IT Journal Research and Development (ITJRD)
Vol.9, No.2, March 2025, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2025.18031

Forecasting Used Car Prices Using Machine Learning

Eni Khusnul Khotimah[1], Dwiretno Istiyadi Swasono[2], Gama Wisnu Fajarianto[3]
Department of Informatics, Faculty of Computer Science, University of Jember[1,2,3]
202410103006@mail.unej.ac.id[1], istiyadi@unej.ac.id[2], gamawisnuf@unej.ac.id[3]

Article Info

Article history:

Received Jul 11, 2024
Revised Oct 22, 2024
Accepted Mar 3, 2025

Keyword:

Car price prediction
Machine Learning
Artificial Neural Network
Random Forest Regression
Mean Absolute Error (MAE)

ABSTRACT

In an increasingly competitive era, it is crucial for car dealers and retailers to address the challenges of accurately determining the prices of used cars. To tackle these challenges, this study implements Machine Learning models to predict used car prices accurately. By applying the Artificial Neural Network (ANN) and Random Forest Regression algorithms, this research aims to evaluate the performance of these methods in predicting used car prices. The used car price data was obtained from the Kaggle repository, consisting of 14,657 data entries that provide comprehensive information about used cars. The analysis focuses on six main columns, including Brand, Model, Variant, Year, and Mileage, to estimate used car prices. Model evaluation was conducted using Mean Absolute Error (MAE) as the primary metric. The results show that the ANN model achieved a lower MAE (0.035) compared to the Random Forest Regression (0.047), indicating better performance in predicting used car prices. These findings demonstrate the effectiveness of ANN in handling data complexity and the non-linear relationships between variables involved in forecasting used car prices. Additionally, this contributes to the implementation of more accurate used car price predictions, enabling automotive companies to improve operational efficiency and provide greater benefits to the community.

Corresponding Author:

Eni Khusnul Khotimah
Department of Informatics
University of Jember
Jl. Kalimantan, Jember, Indonesia
Email: 202410103006@mail.unej.ac.id

1. INTRODUCTION

As times evolve, car manufacturers compete to release the latest models and innovate to create technology that attracts buyers. This situation impacts the availability of used cars that are still in good condition, as upper-class buyers seek cars with up-to-date features and technology. Dealers and retailers face the main challenge of accurately pricing the used cars they sell. Price plays a crucial role in the sales process, and to address this, companies need the ability to predict used car prices accurately based on their specifications. Implementing this price prediction capability not only enhances operational efficiency and achieves substantial profits but also benefits society at large [1]. To tackle this challenge, car dealers and retailers must leverage technological advancements, including the application of Machine Learning techniques, to address issues related to used car pricing. In this context, Machine Learning can be used to analyze various relevant factors, including estimating used car prices. By applying Machine Learning, companies can build accurate price prediction models using historical data and significant variables. This model will contribute to

determining the appropriate selling price for used cars based on their characteristics. It is well-known that the value of a used car is influenced by various factors. The most crucial factors typically include Brand, Model, Variant, Year, and the number of kilometers driven [2].

According to research conducted by [3] titled "Short Term Load Forecasting Based on ARIMA and ANN Approaches," it examines electricity demand forecasting that requires an accurate and sustainable data acquisition system using smart grids. This study compares the performance of ARIMA and Artificial Neural Network (ANN) in predicting daily electricity demand for 709 households in Ireland over an 18-month period, with results showing that ANN outperforms in handling non-linear load data.

In a similar vein, research conducted by [4] titled "Artificial Neural Network: An Innovative Approach in Air Pollutant Prediction for Environmental Applications - A Review," it discusses the use of Artificial Neural Network (ANN) for accurately predicting air pollutants based on various forecasting intervals. This study shows that ANN is superior in predicting air contaminants compared to traditional methods, as it can better handle various input meteorological parameters.

Furthermore, research by [5] titled "COVID-19 Prevalence Forecasting Using Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN): Case of Turkey," it examines the forecasting of COVID-19 prevalence in Turkey. This study compares the performance of ARIMA and Artificial Neural Network (ANN) in predicting cases of infection, death, and recovery, with results indicating that both techniques are highly successful in estimating prevalence in Turkey.

In the context of the automotive industry, research by [6] titled "Car Resale Price Forecasting: The Impact of Regression Method, Private Information, and Heterogeneity on Forecast Accuracy," it compares 19 regression methods, including Lasso, Random Forest, ANN, and SVR, to predict the resale value of used cars. The results indicate that Random Forest is very effective, with an approximate 7% increase in prediction accuracy under low specification conditions. RF and Exponential Smoothing (ES) are identified as highly suitable modeling techniques, with RF being easier to implement. This research has important implications for management and decision-making in the used car industry.

Lastly, research by [7] titled "Forecasting Stock Prices of Fintech Companies of India Using Random Forest with High-Frequency Data," it develops a stock forecasting model for three leading fintech companies in India. The results indicate that the Random Forest Regression method is the most effective, with a prediction accuracy rate exceeding 80% and an accuracy increase of about 85-90% for price forecasts over a 20-day period.

Based on the explanation above, this research proposes the title "Forecasting Used Car Prices with Machine Learning" because it aims to develop a Machine Learning model with a low error rate capable of estimating used car prices based on their specifications. This research is expected to provide insights to used car market participants and facilitate smarter decision-making in the automotive industry. Predictions do not have to be absolutely precise, but close enough to the actual value according to the existing situation. By comparing two Machine Learning algorithms, namely Artificial Neural Network (ANN) and Random Forest Regression, for the task of predicting used car prices, ANN is an attractive choice due to its ability to handle complex data and mimic the workings of the human nervous system [8]. Meanwhile, Random Forest Regression is chosen for its ability to handle overfitting and its good performance in used car price regression tasks [9]. The used car price data is obtained from the Kaggle repository, an online data-sharing platform. The model will be evaluated using the Mean Absolute Error (MAE) score, and the data will be divided into training and testing sets to achieve the best used car price regression results. This research is expected to contribute to the understanding of the ever-evolving automotive industry.

2. THEORETICAL BASIS

In this section, several theories related to the topic will be discussed to provide a more comprehensive understanding of the subject, specifically focusing on their application to the case study of used car price prediction.

2.1 Forecasting

Forecasting refers to the process of predicting future needs, including the quantity, quality, time, and location required to meet the demand for goods or services [10]. It plays a critical role in decision-making, where the challenge is to estimate what will happen in the future based on current or historical data [11]. In the context of used car price prediction, forecasting is used to predict the future prices of vehicles based on past sales data, market trends, and various car features.
To achieve accurate forecasting results, two essential components must be considered:

1. Data Validity: The accuracy of the forecasting model relies heavily on the validity of the data used. In this study, car features such as brand, year of manufacture, mileage, and condition are essential inputs for predicting prices.
2. Appropriate Forecasting Methods: Machine Learning algorithms, such as Random Forest Regression and Artificial Neural Networks (ANN), are employed to predict car prices, as they are capable of capturing complex patterns in data that traditional statistical methods may miss [12].

Used car price forecasting typically follows a trend pattern due to the depreciation of car value over time, along with seasonality that might arise from fluctuations in demand during certain months of the year. The following are the types of forecasting patterns that can be seen in the figure:

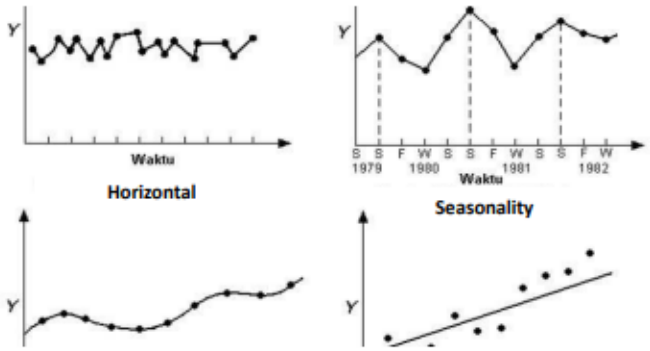

Figure 1. Types of Forecasting Patterns

Here are several data patterns:

1. Trend (T) occurs when there is a gradual increase or decrease in the data movement over a long period of time.
2. Seasonality (S) refers to a recurring pattern in the data after a certain period, such as daily, weekly, monthly, quarterly, or yearly.
3. Cycles (C) represent a data pattern that occurs every few years, usually influenced by long-term economic fluctuations related to business cycles.
4. Horizontal (H) / Stationary occurs when data values fluctuate around a stable average, remaining stationary relative to the average.

2.2 Artificial Intelligence

Artificial Intelligence (AI) is a field of science focused on creating machines capable of solving complex problems with behavior similar to that of humans. AI leverages characteristics of human intelligence and implements them into machine algorithms. AI programs utilize heuristic search techniques to enhance efficiency and shorten the search process for problem-solving. AI systems must possess extensive knowledge to apply it effectively. AI can perform various tasks that are too complex or time-consuming for humans, thereby providing significant assistance in human work [13]. In this study, AI methods are used to analyze large datasets of used car sales and identify patterns that influence price changes. By applying AI techniques, such as Machine Learning, the prediction models are capable of adapting to varying market conditions and customer preferences.

2.3 Machine Learning

Machine Learning (ML) is a component of the field of Artificial Intelligence (AI) that utilizes data, statistical methods, and trained algorithms to perform classification, prediction, or clustering tasks [14]. Machine Learning is a subset of Artificial Intelligence that enables systems to complete given tasks efficiently. Machine Learning involves learning from existing data to generate algorithms capable of making predictions. The success of Machine Learning heavily depends on the quality of the data used. Machine Learning algorithms help detect data relationships between different attributes associated with a particular attribute, thus enhancing understanding through predictive models. For the used car price prediction case, ML algorithms are used to analyze historical data on car prices and detect patterns between different car attributes (e.g., brand, mileage, condition, etc.) and the final sale price.

The Machine Learning process involves seven steps: data collection, data preprocessing, model selection, training phase, performance evaluation, optimization, and testing for prediction analysis [15]. The steps involved in developing a machine learning model for this study include data collection, preprocessing, model selection, training, performance evaluation, and prediction. The quality of the input data, such as the completeness and accuracy of car feature information, greatly influences the model's ability to predict prices with precision.

2.4 Random Forest Regression

Random Forest Regression (RFR) is an ML technique that can be used for both classification and regression problems. Random Forest consists of a combination of multiple decision trees. The given data is split into subsets, and each subset is used to build a decision tree that makes its own decisions. The final decision or prediction is determined based on the majority vote from all the trees [14]. Random Forest Regression is also a classification method that employs an ensemble approach, consisting of several decision trees with different characteristics. The ensemble method is a technique where multiple different models are trained to solve similar problems, and their results are combined to achieve the best outcome. In the application of Random Forest Regression, multiple decision trees are created. Each tree is built using different observations and predictors obtained through sampling, and the best results are sought. Additionally, the implementation of Random Forest is relatively straightforward, making the process fairly fast. For these reasons, this model is often used by researchers in the development of Machine Learning applications [16]. The random forest algorithm is an ensemble classification method that combines decision trees by taking a majority vote to predict classifications, thereby preventing overfitting [17]. Nodes from the decision trees are randomly selected from a subset of variables and then used as candidates to find the best split [18]. Random forest has several advantages, including high accuracy and effectiveness when working with large databases [19]. In addition, Random Forest Regression is a widely-used machine learning technique for regression tasks, including predicting used car prices. This algorithm combines multiple decision trees to improve the accuracy of predictions by averaging the results from several trees. In the context of predicting used car prices, each tree in the random forest is trained on different subsets of the data, capturing various aspects of the car market. This allows the algorithm to provide more accurate predictions than a single decision tree model by reducing overfitting. Key advantages of Random Forest include its ability to handle large datasets, efficiently process many input variables (such as car features), and manage missing data without compromising performance. For example, in predicting used car prices, Random Forest can efficiently manage a large dataset with various car models and attributes while reducing errors from overfitting, leading to more reliable predictions.

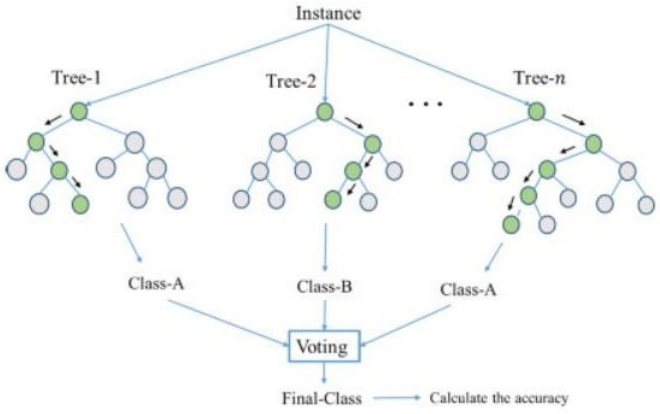

Figure 2. Random Forest

Random Forest offers several advantages, including high accuracy, effective handling of large databases, and the ability to manage thousands of input variables quickly and efficiently. This method minimizes errors, provides satisfactory classification performance, and can handle missing values in the data. Random Forest also provides techniques for estimating incomplete data and identifying important variables in the classification process [20]. It is more resistant to overfitting, typically when minority data is accidentally duplicated into the majority data area [21]. The Random Forest algorithm involves the following steps:

1. Determine the number of trees (k) to be created by randomly selecting from the total available features, where k must be less than the total number of features (m).
2. Take random samples based on the number N in the dataset for each tree to be created.
3. Each tree randomly selects a subset of predictors, with the number of predictors chosen being less than the total number of predictors (p).
4. The sampling and subset selection processes are repeated for the predetermined number of k trees.
5. The final classification prediction is obtained by taking the majority vote from all the trees created.

2.5 Artificial Neural Network

An Artificial Neural Network (ANN) is a model inspired by the structure of the brain, consisting of three main layers: the input layer, hidden layers, and the output layer. ANN processes data from the input layer through the hidden layers using trial-and-error techniques to generate predictions in the output layer. Each neuron in these layers is connected through weights that influence the information processing, mimicking the way the human brain learns from experience [22]. For predicting used car prices, the ANN learns from historical data by adjusting the weights between nodes to minimize prediction errors. A Multi-Layer Perceptron (MLP) is a type of ANN that includes an input layer, several hidden layers, and an output layer. MLP operates through forward propagation, where input data is passed through the hidden layers using non-linear activation functions like ReLU for the hidden layers and a linear activation function for the output layer. Each neuron in the hidden layers applies a non-linear activation function, such as ReLU (Rectified Linear Unit), to capture complex patterns in the data. This is particularly useful in used car price prediction, as car prices are influenced by multiple interacting factors, such as age, mileage, brand, and condition.

1. **Forward Propagation**

   The process begins with input data (e.g., car features) passing through the network layers, and predictions are generated by the output layer. The process involves calculating the net input value $z_j^{(I)}$ using the formula:

$$z_j^{(I)} = \sum_{i=1}^{n^{(I+1)}} w_{ij}^{(I)} a_i^{(I-1)} + b_j^{(I)}$$

and applying the activation function:

$$a_j^{(1)} = \mathrm{ReLU}(z_j^{(1)})$$

The final output is calculated to determine the model's prediction. For example, when predicting a used car price, inputs like the car's brand, age, and mileage are transformed through several layers of neurons to produce the predicted price.

2. Backpropagation

After forward propagation, MLP uses backpropagation to update the weights and biases with the goal of minimizing the loss function, such as Mean Absolute Error (MAE), which is defined as:

$$L = \frac{1}{m} \sum_{i-m}^{m} |\hat{y}_i - y_i|$$

The gradient is calculated using the formula:

$$\delta_j^{(L)} = \hat{y}_j - y_j$$

and for the hidden layers:

$$\delta_j^{(L)} = \left(\sum_{k=1}^{n^{(I+1)}} \delta_k^{(I+1)} w_{jk}^{(I+1)} \right) f'(z_j^{(I)})$$

where is the derivative of the ReLU activation function. The weights and biases are updated using:

$$w_{ij}^{(I)} = w_{ij}^{(I)} - \eta \frac{\partial L}{\partial w_{ij}^{(I)}}$$

and :

$$b_j^{(I)} = b_j^{(I)} - \eta \delta_j^{(I)}$$

where η is the learning rate. This process is repeated over several epochs to train the model until convergence is achieved.

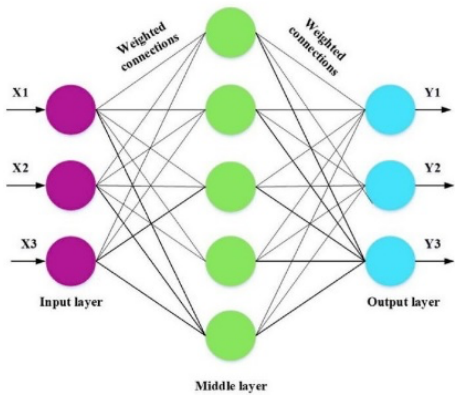

Figure 3. ANN Model Structure

The goal of optimization is to find the best solution to engineering problems through mathematical functions, aiming for either a minimum or maximum objective. Traditional methods such as gradient descent have been replaced by metaheuristic approaches like Genetic Algorithm,

Spotted Hyena Optimizer, Multi-verse Optimizer, and Red Fox Optimizer (RFO). RFO, as a nature-inspired algorithm, offers ease of implementation and effectiveness in managing both continuous and discrete parameters. This method is used in this study to enhance energy demand prediction [22].

3. RESEARCH METHOD

This study employs both qualitative and quantitative research methods. Qualitative research is applied during the stages of identifying research needs, conducting literature reviews, and analyzing research journals. Quantitative research is implemented during the stages of data calculation and processing based on the case study dataset. The calculations use Artificial Neural Network and Random Forest Regression algorithms to measure the error rates in implementing used car price prediction through predetermined attribute parameters. The study also involves experimental testing of regression algorithms and comparing several models to find the one with the lowest error rate for predicting car prices.

3.1 Data Collection

This study uses a dataset from Kaggle, consisting of 14,657 data entries with 16 columns. To focus on the six main variables considered important for price prediction, we selected the following columns: Price, Brand, Model, Variant, Year of Manufacture, and Mileage. This data spans used cars from 2010 to 2019. These six variables were chosen due to their significance in determining the resale value of cars. Although variables such as brand and model may have a smaller impact compared to other factors, they are still important as they contribute to buyer preferences and market value. Through this selection process, we filtered the data down to 308 more specific entries, facilitating a more accurate analysis for predicting used car prices. With this approach, we aim to improve the accuracy and relevance of used car price predictions in the market.

Table 1. Dataset Attributes

Attribute	Count
Brand	5
Model	5
Variant	67
Year	10
Mileage	46

3.2 Pre-Processing

This study begins with data cleaning, data transformation, and normalization to ensure uniform scaling, which supports optimal model performance, especially for algorithms sensitive to attribute scales. The combination of these three steps ensures optimal training data and produces a model capable of providing accurate and reliable predictions.

1. Data cleaning: Data cleaning is the first step in the data preprocessing process. Its purpose is to ensure data quality before it is used in analysis models or Machine Learning. One key aspect of data cleaning is handling missing values, as these can impact the model's accuracy.
2. Data transformation: Data transformation is the second step in preprocessing, aimed at obtaining data that is more aligned with the research needs. One important step in data transformation is feature selection, which involves evaluating the features in the dataset using algorithms such as SelectKBest and Recursive Feature Elimination (RFE) to select the best features. The goal is to enhance modeling efficiency and avoid overfitting. In addition to feature selection, data transformation also involves encoding, which converts categorical data into numerical representations that can be processed by Machine Learning models. The combination of feature selection and encoding ensures that the data used in modeling is more optimal and suitable for analysis needs.

a. SelectKBest is a score-based feature selection method in which features are selected based on scores calculated using a specific metric. The goal is to choose the top k features according to the scores provided by the given metric.

```
from sklearn.feature_selection import SelectKBest, mutual_info_regression
import pandas as pd

# Pisahkan antara fitur dan target variable
X = df.drop(['Harga'], axis=1)  # Menghapus kolom 'Harga' dari dataframe dan menyimpannya sebagai fitur
y = df["Harga"]  # Menyimpan kolom 'Harga' sebagai target variabel (y)

# Seleksi fitur dengan mutual information
selector = SelectKBest(score_func=mutual_info_regression, k=5)  # Membuat objek SelectKBest dengan score
X_selected = selector.fit_transform(X, y)  # Fit dan transform data untuk hanya menyertakan fitur yang t

# Mendapatkan indeks fitur yang terpilih
selected_feature_indices = selector.get_support(indices=True)  # Mendapatkan indeks fitur yang terpilih

# Mendapatkan nama fitur yang terpilih
selected_feature_names = X.columns[selected_feature_indices]  # Mendapatkan nama fitur yang terpilih ber

# Print nama fitur yang terpilih
print("Selected Feature Names:", selected_feature_names)  # Mencetak nama fitur yang terpilih

Selected Feature Names: Index(['Merek', 'Model', 'Varian', 'Tahun', 'Jarak tempuh'], dtype='object')
```

Figure 4. SelectKBest

Based on figure 4 is the implementation code for feature selection using shared information regression with the scikit-learn library. First, the features and target variable are separated from the dataframe. Then, SelectKBest is used to select the top 5 features based on mutual information scores. The `fit_transform(X, y)` method of the selector object is applied to include only the selected features. `selected_feature_indices` is used to obtain the indices of the selected features, and `selected_feature_names` retrieves the names of the features based on those indices for printing.

b. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a feature selection method that works iteratively by removing features one by one based on their contribution to the resulting Machine Learning model. This process is repeated until the desired number of features is achieved. RFE is generally used with models that can estimate the relative importance of each feature, such as decision trees, linear regression, or logistic regression.

```
# Pisahkan antara fitur dan target variable
X = df.drop(['Harga'], axis=1)  # Menghapus kolom 'Harga' dari dataframe dan menyimpannya sebagai
y = df["Harga"]  # Menyimpan kolom 'Harga' sebagai target variabel (y)

# Inisialisasi model RandomForestRegressor
model = RandomForestRegressor()

# Inisialisasi RFE dengan model RandomForestRegressor dan memilih 5 fitur terbaik
selector = RFE(model, n_features_to_select=5)

# Fit model RFE pada data dan transformasi data untuk hanya menyertakan fitur yang dipilih
X_selected = selector.fit_transform(X, y)

# Mendapatkan indeks fitur yang terpilih
selected_feature_indices = selector.get_support(indices=True)

# Mendapatkan nama fitur yang terpilih berdasarkan indeks fitur
selected_feature_names = X.columns[selected_feature_indices]

# Print hasil
print("Selected Features:", selected_feature_names)

Selected Features: Index(['Merek', 'Model', 'Varian', 'Tahun', 'Jarak tempuh'], dtype='object')
```

Figure 5. Recursive Feature Elimination

Based on figure 5 is the code for Recursive Feature Elimination (RFE) with the RandomForestRegressor model to select the top 5 features from the dataframe. After separating the features and target variable from the dataframe, the RandomForestRegressor model is initialized. The RFE selector is then used with `n_features_to_select=5` to choose the best features. The selected data, `X_selected`, is generated using `fit_transform(X, y)` to

include only the features selected by RFE. The result is the names of the selected features, which are printed using a print statement.

Table 2. Dataset with Feature Selection

Attribute	SelectKBest	(RFE)
Location		
Seller		
Brand	✓	
Model	✓	
Variant		✓
Year	✓	✓
Mileage		✓
Fuel Type		
Color		
Transmission		
Body Type	✓	✓
Engine Capacity	✓	✓
Seller Type		
Drive System		
Car Auction Name		

Based on table 2, it can be concluded that for feature selection, SelectKBest selects 5 attributes: Brand, Model, Year, Body Type, and Engine Capacity, while RFE selects 5 attributes: Variant, Year, Mileage, Body Type, and Engine Capacity. In feature selection, the five most relevant attributes are chosen using SelectKBest and RFE, namely Brand, Model, Variant, Year, and Mileage. This decision is based on the combination of the results from both methods, which are considered the most significant. The next step is Label Encoding, which converts categorical variables into numerical form by assigning a unique integer to each category.

3. Normalization: Normalization is the process of adjusting values in a dataset to achieve a uniform or standardized scale. The goal of normalization is to ensure that variables with different scales do not dominate the modeling or data analysis process. Normalization is commonly applied to numerical data, particularly features with a wide range of values. A commonly used normalization technique is Min-Max Scaling. In Min-Max Scaling, each value in the dataset is adjusted to fall within a specific range, usually between 0 and 1.

3.3 Model Building

The evaluation phase of this study uses Mean Absolute Error (MAE) as the primary metric to measure the forecasting errors of used car prices, calculated by taking the absolute difference between predicted values and actual values for each observation. The forecasting performance is assessed based on the lower MAE value, reflecting the closeness of predictions to the actual values. This research compares the predicted values of the model with the actual values and focuses on methods that yield lower MAE to determine the best method for forecasting used car prices. The evaluation is conducted using the Random Forest Regression and Artificial Neural Network (ANN) algorithms in Google Colab, involving a single test with a dataset consisting of five attributes and one label, and focuses the analysis on the two main algorithms according to the specific characteristics of the dataset relevant to the experimental objectives.

3.4 Feature Evaluation

Feature evaluation is the process of analyzing the contribution of each feature in the dataset to the performance of the model or the prediction results. The main objective is to identify the most relevant features for analysis or prediction, as well as to reduce the dimensionality of the dataset by retaining the most important features.

```
import matplotlib.pyplot as plt
import numpy as np

feature_names = np.array(['Merek', 'Model', 'Varian', 'Tahun', 'Jarak tempuh'])

# Normalisasi fitur menggunakan MinMaxScaler
scaler_X = MinMaxScaler()
X_scaled = scaler_X.fit_transform(X)

# Normalisasi target menggunakan MinMaxScaler
scaler_y = MinMaxScaler()
y_scaled = scaler_y.fit_transform(y.reshape(-1, 1)).flatten()

model_rf = RandomForestRegressor()
model_rf.fit(X_scaled, y_scaled)

# Mengambil feature importance dari model dan mengubahnya menjadi persentase
feature_importances = model_rf.feature_importances_ * 100  # Mengubah ke persentase

# Menampilkan feature importance dengan nama kolom aslinya pada sumbu x
plt.figure(figsize=(10, 6))  # Sesuaikan ukuran gambar sesuai kebutuhan
bars = plt.bar(feature_names, feature_importances)

# Tambahkan teks persentase pada setiap bar
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2.0, yval, f"{yval:.2f}%", va='bottom')  # Format persentase

plt.xlabel('Fitur')
plt.ylabel('Importance (%)')
plt.title('Feature Importance in Percentage')
plt.show()
```

Figure 6. Feature evaluation

Based on figure 6 feature evaluation is conducted using the Random Forest Regressor model to calculate feature importance after training on the training data (X_scaled, y_scaled). Feature importance indicates how important each feature is in making predictions, and the bar plot visualizes these values. Features with higher importance are considered more significant in the prediction results. This evaluation provides insights into the relative contribution of features and aids in decision-making related to feature selection.

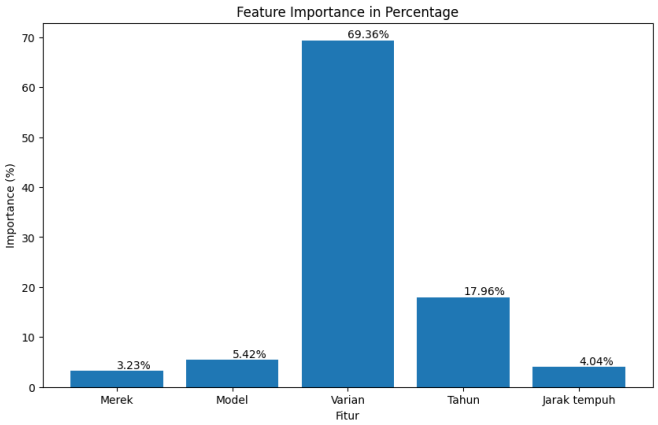

Figure 7. Feature Importance

In the feature importance analysis, the results indicate that variant and year have a significant impact on the prediction model, with weights of 69.36% and 17.96%, respectively. Meanwhile, brand, model, and mileage have lower influences, with weights of approximately 3.23%, 5.42%, and 4.04%, respectively. Thus, the primary focus in developing the prediction model should be on variant and year, while brand, model, and mileage have a smaller impact on the prediction results.

3.5 Methodological Limitations.

This study acknowledges several limitations related to the methodologies used. One potential limitation is the quality and completeness of the dataset obtained from Kaggle, which could introduce bias, especially in predicting car prices across different brands and models. Additionally, while feature selection methods such as SelectKBest and Recursive Feature Elimination were employed to optimize the model, the algorithms chosen Random Forest Regression and Artificial Neural Network also have their limitations. For example, Random Forest may not perform well on small datasets with noisy features, and the neural network can be prone to overfitting if not properly regularized. These issues could affect the generalizability of the model's predictions to different datasets or markets.

4. RESULTS AND ANALYSIS

The purpose of the performance evaluation is to assess how well the constructed model fits the data used. One common metric for evaluation is the Mean Absolute Error (MAE), which measures the average of the absolute differences between the model's predictions and the actual values. The results of the performance evaluation for the two different algorithms, Random Forest Regression and Artificial Neural Network, are as follows:

Table 3. Mean Absolute Error

Algoritma	Mean Absolute Error (MAE)
Random Forest Regression	0.047
Artificial Neural Network	0.035

The conclusion based on table 3 the performance evaluation of the models shows that the Artificial Neural Network (ANN) with an MAE of 0.035 has a lower average prediction error compared to the Random Forest Regression with an MAE of 0.047. This means that ANN predictions are closer to the actual values than those from Random Forest Regression. Further evaluation based on K-Fold cross-validation shows the following results:

Table 4. K-Fold Cross-Validation Results

Algorithm	K=3	K=5	K=7	K=10
Random Forest Regression	0.050	0.046	0.047	0.047
Artificial Neural Network	0.047	0.037	0.037	0.035

Based on table 4 the K-Fold cross-validation results show that ANN consistently delivers better performance across different K values, particularly at K=10, where it achieves the lowest MAE of 0.035. This further confirms ANN as the superior model for this task.

Performance Analysis Between Random Forest Regression and Artificial Neural Network (ANN):

1. Model Structure: One of the key differences between Random Forest Regression and ANN lies in their structural approach to learning from data. Random Forest Regression is an ensemble method that combines predictions from multiple decision trees, which makes it better at handling data with complex relationships but also more sensitive to noisy or extreme data points. On the other hand, ANN leverages interconnected layers of neurons to learn non-linear relationships in the data. This layered architecture allows ANN to capture intricate patterns in the data more effectively, making it well-suited for tasks with high-dimensional and non-linear relationships, such as predicting car prices.
2. Sensitivity to Outliers: Random Forest Regression tends to be more sensitive to outliers because it relies on decision trees, where extreme values can disproportionately affect the final outcome from several trees. Outliers may influence the decision-making process within the Random Forest, leading to less accurate predictions for some data points. In contrast, ANN, as a more flexible model, can absorb and reduce the impact of outliers more effectively. ANN's gradient-based optimization mechanism enables the model to adjust to extreme values, resulting in more stable and reliable predictions.

3. Learning Process: Random Forest models use a decision-making process that iteratively splits data based on certain features, which can lead to high variance if the data is not large enough or if there is noise. In contrast, ANN utilizes backpropagation to iteratively adjust weights based on errors, enabling it to generalize better on complex datasets. This learning mechanism helps ANN reduce bias in its predictions and improve accuracy, as seen in the lower MAE values across different K values in cross-validation.
4. Interpretability vs. Performance: Random Forest Regression is generally easier to interpret, as it can provide insights into which features are most important for making predictions. However, this interpretability comes at the cost of flexibility in handling complex non-linear relationships. ANN, although it provides higher prediction accuracy, is often viewed as a "black box" since the inner workings of the network and learned weights are less intuitive to interpret.
5. Computational Complexity: ANN models tend to require more computational power due to the large number of parameters and the need for extensive training iterations, especially for deep networks. While Random Forest, despite being computationally intensive due to the combination of multiple decision trees, usually requires less tuning and can be faster for medium-sized datasets. However, in cases where computational resources are available, the increased complexity of ANN can deliver better results, as demonstrated in this study.

Table 5. Example of Prediction Results from Both Methods

Brand	BMW	Toyota	Mitsubishi	Honda	Daihatsu
Model	Serie 3	Avanza	Pajero Sport	Mobilio	Xenia
Variant	320i Sport	E STD	Dakar 2.4	E	X
Year	2019	2019	2019	2018	2018
Mileage	0-5.000	0-5.000	0-5.000	0-5.000	10.000-15.000
Actual Price	715000000	175500000	475000000	173000000	139000000
Predicted ANN	718538513	186037769	491162834	176553764	137505645
Predicted RF	694728062	166819000	507755316	171625000	134647706
Mae ANN	3538513.0	10537769.0	16162834.0	3553764.0	1494355.0
Mae RF	20271938.0	8681000.0	32755316.0	1375000.0	4352294.0
% ANN	0.49%	6.00%	3.40%	2.05%	1.08%
% RF	2.84%	4.95%	6.90%	0.79%	3.13%

Table 5 shows a comparison of the predicted car prices from the two models (Artificial Neural Network and Random Forest Regression) against the actual values, along with the Mean Absolute Error (MAE) and the percentage of error for each model. The predictions from the Artificial Neural Network are generally closer to the actual values, especially for BMW and Daihatsu cars, with the lowest percentage errors of 0.49% and 1.08%, respectively. In contrast, the Random Forest Regression predictions exhibit higher percentage errors for some cars, such as Mitsubishi with an error of 6.90%. Therefore, it can be concluded that the Artificial Neural Network demonstrates a more consistent performance in predicting car prices compared to Random Forest Regression.

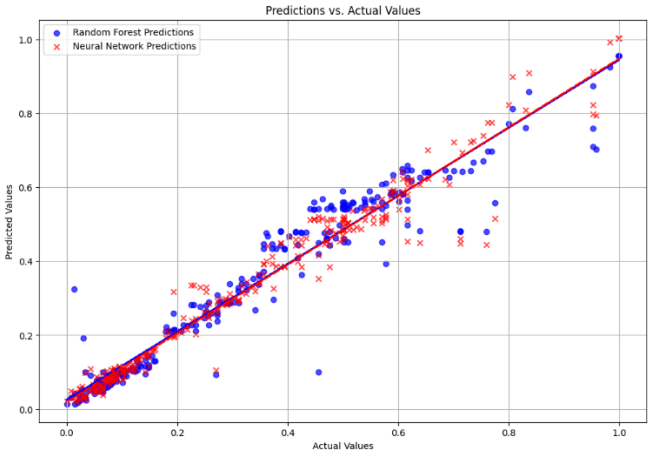

Figure 8. Comparison Chart

Based on figure 8 we can see a scatter plot comparing the predictions from the Random Forest and Artificial Neural Network models against the actual values. The x-axis represents the actual values while the y-axis represents the predicted values. The blue points represent the Random Forest Regression predictions and the red cross marks represent the Artificial Neural Network predictions. The blue and red dashed lines are the regression lines for each model, showing the best linear relationship between actual and predicted values. The gradients of the lines are rf = 0.9197 x + 0.02485 and ann = 0.9283 x + 0.01937. The closer these points are to the diagonal line, the more accurate the predictions are. The plot shows that both models perform well, but there is variation in their prediction distributions, with Random Forest exhibiting a wider spread compared to the Artificial Neural Network.

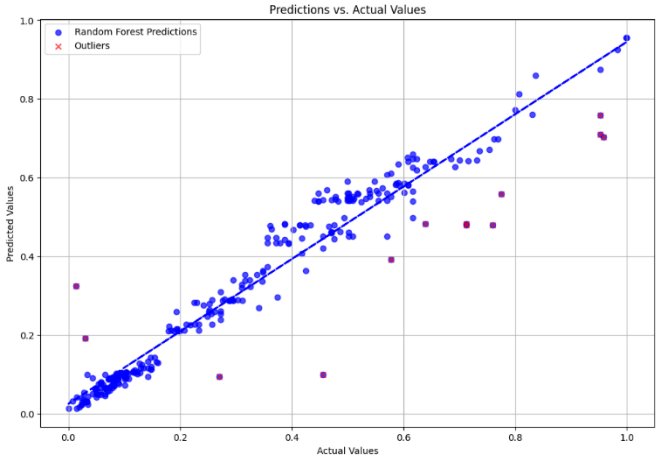

Figure 9. Random Forest Regression Chart

Based on figure 9 shows the scatter plot of the Random Forest Regression model's predictions versus actual values, with the x-axis representing the actual values and the y-axis representing the predicted values. The blue points represent pairs of actual and predicted values, and the blue dashed line is the regression line with the gradient p_rf = 0.9197 x + 0.02485. ost of the points are close to the diagonal line, indicating the accuracy of the model, although there are some outliers. Outliers are identified if | residual rf – mean residual | > 2 * standard deviation residuals, where the residual is the difference between the actual value and the predicted value. The indices of these outliers are used to obtain the standardized target values from the original data. The Random Forest Regression model identifies 13 outliers.

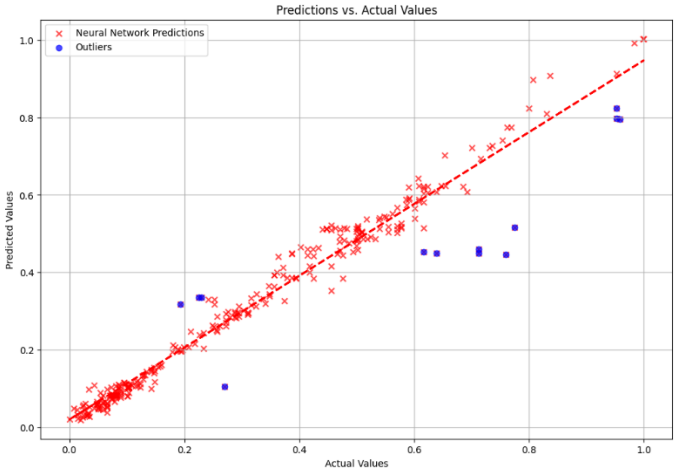

Figure 10. Artificial Neural Network Chart

Based on figure 10 shows the scatter plot of the Artificial Neural Network model's predictions versus actual values. The red cross points represent the pairs of actual and predicted values, and the red dashed line is the regression line with the gradient p_nn: 0.9283 x + 0.01937. Most of the points are close to the diagonal line, indicating the accuracy of the model, though there are some outliers, particularly at higher values. Outliers are identified as points where the residual is more than 2 standard deviations from the mean residual in the ANN model, where the residual is the difference between the actual value and the predicted value from the ANN model. Thus, outliers are data points with residuals significantly different from the predicted values, exceeding the set threshold. In this Artificial Neural Network model, there are 13 outliers.

Table 6. Outlier Analysis Table

Specifications	Analysis
BMW Serie 3 320i 2.0 2014 25.000-30.000 km 490.000.000	The car price in the dataset was recorded at 490,000,000 in 2014, while in 2015 it was 435,000,000. The car price appears to be quite high for its production year. In 2015, the price of a similar car dropped to 435,000,000, indicating a significant price decrease. There may be additional factors contributing to this price reduction, such as changing market conditions, decreased demand, or specific factors related to the condition of the car.
BMW Serie 3 320i E90 LCI 2012 35.000-40.000 km 222.000.000 219.000.000	There is a significant price difference, with prices of 222,000,000 and 219,000,000 in the same year. This noticeable price variation could be due to factors such as differences in the condition of the cars, as well as a good or regular maintenance history, which can increase a car's value.

Outliers represent significant residuals and may indicate unique market conditions, car conditions, or data recording errors. In this analysis, outliers were identified using the following method:

Outlier Identification:

1. Residual Calculation: Residuals are calculated as the difference between the actual value and the predicted value from the model.
2. Threshold Determination: A data point is considered an outlier if its residual exceeds twice the standard deviation of the average residual.
3. Identification Process: All residuals are tested against this threshold, and data points that meet this criterion are marked as outliers.

Impact of Outliers on Prediction Results:

1. Effect on Model Accuracy: Outliers can affect the overall accuracy of the model because they represent unusual or extreme data, which can lead to less accurate predictions for certain data points.
2. Model Robustness: More robust models, such as the ANN in this study, tend to provide more stable and accurate predictions even when extreme data is present.
3. Model Improvement: By identifying and understanding outliers, corrective actions such as model adjustment or data cleaning can be taken to improve prediction performance.

Both models, Random Forest Regression and Artificial Neural Network (ANN), detected 13 outliers. The presence of these outliers indicates that certain factors in the data cause the model's predictions to deviate from actual values. Although the number of outliers is the same for both models, their distribution and impact on model performance may differ. ANN demonstrates better resilience to outliers, maintaining higher prediction accuracy compared to Random Forest Regression.

5. CONCLUSION

The implementation of Machine Learning using the Artificial Neural Network (ANN) and Random Forest Regression algorithms in a regression model for used car price prediction demonstrates that both methods can produce accurate predictions. Random Forest Regression, with parameters set to n_estimators=100, max_depth=5, min_samples_split=2, min_samples_leaf=1, and random_state=0, along with the Artificial Neural Network (ANN), using a Sequential model with several layers, successfully predicted used car prices. The ANN model consists of several Dense layers with varying units, where the first layer has 256 units, followed by three layers, each with 128 units. Additionally, all layers use the ReLU activation function, except for the output layer, which uses a linear activation function. Based on K-Fold cross-validation with K=10, Random Forest Regression produced competitive results with a Mean Absolute Error (MAE), while the ANN also showed good performance with its optimized network architecture. Based on the study of used car price prediction using the two algorithm methods, ANN and Random Forest Regression, it can be concluded that both methods have different capabilities. ANN achieved the best MAE of 0.035, which is lower than the MAE of Random Forest Regression, which was 0.047. This indicates that ANN performs better in predicting used car prices, showing a higher level of accuracy. Further research is recommended to explore other algorithms, such as Support Vector Regression or Gradient Boosting, which may yield better predictions. Additionally, these methods could be applied in different used car market contexts to understand the dynamics that may influence pricing. However, it is important to note that this study has limitations, including the dataset size, which may not be fully representative of the entire used car market, and potential biases in the data used. By acknowledging these limitations, the results provide a more balanced and realistic perspective on the accuracy of used car price predictions using the analyzed methods.

REFERENCES

[1] E. Surya Negara, J. Jenderal Ahmad Yani, K. I. Seberang Ulu, and S. Selatan, “Sulaiman et al, Komparasi Algoritma K-Nearest Neoghbors dan Random Forest …….. 337 Komparasi Algoritma K-Nearest Neighbors dan Random Forest Pada Prediksi Harga Mobil Bekas,” 2023. [Online]. Available: www.cardekho.com.

[2] S. Laxmaiah, K. Shireesha, and B. Prathima, “Prediction of Used Car Prices Using Artificial Neural Networks and Machine Learning”, doi: 10.32628/IJSRCSEIT.

[3] C. Tarmanini, N. Sarma, C. Gezegin, and O. Ozgonenel, “Short term load forecasting based on ARIMA and ANN approaches,” *Energy Reports*, vol. 9, pp. 550–557, May 2023, doi: 10.1016/j.egyr.2023.01.060.

[4] V. Yadav, A. K. Yadav, V. Singh, and T. Singh, “Artificial neural network an innovative approach in air pollutant prediction for environmental applications: A review,” Jun. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.rineng.2024.102305.

[5] G. Toğa, B. Atalay, and M. D. Toksari, "COVID-19 prevalence forecasting using Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN): Case of Turkey," *J Infect Public Health*, vol. 14, no. 7, pp. 811–816, Jul. 2021, doi: 10.1016/j.jiph.2021.04.015.

[6] S. Lessmann and S. Voß, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," *Int J Forecast*, vol. 33, no. 4, pp. 864–877, Oct. 2017, doi: 10.1016/j.ijforecast.2017.04.003.

[7] B. K. Meher, M. Singh, R. Birau, and A. Anand, "Forecasting Stock Prices of Fintech Companies of India using Random Forest with High-frequency Data," *Journal of Open Innovation: Technology, Market, and Complexity*, p. 100180, Mar. 2023, doi: 10.1016/j.joitmc.2023.100180.

[8] M. Mehdi Karakoç, G. Çelik, and A. Varol, "Car Price Prediction Using An Artificial Neural Network," 2020.

[9] B. Kriswantara, K. Kurniawati, and H. F. Pardede, "Prediksi Harga Mobil Bekas dengan Machine Learning," *Syntax Literate ; Jurnal Ilmiah Indonesia*, vol. 6, no. 5, p. 2100, May 2021, doi: 10.36418/syntax-literate.v6i5.2716.

[10] A. Lusiana and P. Yuliarty, "PENERAPAN METODE PERAMALAN (FORECASTING) PADA PERMINTAAN ATAP di PT X."

[11] L. Purwati Ayuningtias and M. Irfan, "Analisa Perbandingan Logic Fuzzy Metode Tsukamoto, Sugeno, Dan Mamdani (Studi Kasus : Prediksi Jumlah Pendaftar Mahasiswa Baru Fakultas Sains Dan Teknologi Universitas Islam Negeri Sunan Gunung Djati Bandung)," 2017.

[12] E. R. Habibi, "peramalan harga garam konsumsi menggunakan artificial neural network feedforward-backpropagation (studi kasus : pt. Garam mas, rembang, jawa tengah)," 2017.

[13] Gupta G, "Artificial Intelligence and Expert Systems. In Mercury Learning & Information. Mercury Learning & Information.," 2020.

[14] A. Al Miaari and H. M. Ali, "Batteries temperature prediction and thermal management using machine learning: An overview," Nov. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.egyr.2023.08.043.

[15] K. P. P. J. K. P. S. Shikha Jain, *Artificial Intelligence, Machine Learning, and Mental Health in Pandemics A Computational Approach. Academic Press.* 2022.

[16] B. Kriswantara and R. Sadikin, "Used Car Price Prediction with Random Forest Regressor Model," *Journal of Information Systems, Informatics and Computing Issue Period*, vol. 6, no. 1, pp. 40–49, 2022, doi: 10.52362/jisicom.v6i1.752.

[17] S. Dasariraju, M. Huo, and S. McCalla, "Detection and Classification of Immature Leukocytes for Diagnosis of Acute Myeloid Leukemia Using Random Forest Algorithm," *Bioengineering*, vol. 7, no. 4, p. 120, Oct. 2020, doi: 10.3390/bioengineering7040120.

[18] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *J Biomed Inform*, vol. 115, p. 103690, Mar. 2021, doi: 10.1016/j.jbi.2021.103690.

[19] F. R. Aszhari, Z. Rustam, F. Subroto, and A. S. Semendawai, "Classification of thalassemia data using random forest algorithm," *J Phys Conf Ser*, vol. 1490, no. 1, p. 012050, Mar. 2020, doi: 10.1088/1742-6596/1490/1/012050.

[20] N. Abdulkareem and A. Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," 2021, doi: 10.5281/zenodo.4471118.

[21] P. Wibowo and C. Fatichah, "An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 63–71, 2021, doi: 10.26594/register.v7i1.2206.

[22] Y. Liu, G. Yan, and A. Settanni, "Forecasting the transportation energy demand with the help of optimization artificial neural network using an improved red fox optimizer (IRFO)," *Heliyon*, vol. 9, no. 11, Nov. 2023, doi: 10.1016/j.heliyon.2023.e21599.