Data Science e Tecnologie per le Basi di Dati

Esercitazione #3 – Data mining

Obiettivo

Applicare algoritmi di data mining per la classificazione al fine di analizzare dati reali mediante l'utilizzo dell'applicazione RapidMiner.

Dataset

Il dataset denominato Utenti (Utenti.xls, scaricabile dalla pagina del corso all'indirizzo http://dbdmg.polito.it/wordpress/teaching/data-science-e-tecnologie-per-basi-dati/#Laboratori, raccoglie dati anagrafici e lavorativi relativi agli utenti americani di un'azienda. Gli utenti sono classificati come "basic" o "premium" in base alla tipologia di servizi richiesti. Ciascun record del dataset fa riferimento ad un utente distinto. Il dataset contiene circa 32,000 utenti differenti e per ciascuno di essi sono riportati alcuni dati personali relativi all'utente (ad es., età, sesso, settore lavorativo principale) e la classe a lui assegnata ("basic" o "premium"). L'attributo relativo alla classe dell'utente, usato come attributo di classe durante l'esercitazione, è riportato come ultimo attributo di ciascun record.

La lista completa degli attributi del dataset da analizzare è riportata di seguito.

(1) Age
(2) Workclass
(3) FlnWgt
(4) Education record
(5) Education-num
(6) Marital status
(7) Occupation
(8) Relationship
(9) Race
(10) Sex
(11) Capital Gain
(12) Capital loss
(13) Hours per week
(14) Native country
(15) **Class (attributo di classe)**

Contesto di analisi

Gli analisti della compagnia vogliono predire la classe di un nuovo utente sulla base delle caratteristiche degli utenti che hanno sottomesso ciascuna richiesta. A tale scopo, gli analisti decidono di utilizzare tre differenti algoritmi di classificazione: un albero di decisione (Decision Tree), un classificatore Bayesiano (Naïve Bayes), e un classificatore di tipo distance-based (K-NN). Il dataset Utenti è utilizzato per la generazione dei modelli di classificazione e per la validazione delle loro performance.

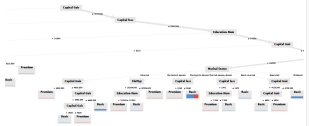
Finalità dell'esercitazione

Lo scopo dell'esercitazione è generare e analizzare diversi modelli di classificazione e validarne le performance sul dataset Utenti mediante l'ausilio del tool Rapid Miner. All'interno di Rapid Miner saranno generati diversi processi. Per valutare le performance dei classificatori saranno testate diverse configurazioni e i rispettivi risultati saranno confrontati tra loro. Per la validazione delle performance sarà applicato un processo di tipo 10-fold Stratified Cross-Validation. I risultati ottenuti saranno analizzati al fine di capire l'impatto dei principali parametri di input di ciascun algoritmo sulle performance di classificazione.

Domande

Rispondere alle seguenti domande:

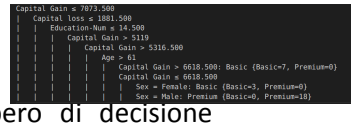
1. Generare un albero di decisione usando l'intero dataset per il training e la configurazione di default per l'algoritmo Decision Tree.



   (a) Quale attributo è considerato dall'algoritmo il più selettivo al fine di predire la classe di un nuovo dato di test? Capital Gain

   (b) Qual è l'altezza dell'albero di decisione generato?

   (c) Trovare un esempio di partizionamento puro all'interno dell'albero di decisione generato.



maximal depth-1 perche per definizione di altezza si usa numero di archi attraversati, mentre rapidminer intende numero di nodi

2. Analizzare l'impatto del minimal gain (considerando il gain ratio come criterio di splitting) e del maximal depth sulle caratteristiche dell'albero di decisione generato dall'intero dataset.
3. Cosa accade se modifichiamo l'attributo di classe da "Service Class" a "Native Country"? Rispondere nuovamente alla domanda (1) in questo nuovo scenario. Ora Education Number è l'attributo più selettivo.
4. Considerando di nuovo l'attributo "Service Class" come attributo di classe e applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del minimal gain e del maximal depth sull'accuratezza media ottenuta da Decision Tree? Confrontare le matrici di confusione ottenute usando diverse configurazioni per i parametri sopra citati (mantenere la configurazione di default per tutti gli altri parametri).
5. Considerando il classificatore K-Nearest Neighbor (K-NN) e applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del parametro K sulle performance del classificatore? Confrontare le matrici di confusione ottenute usando diversi valori di K. Applicare un 10-fold Stratified Cross-Validation con il classificatore Naïve Bayes. K-NN ottiene mediamente prestazioni superiori o inferiori a Naïve Bayes classifier sul dataset analizzato?
6. Analizzare la matrice di correlazione per valutare la correlazione tra coppie di attributi del dataset. Alla luce dei risultati ottenuti, l'ipotesi d'indipendenza Naïve risulta valida per il dataset Utenti?

La correlazione calcolata dalla matrice di correlazione non e' quella da te citata, ma invece quella di Pearson, calcolata come $Cov(X,Y)/\sigma X \sigma Y$. Questo coefficiente, calcolato per ogni coppia di variabili, ha le seguenti caratteristiche:

- e' normalizzato nel range [-1, 1] (dividiamo per le deviazioni standard delle due variabili)
- ha significato analogo alla covarianza (>> 0 -> c'e' correlazione positiva, << 0 -> correlazione negativa, 0 -> non c'e' correlezione), con la differenza che la covarianza non e' normalizzata
- ha solo senso nel momento in cui si usano variabili numeriche: per coppie di variabili in cui almeno una e' categorica (e.g. marital status), non e' possibile calcolare la correlazione (con questo approccio). Quindi, RapidMiner segna con "?" tutte queste situazioni.

Quindi , quello che devi guardare nella matrice di correlazione e' se i valori (non sulla diagonale, che ovviamente avranno correlazione 1) sono ~ 0

Gli unici valori di correlazione presenti sono tutti molto prossimi allo 0 -> Ipotesi di indipendenza naive risulta valida per il dataset Utenti.

-0.2293091490264182 -0.08883173120974495 -0.07664586787504137 -0.04847964686869168 -0.04556735467868451 -0.043194632733023694 -0.03161506295221493
-0.026858045269873267 -0.018768490610746827 -0.012280054339657414 -0.010251711675329405 4.3188579188687987E-4 0.03652718946410697 0.05425636227262286
0.057774539478969864 0.06875570750956339 0.077674498166006 0.07840861539012316 0.07992295668663368 0.1226301146922376 0.14812273262292563

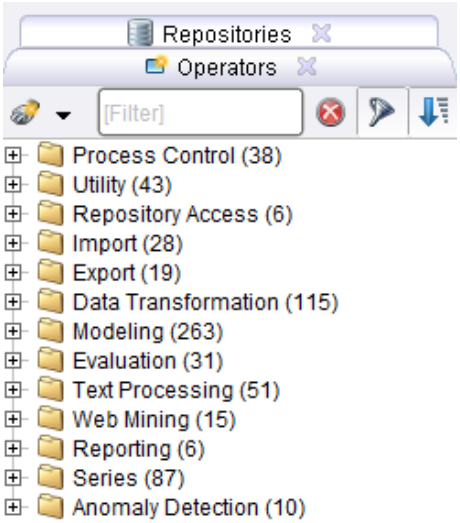
Esercitazione

Installazione e configurazione del programma

- Lanciare RapidMiner in ambiente Windows

Generazione e analisi del processo

- Creare un nuovo processo in Rapid Miner.
- Comporre il flusso del processo di data mining da eseguire selezionando e trascinando gli operatori disponibili sul menu a sinistra all'interno della finestra relativa al processo principale.



-

Figura 1. Operatori.

- Per gestire l'esecuzione del processo usare i pulsanti Start/Stop/Pause. Per visualizzare I risultati del processo, cambiare la prospettiva di visualizzazione da *Design* a *Results*.

Figura 2. Pulsanti di esecuzione / modifica della prospettiva di visualizzazione.

- Visionare il contenuto del dataset Utenti, disponibile in formato Excel (.xls).
- Importare i dati d'ingresso all'interno del processo di Data Mining principale mediante l'uso dell'operatore "Read Excel". Per importare i dati correttamente usare il *Data Import Wizard* configurando l'operatore come segue:
    - Selezionare il file sorgente desiderato (Step 1).
    - Selezionare tutte le celle del foglio di lavoro in cui sono contenuti i dati (Step 2).
    - Annotare la prima riga come quella contenente i nomi degli attributi (etichetta "name"), mantenendo non etichettate ("-") le righe dei dati allo Step 3.
    - Collegare il blocco di data import con il data source. Identificare il ruolo dell'attributo "Service class" come "label", ovvero "etichetta di classe" (Step 4). Data set meta data information
- Includere, in coda al processo di mining, l'operatore relativo al classificatore "Decision Tree". Il processo costruito finora sarà analogo al seguente:

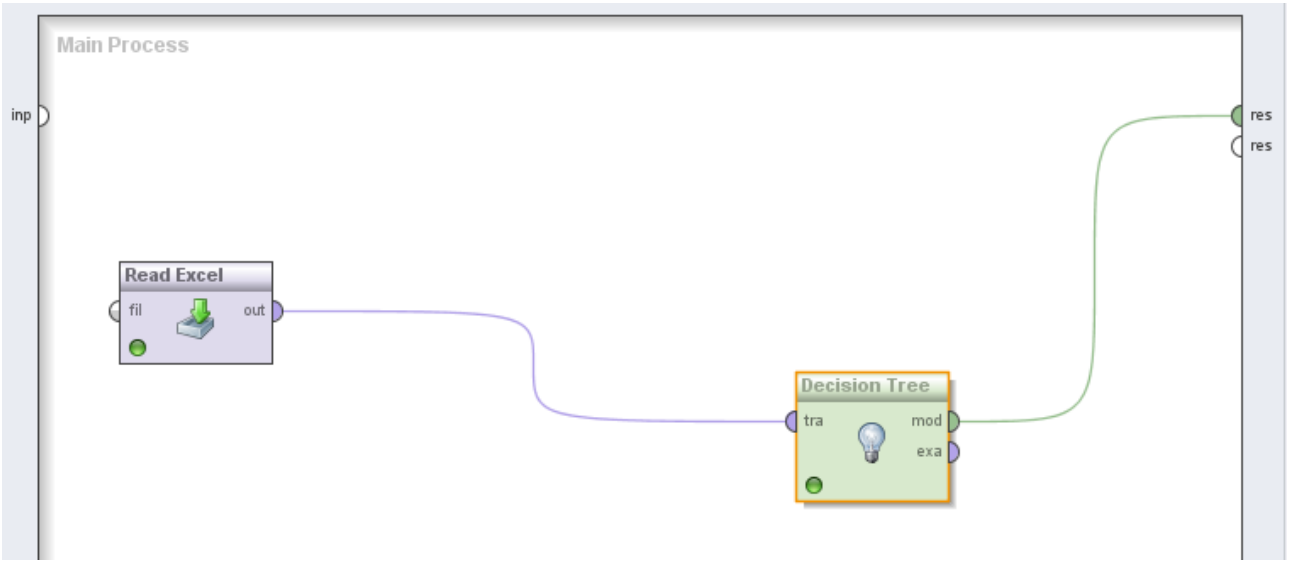


Figura 3. Processo di classificazione – Albero di decisione

- Eseguire il processo e analizzare l'albero di decisione generato mediante la Results perspective.
- Cambiare la configurazione dei parametri d'ingresso del classificatore cliccando sull'operatore Decision Tree e modificando le impostazioni relative nel menu posto sulla destra all'interno della Design perspective. In particolare, modificare i valori di maximal depth e di minimal gain, mantentendo inalterati tutti gli altri parametri, per analizzare il loro effetto sulle caratteristiche principali del modello di classificazione generato.
- Cliccare sull'operatore "Read Excel" al fine di poter modificare le impostazioni relative. Cambiare l'attributo di classe all'interno tra le "Data set metadata information" da "Service Class" a "Native Country" (altrimenti, rieseguire l'intero processo di data import mediante il wizard selezionando il nuovo attributo di classe allo Step 4).
- Rieseguire il processo per generare un nuovo albero di decisione.
- Modificare il flusso del processo principale per poter eseguire una 10-fold Stratified Cross-Validation. A tale scopo, come primo passo includere il blocco "Validation" al posto di Decision Tree nel processo principale.

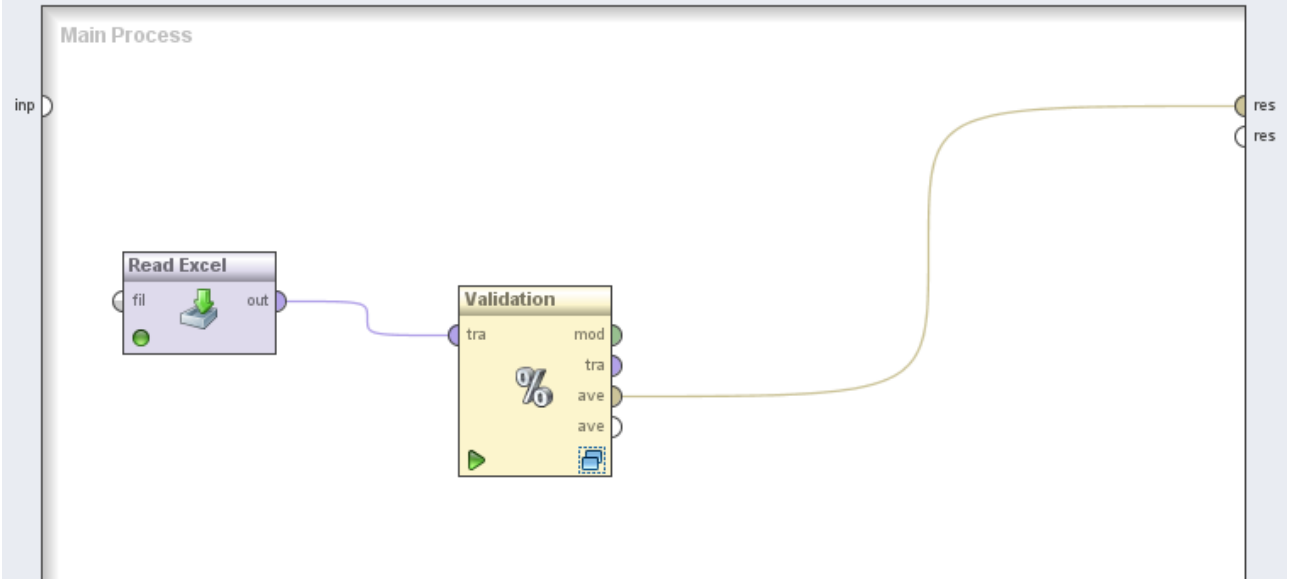


Figura 4. 10-Fold Cross-Validation.

Come passo successivo, fare doppio click sull'operatore "Validation" e creare un processo innestato analogo a quello sotto riportato:

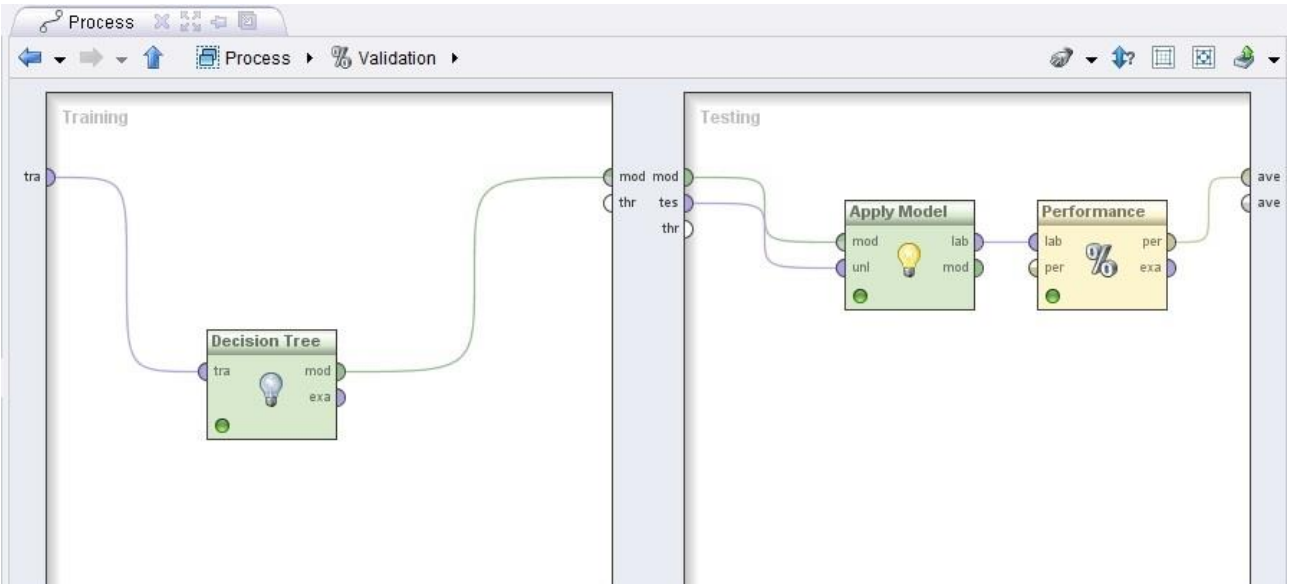


Figure 5. Validation subprocess.

- Tornare al Results perspective e analizzare la matrice di confusione generata dal processo di validazione.
- Disabilitare temporaneamente l'operatore Decision Tree (cliccando col tasto destro sull'operatore e eliminando il segno di spunta a lato di "Enable Operator"). Sostituire l'operatore Decision Tree con Naïve Bayes prima e con K-NN successivamente.
- Confrontare le performance di K-NN e Naïve Bayes, in termini di accuratezza media, precisione e richiamo, analizzando le rispettive matrici di confusione. Per il classificatore K-NN, variare i valori del parametro K usando il menu sul lato destro nella Design perspective.
- Per analizzare la matrice di correlazione associata al dataset in esame tornare al processo principale (click sul pulsante "Process"), disabilitare temporaneamente l'operatore Validation (cliccando col tasto destro sull'operatore e eliminando il segno di spunta a lato di "Enable Operator"), inserire l'operatore "Correlation Matrix" in coda al processo e visualizzare la rispettiva matrice collegando il plug-in del blocco denominato "mat" al plug-in "Result" sulla destra della finestra del processo principale. Il processo così generato sarà analogo al seguente:

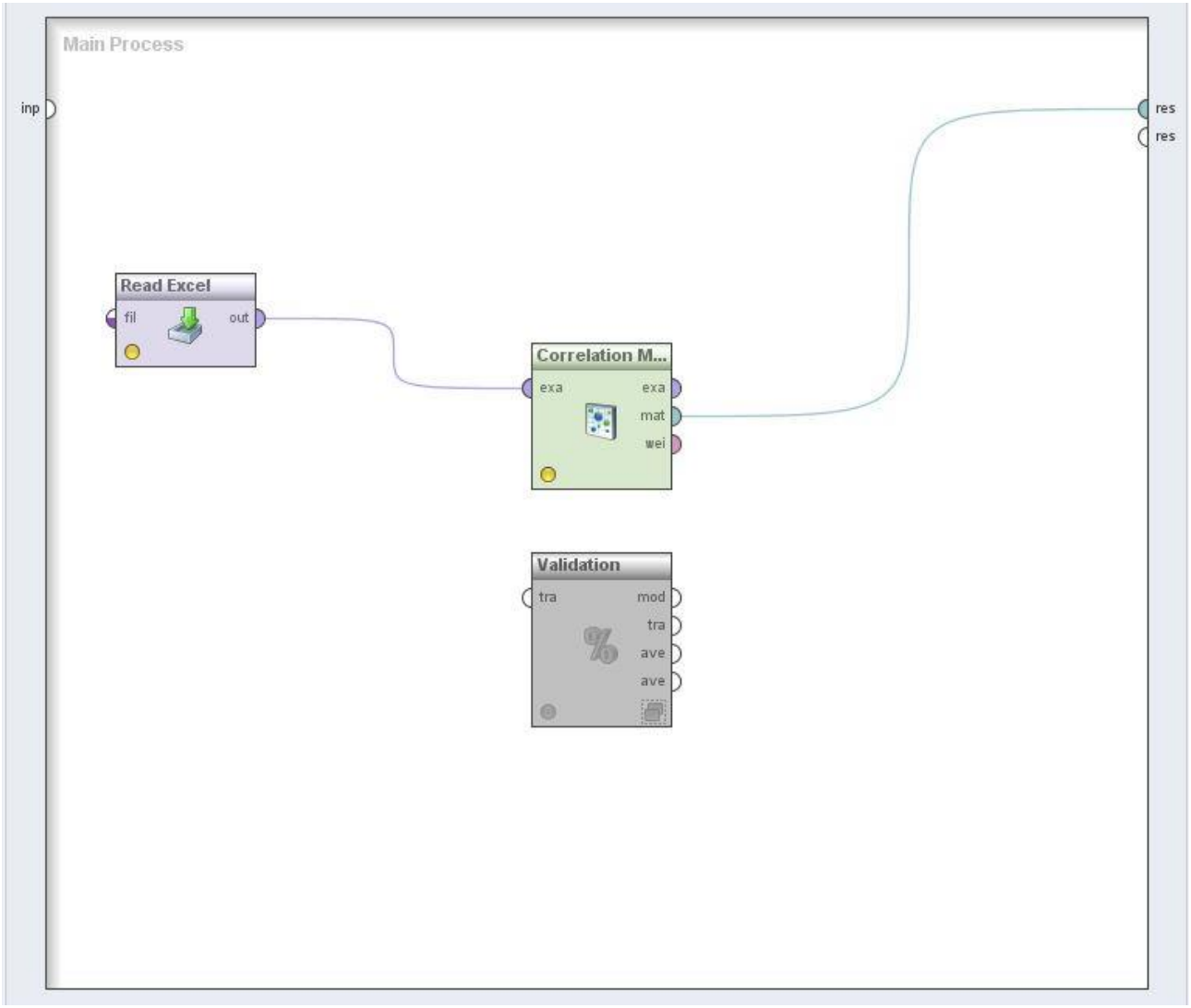


Figure 6. Matrice di correlazione tra attributi

Tornando al Results perspective, per ordinare le correlazioni trovate tra coppie di attributi in ordine decrescente selezionare la “Pairwise Table view” e cliccare sul campo “Correlation” della tabella visualizzata.