Naive Bayes Model Report

1. Dataset Used

The dataset used for this analysis is the Housing Dataset, which was obtained from:

GitHub Repository:
https://raw.githubusercontent.com/sorif95/ML-Assignment/main/Housing.csv

Dataset Description

The dataset contains various features related to housing properties, including:

- price: The price of the house.

- area: The area of the house in square feet.

- bedrooms, bathrooms, stories, parking: Structural features.

- Several categorical features such as mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea, furnishingstatus.

2. Preprocessing Steps

Handling Missing Values

- Mean Imputation for price.

- Median Imputation for area and stories.

- Mode Imputation for bedrooms and parking.

- KNN Imputation for bathrooms to maintain logical relationships.

Encoding Categorical Variables

- Label Encoding was applied to categorical features to convert them into numerical values.

Feature Scaling

- StandardScaler was used to normalize numerical features for better model performance.

Feature Engineering

- A new binary classification target variable, parking_binary, was created based on the parking column.

3. Model Performance & Explanation

3.1 Naive Bayes Classification Model

Goal: Predict whether a house has parking (binary classification).

Training Approach

- Stratified K-Fold Cross-Validation (5 folds) was used.

- The Gaussian Naive Bayes classifier was trained on the data.

Results

Cross-Validation Accuracy: 0.98 ± 0.01

Test Accuracy: 0.99

Precision: 0.96

Recall: 1.00

F1-Score: 0.98

4. Performance Changes with Parameter Modifications

Regression Metrics (Naive Bayes with Discretization):

MSE: 0.73
RMSE: 0.86
MAE: 0.65
R^2 score: 0.47

Comparison with Random Forest Regressor

Naive Bayes (Discretized): R^2 Score = 0.47

Random Forest Regressor: R^2 Score = 0.89

5. Conclusion & Recommendations

- Naive Bayes performed exceptionally well for classification (99% accuracy).

- For regression, discretization helped adapt Naive Bayes, but it performed suboptimally (R^2 = 0.47).