# CSC 261

Artificial Intelligence Programming Languages

Project 1 Report

**Ibrahim Binnafisah – 444101505**

**Overview**

• This project analyzes a real used-car dataset.
• The objective is to predict the selling price of cars using **Linear Regression**.
• Additionally, the project classifies cars into **Automatic or Manual transmission** using **Logistic Regression**.
• The complete machine-learning pipeline was followed:

– Exploratory Data Analysis (EDA)
– Data Cleaning & Preprocessing
– Feature Engineering
– Model Training
– Model Evaluation
– K-Fold Cross Validation

**Design Choices**

• Numerical and categorical features were separated for correct preprocessing.
• A new engineered feature, **Vehicle Age = 2025 – Year**, was added.
• **One-Hot Encoding** was used for categorical variables such as Fuel Type and Seller Type.
• **StandardScaler** was used to normalize numerical features.
• Linear Regression was chosen for price prediction because it models continuous values efficiently.
• Logistic Regression was chosen for

**Experiment Setup**

- Programming Language: **Python 3.11**
- Libraries Used: Pandas, NumPy, Scikit-Learn, Matplotlib
- Development Environment: **Visual Studio Code**
- Hardware Used:
  - Windows 10
  - Intel i7 CPU
  - 16 GB RAM
- Dataset contains features such as:
  - Selling Price
  - Present Price
  - Kms Driven
  - Fuel Type
  - Seller Type
  - Transmission
  - Owner count

transmission prediction because it outputs probabilistic binary classes.

- Evaluation metrics used:
  - Regression → MAE, MSE, RMSE, $R^2$
  - Classification → Accuracy, Precision, Recall, Confusion Matrix

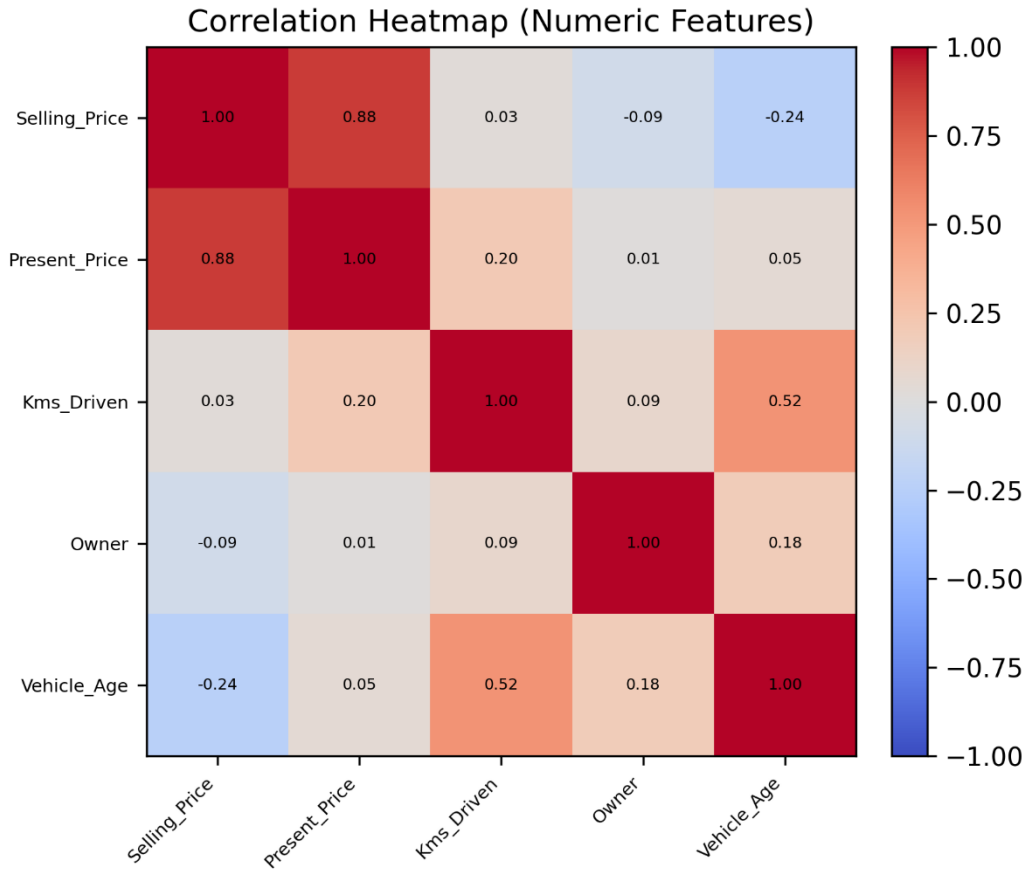
# Results – Linear Regression


## Important Features

- Present Price
- Vehicle Age
- Kms Driven

## Metrics

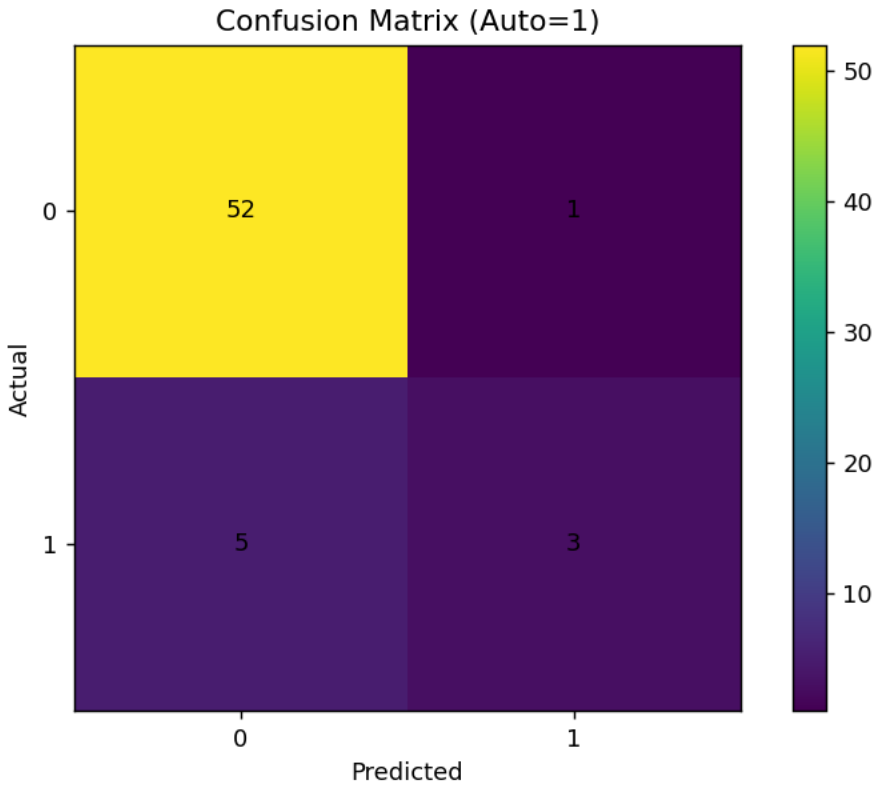
| Metric | Value |
|---|---|
| MAE | 2.0349434490300293 |
| MSE | 9.22566364119091 |
| RMSE | 3.037377574070218 |
| $R^2$ Score | 0.5995038184047492 |



## Results – Logistic Regression

### Metrics

| Metric | Value |
|---|---|
| Accuracy | 0.8852459016393442 |
| Precision (Auto) | 0.6 |
| Recall (Auto) | 0.375 |



**Interpretation**

• High precision means the model is usually correct when predicting “Automatic”.
• High recall means the model successfully detects most automatic cars.
• Coefficient analysis shows which features increase or reduce the probability of being automatic.

**Final Notes**

• Linear Regression performed well for estimating continuous values like car prices.
• Logistic Regression showed good performance for binary classification tasks.
• Data quality strongly affects model accuracy; additional features such as engine specs or maintenance history could improve predictions.
• K-Fold Cross Validation reduced variance and increased reliability.
• All coding, analysis, and report writing were completed independently by the student.