PREDICTIVE ANALYTICS PROJECT REPORT

PREDICTIVE ANALYTICS INT 234

PROJECT REPORT

(Project Semester August-January 2025)

Car Price Prediction Using Machine Learning

Submitted by: Anshu Yadav

Registration No: 12307379

Programme and Section: BTech CSE K23KS

Course Code: INT234

Under the Guidance of

Dr. Madhu Bala

Discipline of CSE/IT

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Anshu Yadav** bearing Registration no. **12307379** has completed INT-234 project titled, “PREDICTIVE ANALYTICS PROJECT REPORT ” Under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

School of Computer Science & Engineering
Lovely Professional University Phagwara,
Punjab.

DECLARATION

I, **Anshu Yadav** student of B. Tech under **CSE/IT** Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:12-12-2025 Anshu Yadav

Registration No.: 12307379 Signature

1. INTRODUCTION

1.1 Overview of Predictive Analytics

Predictive Analytics is an advanced branch of data analytics that focuses on using historical data, statistical methods, and machine learning algorithms to predict future outcomes. Unlike descriptive analytics, which explains what has already happened, predictive analytics aims to forecast what is likely to happen next. This is achieved by identifying patterns, correlations, and trends in historical data.

With the rapid growth of digital data and computational power, predictive analytics has become a core component of decision-making systems across various industries such as finance, healthcare, retail, transportation, and automotive analytics. Techniques such as regression, classification, clustering, and dimensionality reduction enable organizations to make accurate, data-driven decisions.

1.2 Importance of Prediction in the Automotive Domain

The automotive industry is highly influenced by market demand, technological advancements, fuel efficiency, and customer preferences. Car pricing is affected by multiple factors including manufacturing year, mileage, engine capacity, transmission type, fuel type, and ownership history. Accurately predicting car prices is essential for:

1. Car dealers to determine competitive resale prices
2. Buyers to evaluate fair market value
3. Insurance companies to assess vehicle worth
4. Financial institutions to calculate loan values

Manual estimation or rule-based pricing methods often fail to capture complex, non-linear relationships between variables. Machine learning models provide a more robust and scalable solution by learning directly from data.

1.3 Problem Statement

Car price estimation is a challenging task due to the presence of heterogeneous data and complex dependencies among vehicle attributes. Traditional pricing techniques are subjective and prone to errors. Therefore, there is a need for an intelligent system that can analyze historical vehicle data and accurately predict selling prices.

This project aims to build a predictive analytics system using machine learning that can:

- Predict continuous car prices (regression)
- Classify cars into price categories (classification)
- Identify market segments using clustering techniques

1.4 Objectives of the Study

The main objectives of this project are:

- To analyse a real-world used car dataset from Kaggle
- To preprocess and clean the dataset for machine learning
- To perform exploratory data analysis and feature correlation
- To develop multiple regression and classification models
- To compare models using standard evaluation metrics
- To identify the most accurate predictive model
- To explore unsupervised learning for vehicle segmentation

1.5 Scope of the Project

The scope of the project includes:

- Predictive modelling using supervised learning
- Comparative analysis of multiple ML algorithms
- Clustering and dimensionality reduction techniques
- Offline analysis using historical data

The project does not include real-time data streaming or commercial deployment.

1.6 Expected Outcomes

The expected outcomes of this project are:

- Accurate prediction of used car prices
- Identification of key features influencing car value
- Selection of the best machine learning model
- Market segmentation insights for automotive analysis

2. SOURCE OF DATASET

2.1 Name of Dataset

Vehicle Dataset from Car Dekho (Car details v3.csv)

2.2 Source of Dataset

The dataset was collected from Kaggle, a publicly available data science platform.

DatasetLink:
https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?utm_source=chatgpt.com

2.3 Description of Attributes

The dataset contains 13 attributes, categorized as follows:

Numerical Attributes:

- year – Year of manufacture
- selling_price – Selling price of the car (target variable)
- km_driven – Total kilometers driven
- mileage – Mileage of the car

- engine – Engine capacity
- max_power – Maximum engine power
- torque – Torque output
- seats – Seating capacity

Categorical Attributes:

- name – Car model name
- fuel – Fuel type (Petrol/Diesel/CNG)
- seller_type – Individual or Dealer
- transmission – Manual or Automatic
- owner – Ownership history

2.4 Dataset Size

- Total Records: 8,128
- Total Features: 13

2.5 Type of Prediction

- Regression – Continuous car price prediction
- Classification – Binary price category prediction

3. DATASET PREPROCESSING

Data preprocessing is a critical step in predictive analytics. Raw datasets often contain missing values, inconsistencies, and irrelevant information that can negatively impact model performance.

The following preprocessing steps were applied:

3.1 Data Cleaning

- Removal of duplicate records
- Handling inconsistent formatting

3.2 Feature Extraction

- Extracted numerical values from string-based features such as mileage, engine, and power using regular expressions

3.3 Handling Missing Values

- Numerical features filled using median imputation
- Categorical features filled using mode imputation

3.4 Encoding

- Label Encoding applied to categorical variables

3.5 Feature Scaling

- StandardScaler used to normalize features for distance-based algorithms

3.6 Train-Test Split

- Dataset split into training and testing sets

4. DATASET ANALYSIS

4.1 Exploratory Data Analysis (EDA)

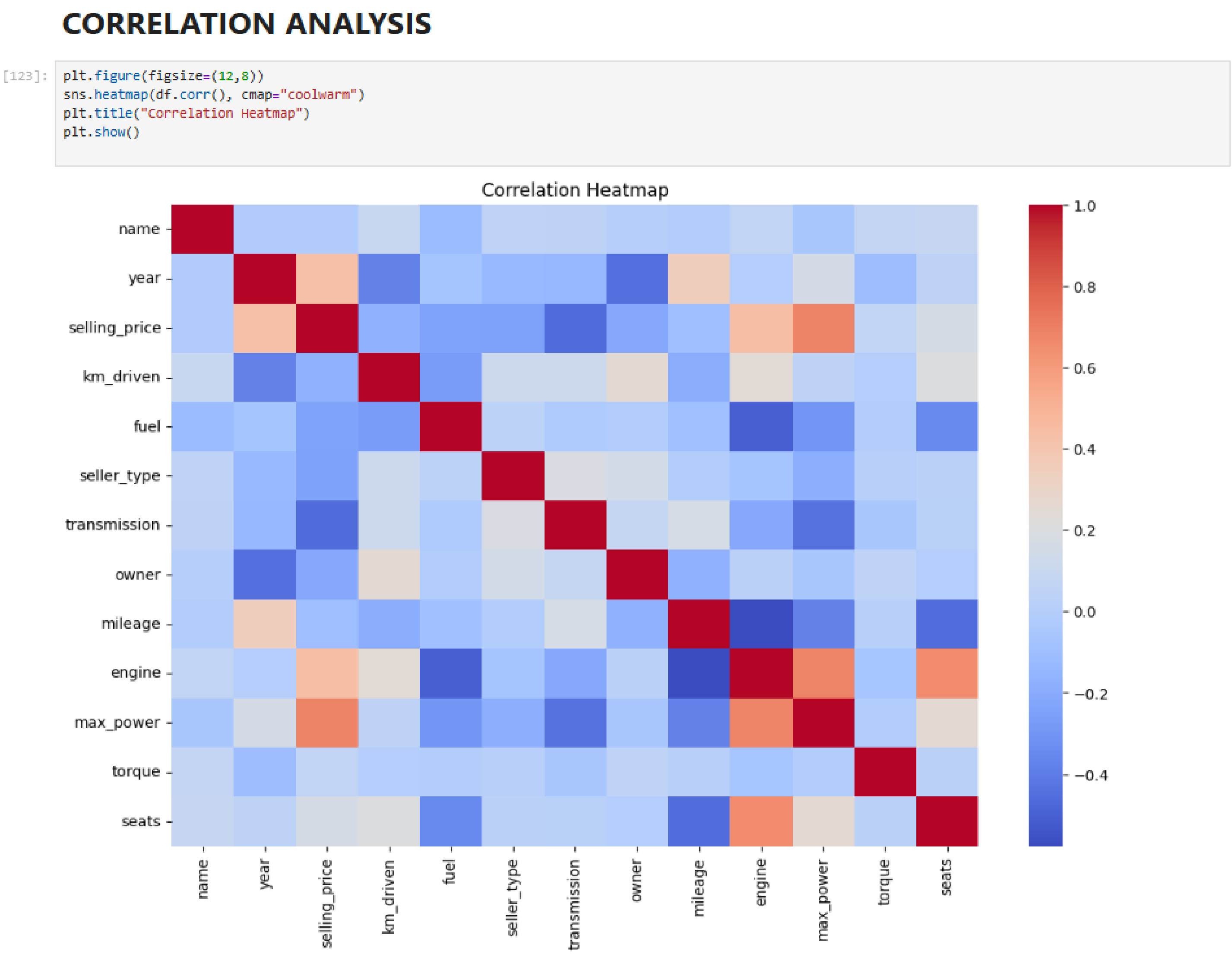
EDA was performed to understand:

- Distribution of features
- Presence of outliers
- Relationships between variables

4.2 Correlation Analysis

Correlation analysis revealed that:

- Selling price has a strong positive correlation with year, engine size, and max power



CORRELATION ANALYSIS

```
[123]: plt.figure(figsize=(12,8))
       sns.heatmap(df.corr(), cmap="coolwarm")
       plt.title("Correlation Heatmap")
       plt.show()
```

- Kilometres driven negatively impacts selling price

A correlation heatmap was used to visualize these relationships.

4.3 Visualization

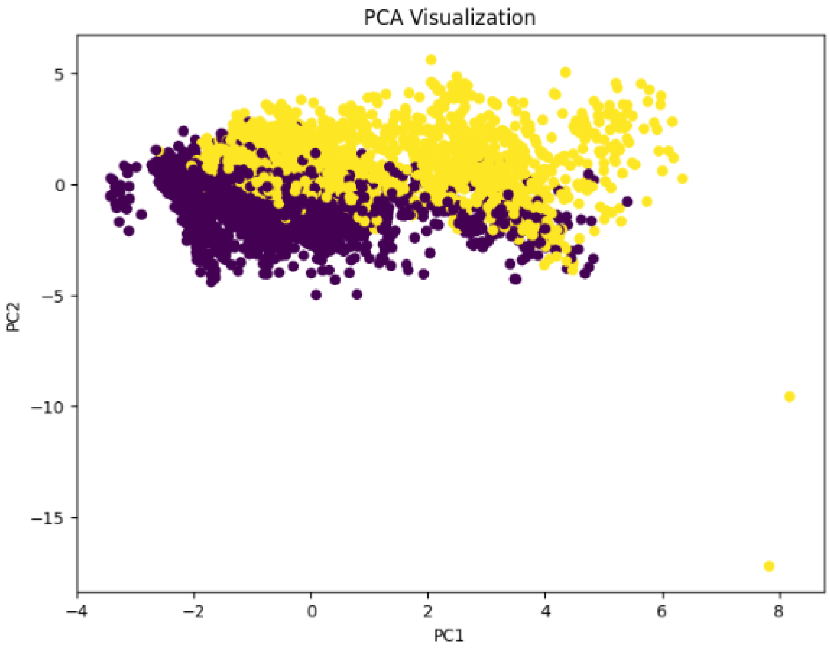
The following visualizations were created:

- Histograms
- Box plots
- Correlation heatmaps
- PCA scatter plots

PCA VISUALIZATION

```
[140]: pca = PCA(n_components=2)
       X_pca = pca.fit_transform(X_scaled)

       plt.figure(figsize=(8,6))
       plt.scatter(X_pca[:,0], X_pca[:,1], c=y_clf, cmap='viridis', s=30)
       plt.title("PCA Visualization")
       plt.xlabel("PC1")
       plt.ylabel("PC2")
       plt.show()
```



5. MODEL DEVELOPMENT

5.1 Regression Models

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

5.2 Classification Models

The selling price was converted into binary categories using the median price.

- Logistic Regression
- K-Nearest Neighbours
- Naive Bayes
- Decision Tree
- Support Vector Machine
- Random Forest
- Multi-Layer Perceptron

5.3 Unsupervised Learning

- K-Means Clustering
- Hierarchical Clustering
- Principal Component Analysis (PCA)

Model	**MAE**	**RMSE**	**R²**
Simple Linear	218,024	414,326	0.217
Multiple Linear	169,930	303,190	0.581
Polynomial	—	208,531	—

Classification Performance

Model	**Accuracy (%)**
Logistic Regression	86.7
KNN	88.5
Naive Bayes	76.6
Decision Tree	89.5
SVM	89.9
Random Forest	**91.9**

7. BEST MODEL IDENTIFICATION

- **Multiple Linear Regression** performed best for price prediction
- **Random Forest** achieved the highest classification accuracy

8. CONCLUSION

This project successfully demonstrates how predictive analytics and machine learning can be applied to real-world automotive pricing problems. The models developed provided accurate predictions and valuable insights into pricing behaviour.

9. FUTURE SCOPE

- Deep learning models
- Integration of real-time market data
- Explainable AI techniques
- Web-based deployment

10. Link:

- LinkedIn: https://www.linkedin.com/posts/anshu-yadav-1a7b8b290_predictiveanalytics-machinelearning-datascience-activity-7406318154957815809-u4E7?utm_source=social_share_send&utm_medium=member_desktop_web&rcm=ACoAAEanV08BkGxUJEvFhB468lil0zowfE5IM3U