Domande a Crocette

Se l'indice di Gini misurato per la variabile "Categoria Acquistata" è pari a 1, vuol dire che l'utente:

- acquista categorie di prodotto sempre diverse.

Per Immobiliare.it, l'aumento dei tassi di interesse è:

- forza macro-economica.

Per un campo rappresentativo della zona di residenza in Italia, che assume i valori "nord" e "sud", un valore mancante può essere sostituito con:

- il valore più frequente.

Un metodo di apprendimento supervisionato presuppone la presenza di una:

- variabile target.

Nell'ipotesi di avere un database contenente il meteo dell'ultimo anno, per prevedere la temperatura posso utilizzare:

- una regressione.

Se l'indice di Skewness del campo "Soddisfazione" (1 = non soddisfatto; 10 = soddisfatto) è pari a 0.8, vuol dire che:

- è maggiore la frequenza di utenti non soddisfatti.

Se prendo in gestione un punto vendita di Starbucks, nel modello di business di Starbucks sono:

- un cliente.

Posso trasformare una variabile continua in una categorica:

- Si, discretizzando la variabile.

Se voglio migliorare l'accuratezza di un albero decisionale, posso provare ad:

- usare la cross-validation.

Per verificare se l'età di una popolazione è molto concentrata attorno alla media, posso utilizzare:

- l'indice di Kurtosis.

Se la coppia di variabili a-b ha una covarianza di 0.8 e quella c-d ha una covarianza pari a -0.9, vuol dire che:

- la coppia c-d presenta una discordanza.

Per la Playstation, l'ingresso di Apple nel mercato dei videogiochi è:

- forza industriale.

Il grado di valutazione, espresso con un voto da 1 a 10, è una variabile:

- ordinale.

Se ho un database contenente le campagne promozionali degli anni precedenti, per identificare gli utenti a cui inviare uno sconto del 10%, 15%, 20% posso usare:

- classificazione.

Se ho una variabile target categorica, posso utilizzare una regressione?

- no.

La value proposition principale di Netflix è:

- ampio catalogo;
- permettere di visualizzare contenuti in streaming.

Se riduco il numero di regole di un albero decisionale, posso aspettarmi:

- una riduzione dell'accuratezza dell'albero.

Per verificare se l'età di una popolazione è prevalentemente bassa o alta, posso usare:

- l'indice di Skewness.

Se la coppia di variabili a-b presenta un indice di Pearson pari a 0.8 e la coppia c-d lo ha di -0.9, la coppia con la relazione più intensa è:

- c-d.

Nel modello di business di Spotify, concentrandoci sul ruolo che hanno le case produttrici di musica, si segue un pattern:

- multi-sided.

La value proposition dipende da:

- i segmenti di clienti che voglio raggiungere

Per individuare il miglior bundle di prodotti da vendere, posso usare:

- regole associative.

Airbnb segue un pattern long tail?

- No.

Per una testata giornalistica il proliferare di piattaforme AI è:

- Trend.

Per un produttore di telefoni, i sistemi operativi sono:

- risorse chiave.

Due variabili con la stessa media hanno uguale dispersione?

- No.

Posso discretizzare il campo mese (che assume 12 valori)?

- Si.

Un metodo di apprendimento unsupervised presuppone la presenza di una variabile target?

- No.

Per prevedere lo stipendio mensile di un lavoratore libero professionista, posso utilizzare una regressione?

- Si.

Se l'indice di Kurtosis del campo "età" (18 - 100 anni) è pari a 0.8, vuol dire che:

- sono più gli utenti attorno ai 40 anni.

Se sky sponsorizza una squadra di calcio, questa diventerebbe per Sky un:

- Partner.

Posso trasformare una variabile categorica in continua?

- No.

Per aumentare la dimensione di un albero decisionale posso provare ad:

- abbassare il numero minimo di istanze per la generazione di un nodo.

Per verificare se c'è una maggiore frequenza di studenti con una media voto alta, posso utilizzare:

- indice di Skewness.

La value proposition di Nike è:

- alte performance dei propri prodotti.

Uno sportivo sponsorizzato da Nike e per Nike:

- partnership.

Un comparatore di prezzi può configurarsi come:

- multi-sided.

Per profilare gli utenti che guidano l'auto possiamo utilizzare:

- classificazione.

Se decidessi di sostituire i valori mancanti della variabile "peso" per identificazione, potrei inserire:

- -1.

Se ho una variabile targe continua, posso utilizzare un classificatore?

- Si, discretizzando la vaiabile.

Per misurare se degli acquisti vengono effettuati sempre in giorni diversi o nello stesso giorno, posso usare:

- indice di Gini.

Se voglio migliorare l'accuratezza di un albero decisionale:

- aumento il numero di regole.

Per verificare se la maggior parte di una popolazione ha la stessa altezza posso usare:

- indice di Kurtosis.

Per capire cosa, tra titolo di studio ed età, influenza maggiormente il reddito, uso:

- indice di Pearson.

Se R-squared è prossimo ad 1:

- il modello è buono.

Se utilizzo un classificatore per configurare la campagna sconti per cercare di trattenere i clienti che potrebbero chiudere il conto, monitorerò principalmente:

- matrice di confusione.

In una classificazione, se riduco la numerosità dei valori della variabile target posso aspettarmi:

- una riduzione del numero di regole.

Ho un database con informazioni sugli accessori montati sulle auto e voglio capire quali di questi vengono normalmente acquistati assieme. Quale variabile posso eliminare?

- Colore dell'accessorio.

Se voglio cercare regole più comuni:

- aumento il livello di supporto minimo.

Se Sky volesse adottare un modello di business Long Tail, dovrebbe:

- ridurre i costi di acquisto e gestione dei contenuti.

Le attività chiave sono legate direttamente a:

- value proposition.

Se un ristorante cambia i fornitori di materie prime, sta cambiando:

- partnership chiave.

Se voglio cercare regole meno ovvie:

- abbasso il livello del supporto minimo.

Canvas e Patterns

BNL (Banca Nazionale del Lavoro)

Value Proposition: brand/status; assistenza 24/7; sicurezza ed affidabilità (banca importante); controllare il conto corrente tramite app.
Customer Segment: mass market (tutte le persone a cui serve un prestito, un conto corrente o altri servizi di base); segmented market (imprese o clienti specifici che richiedono servizi o consulenze speciali).
Channels: filiali fisiche; sportelli automatici (ATM); web app (“Hello Banking”).
Customer Relationships: assistenza personale (call center o filiale fisiche); self-service (ATM).
Revenue Streams: subscription fees (gestione del conto, emissione di carte di credito/debito, bonifici); brokerage fees (interessi su prestiti, mutui, investimenti e servizi assicurativi).
Key Resources: risorse finanziarie (il capitale è la risorsa fondamentale); risorse umane (personale qualificato); risorse materiali (filiali, ATM, data center); risorse intellettuali (software proprietari, algoritmi per l’analisi dei dati).
Key Activities: gestione depositi, prestiti e rischio finanziario; servizio clienti; sviluppo e manutenzione IT; marketing (sponsorizzazione).
Key Partners: circuiti di pagamento (Visa, Mastercard); fornitori tecnologici; istituzioni finanziarie (per permettere le operazioni interbancarie).
Cost Structure: costi manutenzione IT; personale; infrastrutture; pubblicità; oneri finanziari.

Non segue nessuno dei **Pattern** studiati.

Sky

Value Proposition: visione di contenuti esclusivi (sport, film); comodità e possibilità di vedere i contenuti sul telefono tramite APP mobile; brand/status; abbonamenti personalizzati.
Customer Segments: mercato di nicchia (clienti abbonati che hanno acquistato un decoder)
Channels: call-center; advertising; piattaforma online e app; stand in negozi.
Customer Relationships: personali (personale che installa e fa manutenzione del modem a casa, assistenza tramite call center per fare upselling e retaining).
Revenue Streams: subscription fees (abbonamenti); usage fees (contenuti pay-per-view); asset sale (vendita dei decoder); advertising.
Key Activities: produzione di contenuti originali; distribuzione e trasmissione di contenuti; marketing (campagne pubblicitarie).
Key Resources: proprietà intellettuali (diritti sui contenuti e prodotti); risorse umane (tecnici, registi, montatori, giornalisti); infrastrutture tecnologiche (reti di trasmissione); risorse materiali (telecamere, studi di registrazione).
Key Partners: produttori di contenuti; fornitori di tecnologia ed infrastrutture.
Cost Structure: costi di produzione dei contenuti; costi di acquisto di licenze (contenuti/diritti); coasti di infrastrutture (manutenzione reti, server); costi di R&D (ricerca e sviluppo); costi di marketing; costi del personale.

Segue il Pattern **Unbuilding**, perché gestisce principalmente la relazione con il cliente e l'innovazione del prodotto e servizio. Per le infrastrutture si appoggia ad operatori satellitari (pur avendo anche reti di trasmissioni proprie); inoltre acquista contenuti da prodotti esterni.

Amazon

Value Proposition: consegna veloce; brand/status; affidabilità; ampia gamma di prodotti; buon rapporto qualità prezzo (prezzi bassi); vendità di prodotti sia nuovi che usati; servizi di musica; cloud services; facile da usare.
Customer Segments: mercato di massa; heavy users (abbonati che coprano prodotti frequentemente); venditori; imprese.
Channels: sito web; app; social (mass media); e-mail; call center.
Customer Relationships: automatiche [recensioni (co-creation), community (FAQ), sistemi di raccomandazione]; personali [assistenza telefonica (call center)].
Revenue Streams: asset sale (vendita di prodotti e serivizi); brokerage fees; subscription fees (abbonamenti a prime); usage fee (per i servizi cloud); advertising.
Key Activities: logistica; gestione piattaforma/servers (manutenzione); ricerca e sviluppo; marketing e advertising.
Key Resources: risorse fisiche (magazzini warehouse, impianti di produzione, data center, servers); risorse umane (specializzate per la R&D, manutenzione e programmazione del sito e app, corrieri); risorse intellettuali (logo, software e tecnologie proprietarie).
Key Partners: logistici (bartolini, gls, dhl, brt); SIAE; partner finanziari per le modalità di pagamento (Paypal. Visa, etc.).
Cost Structure: personale; costo manutenzione; costo di storage; costi pubblicitari; advertising.

Il pattern che segue è **Long Tail** perché vende molti prodotti di nicchia e abbatte i costi di inventario; **Multi-Sided** perché mettere in relazione clienti indimendenti tra loro (i venditori e i compratori); **Freemium** perché è possibile usare la piattaforma in modo gratuito ma è presente la versione prime a pagamento in cui risparmi i costi di spedizione ed hai un servizio più veloce oltre ad altri vantaggi (video, musica).

McDonald

Value Proposition: prezzi bassi; servizi veloci; facilità d'uso; brand; adatto a famiglie; McDrive.
Customer Segments: giovani; famiglie; viaggiatori; imprenditori (persone che volgiono aprire un McDonald's, attratte dal brand).
Channels: negozio fisico; social; mass media; canale dedicato imprenditori.
Customer Relationships: automatica (totem); personale (negli store).
Revenue Streams: asset sale (vendità di menù, panini, bibite); enetring ee (tassa niziale che un imprenditore paga per iniziare l'attività ed ottenere il brevetto); renting fee (affitto mensile pagato dagli imprenditorei per detenere lo store).
Key Resources: Negozi fisici; materiali economici (per mantenere i rezzi bassi); brand; personale (esperti marketing; impiegati per la produzione standardizzata).
Key Activities: produzione di massa; marketing/advertising; ricerca e sviluppo; istruzione e formazione degli impiegati.
Key Partners: fornitori dei giocattoli happy meals; coca-cola ed altre bibite che non sono competitor ed affiancano le bevande ai panini.
Cost Structure: costi di produzione del cibo; costi delle risorse umani; costi per advertising; costi partnership.

Può essere visto come un **Unbuilding** perché ha maggior flusso di ricavo derivante dagli affitti e percentuali sulle vendite dei franchisee. Quindi ha una focalizzazione sull'infrastruttura e la gestione quotidiana dello sviluppo dei prodotti sono delegate ai franchisee (partner nell'innovazione locale e relazione con il cliente).

Candy Crush

Value Proposition: intrattenimento; accessibilità; facile da usare; novità (sempre nuovi livelli); riduzione del rischio (è possibile giocare senza spendere soldi).
Customer Segments: mercato di massa; giocatori paganti.
Channels: social (facebook); app store (google ed apple).
Customer Relationships: automatica (tutorial in game); comunità (tramite social, puoi inviare vite o mosse agli amici).
Revenue Streams: asset sale (delle mosse speciali); advertising.
Key Activities: sviluppo e manutenzione dell'app; creazione di nuovi contenuti; marketing.
Key Resources: proprietà intellettuali (codice di gioco, design, brand); risorse umane (team di sviluppatori); infrastrutture tecnologiche (i server e gli strumenti di sviluppo sotware necessari); grande numero utenti giocanti.
Key Partners: social media (facebook in particolare).
Cost Structure: costi di sviluppo e manutenzione; costi di advertising; costi partnership.

Segue un pattern **Freemium** in quanto offre il gioco gratuitamente a tutti ma permete degli acquisti in app per avere dei vantaggi ed aiuti nel superare i livelli.

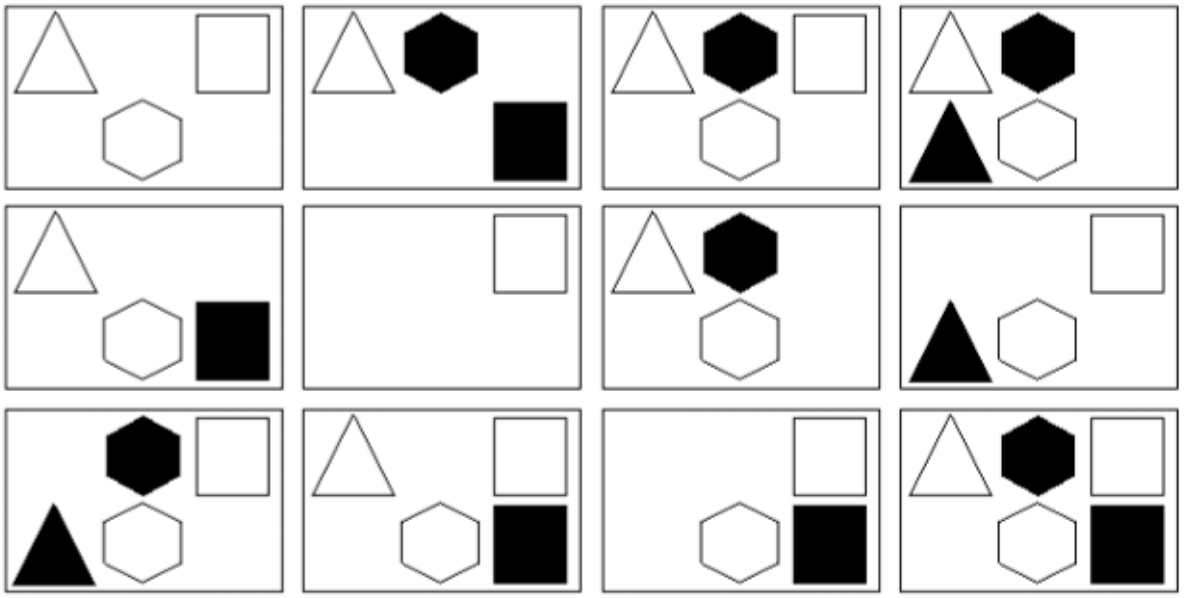
Youtube

Value Proposition: accesso gratuito a molti contenuti (per gli utenti); strumenti di monetizzazione e sponsorizzazione, distribuzione di contenuti (per i creator); ampia audience raggiungibile (per gli inserzionisti).
Customer Segments: mass market globale; creatore di contenui; iserzionisti.
Channels: sito web; app; API ed integrazioni.
Customer Relationships: automatica (algoritmi di raccomandazione, commenti); community/co-creation (gli utenti interagiscono tra loro e con i creatori attraverso i commenti, contribuendo al valore della piattaforma).
Revenue Streams: advertising; subscription fees (youtube premium che funziona senza publicità, ti dà accesso a contenuti esclusivi); data purchase (vendita di dati aggregati e profilazioni utente a terze perti per analisi di marketing).
Key Activities: sviluppo e manutenzione della piattaforma; moderazione dei contenuti; marketing e pubblicità; ricerca e sviluppo; gestione della relazione con creatori ed inserzionisti.
Key Resources: piattaforma e server; contenuti; brand; personale (sviluppatori, team marketing, controllori dei contenuti); dati utente (pubblicità mirata).
Key Partners: google (società madre) condivide le infrastrutture dati e tecnologie; creatori di contenuti; iserzionisti; produttori di dispositivi (integrazione di youtube in sistemi operativi, console, smartTV).
Cost Structure: R&S; personale; infrastruttura; marketing; copyright; partners.

Segue diversi patterns: **Long Tail** perché ha disposizione moltissimi video di nicchia; **Multi-Sided** perché collega i creatori di contenuti agli spettatori e gli insrzionisti; **Free as a BM** perché ha a disposizione due versioni (quella gratis e quella premium).

Business Intelligence

Esercizio 1



Quali sono i valori di **support** e **confidence** per la seguente regola?

(ogni retangolo è una transazione).

Supporto: numero di transazioni contenenti sia il quadrato che il triangolo diviso il numero di transazioni totali = 4/12 = 0,3.
Confidenza: numero di transazioni contenenti sia il quadrato che il triangolo diviso il numero di transazioni che contengono il quadrato = 4/5 = 0,8.

Esercizio 2

coppie	Covarianza	Correlazione
(a,b)	4.500	0.43
(b,c)	9.800	0.87
(c,a)	-5.400	-0.89

Quale coppia ha la relazione più forte e perchè?

Il coefficiente di Pearson (Correlazione) è quello che ci permette di capire se due varabili hanno una relazione lineare o meno. La coppia (c,a) in questo caso ha coefficiente di correlazione maggiore quindi è quello con relazione più forte.

Esercizio 3

```
J48 unpruned tree
------------------

Mese = Novembre: Natale (31.0)
Mese = Ottobre: Natale (0.0)
Mese = Dicembre: Natale (83.0)
Mese = Settembre: Natale (0.0)
Mese = Luglio: Natale (0.0)
Mese = Gennaio: Natale (8.0)
Mese = Marzo: Pasqua (23.0)
Mese = Maggio: Natale (0.0)
Mese = Aprile: Pasqua (14.0/1.0)
Mese = Febbraio: Pasqua (22.0)
Mese = Giugno: Natale (0.0)
Mese = Agosto: Natale (0.0)

Number of Leaves  :     12

Size of the tree :      13


Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         180               99.4475 %
Incorrectly Classified Instances         1                0.5525 %
Kappa statistic                          0.9874
Mean absolute error                      0.0103
Root mean squared error                  0.0716
Relative absolute error                  2.3521 %
Root relative squared error             15.3489 %
Total Number of Instances              181
Ignored Class Unknown Instances                232
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,992     0,000     1,000       0,992   0,996      0,987    0,893     0,671     Natale
               1,000     0,008     0,983       1,000   0,991      0,987    0,991     0,901     Pasqua
```

Considera il database di un e-commerce contenente informazoni sugli acquisti come la categoria di prodotti acquistati, la data dell'acquisto, il denaro speso ed il metodo di pagamento utilizzto. Per ciascun acquisto è indicato se è stato effettuato per un regalo di Natale o di Pasqua. L'obiettivo dell'analisi è quello di classificare gli acquisti per distringuere quelli fatti per Natale e quelli per Pasqua. Mondificheresti qualcosa o lo lascieresti così com'è, perché?

Le istanze classificate correttamente sono 99,5% rispetto a quelle classificate erroneamente (0,5%), quindi il modello ha un accuratezza alte (è un buon modello).
La precisione ed il Recall del modello sono entrambe alte e forniscono quindi un ottimo valore di F-measure. L'area della curva ROC è prossima all'1 il che indica una buona capacità del classificatore nel distinguere tra le due classi.

Questi risultati sono ottimi, il che suggerisce che è stata già stata fatta una preparazione dei dati in modo efficace; quindi lo lascerei così.

Esercizio 4

```
Number of Leaves  :     38

Size of the tree :     52


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1148               84.7858 %
Incorrectly Classified Instances       206               15.2142 %
Kappa statistic                          0.6621
Mean absolute error                      0.1282
Root mean squared error                  0.2603
Relative absolute error                 42.9337 %
Root relative squared error             67.3947 %
Total Number of Instances             1354

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,933    0,118    0,950      0,933    0,941      0,805    0,966     0,986     unacc.
               0,741    0,119    0,634      0,741    0,683      0,590    0,914     0,686     acc.
               0,359    0,026    0,529      0,359    0,428      0,399    0,941     0,478     vgood_good
Weighted Avg.  0,848    0,111    0,849      0,848    0,846      0,727    0,953     0,882

=== Confusion Matrix ===

   a   b   c   <-- classified as
 893  60   4 |   a = unacc.
  47 218  29 |   b = acc.
   0  66  37 |   c = vgood_good
```

Considera i seguenti risultati. L'obiettivo è prevedere se un'auto riceverà una recensione molto buona, accettabile o inaccetabile; in funzione delle caratteristiche dell'auto stessa. Consideri questi risultati come positivi o negativi? Spiega perché. Infine, descrivi se ci sono operazioni che faresti dopo aver ottenuto questi risultati.

La precisione del modello è alta perché le istanze classificate correttamente sono 84,8% mentre quelle classificate erroneamente sono solo il 15,2%. L'algoritmo ha una precisione e recall alti per la classificazione della classe di recensioni "inaccettabile"; abbastaza buoni per la classe "accettabili" mentre per la classe "molto buona" si fa fatica a trovare i veri positivi di questa classe (precisione e recall bassi pari a 0,43 e 0,36). Nel complesso l'AUC è maggiore di 0,9 in tutte le classi sono quindi distinte bene. Molto probabilmente la classe "molto buona" è minoritaria (abbiamo una class imbalance, la maggior parte dei casi appartiene alla classe "inaccettabile", il che spinge il modello a sottovalutare le altre classi). Per migliorare ulteriormente i risultati potremmo effettuare *undersampling* (rimuovere dati dalla classe più numerosa per bilanciare il dataset).

Esercizio 5

```
FATT_CLIENTE =

    -32.8032 * TRANSAZIONI_PER_CLIENTE +
      5.1552 * FATT_PER_TRANS +
     31.1331 * TOT_QUAN +
    -47.8107 * PROD_MEDIO_PER_TRANSAZIONE +
  -1258.0722 * CATEGORIA_PIU_ACQUISTATA=ABBIGLIAMENTO,SCARPE +
   3404.0641 * CATEGORIA_PIU_ACQUISTATA=SCARPE +
    344.133

Regression Analysis:

Variable                                      Coefficient     SE of Coef        t-Stat
TRANSAZIONI_PER_CLIENTE                         -32.8032         3.6816        -8.91
FATT_PER_TRANS                                    5.1552         0.6996         7.3691
TOT_QUAN                                         31.1331         0.5402        57.6278
PROD_MEDIO_PER_TRANSAZIONE                      -47.8107         9.7651        -4.8961
CATEGORIA_PIU_ACQUISTATA=ABBIGLIAMENTO,SCARPE  -1258.0722       772.1747        -1.6293
CATEGORIA_PIU_ACQUISTATA=SCARPE                 3404.0641       632.0477         5.3858
const                                            344.133        695.0747         0.4951

Degrees of freedom = 614
R^2 value = 0.9497
Adjusted R^2 = 0.94917
F-statistic = 1930.7311

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9671
Mean absolute error                   3014.4211
Root mean squared error               7532.2204
Relative absolute error                 21.0251 %
Root relative squared error             25.4987 %
Total Number of Instances              621
```

Considerando i seguenti risultati, ottenuti tramite regressione, spiega e giustifica perché puoi considerare questi risultati come “buoni” o “cattivi”. Perché?

I seguenti risultati sono buoni perchè l’R-squared è prossimo ad 1 e l’F-statistic è elevato (modello predittivo robusto). Il coefficiente di correlazione pari a 0,97 sugerisce una forte relazione lineare positiva. Una tra le sei variabili ha un t-statistic pari a -1,6; il che sugerisce che questa non è significativa nell’analisi (il modello potrebbe essere affetto da multicollinearità). Potrei rimuovere questa variabile collineare per ottenere una modello più robusto (a livello di singole relazioni).

Esercizio 6

Descrivi la matrice di confusione del caso in cui si vogliono classificare i clienti che possiedono un abbonamento Sky in base alla loro probabilità di cancellazione dell’abbonamento (che può essere “alta”, “media” o “bassa”). Nell’ipotesi di

suggerire uno sconto alto ai clienti con "media" probabilità ed uno sconto basso ai clenti con "bassa" probabilità di disiscrizione, per azzerare il rischio di perderli. Descrivi le ipotesi e i costi associati con ciascun caso della matrice stessa.

In questo caso abbiamo una matrice di confusione 3x3, in cui gli elementi sulla diagonale rappresentano le previsioni corrette del modello mentre quelli al di fuori della diagonale rappresentano gli errori di classificazione che si dividono in:

- Errori di sottostima (False Negative): il modelo prevede una probabilità di abbandono inferiore a quella reale;
- Errori di sovrastima (False Positive): il modello prevede una probabiltà di abbandono superiore a quella reale.

Per raggiungere il nostro obiettivo conviene *minimizzare i False Negative*.

Esercizio 7

Considera la seguente regressione dove Y rappresenta i ricavi previsti dall'azienda, X1 rappresenta l'ammontare di investimenti in marketing, X2 rappresenta il numero di aziende concorrent, X3 rappresenta l'ammontare di investimenti in ricerca e sviluppo:

Y = 1,3*X1 - 0,9*X2 + 3*X3 + 3,9

Come utilizzeresti questa equazione per finalità di business? Quali considerazioni puoi fare e che decisioni potresti prendere?

Dall'equazione possiamo ricavare informazioni importanti riguardo le decisioni strategiche ed operative dell'azienda. Notiamo che il coefficiente di X3 è più alto, il che suggerisce che dovremmo dare una proprità maggiore nella R&D. Il coefficiente di X1 è comunque positivo quindi dovremmo comunque mantenere una strategia di marcheting. Non possiamo agire direttamente sulle aziende ma il coefficiente negativo di X2 ci fà capire che dobbiamo stare attenti alla concorrenza e dovremmo adottare strategie per fidelizzare i clienti e differenziare i prodotti/servizi.

Esercizio 8

Ipotizza di avere un franchising composto da 5 punti vendita e di voler profilare il comportament dei clienti di ciascuno di questi per poi impostare una strategia commerciale opportuna per far crescere tutti i punti vendita. Ipotizza di possedere una serie di informazioni per ciascun cliente (età, genere, spesa, modalità di pagamento, etc.) e in più una variabile target che indica in quale punto vendità acquista (1, 2, etc.). Lanciando un classificatore quale sarebbe il primo risultato (non tutti) che controlleresti e perché?

L'Accuracy è il primo indicatore da verificare perché è la perecntuale di istanze correttamente classificate dal modello, quindi se l'accuratezza è bassa il modello non è utile. Se dovessimo comparare diversi modelli, l'accuratezza fornisce una metrica immediata per un confronto preliminare.

Esercizio 9

Immagina di avere un database storico delle transazioni bancarie e che, per ogni bonifico, conosci se è stato fraudolento o meno. Quale sarebbe la matrice di confusione nel caso in cui volessi prevedere gli utenti che effettuano un bonifico fraudolento? Fornisci anche indicazione in merito ai costi associati a ciascun errore della matrice così definita.

Assegnando i seguenti valori alla variabile target:

- 1: bonifico fraudolento;
- 0: bonifico non fraudolento.

Otteniamo una matrice di confusione contenente i seguenti elementi:

- True Positive (v): il bonifico è fraudolento ed il modello lo predice correttamente;
- True Negative (p): il bonifico non è fraudolento ed il modello lo predice correttamente;
- False Positive (q): l bonifico non è fraudolento ma il modello lo predice erroneamente come fraudolento;
- False Negative (u): il bonifico è fraudolento ma il modello lo predice erroneamente come non fraudolento.

L'obiettivo è *minimizzare i False Negative* anche a costo di aumentare i False Positive.

Esercizio 10

Immagina di avere il seguente database contenente una serie di informazioni relative ad una attività di promozione di un prodotto e alla succesiva eventuale accettazione dell'offerta e spesa effettuata dall'utente. Immagina che ti venfa richiesto di profilare i clienti che accettano l'offerta. Descrivi tutte le attività che condurresti per arrivare ad un risultato utile per rispondere all'obiettivo posto e quali performance controlleresti per verificare che il risultato sia positivo.

User ID	Numero di chiamate promozionali fatte	Tempo medio trascorso al telefono con l'operatore	Sconto proposto (%)	ID Categoria di prodotto proposta	Prezzo del prodotto proposto	Accettazione dell'offerta	Spesa dell'utente
1	5	8	20	1		SI	8
2	6		24	4		NO	0
3		5	6	2	4	SI	16
4	10		30	1		NO	0
5	4		7	2		SI	8
6	5	3	0	4	16	NO	0
7			5	3	8	SI	24
8	8	4	23	1		NO	0
9	10		27	1	2	SI	10
10	4		4	1		NO	0
11	10	6	8	3		SI	8
12	8		13	1		NO	0
13	7	9	10	2		SI	20
14			30	4	16	NO	0
15	8		5	4		SI	32
16	2		27	3	8	NO	0
17	5	10	16	1	2	SI	6
18	8		20	1		NO	0
19	1	4	22	2	4	SI	28
20	2		20	4		NO	0

Data Preparation: rimuovo le colonne non necessarie all'analisi, ad esempio il "tempo medio trascorso al telefono con l'operatore" che è troppo sparsa, così come il "prezzo del prodotto proposto". I valori mancanti per la colonna "numero di chiamate proporzionali fatte" possono essere sostituiti con i valore medio. La colonna "spesa utente" può essere eliminata perché segue lo stesso comportamento della variabile target e non aggiunge nuove informazioni.
Data Mining: Binary Classification Tree, perché la variabile target "accettazione offerta" può assumere solo due valori (SI e NO).
Valutazione: per valutare il modello dobbiamo verificare l'accuracy; la confusion matrix e la curva ROC.

Esercizio 12

Si consideri il caso di un supemercato elettronico che possiede un database come quello in tabella. L'azienda vuole raggiungere due obiettivi: definire un set di promozioni per incrementare la vendita dei suoi prodotti; studiare il comportamento dei consumatori e distinguere quelli leali da quelli non leali.
Descrivere tutte le attività di pre-process, incluse quelle necessarie per poter lanciare la metodologia di data mining scelta che andrebbero realizzate prima di avviare alcuna analsi per raggiungere il primo obiettivo ed per il secondo. Spiega quale metodologia di data mining useresi e perché per raggiungere il primo ed il secondo obiettivo.

Transaction ID	Customer ID	Product ID	Price	Quantity	Loyalty card
1	1	47	6	2	Y
1	1	22	2		Y
1	1	28	6	3	Y
2	2	3	5	1	N
2	2	34	6	3	N
2	2	12	1	5	N
3	3	1	3		Y
3	3	40	9	6	Y
4	1	39	10	4	
4	1	28	1	5	
5	4	34	8	4	N
5	4	33	2	5	N
6	5	6	9		Y
6	5	48	2	4	Y
7	2	41	3	6	N
7	2	39	7	1	N
8	3	27	4	5	
8	3	31	10	3	
8	3	10	1	6	

Data Preparation: possiamo ricavare alcuni valori mancanti nella colonna "loyalty card" dal "CustomerID", per gli altri posso riempire il database utilizzando il valore più frequente, mentre per "quantity" i valori mancanti possono essere ricavati dal "prezzo", "loyalty card" e "Product ID"; per i valori rimasti vuoti possiamo inserire la media dei valori.
Per il primo obiettivo possiamo operare una data reduction, andando a rimuovere tutte le colonne inutili (tutte tranne "Tansaction ID" e "Product ID").
Data Mining: Tramite l'algoritmo Apriori sviluppo una serie di regole associative (non banali,

che hanno confidenza alta e basso supporto) che mi permettano di scoprire quali sono i prodotti che vengono acquistati insieme [market basket analysis].

Per il secondo task considero come variabile target la colonna “loyalty card” e costruisco un dataset aggregato per cliente. Creo nuove variabili come il numero di transazioni, la spesa totale (sommo tutti i prezzi per ogni transazione fatta del cliente). Utilizzo una tecninca di data mining: Binary Classification Tree con due regole (il numero di transazioni e la spesa totale, le altre colonne dopo aver fatto data preparation le posso rimuovere). Se non ottengo un buon modello si può effettuare una standardizzazione delle variabili.

Esercizio 13

Considera di avere il seguente database. Nell’ipotesi che tu voglia profilare gli utenti in base al fatto che essi abbiano una spesa media mensile bassa, media oppure alta, descrivi nel dettaglio tutti gli step che seguiresti partendo dal database fino ad arrivare ad avere dei risultati utili.

Customer ID	Gender	Age	Year of birth	Education	Job	Average monthly expense	Average monthly number of access to the store
1	M	82	1938	Medium	Unemployed	32	0
2		83	1937	Advanced		216	0
3	F	61	1959	Medium	Medical Doctor	151	2
4				Advanced	Professor	188	3
5		52	1968	Advanced	Unemployed	168	0
6		43	1977	Medium		244	3
7	M	68	1952	Medium	Medical Doctor	248	3
8	F	29	1991	Advanced	Professor	294	5
9				Advanced	Engineer	147	0
10		29	1991	Advanced		204	4
11		21	1999	Medium	Professor	261	4
12	F	31	1989	Medium	Medical Doctor	170	0
13				Advanced	Professor	36	2
14		58	1962	Advanced		276	5
15		87	1933	Medium	Engineer	84	4
16	M	76	1944	Advanced	Medical Doctor	207	1
17		39	1981	Medium	Unemployed	176	3
18				Medium	Professor	132	5
19		56	1964	Advanced		94	0
20	M	58	1962	Medium	Unemployed	133	0

Data Preparation: Rimuovo la colonna “gender” perché è sparsa (troppi valori mancanti). Per quanto riguarda la colonna “age” posso riempire i valori mancanti con il valore medio. Posso rmuovere “year of birth” perchè ha correlazione inversa perfetta con “age”. Posso riempire i valori mancanti di “job” con il valore più ricorrente. Discretizzo “average monthly expense” che diventerà la variabile target. La discretizzo basandomi su delle soglie ad esempio:

- 0 - 50, bassa;
- 51 - 160 media;
- > 160 alta.

Rimozione della colonna “customer id” perché è nominale e non dà valore al metodo di classificazione.
Data Mining: Classificazione multiclass, tramite Albero.