Bortolussi Lorenzo SM3201370 18/11/2025

CHALLENGE 1 – REPORT

Il presente report contiene un resoconto di una analisi nell'ambito dell'autenticazione di banconote.

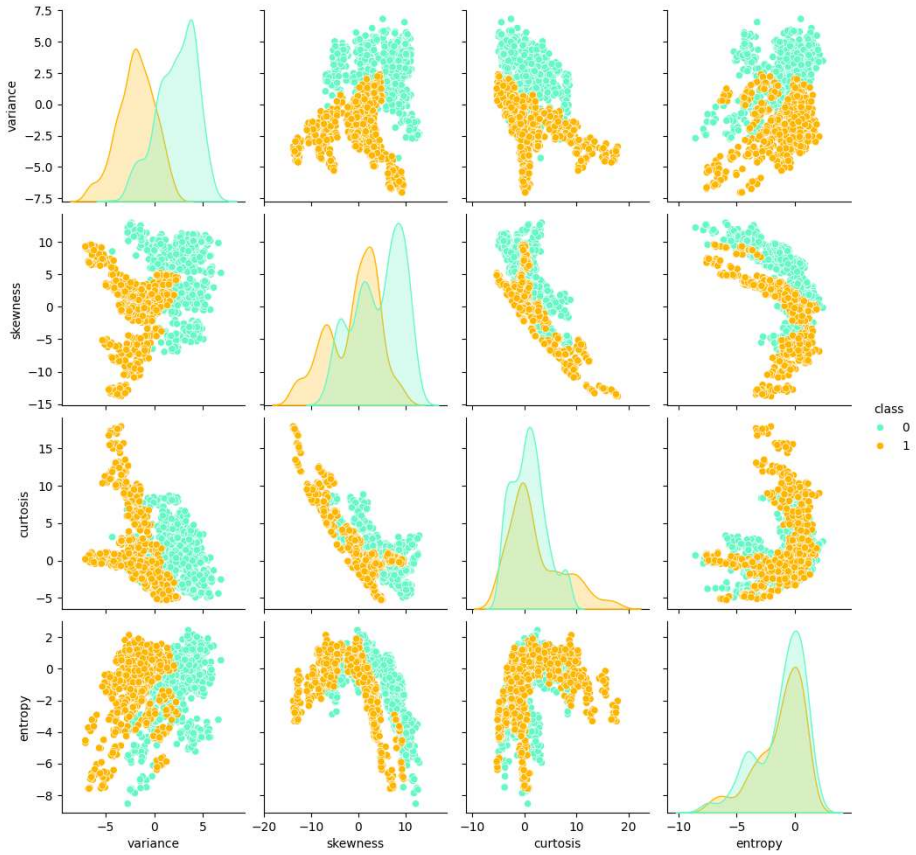
Vengono forniti dati relativi a immagini di banconote reali e contraffatte, dati utilizzati in tre parti principali del progetto:

1. Caricamento e pulizia dei dati forniti;
2. Esplorazione dei dati tramite tecniche di apprendimento non supervisionato;
3. Costruzione di modelli di apprendimento supervisionato.

Preparazione dei dati

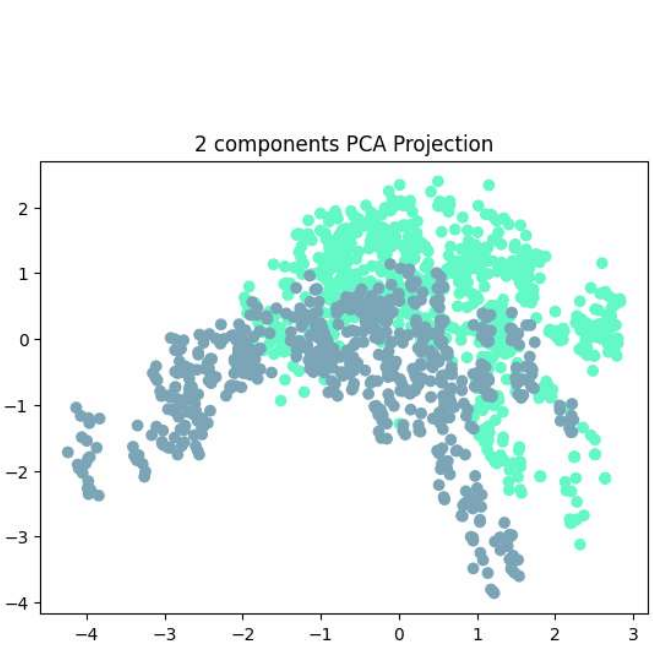
Il dataset contiene informazioni su 1372 banconote, ognuna descritta da quattro caratteristiche e dalla classe di appartenenza. Non sono presenti valori nulli, quindi è stato sufficiente convertire i dati da pandas dataframe ad array di numpy, separando le quattro feature dalla classe conosciuta.

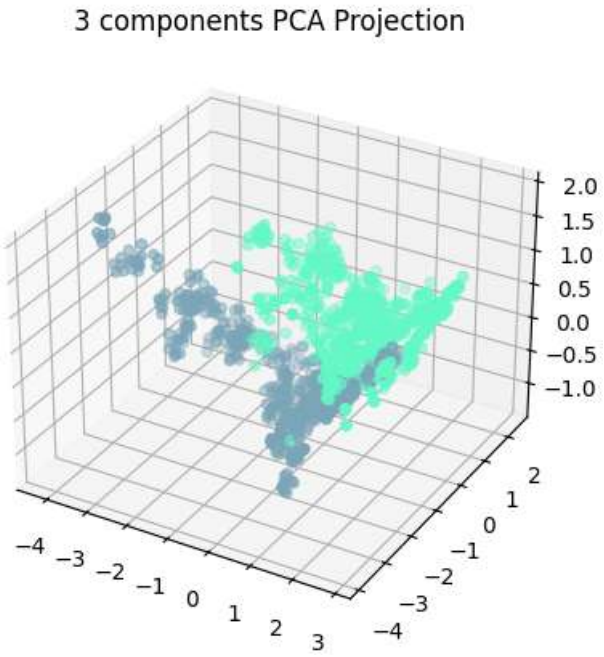
Le feature, tutte contenenti valori numerici continui, hanno ordini di grandezza diversi fra di loro, quindi si decide di standardizzare i valori all'interno delle classi. Di seguito viene riportata la rappresentazione dei dati così ottenuti:



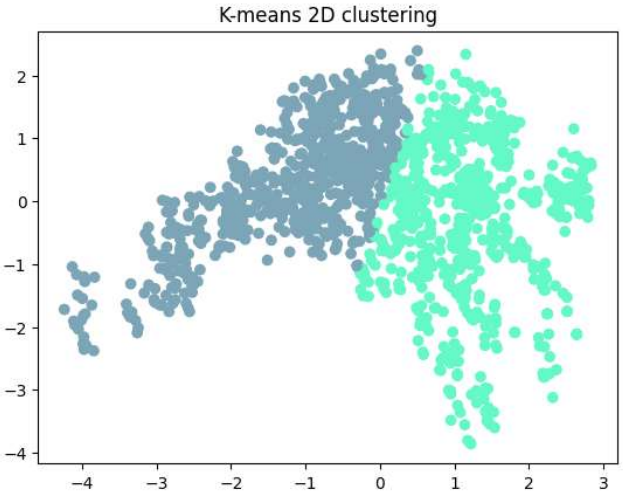
Apprendimento non supervisionato

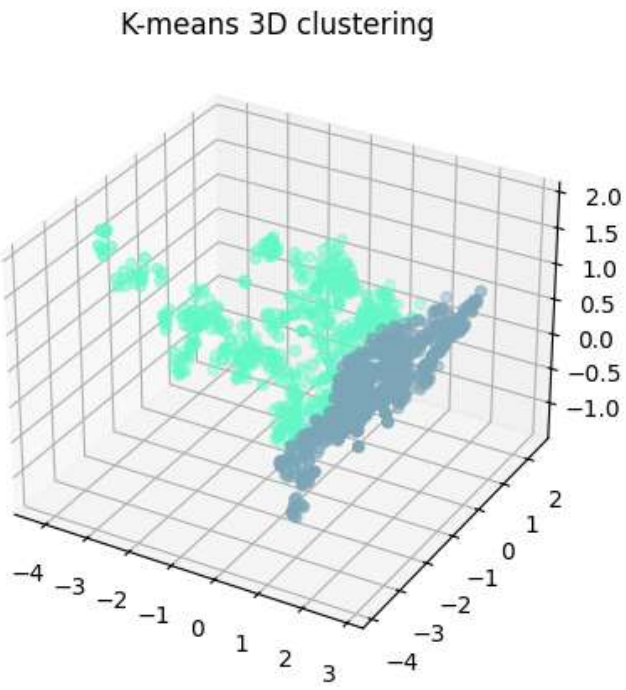
Utilizzando **PCA** con I primi due componenti principali, si nota come le due classi sembrano avere la stessa forma e sono leggermente sfasate, ma non sembrano separabili in maniera lineare. Aggiungendo una componente principale, i due cluster sono essere meglio identificabili, con una perdita della varianza dei dati originali minore del 5%.



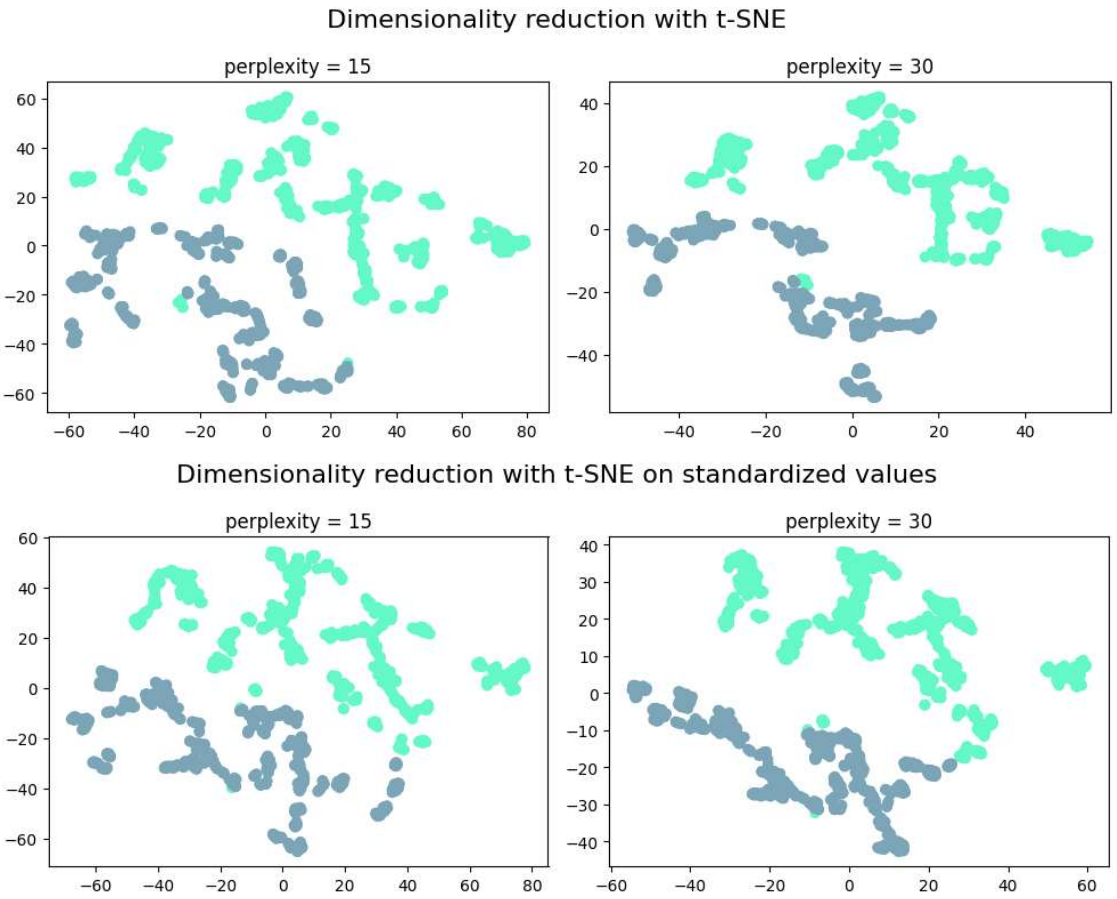


Utilizzare **k-means** Non produce risultati soddisfacenti, dato che i dati non sembrano avere una struttura sferoidale, per la quale avrebbe potuto funzionare questo strumento.

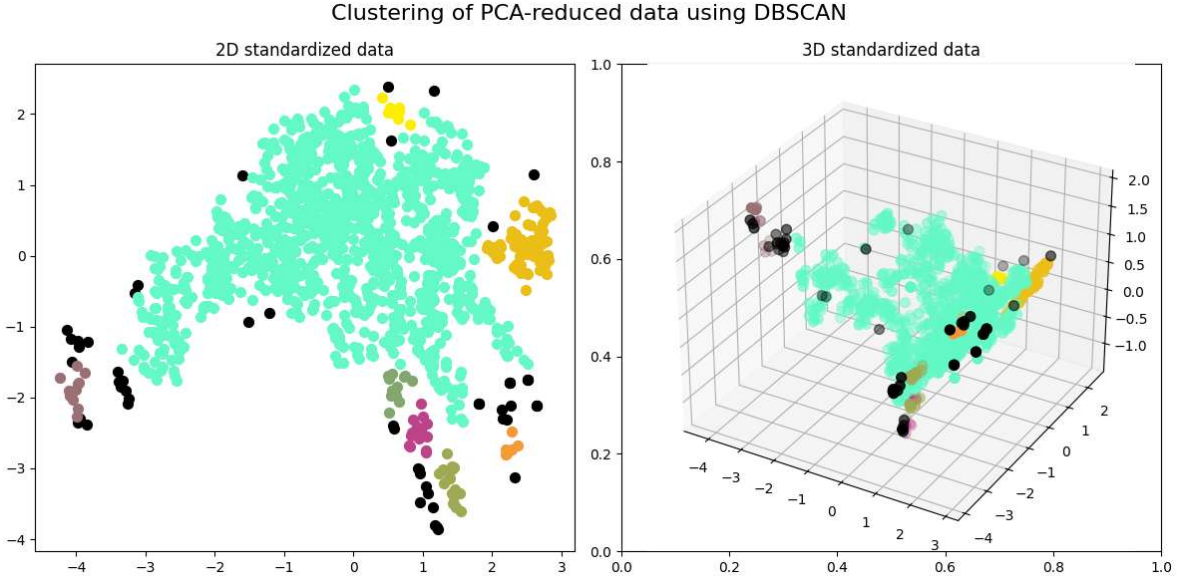


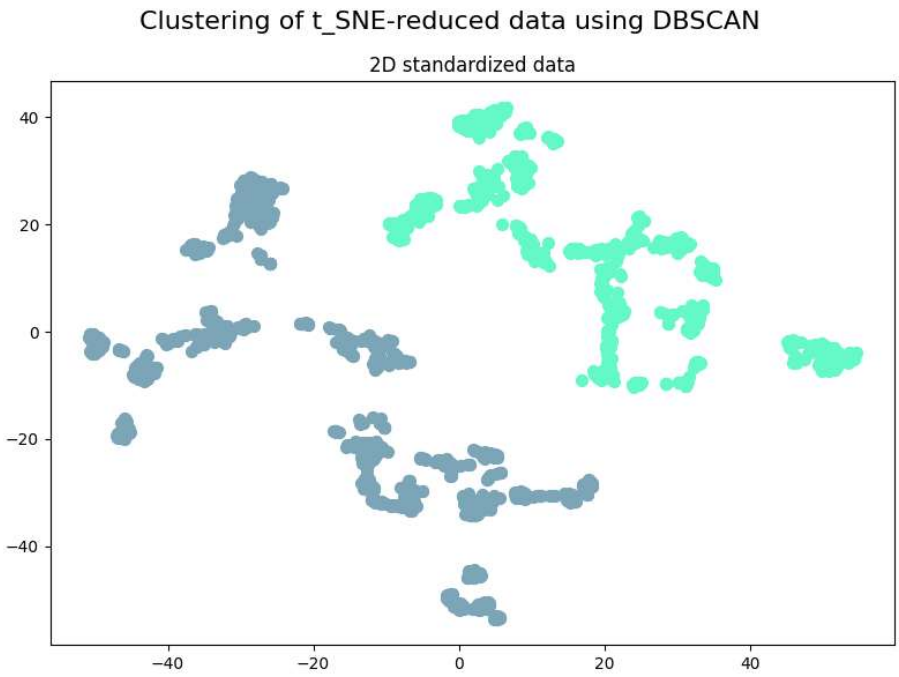


In parallelo a PCA, è stato utilizzato **t_SNE** per proiettare i dati in basse dimensioni prima utilizzando i dati originali e poi quelli standardizzati, ottenendo una buona separazione delle due classi, con dei risultati leggermente migliori sui dati scalati:



E' stato poi utilizzato **DBSCAN** per provare a ricreare le due classi dai dati originali, partendo dai dati standardizzati e abbassati di dimensionalità sia da PCA sia da t-SNE. Come si poteva immaginare i risultati da clustering sui dati di PCA sono molto lontani dalle due classi originali, sia con due sia con tre componenti principali. DBSCAN applicato dopo t-SNE invece ha dei risultati praticamente identici alle classi originali.





Apprendimento supervisionato

Per questa sezione, sempre utilizzando la libreria sklearn, si ha separato i dati scalati in train e test set, per poi allenare e testare i seguenti modelli per la classificaizone binaria: Regressione Logistica, Albero Decisionale (ID3), Naive Bayes, k-NN. Di seguito sono riportate le performance misurate:

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Log. Regression	0.994624	0.987179	1.0	0.993548
Log. Regression L1	0.994624	0.987179	1.0	0.993548
Log. Regression L2	0.986559	0.968553	1.0	0.984026
Log. Regression elastic net	0.986559	0.968553	1.0	0.984026
ID3	0.991935	0 .980892	1.0	0.990354
Naive Bayes	0.852151	0.786127	0.883117	0.831804
K-nn	0.994624	0.987179	1.0	0.993548

La regressione logistica è stata studiata anche con regolarizzazioni diverse, ma il risultato di base è variato poco, essendo già prossimo a predizioni perfette nel test set. Il parametro di k-nn è stato selezionato facendo una ricerca su griglia unita a cross validation, ma anche in questo caso i risultati del modello non sono variati molto dalle predizioni quasi perfette. L'unico modello che ha performato significativamente peggio rispetto agli altri è quello basato su naive bayes, nonostante la ricerca di una distribuzione a priori tramite grid search, forse perché l'assunzione di scorrelazione fra le categorie non è molto realistica in questo caso.