**International Journal of Mathematics and Computer Research**

**ISSN: 2320-7167**

**Volume 14 Special Issue 03 March 2026, Page no. – 172-179**

**Impact Factor: 8.751**

DOI: 10.47191/ijmcr/v14iSPC3.34


# Price Prediction of Used Cars Using Machine Learning Techniques

**Arati Y. Dange, Rupali L. Kamthe, Akshata V. Lembhe**

Department of Statistics, Dr. D. Y. Patil, Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

**ARTICLE INFO**

**Published Online:**
**17 March 2026**

**Corresponding Author:**
**Arati Y. Dange**

**ABSTRACT**

The used-car market in India has grown rapidly due to rising affordability, greater digital accessibility, and increasing consumer demand, making accurate pricing predictions essential for fair and transparent valuation. This study proposes a data-driven framework for forecasting used car prices using machine learning (ML) techniques. The dataset, collected via web scraping from Cars24.com, includes listings from major Indian cities and key vehicle attributes such as manufacturing year, kilometres driven, fuel type, transmission type, and brand. Five ML algorithms, Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, were trained and evaluated to identify the most effective model. Among them, the Random Forest Regressor achieved the best performance with an $R^2$ score of 92.04% and a Mean Absolute Error (MAE) of 2.87%, demonstrating strong capability in capturing complex non-linear patterns. Machine learning, a core component of Artificial Intelligence, enables systems to learn autonomously from data and is widely applied across domains requiring predictive analysis. In the automotive sector, ML offers an objective and scalable approach to estimating resale values by leveraging large datasets and market trends. The primary aim of this research is to identify key factors influencing used car prices and develop a robust predictive model. The findings highlight the effectiveness of ensemble learning methods in improving pricing accuracy and supporting data-driven decision-making in the expanding used car market.

**KEYWORDS:** 1) Used Car Price Prediction, 2) Machine Learning, 3) Random Forest Regressor, 4) Web Scraping, 5) Ensemble Learning, 6) Automotive Market Analysis

## 1. INTRODUCTION

The automotive industry is expanding rapidly worldwide, with the used car market emerging as one of its most dynamic segments. In recent years, the demand for pre-owned vehicles has grown significantly due to factors such as affordability, the rise of the middle-income population, and the increasing convenience of online platforms [1–5]. For many consumers, purchasing a used vehicle provides a practical and cost-effective alternative to buying a new one, making the second-hand automobile sector an essential contributor to the global automotive economy. This growth benefits both buyers and sellers by offering greater accessibility, lower costs, and a wider range of options [2,5–7]. Despite this progress, determining the fair resale value of a used car remains a major challenge. Vehicle prices depend on numerous interrelated factors, including brand, model, manufacturing year, mileage, fuel type, transmission, ownership history, and physical condition. These diverse attributes make traditional valuation methods, often based on human expertise and subjective judgment, prone to inconsistencies and biases. Consequently, there is a strong need for automated, objective, and data-driven pricing systems that can support fair and transparent used car transactions [2,4,8–10]

Machine Learning (ML), a prominent branch of Artificial Intelligence (AI), offers a powerful solution to this complex predictive task. ML algorithms are capable of processing large datasets, identifying hidden patterns, and generating accurate predictions without explicit rule-based programming [1,11,12] Unlike traditional appraisal techniques, ML models can continuously learn and adapt as new data becomes available, making them highly effective for real-time market analysis. Previous studies have shown that regression and ensemble-based ML approaches perform particularly well in modelling the non-linear relationships involved in used car price estimation [1,2,5,11,13]

This study develops a machine learning model to accurately predict used car prices based on key features such as purchase year, kilometers driven, fuel type, and brand. Several regression algorithms were implemented and compared to identify the model with the lowest prediction error and highest accuracy. The dataset was collected directly from the Cars24 website using Beautiful Soup and Selenium, covering listings from four major cities and seven popular car brands. The data was cleaned, structured, and stored in CSV format for analysis. The workflow involved preprocessing, splitting the data, training the models, and evaluating their performance. The results show that ML-based prediction provides accurate, unbiased, and efficient valuation of used cars, supporting better decision-making for buyers, sellers, and online resale platforms.

## 2. MATERIALS AND METHODS

### 2.1 Methodology

#### A. Dataset collection

The dataset for this study was collected from the Cars24 website using Beautiful Soup and Selenium web-scraping tools in Python, as shown in Table 1. This process extracted real-time vehicle listings containing key attributes such as brand, model, year of manufacture, kilometers driven, fuel type, transmission type, owner type, and selling price. A total of 18,180 records were gathered from four major Indian cities Bangalore, Mumbai, New Delhi, and Kolkata-covering seven widely used automobile brands. The data were stored in CSV format for preprocessing and model development. Due to its size and diversity, the dataset provides a reliable foundation for building a machine learning model capable of predicting used car prices under varying market conditions [11,12,14].

#### B. Data pre-processing

Preprocessing is essential in supervised machine learning, as model accuracy depends on data quality. Several steps were performed to clean and prepare the dataset.

**i) Cleaning and handling missing values**

The dataset was checked for missing, duplicate, and inconsistent entries [12]. Missing numerical values were filled using median imputation to reduce the impact of outliers. Redundant and irrelevant entries were removed to improve consistency.

**ii) Feature transformation**

Some numerical fields contained non-numeric characters (e.g., *kmpl*, *cc*, *bhp*). These units were removed programmatically to retain only numeric values. The process involved converting the column into a list, splitting elements to isolate numeric components, and reinserting the cleaned values into the Data Frame.

**iii) Encoding categorical variables**

Categorical attributes such as company, location, fuel type, and transmission were encoded using Label Encoding and One-Hot Encoding to convert them into numerical form suitable for machine learning algorithms.

**iv) Feature scaling and data splitting**

Feature scaling was applied to standardize the numerical variables. The dataset was then split into training (80%) and testing (20%) subsets. The selling price served as the target variable (Y), while all other features were treated as independent variables (X).

#### C. Model implementation

The machine learning workflow outlines the steps from data collection to model evaluation. The cleaned dataset was used to train various regression algorithms, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost. These models were selected for their ability to capture both linear and non-linear patterns. Performance was evaluated using the Coefficient of Determination ($R^2$) and Mean Absolute Error (MAE). The Scikit-learn library was used for model training and analysis.

The model-building process included importing and preprocessing the data, training algorithms, testing them on unseen data, and analyzing accuracy and error distribution [7,12,14] Linear Regression served as the baseline, while ensemble models were evaluated for improved performance [3,7].

#### D. Summary

This framework ensures that the collected data are cleaned, structured, and optimized for reliable machine learning analysis. The combination of real-world data, strong preprocessing, and regression-based modelling supports accurate and scalable prediction of used car prices in the Indian market. The next section presents the performance results of each model.

**Table 1. Sample of a data set**

| Sr No. | Name | Model | Company | Year | Fuel Type | Km Driven | Transmission Type | Price (In Lakh) | Location |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Hyundai Creta Sx Plus At 1.6 Petrol | Creta Sx Plus At 1.6 Petrol | Hyundai | 2017 | Petrol | 98493 | Automatic | 973000 | Bangalore |
| 2 | Renault Kwid 1.0 Marvel Iron Man 1 Edition Amt | Kwid 1.0 Marvel Iron Man Edition Amt | Renault | 2018 | Petrol | 19178 | Automatic | 407000 | Bangalore |
| 3 | Hyundai Eon Era Plus (O) | Eon Era Plus (O) | Hyundai | 2017 | Petrol | 33963 | Manual | 381000 | Bangalore |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | Maruti Swift Vxi | Swift Vxi | Maruti | 2012 | Petrol | 64557 | Manual | 463000 | Bangalore |
| 5 | Hyundai Creta Sx 1.6 Diesel | Creta Sx 1.6 Diesel | Hyundai | 2019 | Diesel | 43987 | Manual | 1150000 | Bangalore |

## 3. RESULTS AND DISCUSSION

### 3.1 Location-wise visualization:

Based on Fig. 1(a) and Fig. 1(b), Bangalore shows the highest number and percentage of used car sales compared to Mumbai, New Delhi, and Kolkata. The bar chart indicates Bangalore leads in total sales, while the pie chart confirms its dominant market share. This suggests strong demand for pre-owned cars, driven by a large working population. Kolkata has the lowest share, indicating relatively lower resale activity.

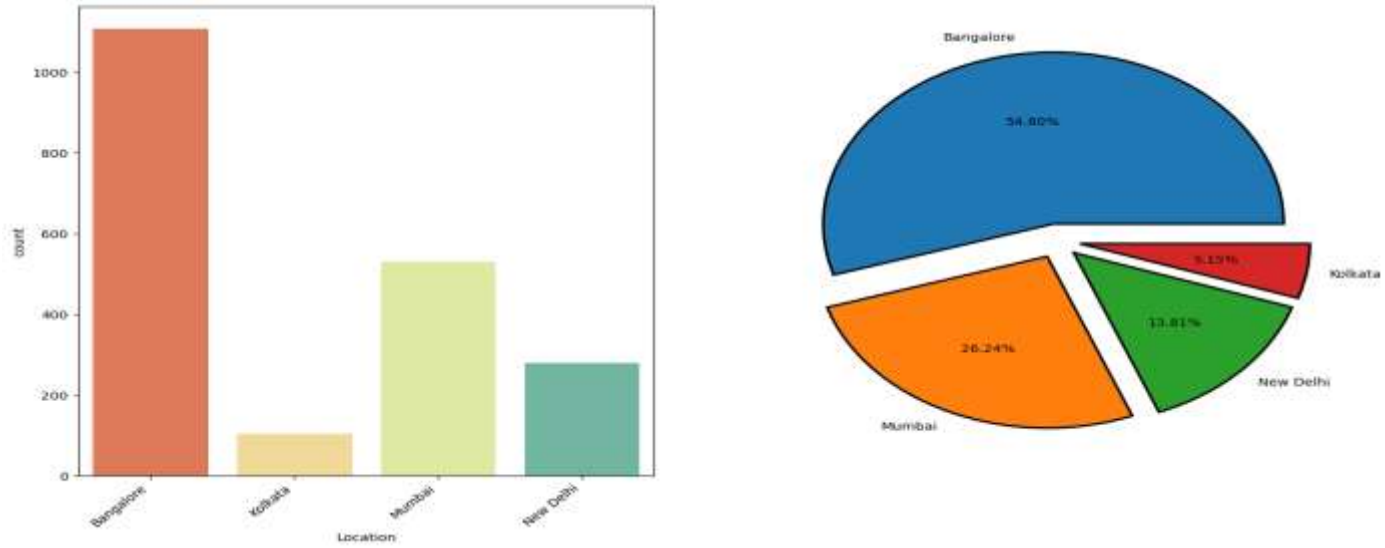


**Figure 1. (a) Bar chart, (b) Pi-chart of location-wise visualization**

### 3.2 Company-wise visualization

According to Fig. 2(a) and Fig. 2(b), Maruti clearly dominates the used car market, recording the highest number and percentage of sales, followed by Hyundai and Honda. The bar chart shows Maruti leading in total sales, while the pie chart confirms its strong market share. This indicates sustained buyer preference for Maruti due to affordability, fuel efficiency, and reliability, with Hyundai and Honda holding solid resale positions.

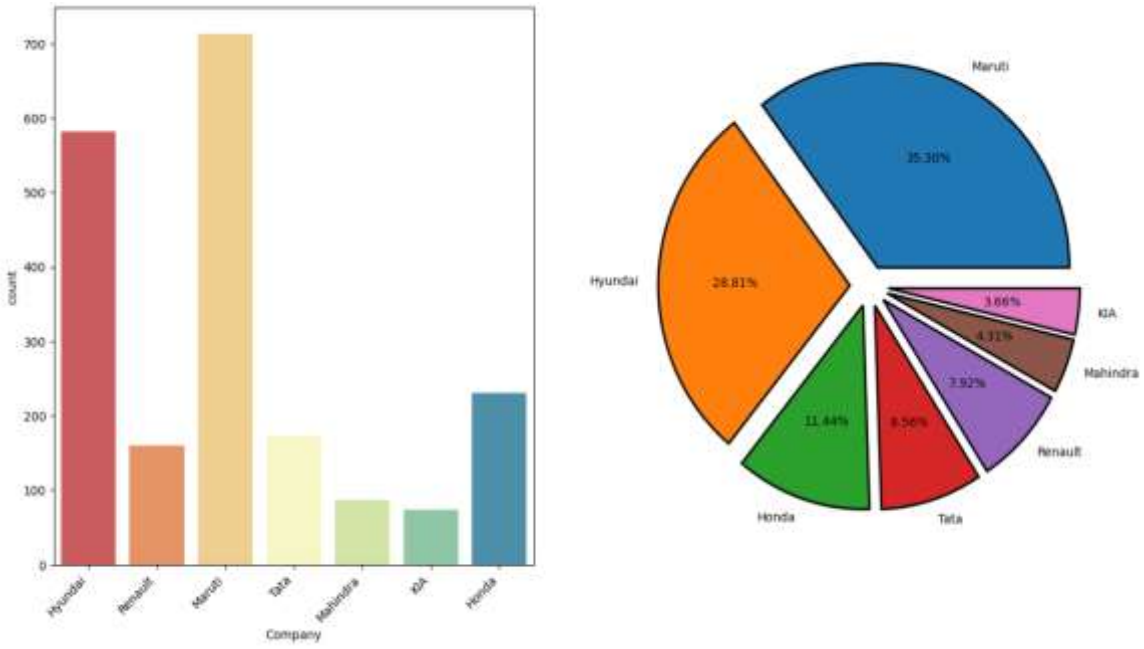


**Figure 2. (a) Bar chart, (b) Pi-chart of company-wise visualization.**

### 3.3 Price of used cars (location-wise)

As shown in Fig. 3, Bangalore has the highest average selling price for used cars, followed by New Delhi and Mumbai. This reflects strong demand, better vehicle upkeep, and higher purchasing power in Bangalore. New Delhi and Mumbai show comparatively lower prices, highlighting regional differences in buyer preferences and market dynamics.

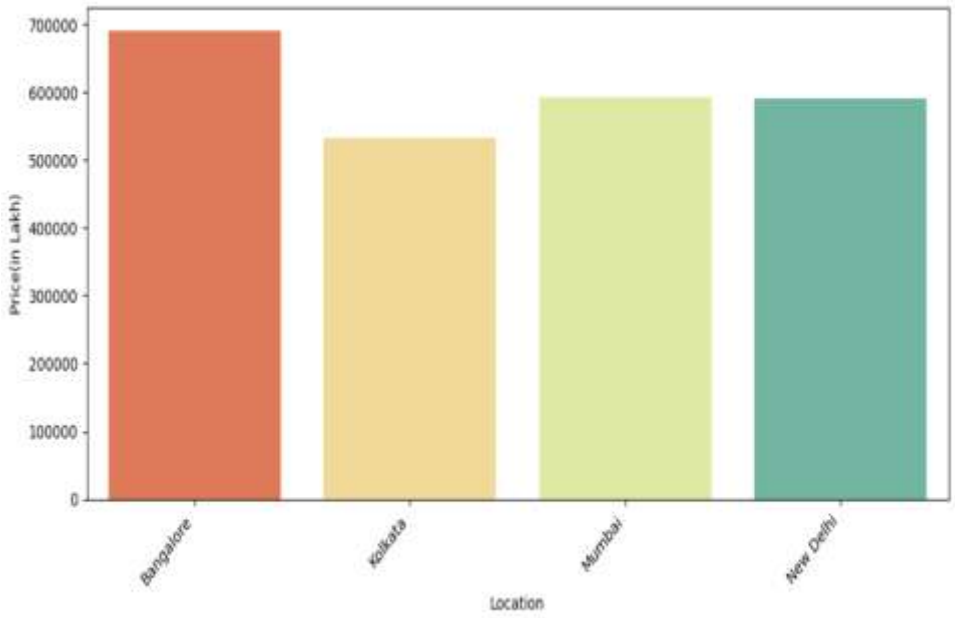


**Figure 3. Location-wise price of used cars**

### 3.4 Transmission types of cars

As illustrated in Fig. 4, Manual transmission cars far outnumber automatic ones, showing clear dominance in the used car market [3,12]. This trend reflects buyer preference for affordability, wider availability, and manual driving control. Overall, the figure highlights transmission type as a key factor influencing used car sales volume.

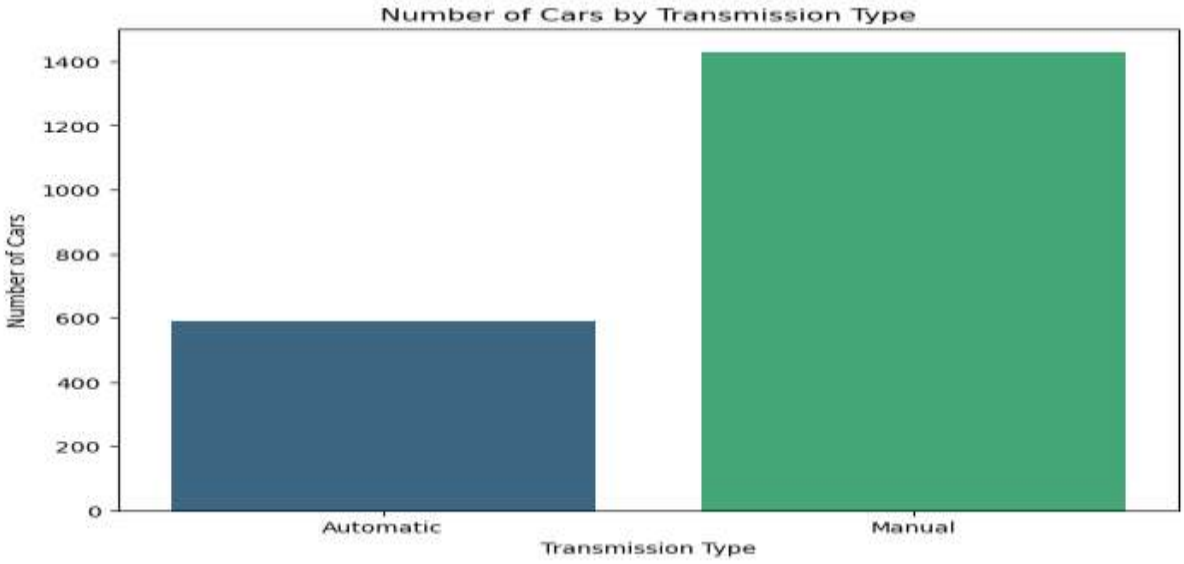


**Figure 4. Comparison of the transmission types of cars**

### 3.5 Cars based on fuel

As shown in Fig. 5(a) and Fig. 5(b), Petrol cars are the most commonly sold in the used market, followed by diesel and CNG vehicles, reflecting petrol's availability, smooth performance, and lower maintenance [3,12]. Diesel cars, though fewer, often sell at higher rates due to their efficiency. Newer models also achieve stronger sales and higher prices, showing vehicle age strongly influences resale value.

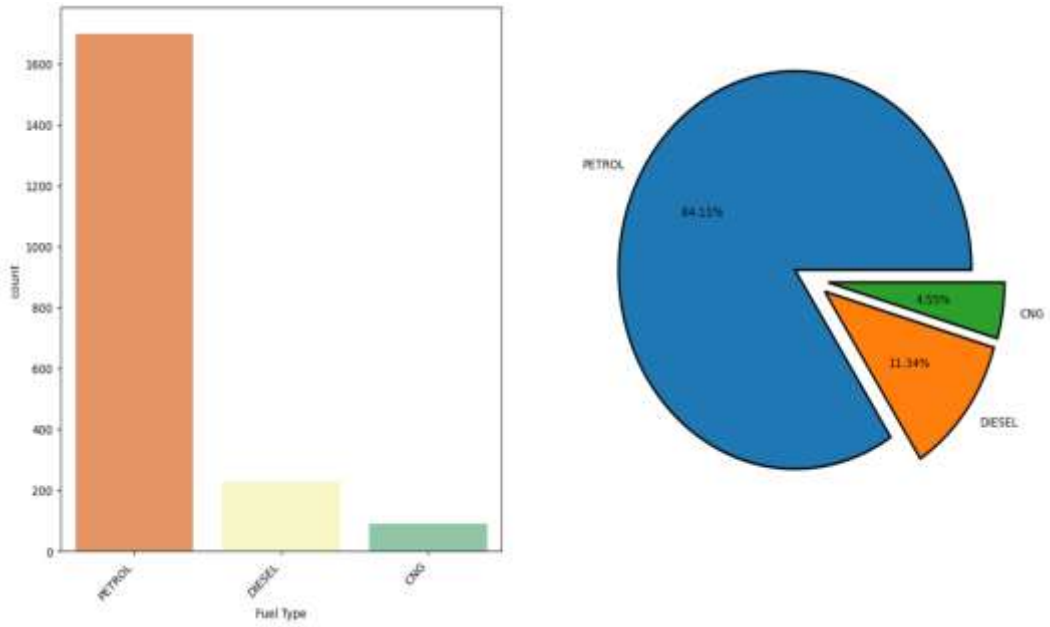


**Figure 5. (a) Bar chart, (b) Pi-chart of comparison of the fuel type of the cars**

### 3.6 Price distribution of various companies

Fig. 6 shows that among the top seven automobile brands, Kia has the most expensive used cars in terms of average selling price. This suggests that Kia vehicles maintain strong resale value, likely due to their modern design, premium features, and relatively recent introduction to the Indian market. The figure highlights how brand reputation and build quality play a key role in determining price distribution within the used car market.

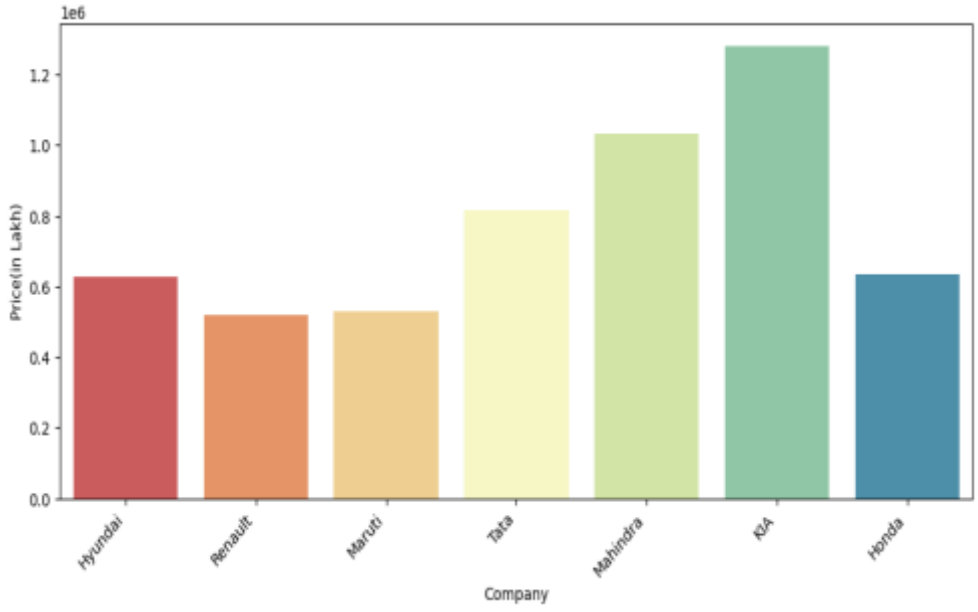


**Figure 6. Price distribution of various companies**

### 3.7 Cars based on kilometer driven:

Fig. 7 shows the link between kilometers (KM) driven and transmission type across brands. Honda records the highest average kilometers, suggesting extensive use and strong reliability [6]. The figure also shows Honda has more manual than automatic cars, indicating a strong market presence and buyer preference for its manual models.

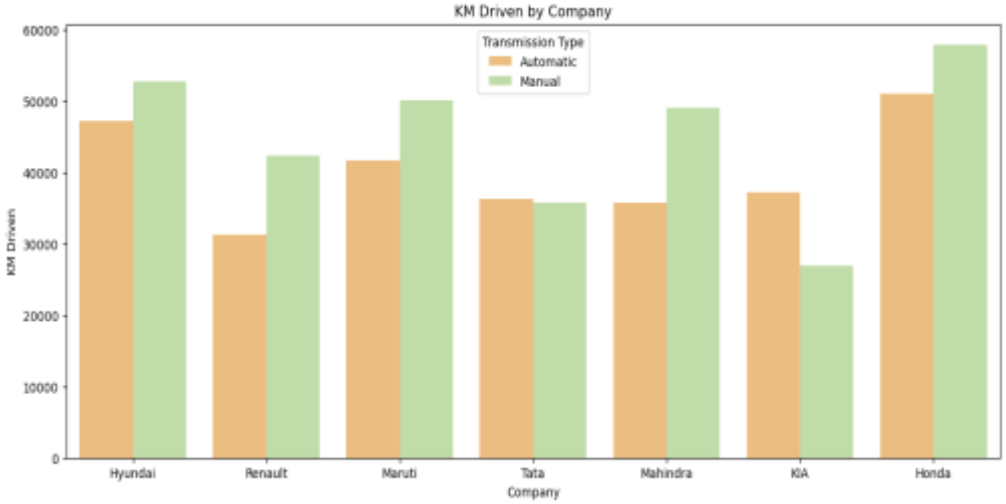


**Figure 7. Cars based on Kilometer driven**

### 3.8 Distribution of year, kilometer driven and price

From the above plots (Figure 8.), it is observed that the year 2017 recorded the highest number of used car sales, indicating strong market activity during that period. Most vehicles sold had been driven between 40,000 and 60,000 kilometers (KM), suggesting this range represents an optimal balance between usage and value retention. Additionally, the selling price of the majority of used cars falls within the range of ₹4-7 lakhs, reflecting the preferred price segment among buyers in the resale market.

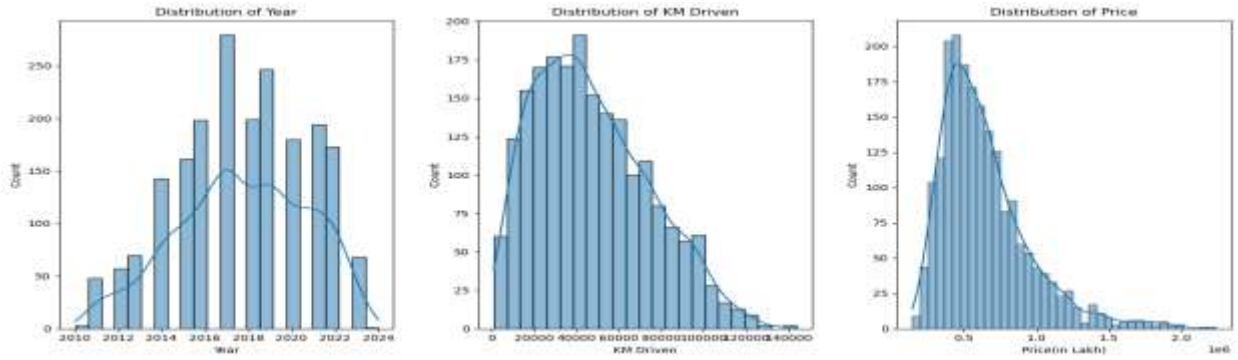


**Figure 8. Distribution of year, KM driven and price**

### 3.9 Descriptive statistics

From Table 2, it is evident that the oldest car sold in the dataset had been used for 14 years, while the highest selling price recorded was ₹22.40 lakhs**.** The statistical summary reveals that most vehicles are relatively new, with an average manufacturing year of 2017, typically ranging between 5 to 7 years old. The cars have been driven an average of 47,236 km, indicating moderate usage, with some reaching over 1.43 lakh km. The average selling price stands at around ₹6.43 lakhs**,** with the majority priced between ₹4.26 and ₹7.8 lakhs**,** reflecting strong demand in the mid-range segment. On average, vehicles were owned for about 6 years, aligning with common replacement trends in the Indian used car market.

**Table 2. Descriptive statistics**

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **Year** | 2020 | 2017.72 | 3.00 | 2010.0 | 2016.00 | 2018.0 | 2020.0 | 2024.0 |
| **Km Driven** | 2020 | 47236.41 | 27135.17 | 1011.0 | 25748.75 | 43399.5 | 65590.0 | 143991.0 |
| **Price (In Lakh)** | 2020 | 642743.06 | 307385.77 | 144000 | 426750 | 576000 | 780000 | 2240000 |
| **No of Years** | 2020 | 6.27 | 3.00 | 0.0 | 4.00 | 6.0 | 8.0 | 14.0 |

### 3.10 Correlation Matrix

The heatmap in Fig. 9 reveals that the year of manufacture and number of years is strongly and inversely related, meaning newer cars have been used for fewer years. A clear positive correlation exists between kilometers driven and number of years, indicating that older cars tend to cover

more distance. The price of used cars shows a positive correlation with the year and a negative correlation with age, confirming that newer cars command higher resale values [3,12]. Overall, factors such as **a**ge**,** year**,** and mileage have the most significant impact on used car pricing.

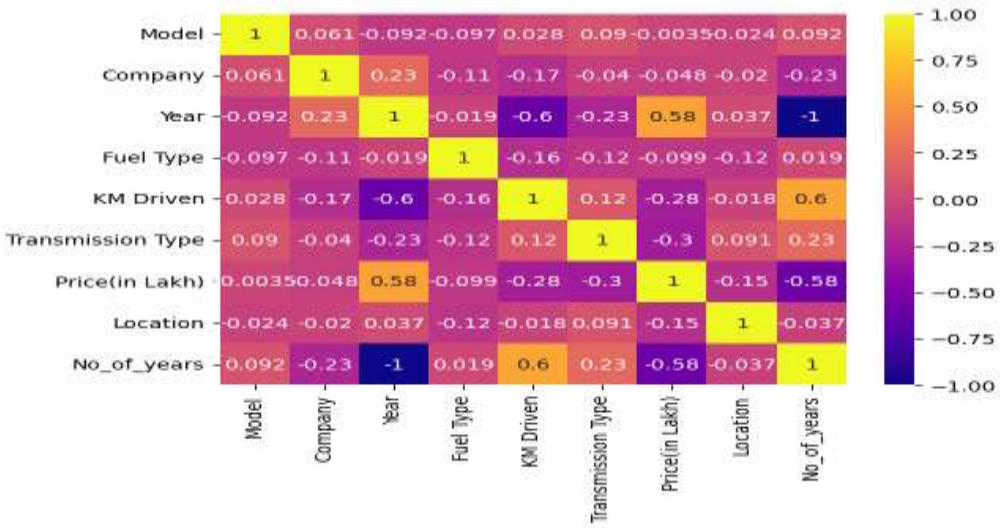


**Figure 9. Correlation matrix of every attribute**

The dataset analysis offers key insights into the dynamics of the used car market. Among the major cities, Bangalore recorded the highest number of listings and the highest average selling prices, indicating strong market activity and greater purchasing power. Maruti and Hyundai emerged as the most frequently listed brands, reflecting their widespread popularity, reliability, and affordability. Feature-wise, petrol cars and manual transmissions dominated the listings, suggesting continued consumer preference for fuel-efficient and low-maintenance vehicles. Correlation analysis highlighted important relationships between attributes. A positive correlation between manufacturing year and price shows that newer vehicles hold higher resale value. Similarly, a moderate positive correlation between kilometers driven and vehicle age indicates that older cars generally accumulate more mileage [8]. These observations reinforce the influence of both age and usage on resale pricing.

### 3.11 Machine learning models

As shown in Table 3, Model evaluation revealed that the Random Forest Regressor achieved the highest accuracy, with an $R^2$ score of 92.04% and a Mean Absolute Error (MAE) of 2.87%, showing its strong ability to capture complex non-linear patterns. The Decision Tree Regressor performed reasonably well ($R^2$ = 88.64%, MAE = 3.5%), while Gradient Boosting produced slightly lower accuracy ($R^2$ = 84.37%, MAE = 3.8%). In contrast, Linear Regression and AdaBoost showed weaker performance due to their limited capability to model non-linearity. Overall, the results demonstrate that ensemble learning techniques, particularly Random Forest, provide a highly effective and reliable approach for predicting used car prices. These models support fair valuation and data-driven decision-making for buyers, sellers, and online resale platforms [3,12,14].

**Table 3. Machine learning models and their accuracy**

| **Model** | **$R^2$ Score (%)** | **MAE (%)** |
|---|---|---|
| Linear Regression | 44.55 | 8.30 |
| Decision Tree | 88.64 | 3.50 |
| Random Forest | 92.04 | 2.87 |
| Gradient Boosting | 84.37 | 3.80 |
| AdaBoost | 59.15 | 7.88 |

## 4. CONCLUSION

The findings of this study demonstrate the strong effectiveness of ensemble learning techniques, particularly

Random Forest and Gradient Boosting, in modelling the complex, non-linear factors that determine used car prices. Random Forest, by aggregating multiple decision trees, reduces overfitting and provides stable, generalizable predictions. Compared to traditional linear approaches, these models offer higher accuracy, greater robustness, and improved adaptability to varying market conditions. These results have practical value for online resale platforms and automobile dealerships, as they highlight the potential of machine learning to automate and enhance vehicle valuation. Integrating such predictive models can support fair pricing, strengthen buyer trust, and streamline resale operations. Future work may involve exploring deep learning methods or incorporating additional variables such as market demand trends, vehicle condition, and regional economic factors to further increase prediction accuracy and real-world relevance.

**Acknowledgment**

The author extends heartfelt appreciation to the Head of the Department of Statistics and the Principal of D. Y. Patil College, Pimpri, Pune-18, for their valuable support throughout the course of this research.

**Compliance with ethical standards**

Conflict of interest: The authors declare that they have no conflict of interest.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships, which may be considered as potential competing interests:

**Author contributions**

**Arati Y. Dange**- Investigation and data collection, and manuscript writing;
**Rupali L. Kamthe**- Formal analysis;
**Akshata V. Lembhe**- Formal analysis;
All authors read and approved the final manuscript.

**Declarations**

Conflict of interest- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**REFERENCES**

1. P. Venkatasubbu, M. Ganesh, Used Cars Price Prediction using Supervised Learning Techniques, (2019) 216–223. https://doi.org/10.35940/ijeat.A1042.1291S319.
2. M. Collard, Price Prediction for Used Cars A Comparison of Machine Learning Regression Models, (2022) 1–45.
3. N. Monburinon, Prediction of Prices for Used Car by Using Regression Models, 2018 5th Int. Conf. Bus. Ind. Res. (2018) 115–119. https://doi.org/10.1109/ICBIR.2018.8391177.
4. R. Yohanes, D. Lasut, Web-Based used Car Price Prediction Application with Linear Regression Method, 7 (2025). https://doi.org/10.32877/bt.v7i3.1722.
5. K. Samruddhi, R.A. Kumar, Used Car Price Prediction using K-Nearest Neighbor Based Model, 4 (2020) 4–7. https://doi.org/10.29027/IJIRASE.v4.i2.2020.629-632.
6. B. Aravind, S. Pillai, P. Architect, A Deep Learning Approach for Used Car Price Prediction, 3 (n.d.) 31–51.
7. A. Alhakamy, A. Alhowaity, A.A. Alatawi, H. Alsaadi, Are Used Cars More Sustainable ? Price Prediction Based on Linear Regression, (2023) 1–17.
8. C. Jin, Price Prediction of Used Cars Using Machine Learning, 2021 IEEE Int. Conf. Emerg. Sci. Inf. Technol. (2021) 223–230. https://doi.org/10.1109/ICESIT53460.2021.9696839.
9. S. Voß, S. Lessmann, Resale Price Prediction in the Used Car Market, (2011).
10. E. Liu, J. Li, A. Zheng, H. Liu, T. Jiang, Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network, (2022).
11. M. Asghar, K. Mehmood, S. Yasin, Z.M. Khan, Used Cars Price Prediction using Machine Learning with Optimal Features, (2021) 113–119.
12. V. Viswanatha, A.C. Ramachandra, B.D. Parameshachari, H. V Vachan, S.S. Shetty, Predicting the Price of used Cars using Machine Learning, 2023 Int. Conf. Evol. Algorithms Soft Comput. Tech. (2023) 1–6. https://doi.org/10.1109/EASCT59475.2023.10393486.
13. T. Doan, Selecting Machine Learning Algorithms Using Regression Models, 2015 IEEE Int. Conf. Data Min. Work. (2015) 1498–1505. https://doi.org/10.1109/ICDMW.2015.43.
14. L. Bukvi, Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning, (2022).