NICOLÒ SIMONATO —— MATRICOLA 146202 —— 274856@STUDENTI.UNIMORE.IT

Census Income Data Set

https://archive.ics.uci.edu/ml/datasets/Census+Income

Machine Learning

Il progetto in analisi richiedeva di prevedere se una persona, date determinate caratteristiche, guadagnasse più o meno di 50k$/all'anno.

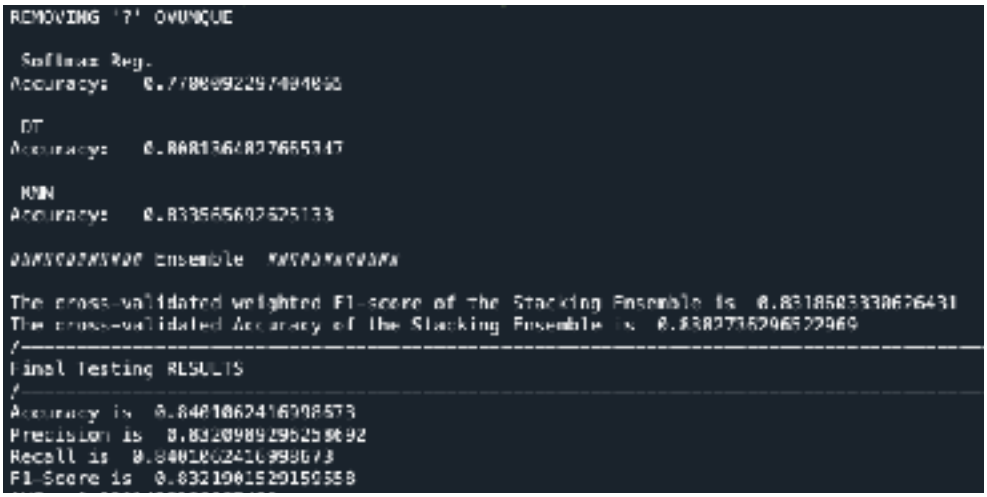
Il nostro Dataset, non richiedeva uno Split in Training e Testing, dato che tra i dati iniziali erano forniti due Dataset distinti.

Census Income Data Set

https://archive.ics.uci.edu/ml/datasets/Census+Income

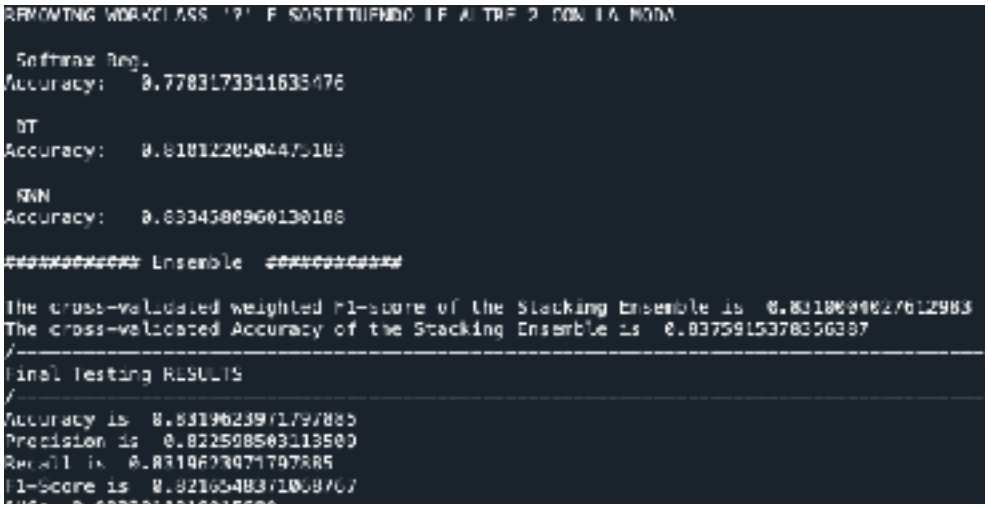
Nella prima fase di osservazione dei dati, ho subito iniziato un'occhiata al Dataset, cercando se vi fossero o meno valori mancanti, ho quindi subito effettuato un check, notando che vi fosse qualche dato mancante, contrassegnato con "?", ho quindi continuato ispezionando il Dataset alla ricerca di altre occorrenze, constatando che il carattere "?" fosse presente in solamente tre features, ovvero in : Occupation, Workclass e Native-Country.
Ho esplorato 4 opzioni:

1- rimuovere tutte le istanze con almeno un "?";

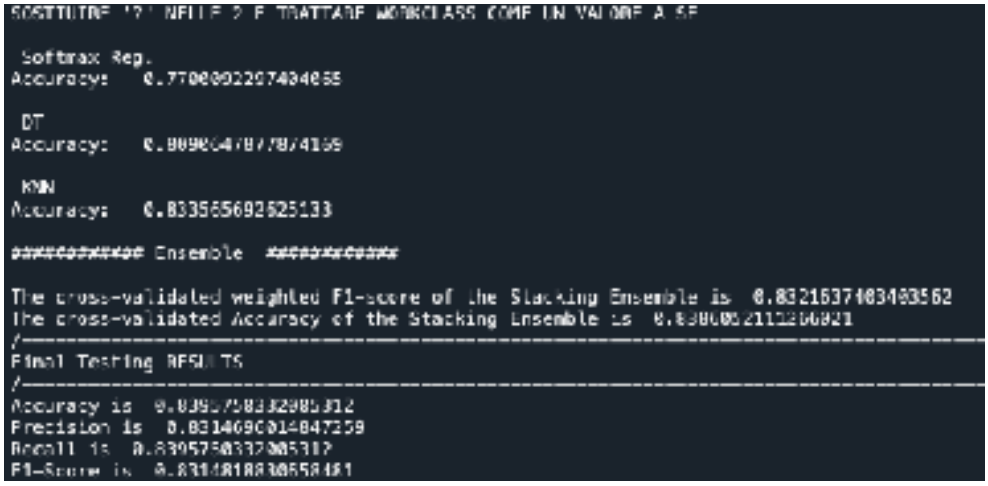


2- sostituire i valori "?" Con le mode di ogni feature;

3- sostituire le istanze ignote di Occupation e Native-Country con la moda e trattare "?" come una categoria di workclass



4- sostituire le istanze ignote di Occupation e Native-Country con la moda e rimuovere le istanze aventi "?" come valore per "Workclass"



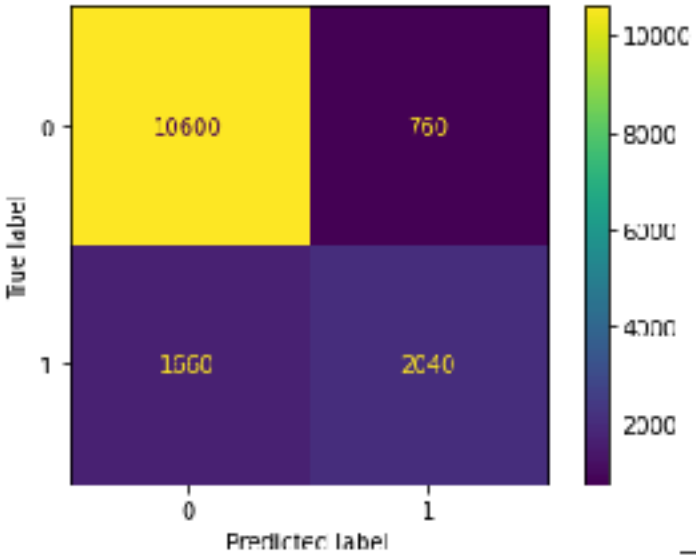
Ho optato per l'opzione 1) essendo l'opzione con le migliori metriche sul Dataset di Testing seppur di poco (circa 0.05% rispetto alla seconda miglior ipotesi).

Proseguendo nella nostra fase di manipolazione dei dati, ho proseguito fattorizzando e successivamente scalando l'intero Dataset, con lo scopo di ottenere un Dataset più uniforme e manipolabile tramite le diverse funzioni fornite dalle librerie come Pandas o sklearn.

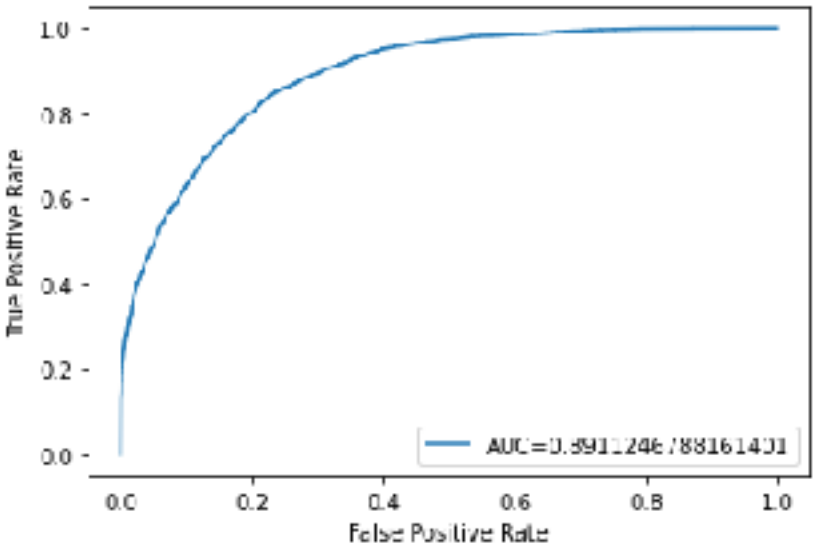
Subito dopo aver standardizzato, ho tentato di verificare se potessi ricavare un dataset migliore, cercando magari di estratte un subset di Feature tramite una Correlation Matrix, ma una volta visualizzata, ho potuto vedere che non vi fossero correlazioni forti tra tutte le Feature e nemmeno tra le diverse Feature verso la variabile target, ho dunque optato per lasciare tutte le Feature inizialmente presenti nel Dataset.

Arrivato a questo punto, ho scelto 3 modelli, i quali saranno poi combinati sia in fase di cross-validation, sia per effettuare un ensemble con lo scopo di ottenere un modello più performante essendo entrambi metodi che combinano i risultati ottenuti al fine di ottenere previsioni migliori.

Procedo dunque ad importare il dataset di Testing e reitero quanto fatto con il Dataset di Training, vado quindi ad eliminare le istanze aventi dati "?", fattorizzo e scalo l'intero Dataset.
Finalmente posso andare ad applicare il modello ottenuto in fase di Training al Dataset di Testing, una volta fatto procedo a visualizzare le diverse metriche di performance fornite dalla libreria sklearn.



La Rappresentazione della confusion Matrix (Nell'immagine a sinistra), utile per ottenere un altra metrica, ovvero la ROC da cui calcoliamo l'AUC (nell'immagine in basso).



I risultati ottenuti sul dataset di testing.