Machine Learning Report Challenge 1

Elisa Coceani

4 Marzo 2024

1 Introduzione

Lo scopo della seguente challenge è di applicare diverse tecniche di unsupervised e supervised learning ai dati contenuti nel file `data_banknote_authentication.csv`. I dati da analizzare rappresentano alcune caratteristiche di immagini scattate da campioni simili a banconote genuine e contraffatte. Queste caratteristiche sono: varianza, asimmetria, curtosi, entropia e se l'immagine ritrae una banconota contraffatta o meno.

2 Unsupervised Learning

Utilizziamo PCA (Principal Component Analysis) e plottiamo le prime due componenti.

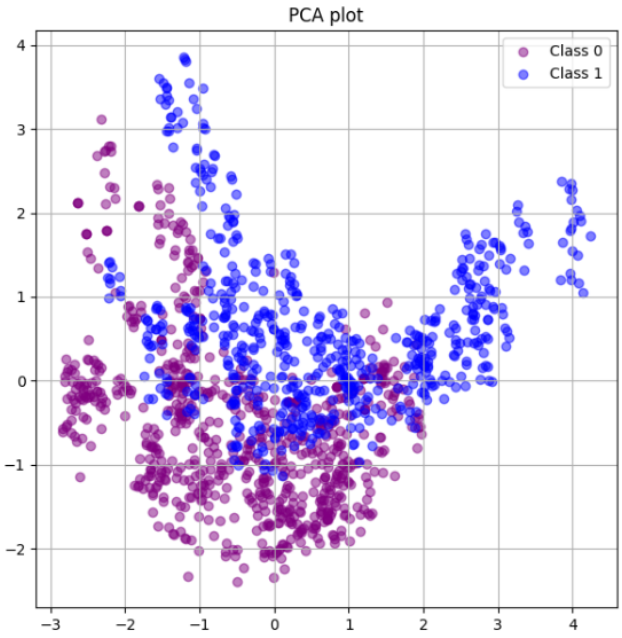

Figura 1: Grafico PCA

La sovrapposizione delle classi indica che i punti dei due gruppi sono distribuiti in modo simile nello spazio delle componenti principali. Questo suggerisce che può essere difficile separare le classi utilizzando solo le componenti principali identificate da PCA.

Proviamo ad applicare K-Means per trovare cluster distinti per le due classi.

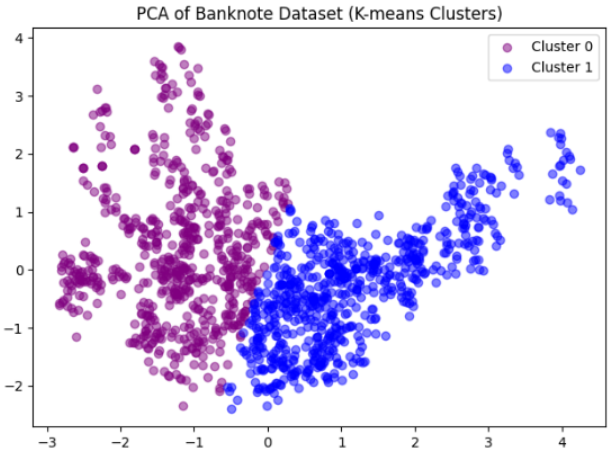

Figura 2: Grafico PCA con K-Means

PCA con K-Means individua due cluster che non corrispondono alle due classi presenti nel dataset. Questo potrebbe indicare che la struttura dei dati non è stata adeguatamente catturata da questa combinazione di algoritmi.

Infine, utilizziamo t-SNE per visualizzare dati di grandi dimensioni in uno spazio a due e tre dimensioni.

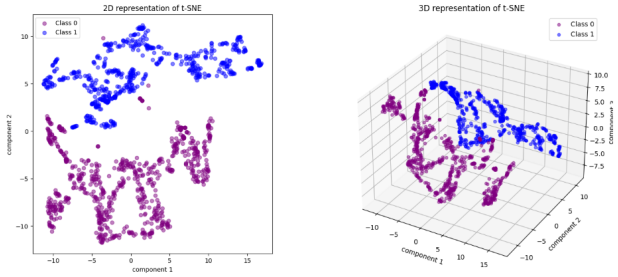

Figura 3: Rappresentazione 2D e 3D con t-SNE

Idealmente, i punti dati appartenenti a classi diverse dovrebbero essere visivamente separati nel grafico t-SNE. Una separazione chiara indica che t-SNE ha catturato efficacemente la struttura sottostante dei dati e che esistono gruppi distinti, più o meno ciò che vediamo nel grafico tridimensionale.

La visualizzazione tridimensionale del clustering utilizzando l'algoritmo DBSCAN mette in luce la sua capacità di identificare cluster densi all'interno del dataset delle banconote, oltre a individuare i punti di "noise" che non fanno parte di alcun cluster definito (rappresentati con il colore nero). DBSCAN, configurato tramite i parametri di densità, organizza i punti in base alla loro vicinanza spaziale, rivelando così gruppi distinti di dati e isolando outlier che non si adattano a nessun cluster predefinito. Si nota che per questo dataset DBSCAN non è efficiente, in quanto non otteniamo dei cluster nettamente separati. Idealmente vorremmo dei gruppi di punti separati che identificano le diverse classi, divisi da regioni vuote.

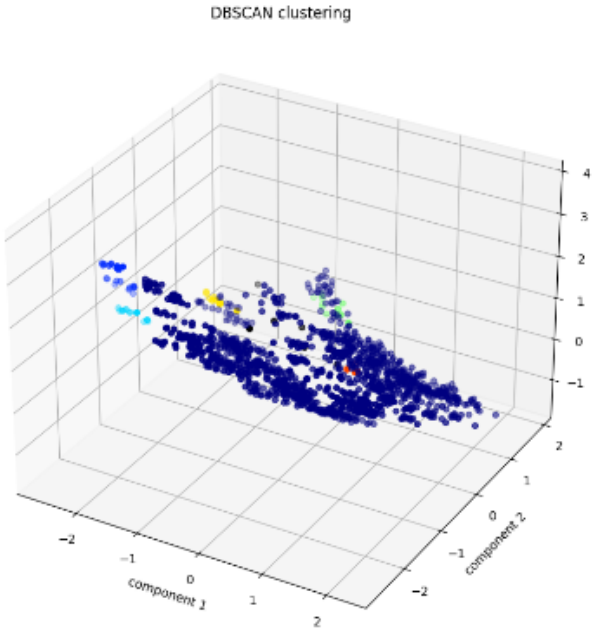

Figura 4: DBSCAN

3 Supervised Learning

Dividiamo il dataset in un training set e un test set di 372 elementi. Questa divisione permette di valutare l'efficacia di diversi modelli su informazioni non utilizzate durante la fase di training.

I modelli che prendiamo in considerazione sono:

- Logistic Regression
- Decision Tree
- Gaussian Naive Bayes
- K-Nearest Neighbour

Ognuno di questi modelli è stato implementato utilizzando la libreria `scikit-learn`, fornita da Python. L'obiettivo principale è valutare le prestazioni di ciascun modello e trarre conclusioni sul loro impatto ed efficacia nel predire correttamente se un'immagine rappresenta o meno una banconota contraffatta. Per valutare i modelli, consideriamo diverse metriche di valutazione, tra cui accuracy, precision, recall e F1-score.

Evaluation metrics:	accuracy	precision	recall	f1_score
Logistic Regression	0.997312	0.994048	1.000000	0.997015
Decision Tree	0.983871	0.976331	0.988024	0.982143
Naive Bayesian	0.852151	0.868421	0.790419	0.827586
k-NN	0.994624	0.988166	1.000000	0.994048

Figura 5: Tabella di valutazione dei modelli

Il modello Logistic Regression è caratterizzato dai parametri `C` e `penalty`, che influenzano la complessità del modello e il tipo di regolarizzazione applicata. Con un valore di `C` pari a 10 e `penalty` impostato su 'l1', il modello si adatta meglio ai dati di addestramento, con il rischio di overfitting. La regolarizzazione L1

penalizza i coefficienti del modello che assumono valori molto elevati, aiutando a prevenirlo e migliorando la generalizzazione.

Il modello Decision Tree ha ottenuto buoni punteggi per tutte le metriche; tuttavia, l'accuratezza è leggermente inferiore rispetto a K-NN e Logistic Regression. Inoltre, si può notare una buona precisione e un recall accettabile, ma leggermente più basso rispetto agli altri due modelli.

È stata effettuata la cross-validation per il Gaussian Naive Bayes. I risultati indicano una notevole variabilità nelle prestazioni del modello tra i diversi fold, in particolare per quanto riguarda il recall. Ciò suggerisce che il modello potrebbe avere difficoltà nell'identificare correttamente tutte le istanze della classe positiva nei vari fold, indicando una possibile scarsa capacità di generalizzazione.

Infine, per il modello K-NN, il parametro ottimale `n_neighbors` risulta essere 1, indicando che il modello considererà solo il punto più vicino per effettuare una previsione.

4 Conclusione

Abbiamo esaminato l'applicazione di algoritmi di unsupervised e supervised machine learning per identificare banconote contraffatte basandoci su caratteristiche derivate dalle immagini delle stesse. Dall'analisi dei dati, abbiamo osservato che le tecniche di confronto delle distribuzioni delle variabili tra le classi, PCA, PCA con K-means, t-SNE e DBSCAN hanno fornito risultati differenti. Ad esempio, la variabile entropia non ha mostrato una differenza significativa tra le classi, mentre sia PCA che PCA con K-means hanno mostrato una mescolanza delle due classi. L'utilizzo di t-SNE ha rilevato due cluster distinti, a differenza di DBSCAN, che non è stato efficace a causa della vicinanza e della mescolanza dei diversi cluster.

In generale, i risultati suggeriscono che le caratteristiche delle immagini delle banconote potrebbero non essere sufficienti per identificare con precisione le banconote contraffatte utilizzando questi approcci di machine learning.

Passando all'analisi dei modelli di supervised learning applicati al dataset, abbiamo ottenuto risultati significativi per il rilevamento delle banconote contraffatte. La logistic regression, con un parametro di regolarizzazione C=10 e una regolarizzazione L1, si è dimostrata uno dei migliori modelli. Possiamo dire che la logistic regression è altamente efficace nel discriminare tra banconote contraffatte.

Il decision tree (ID3) ha mostrato un'accuratezza complessivamente buona, tuttavia, il recall è risultato inferiore rispetto agli altri modelli, indicando una maggiore tendenza a predire erroneamente le banconote genuine come contraffatte.

Il Gaussian Naive Bayes ha prodotto risultati decenti, ma la variabilità tra i risultati ottenuti attraverso la cross-validation con k-fold suggerisce che potrebbe non essere un modello affidabile per il dataset in questione.

Infine, il modello k-NN con un numero di vicini pari a 1 è risultato il migliore tra tutti i modelli valutati. Questo suggerisce che k-NN è estremamente efficace nel rilevare banconote contraffatte nel dataset analizzato.

In conclusione, i modelli di supervised learning si sono dimostrati promettenti per individuare correttamente le banconote contraffatte, con la logistic regression e k-NN che si distinguono come i modelli più efficaci.