# Prediction of Car Prices in Nigeria Using Machine Learning Models

**Samuel Oluyemi Owoeye[1*], Folasade Olayinka Durodola[2], Sikirulahi Opeyemi Abdulkareem[3], Kayode Michael Makinde[4] and Olaniyi Erastus Folaranmi [5]**

[1,2,3,4,5] Department of Mechatronics Engineering, Federal University of Agriculture, Abeokuta, Nigeria
Email: owoeyeso@funaab.edu.ng

**Abstract**

*In Nigeria, where more than 95% of vehicles are used cars, precise car valuation is essential for buyers, sellers, and dealers alike. This research employs machine learning techniques to forecast the prices of pre-owned cars based on various vehicle features. Data were collected from online car sales platforms in Nigeria, encompassing over 8,000 vehicles with attributes such as make, model, transmission type, and overall condition. Following data preprocessing, four machine learning algorithms were evaluated, namely Linear Regression, Random Forest, XGBoost, and Multi-Layer Perceptron (MLP) Regressor. The XGBoost model demonstrated superior performance, achieving a Root Mean Square Error (RMSE) of ₦7, 047, 536.43, a Mean Absolute Error (MAE) of ₦ 3,540,639.15, and an R-squared (R2) score of 0.8612, indicating that it accounts for 86.12% of the variance in car prices. The most effective model was implemented in a web application based on Streamlit, allowing users to enter vehicle details and obtain price estimates and providing a valuable resource for the automotive market in Nigeria.*

**Keywords:** Car prices; Machine Learning; Linear Regression; Multi-Layer Perceptron; Random Forest; XGBoost

## 1.0 INTRODUCTION

The automotive industry in Nigeria is a critical component of the nation's economy, serving as a key driver of transportation, commerce, and personal mobility. However, the pricing of cars in Nigeria is influenced by a complex array of factors, including import tariffs, foreign exchange rates, fuel prices, vehicle age, mileage, and regional market dynamics (Aderibigbe *et al.,* 2024; Agarwal *et al.,* 2019). These factors create a highly volatile and often opaque pricing environment, making it challenging for buyers and sellers to determine fair market values. Traditional methods of car price estimation, which rely on manual appraisal or simplistic statistical models, are often inadequate in capturing the intricate relationships between these variables, leading to inaccurate predictions and suboptimal decision-making.

In recent years, machine learning (ML) has emerged as a transformative tool for predictive analytics, offering the ability to model complex, non-linear relationships in large datasets. Techniques such as random forests, gradient boosting, and neural networks have been successfully applied to price prediction tasks in various domains, including real estate, stock markets, and e-commerce (Rane *et al.,* 2024; Hatta *et al.,* 2024). These methods have demonstrated superior performance compared to traditional statistical approaches, particularly when dealing with high-dimensional and heterogeneous data. Despite these advancements, the application of machine learning to car price prediction in Nigeria remains limited, with few studies exploring the development of accessible and scalable tools tailored to the local market.

This study addresses this gap by developing a machine learning-based web application designed to predict car prices in Nigeria. The application leverages a comprehensive dataset comprising vehicle attributes (e.g., make, model, year, mileage), economic indicators (e.g., inflation rates, exchange rates), and historical transaction data to train and deploy predictive models. By employing state-of-the-art machine learning algorithms, the system aims to provide accurate and reliable price estimates while accounting for the unique characteristics of the Nigerian automotive market. Furthermore, the web-based interface ensures ease of access and usability, enabling a wide range of stakeholders, including individual buyers, dealerships, and financial institutions, to make data-driven decisions. The development of this application is guided by principles of software engineering and machine learning best practices, including data pre-processing, feature engineering, model selection, and performance evaluation (Pan *et al.,* 2022; Watson *et al.,* 2019). The system is designed to be scalable and adaptable, allowing for continuous updates and improvements as new data becomes available. Additionally, the interpretability of the models is prioritized to enhance user trust and facilitate informed decision-making.

Various studies and research initiatives have investigated several methodologies for predicting automobile prices, employing diverse techniques aimed at enhancing the precision and dependability of these predictions. The advent of online platforms has facilitated easier access for both buyers and sellers to comprehensive information regarding the determinants of a vehicle's market value. This paper will examine a forecasting method for automobile prices. Agrahari *et al.* (2021) indicated that Lasso Regression outperformed Linear Regression, particularly in scenarios characterized by high multicollinearity or when feature selection is of paramount importance. The L1 regularization inherent in Lasso helps to alleviate overfitting by driving certain coefficients to zero, thereby fostering improved feature selection and the development of more resilient models. In another study, Pal *et al.* (2019) utilized a random forest model to predict used car prices, which consisted of 500 decision trees and achieved a training accuracy of 95.82%, with a testing accuracy of 83.63%. Samruddhi and Kumar (2020) analysed the Indian used car market, creating a model that estimates used vehicle prices through the K-Nearest Neighbor algorithm, which achieved an accuracy of 85%, surpassing the 71% accuracy of linear regression. Additionally, Putra *et al.* (2024) developed an artificial neural network (ANN) model for predicting used car prices, which demonstrated superior performance relative to other regression models, achieving a Mean Absolute Error (MAE) of 1060, a Mean Absolute Percentage Error (MAPE) of 11%, a Root Mean Square Error (RMSE) of 2104, and an R-squared value of 0.96. Among the baseline models, the Random Forest Regressor yielded the most favourable results, with a MAE of 1382 and a MAPE of 14%, while the Linear Regression model produced the least satisfactory outcomes. The elevated R-squared and diminished RMSE values signify the ANN model's enhanced precision, reflecting a strong correlation between the predicted and actual outcomes. Peerun *et al.* (2015) undertook a study to assess the efficacy of neural networks in forecasting the prices of pre-owned vehicles. Their investigation revealed that, especially for vehicles with higher price points, the predicted values did not closely correspond with the actual market prices. In the context of used car price prediction, it was found that support vector machine regression slightly outperformed both neural networks and linear regression in terms of effectiveness.

Recent researches have examined a variety of machine learning methodologies for predicting the prices of used cars, with several yielding encouraging results. Comparative studies indicated that nonlinear methods, particularly ensemble techniques that integrate neural networks, random forests, and gradient boosting, deliver outstanding results, achieving a mean absolute percentage error of 14.34% (Kovpak *et al.,* 2019). Numerous studies have investigated the factors influencing automobile pricing through the use of regression analysis and machine learning techniques. The most significant determinants identified were mileage and the manufacturing year, which have a profound impact on the prices of used cars (Sun, 2024). Furthermore, vehicle specifications such as size, horsepower, and the number of cylinders play a crucial role in determining the costs of both new and used vehicles, the brand and fuel type are also essential factors to consider (Chandak *et al.,* 2019). By incorporating these variables, machine learning algorithms can provide more accurate pricing predictions, benefiting both buyers and sellers in the increasingly competitive used car market (Pal *et al.,* 2019).

This present research contributes to the growing body of knowledge on the application of machine learning in emerging markets, particularly in the context of price prediction. By addressing the specific challenges of the Nigerian automotive market, the study provides a practical solution to a pressing problem while also offering a framework that can be adapted to other regions with similar economic and market dynamics. Furthermore, the development of an open-access web application represents a significant step toward democratizing access to advanced predictive tools, fostering transparency, and promoting efficiency in the car market.

## 2.0 METHODOLOGY

### *2.1 Data Collection*

The dataset for this study, aimed at predicting car prices in Nigeria, was sourced through web scraping from online car sales platform, with a focus on cars listed in Nigeria. Initially, the dataset comprised approximately 7,500 entries across 36 car brands, covering a range of models under each brand. There are 39 columns of the data containing extensive details on each car, captured across several attributes of the car. The data includes Fuel type, Transmission type, model, brand, Item condition etc.

### *2.2 Data Cleaning, Preprocessing and Analysis*

The initial pre-processing phase began with a systematic review to identify columns most relevant to car price prediction. Columns like id, seller info, seller name, promotion, which contributed limited value to the model, were removed to streamline the dataset and focus on key features. For the remaining columns, missing values were carefully handled to maintain data consistency. Where possible, missing values were addressed by grouping similar car models and observing patterns within each group, which allowed for accurate filling based on shared characteristics. For columns with significant missing data that had minimal impact on price prediction, these entries were removed to maintain data quality without compromising the dataset's integrity. Certain columns required additional processing due to unstructured data that needed to be reorganized for analysis. Two such columns, further attributes and second condition, contained valuable yet concatenated information that could yield insights if properly parsed. Using regular expressions (regex), these columns were parsed, reorganized, and formatted to create a more readable, structured layout, enhancing their usability in the model. Following the data-cleaning and restructuring process, the refined dataset consisted of 56 columns and approximately 7,000 entries. An exploratory data analysis (EDA) was conducted to uncover initial insights and observe patterns within the dataset.

### *2.3 Machine Learning Models*

This study applies four regression models which are Linear Regression, XGBoost, Random Forest, and MLP Regression to predict the outcomes for the training and test sets. The performance of each model is evaluated using three metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Square. These metrics help assess the models' accuracy and predictive power, enabling a comparison of their effectiveness in predicting the car prices.

### *2.3.1 Linear Regression*

Linear regression is a statistical method used to determine the value of a dependent variable based on one or more independent variables. This technique assesses the relationship between two variables and serves as a modeling approach to predict the dependent variable. In this study, the independent variables include the car's name, year of use, and mileage, while the dependent variable is the car's price. The model assumes a linear relationship as shown in equation 1: (Alita *et al.,* 2021; Singh *et al.,* 2024).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \qquad [1]$$

where:
$y$ = predicted car price
$\beta_i$ = model coefficients
$x_i$ = input features
$\epsilon$ = error term

### *2.3.2 Random Forest Regressor*

Random Forest is an ensemble machine learning technique utilized for both classification and regression purposes. It generates numerous decision trees by employing randomly chosen subsets of data and features, subsequently aggregating their results through majority voting for classification tasks or averaging for regression tasks. This methodology improves predictive performance and mitigates the risk of overfitting in comparison to standalone decision trees. The model mathematical expression is shown in equation 2: (Jain and Jana, 2023; Zhang *et al.,* 2022; Mienye and Sun, 2022).

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \qquad [2]$$

where:
$T$ = number of decision trees
$f_t(x)$ = prediction of the t-th tree
Ŷ = final predicted price

### *2.3.3 XGBoost Regressor*

XGBoost represents a cutting-edge gradient boosting framework optimized for both speed and performance in machine learning applications. It enhances the gradient boosting algorithm by assembling a collection of weak predictive models, usually in the form of decision trees, to generate a robust predictive model. Its features, including scalability, portability, and the ability to process data in a distributed manner, have contributed to XGBoost's widespread adoption in machine learning competitions and practical applications. The model is governed by the equations 3 to 5: (Heitz *et al.,* 2025; Maleki *et al.,* 2023).

$$\hat{y}_i = \sum_{k=1} f_t(x_1), \quad f_k \in \mathcal{F} \qquad [3]$$

$$\mathrm{l}(\emptyset) = \sum_{i=1}^{n} l(y_1, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \qquad [4]$$

$$\Omega(f) = \ _{\Upsilon}T + \frac{1}{2}\lambda \, || \, \omega||^2 \qquad [5]$$

where:
$\iota$ = overall objective function
$l(y_1, \hat{y}_i)$ = loss function (e.g., squared loss)
$\Omega$ = regularization term
$T$ = number of leaves in a tree
$\lambda, \Upsilon$ = regularization coefficients

### *2.3.4 Multilayer Perceptron (MLP) Regressor*

The Multi-Layer Perceptron (MLP) Regressor is a supervised machine learning model that utilizes artificial neural networks to forecast results by transforming input data into output values via a sequence of interconnected layers. The architecture of the MLP comprises an input layer, one or more hidden layers, and an output layer. Each layer is made up of interconnected nodes, where the weighted inputs are processed using nonlinear activation functions: The model's mathematical formulation are shown in equations 6 and 7: (Mienye and Sun, 2022; Heitz *et al.,* 2025).

$$a^{(l)} = \sigma\left(W^{(l)}a^{(l-1)} + b^{(l)}\right) \qquad [6]$$

$$\hat{y} = W^{(L)}a^{(L-1)} + b^{(L)} \qquad [7]$$

*where*
$a^{(l)}$ = activation at layer $l$
$W^{(l)}$, $b^{(l)}$ = weights and biases
$\sigma$ = nonlinear activation function (e.g., ReLU)

### *2.4 Model Training and Evaluation*

Each model was trained using an 80:20 train-test split. Hyperparameters for Random Forest, XGBoost, and MLP were tuned using grid search with 5-fold cross-validation. The evaluation used the following performance metrics are:

### *2.4.1 Root Mean Squared Error (RMSE)*

Root Mean Squared Error (RMSE) is a widely used metric for evaluating the accuracy of a model's predictions. It measures the average magnitude of the errors between predicted values and actual values, providing a sense of how far off predictions are from the true values. The RMSE is calculated using equation 8:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad [8]$$

### *2.4.2 Mean Absolute Error (MAE)*

Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of a model's predictions by measuring the average absolute differences between predicted values and actual values. It provides a straightforward assessment of prediction accuracy without being influenced by the magnitude of the errors. The MAE is calculated using equation 9:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad [9]$$

### *2.4.3 Coefficient of Determination (R²)*

The Coefficient of Determination, denoted as $R^2$, is a statistical measure that indicates the proportion of the variance in the dependent variable that can be explained by the independent variable(s) in a regression model. It provides insight into the goodness of fit of the model. The $R^2$ value is calculated using equation 10

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \acute{y}_i)^2} \qquad [10]$$

where:

$y_i$ = actual price
$\hat{y}_i$ = predicted price
$\acute{y}$ = mean of actual prices
n = number of observations

## 3.0 RESULTS AND DISCUSSION

Figure 1 shows a sharp upward trajectory in prices for vehicles manufactured after 2010. This rise reflects the growing market demand for newer cars equipped with modern technology, advanced features, and better fuel efficiency. Newer vehicles also offer greater reliability, which justifies their premium pricing. In contrast, cars manufactured before 1980 exhibit significantly lower prices, due to their age and diminished market appeal.

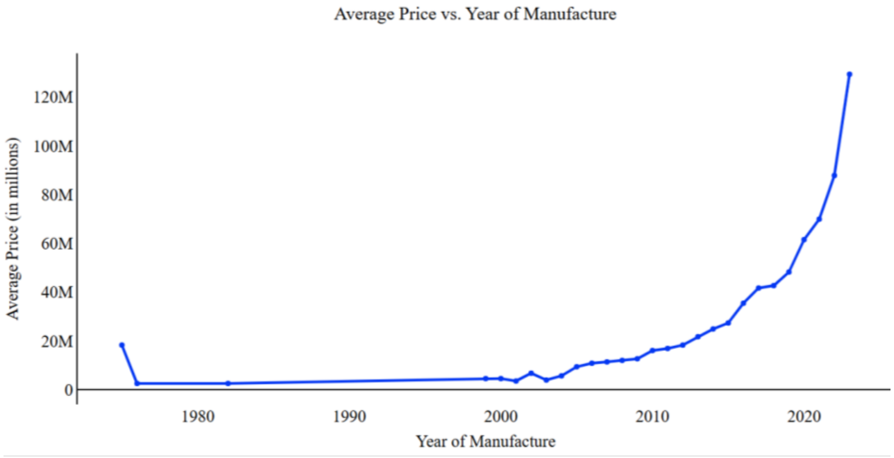

**Figure 1: Plot of Price against Year of Manufacture**

As shown in Figure 2, the dataset indicates that Toyota, Lexus, and Mercedes-Benz dominate the market listings, suggesting these brands are in high demand and widely available within the Nigerian market. In contrast, brands such as Volvo, Mitsubishi, Scion, Lincoln, Peugeot, and Mazda are less popular, with comparatively lower demand and availability.

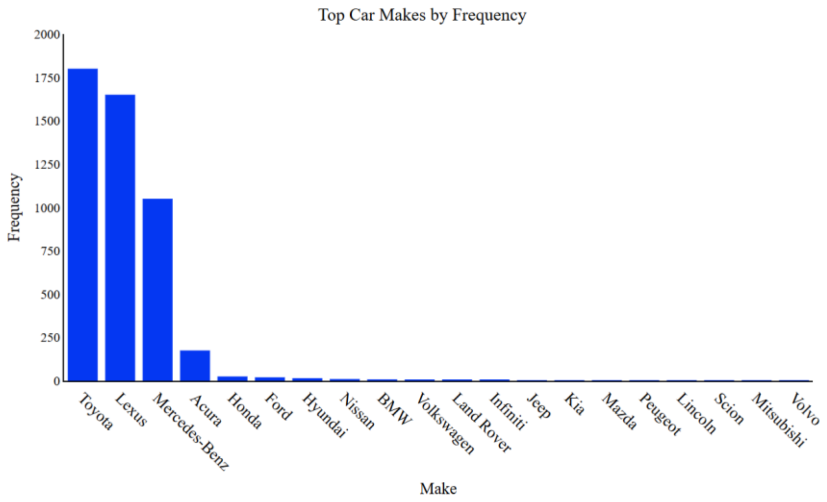

**Figure 2: Frequency of the various car brands**

Figure 3 shows the distribution of cars based on the item condition and the registration status of cars. Figure 3(a) illustrates the distribution of cars in the dataset based on their condition. The majority of cars fall under the Foreign Used category, representing vehicles that were previously used outside Nigeria before being imported into the country. It shows that the foreign used category dominates the market, with over 4,000 cars listed, far surpassing other categories. This preference can be attributed to the affordability of foreign-used cars compared to brand-new vehicles and the perception of higher quality over locally used cars. Conversely, brand new cars form the smallest segment, reflecting their high cost and limited affordability for most buyers. The locally used category also shows limited representation, indicating lower demand due to concerns over maintenance and quality compared to imported vehicles. According to Figure 3(b), registration status further highlights market trends. The majority of cars are unregistered, with over 4,000 cars listed in this category. This suggests a strong preference for freshly imported vehicles, offering buyers the flexibility to handle the registration process themselves. In contrast, registered cars form a much smaller segment, pointing to their reduced appeal in the current market.

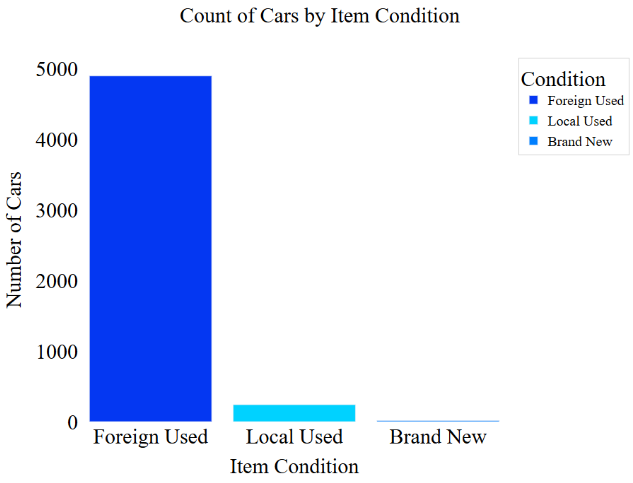

**Figure 3(a): Counts of Cars based on Used Condition**

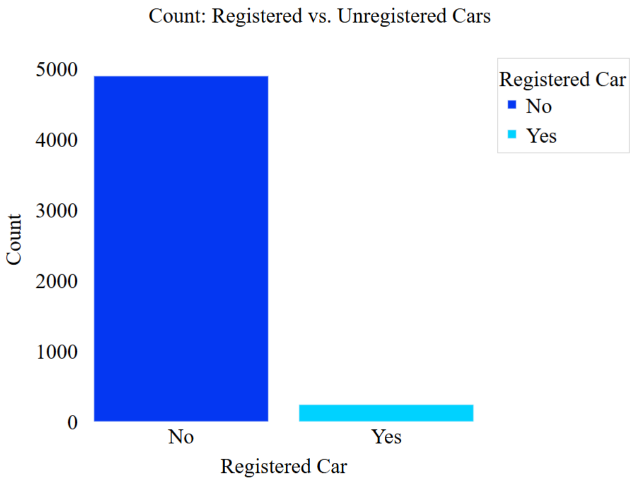

**Figure 3(b): Counts of Cars based on Registration**

Figure 4(a) illustrates how car prices vary with specific attributes. It explores the price distribution across different transmission types: Automatic, Manual, CVT, and AMT. Automatic transmissions exhibit the widest price range, with several high-price outliers, highlighting their association with luxury and high-performance vehicles favoured by premium buyers. In contrast, Manual transmissions have a narrower price range and a lower median price, reflecting their affordability and appeal to budget-conscious consumers. CVT and AMT transmissions, while less common, show distinct trends—AMT vehicles tend to have slightly higher prices, indicating growing adoption in mid-range markets.

Figure 4(b) shifts the focus to the effect of fuel type on car prices. Petrol vehicles dominate the dataset with prices clustered at lower ranges, though some high-end outliers point to luxury or sports cars. Diesel vehicles, on average, are slightly more expensive but maintain a similar price spread, appealing to buyers seeking fuel efficiency and long-term durability. Meanwhile, Hybrid vehicles stand out with higher price ranges and fewer outliers, underscoring their position as premium options due to advanced technology and eco-friendly features. Together, these insights reveal how transmission and fuel types significantly influence car pricing dynamics.

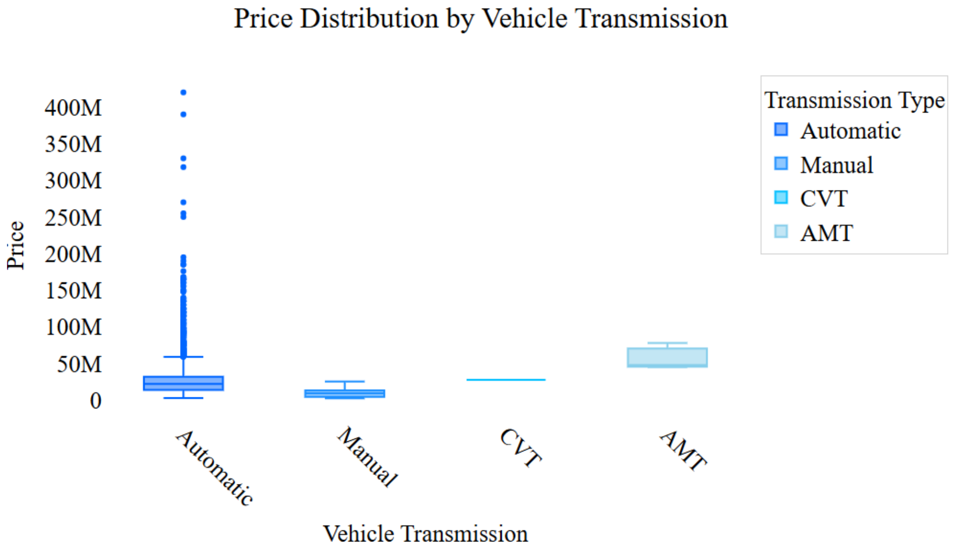

**Figure 4(a): Graph of price against vehicle transmission type**

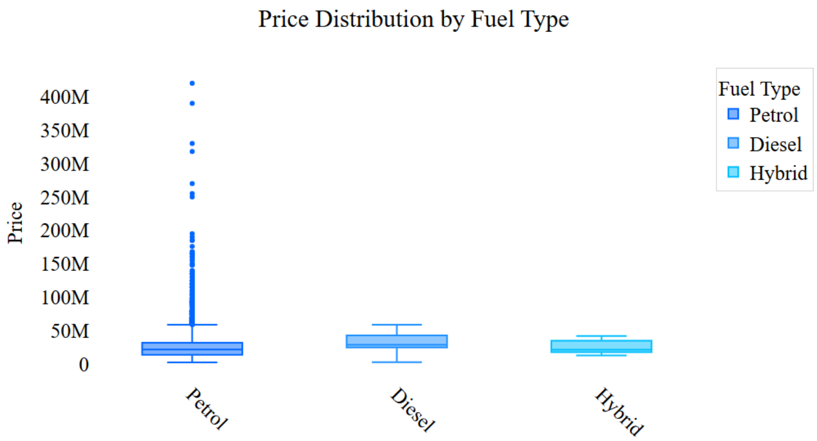

**Figure 4(b): Graph of price against fuel type**

Figure 5(a) illustrates the price distribution across various horsepower ranges, shedding light on how horsepower influences car pricing. Cars with lower horsepower (0-100 hp) exhibit a narrow price range and lower median values, highlighting their association with entry-level or economy vehicles. As the horsepower increases, there is a noticeable rise in both the median price and price variability. From the 251-300 hp range onward, the price distribution widens significantly, marked by several high-price outliers that suggest the presence of high-performance and luxury models. A steep increase is particularly evident in the 351-400 hp range, which aligns with premium vehicles aimed at professional use. Vehicles with horsepower exceeding 451+ hp occupy an exclusive segment of high-end luxury or sports cars, as reflected by their elevated median prices and the lack of lower-priced models in this category.

Figure 5(b) shifts focus to the price distribution across different engine size ranges. Cars equipped with smaller engines, such as those in the 0-1000cc and 1001-1500cc ranges, show a more compact price spread and lower medians, reflecting their appeal as economical and budget-friendly options. However, as the engine size increases to 2001-2500cc and beyond, the price variability grows significantly. The 3501-4000cc range, in particular, stands out with a notable number of high-price outliers and a higher median price, signifying their alignment with premium and performance-oriented segments. Meanwhile, cars with engine sizes exceeding 4500cc cater to a niche market of specialized vehicles, often commanding the highest price points in the dataset.

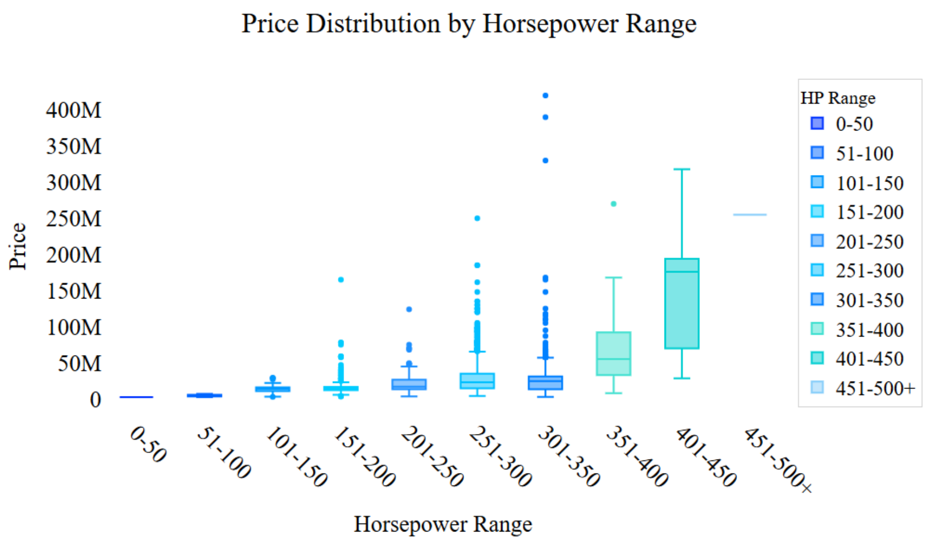

**Figure 5(a): Graph of price against horsepower range**

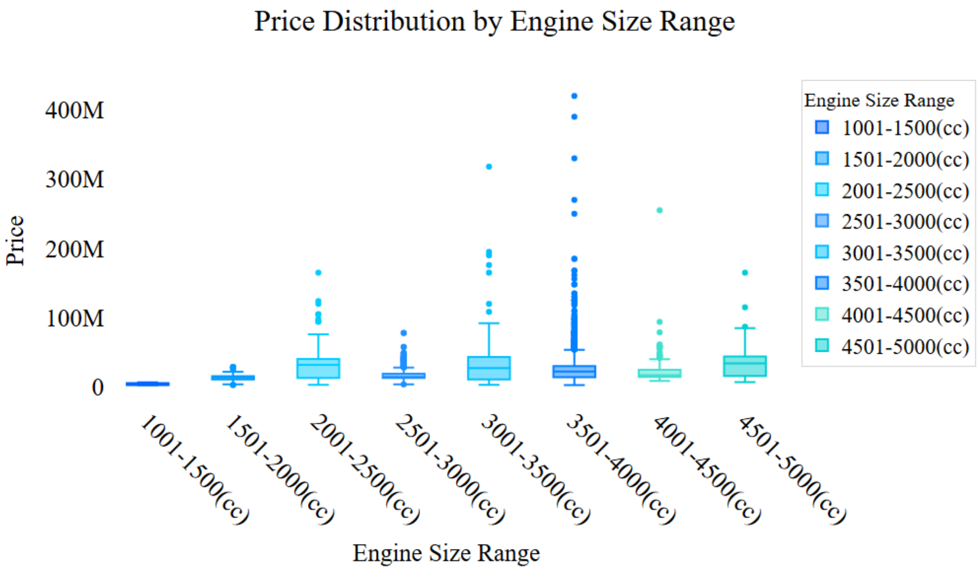

**Figure 5(b): Graph of price against engine size range**

In this study, the dataset contains several categorical variables such as brand, model, and others, which need to be transformed into numerical values for machine learning models to effectively predict car prices. To achieve this, One-Hot Encoding was applied to convert these categorical features into numerical form. The dataset was then split into training and testing sets, with 70% used for training and 30% for testing. The next step involved analyzing the correlation between different features using Pearson's correlation coefficient, visualized through heatmaps as shown in Figure 6.

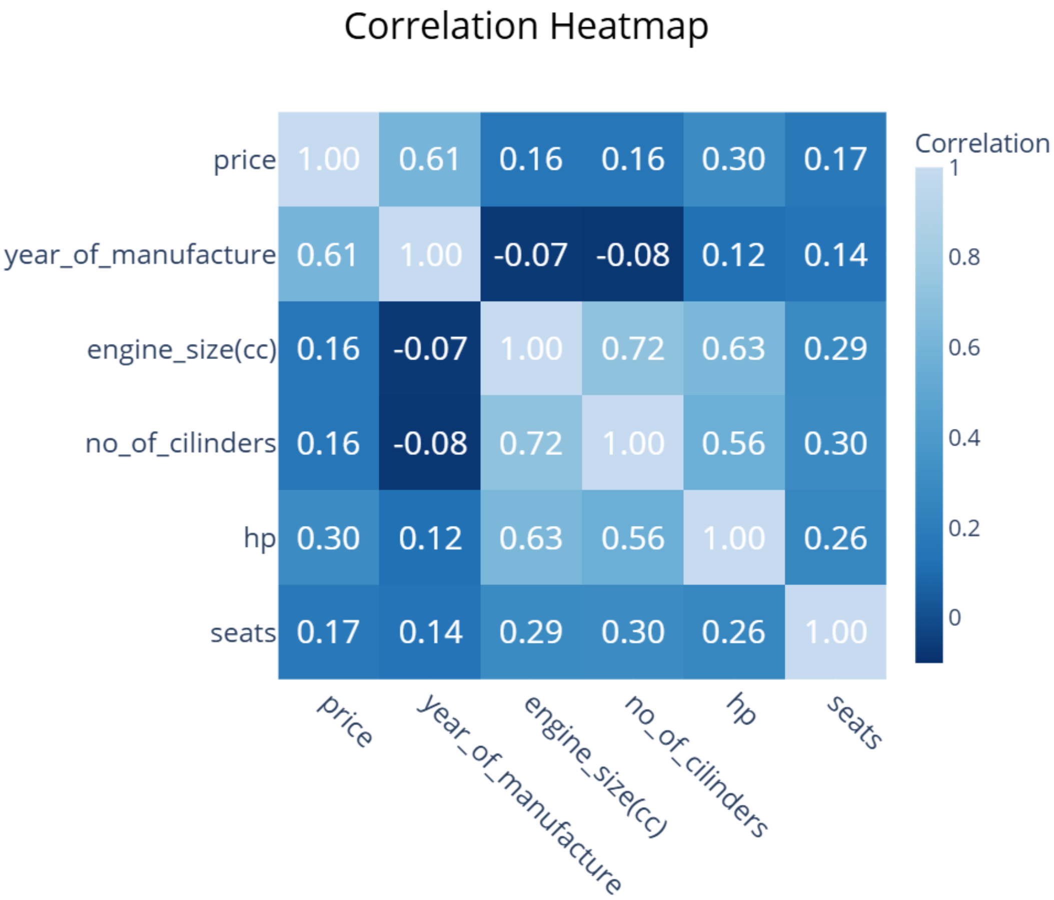

**Figure 6: Heatmap between the important features in the data**

The correlation heatmap reveals insightful relationships between the features in the dataset. As indicated in Figure 6, the price is positively correlated with several variables, such as the year of manufacture, engine size, and horsepower (hp). Among these, the strongest positive correlation is observed with the year of manufacture (0.61), indicating that newer cars tend to have higher prices. This aligns with expectations, as newer models often feature advanced technologies and better overall performance, making them more valuable in the market. Similarly, engine size and horsepower demonstrate moderate positive correlations with price, at 0.30 and 0.63, respectively. These findings suggest that vehicles with larger engines and higher power outputs are generally priced higher, reflecting consumer preferences for performance-oriented cars. On the other hand, features like the number of seats and cylinders show weaker correlations with price, implying that while these attributes might influence consumer choices, they have a relatively limited impact on the overall valuation of a car.

Interestingly, the relationship between some features, such as engine size and the number of cylinders, shows a strong interdependence (0.72). This highlights a potential redundancy, as these features may provide overlapping information about the vehicle's performance characteristics. Therefore, dimensionality reduction techniques or careful feature selection may be considered to streamline the model's input without compromising predictive accuracy. The heatmap helps us understand which features are most important for predicting car prices. Features like the year of manufacture, engine size, and horsepower are more important, while others have less impact. This analysis makes it easier to choose the right features for building accurate models, reducing irrelevant information, and improving predictions.

The car price prediction model was deployed as an interactive web application using Streamlit, chosen for its simplicity and ability to create user-friendly interfaces. The app allows users to input vehicle attributes through numerical fields, dropdowns, and checkboxes for binary options (e.g., "Has Sunroof: Yes/No"), organized for easy navigation. User inputs are processed into a format compatible with the XGBoost model, utilizing one-hot encoding for categorical features, and predictions are generated in real time. This deployment shows how the model can be used in real life, making it easy for both individuals and businesses to get car price predictions. The simple web-based interface makes the tool practical and easy to use.

In analyzing the key factors influencing car value, feature importance was determined using both the XGBoost and Random Forest models, as shown in Figures 7 and 8. Each model assessed the relative significance of various features, assigning scores that reflect their contribution to predictive performance. These importance scores provide insight into which attributes most impact car valuation, offering a great understanding of the model's decision-making process. The XGBoost and Random Forest results align on several critical features, reinforcing the robustness of these attributes in forecasting car prices.

The analysis revealed several significant insights as highlighted below:

i. **Year of Manufacture:** As shown in Figure 7, this emerged as the most influential feature with an importance score of 0.12. Newer cars generally have higher prices, likely because they retain more value over time, making this a crucial factor in car price determination.

ii. **Horsepower:** With a high importance score of 0.10, horsepower is the second most critical feature. Higher horsepower generally correlates with increased performance, which positively influences a car's market value, especially for sports or luxury models.

iii. **Engine Size (cc):** This feature ranks as the third most important factor in predicting car price. It has an importance score of 0.05. Cars with larger engine volumes, typically associated with higher power and more premium models, tend to have higher prices.

iv. **Seats:** The number of seats, also important with a score of 0.04. Vehicles with greater seating capacity, such as SUVs and minivans, often have different market values compared to standard 4- or 5-seat vehicles.

v. **No Faults (Condition):** This feature has an importance score of 0.035. It indicates whether the car has "No Faults," suggesting a binary condition (Yes/No). Cars without reported faults are likely to command higher prices, making this a significant predictor in determining car value.

Additionally, other features such as colour, interior condition, and specific model details showed varying degrees of importance in predicting car prices, though their impact was relatively lower compared to the key features mentioned above. The convergence of feature importance scores across ensemble methods (XGBoost and Random Forest) indicates model robustness and confirms the validity of these features across non-parametric, tree-based algorithms. Tables 1(a) and 1(b) list the five most significant and five least significant features. Many of the most important features are commonly inquired about by potential car buyers in Nigeria, representing fundamental aspects of a car. In contrast, the least important features are more abstract, such as colour and some less common car models.

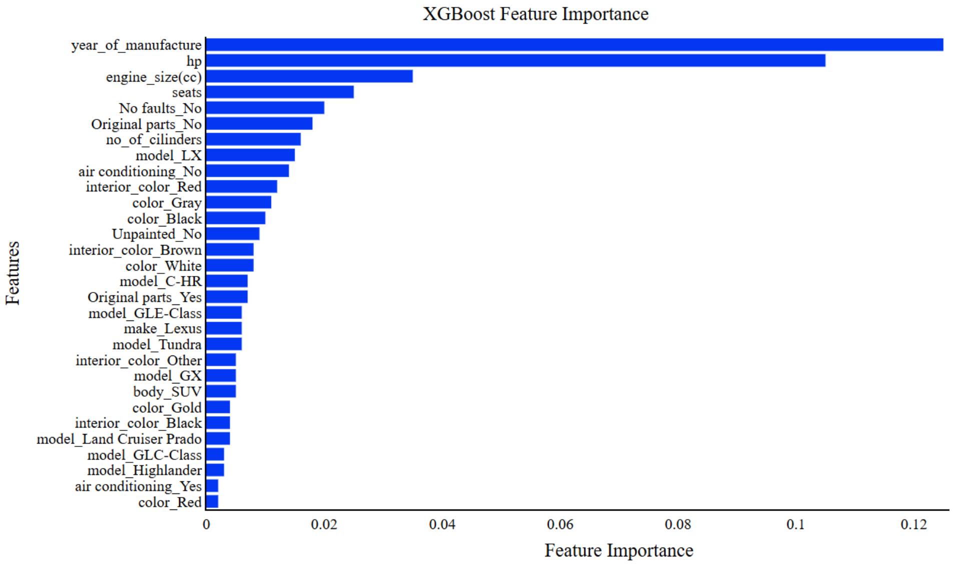

**Figure 7: XGBoost Features**

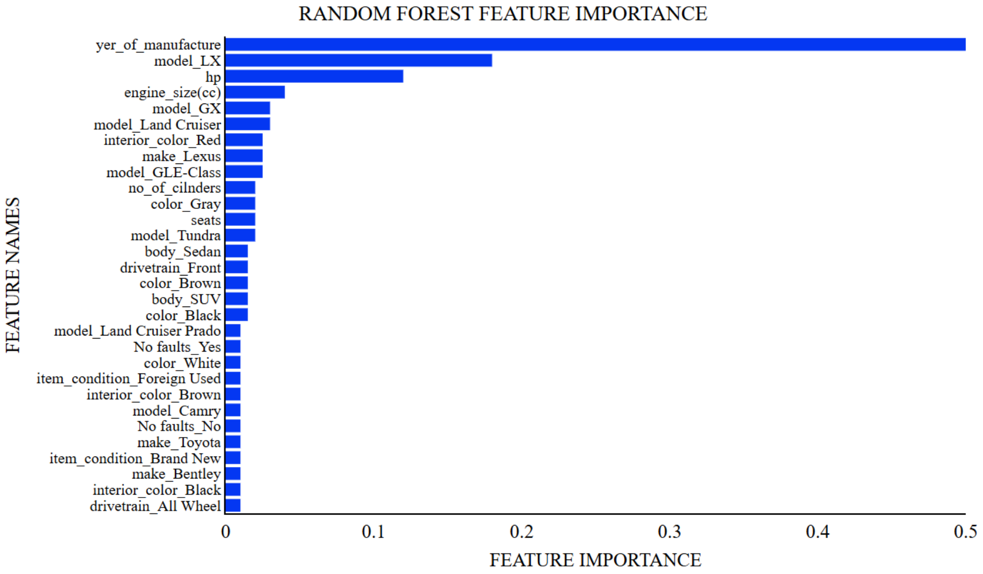

**Figure 8: Random Forest Feature Importance**

Table 1a: Most Significant Features

| Feature | Importance |
|---|---|
| Year of Manufacture | 0.12 |
| Hp | 0.10 |
| Engine Size | 0.05 |
| Seats | 0.04 |
| Number of Faults | 0.035 |

Table 1(b): Least Significant Features

| Feature | Importance |
|---|---|
| Color: Red | 0.005 |
| Air Conditioner | 0.005 |
| Model: High Lander | 0.005 |
| Model: GLC: Class | 0.005 |
| Model: Land Cruiser | 0.005 |

This study evaluates the performance of four regression models Linear Regression, Random Forest Regressor, XGBoost, and MLP Regressor on a car price prediction task. The models were compared using performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ($R^2$). As shown in Table 2, XGBoost achieved the best overall performance with the lowest RMSE and MAE values, alongside the highest $R^2$ score.

Table 2: The performance of different models on the car dataset

| **Methods** | **Linear Regression** | **Random Forest Regressor** | **XGBoost Regressor** | **MLP Regressor** |
|---|---|---|---|---|
| **Root Mean Squared Error** | 13,621,485.06 | 7,187,120.55 | 7,047,536.43 | 15,688,873.41 |
| **Mean Absolute Error** | 7,214,925.08 | 3,618,109.55 | 3,540,639.15 | 9,828,218.22 |
| **R-squared Error** | 0.4774 | 0.8556 | 0.8612 | 0.3121 |

Linear Regression demonstrated poor expressiveness, as it assumes a strictly linear relationship between the features and the target variable, failing to capture the nonlinear complexities in the dataset. The Random Forest and XGBoost models, however, achieved significantly better results by leveraging their ability to model nonlinear relationships and generalize effectively to unseen data. Among these, XGBoost showed superior performance due to its robust optimization algorithms and regularization capabilities. Although, the MLP Regressor is capable of capturing nonlinearities, its performance was suboptimal in this case. This could be attributed to insufficient hyperparameter tuning or the need for more extensive training data to better capture the underlying relationships.

## 4.0 CONCLUSION

This study has demonstrated the effectiveness of machine learning techniques in accurately predicting car prices in Nigeria, using real-world data and regression models. Four models—Linear Regression, Random Forest Regressor, XGBoost Regressor, and MLP Regressor—were evaluated based on their predictive performance. Among them, the XGBoost Regressor achieved the best results, recording a Root Mean Squared Error (RMSE) of 7,047,536.43, a Mean Absolute Error (MAE) of 3,540,639.15, and an R-squared value of 0.8612. These metrics reflect its superior ability to capture nonlinear patterns and deliver highly accurate price predictions.

Key features identified as critical to price determination—such as Year of Manufacture (importance score: 0.12), Horsepower (0.10), and Engine Size (0.05)—were consistently ranked highest in importance across both XGBoost and Random Forest models. These attributes directly reflect what consumers value in the Nigerian car market: newer models, performance capabilities, and utility. The inclusion of a condition indicator such as No Faults (0.035) also underscores the market’s sensitivity to vehicle reliability and maintenance history. The consistent ranking of these features strengthens the reliability and interpretability of the models.

The comparison with other models further emphasizes XGBoost’s advantage. While the Random Forest Regressor also performed well with an RMSE of 7,187,120.55 and $R^2$ of 0.8556, the MLP Regressor and Linear Regression models lagged significantly. The MLP Regressor posted a high RMSE of 15,688,873.41 and $R^2$ of 0.3121, indicating insufficient generalization, likely due to under-tuning and limited data. The Linear Regression model performed the worst, with an RMSE of 13,621,485.06, reflecting its inability to model the nonlinear nature of the dataset and signalling a breakdown in its assumptions.

In conclusion, the study successfully validates the application of machine learning, particularly XGBoost, in predicting car prices in Nigeria with both accuracy and interpretability. These findings have practical implications for car dealerships, digital valuation tools, and online sales platforms seeking data-driven pricing solutions. Future work could enhance model performance through deeper neural networks, more diverse datasets, and real-time data integration. Ultimately, this research contributes valuable insights to both the technical and commercial domains of vehicle price estimation in emerging markets.

## ACKNOWLEDGEMENT

The authors acknowledge the staff of Mechatronics Engineering Department of the Federal University of Agriculture, Abeokuta for providing the providing the platform for the research.

## AUTHORS’ CONTRIBUTIONS STATEMENT

SOO: Conceptualization, Investigation and Research supervision. FOD: Review & editing of initial write-up. SOA: Data collection and Methodology. KMM: Formal Analysis. OEF: Writing original draft. All authors read and approved the final manuscript.

## DATA AVAILABILITY

Datasets generated or analysed during the current study will be made available on request.

## STATEMENTS AND DECLARATIONS

The authors declare that this manuscript is original and has not been published previously nor is it under consideration for publication elsewhere. All data, results, and interpretations presented in this work are the outcome of the authors' independent research efforts.

## ETHICAL

The current study did not include any human or animal subjects. Thus, this study is not subject to an ethics review committee and does not require any informed consent.

## COMPETING INTERESTS

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

## REFERENCES

Aderibigbe O. O., Fadare S. O., and Gumbo T. (2024). "Transport situation in the Global South: Insights from Nigeria, South Africa and India," in Emerging Technologies for Smart Cities: Sustainable Transport Planning in the Global North and Global South, Cham: Springer Nature Switzerland, pp. 43-77.

Agarwal P., Abudu D., Calabrese L., and Chukwurah O. (2023). "The automotive sector in Nigeria: Opportunities under the AfCFTA," ODI Research Report.

Agrahari K., Chaubey A., Khan M., and Srivastava M. (2021). "Car price prediction using machine learning," *Int. J. Innov. Res. Technol.,* vol. 8, no. 1, pp 572-575.

Alita D., Putra A. D., and Darwis D. (2021). "Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* vol. 15, no. 3, pp. 295-306.

Chandak A., Ganorkar P., Sharma S., Bagmar A., and Tiwari S. (2019) "Car price prediction using machine learning," *International Journal of Computer Sciences and Engineering,* vol. 7, no. 5, pp. 444-450.

Hatta M., Wahid W. N., Yusuf F., Hidayat F., Santoso N. A., and Aini Q. (2024). "Enhancing predictive models in system development using machine learning algorithms," *International Journal of Cyber and IT Service Management*, vol. 4, no. 2, pp. 80-87.

Heitz T., He N., Ait-Mlouk A., Bachrathy D., Chen N., Zhao G., and Li L. (2025). "Investigation on eXtreme Gradient Boosting for cutting force prediction in milling," *Journal of Intelligent Manufacturing*, vol. 36, no. 1, pp. 285-301.

Jain N. and Jana P. K. (2023) "LRF: A logically randomized forest algorithm for classification and regression problems," *Expert Systems with Applications,* vol. 213, 119225.

Kovpak E. and Orlov F. (2019). "Comparative analysis of machine learning models and regressions for car price prediction," Bulletin of VN Karazin Kharkiv National University Economic Series, no. 97, pp. 31-40, 2019.

Maleki A., Raahemi M., and Nasiri H. (2023). "Breast cancer diagnosis from histopathology images using deep neural network and XGBoost," Biomedical Signal Processing and Control, vol. 86, 105152, 2023.

Mienye I. D. and Sun Y. (2022). "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access,* vol. 10, pp. 99129-99149.

Pal N., Arora P., Kohli P., Sundararaman D., and Palakurthy S. S. (2019). "How much is my car worth? A methodology for predicting used cars' prices using random forest," in Advances in Information and Communication Networks: *Proceedings of the 2018 Future of Information and Communication Conference (FICC)*, vol. 1, Springer International Publishing, pp. 413-422.

Pan R., Bagherzadeh M., Ghaleb T. A., and Briand L. (2022). "Test case selection and prioritization using machine learning: A systematic literature review," *Empirical Software Engineering*, vol. 27, no. 2, p. 29.

Peerun S., Chummun N. H., and Pudaruth S. (2015). "Predicting the price of second-hand cars using artificial neural networks," *in The Second International Conference on Data Mining, Internet Computing, and Big Data* (BigData2015), vol. 17.

Putra P. H., Azanuddin A., Purba B., and Dalimunthe Y. A. (2024). "Random forest and decision tree algorithms for car price prediction," *Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA)*, vol. 4, no. 1, pp. 81-89.

Rane N. L., Paramesha M., Choudhary S. P., and Rane J. (2024). "Machine learning and deep learning for big data analytics: A review of methods and applications," *Partners Universal International Innovation Journal*, vol. 2, no. 3, pp. 172-197.

Samruddhi K. and Kumar R. A. (2020). "Used car price prediction using K-nearest neighbor based model," *Int. J. Innov. Res. Appl. Sci. Eng. (IJIRASE),* vol. 4, no. 3, pp. 686.

Singh P., Adebanjo A., Shafiq N., Razak S. N. A., Kumar V., Farhan S. A., and Sergeevna M. T. (2024). "Development of performance-based models for green concrete using multiple linear regression and artificial neural network*," International Journal on Interactive Design and Manufacturing* (IJIDeM), vol. 18, no. 5, pp. 2945-2956.

Watson C., Cooper N., Palacio D. N., Moran K., and Poshyvanyk D. (2022). "A systematic literature review on the use of deep learning in software engineering research," *ACM Transactions on Software Engineering and Methodology (TOSEM),* vol. 31, no. 2, pp. 1-58.

Zhang Y., Liu J., and Shen W. (2022). "A review of ensemble learning algorithms used in remote sensing applications*," Applied Sciences*, vol. 12, no. 17, 8654.