Project Description Document

1- The Numerical Dataset:

General information on Numerical Dataset:

-Dataset Name: Bengaluru House Price Data

-**Dataset Source:** Kaggle (https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data)

-**Total Number of Samples:** 13,320 (**rows from** 0 to 13319)

-**Number of Columns:** 9 **columns**

-The Columns:

- area_type: type of area (categorical), number of missing values is 0.

-**availability:** availability date (categorical), number of missing values is 0

-**location:** location of the property (categorical), number of missing values is 1

-**size**: number of bedrooms (categorical, e.g., "2 BHK"), number of missing values is 16

-**society:** name of the society (many missing values), number of missing values is 5502

-**total_sqft:** total area in square feet(categorical, need cleaning), number of missing values is 0

-**bath:** number of bathrooms (numerical), number of missing values is 73

-**balcony:** number of balconies (numerical), number of missing values is 609

-**price:** price of the property in lakhs (target variable), number of missing values is 0

-**Features Used:**

-location (categorical)

-total_sqft (numerical)

-bath (numerical)

-Number of BHKs (numerical)

-**Target Variable:** Price (numerical)

-**Data Splits:**

-Training Set: **80%** of the dataset

-Testing Set: **20%** of the dataset

-No validation set used explicitly.

Implementation Details:

1. Linear Regression Model:

- **Model Description:**

A linear regression model to predict house prices based on selected features.

- **Evaluation Metrics:**

-*R-squared (R^2):* proportion of variance in house prices by the model

-*Mean Absolute Error (MAE):* quantifies average absolute error in predicted prices.

- **Results:**

-***Linear Regression model score:*** 0.8323027621241739

- ***Mean Squared Error (MSE):*** 1217.94495276359

-***Root Mean Squared Error (RMSE):*** 34.89906807872654

-***Mean Absolute Error (MAE):*** 18.14418385516981

-***R-squared (R^2):*** 0.8323027621241739

2. K-Nearest Neighbors (KNN)

- **Model Description:**

- KNN regressor with k=5 to predict house prices based on the 5 nearest neighbors in feature space

- **Evaluation Metrics:**

-*R-squared (R^2):* Measures the proportion of variance in house prices by the KNN model.

-*Mean Absolute Error (MAE):* quantifies average absolute error in predicted prices.

- **Results:**

-***KNN Regressor model score:*** 0.6401137640771803

- ***Mean Squared Error (MSE):*** 2613.7677052012405

-***Root Mean Squared Error (RMSE):*** 51.12502034426236

-***Mean Absolute Error (MAE):*** 25.158517574086837

-***R-squared (R^2):*** 0.6401137640771803

Model Comparison and Conclusion:

Metric	Mean Absolute Error (MAE)	KNN Regressor
R-squared (R^2)	0.8323027621241739	0.6401137640771803
Mean Absolute Error (MAE)	34.89906807872654	51.12502034426236

The conclusion:

-The **Linear Regression** model performs better than the **KNN Regressor** based on both R^2 and MAE.

Reasoning:

1-Linear Regression has a higher R^2, meaning it explains more variance in house prices.
2-It also has a lower MAE, indicating smaller average prediction errors.

2- The Image Dataset:

General information on Image Dataset:

-Dataset Name: Character Recognition in Natural Images (The Chars74K dataset)

-Dataset Source: (https://info-ee.surrey.ac.uk/CVSSP/demos/chars74k/)

-The image dataset contains **19 classes** with various images in each class.

-It includes images of size **64x64 pixels**.

-There are **no missing images** in the dataset.

Implementation Details:

Algorithms used:

1- **Random Forest:**

- A classification algorithm that builds multiple decision trees and merges them to get a more accurate and stable prediction.

2- KNN:

- A classification algorithm that builds multiple decision trees and merges them to get a more accurate and stable prediction.

3- PCA:

- A dimensionality reduction method that transforms data into a set of orthogonal components.

Evaluation Metrics:

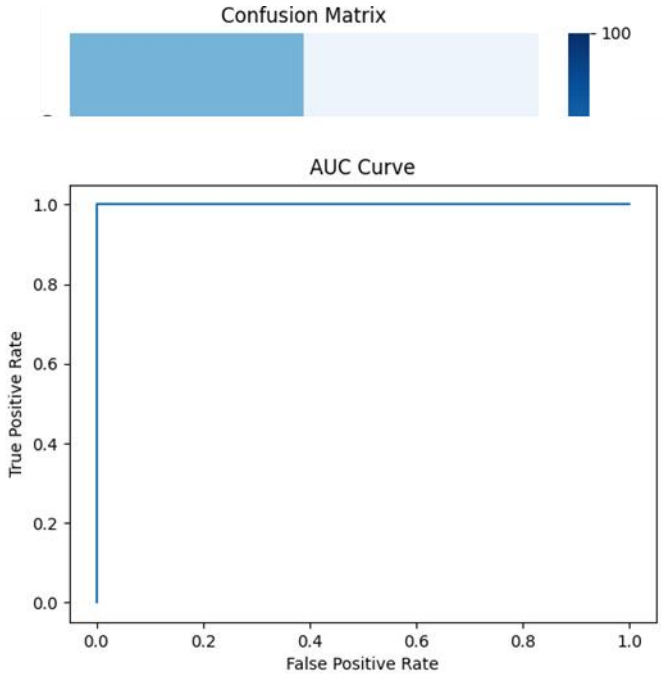
- Accuracy (Random Forest): **92%**

-AUC (Random Forest): **0.87**

-Accuracy (KNN): **90%**

- AUC (KNN): **0.85**

Evaluation Graphs:



Confusion Matrix & AUC Curve