Progetto Machine Learning

Davide Pietrasanta, 844824
Fabio D’Elia, 829937

Descrizione del dominio di riferimento e obiettivi dell’elaborato

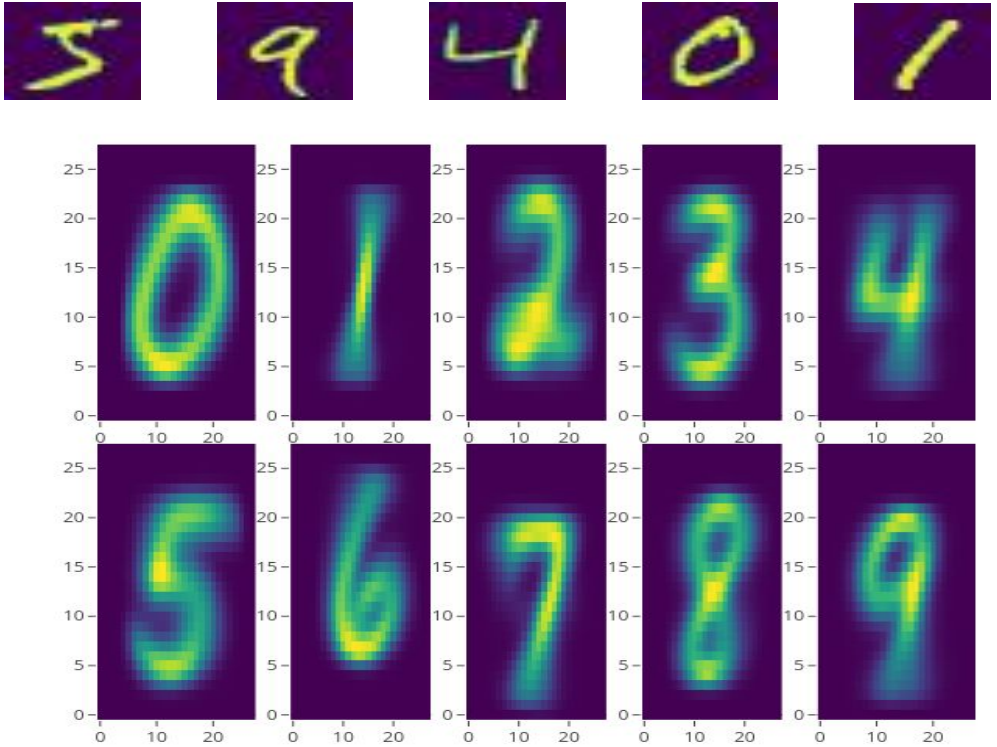
Implementazione di due algoritmi di apprendimento supervisionato per la classificazione di numeri scritti a mano.

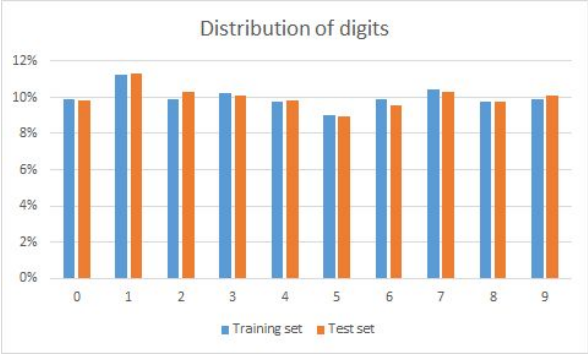
Dataset

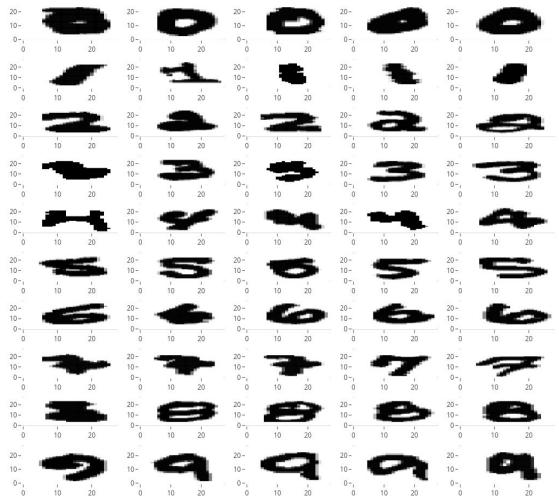
Dataset MNIST

Il database MNIST è composto da 60,000 esempi di numeri scritti a mano nel train-set e 10,000 nel test-set.

Le immagini sono state normalizzate e centrate in una matrice 28x28.







Analisi dei dati

Abbiamo verificato la distribuzione uniforme delle classi nel dataset usando il chi-quadro test.

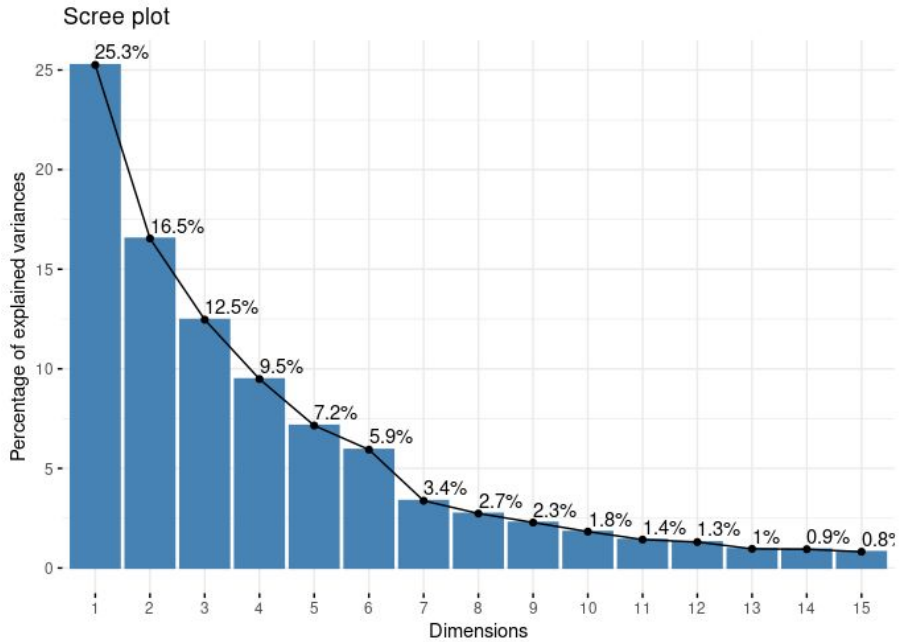
Un attento studio della distanza di ogni campione dal centroide della classe di appartenenza, ci ha fatto notare i possibili problemi di classificazione.

Principal Component Analysis

PCA

Abbiamo deciso di ridurre il numero di dimensioni del dataset in input, utilizzando la PCA.

Abbiamo deciso di mantenere 20 PCs che spiegavano il 95% della varianza cumulata.



Modelli

NAIVE BAYES

Si pensa che l’uso della probabilità possa aiutare nella predizione delle cifre.
Per esempio, una istanza con il centro bianco avrà poche probabilità di essere uno zero.

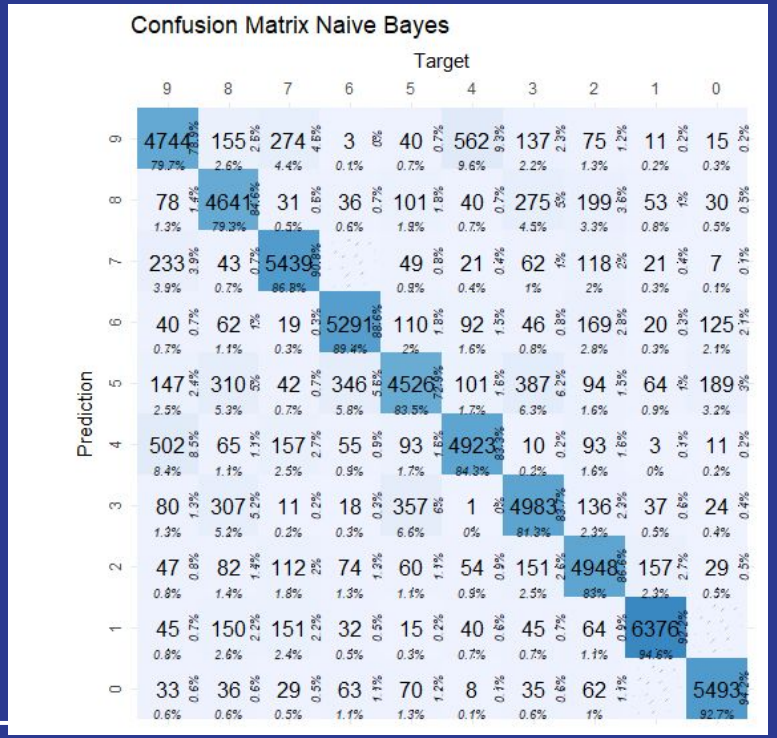
NEURAL NETWORK

La scelta dell’utilizzo di una rete neurale è dovuta dal fatto che una rete neurali si presenta come un modello flessibile e molto simile alle CNN.

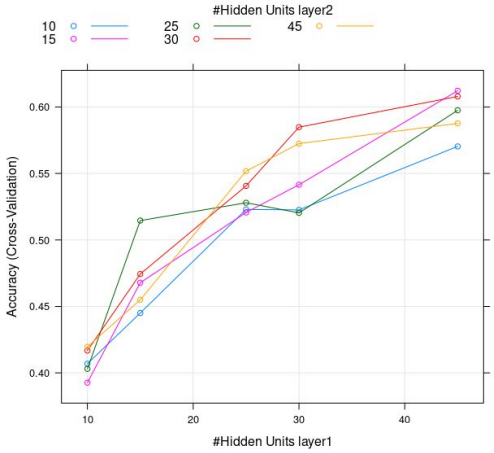
NAIVE BAYES

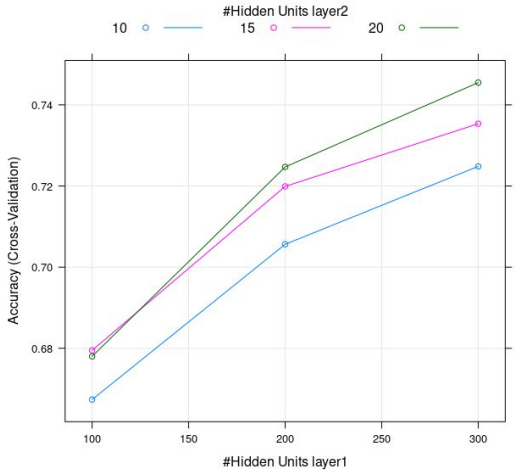
È stata applicata la 10-fold cross validation, in modo da valutare le performance del modello.

- **Accuracy:** 0.8520
- **Precision:** 0.8556952
- **Recall:** 0.8546793
- **F-measure:** 0.8546174
- **AUC:** 0.92027810
- **Accuracy of test:** 0.8621



NEURAL NETWORK



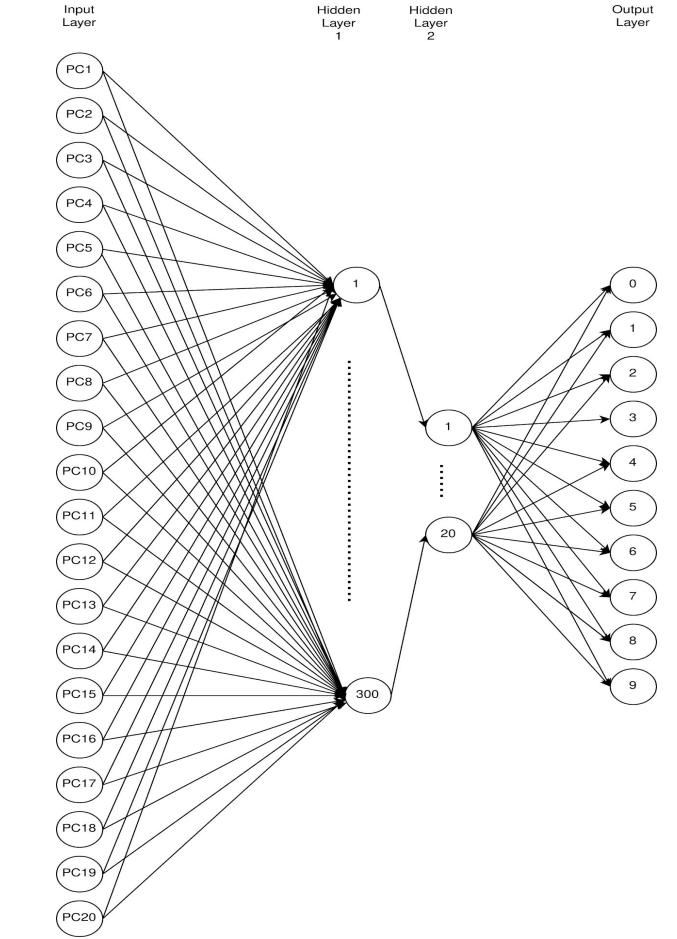


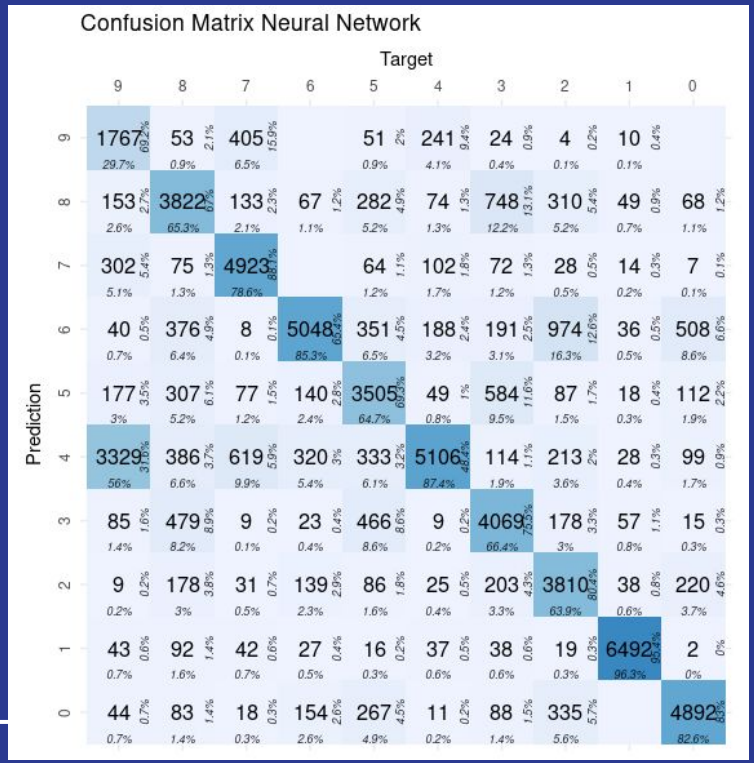
È stata applicata la 10-fold cross validation, in modo da valutare le performance del modello.

Attraverso 2 fasi di tuning differenti si è deciso di utilizzare una NN con un layout composto da 2 strati nascosti, con 300 nodi nel primo e 20 nel secondo.

- **Accuracy:** 0.7449164
- **Precision:** 0.742
- **Recall:** 0.72
- **F-measure:** 0.715
- **AUC:** 0.8453
- **Accuracy of test:** 0.7226

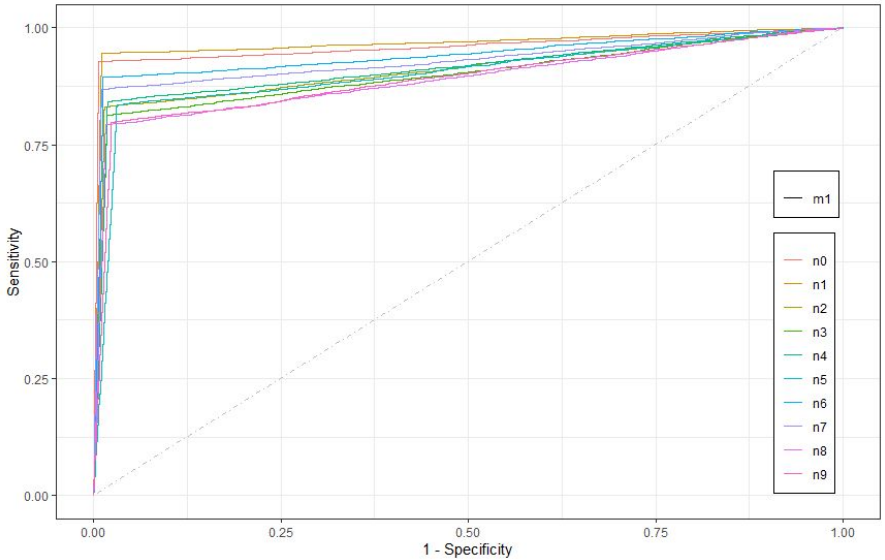
NEURAL NETWORK



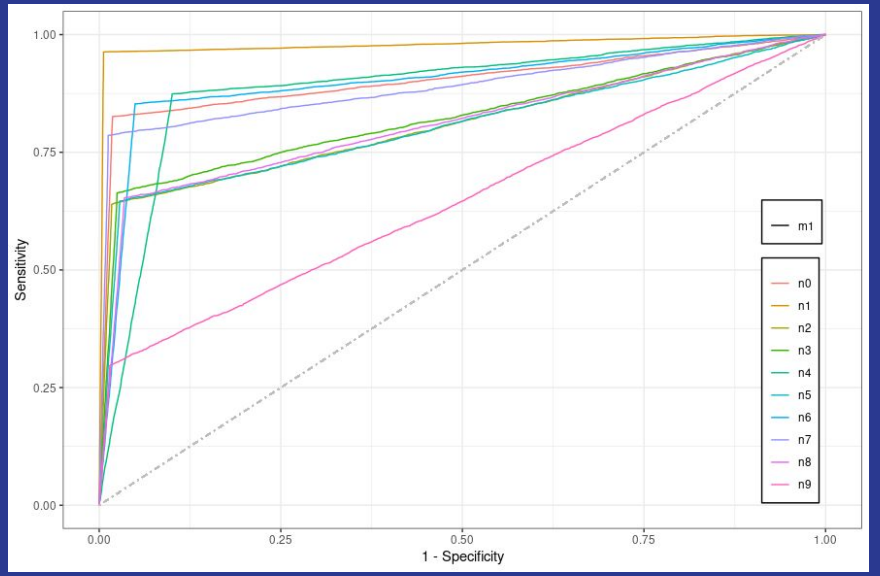


ROC e AUC

Naive Bayes AUC: 0.920278



Neural network AUC: 0.8453



Conclusioni

Entrambi i modelli, nonostante non siano modelli ottimali per il problema definito, forniscono delle discrete performance, anche senza usare l'informazione spaziale.

Si può notare che Naive Bayes fornisce predizioni in modo più accurato, ma con tempi maggiori, mentre Neural Network fornisce predizioni più veloci ma in modo meno accurato.

Pensiamo inoltre che con un miglior tuning degli iperparametri della Neural Network si possano raggiungere migliori risultati.

	NAIVE BAYES	NEURAL NETWORK
Train (s)	1.37	272.79
Predict test (s)	42.79	0.482
Predict train (s)	248.4	5.321
Accuracy (10-f CV)	0.852	0.745
Accuracy (test)	0.862	0.723