Homework 2

Studente: Santaera Alessandro, 1000061221

Corso: Ingegneria informatica, II anno, canale M-Z

Descrizione del dataset CIFAR-10

- Il dataset è costituito da 60.000 immagini a colori.
- Appartengono a 10 classi, le quali, singolarmente, ne contengono 6.000.
- Ogni immagine presenta una dimensione 32x32 pixel ed una struttura a tre canali (RGB).
- Le immagini sono suddivise in due insiemi:
  1. 50.000 per il training set
  2. 10.000 per il test set

Descrizione della metodologia adottata

- **CARICAMENTO DEL DATASET**

1. Importo il dataset attraverso la libreria `tensorflow.keras.datasets`, ottenendo un set di allenamento (x_train, y_train) ed uno di test (x_test, y_test).

2. Uso la funzione `train_test_split` per velocizzare l'allenamento, considero solo una porzione del dataset (20%) e riformatto ogni immagine in un vettore unidimensionale.

3. Applico la `PCA` , ad entrambi i set, per ridurre la dimensione, mantenendo il 95% di varianza.

- Definisco una funzione `plot_confusion_matrix`, utile successivamente per rappresentare le matrici di confusione dei singoli modelli di classificazione.

Descrizione della metodologia adottata

- **APPLICAZIONE MODELLI DI CLASSIFICAZIONE**

  Analizzo singolarmente i vari modelli di classificazione scelti:
  - Regressione logistica,
  - k-NN,
  - SVM,
  - decision tree.

  Ottengo per ognuno:
  - un valore distinto di accuracy mediante la funzione `accuracy_score`
  - una matrice di confusione che rappresento graficamente grazie alla funzione (dichiarata precedentemente) `plot_confusion_matrix`.

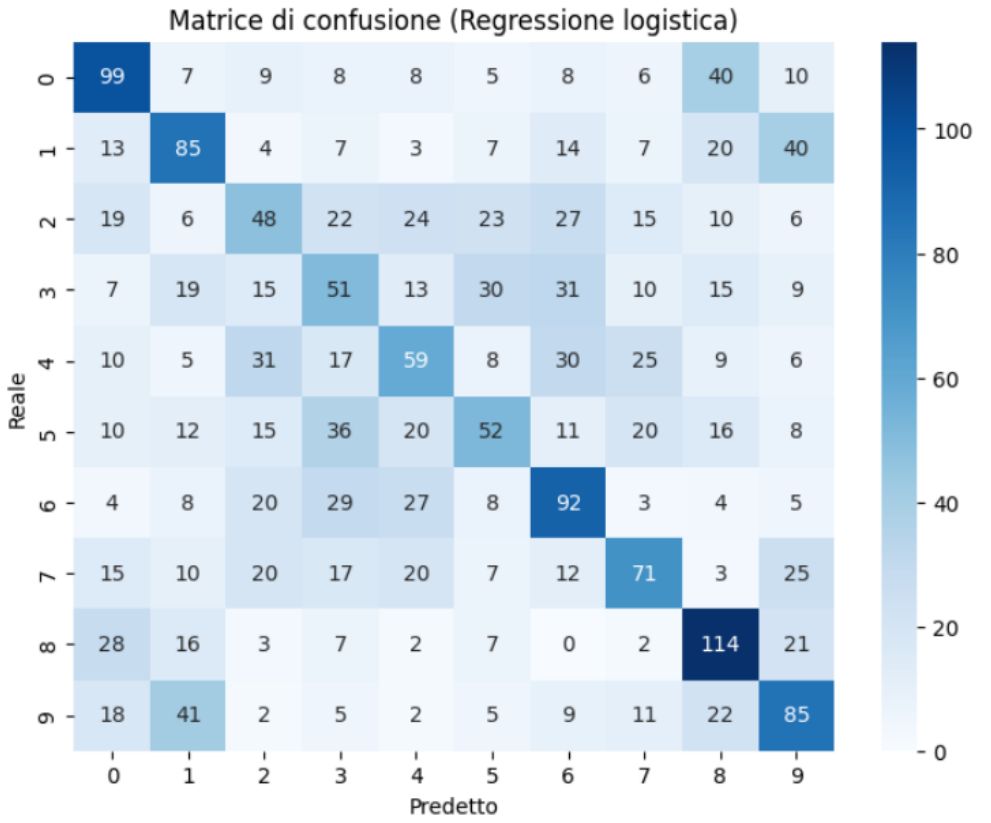
- **MODEL SELECTION:**

  1. Creo un dizionario contenente le varie accuracy ottenute (`model_accuracy`)
  2. Uso la `Grid search` per ogni modello, utile per la ricerca dei parametri ottimali, andando a testare tutte le possibili configurazioni e selezionare la migliore.
  3. Termino la model selection trovando il modello con la migliore accuratezza

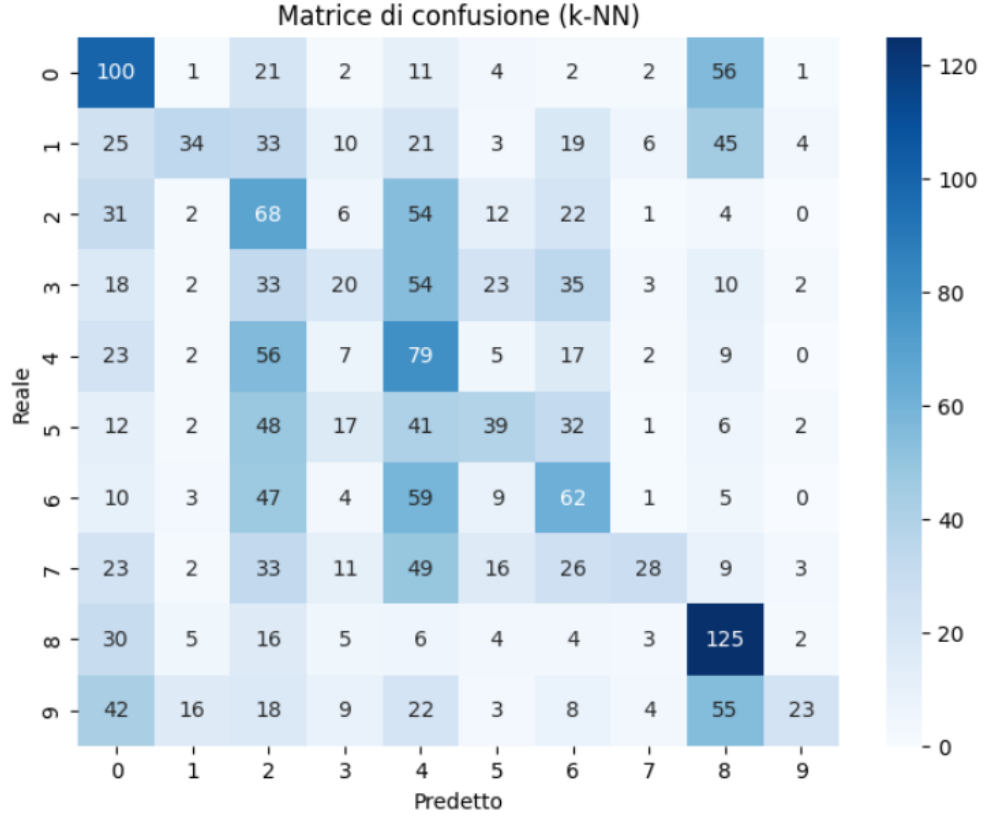
Descrizione dei risultati

- Riporto i risultati prima della **Grid Search**

```
Logistic Regression accuracy: 0.378
 Matrice di confusione per Regressione logistica:
```



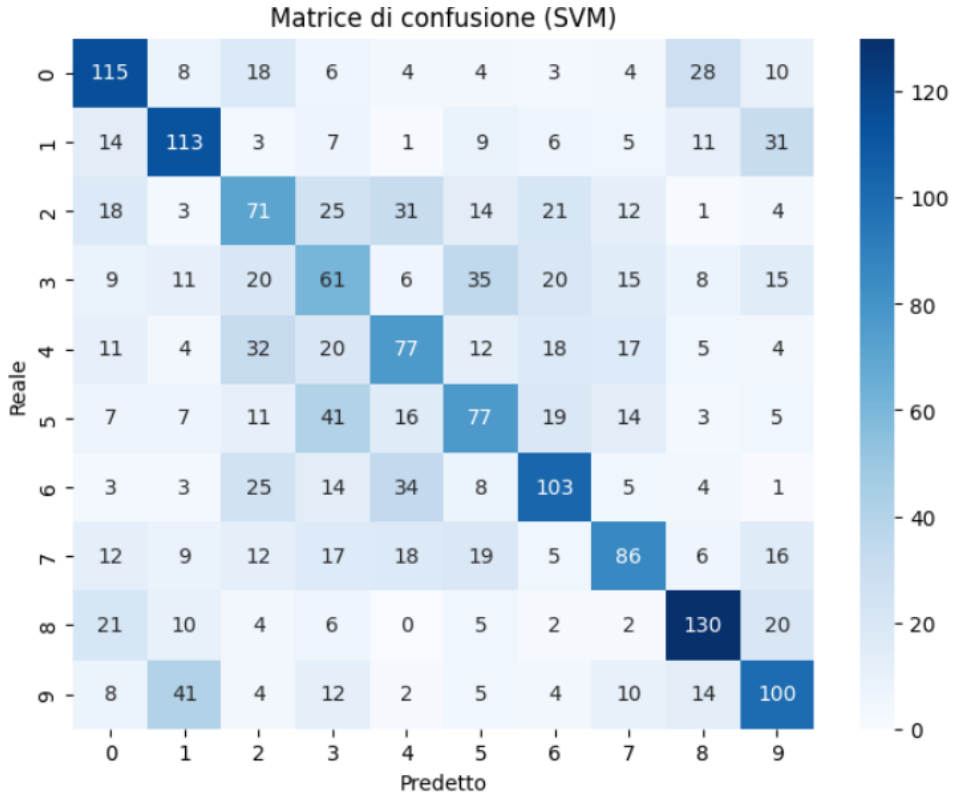

```
k-NN accuracy: 0.289
 Matrice di confusione per k-NN:
```



Descrizione dei risultati

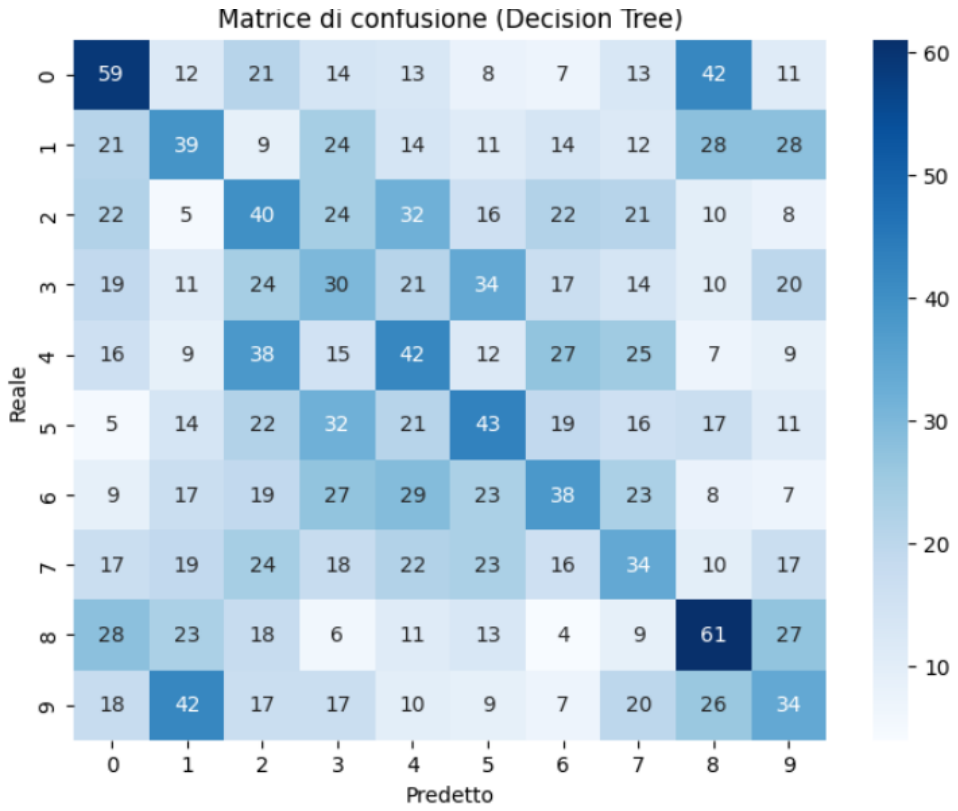
Accuracy SVM: 0.4665

Matrice di confusione per SVM:



Accuracy Decision Tree: 0.21

Matrice di confusione per Decision Tree:



Descrizione dei risultati

- Attraverso **Grid Search**, trovo:
  - Migliori parametri per ogni singolo modello di classificazione
  - Accuracy di ogni miglior modello

```
Migliori parametri per Regressione Logistica: {'C': 0.01, 'max_iter': 3000, 'penalty': 'l2', 'solver': 'saga'}
Accuracy della miglior Logistic Regression: 0.3765

Migliori parametri per k-NN: {'metric': 'euclidean', 'n_neighbors': 7, 'weights': 'distance'}
Accuracy del miglior k-NN: 0.3125

Migliori parametri per SVM: {'C': 10, 'kernel': 'rbf'}
Accuracy del miglior SVM: 0.4665

Migliori parametri per Decision Tree: {'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 10}
Accuracy del miglior Decision Tree: 0.264
```

Descrizione dei risultati

- Confronto i singoli modelli, notando che le accuracy sono:
  - 37,65% Regressione Logistica
  - 31,25% k-NN
  - 46,65% SVM
  - 26,4% Alberi Decisionali

- Scelgo il modello migliore in base all'accuracy:

```
Il modello migliore è: SVM con accuratezza =  0.4665
```