Project Title: Car Price Classification

1. Introduction:

The automotive industry increasingly relies on data-driven strategies to assess market dynamics and customer preferences. In this project, we develop a machine learning-based system to classify car prices as either **high** or **low**, helping dealerships, platforms, and customers make smarter pricing decisions. Our solution employs advanced data preprocessing techniques and evaluates multiple classification algorithms to determine the most accurate model.

2. Objective:

To design and evaluate machine learning models capable of predicting whether a car's price is above or below the median value based on features such as brand, model, year, mileage, and condition.

3. Dataset Description:

- **Source**: Internal dataset CarPricesPrediction.csv
- **Records**: 1000 entries
- **Features**:
    - Make: Manufacturer of the car (e.g., Toyota, Ford)
    - Model: Model name (e.g., Civic, Altima)

  - Year: Manufacturing year
  - Mileage: Distance driven in miles
  - Condition: Condition of the vehicle (e.g., Excellent, Good)
  - Price: Numerical price in USD

4. Data Preprocessing:

4.1. Exploratory Data Analysis (EDA)

- Checked for missing values and data types
- Reviewed statistical summary and feature distributions

4.2. Outlier Handling

- Applied Interquartile Range (IQR) method to remove outliers in Mileage and Price

4.3. Normalization

- Used MinMaxScaler to scale Mileage and Price to a [0, 1] range

4.4. Label Encoding

- Transformed categorical columns (Make, Model, Condition) into numeric labels using Label Encoder

4.5. Target Variable

- Created a new column Price Class:
  - 0: Low (below median price)
  - 1: High (above median price)

4.6. Class Balancing

- Applied **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the binary classes

5. Model Development:

5.1. Train-Test Split

- Dataset was split into 80% training and 20% testing with stratified sampling

5.2. Models Evaluated

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **K-Nearest Neighbors (KNN)**

6. Performance Evaluation Metrics:

Each model was evaluated using:

- **Accuracy**
- **Precision, Recall, F1-Score**
- **Confusion Matrix**
- **ROC-AUC Score**
- **Matthews Correlation Coefficient (MCC)**

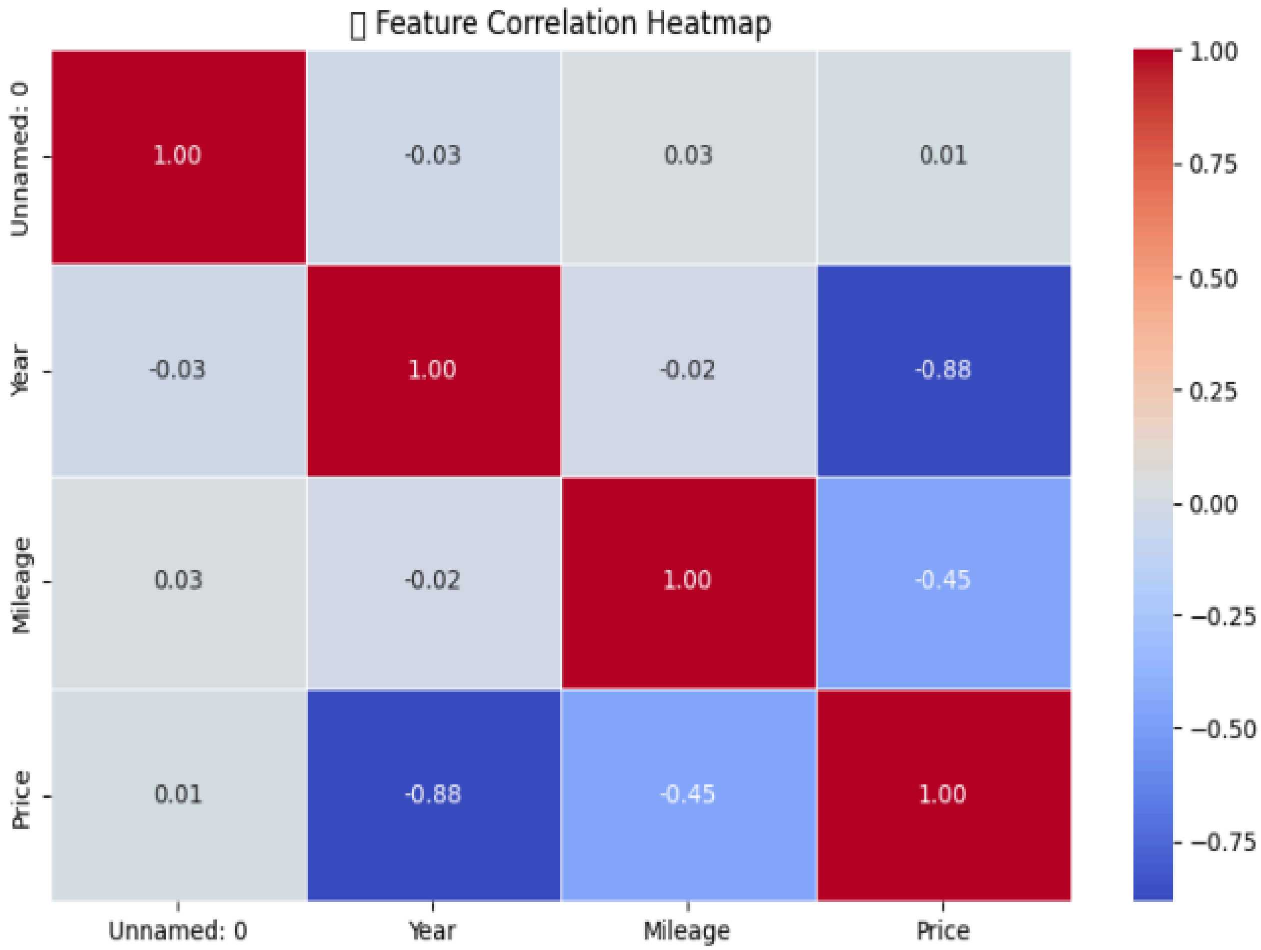
7. Results:

Model	**Test Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	100%	1.00	1.00	1.00	1.00
SVM	90%	0.90	0.90	0.90	0.50
KNN	93%	0.93	0.93	0.93	0.97

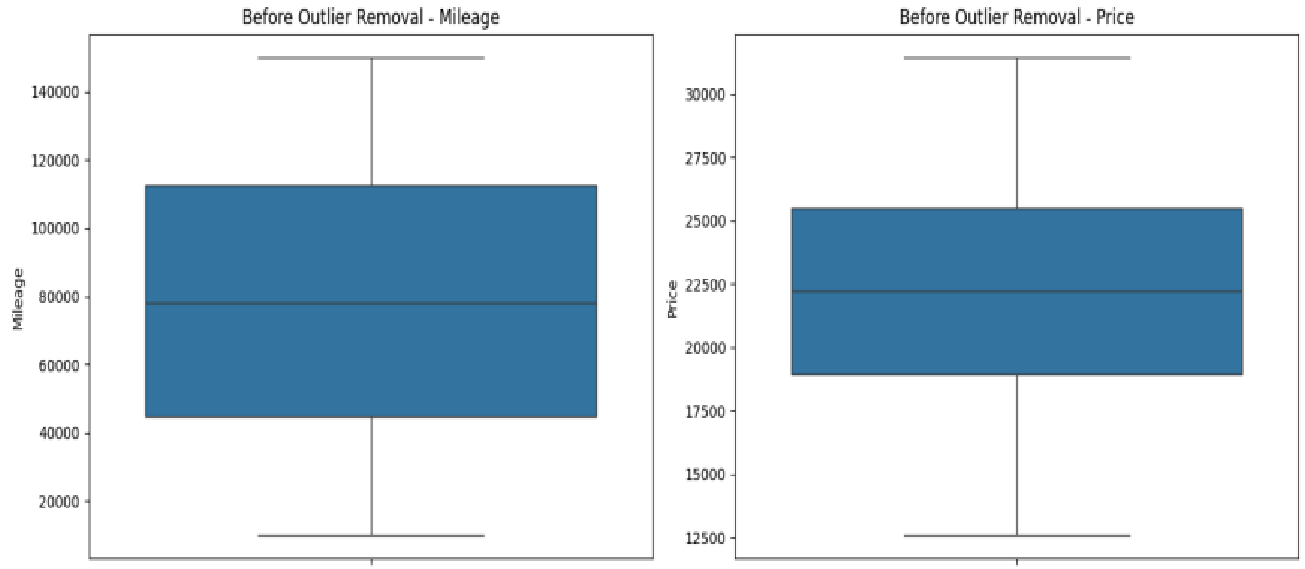
- **Best Model**: Logistic Regression with perfect classification metrics
- **KNN** showed strong performance as a non-parametric alternative
- **SVM** had a relatively lower ROC AUC due to class probability estimation limitations

Visualizations

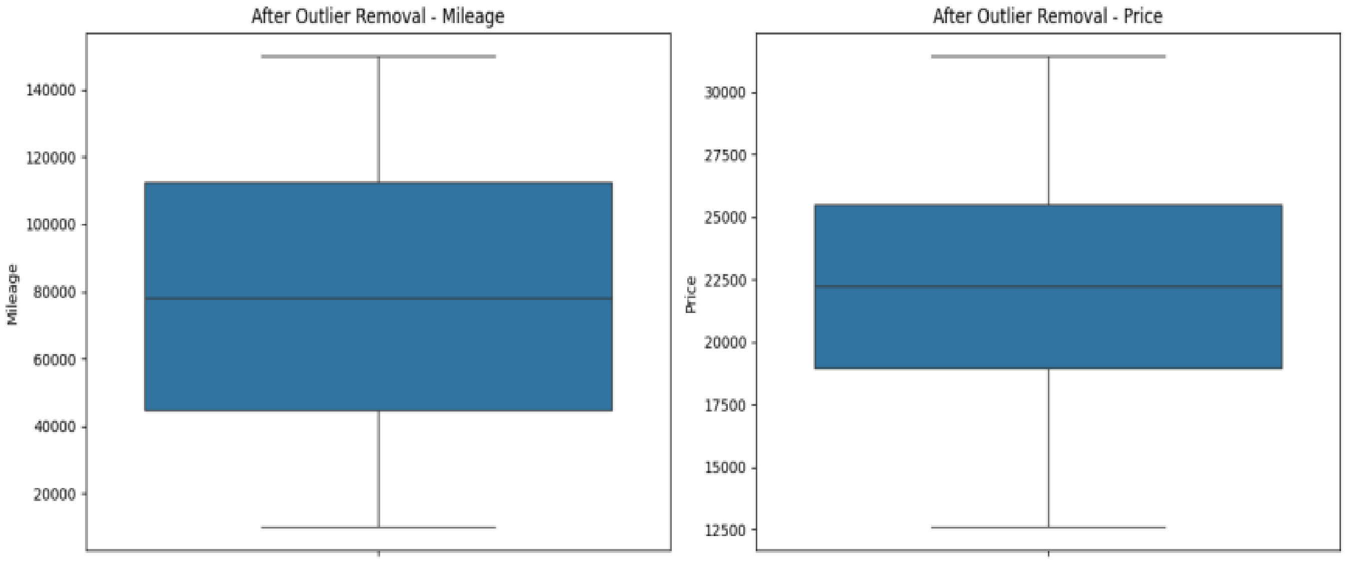
- **Correlation Heatmaps**



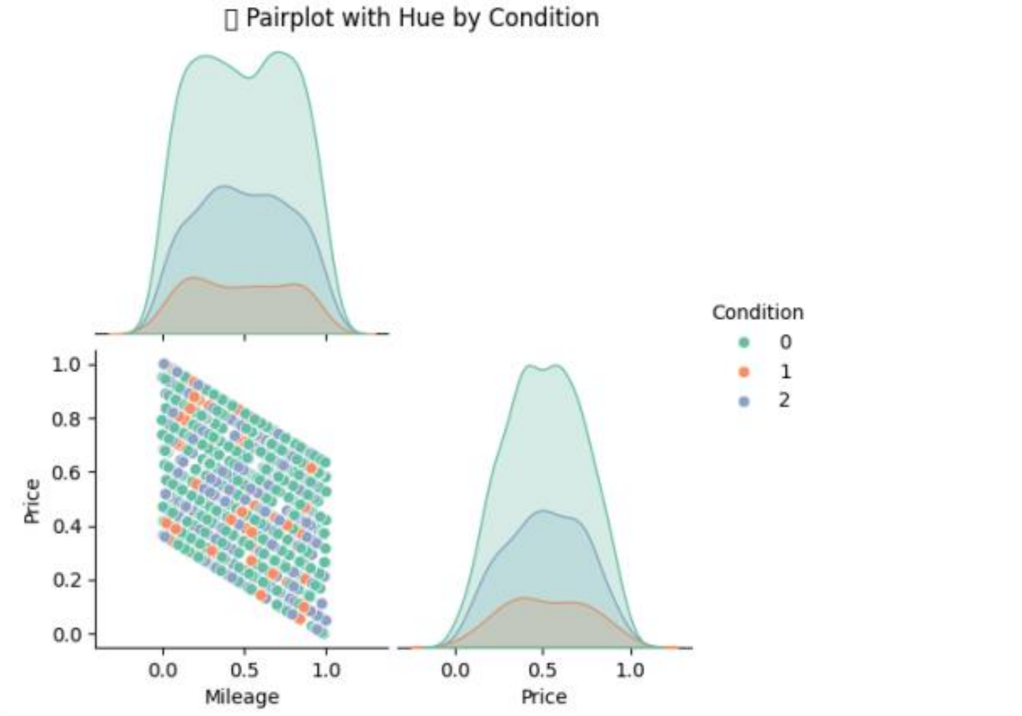
- **Boxplots (Before Outlier Removal)**



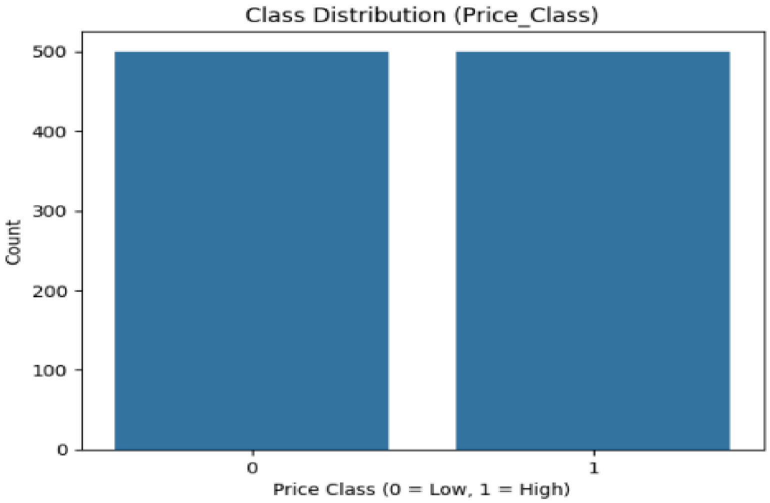
- **Boxplots (after outlier removal)**



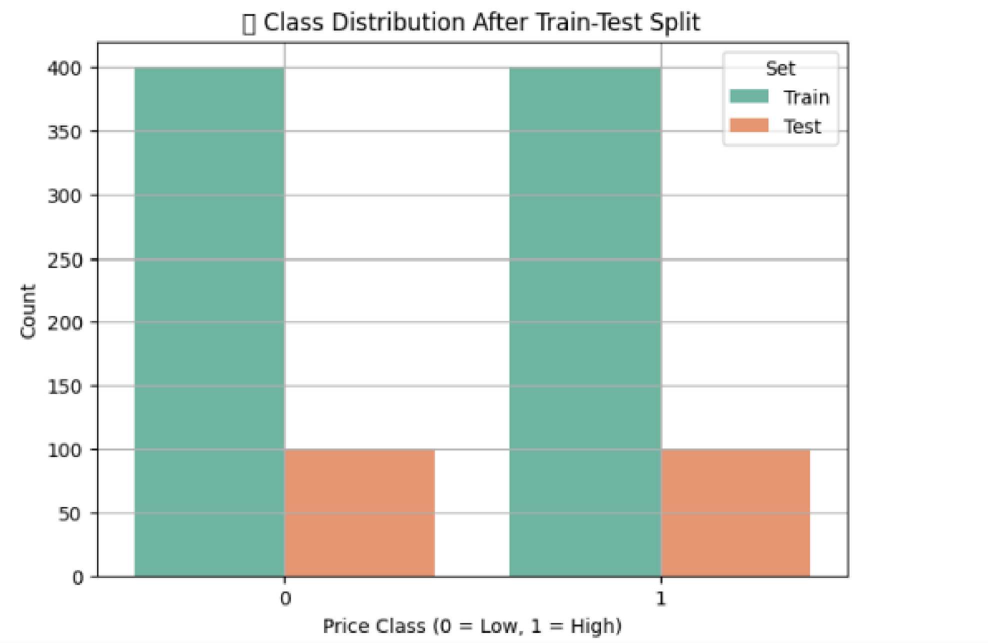
- **Pair Plot colored by Condition**



- **Class Distribution Graph (Pre-SMOTE)**

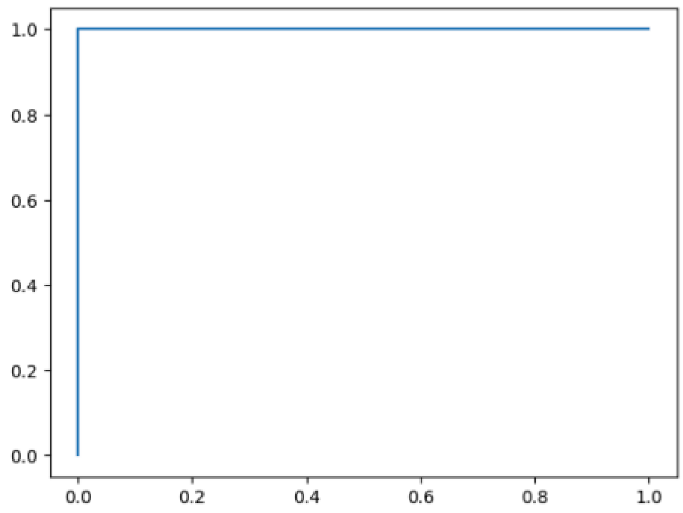


- **Class Distribution Graph (Post-SMOTE)**

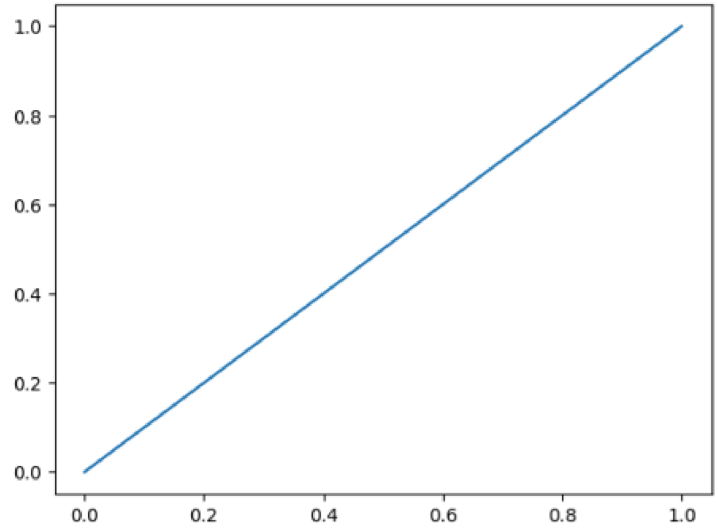


- **Confusion Matrices and ROC Curves for each model**

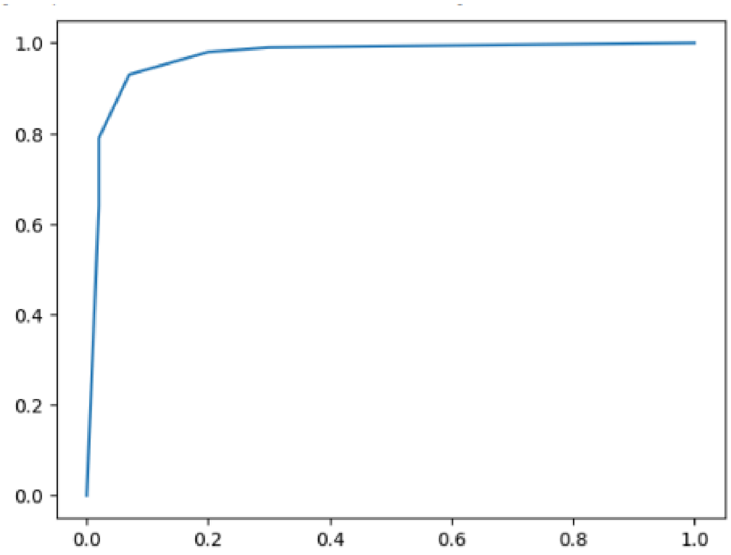
1. Logistic Regression



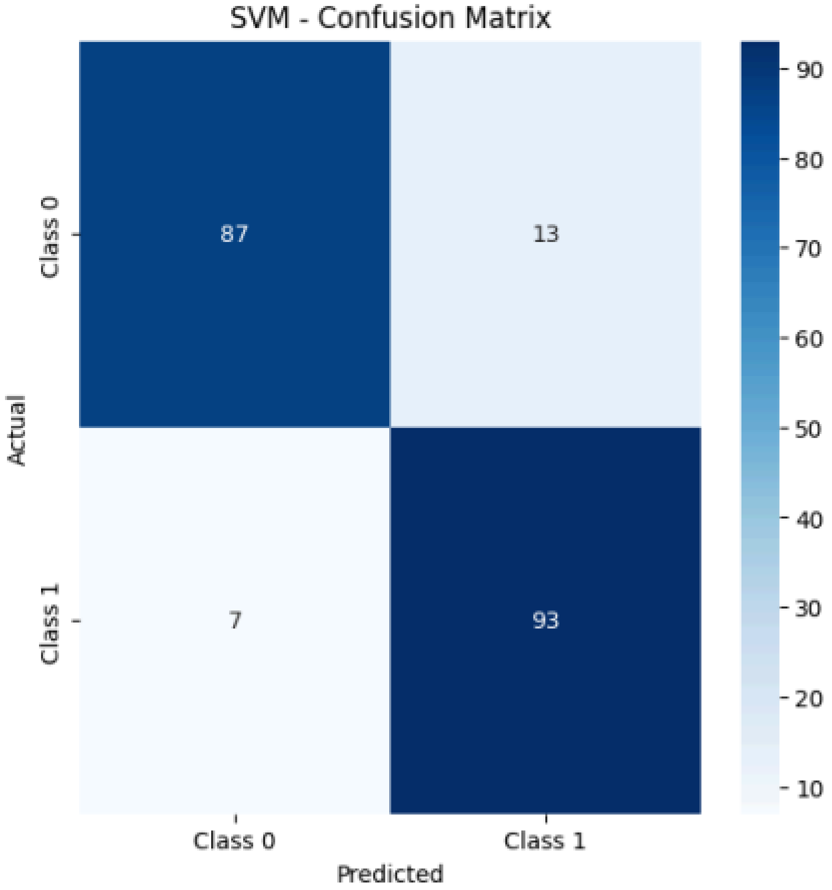
2. Support Vector Machine



3. K-Nearest Neighbors Model



4. **SVM-Confusion Matrices**



9. Conclusion

This project successfully developed a binary classification model to distinguish between high and low-priced cars with outstanding performance. Logistic Regression emerged as the top performer, achieving perfect precision and recall after robust preprocessing and balancing techniques.