Comparative Analysis of Machine Learning Algorithms in Predicting Smartphone Prices

Rahmi Azizi[1*], Muhammad Halmi Dar[2], Mila Nirmala Sari Hasibuan[3]

[1,2,3] Faculty of Science and Technology, Universitas Labuhanbatu**,** Sumatera Utara Indonesia.
[*]Corresponding Author:
Email: rahmiazizi2018@gmail.com

Abstract.

Rapid technological developments have made the smartphone market very competitive and dynamic. Consumers now have a variety of choices with various specifications and prices. Smartphone price prediction is important for helping consumers make purchasing decisions and for manufacturers to determine the right pricing strategy. Machine learning algorithms offer a potential solution to predict prices by utilizing specification data and other features. This study proposes the use of two machine learning algorithms, namely K-Nearest Neighbors and Random Forest, to predict smartphone prices. This study aims to analyze and compare the performance of the two algorithms in predicting smartphone prices, as well as provide recommendations on which algorithm is more effective based on the accuracy and error generated. This study employs a methodology that includes several main steps: data collection, data pre-processing, application of the proposed model, and model testing and evaluation. The results show that the Random Forest algorithm is significantly superior to K-Nearest Neighbors. Random Forest achieved an accuracy of 96.38% with a train error of 0.001003 and a test error of 0.003206, while K-Nearest Neighbors only achieved an accuracy of 59.17% with a train error of 0.009817 and a test error of 0.044094. These results indicate that Random Forest is able to handle data complexity well and provide more accurate and reliable predictions. Random Forest is a more effective algorithm than KNN for smartphone price prediction. Random Forest has a strong generalization ability and does not show any significant signs of overfitting. The results of this study can be a reference for researchers and practitioners in choosing the right machine learning algorithm for price prediction or similar problems. In addition, this study also provides insight into the importance of data preprocessing and hyperparameter tuning to obtain optimal results.

Keywords: *K-Nearest Neighbors, Machine Learning, Prediction, Random Forest, Smartphone.*

1. INTRODUCTION

Smartphones are a necessity in modern society to complement daily activities. Many smartphone brands have emerged because electronics companies are aware of this need. Public demand for high-quality goods and services has increased as a result of improving people's living standards and advances in information technology. Most modern people believe that having a smartphone is a necessity. The emergence of numerous smartphones, which cater to people's communication technology needs, supports this phenomenon. As the times change, smartphone manufacturers are competing to release the latest series and develop technology that appeals to consumers

with higher purchasing power. This has an impact on the number of used smartphones that are still suitable for use because upper-class consumers are looking for smartphones that have features and technology that are qualified at this time. Distributors and retail companies operate a business model where they purchase used smartphones from manufacturers and subsequently sell them to consumers. The company also accepts resale of used smartphones from consumers. The company faces the challenge of determining the appropriate price for used smartphones [1]. Price is the first benchmark in purchasing and selling. Accordingly, the company must estimate the selling and buying price of used smartphones based on their specifications. The company hopes to make a significant profit [2].

Machine learning is the process of finding intriguing patterns and information about selected data using certain methods [3]. Machine learning assists in various tasks such as classification, clustering, association, regression, forecasting, sequence analysis, and deviation analysis [4]. Companies can use machine learning to uncover crucial information from their own data warehouses [5]. Machine learning uncovers previously unknown important information through modeling techniques, a process known as prediction. Research related to smartphone price prediction with machine learning using the Decision Tree, Deep Neural Network (DNN), and Random Forest models yielded results where the accuracy of the Decision Tree was 72%, the DNN got an R2 result of 0.88, and the Random Forest got an accuracy of 84% [6]. Based on the above results, we can conclude that Random Forest performs faster during the training process and achieves the highest accuracy. The study conducted a comparison between the fuzzy mamdani method and the artificial neural network method for determining smartphone prices. The results showed that the artificial neural network method is more accurate in predicting smartphone prices, with a level of truth of 97.910098% for the fuzzy mamdani method and 97.93914% for the artificial neural network method [7].

The goal of this study is to create a machine model with a small error value that can predict smartphone prices according to existing specifications. We expect to minimize errors in the prediction process by utilizing machine learning technology. Predictions do not have to provide a definite answer about events that will occur, but rather seek answers that are as close as possible to what will happen [5]. The use of predictions in business is an important tool and factor in consumer decisions [8]. This study compares 2 algorithms, namely K-Nearest Neighbors and Random Forest, as a comparison of the results of the comparison and the accuracy and error values obtained. This algorithm works by looking for patterns in each data set and then predicting smartphone prices [9], [10]. The two studies above clearly demonstrate differences in the models' prediction results, particularly in the methods employed. Therefore, we conducted this study to compare the performance of the prediction models, aiming to identify the most suitable model based on its accuracy level, error value, and generated prediction outcomes. We calculate the model's error value using the MSE evaluation metric, as it tends to have minimal bias and serves as an

underestimate estimator [11]. This study will compare K-Nearest Neighbors and Random Forest models.

II. METHODS

This study uses an experimental method. The purpose of the experimental research method is to determine the impact of specific treatments [12]. The method divides into several stages, which include data collection, data preprocessing, proposed models, testing, and evaluation. Figure 1 displays the flow diagram.

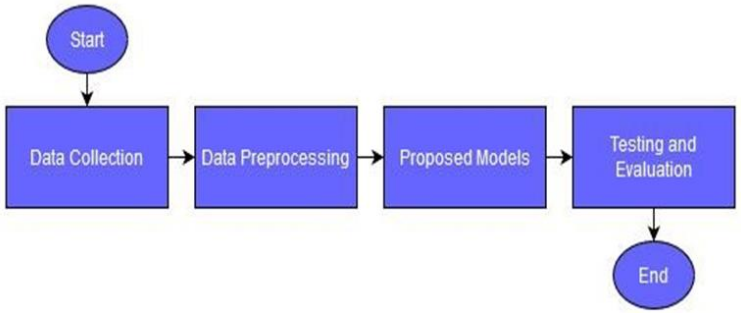

Fig. 1. Research stages

The data used in this study comes from Kaggle.com. This dataset contains smartphone feature data. You can use this dataset for various purposes, including price predictions. The data preprocessing stage transforms the data into a form appropriate for the modeling process. The process includes encoding data using the one-hot-encoding technique and dividing the dataset using the train_test_split function with a proportion of 0.2. Google Colab tools and the Python programming language assist in carrying out the modeling process for the proposed models. We will compare K-Nearest Neighbors and Random Forest models. We will calculate the error and accuracy values for each model to serve as a reference for the comparison process.

The K-nearest neighbors (KNN) algorithm is a classification method that groups new data based on the distance of the nearest data or new data with neighbors [13]. KNN receives input in the form of a vector [14]. KNN is relatively simple compared to other algorithms. This algorithm uses feature similarities to predict values for all new data. KNN, a component of supervised machine learning, trains a labeled data set [14]. The distance measurement metrics used in KNN include Euclidean distance, Hamming, Minkowski, cosine, jaccard, and Manhattan distance [15]. Regression and classification cases can utilize Random Forest, a supervised learning algorithm. People often use Random Forest because of its simplicity and good stability. Ensemble machine learning incorporates Random Forest as a prediction model, a collaborative effort between several models. We build random forest decisions based on random vectors, and form this tree by selecting a random F value. The parameters determine the random forest's intensity based on the selected F value and the number of trees to construct [3]. The selected F value forms a correlation; a small value leads to a small correlation [16].

During the testing and evaluation stage, we calculate the error value using the Mean Squared Error (MSE) evaluation metric. This stage evaluates the data mining output from the previous phase [17]. MSE works by measuring the accuracy of the model's estimated value expressed in the average square of the error. It can also be used to compare the accuracy of predictions between different forecasting methods [18].

III. RESULT AND DISCUSSION

We execute this step in multiple phases, including One Hot Encoding and partitioning the dataset with the train_test_split function. In addition, we manually convert the categorical feature to numeric form and add the age feature to represent age on the smartphone. This process is used to change each value in the column into a new column and fill it with binary values, ranging from categorical variables to numeric variables to 0 and 1. We use one-hot-encoding functions to generate new features that accurately represent categorical variables. The train_test_split function's dataset division process is crucial in machine learning, as it determines the model's performance on unfamiliar data. Therefore, it is necessary to train and test the model using two different datasets. Figure 2 displays the correlation of features in the data.

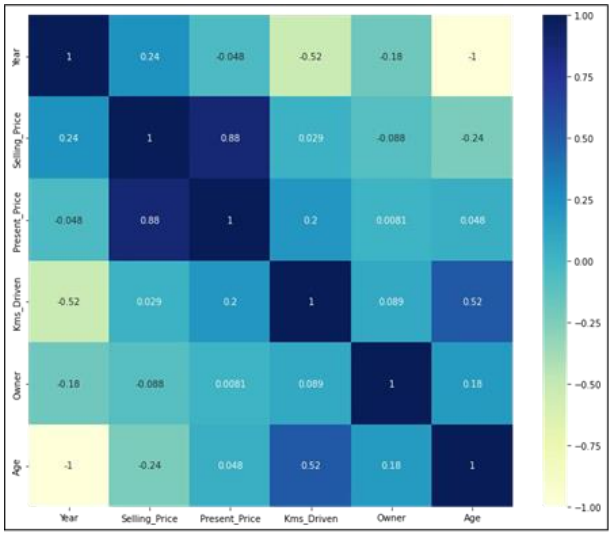

Fig. 2. Fitur Correlation

Figure 2 shows the correlation coefficient, which ranges between -1 and +1. The correlation coefficient measures the strength of the relationship between two variables and its direction, either positive or negative. Regarding the strength of the relationship between variables, the closer the value is to 1 or -1, the stronger the correlation. Meanwhile, the closer the value is to 0, the weaker the correlation.

After completing the data preparation process, the next step is to enter the modeling stage using two algorithms, namely K-Nearest Neighbors and Random

Forest. The results of this training process will then be used to calculate the error value and conduct trials on the test data. The first experiment compares the accuracy of each model; the second experiment calculates the error value using the MSE metric and compares the results; and the last experiment compares the results obtained by the two models using the predict function. We execute the modeling procedure by selecting the hyperparameter tuning neighbors that exhibit the highest performance on the data. Figure 3 displays the results.

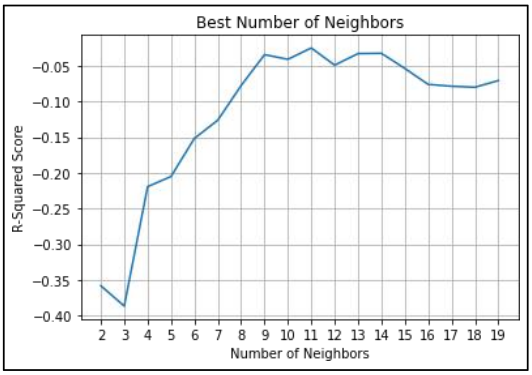

Fig. 3. Best neighbors

Figure 3 shows the results obtained with a value of n = 2. If we refer to the image above, 3 should be the best value. However, upon testing the value of 3 as a parameter n, we found that it yielded lower accuracy compared to the value of 2, leading us to opt for the value of 2.

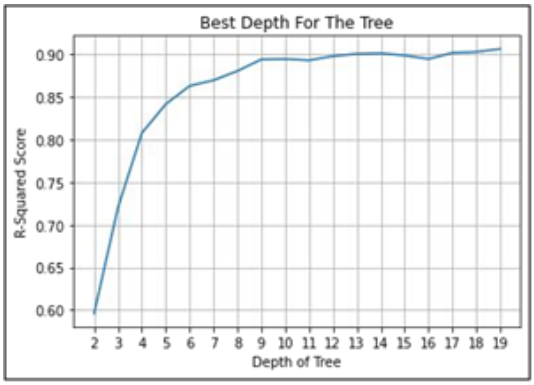

Fig. 4. Best depth

Figure 4 displays the outcomes of the modeling process, which utilized the optimal hyperparameter tuning for max_depth on the data. We apply a max_depth value of 12 to the model during this process. Table 1 displays the accuracy and error values obtained for both models.

Tabel 1. KNN and Random Forest model test results

Model	Accuracy	Error	
		Train	Test
KNN	59.17%	0.009817	0.044094
Random Forest	96.38%	0.001003	0.003206

Table 1 shows the results of testing the K-Nearest Neighbors (KNN) and Random Forest models in terms of accuracy and error, both for train and test data. K-Nearest Neighbors (KNN) is an algorithm that uses a nearest neighbor-based approach for classification or regression. We calculate accuracy and error based on how well the model predicts smartphone prices in the available data. Random Forest is an ensemble algorithm that uses multiple decision trees to improve prediction performance. It often produces better results because it reduces overfitting. The KNN model was able to correctly predict 59.17% of the entire test data. This accuracy is relatively low, indicating that KNN is less effective for this dataset or may require further tuning of its hyperparameters. The Random Forest model has a very high success rate of correctly predicting smartphone prices (96.38%). This figure shows that Random Forest is much more effective than KNN for this dataset.

Train Error measures how well the model fits the training data. The low train error (0.009817) indicates that the KNN model is able to adapt well to the training data, but the higher test error (0.044094) indicates the possibility of overfitting, where the model performs well on the training data but poorly on the test data. The very low train error (0.001003) indicates that Random Forest fits the training data very well. The low test error (0.003206) also indicates that the model generalizes well on the test data, showing no signs of overfitting like in KNN. The much higher accuracy of Random Forest (96.38%) compared to KNN (59.17%) indicates that Random Forest is better able to handle the complexity of the data in this study. This could be due to Random Forest's ability to combine multiple decision trees, thereby reducing bias and variance. The significant difference between the train error and test error in KNN suggests that KNN may be suboptimal for this dataset. The need for tuning the K parameter in KNN becomes clear. Random Forest, with very low error on both train and test data, shows that this model not only fits well but also has strong generalization ability. This is often one of the main advantages of ensemble models such as Random Forest. Based on these results, we conclude that Random Forest outperforms other models for smartphone price prediction in the context of this study's dataset. Researchers or practitioners may choose Random Forest for more accurate and reliable predictions. This table clearly shows the superiority of the Random Forest model over KNN in predicting smartphone prices, with much higher accuracy and lower error on both train and test data.

We recommend performing further hyperparameter tuning for the KNN model, such as determining the optimal K value and choosing an appropriate distance metric. You can use grid search or random search to determine the optimal combination of hyperparameters. While the Random Forest model yields excellent results, we can

refine the number of trees (n_estimators), maximum tree depth (max_depth), and other parameters to achieve the best outcomes. Conducting a deeper analysis of the features used in the model can help improve accuracy. We should remove or further process less relevant or redundant features to enhance the model's performance. Cross-validation techniques, such as k-fold cross-validation, can provide a more accurate picture of model performance on unseen data. This will assist in identifying and mitigating overfitting. Using data augmentation techniques to improve the quality and quantity of data can help the model learn better. In addition, better preprocessing, such as normalization or standardization of features, can also improve model performance. Although Random Forest is already an ensemble method, trying other ensemble methods such as Gradient Boosting Machines (GBM), AdaBoost, or XGBoost can provide better results. Combining predictions from multiple ensemble models can also be an effective approach. In addition to accuracy and error, using other evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), or R-squared can provide a more comprehensive picture of model performance. We hope that by taking these suggestions into account, future research can create a smartphone price prediction model that is more accurate and reliable, while also addressing various challenges that may arise during the prediction process.

IV. CONCLUSION

The comparative analysis of machine learning algorithms in smartphone price prediction yields several key conclusions. We have proven the superiority of the Random Forest algorithm over KNN in terms of accuracy and error. Random Forest achieved an accuracy of 96.38% with a very low error on both training data (0.001003) and testing data (0.003206). This shows that Random Forest is able to handle data complexity and provide more reliable and accurate predictions. In contrast, KNN only achieved an accuracy of 59.17% with a higher error on testing data (0.044094). For this study's dataset, KNN is less effective than Random Forest. The KNN model shows signs of overfitting, with a very low error on training data (0.009817) but much higher on testing data (0.044094). This shows that KNN is able to adapt well to training data but is less able to generalize well to new data. Random Forest, on the other hand, shows excellent generalization ability with low and consistent error on both training and testing data. This shows that Random Forest not only fits well but is also able to make accurate predictions on new data. Random Forest, as an ensemble model, has the advantage of reducing overfitting and increasing accuracy by combining predictions from multiple decision trees. This advantage is evident in this study's results, where Random Forest significantly outperforms KNN.

REFERENCES

[1] Amalia, M. Radhi, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, “Prediksi Harga Smartphone menggunakan Algoritma Regressi dengan Hyper-Parameter Tuning,” *JUSIKOM PRIMA*, vol. 4, no. 2, pp. 28–32, 2021, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2479.

[2] V. W. Siburian and I. E. Mulyana, “Prediksi Harga Ponsel menggunakan Metode Random

Forest," in *Annual Research Seminar (ARS) 2018*, 2018, pp. 144–147.
[3] R. Amanda and E. S. Negara, "Analysis and Implementation Machine Learning for YouTube Data Classification by Comparing the Performance of Classification Algorithms," *J. Online Inform.*, vol. 5, no. 1, pp. 61–72, 2020, doi: 10.15575/join.v5i1.505.
[4] A. Saleh and F. Nasari, "Penerapan Equal-Width Interval Discretization dalam Metode Naive Bayes untuk Meningkatkan Akurasi Prediksi Pemilihan Jurusan Siswa (Studi Kasus: MAS PAB 2 Helvetia, Medan)," *J. Masy. Telemat. dan Inf.*, vol. 9, no. 1, pp. 1–12, 2018, doi: 10.17933/mti.v9i1.113.
[5] B. Kriswantara and R. Sadikin, "Used Car Price Prediction with Random Forest Regressor Model," *J. Inf. Syst. Informatics Comput.*, vol. 6, no. 1, pp. 40–49, 2022, doi: 10.52362/jisicom.v6i1.752.
[6] B. Kriswantara, Kurniawati, and H. F. Pardede, "Prediksi Harga Smrtphone dengan Machine Learning," *Syntax Lit. J. Ilm. Indones.*, vol. 6, no. 5, p. 6, 2021.
[7] S. S. Winarto and T. Sutojo, "Menentukan Harga Smartphone dengan Menggunakan Metode Fuzzy Mamdani dan Metode Jaringan Syaraf Tiruan," *Techno.COM*, vol. 11, no. 3, pp. 134–141, 2012.
[8] G. N. Ayuni and D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ," *J. Telemat.*, vol. 14, no. 2, pp. 79–86, 2020, doi: 10.61769/telematika.v14i2.321.
[9] M. A. A. Syukur and M. Faisal, "Penerapan Model Regresi Linear Untuk Estimasi Smartphone Bekas Menggunakan Bahasa Python," *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 11, no. 2, pp. 182–191, 2023, doi: 10.37905/euler.v11i2.20698.
[10] F. Rahmawati and N. Merlina, "Metode Data Mining Terhadap Data Penjualan Sparepart Mesin Fotocopy Menggunakan Algoritma Apriori," *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, vol. 6, pp. 9–20, Mar. 2018, doi: 10.33558/piksel.v6i1.1390.
[11] Y. Suparman, "Perlukah Cross Validation dilakukan? Perbandingan antara Mean Square Prediction Error dan Mean Square Error sebagai Penaksir Harapan Kuadrat Kekeliruan Model," in *Seminar Nasional Matematika dan Pendidikan Matematika*, 2012, pp. 833–839.
[12] Z. Arifin, "Education Research Methodology," 2017, doi: 10.4324/9781315149783.
[13] B. Santoso, *Data mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, 1st ed. Yogyakarta: Graha Ilmu, 2007.
[14] N. M. Putry and B. N. Sari, "Komparasi Algoritma KNN dan Naïve Bayes untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus," *EVOLUSI J. Sains dan Manaj.*, vol. 10, no. 1, 2022, doi: 10.31294/evolusi.v10i1.12514.
[15] K. M. A. Pasaribu, R. E. Saputra, and C. Setianingsih, "Sistem Informasi Monitoring Bencana Alam dari Data Media Sosial menggunakan Metode K-Nearest Neighbor," in *e-Proceeding of Engineering*, 2021, pp. 6684–6693.
[16] Mambang and A. Byna, "Analisis Perbandingan Algoritma C.45, Random Forest dengan Chaid Decision Tree untuk Klasifikasi Tingkat Kecemasan Ibu Hamil," in *Seminar Nasional Teknologi Informasi dan Multimedia*, 2017, pp. 103–108.
[17] N. L. W. S. R. Ginantra *et al.*, *Data Mining dan Penerapan Algoritma*, 1st ed. Yayasan Kita Menulis, 2021.
[18] S. M. Robial, "Perbandingan Model Statistik pada Analisis Metode Peramalan Time Series (Studi Kasus: PT. Telekomunikasi Indonesia, Tbk Kandatel Sukabumi)," *J. Ilm. SANTIKA*, vol. 8, no. 2, pp. 1–17, 2018.