International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 05, Issue 01, March 2024
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Used Car Price Prediction Model: A Machine Learning Approach

1st Daniel Aprillio Budiono, 2nd Kevin Sander Utomo, 3rd Kenny Jinhiro Wibowo, 4th Marcell Jeremy Wiradinata
School of Information Technology Universitas Ciputra Surabaya
Surabaya, Indonesia
[1]danielaprillio02@gmail.com, [2]kevinsanderutomo@gmail.com, [3]kenny.jinhiro@gmail.com,[4]mjwiradinata@gmail.com

Abstract— The impact of the Covid-19 pandemic over the past two years has slowed down the economy, including the market of used cars. However, the recent decline in the number of cases infected with Covid-19 has reignited interest in the used car market. One of many persisting issues found in the used car market is that sellers want the highest price possible; but, buyers and used car dealers bid the lowest price due to economic stability uncertainty. To accelerate the recovery of the used car industry, various innovations are required. This study proposes the use of the K-Nearest Neighbors (KNN) regression model to predict used car prices to address this issue. The proposed KNN model is a machine learning algorithm which is capable of handling multi-dimensional data and its robustness to noisy data, making it suitable for predicting used car prices based on multiple factors. By analyzing collected data on used car prices, a machine learning-based regression model can be developed to predict used car prices based on factors commonly used in the used car industry, such as year of production, car type, car condition, and others. This study makes use of 504 used car data collected through web scraping as a secondary data collection method. With a relatively small error rate of 8.3% and an R^2 value of 98.8%, the results of this analysis can provide insight for used car buyers and sellers, to better gauge the price of used cars in the market.

Keywords: Used Cars, Machine Learning, Multiple Linear Regression, Polynomial Regression, K-Neighbors Regression.

I. INTRODUCTION

The COVID-19 pandemic has presented several obstacles to the used car market over the past two years. While the pandemic persists, there are signs of gradual recovery. The resurgence of the domestic economy has contributed to the revival of the used car industry, including the Indonesian market [1].

Since fuel prices have increased, the sales of used cars have also increased by around 10 percent [2]. There is often a perception that used cars are not worth buying or selling. In reality, used cars do not mean they cannot be used. There are also used cars that are still of good quality and still suitable for use. Several factors can usually affect the selling price in the purchase of used cars, such as fuel type, total kilometers traveled, registration type, and others [3].

However, a problem arises in the sale of used cars. Owners of used cars may feel that showrooms offer deficient prices, on one side showrooms believe that their offers are fair. However, this creates a trust issue between the owner of the used car and the showroom who are offering the prices [4]. To address this issue, a third party is crucial. This program for predicting used car prices aims to serve as a third-party application to tackle this problem. The dataset used for this case is related to used cars and was obtained by scraping data from the carsome.id website independently.

II. RESEARCH METHODS

In recent years, there has been a growing interest in using machine learning techniques for predicting used car prices. Numerous studies have proposed various models to address the task of price prediction based on multiple parameters. In a study done in [5], the authors examine the increasing trend of Machine Learning and its applications in different fields, one of which is predicting house prices. The suggested model employs certain features such as the number of bedrooms, age of the house, availability of transportation and educational facilities, as well as nearby shopping malls to estimate house values. The research employs several methods including decision tree classification, decision tree regression, and multiple linear regression, and it is executed through the Scikit-Learn Machine Learning Tool in a particular area of Andhra Pradesh, India. The created model can potentially assist buyers in finding a suitable house based on their requirements.

In another study from [6], the authors address the surging cases of COVID-19 and the necessity of advanced technologies like Machine Learning (ML) to monitor and forecast the virus's spread. The proposed approach endeavors to estimate the number of individuals infected with COVID-19 by utilizing four ML models, namely Neural Network (NN), Support Vector Machines (SVM), Bayesian Network (BN), and Polynomial Regression (PR). Five performance parameters were employed to assess the efficiency of each model. The findings indicate that NN exceeded the other models, while SVM exhibited poor performance in all forecasts. The research provides optimism owing to the reduced death rate and proposes an ongoing need for the combination of data and case interpretation for future prospects.

Further study in [7], the author proposed the surge in the popularity of bicycle exercise during the

pandemic, leading to increased sales and bike shortages. However, potential buyers are still unsure about which type of bicycle to purchase. The study gathered data from 242 bicycle users in Java Island, Indonesia, using various predictors to determine the type of bike purchased. The study utilized several classification methods to create a predictive model, which resulted in the Support Vector Machine and Decision Tree models having the highest accuracy of 90%, while Naive Bayes has the lowest accuracy of 73%. The model can assist potential buyers in choosing the appropriate bike.

Much success has been made in utilizing machine learning methods for value or number prediction based on specified parameters. The above studies not only prove the possibility of predicting values accurately and reliably but also that such methods could be implemented for the price prediction of a myriad of interests, such as used cars.

A study conducted on the Croatian used-car market in [17] implemented data mining and analysis techniques to achieve a 95% prediction or testing accuracy by training a linear regression model. To further validate the model's performance, the authors introduced a second dataset taken 3 months later and successfully identified the contrasting relationship between the decrease in average kilometers traveled and the increase in average prices in the market.

The study in [15] presents a system for predicting a fair price for any pre-owned car in the Mumbai region of India by comparing the performance of a Random Forest (RF) algorithm to an eXtreme Gradient Boost (XGBoost) algorithm. The study found that the XGBoost algorithm achieved an RMSE of 0.53 while the RF algorithm achieved an RMSE of 3.44; hence, concluding that the XGBoost algorithm outperformed the RF algorithm in accurately predicting used car prices.

A nuanced study in [16], through exploratory analysis and the testing of various machine learning algorithms for predicting used car prices, found that the XGBoost algorithm achieved an R^2 score of 91% while the RF model came in second. The study highlights that the RF model was capable of outperforming the XGBoost model based on LMAE evaluation methods.

Contrary to the previous studies, a study conducted in Bangladesh in [18], successfully implemented an RF machine learning model to achieve a 99.59% testing accuracy in predicting the future costs of automobiles in Bangladesh by analyzing and identifying patterns in the dataset.

Various machine learning models have been proposed for predicting prices, especially for used cars. Based on all the above studies, it is reasonable to conclude that the best ML model depends on the specific problem and the available dataset at hand, as well as the desired trade-off between accuracy and interpretability when evaluating the model's performance.

III. RESULT AND ANALYSIS

Machine learning is a field of artificial intelligence that deals with the development of algorithms that allow computers to learn from data and make predictions or take actions without being explicitly programmed [19], [20]. There are three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning [21].

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables [22]. Linear regression aims to find the line of best fit that minimizes the sum of squared differences between the predicted and actual values [22]. Hence, linear regression performs better as the linearity between the independent variables increases, hence the name "linear regression". Another type of regression is non-linear regression, a regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an n^{th} degree polynomial [23]. Unlike linear regression, non-linear regression models can capture complex relationships between the variables. Non-linear regression models are more flexible than linear regression models and are used when the relationship between the variables is not linear.

K-Nearest Neighbors (KNN) is a supervised learning algorithm that is commonly used for classification tasks but can also be used for regression tasks [24]. In a regression problem, KNN works by finding the K nearest points in the training data to a given point and then calculating the average of those K points as the predicted value for the given point [24]. There are several reasons why KNN can be a good choice for regression problems, including:

1. Simplicity: KNN is a simple algorithm that requires little prior knowledge of statistical models or feature engineering. It is easy to implement and can be used for a variety of regression problems.
2. Flexibility: KNN can model non-linear relationships between variables, making it suitable for a wide range of regression problems.
3. Handling Missing Data: KNN resiliently adapts to missing data, as it only predicts based on the nearest neighbors and not the entire dataset.

Compared to other models, KNN is an excellent algorithm for resolving regression problems, especially

when compared to more complex models such as deep learning algorithms. KNN is often faster and easier to train than deep learning algorithms, and it provides local interpretability, which is often not the case for deep learning algorithms. Additionally, KNN does not require large amounts of data to perform well, making it suitable for small or medium-sized datasets.

3.1. CRISP-DM (Cross-Industry Standard Process for Data Mining) Cycle

The CRISP-DM (Cross-Industry Standard Process for Data Mining) cycle is a widely used methodology for data mining or analytics projects. It is a structured approach that provides a framework for managing and executing data mining projects from start to finish [8]. The CRISP-DM cycle consists of six phases [9]:

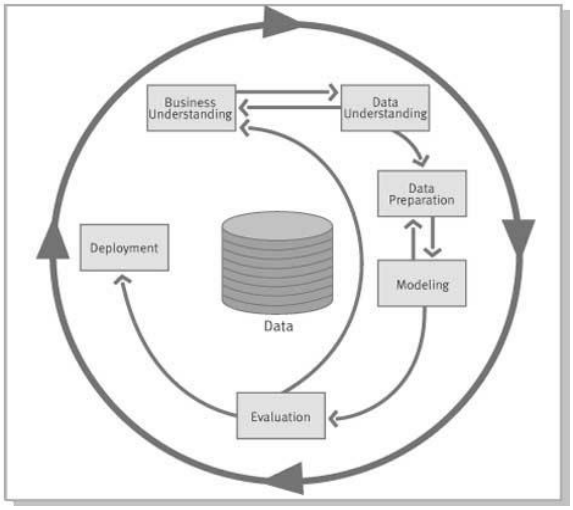

Figure 1: The CRISP-DM Cycle. Adapted from [10].

1. Business Understanding

During this phase, it is necessary to have knowledge about the business object, such as how to obtain or build data, and how to align modeling goals with business objectives to create the best possible model. The activities carried out in this project include clearly defining the main goals and problems in a general sense, translating these goals, and determining limitations when formulating data mining problems. Finally, an initial strategy is prepared to solve the problems identified.

2. Data Understanding

The Data understanding phase focuses on collecting and exploring data to gain insights into its characteristics and quality [11]. In this project, the data exploration was conducted using EDA charts to determine interesting facts about the data. EDA (Exploratory Data Analysis) is a statistical approach to analyzing data that focuses on discovering patterns, relationships, and anomalies in the data.

3. Data Preparation

The focus of the data preparation phase is to clean, transform, integrate, and reduce data to make it suitable for modeling. The main objective is to ensure that the data is relevant, accurate, complete, and consistent with the business problem. In this project, it is essential to prepare the data so that it can be used in the subsequent steps, such as converting strings to numbers. Furthermore, it may be necessary to clean the data by eliminating duplicates, missing values, and outliers. Additionally, integrating the data may be necessary too, which involves merging data from various sources and ensuring consistency. Finally, it may be necessary to reduce the data by selecting the most relevant features, to prevent the model from being overly complex or biased. These tasks enable the data to be prepared for modeling in the next phase of the CRISP-DM cycle.

4. Modeling

The modelling phase involves selecting and applying appropriate modeling techniques to build predictive or descriptive models. This project uses various models of regression, which involves fitting the data to the chosen model and determining the model's accuracy in predicting the dependent variable based on the independent variables. The modeling process may also involve selecting the best model based on its performance and fine-tuning its hyperparameters to improve its accuracy.

5. Evaluation

The evaluation phase focuses on evaluating the performance of the models and assessing their effectiveness in meeting project objectives. The previously created model is assessed and validated. Various evaluation techniques are employed to assess the performance of the model, such as comparing the model's predictions to actual data, assessing the model's accuracy, precision, recall, F1 score, RMSE, MAPE, R2, and other relevant metrics to ensure its effectiveness and applicability [12].

6. Deployment

The deployment phase involves deploying the models into a production environment and integrating them into the business processes to deliver the desired results. This refers to the process of implementing the model in a real-world scenario and making it accessible to end-users or stakeholders. During this phase, several tasks need to be performed, such as creating documentation and user manuals, providing training to stakeholders or users on how to use the model, integrating the model with existing systems or processes, and ensuring that the model is regularly maintained and updated to keep it relevant and effective.

3.2. Data Scraping Process

The data scraping process is done by utilizing a website extension from webscraper.io to scrape data from the carsome.id website. Here are the steps:

1. Use the "Inspect Element" function from a browser. Then, enter the "Web Scraper" tab and create a new sitemap.
2. Enter selectors based on the carsome.id selected html tag.
3. Choose the type of selector based on the chosen html tag:: such as link, text, or image.
4. After all the data is properly selected, run the "Scrape" function and the extension will perform the scraping automatically.
5. Wait for the extension to finish scraping the data. After it's done, the data can be exported in .csv format.

3.3. Data Understanding & Data Preparation

The used car dataset has 13 features and 504 rows of data. This dataset includes the following features: Brand, Fuel Type, Current Color, Seat, Registration Age, Registration Type, Current Mileage, Spare Key, Service Book, Principal Warranty, Road Tax Exp Age, Location, and Price.

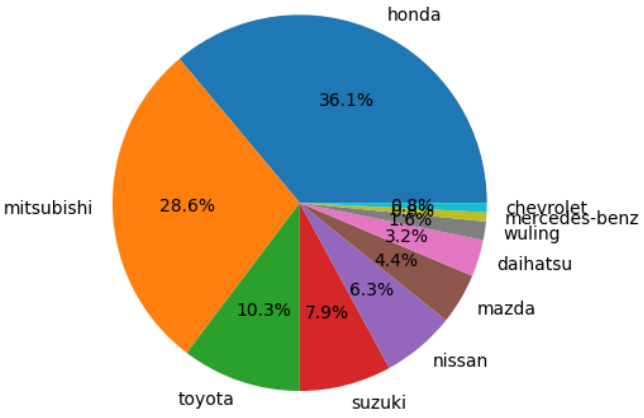

Figure 2: Brand's Feature Pie Chart

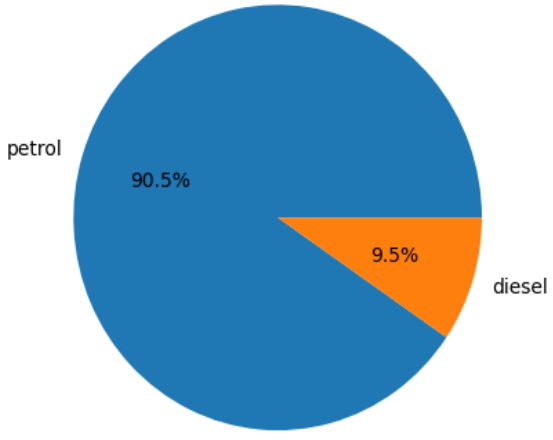

Figure 3: Fuel Type's Feature Pie Chart

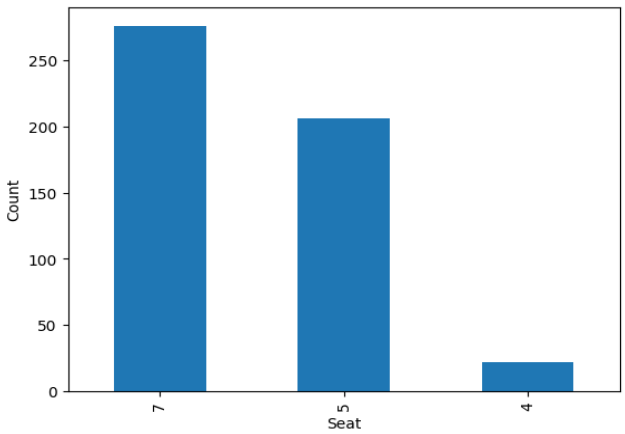

Figure 4: Seat's Feature Bar Plot

According to the problem above, the target variable for the regression analysis will be the "Price" column. Before moving forward, it is suggested to lowercase all objects/categorical to minimize any variations in spelling.

All date-related variables in the dataset are converted into ages value while data splitting and formatting is implemented on some features so data processing could be done later. In detail, features which contain date-related information will be transformed to age. Continuing the process, all numerical features were made sure to have an 'Integer' data type. Then, all remaining categorical features were then converted into numerical features. Lastly, dataset information is checked to see if there are any empty rows/columns or if the data types are already in integer format.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Brand               504 non-null    int64
 1   Fuel Type           504 non-null    int64
 2   Current Color       504 non-null    int64
 3   Seat                504 non-null    int32
 4   Registration Age    504 non-null    int32
 5   Registration Type   504 non-null    int64
 6   Current Mileage     504 non-null    int32
 7   Spare Key           504 non-null    int64
 8   Service Book        504 non-null    int64
 9   Principal Warranty  504 non-null    int32
 10  Road Tax Exp Age    504 non-null    int32
 11  Location            504 non-null    int64
 12  Price               504 non-null    int32
dtypes: int32(6), int64(7)
memory usage: 39.5 KB
```

Figure 5: Checking Dataset Information

Exploratory Univariate

With regards to exploring univariate variables, data count, mean, standard deviation, minimum and maximum values, and quartiles of the "price" column,

is displayed with the python's 'describe' function. It can be seen from the output that the range of car prices ranges from 78 million rupiah to 1.163 billion rupiah.

	Price
count	5.040000e+02
mean	2.378742e+08
std	1.455192e+08
min	7.800000e+07
25%	1.460000e+08
50%	1.995000e+08
75%	2.615000e+08
max	1.163000e+09

Figure 6: Displaying "Price" Data Description

Next, outliers from all numeric features of "Seat" and "Registration Age" were able to be found, by displaying boxplots. However, outlier removal was not necessary as outliers are still valid in this case.

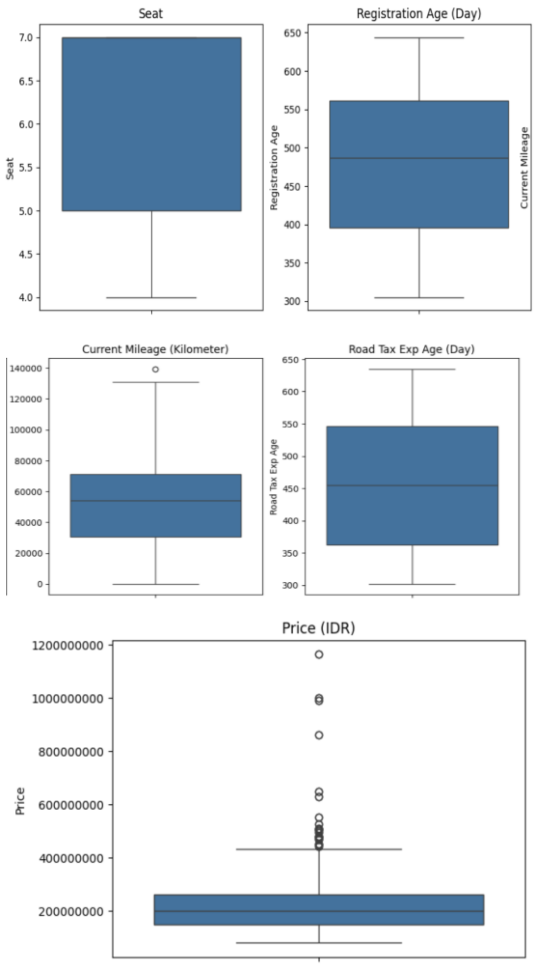

Figure 7: Searching Outliers

Exploratory Bivariate

Displaying the correlation between variables using a heatmap.

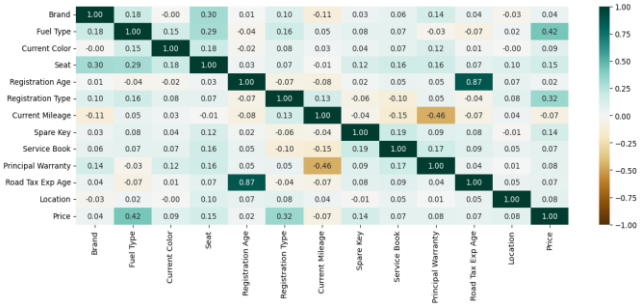

Figure 8: The heatmap shows the correlation between different variables

According to the heatmap above, several features are strongly correlated with the target variable "Price". The "Fuel Type" and "Registration Age" features have the highest correlation among other variables and thus have a greater influence in determining the used car prices.

Binning and Resampling Data

Based on the result of the "Price" feature boxplot, the distribution of the obtained data is not even. Therefore, the binning process is needed to divide the "Price" feature into several parts. This feature is divided into 3 parts: "lite", "standard", and "premium". Afterward, the data needs to be rechecked for its data distribution. To resolve an uneven distribution of data, a resampling process is necessary to balance the distribution of data without reducing the existing information.

3.4. Data Modelling

3.4.1. Train Test Split

To start the data modelling process, separation of the dataset into independent variables, identified as X and dependent variables, identified as Y, were done using 'loc'. The y variable is filled with the 'Price' column in accordance with the target variable at the beginning, while the X variable is filled with the other columns except 'Price'.'MinMaxScaler' is used to normalize the data which reduces the minimum data range to 0 and the maximum data range to 1.

After going through all the processes and normalization above, data splitting is done in which 70% of the data is used for training and the rest 30% is used for testing. The random_state value can be set to any number, but in this case, it will be set to 0 to ensure its consistent results when run repeatedly.

3.4.2. Regression Model Analysis

After preparing the data, regression analysis was performed using several models such as Linear Regression (LinearRegression), Polynomial Regression (Polynomial Features), and K-Nearest Neighbors Regression (KNeighborsRegressor).

3.4.3. Multiple Linear Regression

Almost all problems in the current reality have multiple variables. Linear regression which involves more than one variable is called Multiple Linear Regression [5]. Its operation is almost the same as Simple Linear Regression, the difference lies in the evaluation of the model. Implementation-wise, a multiple linear regression model was created using LinearRegression. Variables are then created to store the prediction results based on the test data.

3.4.4. Polynomial Linear Regression

The Polynomial Regression model is commonly used to analyze the relationship between the independent variable (X) and the dependent variable (y). The evaluation model produced by polynomial regression always uses a polynomial degree based on the independent variable [6].

As for implementation, a polynomial regression model using PolynomialFeatures was created. X variables were then re-trained in the polynomial regression model. Lastly, a variable is initialized to store the prediction results based on the test data.

3.4.5. K-Nearest Neighbors Regression

K-Nearest Neighbors (KNN) is a method with a lazy learning or instant-based learning algorithm approach, resulting in less time required for training [13]. KNN can work by using the similarity between new data and other existing data (k) in the nearest positions [7]. There are several formulas for finding the distance between neighbors (k), such as euclidean, manhattan, etc.

GridSearchCV was used for determining the best number of k-neighbors, or in other words, parameter tuning. Then, the model KNeighborsRegressor was used to train the dataset to make the most suitable K-Nearest-Neighbors Regression model for the dataset. Lastly, a variable was initialized to hold the prediction results based on the test data.

3.5. Model Evaluation

Regarding acknowledging the performance of the models, Goodness-of-Fit of all regression models was displayed. The Goodness-of-fit of all the models can be obtained by comparing the test data results and prediction results.

	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)	Coefficient of Determination (R2)
LinearRegression	75972799.352649	0.177493	0.951439
PolynomialRegression	47189102.227389	0.095499	0.981265
KNeighborsRegressor	36561034.635417	0.083127	0.988754

Figure 9: Model Evaluation

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a commonly used metric to evaluate the performance of a regression model [14]. It measures the square root of the average of the squared differences between the predicted values and the actual values. In other words, it measures the deviation of the predicted values from the actual values and provides a single number that represents the average error of the model [14]. The smaller the value, the better, as it concludes that the model has a small margin error [14]. In this case, the smallest error for the price is 36 million.

Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) measures the average absolute percentage difference between the actual and predicted values of a variable [14]. MAPE is useful in measuring the accuracy of a forecasting model, as it provides a sense of the magnitude of the errors in the predictions relative to the actual values [14]. MAPE ranges from 0 to 1 (valued as a percentage) [14]. The closer to zero the value, the more accurate the prediction results [14]. MAPE indicates the average percentage rate of absolute errors [14]. In this case, the smallest error for the prediction results is 8.3%.

Coefficient of Determination (R2)

Coefficient of Determination (R2) is used to determine the proportion of the variance in the dependent variable that can be explained by the independent variable(s) [14]. R2 ranges from 0 to 1 (valued as a percentage) [14]. The closer to 1 the R2 value is, the more the prediction results match the actual results [14]. In this case, the highest matching percentage is 98.8%.

Based on the above evaluation, The k-neighbors regression model produces the lowest rate of error in predicting the prices of used cars. KNN regression model can determine where new data should be placed. This dataset of used cars only consists of 504 data points, which is considered a small amount, making the k-neighbors regression model well-suited for use.

In addition, it was observed that the k-neighbors regression model's performance is significantly affected by the number of neighbors used, as a higher number of neighbors led to lower accuracy and a lower number of neighbors produced better results. This

emphasizes the importance of tuning the model's hyperparameters to achieve optimal performance.

Overall, the k-neighbors regression model proved to be effective in predicting the prices of used cars. However, further research and experimentation can be conducted to improve the model's accuracy and robustness, such as exploring other regression models and incorporating more relevant features.

VI. CONCLUSION

In conclusion, this study has successfully demonstrated the effectiveness of the K-Nearest Neighbors (KNN) regression model in predicting used car prices. The model's performance was evaluated using three key metrics: Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2). The model achieved a relatively small RMSE of 36 million, indicating a small margin of error. The MAPE was 8.3%, suggesting a high level of accuracy in the prediction results. Furthermore, the R^2 value was 98.8%, indicating a strong correlation between the predicted and actual results.

This model can be a valuable tool for both buyers and sellers to gain insights into the possible price range for buying or selling used cars while also helping showrooms offer more accurate prices. The study ultimately demonstrates the significance of hyperparameter tuning and the inclusion of relevant features to enhance the model's accuracy. While the dataset in this study was limited, the k-neighbors regression model can be applied to larger and more diverse datasets to further enhance its performance. Ultimately, the K-neighbors regression model holds great potential for aiding individuals and businesses in the automotive industry in making informed decisions regarding the purchase and sale of used cars.

	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)	Coefficient of Determination (R2)
LinearRegression	75972799.352649	0.177493	0.951439
PolynomialRegression	47189102.227389	0.095499	0.981265
KNeighborsRegressor	36561034.635417	0.083127	0.988754

Figure 10: Model Evaluation

THANK-YOU NOTE

Thank you to the IJCIS Team for taking the time to create this template.

REFERENCES

[1] I. Infootomotif, "Manfaat Mobil untuk Aktivitas Manusia," *Kumparan*, Nov. 05, 2021. Accessed: Feb. 25, 2023. [Online]. Available: https://kumparan.com/info-otomotif/manfaat-mobil-untuk-aktivitas-manusia-1wqysMbGRz0

[2] A. Ferdian, "Harga BBM Naik, Penjualan Mobil Bekas Meningkat 10 Persen," *KOMPAS.com*, Oct. 04, 2022. Accessed: Feb. 25, 2023. [Online]. Available: https://otomotif.kompas.com/read/2022/10/04/144100015/harga-bbm-naik-penjualan-mobil-bekas-meningkat-10-persen

[3] A. Admin, "Pilih Membeli Mobil Bekas atau Mobil Baru?" Accessed: Feb. 25, 2023. [Online]. Available: https://www.otonesia.co.id/blog/post/pilih-membeli-mobil-bekas-atau-mobil-baru

[4] S. Y. Ahmed, B. J. Ali, and C. Top, "Understanding the Impact of Trust, Perceived Risk, and Perceived Technology on the Online Shopping Intentions: Case Study in Kurdistan Region of Iraq," *Journal of Contemporary Issues in Business and Government*, vol. 27, no. 3, Apr. 2021, doi: 10.47750/cibg.2021.27.03.264.

[5] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," *International Journal of Information Engineering and Electronic Business*, vol. 12, no. 2, pp. 15–20, Apr. 2020, doi: 10.5815/ijieeb.2020.02.03.

[6] A. H. M. Hassan, A. A. M. Qasem, W. F. M. Abdalla, and O. H. Elhassan, "Visualization & Prediction of COVID-19 Future Outbreak by Using Machine Learning," *International Journal of Information Technology and Computer Science*, vol. 13, no. 3, pp. 16–32, Jun. 2021, doi: 10.5815/ijitcs.2021.03.02.

[7] T. Wiradinata, "Folding Bicycle Prospective Buyer Prediction Model," *International Journal of Information Engineering and Electronic Business*, vol. 13, no. 5, pp. 1–8, Oct. 2021, doi: 10.5815/ijieeb.2021.05.01.

[8] P. Chapman *et al.*, "CRISP-DM 1.0: Step-by-step data mining guide," 2000.

[9] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *Jurnal Teknologi Informatika dan Komputer*, vol.

6, no. 2, pp. 12–20, Sep. 2020, doi: 10.37012/jtik.v6i2.299.

[10] C. Soares, Y. Peng, J. Meng, T. Washio, and Z. H. Zhou, "Applications of data mining in E-business finance: Introduction," *Frontiers in Artificial Intelligence and Applications*, vol. 177, no. 1, pp. 1–9, 2008, doi: 10.3233/978-1-58603-890-8-1.

[11] M. Guha Majumder, S. Dutta Gupta, and J. Paul, "Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis," *J Bus Res*, vol. 150, pp. 147–164, Nov. 2022, doi: 10.1016/j.jbusres.2022.06.012.

[12] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[13] A. Singh, M. N., and R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, 2017, doi: 10.14569/IJACSA.2017.081201.

[14] V. Plevris, G. Solorzano, N. Bakas, and M. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *8th European Congress on Computational Methods in Applied Sciences and Engineering*, CIMNE, 2022. doi: 10.23967/eccomas.2022.155.

[15] Longani, C., Potharaju, S. P., & Deore, S. (2021). Price prediction for pre-owned cars using ensemble machine learning techniques. Advances in Parallel Computing, 39, 178–187. https://doi.org/10.3233/APC210194

[16] F. R. Amik, A. Lanard, A. Ismat, and S. Momen, "Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh," *Information (Switzerland)*, vol. 12, no. 12, Dec. 2021, doi: 10.3390/info12120514.

[17] L. Bukvić, J. Pašagić Škrinjar, T. Fratrović, and B. Abramović, "Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning," *Sustainability (Switzerland)*, vol. 14, no. 24, Dec. 2022, doi: 10.3390/su142417034.

[18] F. Abdullah, Md. A. Rahman, M. Shidujaman, M. Hasan, and Md. T. Habib, "Machine learning modeling for reconditioned car selling price prediction," SPIE-Intl Soc Optical Eng, Sep. 2023, p. 100. doi: 10.1117/12.2689745.

[19] Amit Kumar Tyagi and P. Chahal, "Artificial Intelligence and Machine Learning Algorithms," IGI Global eBooks, pp. 421–446, May 2022, doi: https://doi.org/10.4018/978-1-6684-6291-1.ch024.

[20] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," Journal of Physics: Conference Series, vol. 1142, no. 012012, Nov. 2018, doi: https://doi.org/10.1088/1742-6596/1142/1/012012.

[21] E. F. Morales and H. J. Escalante, "A brief introduction to supervised, unsupervised, and reinforcement learning," Biosignal Processing and Classification Using Computational Learning and Intelligence, pp. 111–129, 2022. doi:10.1016/b978-0-12-820125-1.00017-8.

[22] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis. Estats Units d'Amèrica: Wiley, 2021.

[23] E. Ostertagová, "Modelling using polynomial regression," *Procedia Engineering*, vol. 48, pp. 500–506, 2012. doi:10.1016/j.proeng.2012.09.545

[24] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, pp. 218–218, Jun. 2016, doi: https://doi.org/10.21037/atm.2016.03.37.