

# **PREDICTIVE ANALYTICS PROJECT REPORT**

(Project Semester August-December 2024)

## ***Used Car Price Prediction and Model Comparison***

Submitted by

Abhishek Kumar

Registration No: 12201101

Bachelor of Technology in Computer Science and Engineering  
INT-234

Under the Guidance of

**Savleen Kaur (18306)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**



**L** OVELY  
**P** ROFESSIONAL  
**U** NIVERSITY

---

## **CERTIFICATE**

This is to certify that Abhishek Kumar bearing Registration no. 12201101 has completed Predictive Analytics (Int-234) project titled, “Used Car Price Prediction and Model Comparison” 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.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date:

## **DECLARATION**

I, Abhishek Kumar student of Bachelor of Technology 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: 10-11-2024

Signature

Registration No.12201101

Abhishek Kumar

## CONTENTS OF THE REPORT

Section Number	Section Title	Description
1	<b>Introduction</b>	Brief overview of the project, objectives, and purpose of the analysis.
2	<b>Scope of the Analysis</b>	Defines the extent of the analysis, objectives, and the specific outcomes aimed for in the project.
3	<b>Existing System</b>	Describes current methods or systems in place for similar analysis.
	- Drawbacks/Limitations	Discusses any limitations or drawbacks of the existing system.
4	<b>Source of Dataset</b>	Information on the dataset used, including source, relevant details, and potential biases or limitations.
5	<b>ETL Process</b>	Details on data Extraction, Transformation, and Loading processes, including preprocessing steps.
6	<b>Analysis on Dataset</b>	Detailed analysis procedures and methodology.
	- Introduction	Brief introduction to the dataset analysis and overall approach.
	- General Description	Describes dataset characteristics, data types, and summary statistics.
	- Specific Requirements	Lists functions, formulas, and specific requirements used for analysis.
	- Analysis Results	Summarizes findings and key insights obtained from the analysis.
	- Visualization	Includes charts, graphs, and visual representations of the analysis results.
7	<b>List of Analysis with Results</b>	Lists the various analyses conducted with brief explanations and results.
8	<b>Future Scope</b>	Discusses potential future improvements or extensions to the project.

# 1. Introduction

The used car market represents a significant segment of the global automotive industry. With an increasing number of buyers and sellers, determining an accurate price for used cars remains a complex task. Traditionally, car prices have been estimated using price guides, dealership appraisals, and manual evaluations. However, these methods do not always take into account the multitude of influencing factors such as market trends, vehicle condition, and economic conditions.

Machine learning, however, offers an opportunity to build predictive models that can incorporate multiple variables and learn complex relationships between the features and car prices. In this project, I have implemented machine learning algorithms to predict the prices of used cars based on various features.

**The project objectives include:**

**Data Preprocessing:** Cleaning the used car sales dataset, handling missing values, and transforming categorical features.

**Model Development:** Implementing four machine learning models:

- **Linear Regression**

Linear Regression is a model that predicts a value based on a straight-line relationship between two or more factors. For example, it can predict a car's price based on its mileage, age, and engine power. It's useful for estimating or understanding how one factor affects another.

- **Decision Tree**

A Decision Tree is a model that makes decisions by following a series of yes/no questions, splitting the data at each question. It sorts the data into groups at each step until it reaches a final prediction. Decision Trees are simple and easy to understand because they look like flowcharts.

- **K-Nearest Neighbors (KNN)**

K-Nearest Neighbors is a model that makes predictions based on the most similar past examples. For example, if we want to predict the price of a car, it will look at prices of the nearest cars with similar characteristics and take an average. This model is straightforward but can be slow with a lot of data.

- **Support Vector Machine (SVM)**

Support Vector Machine is a model that finds the best dividing line (or surface) to separate data into categories. It tries to place this line in a way that leaves the largest possible space on either side, separating the groups as clearly as possible. It's commonly used for complex tasks like sorting images or texts into categories.

## 2. Scope of the Analysis

This project aims to develop a machine learning-based system to predict used car prices with high accuracy. The scope of the analysis covers the following stages:

- **Data Collection:** The dataset used in this project contains a variety of features including mileage, engine size, power, fuel type, and car price. The data was sourced from a publicly available used car sales database.
- **Data Preprocessing:** Cleaning the dataset by handling missing data, transforming categorical variables into numeric format, and scaling the numerical features for consistency.
- **Model Development:** Four machine learning algorithms (Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)) were applied to predict the target variable: car price.
- **Dashboard Development:** An interactive Shiny dashboard was created to allow users to input vehicle features and predict the corresponding price, with real-time model comparisons.
- **Model Evaluation:** The models were evaluated based on their prediction accuracy, using MAE to compare the performance of each model.

## 3. Existing System

Traditionally, used car price predictions have been made through manual evaluations, dealership-based price guides, and rule-based valuation systems. These methods have several key limitations:

- **Manual Evaluation:** Requires human intervention, which is time-consuming and inconsistent.
- **Inaccuracy:** These methods cannot account for the complex, dynamic relationships between car features (e.g., mileage, engine size, and fuel type).
- **Lack of Real-Time Prediction:** Existing systems do not offer the flexibility to provide real-time, data-driven predictions based on changing market conditions.

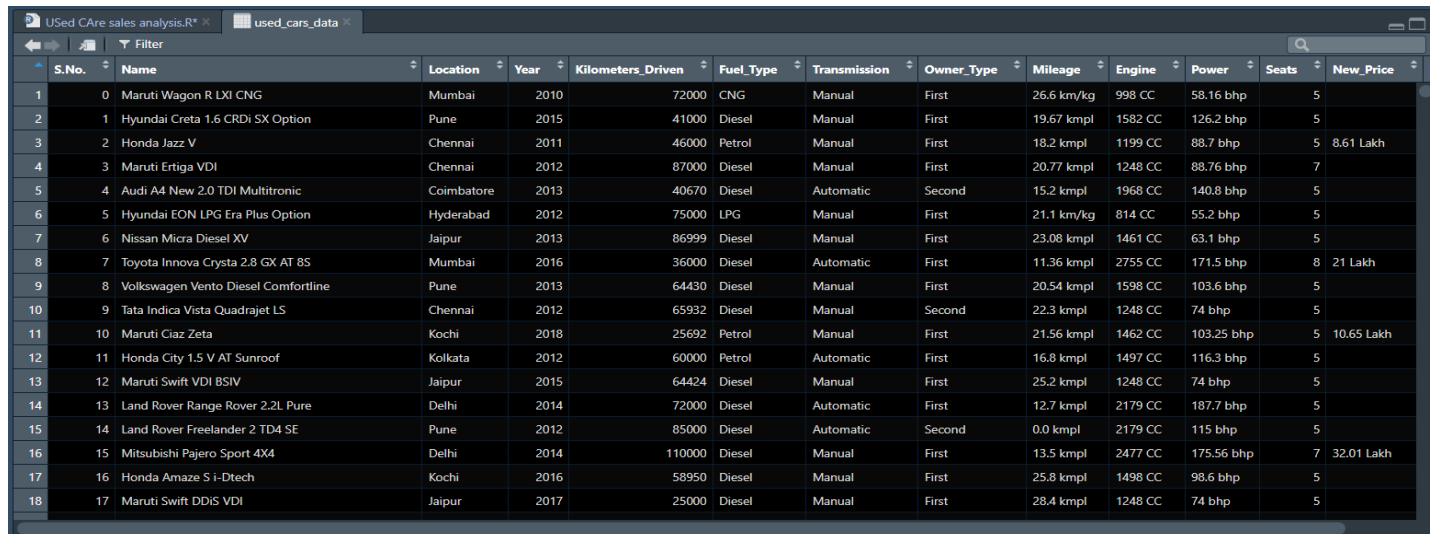
### Drawbacks or Limitations of the Existing System:

- **Limited Scope:** Traditional methods often ignore variables like car condition, market trends, and geographical influences, which can significantly affect car prices.
- **Time-Consuming:** Manually evaluating a used car's price involves time and expertise, often delaying the transaction process.
- **Error-Prone:** Traditional systems depend on human input, which is prone to errors and inconsistencies.

## 4. Source of Dataset

The dataset used for this project is sourced from Kaggle and consists of historical sales data for used cars. It includes the following columns:

- **Mileage:** Total kilometers driven by the car.
- **Engine:** The engine capacity of the car in cubic centimeters (CC).
- **Power:** The power output of the engine in brake horsepower (bhp).
- **Fuel\_Type:** The fuel type (Petrol, Diesel, etc.) used by the car.
- **Price:** The selling price of the car (target variable).



S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price
1	0 Maruti Wagon R LXi CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	
2	1 Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	
3	2 Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	8.61 Lakh
4	3 Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	
5	4 Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	
6	5 Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5	
7	6 Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	
8	7 Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	21 Lakh
9	8 Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5	
10	9 Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.3 kmpl	1248 CC	74 bhp	5	
11	10 Maruti Ciaz Zeta	Kochi	2018	25692	Petrol	Manual	First	21.56 kmpl	1462 CC	103.25 bhp	5	10.65 Lakh
12	11 Honda City 1.5 V AT Sunroof	Kolkata	2012	60000	Petrol	Automatic	First	16.8 kmpl	1497 CC	116.3 bhp	5	
13	12 Maruti Swift VDI BSIV	Jaipur	2015	64424	Diesel	Manual	First	25.2 kmpl	1248 CC	74 bhp	5	
14	13 Land Rover Range Rover 2.2L Pure	Delhi	2014	72000	Diesel	Automatic	First	12.7 kmpl	2179 CC	187.7 bhp	5	
15	14 Land Rover Freelander 2 TD4 SE	Pune	2012	85000	Diesel	Automatic	Second	0.0 kmpl	2179 CC	115 bhp	5	
16	15 Mitsubishi Pajero Sport 4X4	Delhi	2014	110000	Diesel	Manual	First	13.5 kmpl	2477 CC	175.56 bhp	7	32.01 Lakh
17	16 Honda Amaze S i-Dtech	Kochi	2016	58950	Diesel	Manual	First	25.8 kmpl	1498 CC	98.6 bhp	5	
18	17 Maruti Swift DDiS VDI	Jaipur	2017	25000	Diesel	Manual	First	28.4 kmpl	1248 CC	74 bhp	5	

## 5. ETL Process

The ETL (Extract, Transform, Load) process followed to prepare the dataset involved several stages:

### Extract:

- The raw dataset was extracted from a CSV file and imported into R for analysis.

### Transform:

- **Cleaning the Data:** Irrelevant columns, such as the car's name, were removed. Missing values were handled through imputation or removal.
- **Handling Categorical Data:** Categorical features, such as **Fuel\_Type**, were transformed into numeric values using **one-hot encoding** or **factor encoding**.
- **Scaling Numerical Data:** Columns such as **Mileage**, **Engine**, and **Power** were normalized to a consistent range to ensure proper model performance.

### Load:

The cleaned and transformed dataset was split into a training (80%) and testing (20%) set, ensuring a fair evaluation of the models.

## 6. Analysis on Dataset

### i. Introduction

The main goal of this analysis is to predict the price of a used car based on its features. The dataset includes several independent variables that may affect the car's price. I applied four machine learning algorithms: Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) to build predictive models.

```
### Linear Regression
lm_model <- lm(Price ~ ., data = train)
lm_preds <- predict(lm_model, test)
results <- rbind(results, data.frame(Model = "Linear Regression", MAE = mean(abs(test$Price - lm_preds))))

# Plot for Linear Regression
ggplot(data.frame(Actual = test$Price, Predicted = lm_preds), aes(x = Actual, y = Predicted)) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "Linear Regression: Predicted vs Actual", x = "Actual Price", y = "Predicted Price") +
  theme_minimal()

### Decision Tree
tree_model <- rpart(Price ~ ., data = train)
tree_preds <- predict(tree_model, test)
results <- rbind(results, data.frame(Model = "Decision Tree", MAE = mean(abs(test$Price - tree_preds))))

# Plot for Decision Tree
ggplot(data.frame(Actual = test$Price, Predicted = tree_preds), aes(x = Actual, y = Predicted)) +
  geom_point(color = "green") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "Decision Tree: Predicted vs Actual", x = "Actual Price", y = "Predicted Price") +
  theme_minimal()

### K-Nearest Neighbors - only numeric columns scaled
numeric_train_x <- scale(select_if(train, is.numeric) %>% select(-Price))
numeric_test_x <- scale(select_if(test, is.numeric) %>% select(-Price))

knn_preds <- knn(train = numeric_train_x, test = numeric_test_x, cl = train$Price, k = 5)
knn_preds <- as.numeric(as.character(knn_preds))
results <- rbind(results, data.frame(Model = "KNN", MAE = mean(abs(test$Price - knn_preds))))

# Plot for KNN
ggplot(data.frame(Actual = test$Price, Predicted = knn_preds), aes(x = Actual, y = Predicted)) +
  geom_point(color = "purple") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "KNN: Predicted vs Actual", x = "Actual Price", y = "Predicted Price") +
  theme_minimal()

### Support Vector Machine
svm_model <- svm(Price ~ ., data = train)
svm_preds <- predict(svm_model, test)
results <- rbind(results, data.frame(Model = "SVM", MAE = mean(abs(test$Price - svm_preds))))

# Plot for SVM
ggplot(data.frame(Actual = test$Price, Predicted = svm_preds), aes(x = Actual, y = Predicted)) +
  geom_point(color = "orange") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "SVM: Predicted vs Actual", x = "Actual Price", y = "Predicted Price") +
  theme_minimal()
```

### ii. General Description

The dataset consists of the following key features:

- **Mileage:** Distance driven in kilometers.
- **Engine:** Engine size in cubic centimeters (CC).
- **Power:** Engine power in brake horsepower (bhp).
- **Fuel\_Type:** Type of fuel used by the car (e.g., Petrol, Diesel).
- **Price:** The target variable represents the car's price.



### iii. Specific Requirements, Functions, and Formulas

- **Linear Regression:** The formula used is  $\text{lm}(\text{Price} \sim \text{Mileage} + \text{Engine} + \text{Power} + \text{Fuel\_Type})$ .
- **Decision Tree:** Regression tree created using the rpart package.
- **KNN:** The KNN model was implemented with  $k=5$ , considering the nearest neighbors.
- **SVM:** Support Vector Regression was used with a radial basis kernel.

### iv. Analysis Results

Each model was evaluated using Mean Absolute Error (MAE), which measures the average magnitude of the errors in the predictions:

**Linear Regression: MAE = 3.72**

**Decision Tree: MAE = 3.04**

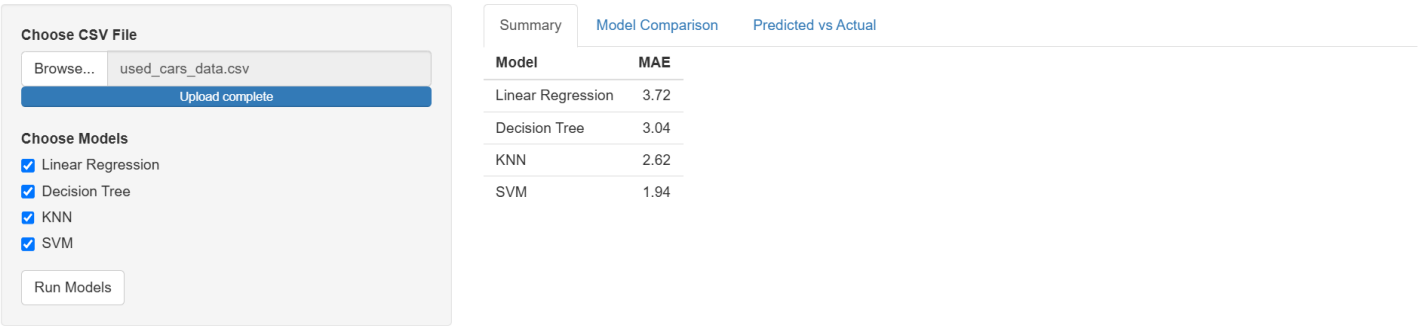
**KNN: MAE = 2.62**

**SVM: MAE = 1.94**

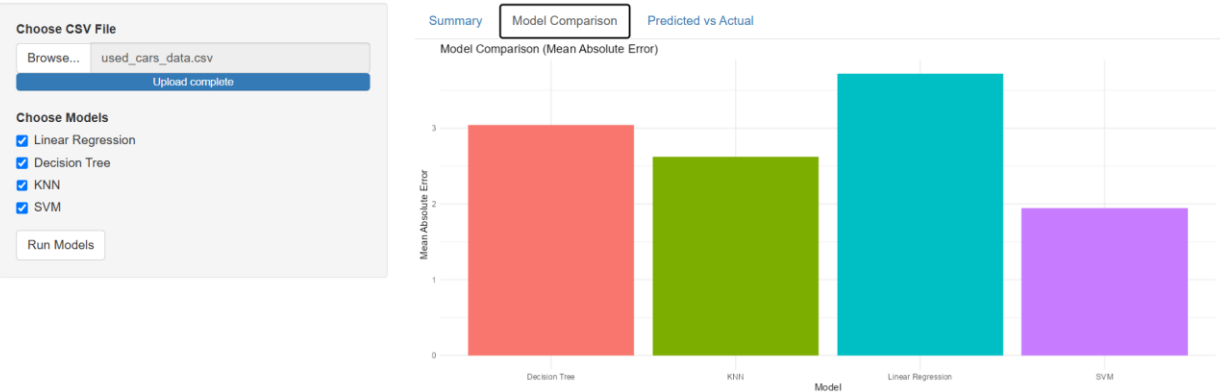
### v. Visualization (Dashboard)

To aid in visualizing the analysis, I developed an interactive Shiny Dashboard. This dashboard allows users to input car details (e.g., mileage, engine size, fuel type) and see the predicted price based on the selected model. The dashboard also provides a visual comparison of the performance of each model using plots for the predicted prices and actual values, helping users assess model accuracy.

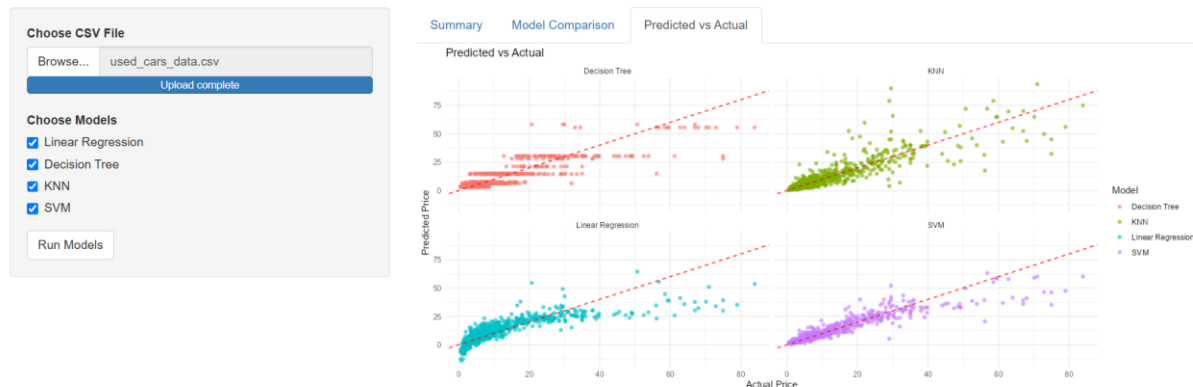
Used Car Price Prediction Model Comparison



Used Car Price Prediction Model Comparison



## Used Car Price Prediction Model Comparison



## 7. List of Analysis with Results

- Linear Regression: Predicted car price using a linear model. MAE = 3.72.
- Decision Tree: Captured non-linear relationships. MAE = 3.04.
- KNN: Used nearest neighbors for prediction. MAE = 2.62.
- SVM: Utilized support vector regression with radial kernel. MAE = 1.94.

Based on the Mean Absolute Error (MAE) values provided, **Support Vector Machine (SVM)** gives the best result for predicting car prices, with the lowest MAE of 1.94. A lower MAE indicates that the SVM model's predictions are closer to the actual car prices, making it the most accurate model among the four options tested.

## 8. Future Scope

The project can be further enhanced in several ways:

- **Incorporating More Features:** Including features such as car age, condition, maintenance history, and geographic location can lead to more accurate predictions.
- **Ensemble Methods:** Advanced techniques such as Random Forest or Gradient Boosting can improve model performance by combining multiple models.
- **Real-Time Prediction:** Integrating real-time market data and continuously updating the model for dynamic price predictions would make this tool more relevant for live transactions.