HOMEWORK 2

RUNDO MICHAEL SEBASTIAN

N.MATRICOLA: 1000058506

TASKS

Dataset	• Descrizione del dataset CIFAR10 • Caricamento dataset

Metodologia adottata	• Spiegare la metodologia adottata
	  • K-NN, SVM, decision tree, Regression Logistica
	  • Model selection

Risultati	• Stima dei risultati • Prestazioni ottenute

DATASET

- Descrizione del dataset CIFAR10
  - 60.000 immagini 32x32 divise in 10 classi tipologia (airplane, automobile, bird..)
  - Training set: 50.000 / 5, 10.000 per ognuna (83,33%)
  - Test set: 10.000 (16,67%)

DATASET

- Caricamento
  - Il dataset in formato pickle viene convertito in dizionario dalla quale prendiamo I dati che ci interessano
  - E' stata presa una parte del test set da usare come Validation set
  - Inoltre è stato eseguito un sotto-campionamento del dataset per problemi di tempistiche di calcolo

METODOLOGIA ADOTTATA

- Spiegare la metodologia adottata
  - I Modelli che andremo a provare sono
    - Regressione Logitistica
    - K-NN (+PCA)
    - SVM (+SVM Kernell)
    - Decision Tree

METODOLOGIA ADOTTATA

- Supposizioni
  - Dall'analisi del dataset mi aspetto che:

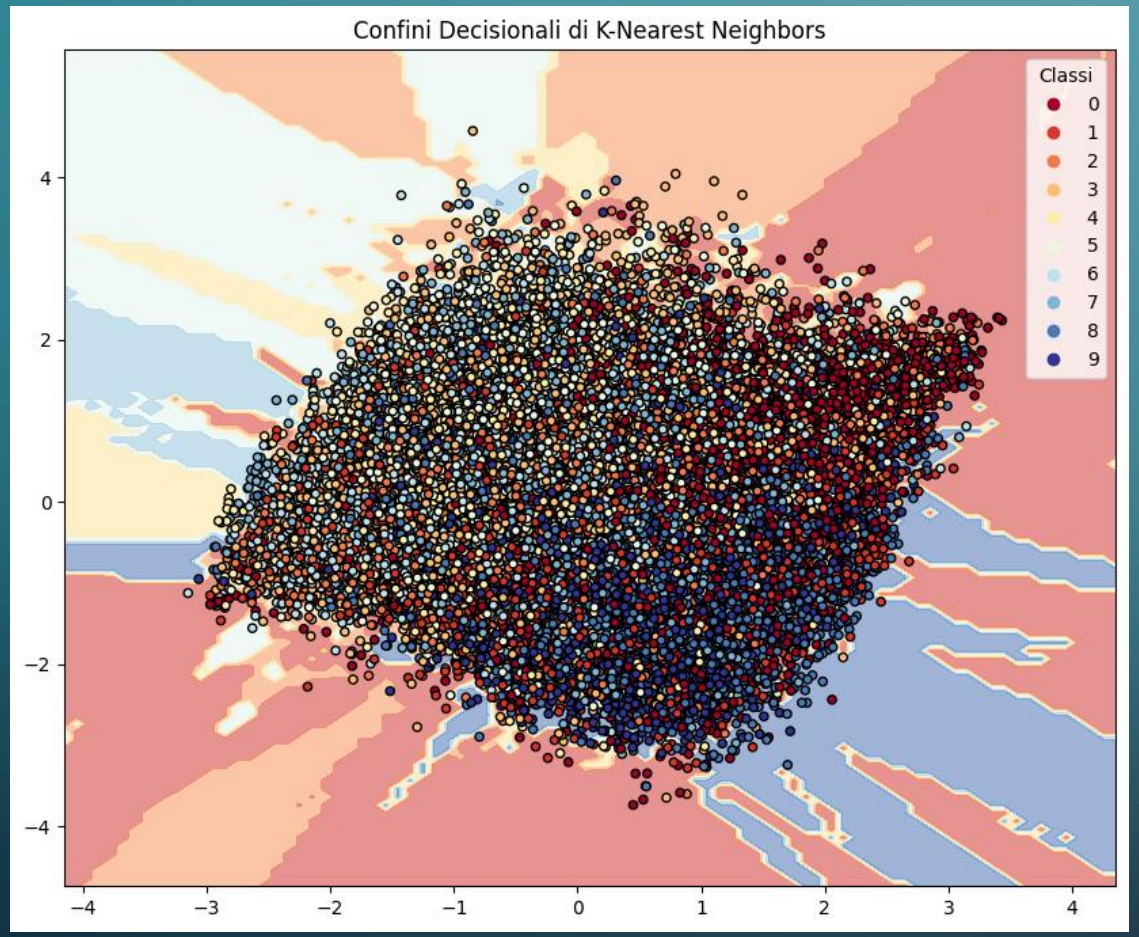
Modello	Accuratezza	Scalabilità
Regr. Logistica	Bassa	Alta
SVM Lineare	Bassa	Media
SVM + Kernell	Alta	Bassa
K-NN + PCA	Media	Bassa
Decision Tree	Bassa	Alta

  - Il Decision Tree è quello meno utile in questo caso, nel codice si troveranno I risultati ma non verrà considerato nel powerpoint in quanto è molto poco performante

METODOLOGIA ADOTTATA

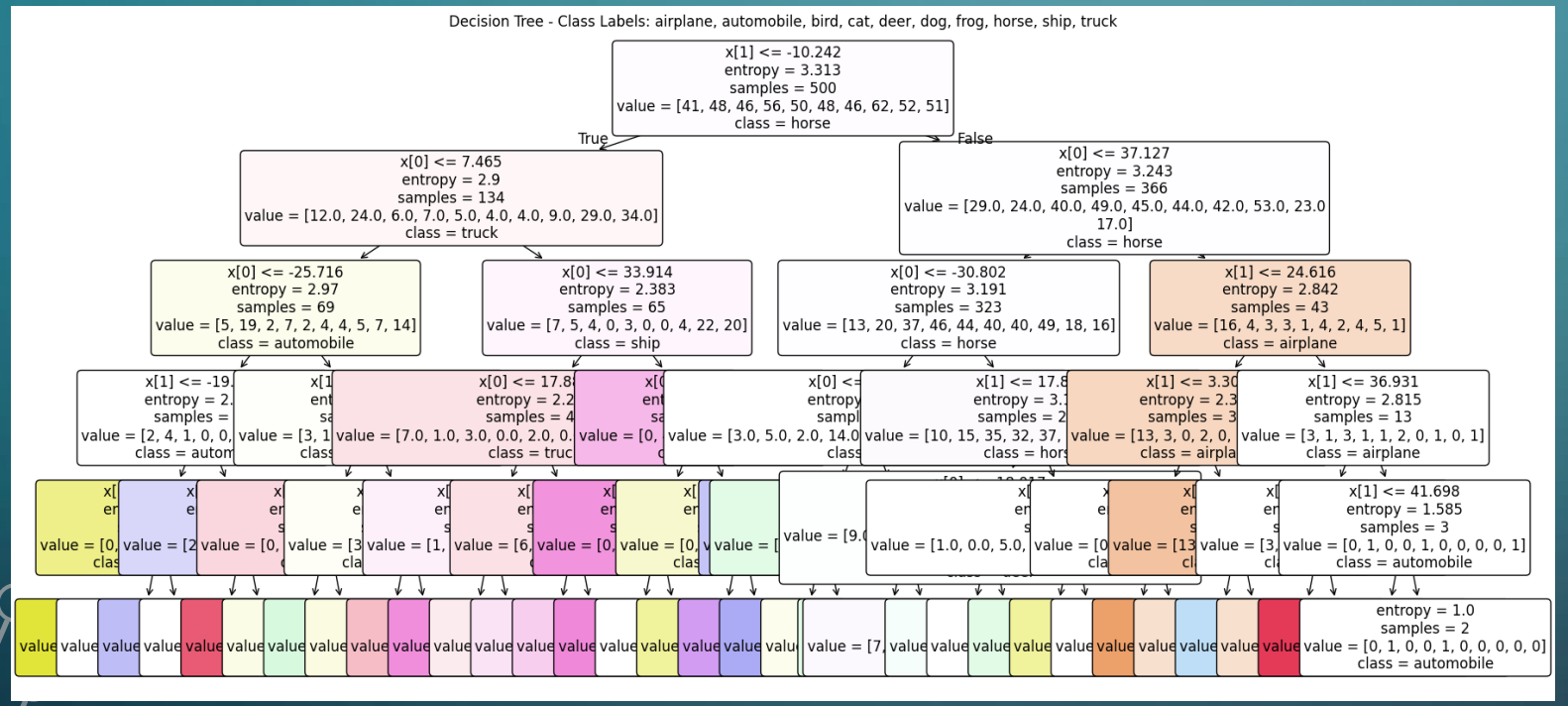
Regressione Logistica	SVM	SVM + Kernell	K-NN
• Test Accuracy: 0.2306 • Validation Accuracy: 0.2112	▪ Test Accuracy: 0.2900 ▪ Validation Accuracy: 0.2688	▪ Test Accuracy: 0.3356 ▪ Validation Accuracy: 0.3378	▪ Test Accuracy: 0.2218 ▪ Validation Accuracy: 0.2288 ▪ L'analisi dei confini decisionali si trova nelle slide successive

METODOLOGIA ADOTTATA



- Analisi confini k-NN
  - Dall'immagine si evince che, anche usando la PCA, non andremo a migliorare di tanto il risultato in quanto il modello è tipicamente sensibile al rumore, anche se dovessimo eseguire una forte riduzione di dimensionalità (PCA) rimarrebbe inefficiente.

METODOLOGIA ADOTTATA



- Cenni al Decision Tree (con PCA)
  - Vediamo come il tree overfitti il train set, dovuto anche ai pochi sample, rischio alto di overfit .
  - E’ bensì inutile tentare tecniche di Pruning in quanto il modello sia limitato e tendi comunque a overfittare

RISULTATI

- Stima dei risultati ottenuti
  - In ordine di performance accuratezza abbiamo:
    1. SVM + Kernel
    2. SVM Lineare
    3. Regressione Logistica
    4. K-NN
    5. Decision Tree

RISULTATI

- SVM + Kernel
  - Considerando le performance computazionali delle macchine di calcolo a disposizione, dunque anche considerando il sotto-campionamento a 500 immagini per train, l'SVM + Kernel, come da precedente analisi, è il più adatto al tipo di dati forniti.
  - La differenza sta nella separazione delle classi, nell'SVM classico avremo una separazione lineare, mentre nell'SVM + Kernel (es. Rbf), proiettiamo I dati in uno spazio ad alta dimensionalità dove le classi sono più separabili.