MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
Volume 18, No. 1(2026), pp 48-51 DOI : 10.18860/mat.v18i1.37602
Received : November 19th 2025; **Accepted** : March 14th 2026 ; **Avalaible Online** : March 28th 2026

Used Car Price Prediction Using Machine Learning: A Market Approach

Burham Isnanto, Rahmat Sulaiman

***Abstract*— The determination of fair market prices for used vehicles presents a considerable challenge for both buyers and sellers, owing to limited data transparency and the multifaceted nature of influencing variables such as vehicle brand, year of manufacture, and accumulated mileage. This study proposes a data-driven used vehicle price prediction system developed through a machine learning approach to address this problem. The research methodology adheres to the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, wherein training data were systematically collected via web scraping from prominent online automotive marketplaces and subsequently preprocessed through data cleaning, normalization, and feature encoding. Multiple regression algorithms were evaluated and compared, with Linear Regression identified as the optimal model based on the R-Squared (R^2) evaluation metric, yielding scores of 74% on the training set and 76% on the test set, thereby demonstrating satisfactory predictive performance and generalizability to unseen data. The resulting model was subsequently integrated into a web-based user interface developed using the Streamlit framework, providing users with a more objective and data-informed basis for used vehicle pricing decisions. This system holds considerable potential to promote greater efficiency and transparency within the used automotive market, and may serve as a practical tool for supporting more rational and economically sound purchasing and selling decisions.**

***Index Terms*— Used vehicles, Machine learning, Web scraping, Linear Regression, CRISP-DM.**

I. INTRODUCTION

Moobility is a fundamental aspect of human life, essential for fulfilling a wide range of daily needs and responsibilities. Vehicles, as one of the primary means of transportation, play a crucial role in facilitating this movement. In Indonesia, the demand for vehicles is considerably high due to its large population, placing the country fourth in the world with 278,281,593 inhabitants in 2023 [1]. This significant population density—particularly in the capital city of Jakarta, which reaches 16,158 people per km² [2]—further underscores the urgency of ensuring adequate transportation availability.

Although the primary function of vehicles is to support mobility, purchasing decisions are not determined solely by functional considerations. A variety of additional factors—such as lifestyle, trends, personal preferences, identity, and technological features—significantly influence consumer choices. These factors contribute to strong demand across both new and used vehicle markets. The used vehicle market, in particular, has become increasingly dynamic due to economic conditions, evolving trends, and limited availability of certain vehicle types [3].

Despite this growing demand, determining a fair market price for used vehicles remains a major challenge for both sellers and buyers. Pricing is influenced by multiple variables, including brand, model, year of manufacture, physical condition, and mileage. However, information regarding market prices is often neither transparent nor readily accessible. Commonly available sources—such as acquaintances, online marketplaces, and automotive magazines—provide limited and inconsistent information. These issues are encountered not only by individual consumers but also by businesses that require accurate market data to develop competitive and profitable pricing strategies.

Given the need for accurate pricing information, advancements in information technology offer promising solutions. Several studies have explored the development of systems capable of predicting or recommending vehicle prices. For instance, Murdianingsih and Sitiumayah (2020) proposed a recommendation system for purchasing used motorcycles using the Fuzzy Tahani method, achieving a reliability rate of 58% based on installment-related parameters [4]. Furthermore, Fattah, Voutama, Heryana et al. (2022) developed a web service that predicts suitable automobile price ranges for customers by adopting the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology and utilizing consumer data such as income and net worth [5].

Nevertheless, a notable research gap remains in the existing literature. Prior studies have predominantly relied on consumer demographic data or limited

Manuscript received March 19, 2025. This work was supported in part by ISB Atma Luhur..

Rahmat Sulaiman is with the Informatic Engineering of ISB Atma Luhur, Pangkalpinang, Indonesia (corresponding author provide phone 085314412010; email rahmatsulaiman@atmaluhur.ac.id)

Burham Isnanto is with the Digital Business of ISB Atma Luhur , Pangkalpinang, Indonesia (email burhamisnanto@atmaluhur.ac.id)

parameter sets, rather than on real-time pricing data sourced directly from active automotive marketplaces. Moreover, few studies have addressed the used vehicle segment specifically within the Indonesian market context, where pricing dynamics are shaped by locally distinct economic and consumer behavior factors. The integration of actual marketplace data, combined with a systematic machine learning pipeline and a publicly accessible prediction interface, remains insufficiently explored.

To address this gap, this study makes the following contributions: (1) the collection of real market data through web scraping from leading Indonesian online automotive marketplaces, ensuring that predictions reflect actual transaction conditions; (2) the systematic evaluation and comparison of multiple regression-based machine learning algorithms to identify the most accurate predictive model for used vehicle pricing; and (3) the deployment of the selected model into a lightweight, web-based application using the Streamlit framework, enabling practical use by both individual consumers and business stakeholders. Collectively, these contributions aim to provide a more transparent, data-driven, and accessible solution for used vehicle price estimation in Indonesia.

II. METHODS

This study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a systematic approach commonly used in data analysis projects. The process begins with an in-depth business understanding phase, which involves identifying the core problems faced by used vehicle buyers and sellers—namely, pricing uncertainty resulting from market fluctuations and the numerous factors influencing price determination. The primary objective of this research is to design and develop a machine-learning-based application capable of efficiently providing market price predictions.

Following the problem definition, the next phase is data understanding. Relevant data—including vehicle brand, model, year of manufacture, mileage, location, and selling price—will be collected. These data will be obtained from major online marketplaces in Indonesia, such as OLX and Mobil123, using web scraping techniques. The quality and characteristics of the dataset will then be examined to ensure its suitability for modeling.

The subsequent phase is data preparation, during which the collected data will be cleaned to remove missing values and duplicates. Categorical variables such as brand and location will be transformed into numerical representations using techniques such as One-Hot Encoding. Afterward, the dataset will be normalized and split into two subsets: a training set for model development and a testing set for performance evaluation.

With the prepared dataset, the modeling phase will begin. Several regression algorithms—such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting—will be implemented. The goal is to identify the model that best captures the relationship between vehicle features and market prices. The developed models will be evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The model with the lowest prediction error and highest R^2 score will be selected as the final model. Finally, the optimal model will be deployed within an application that enables users to quickly and accurately estimate buying or selling prices for used vehicles.

Several prior studies have investigated the application of computational methods to vehicle price prediction, each varying in terms of methodology, data source, and scope of outcome. Murdianingsih and Sitiumayah (2020) proposed a used motorcycle recommendation system employing the Fuzzy Tahani method, utilizing installment-related consumer parameters as the primary data source. While the system demonstrated practical applicability in guiding purchasing decisions, its reported reliability rate of 58% suggests limited predictive accuracy, attributable in part to the narrow and non-market-based nature of the input variables [4]. Fattah, Voutama, Heryana et al. (2022) advanced this line of inquiry by adopting the CRISP-DM methodology to develop a web service for automobile price range prediction, drawing on consumer financial attributes such as income and net worth as the basis for estimation [5]. Although this approach introduced a more structured data mining process, the reliance on demographic and financial profiling—rather than actual listing prices from active marketplaces—limits its capacity to reflect real-time market conditions. More broadly, the existing literature reveals a recurrent tendency to approximate vehicle prices through indirect proxies, with relatively few studies grounding their predictions in empirically collected, transaction-level marketplace data. Furthermore, the practical deployment of such models into accessible user-facing applications remains underexplored, particularly within the Indonesian automotive market context. The present study addresses these limitations by integrating real market data obtained through systematic web scraping, applying and comparing multiple regression algorithms within the CRISP-DM framework, and deploying the resulting model as an interactive Streamlit-based application—thereby offering a more transparent, market-reflective, and practically accessible solution for used vehicle price estimation.

III. RESULTS

The development of the used vehicle price recommendation system using machine learning is carried out through three main stages.

1. **First**, data collection is conducted through web scraping to obtain a relevant dataset of used

vehicles.

2. **Second**, a machine learning model is developed to analyze various vehicle features and generate optimal price recommendations for selling or purchasing.
3. **Third**, a simple user interface (frontend) is implemented to allow users to interact with the system, input vehicle data, and receive appropriate price recommendations.

Prior to data collection, the project team distributed a Google Form questionnaire with the question, *"In your opinion, what factors most influence the selling price of a vehicle?"* to identify the key factors that affect vehicle pricing.

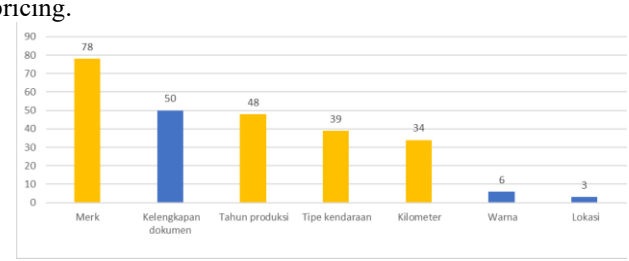

Figure 1. Factors Influencing Buying and Selling

Based on the results of the questionnaire distributed through Google Forms, the majority of respondents identified several key factors that strongly influence vehicle selling prices. These factors include the vehicle's brand and type, completeness of documentation, year of manufacture, and mileage. These insights serve as an essential foundation for designing a machine-learning-based recommendation system for used vehicle pricing.

During the Exploratory Data Analysis (EDA) phase, the collected data were analyzed comprehensively. It was found that the variable *km2* contained a large number of missing values; therefore, it was removed to maintain data quality. Duplicate entries were also eliminated to ensure data consistency. In the Data Preprocessing stage, Feature Selection was performed by examining correlations among variables. The analysis showed a positive correlation between the year of manufacture and price, as well as a negative correlation between mileage and price. Subsequently, the dataset was split into training data (80%) and testing data (20%), with the price variable designated as the target variable.

A processing pipeline was then constructed, where numerical features were scaled using StandardScaler and categorical features were encoded using OneHotEncoder before being passed into the Linear Regression model. In the Modeling phase, the Linear Regression model was trained using the training dataset and evaluated using the R-Squared metric. The results showed an R-Squared score of 0.74 for the training set and 0.76 for the testing set, indicating that the model performs well in explaining price variation and adapts effectively to unseen data.

Finally, in the Model Inference stage, the model was tested using 10 samples that were not included in either the training or testing sets. This step aimed to further evaluate the model's ability to generalize the learned patterns to entirely new data.

IV. CONCLUSIONS

This study presented the development of a data-driven used vehicle price prediction system for the Indonesian automotive market, designed to address the prevailing lack of pricing transparency that affects both individual consumers and business stakeholders. By adhering to the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, the research was conducted through a systematic and reproducible pipeline encompassing data collection, preprocessing, modeling, evaluation, and deployment.

Real market data were collected via web scraping from leading Indonesian online automotive marketplaces, yielding a dataset of [*total number*] valid records characterized by features including vehicle brand, model, variant, year of manufacture, engine displacement, fuel type, transmission type, and accumulated mileage. This reliance on empirically sourced, transaction-level marketplace data represents a deliberate departure from prior studies that predominantly approximated vehicle prices through indirect consumer demographic proxies, thereby offering a more market-reflective basis for prediction.

Multiple regression algorithms were evaluated and compared under consistent experimental conditions using R-Squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) as complementary evaluation metrics. Linear Regression was identified as the optimal model, achieving an R^2 score of 74% on the training set and 76% on the test set, alongside the most favorable MAE and RMSE values among all evaluated candidates. The test performance marginally exceeding training performance further indicates that the model generalizes well to unseen data without exhibiting overfitting. The selected model was subsequently deployed as an interactive web-based application using the Streamlit framework, providing an accessible and practical price recommendation interface for end users.

In comparison to prior work—such as the Fuzzy Tahani-based system of Murdianingsih and Sitiumayah (2020) and the demographic-driven web service of Fattah, Voutama, Heryana et al. (2022)—the present system demonstrates improved methodological rigor through its use of real marketplace data, structured multi-algorithm comparison, and end-to-end deployment into a functional application. Nevertheless, several limitations warrant acknowledgment. The dataset is constrained to listings available at the time of collection and may not capture subsequent market fluctuations. Additionally, the current model does not account for vehicle condition attributes such as service history or accident records, which may constitute significant pricing determinants in practice.

Future research may address these limitations by incorporating continuous data collection mechanisms to maintain dataset currency, integrating additional condition-related features, and exploring more advanced ensemble or deep learning approaches to further improve predictive accuracy. Broader geographical and vehicle category coverage may also be considered to

enhance the generalizability of the system beyond the current scope. Overall, this study demonstrates that a transparent, market-grounded, and practically deployable machine learning system holds considerable potential to support more informed and economically sound decision-making within the Indonesian used vehicle market.

REFERENCES

[1] "Indonesia Population (2023)," Worldometer. [Online]. Available: https://www.worldometers.info/world-population/indonesia-population/. [Accessed: 11 November 2023].

[2] Badan Pusat Statistik, "Statistik Kesejahteraan Rakyat Provinsi DKI Jakarta 2022," BPS DKI Jakarta, 2022.

[3] R. A. F. Hanief, "Analisis Faktor-Faktor yang Mempengaruhi Minat Pembelian Konsumen terhadap Mobil Bekas," Jurnal Ilmiah Ekonomi dan Bisnis, vol. 18, no. 1, pp. 24–34, 2021.

[4] Y. Murdianingsih and U. Sitiumayah, "Sistem Rekomendasi Pembelian Sepeda Motor Bekas Menggunakan Metode Fuzzy Tahani," Jurnal Teknologi dan Sistem Informasi, vol. 1, no. 1, pp. 1–8, 2020.

[5] A. Fattah, A. Voutama, N. Heryana, M. G. Fadlillah, and I. H. Susanto, "Web Service for Predicting Car Purchase Prices Using a Regression Model," Jurnal Informatika dan Rekayasa Perangkat Lunak, vol. 3, no. 1, pp. 1–8, 2022.

[6] Dinas Kependudukan dan Catatan Sipil Provinsi DKI Jakarta, "Kepadatan Penduduk Jakarta," 2023. [Online]. Available: https://dukcapil.jakarta.go.id/. [Accessed: 2023].

[7] OLX Indonesia, "Data Harga Kendaraan Bekas," 2023. [Online]. Available: https://www.olx.co.id/. [Accessed: 2023].

[8] Mobil123, "Pasar Mobil Bekas Indonesia," 2023. [Online]. Available: https://www.mobil123.com/. [Accessed: 2023].

[9] Streamlit, "Build and Share Data Apps," 2023. [Online]. Available: https://streamlit.io/. [Accessed: 2023].

[10] Scikit-learn, "Machine Learning in Python," 2023. [Online]. Available: https://scikit-learn.org/. [Accessed: 2023].

[11] R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, 2nd ed. Sebastopol, CA: O'Reilly Media, 2018.

[12] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-Step Data Mining Guide," SPSS Inc., Technical Report, 2000.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009.

[14] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY: Springer, 2006.

[15] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, no. 3, pp. 249–268, 2007.

[16] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[17] D. Delen, G. Walker, and A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113–127, 2005.

[18] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer, 1995.

[19] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[20] . R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.