**Car Price Prediction and Classification Using Machine Learning**

# Week-02 Task: Sapience Edu Connect Pvt Ltd.

**Name**: Muhammad Yasir

Data Science Internship

**Duration**: 15-june-2025 To 14-july-2025

**Gmail:** yaisikhan111@gmail.com

**Phone No:** +92 03400462508

# 📌 Objective

To predict car prices and classify them as High Price or Low Price using machine learning models such as Linear Regression, Logistic Regression, Decision Tree, and Random Forest, with proper evaluation and tuning.

# 📂 Dataset Overview

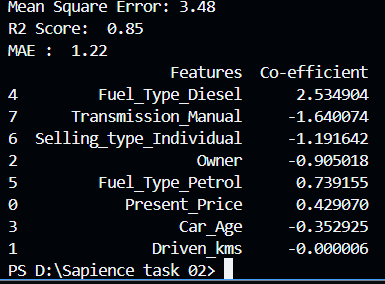
The dataset includes details of used cars with features such as:  
- Year: Manufacturing year  
- Present\_Price: Price when new  
- Selling\_Price: Current market price  
- Driven\_kms: Kilometers driven  
- Owner: Number of previous owners  
- Fuel\_Type, Transmission, Selling\_type

# 🧼 Data Preprocessing

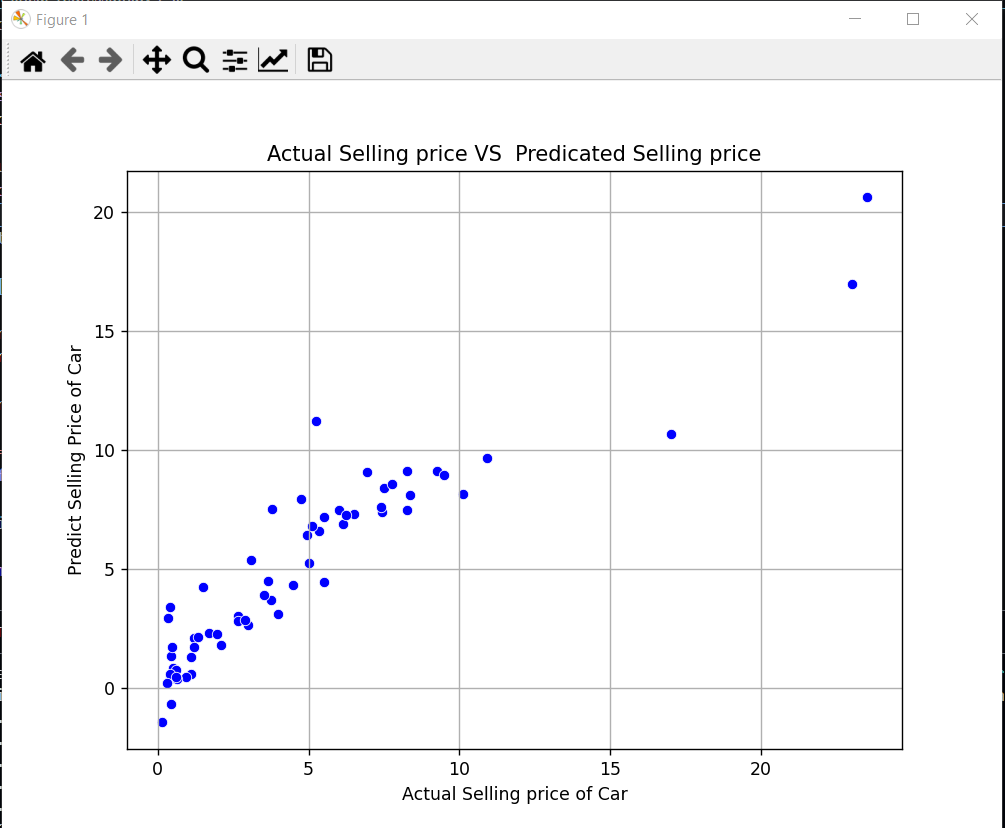
- Converted Year to Car\_Age  
- Encoded categorical variables using one-hot encoding  
- Removed original string columns like Car\_Name  
- Ensured all features are numeric for modeling

# 📈 Model 1: Linear Regression

- Goal: Predict the continuous Selling\_Price  
- Evaluation Metrics:  
 - R² Score  
 - Mean Absolute Error (MAE)  
 - Mean Squared Error (MSE)



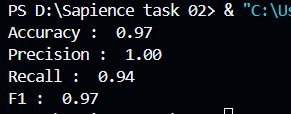
**Figure 1.0:** Linear Regression Model Feature Importance



# Figure 1.1: Linear Regression Model Graph

# 🧪 Model 2: Logistic Regression

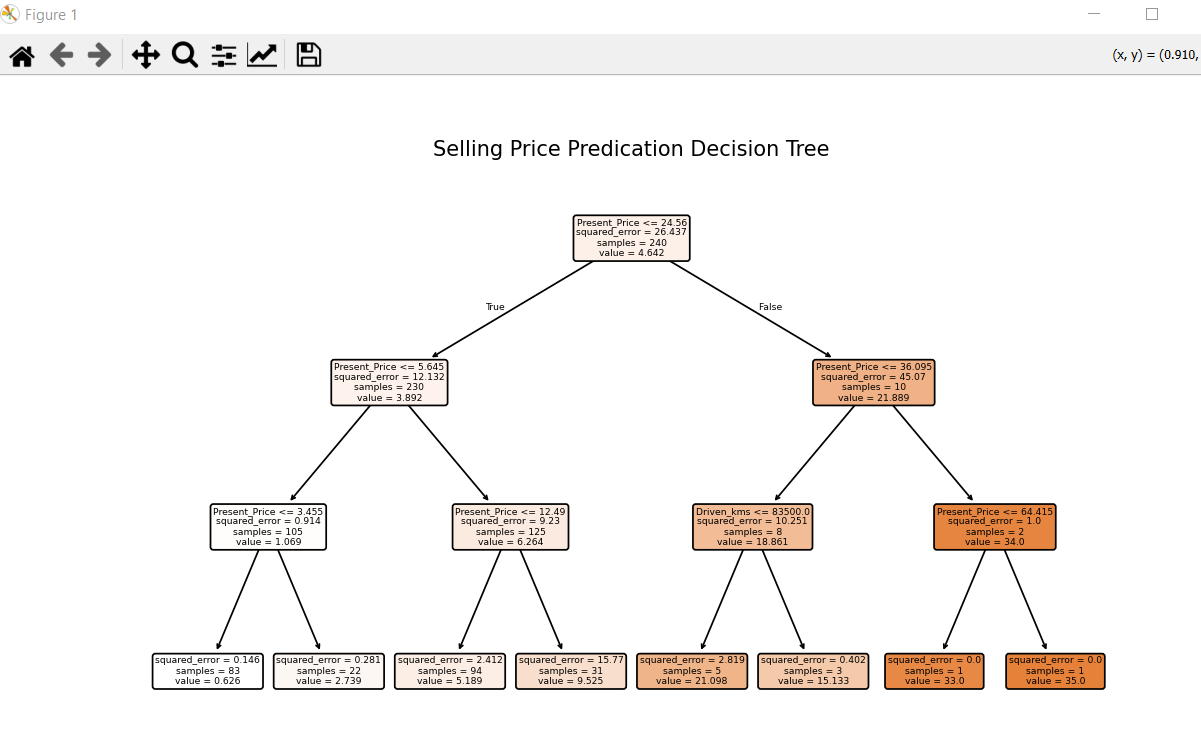
- Goal: Classify cars into High Price or Low Price  
- Created binary target using median price threshold  
- Evaluation Metrics:  
 - Accuracy  
 - Precision  
 - Recall  
 - F1 Score



# Figure 2.0: Logistic Regression

# 🌳 Model 3: Decision Tree Regressor

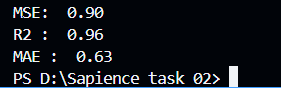
- Applied regression   
- Visualized the decision tree for interpretability  
- Tuned using:  
 - max\_depth  
 - min\_samples\_split  
 - min\_samples\_leaf



# 

# 🌲 Model 4: Random Forest Regressor

- Built ensemble of decision trees to improve accuracy  
- Tuned using:  
 - n\_estimators, max\_depth, min\_samples\_split  
- Outperformed all previous models

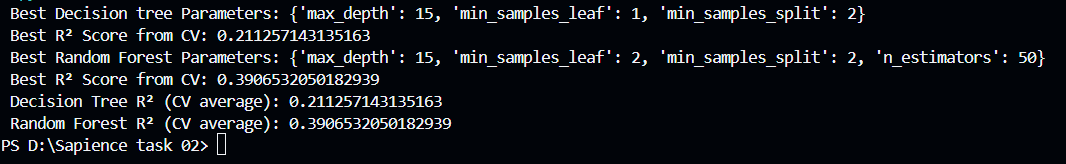


# 🔁 Model Evaluation Summary

| Model | Type | R² | MAE | Accu | F1 Score |   
|-------------------------|------------------|--------|-------|----------|----------|  
| Linear Regression | Regression | 0.85 | 1.22 | — | — |  
| Logistic Regression| Classification| — | — | 0.97 | 0.97 |  
| Decision Tree | Regression | 0.95 | 0.70 | \_\_\_ | \_\_\_ |  
| Random Forest | Regression | 0.96 | 0.63 | \_\_\_ | \_\_\_ |

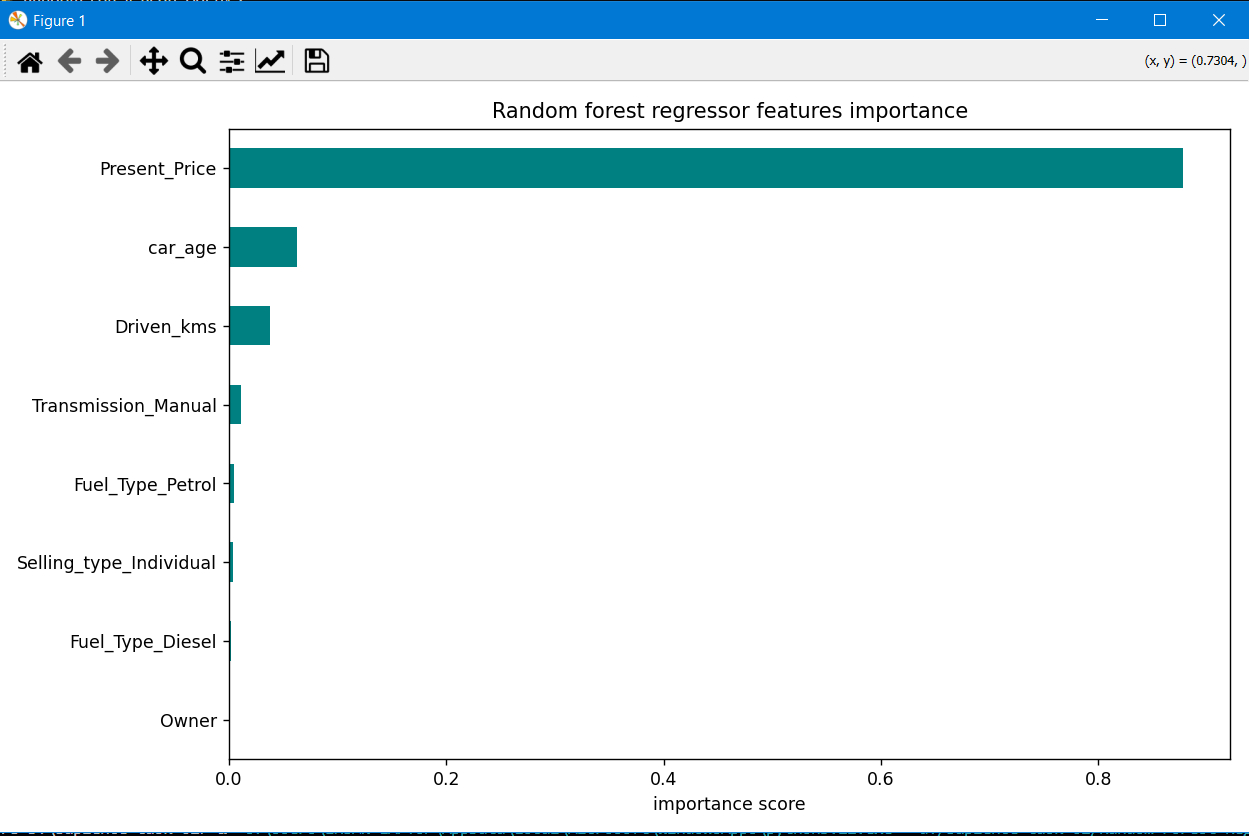
# ⚙️ Hyperparameter Tuning & Cross-Validation

- Used GridSearchCV for:  
 - Decision Tree Regressor  
 - Random Forest Regressor  
- Used scoring='neg\_mean\_absolute\_error' and scoring='r2'  
- Applied 5-Fold Cross-Validation for reliable model evaluation



# 🔍 Feature Importance (Random Forest Regressor and Decision tree Regressor)

Most impactful features:  
- Present\_Price  
- Car\_Age  
- Fuel\_Type\_Diesel  
- Transmission\_Manual



# 

# ✅ Conclusion

- Random Forest performed best overall on both regression and classification tasks.  
- Feature importance showed that Present\_Price and Car\_Age heavily influence the final selling price.  
- Cross-validation and hyperparameter tuning significantly improved performance and reduced overfitting.