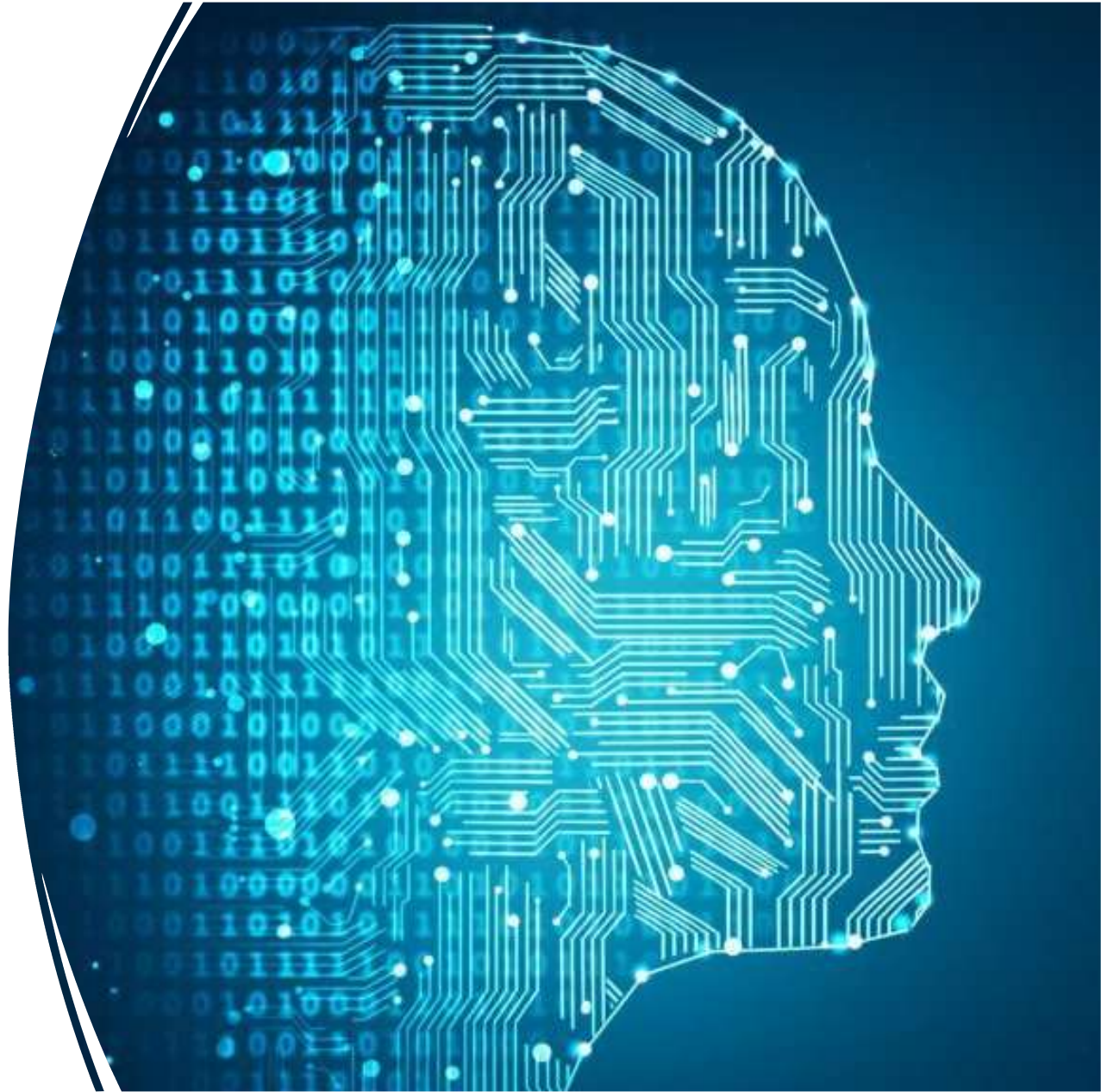


HOMEWORK 2

Nicotra Giulio
1000063414



Dataset CIFAR-10

Il dataset utilizzato in questo progetto si chiama **CIFAR-10**, ed è uno dei più usati per problemi di classificazione di immagini.

Contiene in totale **60.000 immagini a colori**, tutte molto piccole: hanno una risoluzione di **32x32 pixel**, quindi sono abbastanza semplici da elaborare. Queste immagini sono suddivise in **10 categorie**: aerei, automobili, uccelli, gatti, cervi, cani, rane, cavalli, navi e camion.

Ogni classe contiene **6.000 immagini**, suddivise in 5.000 per l'addestramento e 1.000 per il test.

Il set di dati CIFAR-10 è suddiviso in due sottoinsiemi:

- **50.000 immagini** vengono usate per addestrare i modelli (training set)
- **10.000 immagini** vengono usate per valutare i risultati (test set)

L'obiettivo è quindi costruire un modello che, osservando un'immagine, sappia **riconoscere automaticamente a quale delle 10 classi appartiene**.

Metodologia adottata

Per ridurre i tempi di calcolo, è stato utilizzato un **sottoinsieme del 20%** del dataset totale.

Le immagini, inizialmente in formato 32x32x3, sono state **convertite in vettori unidimensionali** tramite operazione di flattening.

Successivamente è stata eseguita la **standardizzazione delle feature** con **StandardScaler** per normalizzare il dataset e garantire una scala uniforme di valori.

Il training set è stato ulteriormente suddiviso in training set (80%) e validation set (20%) per poter effettuare una corretta valutazione dei modelli.

Modelli di classificazione

Sono stati testati quattro modelli differenti:

- **Regressione logistica** → con incremento del numero massimo di iterazioni (max_iter=2000) per migliorarne la convergenza.
- **k-Nearest Neighbors (k-NN)** → con scelta del parametro k tra 3, 5 e 7 per selezionare il miglior compromesso.
- **Support Vector Machine (SVM)** → utilizzata con parametri standard della classe SVC.
- **Decision Tree** → per il quale sono stati testati diversi valori della profondità massima (3, 5, 7 e senza limite).

Risultati

I modelli sono stati valutati sia sul **validation set**, sia sul **test set**, utilizzando l'**accuratezza** come metrica principale.

Il modello che ha ottenuto i risultati migliori è stato la **SVM**, con le performance più alte sia in validation che in test.

Per ogni modello è stata prodotta anche una **matrice di confusione**, utile per visualizzare in dettaglio le classi più frequentemente confuse.

La **SVM**, in particolare, **si è dimostrata più precisa** nel distinguere le classi visivamente simili, mentre il **Decision Tree ha mostrato maggiore incertezza**, commettendo più errori.

