Challenge 1 ML

Matteo Vicenzino

February 2025

Dataset e trattamento dei dati

Nel seguente report, il dataset **Banknote Authentication** è analizzato utilizzando diversi algoritmi di Supervised e Unsupervised Learning. Vengono confrontati i risultati ottenuti e viene valutato il modello più efficiente per classificare le immagini delle banconote in due diverse classi.

Il dataset presenta 4 caratteristiche relative alla Wavelet Transformation dell'immagine: varianza, skewness, curtosis ed entropy. Tutte queste sono variabili quantitative che sono state standardizzate per scalare le caratteristiche e renderle confrontabili. Questo è necessario per alcuni algoritmi utilizzati successivamente, come ad esempio PCA.

Inoltre, si nota che il dataset è ordinato secondo la variabile *class*. Si decide quindi di riordinare i dati in maniera casuale attraverso la funzione `shuffle` di sklearn.

Questo per evitare sia che in una futura divisione tra train e test i dati con lo stesso valore di class siano tutti compresi nel test set, sia che eventuali suddivisioni risultino sbilanciate.

Realizzando il plot della matrice di correlazione, si nota che le coppie di variabili maggiormente correlate negativamente sono *Skewness* con *Curtosis* (-0.79) e *Class* con *Variance* (-0.72).

Inoltre, dallo scatterplot che rappresenta i punti del dataset con le tre variabili più significative sugli assi (Variance, Skewness e Curtosis), si nota che le due classi di banconote sono collocate su due iperpiani distinti e facilmente distinguibili.

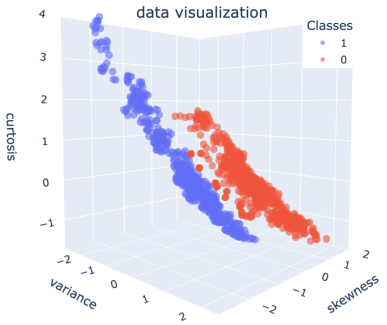

Figure 1: Visualizzazione del dataset in 3 dimensioni

Unsupervised Learning

Si inizia ad analizzare il dataset utilizzando tecniche di apprendimento non supervisionato, partendo dalla riduzione di dimensionalità tramite **PCA**. Si visualizzano i risultati in uno spazio delle componenti principali, con le osservazioni colorate in base alla classe.

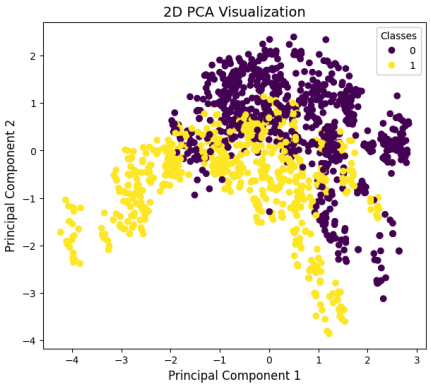

Figure 2: Prime 2 componenti principali, in relazione alle true labels

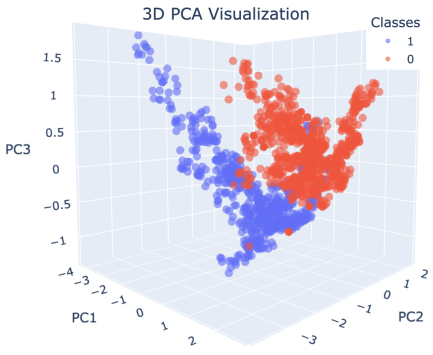

Figure 3: Prime 3 componenti principali, in relazione alle true labels

Si nota che nel grafico con 2 componenti principali (Figura 2) le classi non sono facilmente distinguibili in modo lineare, mentre in quello con 3 componenti (Figura 3) la separazione migliora notevolmente, ma non raggiunge mai il livello di chiarezza descritto in Figura 1.

Inizialmente si applica quindi l'algoritmo **k-means** con k = 2 sulle prime 2 e 3 proiezioni create da PCA, ottenendo uno scarso risultato e un'accuratezza di solo il 55.8%.

Applicando invece k-Means sulle prime 2 variabili del dataset: variance e skewness (le maggiormente correlate con la class) si riesce a ottenere dei risultati migliori e un'accuratezza superiore dell'87.8%, concorde con quanto osservato confrontando il dataset con PCA.

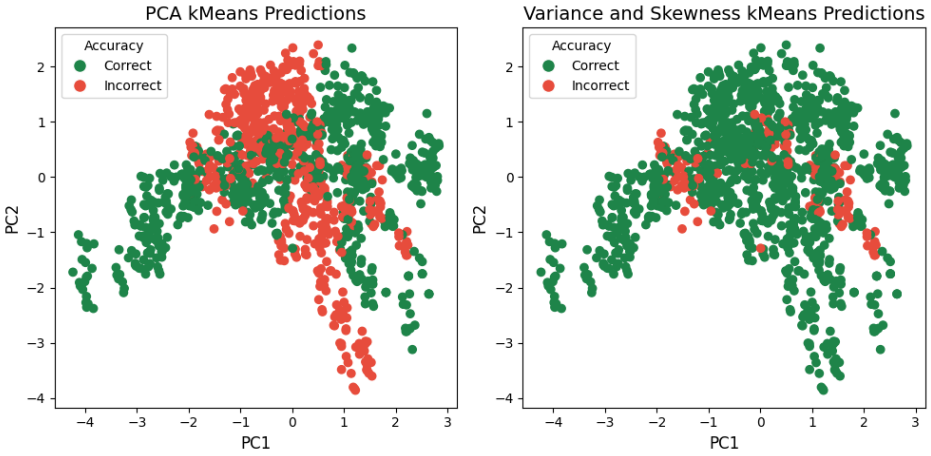

Figure 4: a sinistra l'accuratezza della cassificazione kMeans con le prime 2 componenti PCA, a destra l'accuratezza della cassificazione kMeans sulle variabili *Skewness* e *Variance* del dataset

Si continua con l'algoritmo **t-SNE** utilizzando 2 componenti e una perplessità di 70; successivamente, si applica t-SNE anche con 3 componenti per confrontare i risultati.

Si nota che in entrambi i casi si ottiene un miglioramento rispetto a PCA, poiché le due classi di banconote sono linearmente separabili e presentano un confine decisionale meglio definito. Questo anche grazie alla capacità di t-SNE di analizzare dataset non lineari.

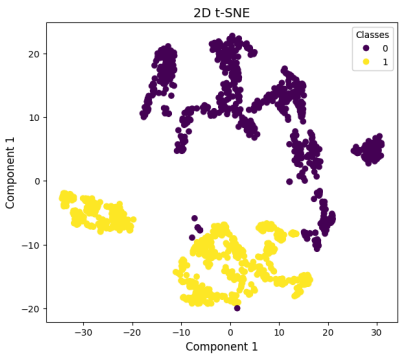

Figure 5: Prime 2 componenti di t-SNE, in relazione alle *true labels*

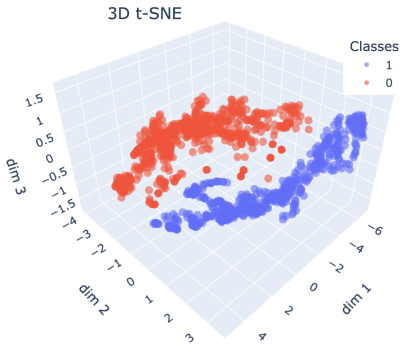

Figure 6: Prime 3 componenti di t-SNE, in relazione alle *true labels*

Dopo aver eseguito con diversi parametri il test del kNN distance, si è scelto un K = 13 e una dimensione del raggio pari a 0.532 per l'algoritmo di clustering **DBSCAN**. Dai risultati ottenuti, si distinguono 3 cluster che coincidono quasi perfettamente con la variabile target. Ci sono alcuni punti di noise. In particolare, la classe 0 è riconosciuta perfettamente dal cluster 1, mentre la classe 1 è inserita nei cluster 0 e 2. La maggior parte dei punti classificati in modo

incorretto non è stata assegnata a una classe errata, bensì è stata inserita nei punti di noise (Figura 7).

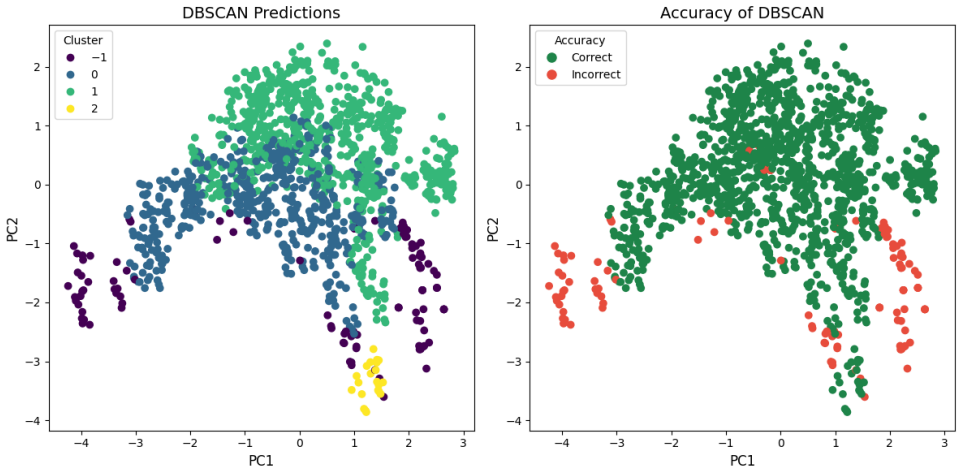

Figure 7: Cassificazione con DBSCAN. sugli assi le prime 2 componenti principali di PCA

Supervised Learning

Prima di applicare le tecniche di Supervised Learning, il dataset viene suddiviso in train e test set come indicato. Successivamente, si utilizza la tecnica del K-fold Cross Validation, scegliendo un valore di K pari a 5 per ottenere un buon trade-off fra la correlazione tra i modelli e il rapporto train / validation set.

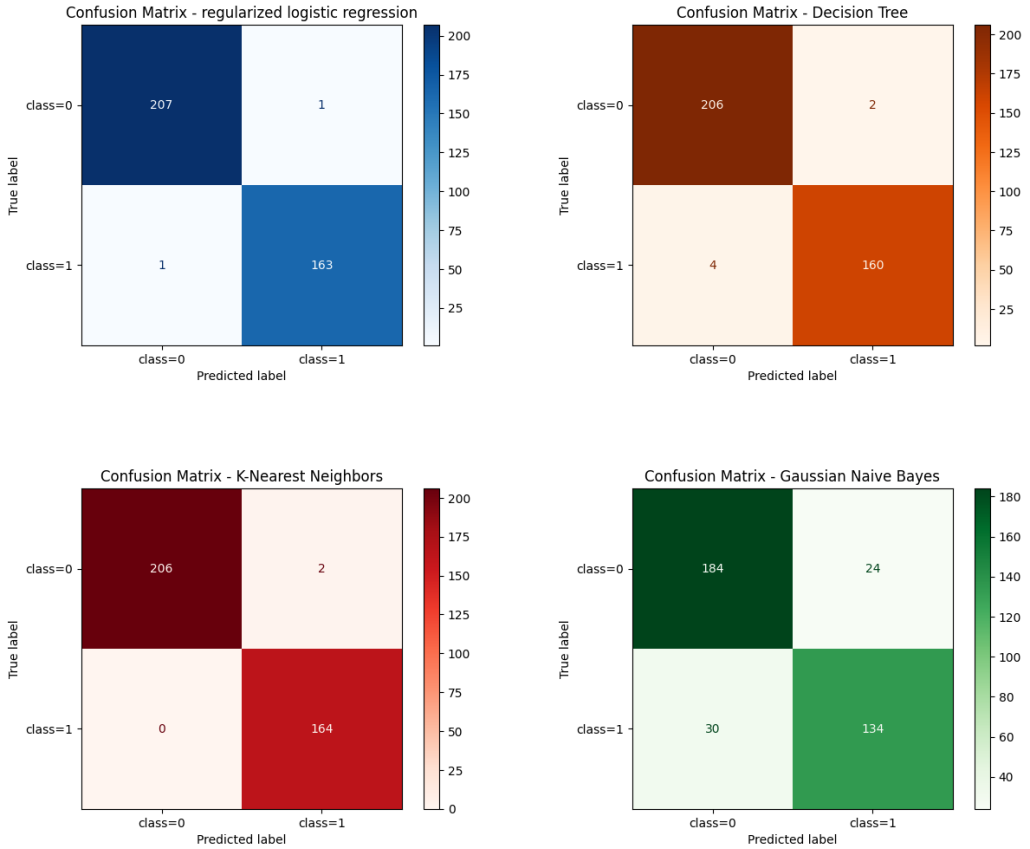

Figure 8: Confusion Matrix dei diversi modelli di apprendimante supervisionato

Dapprima si applica una **Regressione Logistica** senza regolarizzazione, e si nota che il

modello performa già ottimamente in termini di accuratezza e F1-score. Si decide di usare grid search in combinazione con i k-fold creati in precedenza per trovare il miglior tipo e parametri di regolarizzazione. Tuttavia, anche con la Lasso e un parametro $\lambda = 10$ l'accuratezza e lo score F1 rimangono pari a 99.4%.

Si procede valutando l'algoritmo **ID3** con la metrica di entropia rispetto a quella di Gini: ciò serve a creare un albero più bilanciato e, dopo aver testato le due metriche tramite la validazione incrociata k-Fold, si ottiene un'accuratezza migliore. Nonostante i suoi punti di forza, l'algoritmo ID3 presenta prestazioni leggermente inferiori rispetto alla regressione logistica, classificando erroneamente 6 punti del test set. Infatti, l'accuratezza scende al 98.2% e lo score F1 al 97.6%.

Successivamente, si analizza il dataset utilizzando l'algoritmo **Gaussian Naive Bayes**, il quale etichetta erroneamente 54 punti del dataset. Questo modello presenta un'accuratezza e un punteggio F1 rispettivamente di 83,2% e 81,7%, probabilmente per la bassa correlazione presente tra la maggior parte delle variabili.

Infine si utilizza il metodo **K-Nearest Neighbors** per classificare un campione in base alla maggioranza delle etichette dei suoi K vicini più prossimi nel dataset. La scelta di K è cruciale. Infatti, un valore di K troppo elevato porterà a includere nella `neighbourhood` punti di altre classi, diminuendo l'accuratezza del modello. Al contrario, un K troppo basso porterà a un aumento della varianza e renderà il modello troppo sensibile al rumore. i presenta il grafico dell'accuratezza per diversi valori di k (Figura 9). Si sceglie quindi un valore di K = 7.

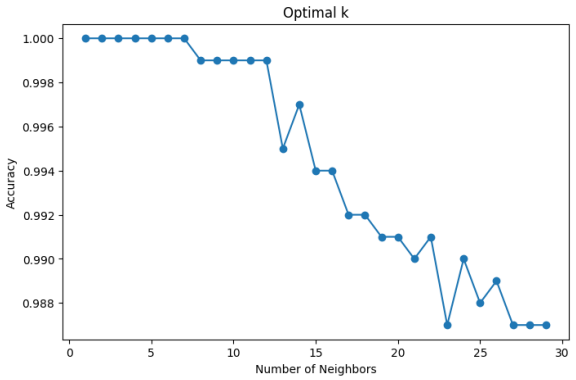

Figure 9: Dal grafico si evince che per valori di K superiori a 12 l'accuratezza cala notevolmente

Performance dei modelli e conclusioni

Stampando il grafico della curva ROC e dei punteggi di accuratezza, precisione, recall e F1, si conclude che i modelli di apprendimento supervisionato presentano performance elevate. Tuttavia, Gaussian Naive Bayes ha un valore di Area Under the Curve di circa 94%.

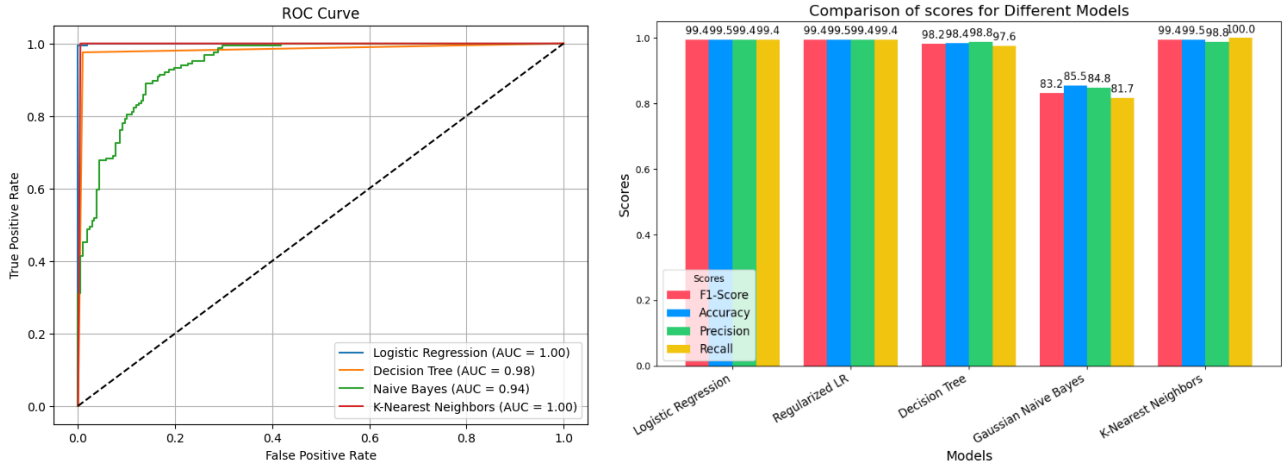

Figure 10: curva ROC e punteggi dei modelli di apprendimanto supervisionato studiati