BIG DATA ANALYSIS
25/07/2023

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

Regole:

1. E' vietato comunicare con altri durante la prova. E' possibile consultare qualsiasi materiale tranne chatGPT.
2. Rispondere a tutte le domande nel notebook, indicando chiaramente a quale punto si sta facendo riferimento.
3. Al termine della prova, lo studente rinomina il notebook con il proprio nome e cognome e le manda via email al docente: francesco.guerra@unimore.it, oggetto: BDTA: 25-7-2023.
4. I risultati sono pubblicati entro il giorno 1/8/2023.

Note:

Durata della prova: 2 ore. Il file csv che si trova al link
bit.ly/2023BDTAS

Parte 0: Il Dataset

Il datataset (preso e modificato da kaggle -- https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023) contiene dati relativi a salari di persone che operano in ambito data science. Il separatore è il ";". L'obiettivo è quello di inferire il livello di esperienza (Experience Level).

Parte 1: Analisi (10 punti)

1. Quante sono le istanze contenute nel dataset? _____ Il dataset è completo (cioè per ogni istanza tutti i valori di ogni attributo sono sempre correttamente specificati – non esistono "missing values")? _____ Il dataset è bilanciato per quanto riguarda la classe da predire? _____ Il livello di conoscenza (Expertise Level) è un attributo significativo nel determinare il livello di esperienza? (punti 2).

2. Verificare se il salario medio varia rispetto alla dimensione dell'impresa ("Company Size"), sia nel complesso sia relativamente al livello di esperienza (punti 2). Quanti sono i data scientist che non risiedono nella nazione della impresa per cui lavorano (punti 1)?

3. Il salario ricevuto dai lavoratori (Salary in USD) è distribuito nello stesso modo nelle imprese di piccola, media e grande dimensione? Rappresentare con il/gli opportuni grafici il concetto. Il salario ha poi la stessa distribuzione all'interno dei livelli di esperienza? (punti 3)

4. Quali sono i 5 lavori (Job Title) più remunerativi? (punti 2).

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il valore di Experience Level sulla base degli attributi presenti nel dataset. Ricaricare il dataset originale, eliminare eventuali attributi inutili (giustificare la scelta), eliminare le istanze che eventualmente contengono valori nulli, rendere tutti gli attributi numerici utilizzando un ordinal encoder, e dividerlo in modo che 3/4 degli elementi siano contenuti in un nuovo dataset "train" e 1/4 nel dataset "test".

Allenare il train con il modello Decision Tree e valutare l'accuracy ottenuta calcolata sia sul dataset train sia sul dataset test. Confrontare i risultati ottenuti con quelli ottenuti con una predizione basata sul modello KNeighborsClassifier. Effettuare alcune considerazioni sui risultati ottenuti, tenendo in considerazione anche l'analisi della confusion matrix e la predizione effettuata da un dummy classifier (punti 3)

2. Trovare i parametri migliori del classificatore DecisionTree. Agire sui parametri criterion e min_samples_leaf. Verificare se l'accuratezza che si ottiene con la nuova configurazione supera quella con i parametri di default ottenuta al punto 1 (punti 4)

3. Creare una pipeline che a partire dal dataset numerico utilizzato nel punto 1 applichi
- il SimpleImputer per sostituire eventuali valori nulli (punti 1)
- divida in 10 bins i valori di Salary in USD (punti 2)
- applichi il DecisionTreeClassifier per effettuare la predizione (punti 1)

4. Estendere la pipeline del punto precedente aggiungendo a ogni feature una nuova feature che rappresenti il valore della feature normalizzato. Applicare il DecisionTreeClassifier per effettuare la predizione

5. Creare una pipeline che a partire dal dataset iniziale (dopo aver tolto le colonne rimosse al punto 1)
- usi il SimpleImputer per inserire i valori nulli (punti 0.5)
- trasformi in vettori booleani (OneHotEncoder, sparse_output=False) le colonne 'Job Title', 'Employment Type', 'Company Location', 'Employee Residence' (punti 2)
- trasformi in valori numerici le colonne 'Year', 'Company Size' (punti 1)
- Applichi lo standard scaler sulla colonna Salary in USD (punti 1)
- applichi il DecisionTreeClassifier per effettuare la predizione (punti 0.5)

6. E' possibile utilizzare un regressore linerare al posto del DecisionTree? In che modo? (punti 2).