIME. Sarà necessario scegliere un'area di lavoro prima di avviare il programma (si può anche lasciare la cartella che appare di default al momento dell'installazione).



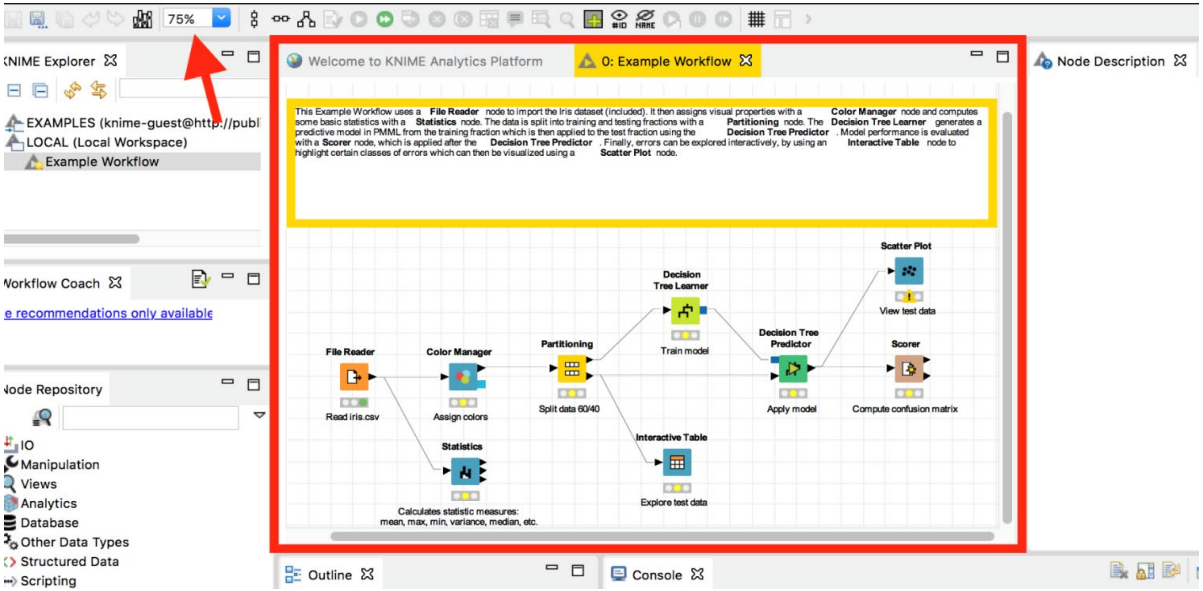
2.2.1. L’area di lavoro: KNIME Explorer

Y ├½ćś ¼ ╠ ľ Ļ, │Ŵ ╩↑ ▓╠▓↑ ś ğĩ ŀ │ │Ĉć▓↑ ś ĩĩ │ś │¼Ļ, ĝĝ │Ő│ ├₄Ŵ▓ć▓ ĝ¼ ├Ŵ₫ ▛ Ő▓Ŵ ĝ │╠ Ĉ▓ćĩ ¼╠▓ ▓ ĝ │ś ĝĈ▓ćĩ ¼╠▓ │¼Ļ, ĝĝ │Ő│ ├₄Ŵ▓ć▓



2.2.2. L'area di lavoro: flusso di lavoro, editor

Questa è l'area di lavoro principale, dove i nodi vengono trascinati, collegati e il flusso di lavoro viene configurato.

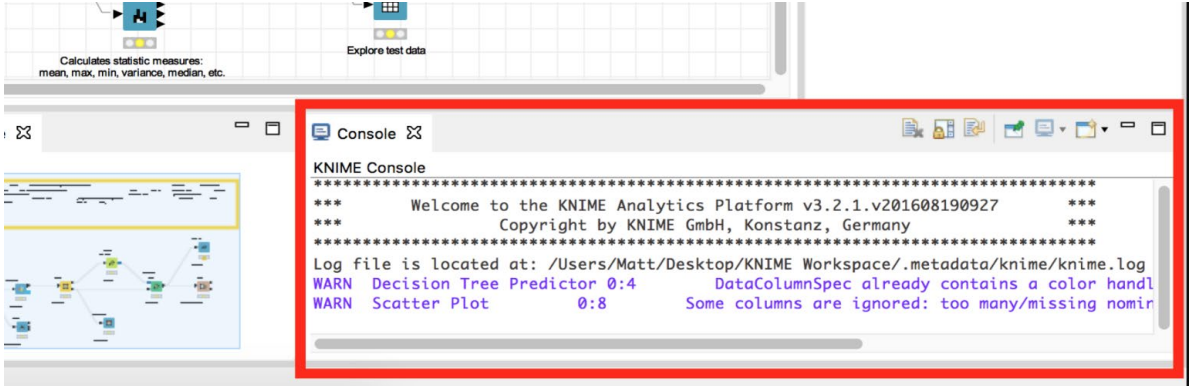


2.2.3. L’area di lavoro: schema

í ▒ğĩ ċ¼ Ļ╩¼ Ĉ¼╩▒ċ¼╩ ╞’¼ Őś ╞⅛Ļ ĝĝ▒ Ő│ ╞Ẃ▒▒ ↩┘’ ╟╞¼╩Ő▒ ╟╞Ĉ½ĝĝ¼↑ ↑ ▒ Ő¼ Ļ╩¼ Ĉ¼ċĩś ¼╟╞¼ċ¼ ĊĻ¼╩Ő▒ ╟╞⅛Ļ ĝĝ▒ Ő│ ╞Ẃ▒▒ Φ ╩ ▒╞▒ ↑ ċ¼╩Őś ♯

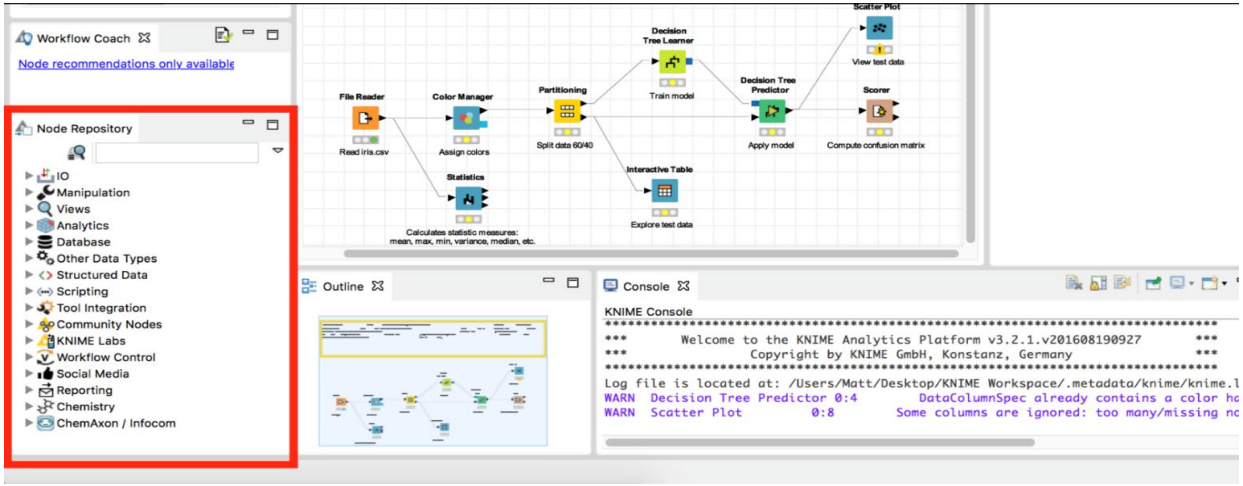
2.2.4. L'area di lavoro: la console

È un campo di output di testo che visualizza gli avvisi e gli errori che si verificano durante l'esecuzione del flusso di lavoro. Visualizza anche informazioni rilevanti sul processo di esecuzione.



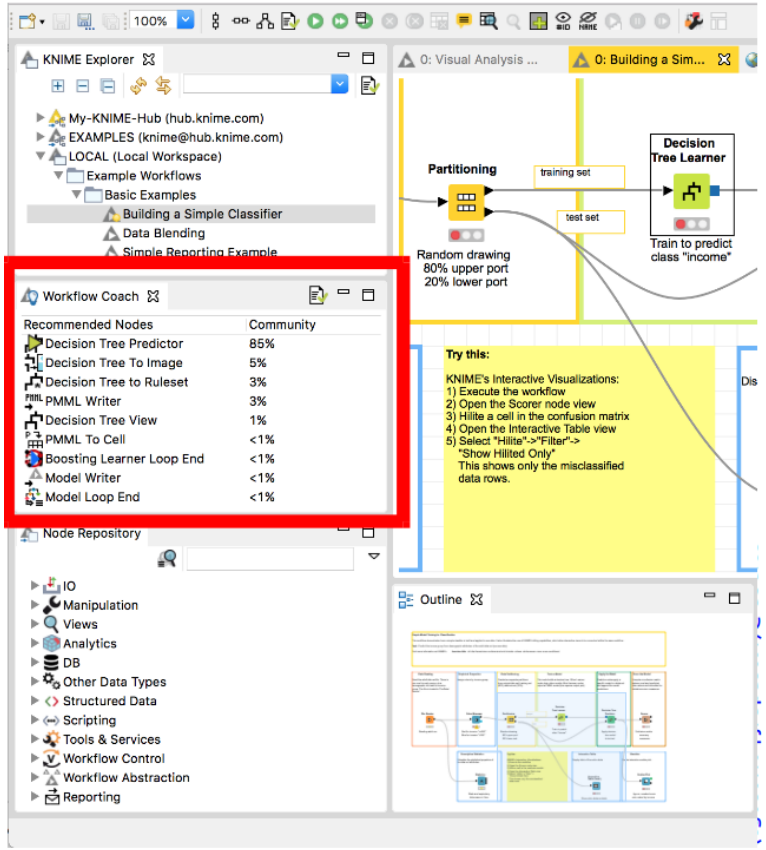
2.2.5. L’area di lavoro: archivio dei nodi

ƒ Ļśĝĩ¼Φ ╞ǽśź¼ ╠ľĻ ||╩▒Ő|ĝ▒╩▒▒ç↑ ¼╩ ╞ẅẅ¼ĩ |Ĉśċľ¼śź↑▒ç|¼ Y ¼╩ľ≡ś Ĉ▒ĝĝ|Ļ ||╞ľśċľ¼ćś |╩▒Ő|Ĉśċ╩▒╩śˇ¬śċĻĭ ||╞ẅẅ¼ćś Ļ╩╩▒Ő▒ Φ ĝĻ↞eľ╞╩ĩś ĝś ╞ẅ▒╩½ć▒ś ĭċ½ĝľ ╠½ć▒╩ś ╞ś Ő|▒ç



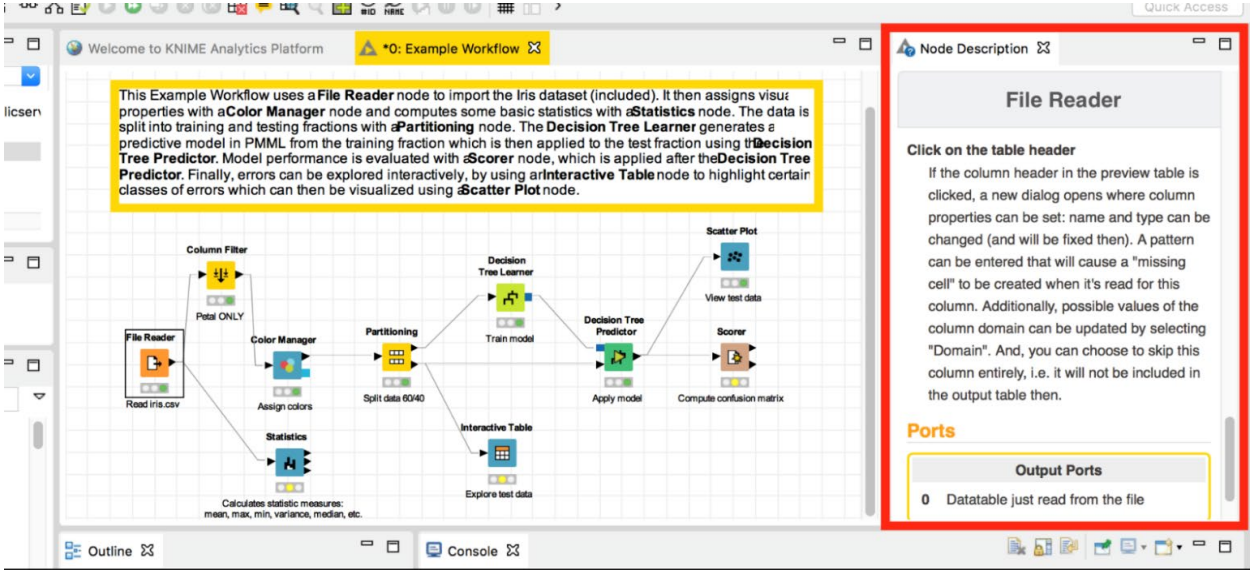
2.2.6. L'area di lavoro: coach del flusso di lavoro

Se abbiamo dato il permesso di raccogliere i nostri dati, questa sezione ci suggerisce quali nodi è più probabile che dobbiamo utilizzare in un determinato momento.

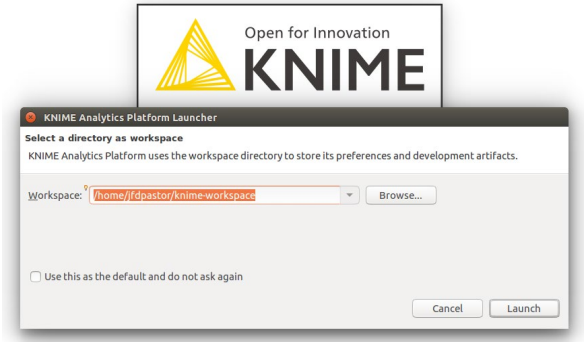


2.2.7. L’area di lavoro: descrizione dei nodi

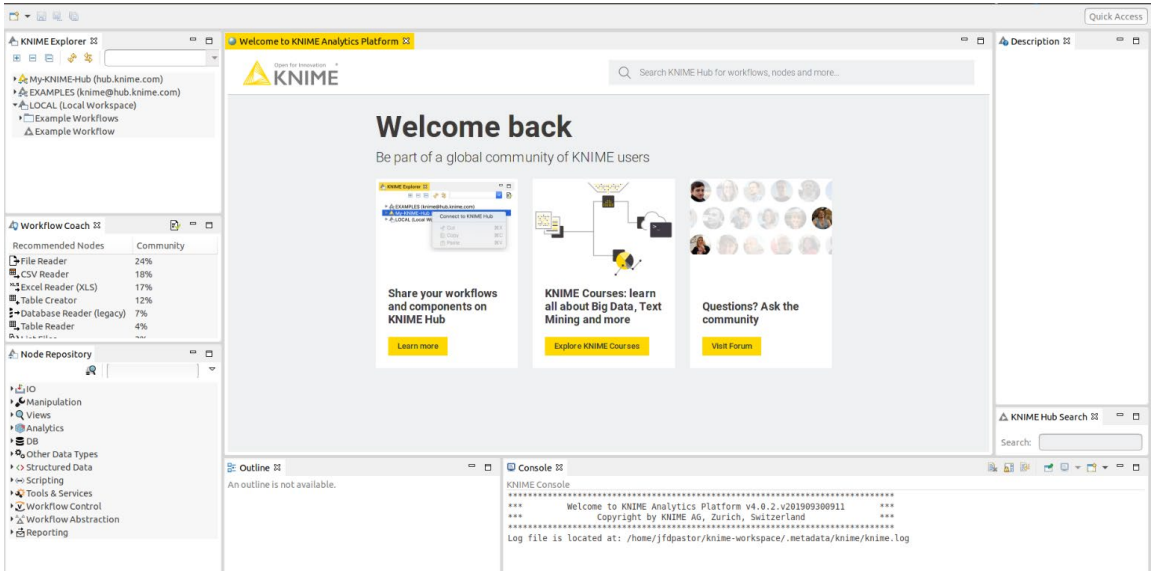
Œ| ï ¢¼ã ï ¼ Ő | Ļ ╩ ¢ |ĊĻ ¼Ő¢▒ ╠═▒¢═ ¼ã |M▒ ľ ≡ ś ¼ĈĈ¼¢ś ĊĻ ¼╩Ő▒ ĝ | ĝś ╞ẅ▒╩¼ Ļ ╩ ╩▒Ő▒ ľ ≡ ś Ĉś ¢═ ś ïïś Ő | Ŵ₿Ļ ¼╠ẅẅ¼¢ś ╠═▒¢═ ¼ẅ▒╩ | ĝĻ | ľ▒═ Ĉ ┠ | Ő¼ś ĝś ↑ Ļ ¢ś ś ĝĻ ╠╞ ĝĻś Ĉ▒¢ïś □╠↑ ¢ś ĝĝ | ś Ļĝľ ┠ś ▯



3. Esempio generico: classificazione delle specie floreali



Supponendo che KNIME sia già stato installato, si deve andare nella cartella in cui si trova ed eseguirlo facendo doppio clic sulla sua icona. All'apertura, ci chiederà la cartella "Workspace". Questa è la cartella in cui si troveranno tutti i nostri progetti.

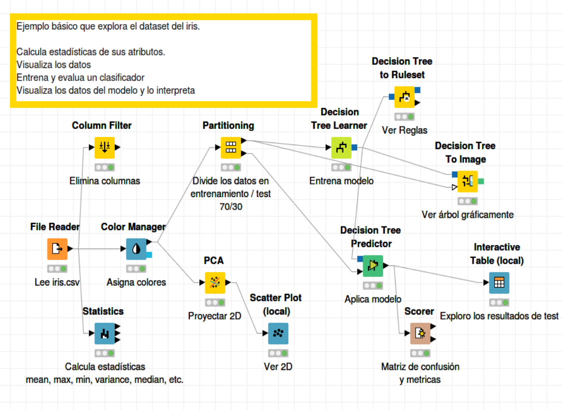


3.1. Esempio generico: classificazione delle specie floreali

- [illegible]
- [illegible]
- [illegible]
- [illegible]

Si tratta di un flusso di lavoro di base, che lavora con il set di dati iris. L'iris è costituito da 150 esempi appartenenti a 3 diverse specie di fiori. Ogni esempio ha 4 attributi che descrivono il fiore: lunghezza del sepalo, lunghezza del petalo, larghezza del sepalo e larghezza del petalo. Con questo set di dati, ci occuperemo di:

- calcolare le statistiche degli attributi.
- visualizzare i dati,
- addestrare e valutare un classificatore.



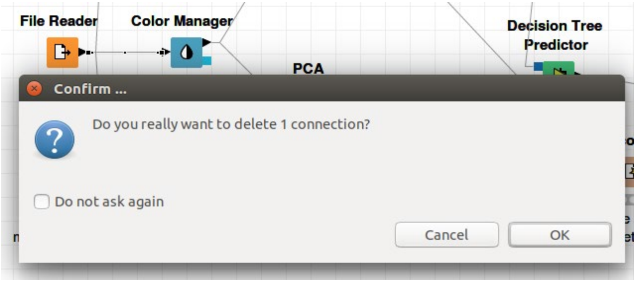
3.2. Esempio generico: classificazione delle specie floreali

Nell'editor si può vedere una serie di nodi interconnessi. In questo editor, i nodi vengono trascinati, uniti, configurati ed eseguiti per eseguire operazioni e analisi sui dati.

Dispone di strumenti di navigazione come lo zoom in/out (ingrandimento o rimpicciolimento) e permette di aggiungere commenti.

È possibile eseguire ogni nodo o l'intero flusso di lavoro con pulsanti simili a "play".

È necessario eseguire i nodi successivi ogni volta che viene apportata una modifica a un nodo. In altre parole, se si modifica un parametro di un nodo che si trova all'inizio del flusso di lavoro, è necessario premere il pulsante play con le due frecce bianche per eseguire nuovamente tutti i nodi successivi.



3.3. Esempio generico: classificazione delle specie floreali

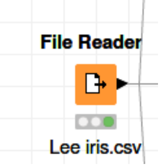
Loading data

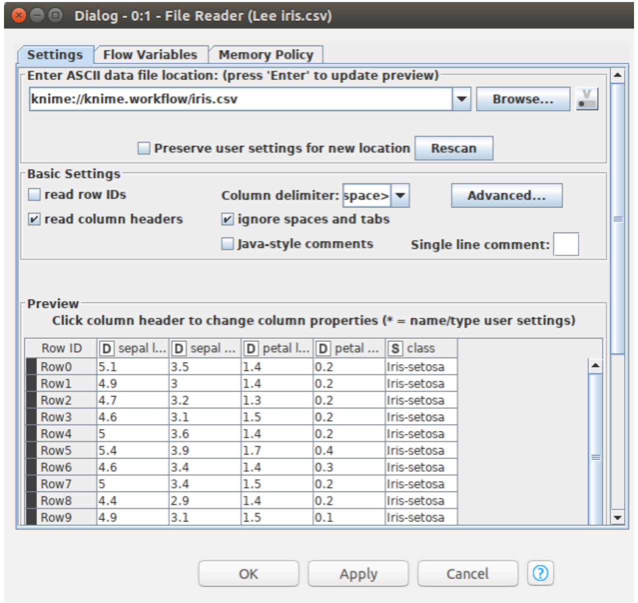
Il nodo "File Reader" è il nodo utilizzato per caricare i set di dati (leggere i dati da qualsiasi punto siano memorizzati). È possibile caricare i dati da un url (Internet) o dal disco rigido del computer.

[illegible]

[illegible]

[illegible]





, ⁑* ⁑□J ĝś ╧ Ĉ ▓□↑ ś ╧ś ċ┠▒□' ╞≠ĝĝ ├e l' ¼ẅ ▓╧ś □Ős ╟╞ □ĝĈś l' ├s □⅛▓ċś ¼╟

Colorazione dei dati

Il nodo "Colour Manager" consente di colorare il dataset in base ai valori di uno dei suoi attributi.

Il risultato è una tabella in cui ogni riga è colorata in base al valore dell'attributo scelto.

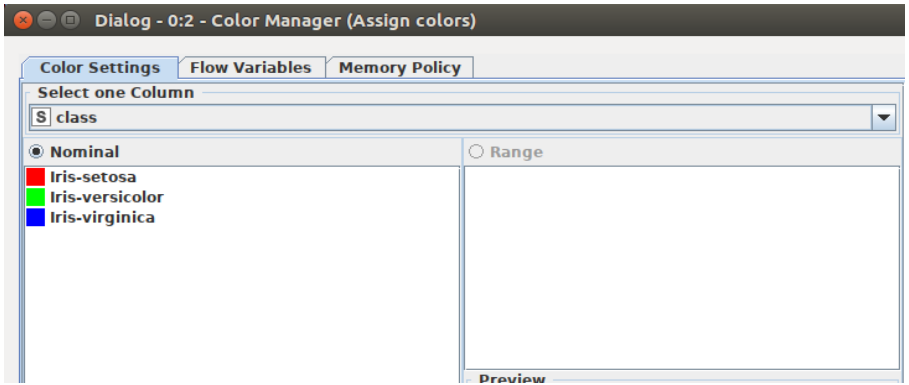


Table with Colors - 0:2 - Color Manager (Asigna colores)

File Hilite Navigation View

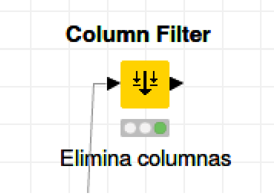
Table "default" - Rows: 150 | Spec - Columns: 5 | Properties | Flow

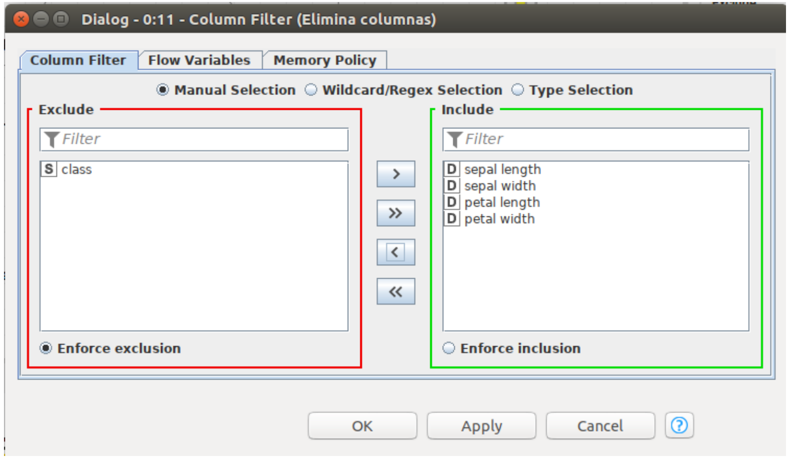
Row ID	D sepal l...	D sepal ...	D petal l...	D petal ...	S class
Row24	4.8	3.4	1.9	0.2	Iris-setosa
Row25	5	3	1.6	0.2	Iris-setosa
Row26	5	3.4	1.6	0.4	Iris-setosa
Row27	5.2	3.5	1.5	0.2	Iris-setosa
Row28	5.2	3.4	1.4	0.2	Iris-setosa
Row29	4.7	3.2	1.6	0.2	Iris-setosa
Row30	4.8	3.1	1.6	0.2	Iris-setosa
Row31	5.4	3.4	1.5	0.4	Iris-setosa
Row32	5.2	4.1	1.5	0.1	Iris-setosa
Row33	5.5	4.2	1.4	0.2	Iris-setosa
Row34	4.9	3.1	1.5	0.2	Iris-setosa
Row35	5	3.2	1.2	0.2	Iris-setosa
Row36	5.5	3.5	1.3	0.2	Iris-setosa
Row37	4.9	3.6	1.4	0.1	Iris-setosa
Row38	4.4	3	1.3	0.2	Iris-setosa
Row39	5.1	3.4	1.5	0.2	Iris-setosa
Row40	5	3.5	1.3	0.3	Iris-setosa
Row41	4.5	2.3	1.3	0.3	Iris-setosa
Row42	4.4	3.2	1.3	0.2	Iris-setosa
Row43	5	3.5	1.6	0.6	Iris-setosa
Row44	5.1	3.8	1.9	0.4	Iris-setosa
Row45	4.8	3	1.4	0.3	Iris-setosa
Row46	5.1	3.8	1.6	0.2	Iris-setosa
Row47	4.6	3.2	1.4	0.2	Iris-setosa
Row48	5.3	3.7	1.5	0.2	Iris-setosa
Row49	5	3.3	1.4	0.2	Iris-setosa
Row50	7	3.2	4.7	1.4	Iris-versic...
Row51	6.4	3.2	4.5	1.5	Iris-versic...
Row52	6.9	3.1	4.9	1.5	Iris-versic...
Row53	5.5	2.3	4	1.3	Iris-versic...
Row54	6.5	2.8	4.6	1.5	Iris-versic...
Row55	5.7	2.8	4.5	1.3	Iris-versic...
Row56	6.3	3.3	4.7	1.6	Iris-versic...
Row57	4.9	2.4	3.3	1	Iris-versic...

3.5. Esempio generico: classificazione delle specie floreali

Rimuovere colonne

[illegible]



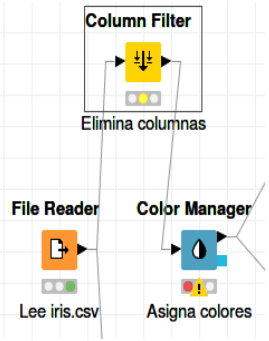


Nell'esempio, la colonna "Classe", che contiene il nome della specie a cui appartiene il fiore descritto in ogni esempio, verrà eliminata.

L'operazione viene eseguita solo per causare un errore nel flusso di lavoro.

Saper identificare i tipi di errore è fondamentale per utilizzare uno strumento come KNIME.

Errori e notifiche

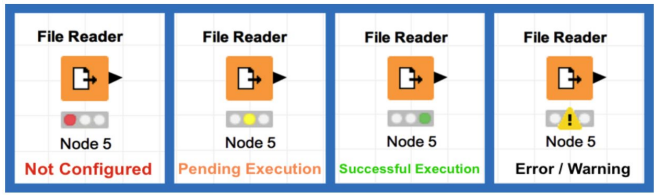


- Rimuovere il collegamento tra "File Reader" e "Color Manager".
- Configurare "Column Filter" per rimuovere la classe.
- Colleghiamo "Column Filter" con "Color Manager".

Ora si ottiene un errore in "Color Manager", perché questo nodo ha usato la classe per dare colore agli esempi. Per continuare, ristabiliamo la connessione tra "File Reader" e "Color Manager".
Un nodo può trovarsi in 4 stati diversi.

- Non configurato. È necessario fare doppio clic su di esso e scegliere qualche parametro importante che lo strumento non può scegliere per noi.
- In attesa. Il pulsante di esecuzione non è ancora stato premuto.
- Eseguito.
- Errore/Avvertimento. Non può essere eseguito.

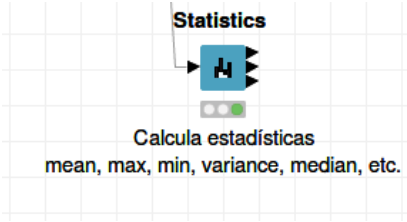
(Come sopra, quando si elimina una colonna utilizzata da un nodo successivo).
(Colonna utilizzata da un nodo successivo).



3.7. Esempio generico: classificazione delle specie floreali

Calcolare statistiche

Il nodo "Statistiche" consente di ottenere statistiche da una tabella di dati. Selezionando il nodo e facendo clic su "Visualizzazione statistiche", si ottiene una tabella con le statistiche per ciascun attributo.



Statistics View - 0:9 - Statistics (Calculates statistic measures:)

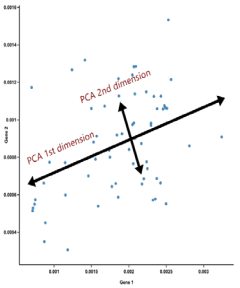
File

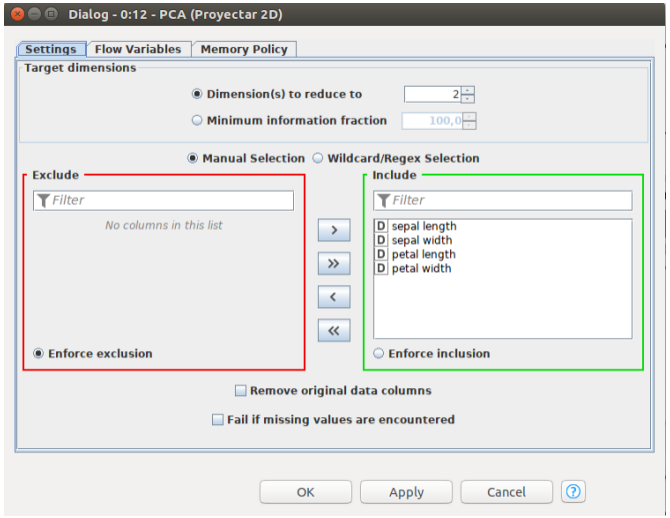
Numeric | Nominal | Top/bottom

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
sepal length	4,3	5,8433	?	7,9	0,8281	0,3149	-0,5521	0	0	0	
sepal width	2	3,0573	?	4,4	0,4359	0,319	0,2282	0	0	0	
petal length	1	3,758	?	6,9	1,7653	-0,2749	-1,4021	0	0	0	
petal width	0,1	1,1993	?	2,5	0,7622	-0,103	-1,3406	0	0	0	

3.8. Esempio generico: classificazione delle specie floreali

Principal Component Analisys (analisi delle componenti principali)





Questa tecnica ci aiuta a visualizzare in 2D gli insiemi di dati che hanno più di 2 attributi, in modo da poter osservare se ci sono outlier, sovrapposizioni tra le classi o se il confine tra le classi è lineare o non lineare.

Nel nodo è possibile configurare il numero di componenti da calcolare.

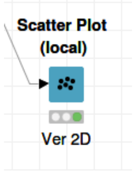
La PCA crea nuovi attributi, non visualizza direttamente; se vogliamo visualizzare, dobbiamo collegare un nodo per creare dei grafici (lo vedremo più avanti).

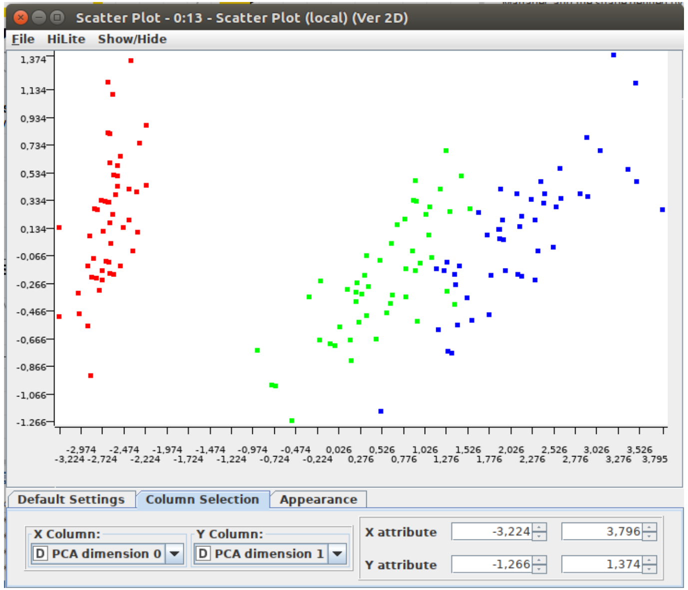
3.9. Esempio generico: classificazione delle specie floreali

Grafico di dispersione

[illegible]
[illegible]

Nell'esempio vengono visualizzate le prime due componenti principali; questa tecnica consente di riassumere e visualizzare (con due attributi) un insieme di dati con diverse colonne.

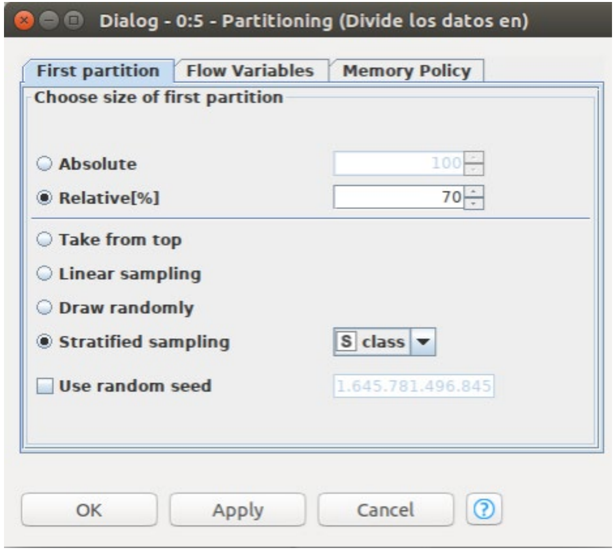


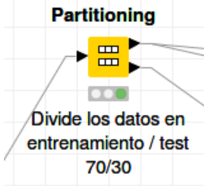


3.10. Esempio generico: classificazione delle specie floreali

Partizionamento dei dati

[illegible]



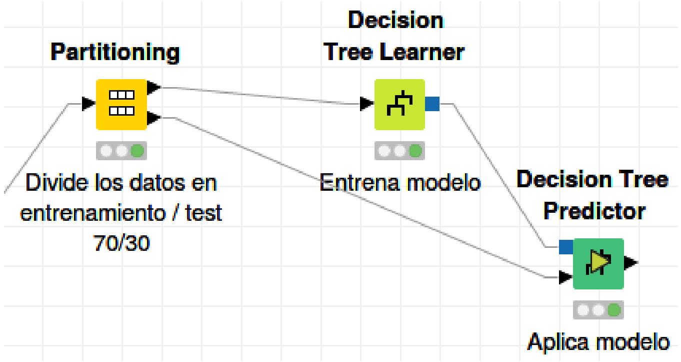


Nel nodo si può scegliere la percentuale di istanze da utilizzare per imparare e la percentuale da utilizzare per il test.

Esistono opzioni per rendere questo partizionamento completamente casuale o "stratificato", in modo da mantenere la proporzione degli spazi di lavoro.

3.11.Esempio generico: classificazione delle specie floreali

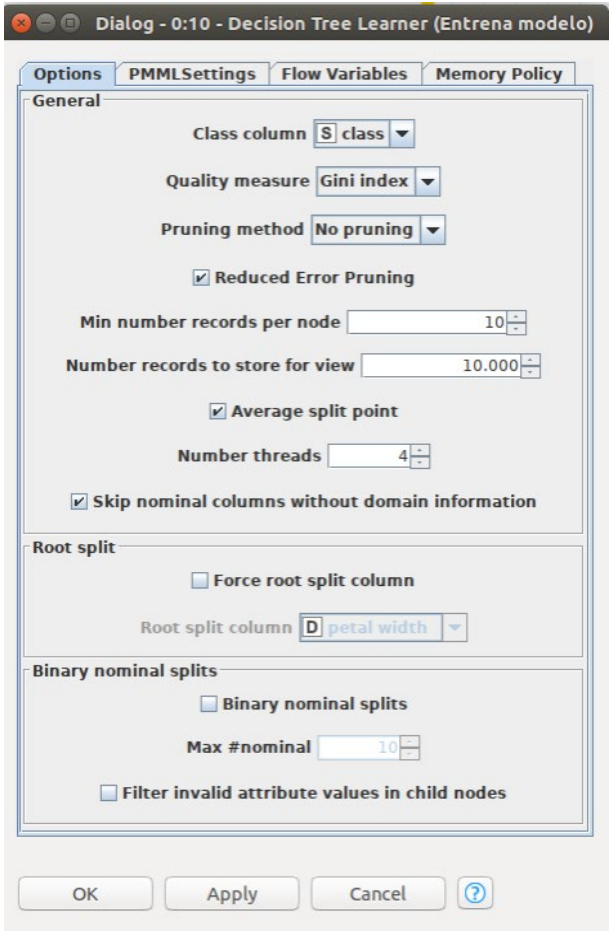
Creazione di modelli di data mining



; ċś ¼ẅ▓╚ś □Ő ├╞ ▓Ős ╟╫Ő ├Ő¼ī ¼╞ ╞╚↑

L'esempio mostra l'interno del nodo KNIME "Learner" di un albero di classificazione.

Permette di configurare, tra l'altro, la misura di qualità degli attributi, se ha un pruning o meno, ecc.



3.12. Esempio generico: classificazione delle specie floreali

Visualizzazione dei risultati

È possibile utilizzare un nodo interattivo di tipo tabella per visualizzare i valori della classe effettiva e della classe prevista, per tutti gli esempi di test.

In questo modo è possibile vedere gli esempi mal classificati.

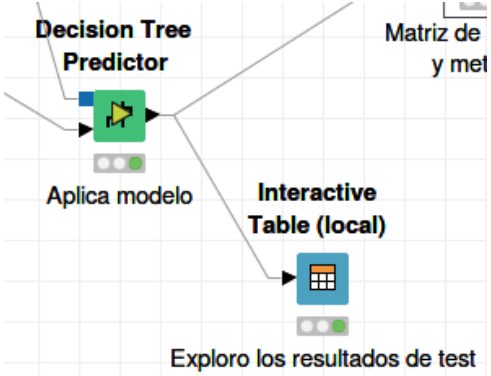
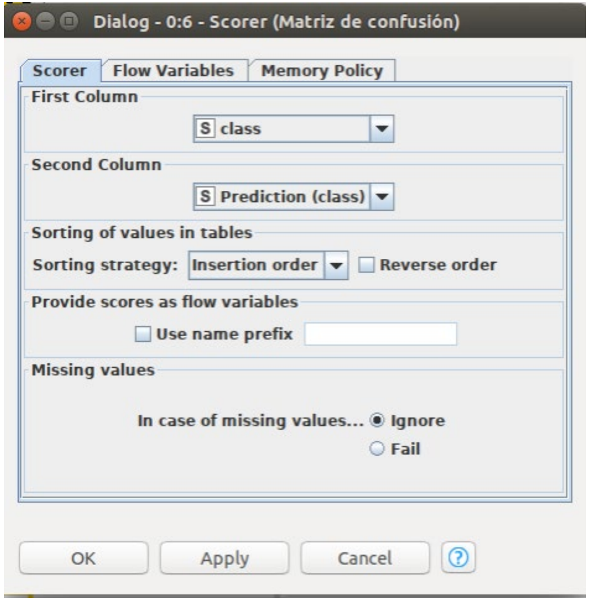


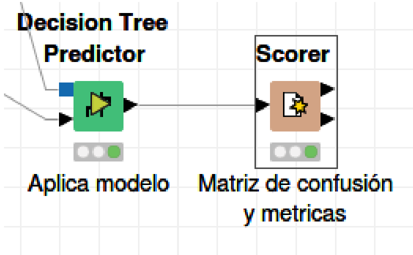
Table View - 0:7 - Interactive Table (local) (Exploro los resultados de test)

File Hilite Navigation View Output

Row ID	D sepal l...	D sepal ...	D petal l...	D petal ...	S class	S Prediction (class)
Row0	5.1	3.5	1.4	0.2	Iris-setosa	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa	Iris-setosa
Row16	5.4	3.9	1.3	0.4	Iris-setosa	Iris-setosa
Row17	5.1	3.5	1.4	0.3	Iris-setosa	Iris-setosa
Row18	5.7	3.8	1.7	0.3	Iris-setosa	Iris-setosa
Row24	4.8	3.4	1.9	0.2	Iris-setosa	Iris-setosa
Row27	5.2	3.5	1.5	0.2	Iris-setosa	Iris-setosa
Row28	5.2	3.4	1.4	0.2	Iris-setosa	Iris-setosa
Row30	4.8	3.1	1.6	0.2	Iris-setosa	Iris-setosa
Row32	5.2	4.1	1.5	0.1	Iris-setosa	Iris-setosa
Row37	4.9	3.6	1.4	0.1	Iris-setosa	Iris-setosa
Row39	5.1	3.4	1.5	0.2	Iris-setosa	Iris-setosa
Row40	5	3.5	1.3	0.3	Iris-setosa	Iris-setosa
Row44	5.1	3.8	1.9	0.4	Iris-setosa	Iris-setosa
Row46	5.1	3.8	1.6	0.2	Iris-setosa	Iris-setosa
Row52	6.9	3.1	4.9	1.5	Iris-versicolor	Iris-virginica
Row59	5.2	2.7	3.9	1.4	Iris-versicolor	Iris-versicolor
Row60	5	2	3.5	1	Iris-versicolor	Iris-versicolor

Creazione di modelli di data mining





Per configurare il nodo "Scorer" dobbiamo definire quale sia la colonna che rappresenta la classe reale e quale quella della classe prevista dal modello.

3.13. Esempio generico: classificazione delle specie floreali

Nell'esempio concreto (set di dati dell'iride, utilizzando un albero di classificazione, il 70% dei dati per imparare e il restante 30% per valutare) viene mostrato il tasso di successo complessivo del modello sul set di dati di test (91%) e altre metriche come i falsi positivi o i falsi negativi per ciascuna delle classi.

Accuracy statistics - 0:6 - Scorer (Matriz de confusión)

File Hilite Navigation View

Table "default" - Rows: 4 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalseP...	TrueN...	FalseN...	Recall	Precision	Sensiti...	Specifity	F-mea...	Accur...	Cohen'...
Iris-setosa	15	0	30	0	1	1	1	1	1	?	?
Iris-versicolor	12	1	29	3	0.8	0.923	0.8	0.967	0.857	?	?
Iris-virginica	14	3	27	1	0.933	0.824	0.933	0.9	0.875	?	?
Overall	?	?	?	?	?	?	?	?	?	0.911	0.867

Visualizzazione dei risultati

Nello stesso nodo si può ottenere anche la matrice di confusione, che mette in relazione la classe reale con la classe prevista, per osservare i tipi di errori nel modello.

In questa tabella:

- righe: sono le classi reali.
- colonne: sono le classi previste.

Nell'esempio:

- 3 esempi di iris-versicolour sono stati classificati erroneamente come iris-virginia.
- 1 esempio di iris-virginia è stato erroneamente classificato come iris-versicolor.

Confusion matrix - 0:6 - Scorer (Matriz de confusión)

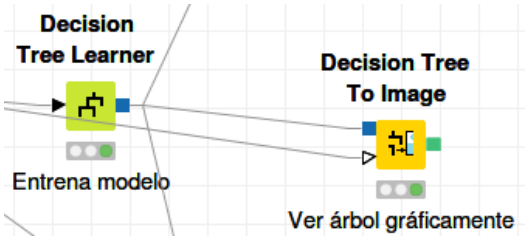
File Hilite Navigation View

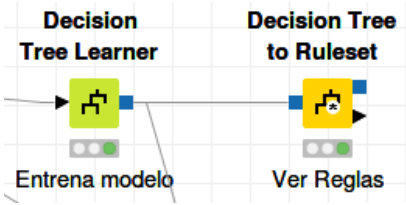
Table "spec name" - Rows: 3 | Spec - Columns: 3 | Properties

Row ID	Iris-set...	Iris-ver...	Iris-vir...
Iris-setosa	15	0	0
Iris-versicolor	0	12	3
Iris-virginica	0	1	14

3.14. Esempio generico: classificazione delle specie floreali

Visualizzazione 1 modello





Visualizzazione dell'albero di classificazione.

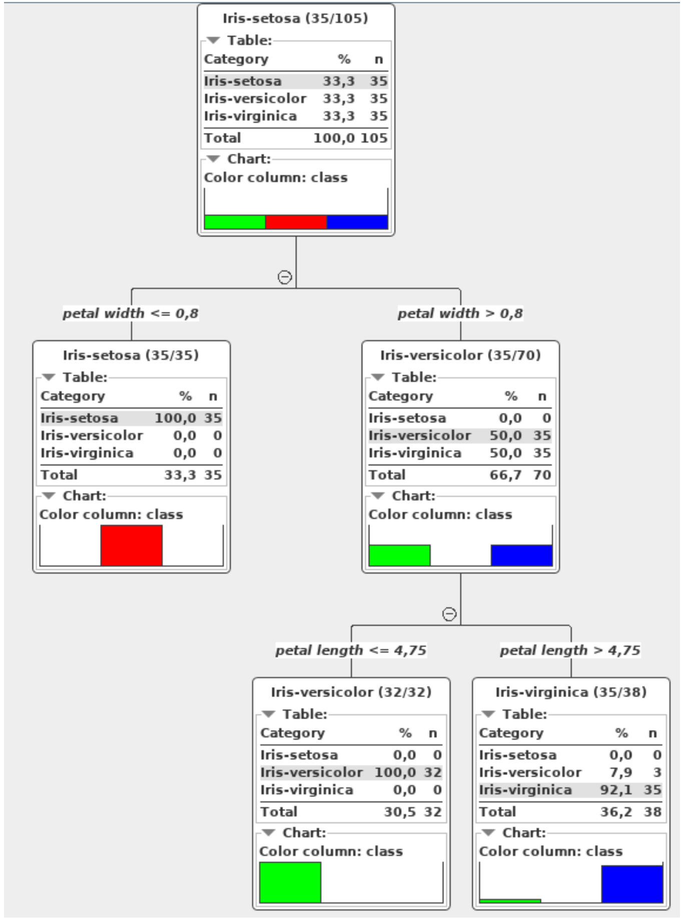
- Se sì

  Se la larghezza dei petali è inferiore a 0,8, classifichiamo l'esempio come setosa.

- Se non

  Se la lunghezza dei petali è inferiore a 4,75 classifichiamo l'esempio come versicolor.

  In caso contrario, classifichiamo l'esempio come virgínica.



Visualizzazione del modello

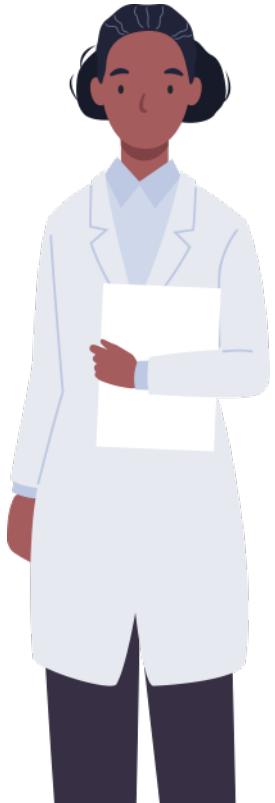
Un albero più grande sarebbe poco pratico da visualizzare graficamente, quindi può essere tradotto in un insieme di regole, in una rappresentazione più compatta.

Rules table - 0:14 - Decision Tree to Ruleset (Ver Reglas)

File Hilite Navigation View

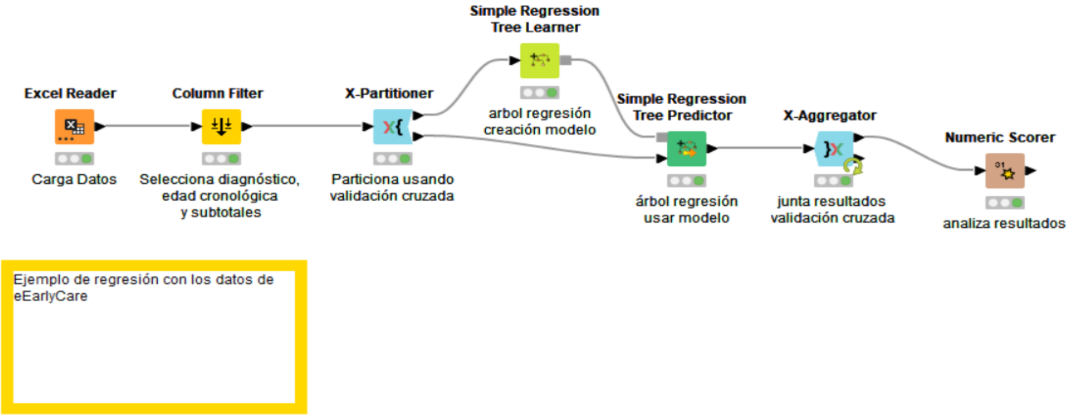
Table "default" - Rows: 3 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	S Rule	D Record count	D Number of correct
Row1	$petal width$ <= 0.8 AND TRUE => "Iris-setosa"	35	35
Row2	$petal length$ <= 4.75 AND $petal width$ > 0...	32	32
Row3	$petal length$ > 4.75 AND $petal width$ > 0....	38	35



- È necessario scaricare il file "eEarlyCare.knwf".
- In KNIME Explorer fare clic con il tasto destro del mouse e poi "Importa flusso di lavoro KNIME ...".
- Quindi l'opzione "Seleziona file" →"Sfoglia".
- Selezionarlo e cliccare su "Ok".

Si tratta di un flusso di lavoro che utilizza gli item della scala eEarlyCare, l'età cronologica e il sesso come variabili indipendenti e la diagnosi principale come variabile dipendente. Esplora il flusso di lavoro.



Riferimenti dal web

https://www.knime.com