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

Data Mining

Docente: Annamaria Guolo

Prova pratica del 30 giugno 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 `cuore`: i dati si riferiscono agli episodi di infarto di soggetti per i quali sono state raccolte informazioni cliniche e informazioni sullo stile di vita.

- `pressione`: valore della pressione sistolica
- `colesterolo`: valore del colesterolo
- `adiposita`: misura del grasso corporeo
- `familiarita`: precedenti episodi cardiovascolari in famiglia? Si/No
- `carattere`: misura del comportamento di tipo A, caratterizzato da impazienza, nervosismo, aggressività, ...
- `bmi`: indice di massa corporea
- `alcolici`: quantità di alcolici assunti
- `anni`: età del soggetto in anni compiuti
- `infarto`: indicatore dell'episodio di infarto 0 (no) / 1 (si)
- `tabacco`: livello di tabacco consumato; Alto/Basso

Lo scopo dell'analisi è valutare quali variabili siano associate alla probabilità di un infarto. In particolare, rispondere alle seguenti domande: la familiarità all'evento coronarico è un fattore di rischio per la probabilità di infarto? Il carattere di tipo A è associato (e, se si, come) alla probabilità di infarto?

1. Si consideri il sottoinsieme di dati costituito dalle variabili `infarto`, `familiarita`, `colesterolo`, `carattere`, `tabacco`. 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.