VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6598

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

PREDICTION MODEL OF HUMAN DEVELOPMENT INDEX (HDI) USING K-NEAREST NEIGHBOR (KNN) ENSEMBLE

Fitri Nuraeni[1]; Siska Nuraeni[1]; Asri Mulyani[1]; Dede Kurniadi[1]

Computer Science Departement[1]
Institut Teknologi Garut[1]
https://www.itg.ac.id[1]
fitri.nuraeni@itg.ac.id*, 2006047@itg.ac.id, asrimulyani@itg.ac.id, dede.kurniadi@itg.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The Human Development Index (HDI) is an essential indicator in measuring the success of human development. Although some regions in Indonesia have experienced increased HDI, inequality between areas makes it difficult to predict future HDI values. This research aims to build an HDI prediction model using the ensemble K-nearest neighbor (KNN) method. The dataset consists of 574 data points with attributes of life expectancy, expected years of schooling, average years of education, and regional income per capita. The method used is SEMMA with z-score normalization, feature selection based on domain knowledge, and validation with 10-fold cross-validation. The results showed that the KNN Ensemble model with the Boosting (Adaboost) technique had the best performance with an average MAPE of 0.58%, which indicates that the model's predictions deviate by less than 1% from actual HDI values, which is considered highly accurate and reliable for policy planning. This model proved better than linear regression, neural networks, single KNN, and double exponential smoothing algorithms. The improved prediction accuracy of the proposed model provides local governments with a reliable tool for scenario-based development planning and policy simulation, contributing to achieving the Golden Indonesia 2045 strategic vision.*

Keywords*: ensemble, human development index, k-nearest neighbor, prediction.*

Intisari— *Indeks Pembangunan Manusia (IPM) merupakan salah satu indikator penting dalam mengukur keberhasilan pembangunan manusia. Meskipun beberapa daerah di Indonesia mengalami peningkatan IPM, namun ketimpangan antardaerah menyebabkan sulitnya memprediksi nilai IPM di masa mendatang. Penelitian ini bertujuan untuk membangun model prediksi IPM menggunakan metode ensemble K-nearest neighbor (KNN). Dataset yang digunakan terdiri dari 574 titik data dengan atribut angka harapan hidup, harapan tahun sekolah, rata-rata tahun sekolah, dan pendapatan per kapita daerah. Metode yang digunakan adalah SEMMA dengan normalisasi z-score, pemilihan fitur berdasarkan pengetahuan domain, dan validasi dengan 10-fold cross-validation. Hasil penelitian menunjukkan bahwa model KNN Ensemble dengan teknik Boosting (Adaboost) memiliki kinerja terbaik dengan rata-rata MAPE sebesar 0,58%, dengan nilai minimum sebesar 0,31% pada lipatan kesembilan. Model ini terbukti lebih baik dibandingkan dengan algoritma regresi linier, neural network, single KNN, dan double exponential smoothing. Akurasi prediksi yang ditingkatkan dari model yang diusulkan memberi pemerintah daerah alat yang andal untuk perencanaan pembangunan berbasis skenario dan simulasi kebijakan, yang berkontribusi pada pencapaian visi strategis Indonesia Emas 2045.*

Kata Kunci*: ensemble, indeks pembangunan manusia, k-nearest neighbour, prediksi.*

INTRODUCTION

To achieve the vision of an "Indonesia Emas 2045", various provinces in Indonesia have established various government programs as an integrated part of the national development strategy. One of them is West Java Province, whose main objective of its local government program is to increase the Human Development Index (HDI), a key indicator of the progress of human development and society's general welfare [1]. During 2010-2022, West Java experienced consistent annual HDI growth with an average of 0.84%. The most significant increase occurred in the aspect of decent living standards. Although the HDI in West Java reached a high category in Java Island, variations and inequalities are still evident in the HDI values at the district level. Therefore, the future state of HDI is still challenging to predict. T

he HDI is also an important metric in assessing multidimensional development across countries that integrates health, education, and economy. [2]. Understanding the development of the HDI is essential for policymakers and researchers to develop interventions and assess progress[3]. Developing countries are increasingly interested in machine learning applications for HDI prediction, including methodological innovations in GCC countries[4]. Research shows the success of machine learning techniques in forecasting HDI values by analyzing economic, health, and education indicators at the country level[5], [6]. This shows the development of machine learning applications in global HDI prediction.

Several studies are relevant to the research to be conducted, such as Akın and Koç [5] showed that the ensemble algorithm can predict the Human Development Index (HDI) value with a very high level of accuracy, but it only uses health indicators and does not consider the education and economic dimensions as essential components of HDI. Then [7] applied double exponential smoothing for Bojonegoro, resulting in two satisfaction clusters, but the double exponential smoothing shows limited capability in handling non-linear patterns and is highly sensitive to abrupt changes in data. Then [8] uses backpropagation in Wonosobo with an accuracy of 99.8%, although highly accurate in small-scale applications, backpropagation requires intensive tuning of hyperparameters and is prone to overfitting, making it less suitable for generalized regional HDI prediction tasks.

Then [9] applied double exponential smoothing in North Toraja with low prediction error, but the approach is limited to linear data and not adaptive enough to complex data patterns. While [10] used random forest classification in Eastern Indonesia, with 4.08% incorrect predictions, indicating that the classification approach is less appropriate for prediction problems with continuous attributes such as HDI. Therefore, this research uses the K-Nearest Neighbor (KNN) machine learning ensemble approach to overcome these limitations. KNN ensemble can handle non-linear data patterns, is adaptive to heterogeneous data, does not require specific data distribution assumptions, and has proven effective in reducing errors due to parameter selection sensitivity [11] Such as determining the value of k, so that it is expected to produce IPM prediction models with more accurate and stable performance [12] Compared to previous methods.

Based on these problems, this research aims to develop an HDI prediction model using an ensemble approach to the KNN algorithm to produce more accurate and stable predictions. This model is expected to positively contribute to more effective and sustainable development planning in various provinces in Indonesia, per the strategic goals towards Golden Indonesia 2045.

MATERIALS AND METHODS

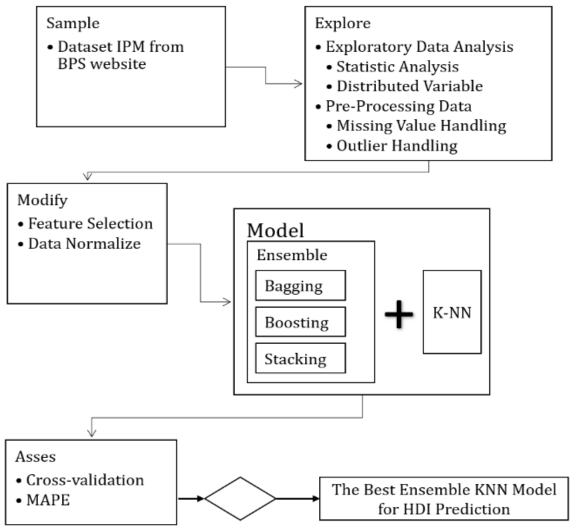
This research uses the Sample, Explore, Modify, Model, and Assess (SEMMA) method, which is a method that describes a procedural framework for carrying out Machine Learning tasks [9]. In addition to prediction accuracy, the evaluation should also include an analysis of the computation time required by each model. The results show that ensemble models, such as AdaBoost-KNN and Stacking-KNN, require longer computation time than single models, such as Linear Regression or KNN. This is due to the large number of estimators involved in the ensemble. However, the significant accuracy improvement of ensemble models can be a valid reason to consider higher computational costs, depending on the needs and context of the application.

Figure 1 shows that this method, as the name suggests, consists of a series of several activities [13]:

1. The first step is *sampling*, which is the process of collecting data to get enough datasets to obtain significant information while still being easy to process.
2. The exploration step involves analyzing the data to identify trends and anomalies and deeply understanding the data

3. In the modify step, the data is transformed by selecting, normalizing, and dividing the data for use in modeling.
4. The next step is the model. At this stage, the data from the exploration stage is processed using the KNN ensemble by applying three techniques, namely bagging, boosting, and stacking, to get the best results. As in the conventional KNN method, several "k" nearest objects from the training data are considered[14], then the closest distance [15] is used for prediction. Furthermore, ensemble learning uses several simple models or "weak learners" that are trained separately and combined to overcome their weaknesses, resulting in a combined model that is more accurate and performs better [16]. In this study, the optimal value of "k" is nine, which was empirically determined by conducting a grid search across a range of k values from 1 to 20, with k=9 yielding the lowest average MAPE during 10-fold cross-validation.
5. The assessment stage considers that prediction is a scientific method that uses historical data to forecast future events, aiming to reduce errors and increase accuracy [11]. In this study, the generated patterns are evaluated to determine whether they are helpful and reliable using Mean Absolute Percentage Error (MAPE) matrix evaluation, a standard measure for assessing prediction accuracy [17]. MAPE helps evaluate the prediction error relative to the actual value.

$$MAPE = \sum_{i=1}^{n} \frac{\frac{|x_i - y_i|}{x_i}}{m} * 100\% \qquad (1)$$



Source: (Research results, 2025)

Figure 1. Research Design

The performance analysis of the prediction process will consider the MAPE value, which is noted in Table 1 as a reference, as previous researchers have done [11].

Table 1. MAPE Value for Prediction Accuracy

MAPE Value	Prediction Accuracy
MAPE ≤ 10%	High
10% < MAPE ≤ 20%	Good
20% < MAPE ≤ 50%	Medium
MAPE >50%	Low

Source: (Research results, 2025)

This study uses a dataset collected from the official website of BPS West Java through the link https://jabar.bps.go.id/, with a dataset of 574 records and six main attributes, namely year, UHH, HLS, RLS, PKD, and IPM. Then, the model was applied to datasets for Lampung, Gorontalo, and Central Java provinces.

Table 2. Dataset Sample

No.	Year	UHH	HLS	RLS	PKD	IPM
1	2021	71.36	12.49	8.31	10410.00	70.60
2	2021	71.21	12.24	7.10	8850.00	67.07
3	2021	70.32	12.00	7.19	8052.00	65.56
4	2021	73.72	12.70	9.07	10307.00	72.73
5	2021	71.59	12.03	7.53	7961.00	66.45
6	2021	69.67	12.54	7.48	7829.00	65.90
7	2021	72.02	14.20	7.90	9259.00	70.93
8	2021	73.78	12.23	7.80	9409.00	69.71
9	2021	72.18	12.27	7.10	10368.00	69.12
...	..	..	..	..	..	..
81	2023	71.80	13.27	8.79	11356.00	73.08

Source: (Research results, 2025)

RESULTS AND DISCUSSION

This research utilizes the K-Nearest Neighbor ensemble method implemented using the Python programming language on the Google Colab platform. The first step in this research is to collect datasets from the official website of BPS West Java (https://jabar.bps.go.id/). This dataset consists of 574 data entries with six main attributes, with an explanation of each attribute:

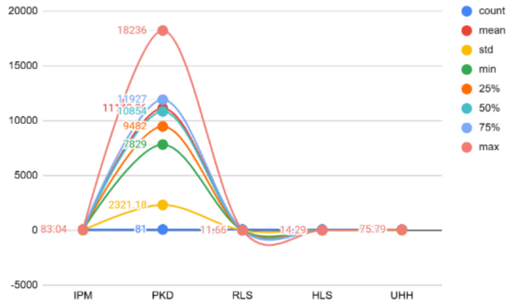
1) Year, the year of data collection.
2) UHH: Life expectancy is the average age people can expect to live.
3) HLS: expected years of schooling, the expected years of schooling for a 7-year-old child.
4) RLS: average years of schooling, the average years of education for the population aged 25 years and over.
5) PKD: regional per capita income, the community's average income, and welfare level.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6598

6) IPM: human development index, an indicator of success in improving the quality of human life.

Attribute selection was based on reference studies and the availability of consistent annual provincial-level data over the 2021-2023 period. In addition, the selected attributes were prioritized based on their proven direct impact on HDI

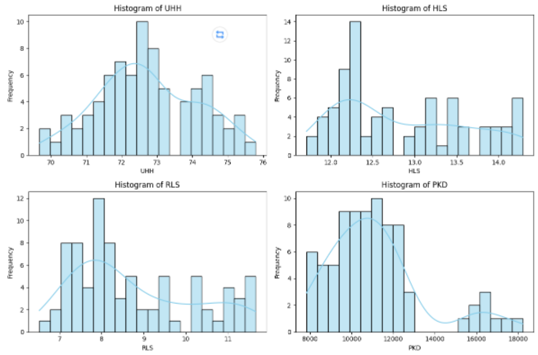
Once the sample is selected, the next step is the data exploration process to understand the dataset's characteristics through descriptive statistics, data visualization, and analysis to identify patterns, outliers, or other important information. This stage is important to understand the dataset's quality after data collection and as an initial step in analysis to uncover significant patterns and characteristics.



Source: (Research results, 2025)

Figure 2. Statistical Data Summary

At this stage, the researcher presents summary statistics to understand the data distribution, including each numerical column's mean, median, standard deviation, and minimum and maximum values. Figure 2 presents the summary descriptive statistics of the five variables UHH, HLS, RLS, PKD, and HDI, each variable showing different mean values and data distributions, reflecting the variation in HDI indicators.

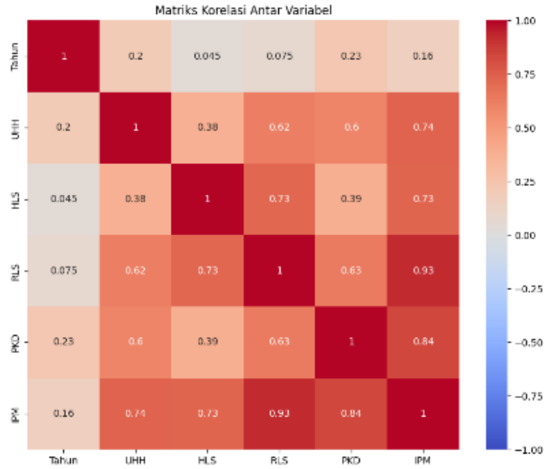


Source: (Research results, 2025)

Figure 3. Visualization of Variable Distribution

Figure 3 shows the histograms for all variables, where UHH, RLS, and PKD have right-skewed distributions, indicating many residents with low values. HLS is close to a normal distribution, indicating that most of the population is well-educated. This condition shows inequality in health and income despite good education.

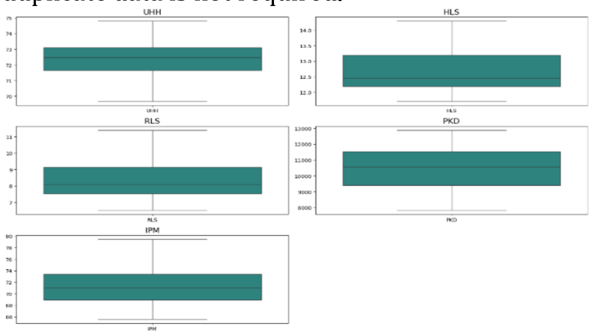
Figure 4 shows the correlation between the variables, where UHH and HDI correlate 0.14 (weak), HLS and HDI 0.75 (strong), RLS and HDI 0.65 (medium), PKD and HDI 0.78 (strong), and Years and HDI 0.92 (very strong). The most substantial relationships with HDI are HLS and PKD, while UHH and RLS have weaker but significant relationships.



Source: (Research results, 2025)

Figure 4. Correlation between the variables

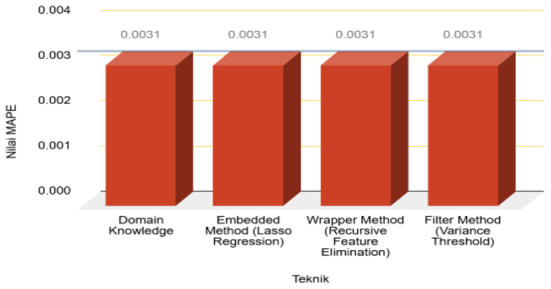
Missing data refers to a condition with incomplete or empty values in one or more criteria [18]. Before prediction begins, preprocessing is necessary to remove duplicate data, incomplete entries, and correct errors in the dataset [19]. The dataset used does not have missing values for all variables, so this study's cleaning process is unnecessary. In addition, the dataset has no duplicate data found on all attributes, so removing duplicate data is not required.



Source: (Research results, 2025)

Figure 5. Boxplot Outlier Data

Data outliers are observations with significant absolute residual values relative to other residuals, which may lead to violations of assumptions in the regression model [20]. From Figure 5, the y-axis shows the attribute values (UHH, HLS, RLS, PKD, HDI), with the x-axis showing the examined attributes. The visualization results show that there are no outliers for all attributes. Therefore, there is no need for special handling of outliers in the dataset examined using the Interquartile Range (IQR) method.

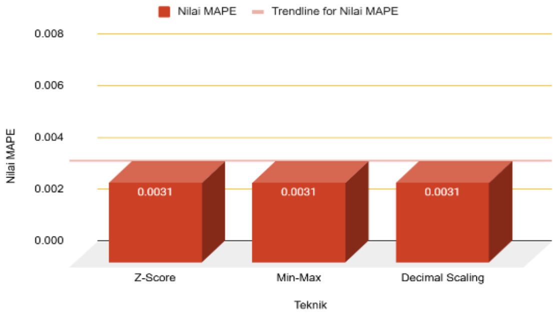


Source: (Research results, 2025)

Figure 6. Comparison of Feature Selection Techniques with MAPE Value

In the modify stage, the first thing to do is attribute selection, which aims to reduce data dimensions and eliminate irrelevant features, improving prediction performance [21]. Based on Figure 6, the feature selection technique does not affect the MAPE value. Then, the technique chosen is domain knowledge because features such as UHH, HLS, RLS, and PKD are the primary indicators that affect HDI.

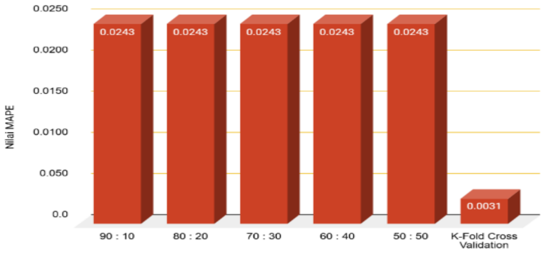
The data normalization process aims to adjust the scale of variables in the dataset, ensure a uniform range of values, and improve the efficiency of analysis and machine learning models [22]. In this study, the normalization process uses techniques such as Z-score, min-max, and decimal scaling to determine the optimal technique for predicting HDI.



Source: (Research results, 2025)

Figure 7. Comparison of Normalization Techniques

From Figure 7, all normalization techniques produce the same MAPE value of 0.0031. However, the Z-score was chosen because it preserves the original data distribution and reduces the influence of extreme values [23]. This selection is crucial to ensuring prediction consistency and maintaining the balanced contribution of each feature in analysis and modeling.

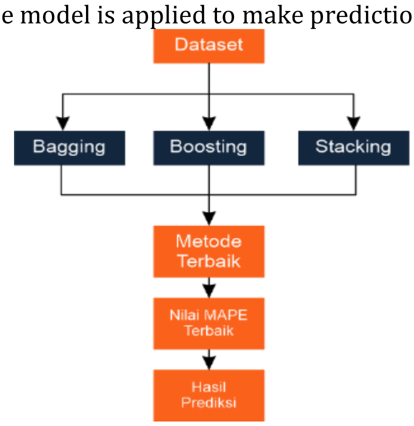


Source: (Research results, 2025)

Figure 8. Comparison of Data Splitting Techniques

At this stage, the dataset is divided into two parts: training data and testing data. Training data is used to build and train the model, while testing data is used to measure performance and evaluate the accuracy of the trained model. As in Figure 8, the dataset is divided into training data and testing data using the train-test separation technique, with variation ratios such as 90:10, 80:20, 70:30, 60:40, 50:50, and using k-fold cross-validation. The selection of data division techniques is based on the lowest MAPE value, where the K-Fold Cross Validation approach is the lowest, so this technique is chosen for data division in this study.

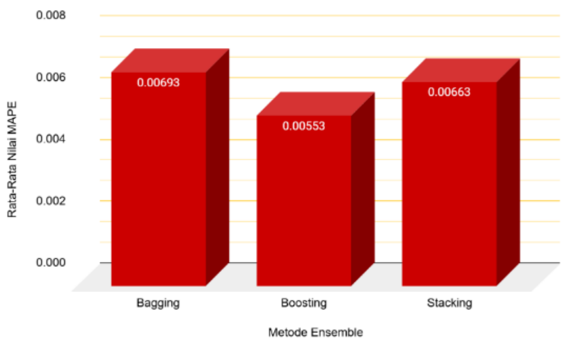
In the modeling stage, the prediction model is built by applying the KNN ensemble, according to Figure 9, which illustrates the workflow of the ensemble KNN modeling process, starting from dataset preparation, application of ensemble techniques (Bagging, Boosting, Stacking), model evaluation, and final prediction. The best ensemble method is selected based on the lowest MAPE, and then the model is applied to make predictions.



Source: (Research results, 2025)

Figure 9. Experimental Flow of KNN Ensemble Technique

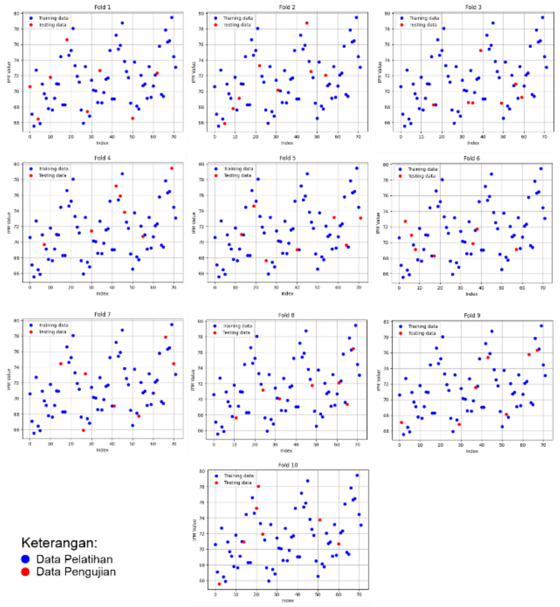
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri



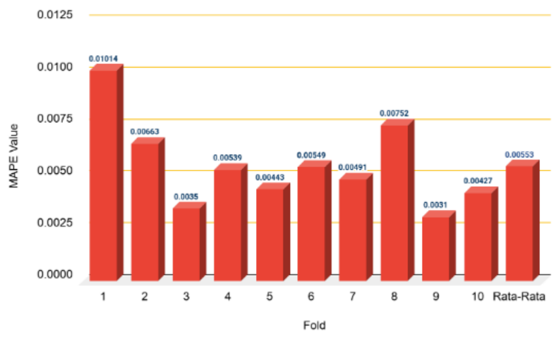
Source: (Research results, 2025)
Figure 10. Comparison of Ensemble Technique Experiment Results

Figure 10 shows a comparison chart of the lowest MAPE value of Boosting, 0.00553, followed by Stacking, 0.00663, and Bagging, 0.00693. This condition shows that boosting can improve accuracy by correcting previous errors, proving more effective than this study's other methods.

After determining the best ensemble technique, KNN with the Boosting (Adaboost) technique, cross-validation is performed by dividing the dataset into 10 folds. Figure 11 shows that each sub-image represents one fold, with the training data marked by blue dots and the testing data by red. Each part of the dataset is used for training and testing, providing a thorough evaluation of the performance of the KNN ensemble model with *Boosting (Adaboost).*

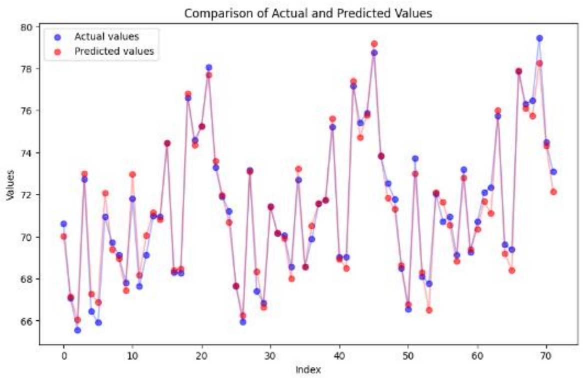


Source: (Research results, 2025)
Figure 11. The 10-*Fold Cross-Validation Result*



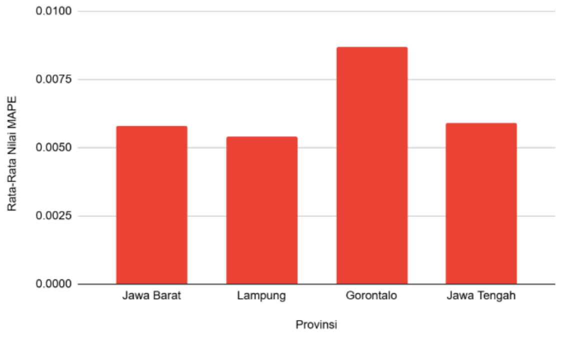
Source: (Research results, 2025)
Figure 12. MAPE Value for Each Fold and Average

Figure 12 displays the MAPE value of each fold, where the first fold has the highest value of 0.01014, while the ninth fold has the lowest value of 0.0031, and the average for all folds is 0.00553, indicating a prediction error rate of about 0.55%. The variation in MAPE values at each fold reflects fluctuations in model performance, but overall, the model shows good consistency. These results indicate that the KNN ensemble model with Boosting can produce predictions with reasonably high accuracy.



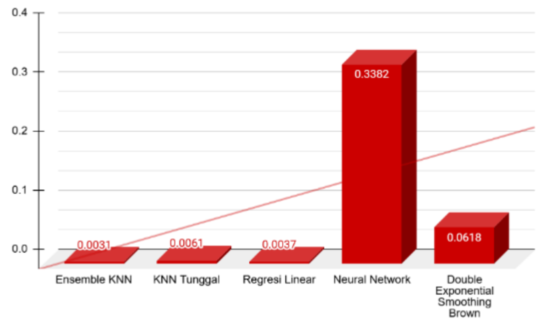
Source: (Research results, 2025)
Figure 13. Comparison of Actual and Predicted Value

Figure 13 shows the comparison between actual values (marked with blue dots) and predicted values (marked with red dots) generated by the KNN ensemble model with Boosting (Adaboost). The horizontal axis represents the data index, while the vertical axis shows the measured values. This graph shows that the KNN ensemble model is quite effective in following the trend of the actual values, with most of the predictions being close to the actual values. Despite some deviations, the model performs well in predicting values with a reasonably high level of accuracy. This condition indicates that the model is reliable in predicting HDI values, although there is room for further improvement to reduce prediction errors.



Source: (Research results, 2025)
Figure 14. Comparison of model evaluation results on datasets from other provinces

This study applied the KNN ensemble model to datasets from several provinces - Lampung, Gorontalo, and Central Java- to assess its consistency. Figure 14 shows that all provinces' MAPE values remain low, indicating good model performance despite data variations. The model achieved an MAPE of 0.0058 (K=9) in West Java, 0.0055 (K=4) in Lampung, 0.0079 (K=6) in Gorontalo, and 0.0059 (K=4) in Central Java. These results show that the KNN ensemble model remains accurate and reliable even when applied to different datasets.



Source: (Research results, 2025)
Figure 15. Comparison of KNN Ensemble Model with Other Models

At last, the proposed model, compared to all baseline models, was trained and evaluated using the same dataset and k-fold cross-validation setting to ensure fair and unbiased comparison. Figure 15 compares the performance of the models in predicting the HDI value in West Java using MAPE. Ensemble KNN has the lowest MAPE of 0.0031, followed by Linear Regression with 0.0037 and Single KNN with 0.0061. Neural Network has the highest MAPE of 0.3382, while Double Exponential Smoothing Brown records a MAPE of 0.0618. This condition shows that Ensemble KNN with Boosting (Adaboost) provides the best prediction and proves that the KNN ensemble model performs better than others.

In this comparison model, the architecture adopts a feedforward neural network configuration designed to model the relationship between five input features and the Human Development Index (HDI) as the target variable. The network architecture consists of an input layer that receives five-dimensional input vectors, followed by a first hidden layer of 64 neurons. The unique choice of the ReLU activation function in this layer is a key factor in the model's performance. To mitigate the risk of overfitting, a dropout layer with a dropout rate of 0.2 is applied. The second hidden layer includes 32 neurons, also utilizing the ReLU activation function, and is followed by another dropout layer with the same dropout rate. Finally, the network concludes with an output layer comprising a single neuron with a linear activation function, suitable for regression tasks. The model was compiled using the Adam optimizer with a learning rate of 0.001.

The training process of the neural network model shows a steady decline in both training and validation losses over 100 epochs, indicating effective learning. The model reached a final MAPE of 0.3382, which suggests moderate predictive capability but lower accuracy compared to the ensemble model. This is due to the relatively small dataset size, where deep learning models generally require more data to outperform classical methods [24], [25]. The consistent drop in validation loss without signs of divergence confirms that the applied architecture, including dropout layers and moderate depth, was able to generalize reasonably well despite not applying early stopping.

The study results show that the KNN ensemble model with Boosting (Adaboost) performs well in predicting HDI, as evidenced by the low average MAPE based on cross-validation. The consistency of the small MAPE values between folds indicates that the model has good generalization and does not experience overfitting. In addition, the low MAPE value in each fold indicates that the model does not experience underfitting, so it can provide accurate predictions. Compared with previous research [11]These results strengthen the finding that applying the KNN ensemble can improve prediction accuracy. With the lowest MAPE value of 0.31%, this model is reliable for predicting HDI.

AdaBoost generally outperforms bagging and stacking in estimation models across various domains, particularly in terms of accuracy and error reduction [26], [27]. This is because boosting combines multiple weak classifiers into a strong classifier by reducing both bias and variance, whereas bagging primarily reduces variance by averaging independent models, and stacking relies heavily on the meta-model's ability to combine base learners [28].

Although the AdaBoost-KNN ensemble demonstrated the best predictive performance among the evaluated models, it required approximately 2.5 times longer training than the single KNN model. This increased computational cost is attributed to the sequential training of multiple weak learners, where each learner is trained to correct the errors of its predecessor [29]. Despite the absence of parallelization, this iterative approach effectively enhances accuracy by reducing bias and variance [30].

Nevertheless, the potential computational burden associated with the ensemble method was mitigated in this study due to the relatively small dataset and low feature dimensionality, comprising only five input variables. This characteristic significantly reduced the training overhead, indicating that ensemble techniques such as AdaBoost-KNN can still be efficiently implemented when applied to structured datasets with simple feature spaces [31]. Consequently, the model balances prediction accuracy and computational efficiency, especially in regression tasks involving limited and low-dimensional data.

CONCLUSION

This research has produced a K-Nearest Neighbor (KNN) ensemble prediction model with z-score normalization technique, Adaboost ensemble booster method, and the best k value at K=9. This KNN ensemble model has an average MAPE value of 0.0058 or 0.58%, indicating that the model falls into the high accuracy category. In addition, by comparing this model with several other algorithms, namely single KNN, Linear Regression, Neural Network, and Double Exponential Smoothing, this KNN ensemble model provides the best performance. Furthermore, the proposed model can be integrated into local government dashboards or decision support systems to provide dynamic and real-time HDI forecasts. This can assist provincial leaders in evaluating the impact of policy interventions and reallocating resources more effectively. A suggestion for future research is to develop the prediction model by adding other relevant variables, such as economic indicators, poverty rate, and infrastructure, to improve the accuracy and generalizability of the model in the future.

REFERENCE

[1] Badan Pusat Statistik, “Indeks Pembangunan Manusia (IPM) Jawa Barat meningkat pada tahun 2022 yang mencapai 73,12,” Badan Pusat Statistik Jawa Barat.

[2] A. A. Vladimirskaya and M. G. Kolosnitsyna, “Factors in Life Expectancy: A Cross-Country Analysis,” *Vopr. Stat.*, vol. 30, no. 1, pp. 70–89, Feb. 2023, doi: 10.34023/2313-6383-2023-30-1-70-89.

[3] A. Anekawati, . P., M. Rofik, and S. Hidayat, “A Spatial Error Model in Structural Equation for the Human Development Index Modeling,” *Int. J. Math. Eng. Manag. Sci.*, vol. 9, no. 3, pp. 537–556, Jun. 2024, doi: 10.33889/IJMEMS.2024.9.3.028.

[4] M. Goldani, “A Numerical Assessment for Predicting Human Development Index (HDI) Trends in the GCC Countries,” Nov. 2024, [Online]. Available: http://arxiv.org/abs/2411.01177

[5] P. Akin and T. Koç, “Prediction of Human Development Index with Health Indicators Using Tree-Based Regression Models,” *Adiyaman Univ. J. Sci.*, vol. 11, no. 2, pp. 410–420, Nov. 2021, doi: 10.37094/adyujsci.895084.

[6] V. Georgiev, S. Hadzhikoleva, and E. Hadzhikolev, “Impact of Global Country Indicators on Life Expectancy,” *Comput. Sci. Interdiscip. Res. J.*, vol. 1, no. 1, Aug. 2024, doi: 10.70862/CSIR.2024.0101-04.

[7] Y. Farida, D. A. Sulistiani, and N. Ulinnuha, “Peramalan Indeks Pembangunan Manusia (IPM) Kabupaten Bojonegoro Menggunakan Metode Double Exponential Smoothing Brown,” *J. Teorema Teor. dan Ris. Mat.*, vol. 6, no. 2, pp. 173–183, 2023, doi: http://dx.doi.org/10.25157/teorema.v6i2.5521.

[8] N. N. Amiroh and D. Avianto, “Prediksi Indeks Pembangunan Manusia di Kabupaten Wonosobo Menggunakan Algoritma Backpropagation,” *Techno.Com*, vol. 22, no. 2, pp. 388–399, 2023, doi: 10.33633/tc.v22i2.7980.

[9] A. C. Ampang, B. Eden, W. Asrul, and M. A. Nur, “Implementasi Metode Double Exponential Smoothing Untuk Prediksi Indeks Pembangunan Manusia (IPM) Di Kabupaten Toraja Utara,” *J. Fokus Elektroda*, vol. 08, no. 03, pp. 197–207, 2023.

[10] A. Arisandi and S. Syarifuddin, “Memprediksikan Indeks Pembangunan Manusia di Wilayah Indonesia Bagian Timur

Menggunakan Random Forest Classification," *J. Math. Theory Appl.*, vol. 5, no. 1, pp. 1–6, 2023, doi: 10.31605/jomta.v5i1.2402.

[11] M. Jusman, N. Nur'eni, and L. Handayani, "Ensemble K-Nearest Neighbors Method to Predict Composite Stock Price Index (CSPI) in Indonesia," *J. Mat. Stat. dan Komputasi*, vol. 18, no. 3, pp. 423–433, 2022, doi: 10.20956/j.v18i3.19641.

[12] M. Fajri, S. Syafriandi, D. Vionanda, and Z. Zilrahmi, "Prediksi Harga Minyak Mentah Dunia Menggunakan Metode Ensemble k-Nearest Neighbor," *J. Pendidik. Tambusai*, vol. 7, no. 2, pp. 17357–17368, 2023.

[13] Iin, R. Supriatna, Mulyawan, and D. Rohman, "Penerapan Natural Language Processing Dalam Analisis Sentimen Cawapres 2024 Menggunakan Algoritma Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 1109–1115, 2024, doi: 10.36040/jati.v8i1.8572.

[14] H. Gunawan, A. Chusyairi, and M. I. Saputra, "Penerapan K-Nearest Neighbor Dengan Metode Euclidean Distance Untuk Klasifikasi Tingkat Ketebalan Cat Di PT XYZ," *J. Teknol. dan Ilmu Komput.*, vol. 01, no. 2, pp. 59–72, 2025, doi: 10.35134/Jutekom.v9i2.1.

[15] J. J. Pangaribuan, A. Maulana, and R. Romindo, "UNLEASHING THE POWER OF SVM AND KNN: ENHANCED EARLY DETECTION OF HEART DISEASE," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 10, no. 2, pp. 342–351, Nov. 2024, doi: 10.33480/jitk.v10i2.5719.

[16] L. M. Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," *J. Masy. Inform.*, vol. 13, no. 1, pp. 33–44, 2022, doi: 10.14710/jmasif.13.1.42912.

[17] S. Bhattacharya, K. Kalita, R. Čep, and S. Chakraborty, "A comparative analysis on prediction performance of regression models during machining of composite materials," *Materials (Basel).*, vol. 14, no. 21, p. 6689, Nov. 2021, doi: 10.3390/ma14216689.

[18] W. Sudrajat and I. Cholid, "K-Nearest Neighbor (K-NN) Untuk Penanganan Missing Value Pada Data UMKM," *J. Rekayasa Sist. Inf. dan Teknol.*, vol. 1, no. 2, pp. 54–63, 2023, doi: 10.59407/jrsit.v1i2.77.

[19] A. Nur Aziziah, A. Irma Purnamasari, and I. Ali, "Penerapan Algoritma C4.5 Untuk Prediksi Stok Bahan Minuman Di Cafe Semanis," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 292–295, 2024, doi: 10.36040/jati.v8i1.8347.

[20] A. Husain and S. R. W. Jamaluddin, "Pemodelan Data Angka Kematian Bayi Menggunakan Regresi Robust," *SAINTEK J. Sains, Teknol. Komput.*, vol. 1, no. 1, pp. 1–7, 2024.

[21] D. Cahya Putri Buani, "Penerapan Algoritma Naïve Bayes dengan Seleksi Fitur Algoritma Genetika Untuk Prediksi Gagal Jantung," *EVOLUSI J. Sains dan Manaj.*, vol. 9, no. 2, pp. 43–48, 2021, doi: 10.31294/evolusi.v9i2.11141.

[22] V. Rapika Sari, E. Buulolo, and K. Kunci ABSTRAK, "Implementasi Algoritma K-Means dengan Normalisasi Sigmoidal Untuk Klastering Data Ternak Sapi," *Jikteks*, vol. 02, no. 01, pp. 30–42, 2023.

[23] I. N. Simbolon, H. D. S. . Siburian, and W. A. Manik, "Prediksi Kualitas Air Sungai Di Jakarta Menggunakan Knn Yang Dioptimalisasi Dengan Pso," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4191.

[24] A. Althnian *et al.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Appl. Sci.*, vol. 11, no. 2, pp. 1–18, Jan. 2021, doi: 10.3390/app11020796.

[25] S. Silvey and J. Liu, "Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study," *J. Med. Internet Res.*, vol. 26, p. e60231, Dec. 2024, doi: 10.2196/60231.

[26] C. W. Teoh, S. B. Ho, K. S. Dollmat, and C. H. Tan, "Ensemble-Learning Techniques for Predicting Student Performance on Video-Based Learning," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 8, pp. 741–745, 2022, doi: 10.18178/ijiet.2022.12.8.1679.

[27] T. Bokaba, W. Doorsamy, and B. S. Paul, "A Comparative Study of Ensemble Models for Predicting Road Traffic Congestion," *Appl. Sci.*, vol. 12, no. 3, p. 1337, Jan. 2022, doi: 10.3390/app12031337.

[28] A. Satoła and K. Satoła, "Performance comparison of machine learning models used for predicting subclinical mastitis in dairy cows: Bagging, boosting, stacking, and super-learner ensembles versus single machine learning models," *J. Dairy Sci.*, vol. 107, no. 6, pp. 3959–3972, Jun. 2024, doi: 10.3168/jds.2023-24243.

[29] K. Abegaz and İ. Etikan, "Boosting the Performance of Artificial Intelligence-

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6598

Driven Models in Predicting COVID-19 Mortality in Ethiopia," *Diagnostics*, vol. 13, no. 4, p. 658, Feb. 2023, doi: 10.3390/diagnostics13040658.

[30] O. Nooruldeen, M. R. Baker, A. M. Aleesa, A. Ghareeb, and E. H. Shaker, "Strategies for predictive power: Machine learning models in city-scale load forecasting," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 6, no. December, p. 100392, Dec. 2023, doi: 10.1016/j.prime.2023.100392.

[31] M. S. H. Rabbi *et al.*, "Performance evaluation of optimal ensemble learning approaches with PCA and LDA-based feature extraction for heart disease prediction," *Biomed. Signal Process. Control*, vol. 101, p. 107138, Mar. 2025, doi: 10.1016/j.bspc.2024.107138.