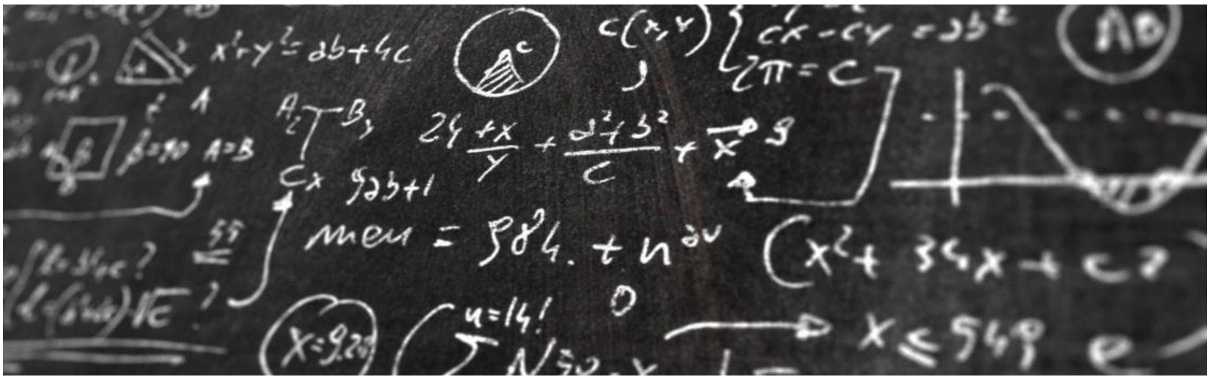


Machine Learning Project Description

General Information

Numerical Dataset

- **Name:** Houses Prices Dataset.

- **Description:** This dataset contains information about houses and their associated sale prices, intended for use in analyzing the factors that influence house prices. The dataset typically includes numerical and categorical variables that describe the physical attributes of the houses, their location, and market-related factors.

- **Total Samples:** 2000.

- **Missing Values:** The **square_feet** feature contains **10%** missing values for this feature. These represent houses for which the size data is not recorded or unavailable.

- **Target Variable: Price_USD.**

- **Dataset Splits:**
- **Training Set:** 80%.
- **Testing Set:** 20%.

Image Dataset

- **Dataset Name:** Traffic Sign Recognition.

- **Dataset Description:** This Dataset is a set of labeled images of traffic signs, like stop signs, speed limits, and warning signs. Each image has a label showing what type of sign it is.

- **Classes:**
- Class 1: **Crosswalk** with lable **0**.
- Class 2: **Speed limit** with lable **1**.
- Class 3: **Stop** with lable **2**.
- Class 4: **Traffic light** with lable **3**.

- **Image Dimensions:** 64 * 64 * 3.
- **Total Samples:** 2,077
- **Dataset Splits:**
  - **Training Set:** 80%.
  - **Testing Set:** 20%.

Algorithms and Models Implemented

- ### Numerical Dataset

1. Preprocessing

- **Filling** missing values in **Square_Feet** feature using the **mean** of this feature.
- **Encoding** categoric features like (Location, Street_Type) into lables.
- **Scaling** features like (Year_Built, Price_USD, Square_Feet, Price_per_Square_Foot)

2. Linear Regression

- **Description:** Linear Regression is a supervised learning algorithm used to predict continuous values. It finds the line of best fit by minimizing the mean squared error between predicted and actual values.

- **Evaluation Metrics:**

  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - R-Squared (R^2)

3. K-Nearest Neighbors (KNN) Regressor

- **Description:** KNN is a non-parametric algorithm that predicts a value based on the average (or weighted average) of the K nearest neighbors in feature space.

- **Evaluation Metrics:**
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - R-Squared (R^2)

Image Dataset

1. Preprocessing

- **Normalization** a grayscale normalization to reduce the effect of illumination's differences.
- **Apply PCA for dimensionality reduction** apply n_components=1672 to PCA to make
  the dataframe have at most 1672 features keeping only necessary ones.

2. Logistic Regression (Image Dataset)

- **Description:** Logistic Regression is a supervised classification algorithm used to predict the probability of a sample belonging to a specific class.

- **Evaluation Metrics:**
  - Accuracy
  - Confusion Matrix
  - Precision
  - Recall
  - ROC-AUC

3. K-Nearest Neighbors (KNN) Classifier (Image Dataset)

- **Description:** KNN classifies images by identifying the K nearest samples in feature space and assigning the most frequent label among them.

- **Evaluation Metrics:**
    - Accuracy
    - Confusion Matrix
    - Precision
    - Recall
- ROC-AUC

Results and Comparisons

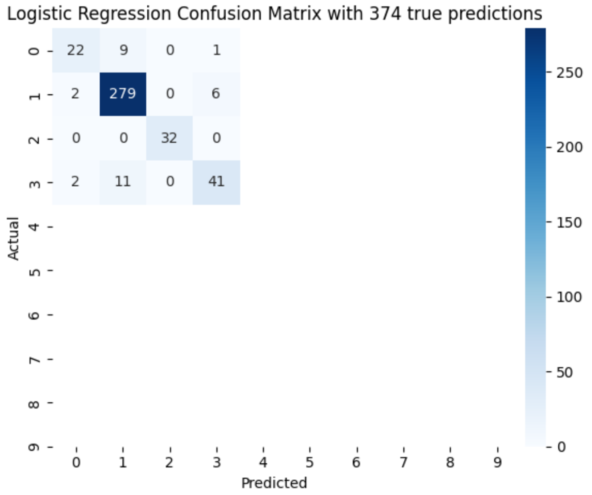
Numerical Dataset Results

Model	MSE	R^2	MAE
Linear Regression	1.042543	-0.005356	0.884773
KNN Regressor	1.288908	-0.242934	0.944548

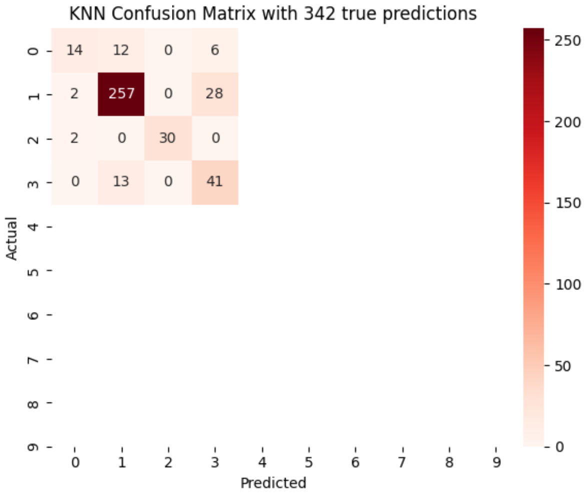
- **Analysis:** Linear Regression outperforms KNN on this numeric dataset. It has lower errors (MSE: 1.04 vs. 1.29, MAE: 0.88 vs. 0.94) and a higher R^2 score (-0.005 vs. -0.243). Linear Regression is the better model.

Image Dataset Results

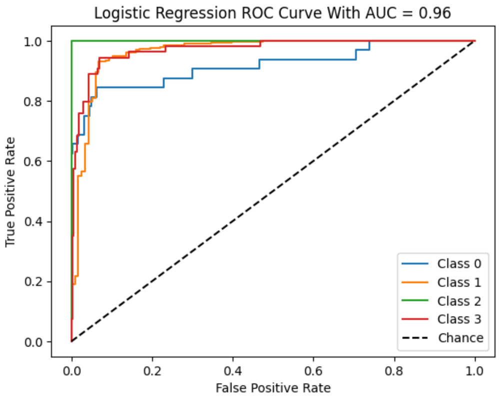
Model	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	0.92	0.92	0.92	0.96
KNN Classifier	0.84	0.84	0.84	0.92



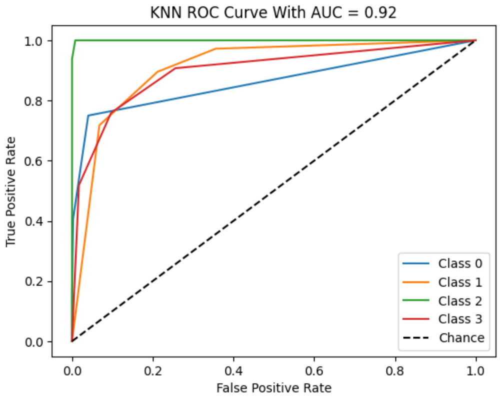
Logistic Regression Confusion Matrix with 374 true predictions



KNN Confusion Matrix with 342 true predictions



Logistic Regression ROC Curve With AUC = 0.96



KNN ROC Curve With AUC = 0.92

- **Analysis:** Logistic Regression outperforms KNN in all metrics, including Accuracy (0.92 vs. 0.84), Precision (0.92 vs. 0.84), Recall (0.92 vs. 0.84), and AUC (0.96 vs. 0.92). Logistic Regression is the better model for this dataset.