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

Data Mining

Docente: Annamaria Guolo

Prova pratica del 22 gennaio 2018

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 `inquinamento`: i dati si riferiscono ad uno studio sulla relazione tra la mortalità in 60 aree metropolitane degli Stati Uniti e l'inquinamento atmosferico. Altre informazioni ambientali e demografiche son state raccolte ai fini di evitare eventuali effetti confondenti.

- `mortalita`: indice di mortalità (morti annuali per 100000 persone)
- `precipitazioni`: precipitazioni medie annuali (in inches)
- `umidita`: percentuale di umidità relativa
- `temperatura.gennaio`: temperatura media di gennaio (in Farenheit)
- `temperatura.luglio`: temperatura media di luglio (in Farenheit)
- `over65`: percentuale di persone con più di 65 anni
- `abitazione`: persone per abitazione
- `istruzione`: numero mediano di anni di istruzione per persone con oltre 25 anni di età
- `comfort`: percentuale di abitazioni dotate di ogni comfort
- `densita`: densità di popolazione (persone per miglio quadrato)
- `impiegati`: percentuale di impiegati (colletti bianchi)
- `poverta`: percentuale di abitanti con reddito annuale minore di 3000 dollari
- `HC`: livello di idrocarburi
- `NOX`: livello dannoso di ossido di azoto?: Si (maggiore di 30 $\mu g/mc$), No (minore o uguale di 30 $\mu g/mc$)
- `SO2`: livello dannoso di anidride solforosa?: Si (maggiore di 125 $\mu g/mc$), No (minore o uguale di 125 $\mu g/mc$)

1. Si consideri il sottoinsieme di dati costituito dalle variabili `mortalita`, `precipitazioni`, `umidita`, `HC`, `NOX`, `SO2`. Costruire il modello che si ritiene più opportuno per gli scopi dell'analisi. 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. 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.