# KNN Model Training Report

## 1. Dataset Used

The dataset used in this project is the Housing.csv file from (<https://raw.githubusercontent.com/sorif95/ML-Assignment/main/Housing.csv>) . This dataset contains various attributes related to housing properties, such as price, area, number of bedrooms, bathrooms, parking spaces, and other categorical features. The goal was to apply K-Nearest Neighbors (KNN) for both classification (binary classification of parking availability) and regression (predicting house prices).

## 2. Data Preprocessing Steps

To ensure the dataset was suitable for training the KNN models, the following preprocessing steps were performed:

### Handling Missing Values

| price | 22 |
| --- | --- |
| area | 54 |
| bedrooms | 54 |
| bathrooms | 54 |
| stories | 54 |
| parking | 54 |

# Handle missing values

- Mean imputation for price.  
- Median imputation for area and stories.  
- Mode imputation for bedrooms and parking.  
- KNN imputation for bathrooms.

### Encoding Categorical Variables

Categorical features were label-encoded using LabelEncoder.

### Feature Scaling

Standardisation was applied to numerical features using StandardScaler to normalise the dataset.

### Binary Classification Target

The parking feature was converted into a binary variable:  
- 0 for properties with 0 or 1 parking spaces.  
- 1 for properties with more than 1 parking space.

### Data Splitting

The dataset was split into training (80%) and testing (20%) sets separately for classification and regression tasks.

## 3. Model Training

## KNN Classifier (Binary Classification of Parking)

The KNN Classifier was trained using KNeighborsClassifier(n\_neighbors=5). The trained model was used to predict labels for the test set.

### KNN Regressor (Predicting House Prices)

The KNN Regressor was trained using KNeighborsRegressor(n\_neighbors=5). The trained model was used to predict house prices on the test set.

## 4. Model Performance & Evaluation

### Classification Results

Accuracy, Precision, Recall, and F1-score were used as evaluation metrics.

| Metric | Score |
| --- | --- |
| Accuracy | 0.89 |
| Precision | 0.93 |
| Recall | 0.54 |
| F1-score | 0.68 |

### Regression Results

Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² Score were used as evaluation metrics.

| Metric | Score |
| --- | --- |
| MSE | 0.49 |
| RMSE | 0.70 |
| R² Score | 0.64 |

## 5. Observations on Performance Changes

Effect of Changing K (Number of Neighbors):  
- Lower values of K resulted in more variance (overfitting).  
- Higher values of K resulted in more bias (underfitting).  
- The optimal value of K=5 was chosen as it provided a balance between bias and variance.

Impact of Feature Scaling:  
- Without standardization, KNN performed poorly because distance-based algorithms are sensitive to different feature magnitudes.

Performance with Imputation Strategies:  
- Using KNN imputation for bathrooms improved classification accuracy compared to median imputation.  
- Different imputation strategies had minor effects on regression performance.

## 6. Conclusion

The KNN classifier successfully categorized parking availability with high accuracy.  
The KNN regressor provided a reasonable approximation of house prices but is sensitive to feature selection and scaling.