Università degli Studi di Padova
Corso di Laurea Magistrale in *Informatica*
a.a. 2016/2017

Data Mining

Docente: Annamaria Guolo

Prova pratica del 21 settembre 2017

ISTRUZIONI: La durata della prova è di 2 ore e 30 minuti. Redigere una breve relazione che riassuma l'analisi dei dati svolta ed i risultati conseguiti e consegnare la stampa in versione cartacea completa di i) nome e cognome, ii) numero di matricola, iii) data. Relazioni senza queste tre informazioni non potranno essere corrette.

Di seguito si riporta la descrizione del file ed una breve traccia dell'analisi da effettuare. È ammesso l'uso del materiale relativo al corso (slides della teoria, dispense del laboratorio, appunti, ...), del libro di testo, ma non di internet.
Dataset `mutuo`: i dati si riferiscono alla concessione di finanziamento ai clienti di una banca, per i quali sono state raccolte le seguenti informazioni.

- `durata`: durata del finanziamento in mesi
- `storico`: il cliente ha restituito finanziamenti precedenti? (regolare/non regolare)
- `scopo`: scopo della richiesta di finanziamento (affari/casa/automobile/istruzione/altro)
- `ammontare`: ammontare della richiesta
- `conto`: il cliente ha un conto al risparmio? (Si/No)
- `genere`: maschio/femmina (M/F)
- `proprieta`: il cliente possiede delle proprietà? (immobiliari/altro/no)
- `eta`: età del cliente (in anni)
- `casa`: in affitto/proprietario
- `disoccupato`: il cliente è disoccupato? (Si/No)
- `persone`: quante persone a carico ha il cliente? (1/ più di 1)
- `straniero`: il cliente è straniero? (Si/No)
- `cliente`: il cliente restituisce il finanziamento oppure no? (Buon cliente=1, Cattivo cliente=0)

Lo scopo dell'analisi è valutare quali variabili siano associate alla probabilità che il cliente restituisca il finanziamento (sia un buon cliente). In particolare, rispondere alle seguenti domande: lo scopo della richiesta di finanziamento è associata ad un'eventuale inadempienza nella restituzione del prestito? Vi è una relazione tra un'eventuale inadempienza nella restituzione del prestito e la sua durata?

1. Si consideri il sottoinsieme di dati costituito dalle variabili `cliente`, `durata`, `storico`, `eta`, `scopo`. Costruire il modello che si ritiene più opportuno per gli scopi dell'analisi e per rispondere alle domande indicate. Riportare analisi grafiche dei dati, output e valutazione grafiche del modello / dei modelli che si ritengono adatti per una spiegazione dell'approccio scelto e dei risultati.

2. Considerare tutte le variabili del dataset. Procedere alla costruzione del modello che si ritiene più opportuno per gli scopi di analisi e per rispondere alle domande indicate. Riportare analisi grafiche dei dati, output e valutazione grafiche del modello / dei modelli che si ritengono adatti per una spiegazione dell'approccio scelto e dei risultati.

$*$ Precisazione: nel caso in cui si svolgano analisi che richiedano la definizione di un seed, specificare quale seed viene scelto.