Fondamenti di Data Science e Machine Learning - Prof. G. Polese - Anno Accademico 2020/21

Prova Scritta (ore 9:00) 21/2/2022

First Name and Last Name: **University ID(Matricola):**

Esercizio 1 (punti 6 su 30)
Dati N documenti da cui vengono estratti 100 shingle.

a) Formulare 3 funzioni hash per simulare altrettante permutazioni dei 100 shingle, al fine di costruire una signature di 3 elementi, cercando di ridurre il numero di collisioni;

b) Supponendo di costruire signature di 20 elementi, suddivise in 5 bande da 4 elementi ciascuno, specificare la migliore soglia di similarità al di sopra della quale le coppie di documenti hanno un'elevata probabilità di essere selezionate per il confronto.

Esercizio 2 (punti 6 su 30)
Utilizzando il paradigma Map Reduce, scrivere le funzioni Map e Reduce di un programma che, letto in input un file contenente parole, emetta in output una lista di coppie in cui il primo elemento è un numero intero $k > 0$, mentre il secondo rappresenta il numero di parole del documento contenenti k caratteri.

Esercizio 3 (punti 8 su 30)
Dato il seguente training set di pazienti:

Soggetto	Range Colesterolo	Fumatore	Trigliceridi	Iperteso	Infarto
1	150-200	SI	60-100	NO	NO
2	250-300	NO	170-200	SI	SI
3	150-200	SI	100-170	NO	NO
4	200-250	NO	200-300	SI	SI
5	150-200	SI	170-200	NO	NO
6	250-300	SI	170-200	LIEVE	SI
7	200-250	NO	100-170	NO	NO
8	150-200	NO	60-100	SI	NO
9	200-250	SI	200-300	NO	SI
10	250-300	NO	200-300	LIEVE	SI

Considerando l'attributo *Infarto* come dipendente, costruire un albero di decisione usando l'indice Gini ed indicando in ogni nodo il valore dell'indice e la distribuzione dei campioni.

Esercizio 4 (punti 5 su 30)

Sulla base dell'albero di decisione costruito nell'esercizio 3, classificare i seguenti soggetti dicendo se sono o meno a rischio infarto:

(150-200, SI, 100-170, NO), (200-250, NO, 200-300, SI), (250-300, SI, 170-200, LIEVE)

Successivamente, classificarli con l'algoritmo KNN utilizzando k=3 e la funzione di similarità di Jaccard.

Esercizio 5 (punti 5 su 30)

Nella tabella dell'esercizio 3 dire da quale dei 4 attributi predittivi l'attributo dipendente dipende maggiormente, utilizzando come criterio il g3 error delle 4 RFD che rilassano sull'extent e che hanno un singolo attributo predittivo sull' LHS e l'attributo dipendente sull' RHS.