Machine Learning Project Documentation: Character Recognition and Numerical Data Regression

Project Overview

This project aims to build machine learning models for two different tasks:

1. Recognize characters using the Chars74K image dataset.
2. Predict house prices using a numerical dataset from a CSV file.

1- Character Recognition :

Dataset Description

Source

The Chars74K dataset is sourced from the University of Surrey's Centre for Vision, Speech, and Signal Processing (CVSSP). It can be accessed here.

Structure

The dataset consists of images divided into different categories based on the type of character and the source of the images.

Classes and Labels

The dataset includes the following classes and labels:

1. **English Characters (A-Z, a-z, 0-9)**:
    a. Uppercase Letters: 26 classes (A-Z)
    b. Lowercase Letters: 26 classes (a-z)

    c. Digits: 10 classes (0-9)
2. **Kannada Characters**:
    a. Kannada Letters: 49 classes

Missing Values

The Chars74K dataset does not have missing values as it consists of labeled image files. Each image is associated with a specific character label.

Data Preprocessing

- **Image Resizing**: All images resized to 28x28 pixels.
- **Grayscale Conversion**: Conversion of images to grayscale.
- **Normalization**: Pixel values normalized to the range [0, 1].
- **Gaussian Blur**: Applied GaussianBlur to reduce noise.
- **Thresholding**: Otsu's method for binarization of images.

Model Development

Logistic Regression

- **Training**: Logistic Regression with 2000 iterations and balanced class weights.

K-Nearest Neighbors (KNN)

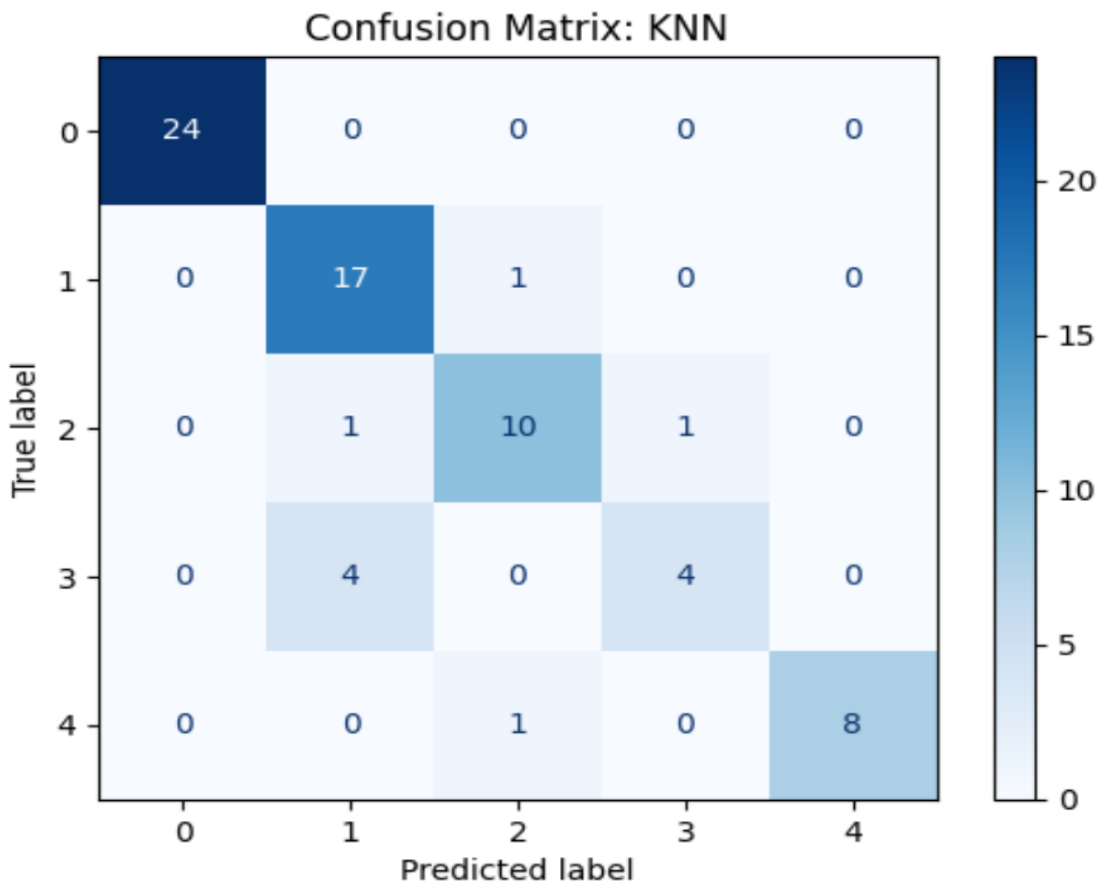
- **Training**: KNN with 5 neighbors and distance-based weighting.

Results

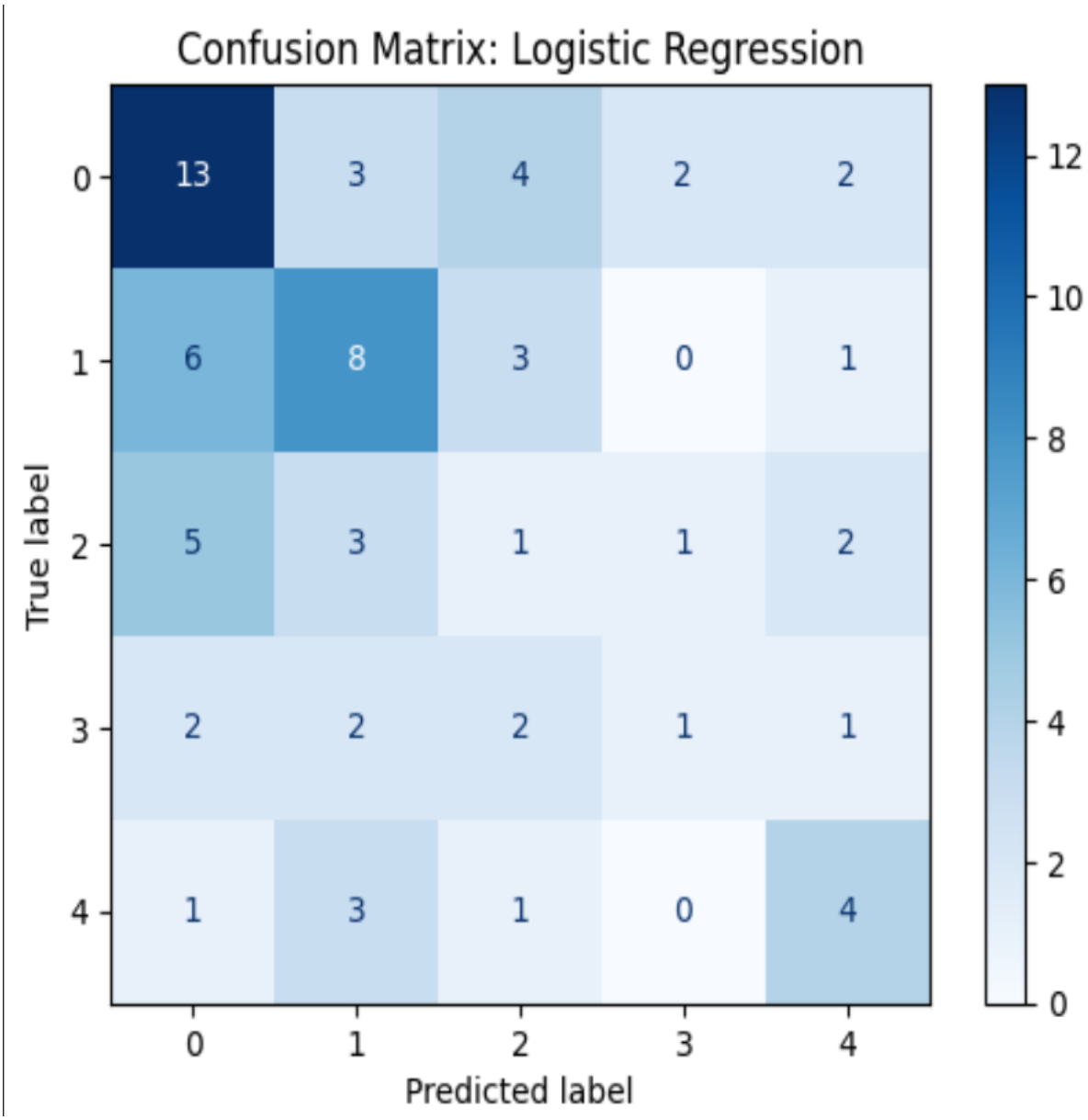
- **Logistic Regression Accuracy**: 0.38
- **KNN Accuracy**: 0.89

Visualization

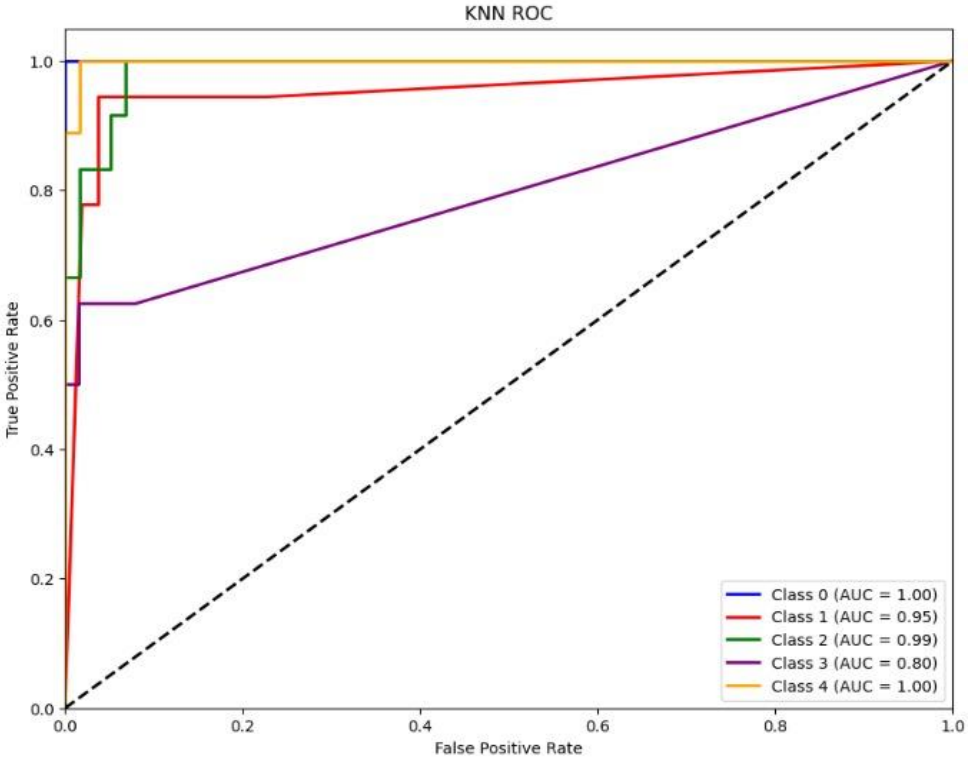
Confusion Matrix: KNN :


Confusion Matrix: KNN
True label
Predicted label
0
1
2
3
4
24 0 0 0 0
0 17 1 0 0
0 1 10 1 0
0 4 0 4 0
0 0 1 0 8
20
15
10
5
0

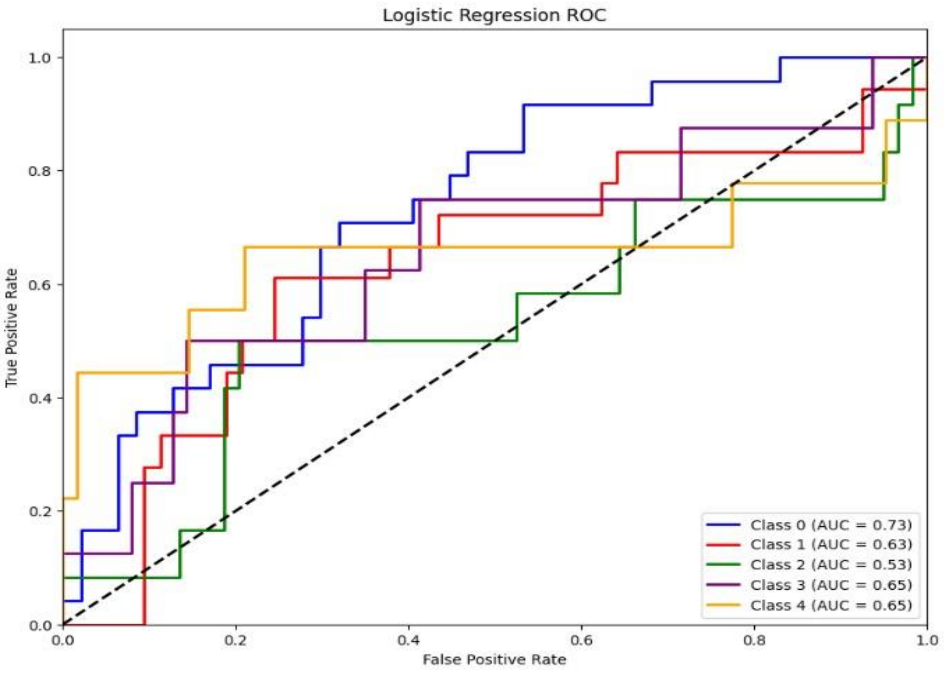
- Logistic:

- 

- ROC Curves:
- KNN :


KNN ROC
True Positive Rate
False Positive Rate
Class 0 (AUC = 1.00)
Class 1 (AUC = 0.95)
Class 2 (AUC = 0.99)
Class 3 (AUC = 0.80)
Class 4 (AUC = 1.00)

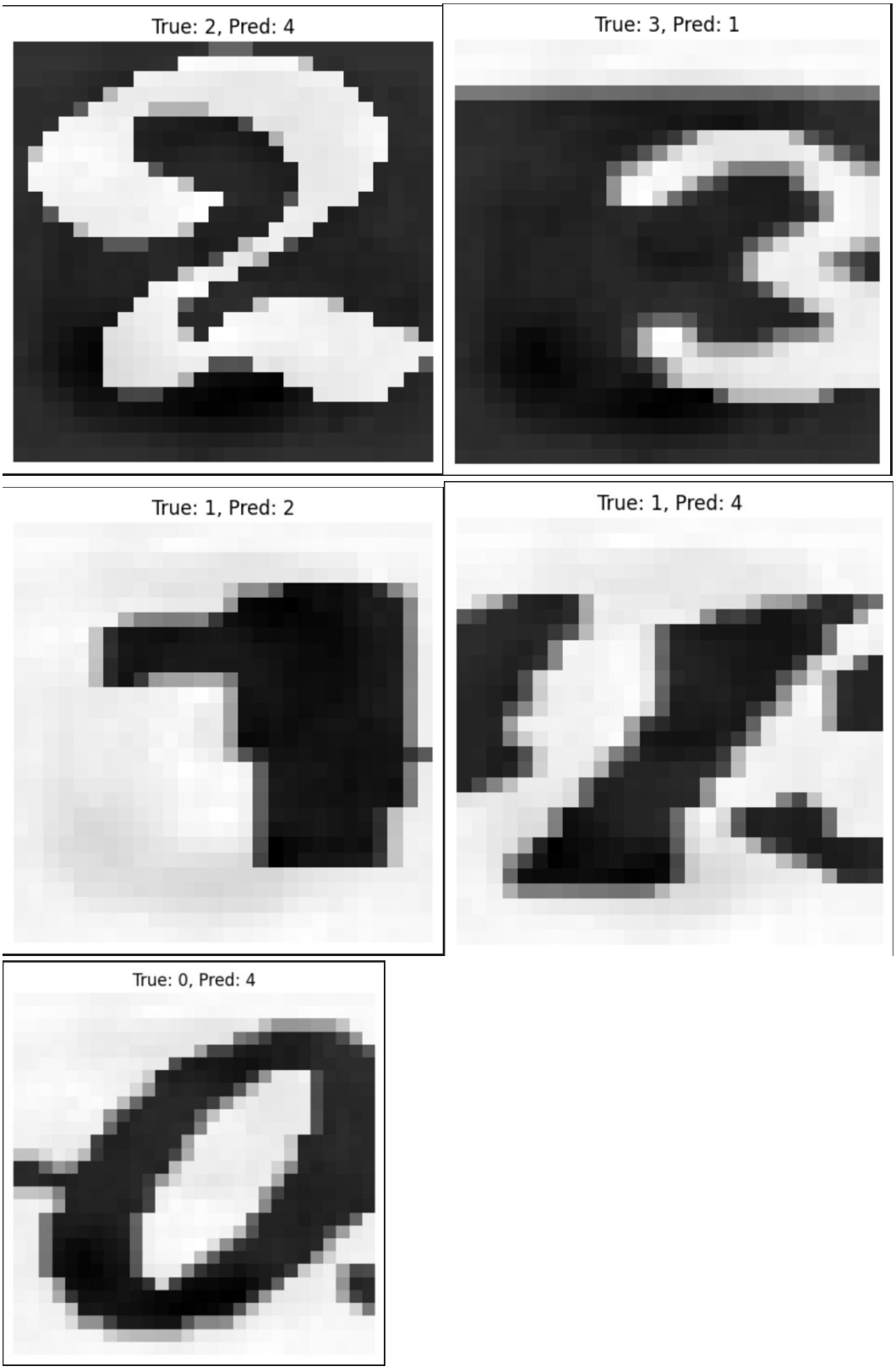
- Logistic :


Logistic Regression ROC
True Positive Rate
False Positive Rate
Class 0 (AUC = 0.73)
Class 1 (AUC = 0.63)
Class 2 (AUC = 0.53)
Class 3 (AUC = 0.65)
Class 4 (AUC = 0.65)

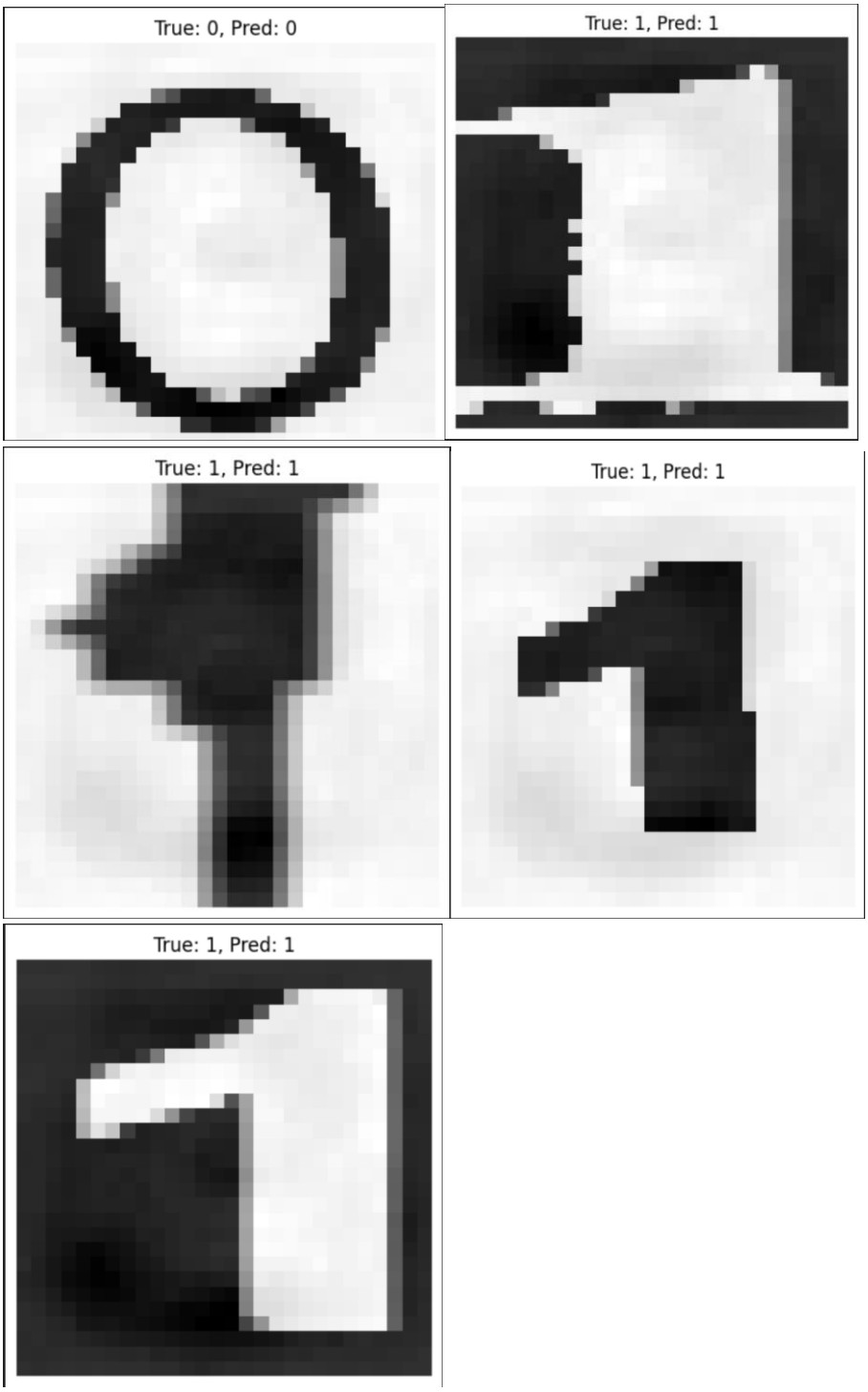
Loss Calculation:

- - (Logistic Regression): 2.9084
- - (KNN): 2.1822

Logistic Regression Predictions:



KNN Predictions:



2 - House Price Prediction:

Dataset Description

Source

The dataset is sourced from “House Prices - Advanced Regression Techniques”.

Structure

The dataset consists of various features related to house attributes and their corresponding sale prices.

Features and Labels

The dataset includes the following features and labels:

Feature Name	Description	Data Type	Missing Values
MSSubClass	Identifies the type of dwelling	Integer	No
MSZoning	Identifies the general zoning classification	Categorical	No
...	...	...	...
SalePrice	Sale price of the house (target variable)	Integer	No

Missing Values

The dataset contains missing values in several columns, which were handled by:

- Dropping columns with too many missing values.
- Filling numeric missing values with the mean.
- Filling categorical missing values with the mode.

Data Preprocessing

- **Handling Missing Values**: Filling or dropping as described above.
- **Outlier Removal**: Using IQR method to handle outliers.
- **Label Encoding**: Encoding categorical features.

- **Scaling**: Standardizing the numerical features.

Model Development

Linear Regression

- **Training**: Linear Regression model trained on the scaled data.
- **Evaluation**: R^2 Score, Mean Absolute Error (MAE), Mean Squared Error (MSE).

K-Nearest Neighbors (KNN) Regressor

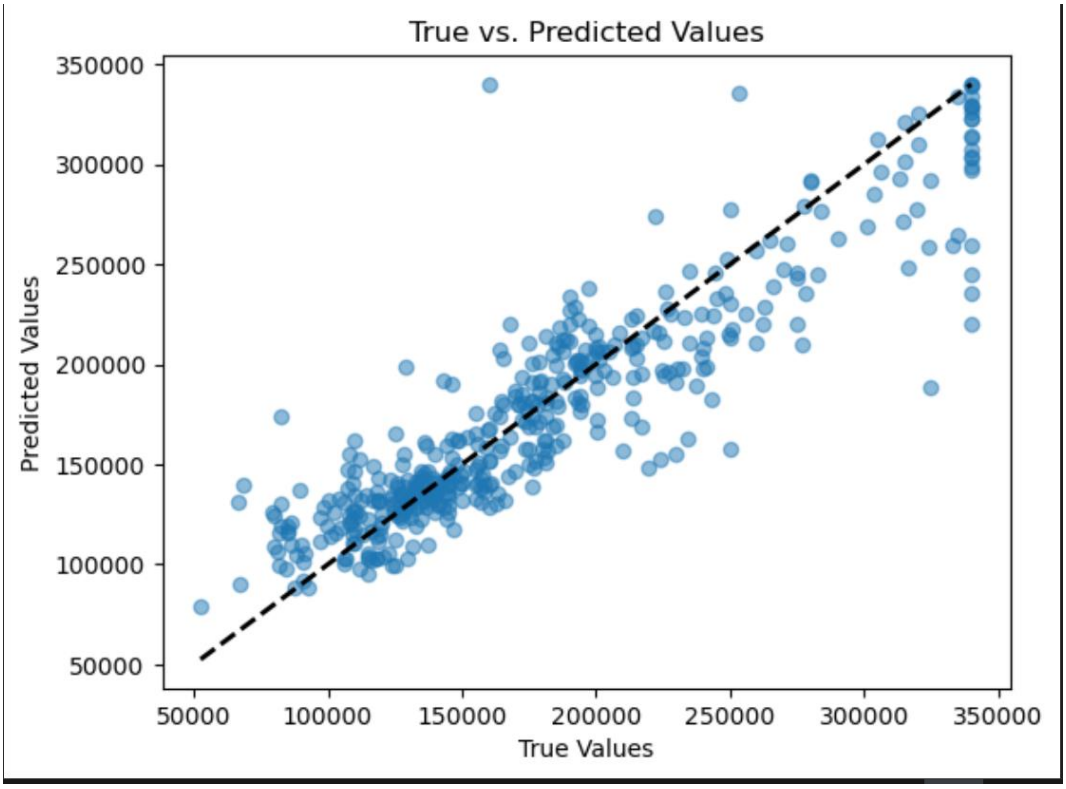
- **Training**: KNN Regressor with 5 neighbors.
- **Evaluation**: R^2 Score, Mean Absolute Error (MAE), Mean Squared Error (MSE).

Results

- **Linear Regression**:
-
    - R^2 Score: 0.862
    - MAE: 17226
    - MSE: 622627228
- **KNN Regressor**:
-
    - R^2 Score: 0.851
    - MAE: 18110
    - MSE: 673299712

Visualization

- **True vs. Predicted Values**:

- 

Linear Regression is the better algorithm in this comparison based on:

A higher R^2 Score (86.2% vs. 85.1%).

A lower MAE (17,226 vs. 18,110).

A lower MSE (622,627,228 vs. 673,299,712).

Linear Regression explains more variance in the data and produces smaller prediction errors on average, both in terms of absolute differences and squared errors.