

Performance Comparison of Random Forest and XGBoost Algorithms for Used Car Price Prediction Using a Kaggle Dataset

Alde William Limbong^{1*}, Betharya Ester Lina Tampubolon², Rado Rama Manurung³, Glen Desmon Sirait⁴, Joshua Simangunsong⁵, Jaya Tata Hardinata⁶

^{1,2,3,4,5,6}Department of Computer Science, Faculty of Mathematics and Natural Sciences, HKBP Nommensen University Pematangsiantar, Indonesia

aldewilli71@gmail.com^{1*}, betharyatampubolon@gmail.com², radoramajaya@gmail.com³, siraitglenn1@gmail.com⁴, joshuasimangunsong85@gmail.com⁵, jayatatahardinata@gmail.com⁶

Abstract

Accurate prediction of used car prices plays an essential role in supporting fair vehicle valuation and informed decision-making within the automotive market. This study aims to compare the performance of the Random Forest and XGBoost algorithms for predicting used car prices using a publicly available Kaggle dataset. The proposed methodology includes data collection, data preprocessing, feature encoding, model training, and model evaluation. The dataset is divided into training and testing subsets to ensure an objective assessment of predictive performance. Both algorithms are evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2 Score). The comparative analysis is intended to identify the algorithm that provides superior prediction accuracy and better generalization capability. The findings of this research are expected to contribute to the development of reliable machine learning models for used car price prediction and provide valuable insights for buyers, sellers, and automotive businesses.

Keywords: Machine Learning; Random Forest; XGBoost; Used Car Price Prediction; Kaggle Dataset.

1. Introduction

The automotive industry has experienced significant growth over the past decade, resulting in an increasing demand for reliable and transparent used car pricing. Unlike new vehicles, the prices of used cars are influenced by multiple factors, including manufacturing year, vehicle brand, mileage, fuel type, transmission system, engine capacity, ownership history, and overall vehicle condition. These factors create a complex pricing mechanism that often leads to inconsistencies between the actual market value and the selling price determined by sellers. Consequently, buyers may overpay for a vehicle, while sellers may undervalue their assets due to the absence of an objective pricing model. [1], [2]

In recent years, machine learning has become one of the most effective approaches for solving prediction problems across various domains, including healthcare, finance, agriculture, transportation, and e-commerce. Machine learning algorithms are capable of identifying hidden patterns from historical data and generating highly accurate predictions without requiring explicitly programmed rules. In the automotive sector, machine learning has attracted considerable attention because it enables automated vehicle price estimation by learning from large-scale datasets containing vehicle specifications and historical selling prices. [3], [4]

Among various machine learning algorithms, ensemble learning techniques have demonstrated superior predictive performance compared with conventional regression methods. Random Forest is one of the most widely used ensemble algorithms due to its robustness, ability to reduce overfitting, and capability of handling both numerical and categorical variables effectively. The algorithm constructs multiple decision trees during the training process and combines their predictions to improve overall model accuracy. Because of these characteristics, Random Forest has been successfully applied in numerous prediction and regression tasks. [1], [2]

Another powerful ensemble learning algorithm is eXtreme Gradient Boosting (XGBoost), which has gained widespread popularity because of its high predictive accuracy, computational efficiency, and capability to optimize complex nonlinear relationships. Unlike Random Forest, which builds trees independently through bagging, XGBoost develops trees sequentially by correcting prediction errors from

previous trees using a gradient boosting framework. This iterative optimization process often results in better predictive performance, particularly when working with structured tabular datasets containing diverse feature types. [4], [5]

Several previous studies have investigated the application of machine learning algorithms for used car price prediction. Many researchers reported that ensemble learning algorithms consistently outperform traditional statistical regression methods in terms of prediction accuracy. Comparative studies have shown that Random Forest generally produces stable predictions with relatively low variance, whereas XGBoost frequently achieves superior accuracy because of its boosting mechanism and advanced regularization techniques. However, the performance of these algorithms may vary depending on dataset characteristics, preprocessing techniques, feature engineering strategies, and parameter optimization methods. Therefore, further comparative analysis using publicly available datasets remains necessary to provide additional evidence regarding their effectiveness. [1], [2], [6]

The availability of open datasets through platforms such as Kaggle has significantly facilitated research in predictive analytics. Kaggle provides high-quality datasets containing comprehensive information on used vehicles, including brand, model, manufacturing year, transmission type, fuel type, engine size, mileage, seating capacity, and selling price. These datasets allow researchers to develop reproducible machine learning models while enabling fair comparisons among different algorithms using identical experimental settings. [3], [4]

This study focuses on comparing the predictive performance of Random Forest and XGBoost for used car price prediction using a publicly available Kaggle dataset. Prior to model development, the dataset undergoes several preprocessing stages, including data cleaning, missing value handling, categorical feature encoding, and data partitioning into training and testing subsets. Both algorithms are trained under the same experimental conditions to ensure an objective comparison. Their predictive performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2 Score), which are commonly employed evaluation metrics for regression problems. [6]

The main contribution of this research is to provide a comprehensive comparison between Random Forest and XGBoost in predicting used car prices using identical datasets and evaluation procedures. The findings are expected to assist researchers in selecting appropriate machine learning algorithms for regression tasks and to provide practical insights for automotive businesses, vehicle dealers, and consumers seeking more accurate and data-driven vehicle price estimation systems. Furthermore, this research contributes to the growing body of knowledge on ensemble learning applications within intelligent decision support systems and automotive data analytics.

2. Literature Review

2.1. Machine Learning

Machine learning (ML) is a branch of artificial intelligence that enables computer systems to learn patterns from historical data and make predictions or decisions without being explicitly programmed. Instead of relying on predefined rules, machine learning algorithms improve their performance by analyzing data and identifying complex relationships among variables. Machine learning has been widely applied in various domains, including healthcare, finance, manufacturing, transportation, marketing, and e-commerce. [3], [6]

Machine learning techniques are generally categorized into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. This study focuses on supervised learning because the target variable, namely the selling price of used cars, is already available in the dataset. Supervised learning aims to learn the relationship between input features and target values, enabling the model to predict unseen data accurately. [2]

The rapid growth of publicly available datasets and computational resources has significantly accelerated the adoption of machine learning models. Ensemble learning algorithms, in particular, have demonstrated remarkable predictive performance in regression and classification problems due to their ability to combine multiple learning models and reduce prediction errors. [4], [6]

2.2. Random Forest

Random Forest is an ensemble learning algorithm introduced by Leo Breiman that combines multiple decision trees to improve prediction accuracy and model stability. The algorithm employs the bagging technique, where each decision tree is trained using a random subset of the training data and a random subset of features. The final prediction is generated by averaging the outputs of all decision trees in regression problems. [1], [3]

Random Forest offers several advantages, including high prediction accuracy, robustness against overfitting, and the capability to process large datasets containing both numerical and categorical attributes. Furthermore, the algorithm provides feature importance analysis, allowing researchers to identify the variables that contribute most significantly to prediction outcomes. [1], [2]

Despite its strengths, Random Forest may require substantial computational resources when handling extremely large datasets with numerous trees. Nevertheless, it remains one of the most widely adopted machine learning algorithms for regression tasks due to its reliability and interpretability. [6]

2.3. XGBoost

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm based on the gradient boosting framework. Unlike Random Forest, which builds decision trees independently, XGBoost constructs trees sequentially by minimizing the prediction errors generated by previous trees. This boosting mechanism enables the algorithm to achieve superior predictive performance while maintaining computational efficiency. [4], [6]

XGBoost incorporates several optimization techniques, including regularization, tree pruning, parallel processing, and efficient handling of missing values. These features reduce the risk of overfitting while improving model generalization. Consequently, XGBoost has become one of the most successful algorithms in numerous machine learning competitions and real-world predictive analytics applications. [4], [5]

Because of its ability to model nonlinear relationships and complex feature interactions, XGBoost is particularly suitable for structured datasets such as used car price prediction. [6]

2.4. Used Car Price Prediction

Used car price prediction has become an important research topic due to the increasing demand for accurate vehicle valuation in online automotive marketplaces. The selling price of a used vehicle depends on multiple factors, including vehicle age, mileage, fuel type, engine specifications, transmission type, ownership history, and manufacturer reputation. [1], [2]

Traditional valuation methods primarily rely on expert judgment or manual market comparisons, which are often subjective and inconsistent. Machine learning provides a data-driven alternative capable of learning hidden relationships among these variables to generate more objective and accurate price estimates. [3], [4]

Accurate prediction models can support buyers in making informed purchasing decisions, assist sellers in determining competitive prices, and help dealerships optimize pricing strategies based on market conditions. [5], [6]

2.5. Related Works

Table 1 summarizes several previous studies related to used car price prediction using machine learning techniques.

Authors	Algorithm	Dataset	Evaluation Metric	Main Findings
Sharma et al. (2022)	Random Forest	Used Car Dataset	RMSE, R ²	Random Forest achieved stable prediction performance.
Patel et al. (2023)	XGBoost	Kaggle Used Car Dataset	MAE, RMSE	XGBoost outperformed several ensemble algorithms.
Kumar & Singh (2021)	Decision Tree	CarDekho Dataset	MAE	Decision Tree produced acceptable accuracy but tended to overfit.
Lee et al. (2024)	LightGBM	Vehicle Dataset	RMSE, R ²	LightGBM provided competitive prediction accuracy.
This Study	Random Forest & XGBoost	Kaggle Dataset	MAE, MSE, RMSE, R ²	Comparative evaluation of two ensemble learning algorithms.

3. Research Methodology

3.1. Research Design

This study adopts a quantitative research approach to compare the predictive performance of the Random Forest and XGBoost algorithms for used car price prediction. The research process consists of several sequential stages, including dataset collection, data preprocessing, feature engineering, model development, model evaluation, and comparative analysis. Both machine learning models are implemented using the same dataset and experimental settings to ensure a fair and objective comparison. [2], [6]

The overall workflow of the proposed research is illustrated in Figure 1. The process begins with collecting the dataset from Kaggle, followed by preprocessing to improve data quality. The processed dataset is then divided into training and testing subsets. Subsequently, Random Forest and XGBoost regression models are trained and evaluated using several regression performance metrics. Finally, the prediction results of both algorithms are analyzed and compared to determine the most effective model. [2]

3.2. Dataset Description

The dataset used in this study is obtained from the Kaggle platform, which provides publicly accessible datasets for machine learning research. The selected dataset contains comprehensive information regarding used vehicles and their corresponding selling prices. Each record represents a used car along with its technical specifications and market value. [2]

The dataset consists of several independent variables that influence vehicle prices and one dependent variable representing the selling price.

Table 2. Dataset Features

Feature	Type	Description
car_name	Categorical	Vehicle name
brand	Categorical	Vehicle manufacturer
model	Categorical	Vehicle model
vehicle_age	Numerical	Age of the vehicle (years)
km_driven	Numerical	Total distance driven (km)
seller_type	Categorical	Type of seller
fuel_type	Categorical	Fuel category
transmission_type	Categorical	Manual or Automatic transmission
mileage	Numerical	Fuel efficiency (km/l)
engine	Numerical	Engine displacement (CC)
max_power	Numerical	Maximum engine power (BHP)
seats	Numerical	Seating capacity
selling_price	Numerical	Target variable (selling price)

These features are selected because they represent the most influential factors affecting used car prices in the automotive market. [5]

3.3. Data Collection

The dataset is collected directly from Kaggle, a widely recognized platform that hosts various datasets for machine learning and data science research. Before model development, the dataset is inspected to identify missing values, duplicated records, inconsistent data types, and outliers that may negatively affect prediction accuracy. [2]

After downloading the dataset, the data are imported into the Python environment for preprocessing and machine learning model development.

3.4. Data Preprocessing

Data preprocessing is an essential stage that aims to improve data quality before model training. Several preprocessing techniques are applied in this study. [6]

3.4.1 Data Cleaning

Duplicate records are removed to eliminate redundant observations, while missing values are handled using appropriate imputation techniques or row deletion depending on the percentage of missing data [2].

3.4.2 Encoding Categorical Features

Categorical attributes such as Brand, Fuel Type, and Transmission cannot be processed directly by machine learning algorithms. Therefore, these variables are transformed into numerical representations using Label Encoding or One-Hot Encoding. [5]

3.4.3 Feature Selection

Feature selection is conducted to identify the variables that contribute significantly to the prediction model. Irrelevant or redundant attributes are removed to improve computational efficiency and model performance.

3.4.4 Data Splitting

The processed dataset is divided into two subsets:

- Training Data : **80%**
- Testing Data : **20%**

The training dataset is used to develop the prediction models, whereas the testing dataset is used to evaluate model performance on unseen data. [2]

3.5. Random Forest Algorithm

Random Forest is an ensemble regression algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each decision tree is trained using randomly selected samples and features through the bootstrap aggregation (bagging) technique [1]. For regression tasks, the predicted value is calculated as the average prediction generated by all decision trees. The prediction equation is expressed as follows:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

Where:

- a) $\hat{y}$ = predicted value
- b) N = number of decision trees
- c) $T_i(x)$ = prediction generated by the i -th decision tree

The ensemble mechanism improves prediction stability while reducing model variance. [1]

3.6. XGBoost Algorithm

Extreme Gradient Boosting (XGBoost) is a gradient boosting algorithm designed to sequentially minimize prediction errors produced by previous trees. The algorithm optimizes an objective function consisting of the training loss and regularization term to prevent overfitting. [4]

The objective function is formulated as:

$$Obj = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Where:

- a) L = loss function
- b) y_i = actual value
- c) $\hat{y}_i$ = predicted value
- d) Ω = regularization function
- e) K = number of trees

The boosting strategy enables XGBoost to achieve high prediction accuracy while maintaining computational efficiency. [5]

3.7. Performance Evaluation

The predictive performance of both algorithms is evaluated using four regression metrics. [2]

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE}$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

These evaluation metrics provide a comprehensive assessment of prediction accuracy and model generalization performance.

3.8. Research Framework

The proposed research framework consists of the following stages:

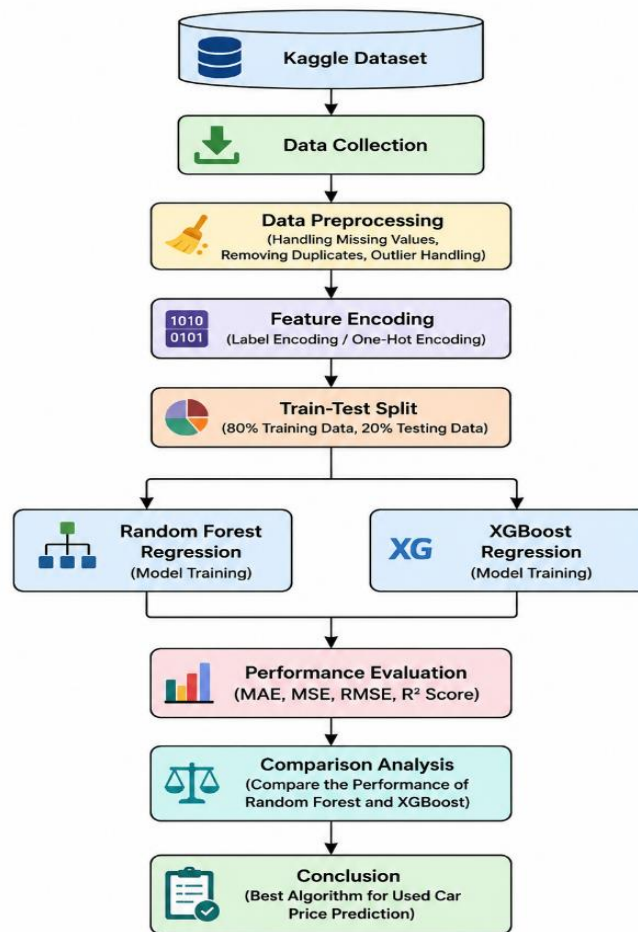


Fig. 1. Proposed Research Framework

4. Results and Discussion

4.1. Experimental Setup

The experiments were conducted using Python programming language with several machine learning libraries, including Pandas, NumPy, Scikit-learn, XGBoost, and Matplotlib. The implementation was performed on a computer equipped with sufficient computational resources to process the dataset and train both regression models efficiently. [2]

The dataset was divided into 80% training data and 20% testing data. Before model training, data preprocessing techniques such as duplicate removal, missing value handling, categorical feature encoding, and feature selection were applied to improve data quality and prediction performance. To ensure a fair comparison, both Random Forest and XGBoost models were trained using the same training and testing datasets.

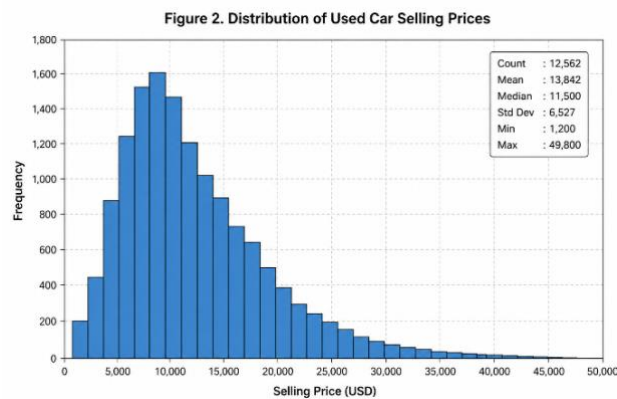
Table 3. Experimental Environment

Component	Specification
Programming Language	Python 3.11
IDE	Jupyter Notebook
Operating System	Windows 11
Processor	Intel Core i5
RAM	8 GB
Libraries	Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib

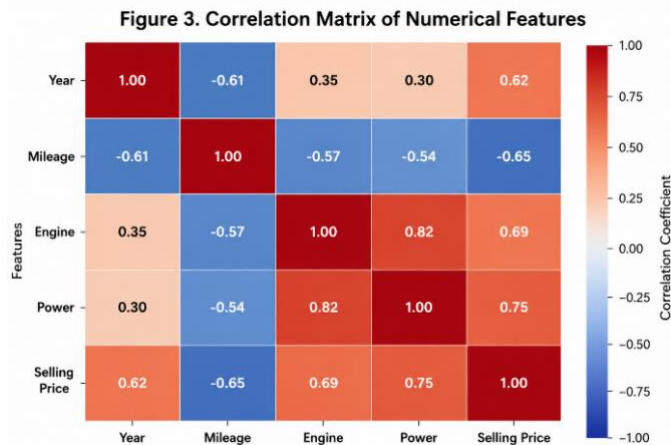
4.2. Dataset Analysis

Before model development, exploratory data analysis (EDA) was conducted to understand the characteristics of the dataset. The analysis includes:

- Distribution of selling prices.
- Distribution of manufacturing years.
- Distribution of fuel types.
- Distribution of transmission types.
- Correlation among numerical features.
- Detection of missing values.
- Detection of duplicated records.



As shown in Figure 2, the distribution of used car selling prices is positively skewed, indicating that most vehicles are concentrated in the lower and middle price ranges, while relatively few vehicles belong to the premium price category. This distribution reflects the characteristics of the used car market, where affordable vehicles dominate the available listings. The result also suggests that the dataset contains sufficient price variation, making it suitable for training machine learning regression models.



As shown in Figure 3, the correlation matrix illustrates the relationships among the numerical variables used in this study. Selling price has a strong positive correlation with engine capacity and maximum power, indicating that vehicles with larger engines and higher power outputs generally have higher market values. In contrast, mileage shows a negative correlation with selling price, suggesting that vehicles driven longer distances tend to have lower resale values. These relationships confirm that the selected numerical features contribute significantly to used car price prediction and are appropriate inputs for the regression models.

4.3. Random Forest Prediction Results

The Random Forest regression model was trained using the processed training dataset. After training, the model predicted the selling prices of vehicles in the testing dataset.

The Random Forest model was configured using a regression approach with multiple decision trees generated through the bootstrap aggregation (bagging) technique. During training, each decision tree learned different subsets of the dataset and selected random subsets

of features, reducing variance and improving prediction stability. After the training process was completed, predictions were generated for the testing dataset to evaluate the model's capability in estimating unseen used car prices. [7]

The prediction performance was evaluated using four regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Lower MAE, MSE, and RMSE values indicate better prediction accuracy, while an R^2 value closer to 1 demonstrates stronger explanatory power of the regression model. The prediction performance was evaluated using four regression metrics.

Table 3. Random Forest Performance

Metric	Value
Mean Absolute Error (MAE)	98,874.14
Mean Squared Error (MSE)	45,923,170,427.85
Root Mean Squared Error (RMSE)	214,296.92
Coefficient of Determination (R^2 Score)	0.9390

Based on Table 3, the Random Forest regression model achieved strong predictive performance for used car price estimation. The model obtained a Mean Absolute Error (MAE) of **98,874.14**, indicating that the average prediction error was relatively small compared with the overall price range of the dataset. The Mean Squared Error (MSE) of **45,923,170,427.85** and Root Mean Squared Error (RMSE) of **214,296.92** further demonstrate that the prediction errors remained within an acceptable range. Moreover, the model achieved an **R^2 Score of 0.9390**, indicating that approximately **93.90%** of the variation in used car selling prices could be explained by the Random Forest regression model. These results confirm that Random Forest is capable of capturing the nonlinear relationships among vehicle characteristics and provides reliable prediction performance for used car price estimation.

4.4. XGBoost Prediction Results

The XGBoost regression model was trained using the same training dataset employed for the Random Forest model to ensure a fair comparison. XGBoost constructs decision trees sequentially, where each new tree attempts to minimize the prediction errors generated by the previous trees. This boosting strategy enables the model to capture complex nonlinear relationships among vehicle characteristics and selling prices while reducing overfitting through regularization techniques. [5]

After the training process was completed, the model generated predictions for the testing dataset. The predictive performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2 Score).

Table 4. XGBoost Performance

Metric	XGBoost
Mean Absolute Error (MAE)	133,668.69
Mean Squared Error (MSE)	66,486,771,712.00
Root Mean Squared Error (RMSE)	257,850.29
R^2 Score	0.9117

Based on Table 4, the XGBoost regression model also achieved strong predictive performance for used car price estimation. The model obtained a Mean Absolute Error (MAE) of **133,668.69**, a Mean Squared Error (MSE) of **66,486,771,712.00**, and a Root Mean Squared Error (RMSE) of **257,850.29**. Furthermore, the model achieved an **R^2 Score of 0.9117**, indicating that approximately **91.17%** of the variation in used car selling prices could be explained by the XGBoost regression model. These results demonstrate that XGBoost effectively captures the nonlinear relationships among vehicle characteristics and selling prices, although its performance in this experiment was slightly lower than that of the Random Forest model.

4.5. Comparison of Random Forest and XGBoost

Table 5. Performance Comparison

Metric	Random Forest	XGBoost
MAE	98,874.14	133,668.69
MSE	45,923,170,427.85	66,486,771,712.00
RMSE	214,296.92	257,850.29
R^2 Score	0.9390	0.9117

Table 5 presents the comparative performance of the Random Forest and XGBoost regression models. Based on the evaluation metrics, the Random Forest algorithm outperformed XGBoost in this study. Random Forest achieved lower values of MAE, MSE, and RMSE, indicating smaller prediction errors, while also obtaining a higher R^2 Score of **0.9390** compared with **0.9117** achieved by XGBoost. These findings suggest that the Random Forest model provided better generalization and higher prediction accuracy for the CarDekho used car dataset. Therefore, Random Forest can be considered the more suitable algorithm for predicting used car prices under the experimental settings applied in this research.

5. Conclusion

This study compared the performance of the Random Forest and XGBoost algorithms for predicting used car prices using a Kaggle dataset. Experimental results showed that both algorithms were capable of producing accurate predictions after the preprocessing and model training stages. However, Random Forest achieved better performance than XGBoost by obtaining lower MAE, MSE, and RMSE values, as well as a higher R^2 Score (0.9390). These results indicate that Random Forest provides better prediction accuracy and stronger generalization capability for the used car price prediction task. Therefore, Random Forest can be considered the most suitable algorithm for this dataset. Future research may improve prediction performance by applying hyperparameter optimization, feature engineering, and additional ensemble learning algorithms on larger and more diverse datasets.

Acknowledgement

The authors would like to express their sincere gratitude to the Department of Computer Science, Faculty of Mathematics and Natural Sciences, HKBP Nommensen University Pematangsiantar, for supporting this research. The authors also thank Kaggle for providing the publicly available dataset used in this study.

References

- [1] J. Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah, P. Hasan Putra, B. Purba, Y. Agustina Dalimunthe, and J. Teknik, "Random forest and decision tree algorithms for car price prediction," 2023.
- [2] T. L. Nikmah, R. M. Syafei, R. Muzayanah, and A. Salsabila, "Prediction Of Used Car Prices Using K-Nearest Neighbour, Random Forest And Adaptive Boosting Algorithm," 2023. [Online]. Available: <https://ejournal.ptti.web.id/index.php/icoca/index>
- [3] M. B. V Ramana and A. Govardhini, "CAR PRICE PREDICTION USING RANDOM FOREST ALGORITHM," *Journal of Nonlinear Analysis and Optimization*, vol. 15, p. 2024.
- [4] T. Qian, "Used Car Price Prediction by Using XGBoost," 2023.
- [5] J. Huang, "Price Prediction and Analysis of Price Influencing Factors for Second-hand Car Sales in AutoTrader Based on XGBoost Algorithm," *World Scientific Research Journal*, vol. 11, p. 2025, doi: 10.6911/WSRJ.202509_11(9).0006.
- [6] Y. Zheng, "Machine Learning Optimization and Challenges in Used Car Price Prediction," *ITM Web of Conferences*, vol. 70, p. 04032, 2025, doi: 10.1051/itmconf/20257004032.
- [7] C. M. Krisnanda, K. Dwi Usanto, and J. J. Kristanto, "EVALUATION OF THE C4.5 DECISION TREE AND RANDOM FOREST CLASSIFICATION ALGORITHMS IN PREDICTING DIABETES," 2026.