

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 "?";

```
REMOVING '?' UNIQUE

Software Reg.
Accuracy: 0.7788992297484850

DT
Accuracy: 0.8081367807655347

KNN
Accuracy: 0.833555602525133

Stacking Ensemble *****

The cross-validated weighted F1-score of the Stacking Ensemble is 0.831858330676431
The cross-validated Accuracy of the Stacking Ensemble is 0.8382735296922969
/-----/
Final Testing RESULTS
/-----/
Accuracy is 0.8481062416098573
Precision is 0.8320989296259692
Recall is 0.84018024216998673
F1-Score is 0.8321901529150558
```

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

```
REMOVING '?' & SOSTITUENDO CON LA MODA

Software Reg.
Accuracy: 0.7710145292489621

DT
Accuracy: 0.8245941909411963

KNN
Accuracy: 0.8375913341542282

Stacking Ensemble *****

The cross-validated weighted F1-score of the Stacking Ensemble is 0.8357686164413355
The cross-validated Accuracy of the Stacking Ensemble is 0.8422962381794719
/-----/
Final Testing RESULTS
/-----/
Accuracy is 0.8369254787175235
Precision is 0.8275178252235789
Recall is 0.8369264787175235
F1-Score is 0.8281587728630589
```

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

```
REMOVING WORKCLASS '?' E SOSTITUIENDO IL VALORE 2 CON LA MODA

Softmax Reg.
Accuracy: 0.7783173311635476

DT
Accuracy: 0.8181228504475183

KNN
Accuracy: 0.8334580960130188

##### Ensemble #####

The cross-validated weighted F1-score of the Stacking Ensemble is 0.8318080027612983
The cross-validated Accuracy of the Stacking Ensemble is 0.8375915378356397
/-----/
Final Testing RESULTS
/-----/
Accuracy is 0.8319623971797883
Precision is 0.822598503113500
Recall is 0.8319673071797885
F1-Score is 0.8216548371068707
```

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

```
SOSTITUIRE '?' NEI F 2 E TRATTARE WORKCLASS COME UN VALORE A SE

Softmax Reg.
Accuracy: 0.7700002207404055

DT
Accuracy: 0.8098047877874169

KNN
Accuracy: 0.833555692525133

##### Ensemble #####

The cross-validated weighted F1-score of the Stacking Ensemble is 0.8321537403403562
The cross-validated Accuracy of the Stacking Ensemble is 0.8380052112360821
/-----/
Final Testing RESULTS
/-----/
Accuracy is 0.8395758332985312
Precision is 0.8314090014847259
Recall is 0.8395758332985312
F1-Score is 0.8314818830558451
```

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).

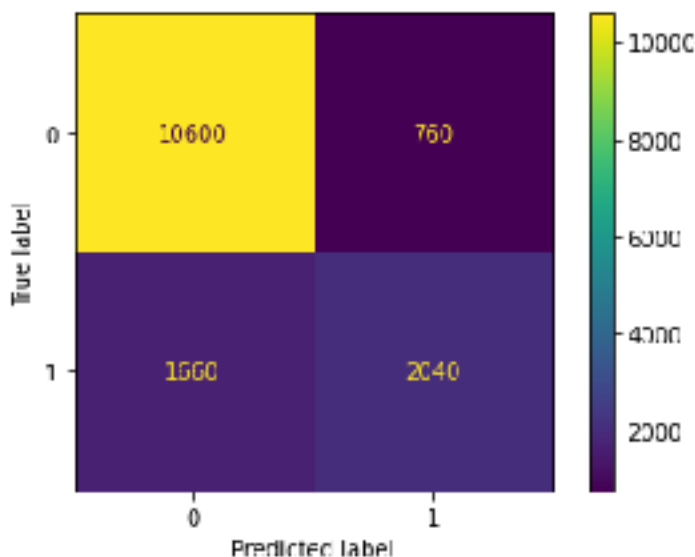
Proseguendo nella nostra fase di manipolazione dei dati, ho proseguito fattorizzando e successivamente scalando l'intero Dataset, con lo scopo di ottenere un Dataset più uniforme e manipolabile tramite le diverse funzioni fornite dalle librerie come Pandas o sklearn.

Subito dopo aver standardizzato, ho tentato di verificare se potessi ricavare un dataset migliore, cercando magari di estrarre 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.

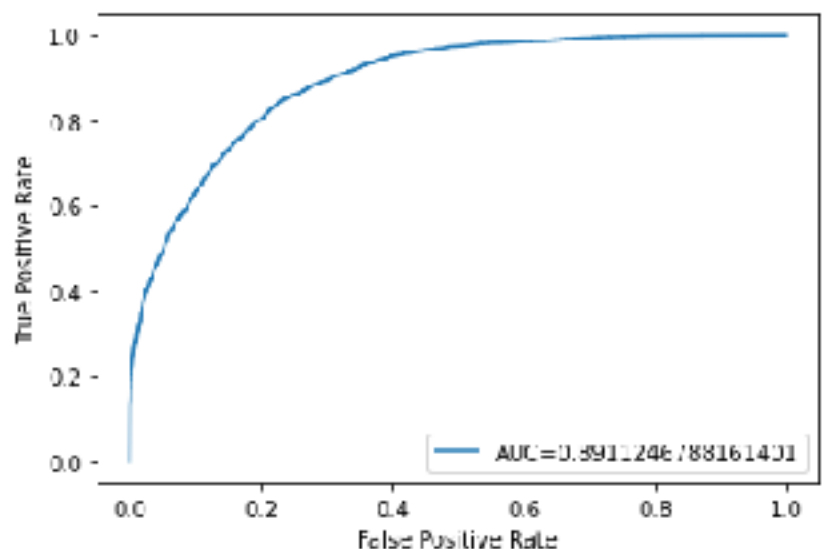
Arrivato a questo punto, ho scelto 3 modelli, i quali saranno poi combinati sia in fase di cross-validation, sia per effettuare un ensemble con lo scopo di ottenere un modello più performante essendo entrambi metodi che combinano i risultati ottenuti al fine di ottenere previsioni migliori.

Procedo dunque ad importare il dataset di Testing e reitro 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.

```
Final Testing RESULTS|  
/-----  
Accuracy is  0.8393094289508632  
Precision is  0.8311802119572669  
Recall is    0.8393094289508632  
F1-Score is  0.8312461178106133  
AUC:  0.8911246788161401
```