Vol 1 No 8 2024 || E-ISSN 2997-7258

The Journal of Academic Science
journal homepage:
https://thejoas.com/index.php/

Application of The K-Nearest Neighbor Ensemble Method in Predicting The Human Development Index and Its Relationship with The Unemployment Rate in Central Sulawesi Province

[1] Selvy Musdalifah, [2] Saputra Hardiansyah, [3] Andri
[1,2,3] Tadulako University, Soekarno-Hatta Street Km. 09 Tondo, Palu 94118, Central Sulawesi, Indonesia
Email: selvymusdalifah@gmail.com, saputrahardiansyah26@gmail.com, andri.math2007@gmail.com

KEY WORDS

Ensemble K-NN, HDI, Predictions, Unemployment

A B S T R A C T

Unemployment has an impact on economic development and also community welfare. One of the indicators to measure the level of development is the Human Development Index (HDI). Therefore, the purpose of this research is to obtain the results of HDI prediction and the relationship with the unemployment rate. The method used in this study is the K-NN Ensemble method. This research was conducted using HDI data consisting of data on school life expectancy, per capita expenditure, average school duration, and life expectancy as well as unemployment rate data. The results of the HDI prediction in Central Sulawesi Province are with an average of 69.03 with a MAPE value of 1.45%. Based on the correlation test, it was found that the relationship between HDI and the unemployment rate with a significant value below 0.05 was 0.001. The resulting correlation value is 0.803 which shows that the closeness is very strong and has a negative correlation coefficient property.

1. Introduction

One of the developing countries that is still experiencing the process of economic development to achieve prosperity is Indonesia. Welfare requires good job opportunities and equal distribution of income in the community. In Indonesia, there is a gap in employment opportunities and the labor force, which means that the increase in the labor force is not proportional to the increase in employment opportunities. As a result, unemployment will have an impact on social life, including an increase in crime and violence, which will certainly hinder economic stability and progress and reduce the number of unemployment cases. In achieving a welfare, equitable development is needed in society.

Development is the government's effort to improve the welfare and prosperity of the community. Development results can be measured by indicators, one of the indicators that can be used is the Human Development Index (HDI), which is a field of human development that can measure success in a country or region. HDI consists of three indicators, namely living standards, health, and education. Health indicators include the average life expectancy, and education indicators include the average expected length of schooling (Devita Rosmadayanti, Niniek Imaningsih, 2021). HDI has a relationship with the

unemployment rate. That is, if the HDI value of a region increases, then the unemployment rate in that region will decrease, and conversely, if the HDI value of a region decreases, then the unemployment rate in that region will increase (Mahroji & Nurkhasanah, 2019).

Based on this, the researcher wants to conduct research using the same method but with a different topic of the problem. This study predicted the HDI level in the coming year, and looked for the relationship between HDI and the unemployment rate in Central Sulawesi Province through a correlation test using SPSS software. This correlation test was carried out to find out how much of a relationship and also the closeness between HDI and the unemployment rate. This study uses the k-nearest neighbor ensemble method to apply prediction testing. One of the drawbacks of the K-NN method is the inability to make predictions for time data sets. In the calculation of distance values, the K-NN method has not been able to determine the value of k that has a result that is close to the actual data value, and the determination of parameters will give better results. Therefore, the K-NN method must be optimized by adding or designing a good method so that it can work consistently. Furthermore, ensemble techniques are added to solve the problem that the author wants to research. The ensemble technique, which can be used to optimize the performance of the K-NN method in making predictions, does not select one best prediction from many prediction candidates and then make a guess from the prediction. Instead, this technique combines various existing predictions to make a final prediction.

2. Methodology

This research is carried out through a systematic and structured series of steps aimed at analyzing the relationship between the Human Development Index (HDI) and the unemployment rate in Central Sulawesi Province, as well as applying mathematical methods to find solutions to the existing problems. The first step in this research is conducting a literature review related to the issues of HDI and the unemployment rate. This review focuses on examining relevant mathematical models that can be used to provide solutions to the problems related to human development and unemployment. Based on this, the researchers aim to understand various approaches that have been applied in previous studies, as well as explore the potential of these models to be applied in the study area.

The second step involves formulating problems related to HDI and the unemployment rate in Central Sulawesi Province. This process is carried out by referring to the results of the previous literature review, where the researchers use information from previous studies to identify specific challenges faced in the region. In this step, problem formulation is done to understand the social and economic dynamics that influence HDI and the unemployment rate, as well as how mathematics can be used as a tool to analyze and solve these issues.

The third step involves collecting relevant data, specifically data on HDI, which consists of four main indicators, and data on the unemployment rate in Central Sulawesi Province. This data is collected to ensure that the research is based on solid and representative information. The HDI data includes indicators such as life expectancy, education level, and standard of living, which will be further analyzed to see how they relate to the unemployment rate in the province.

Next, the collected data will be processed using a mathematical method, namely the Ensemble K-nearest neighbor (K-NN) method, to predict the future HDI. The K-NN method is a machine learning technique used for making predictions based on the proximity of data to other data points. This method was chosen for its ability to handle regression and classification problems with relatively accurate results, as well as its ease of implementation.

In the fifth step, to further explore the relationship between HDI and the unemployment rate, a correlation test will be conducted using SPSS software. The correlation test aims to measure the

extent of the linear relationship between these two variables, allowing the researchers to determine whether there is a significant correlation between HDI and the unemployment rate. SPSS is used because it is a statistical software commonly used in social and economic research, with the ability to perform statistical analysis quickly and easily.

Finally, in the last step, the researchers will obtain the results from the entire process and draw conclusions based on these findings. The conclusions will summarize the relationship between HDI and the unemployment rate in Central Sulawesi Province, and provide recommendations regarding policies that can be implemented to improve human development and reduce the unemployment rate.

Overall, this research combines theoretical approaches through literature review with empirical approaches using data and mathematical methods to find solutions to the problems faced by Central Sulawesi Province. By applying the K-NN method and correlation testing, this research aims to provide a better understanding of the factors affecting HDI and the unemployment rate, as well as offer guidance for development policies in the region.

3. Result and Discussion

3.1. Human Development Index Prediction

1. Data Collection

The data used is data taken through the Central Sulawesi Province BPS website. The data taken were in the form of HDI data, School Length Expectancy (HLS), Per Capita Expenditure (PK), Average School Length (RLS) and Life Expectancy (UHH) from 2020 to 2023, and for the process of finding the relationship between HDI and the unemployment rate, data from 2010 to 2023 was used. This study used 52 pairs of data with 5 variables. The value of unknown test data is predicted by using k objects from the training data closest to the test. Furthermore, data was distributed, namely 39 training data and 13 test data.

2. Data Normalization

Data normalization is used to obtain relatively uniform data, but still retain the information contained in the original data. The purpose of data normalization is to shape data in value positions with the same range and ensure that data remains consistent. The normalization method used is the min-max method, here is the equation:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Where x_i^, is the normalized data, x_i is the data to be normalized, min(x) is the minimum value and max(x) is the maximum value of the data. The following are the results of data normalization

Table 1. Data Normalization

No	(X_1)	(X_2)	(X_3)	(X_4)
1	0.192	0	0.223	0.281
2	0.234	0.263	0.244	0.920
3	0.257	0.441	0.434	0.682
4	0.342	0.170	0.453	0.962
5	0.058	0.065	0.115	0.423
6	0.114	0.060	0.241	0.223
7	0.199	0.0635	0.324	0.623
8	0.053	0.274	0	0

9	0.009	0.022	0.213	0.191
10	0.147	0.086	0.263	0.795
⋮	⋮	⋮	⋮	⋮
11	0.185	0.129	0.272	0.249
12	0.011	0.271	0.342	0.793
13	1	1	1	1

3. Calculating the Euclidean Distance

Euclidean distance is done to measure the closest distance from one data to another. In addition, euclidean distance is the most commonly used technique in the K-Nearest Neighbor (K-NN) method. The following equation is used in calculating the euclidean distance:

$$d(x_{latih,i}, x_{uji,j}) = \sqrt{\sum_{i,j=1}^{n} (x_{latih,i} - x_{uji,j})^2}$$

Where $d(x_{train,i}, x_{test,j})$ is the euclidean distance, $x_{train,i}$ is the value in the training data, $x_{test,j}$ is the value in the test data, n is the amount of data and i, j is the scale of values 1 to n. Here are the results of the calculation of the Euclidean distance:

Table 2. Euclidean distance

No	1	2	3	4	5	6	⋯	52
1	0	0,693	0,636	0,753	0,233	0,115	⋯	1,665
2	0,693	0	0,354	0,256	0,577	0,735	⋯	1,306
3	0,636	0,354	0	0,399	0,591	0,642	⋯	1,133
4	0,753	0,256	0,399	0	0,703	0,808	⋯	1,192
5	0,233	0,577	0,591	0,703	0	0,243	⋯	1,695
6	0,115	0,735	0,642	0,808	0,243	0	⋯	1,686
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
52	1,665	1,306	1,133	1,192	1,695	1,686	⋯	0

4. Single K-NN Prediction

A single K-NN prediction is performed to find the HDI prediction from the test data based on the closest distance from the training data at a certain number of k. The number of k used in this study is 8, namely k=1,2,3,4,5,6,7 and 8. The following equation is used in calculating a single K-NN.

$$\bar{y}_i = \frac{1}{k} \sum_{i=1}^{k} y_i$$

Where y¯_i is the value of the single K-NN prediction result, k is the range of the data (K-NN parameter) and y_i is the sequence of distances based on the euclidean result. Here are the results of the calculation:

Table 3. Single K-NN Prediction

No	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
1.	66,08	66,42	66,15	66,1875	66,034	66,065	66,01143	65,96625
2.	71,08	70,84	70,733	71,0325	71,082	70,7433	70,84429	71,025
3.	72,55	72,42	72,35	72,0325	71,746	71,77667	71,37571	71,085
4.	71,93	71,69	71,5533	71,435	71,268	70,89833	71,09714	70,77125
5.	66,25	65,985	65,8433	65,9025	66,286	66,17333	66,08286	66,1675
6.	66,76	66,53	66,38	66,34	66,194	66,11	66,04429	65,98125
7.	68,72	68,885	68,7533	68,6275	68,526	68,40833	68,41857	68,4875
8.	66,26	66,04	65,84	65,765	65,856	65,645	65,61429	65,505
9.	65,54	65,14	65,68	65,835	65,586	65,69167	65,74714	65,735
10.	69,05	68,77	68,5533	69,065	68,996	69,3433	69,29	69,16
11.	66,22	66,49	66,42667	66,2325	66,202	66,0733	66,00714	65,94875
12.	68,97	68,725	68,60333	68,715	68,67	68,99167	68,95286	69,21875
13.	82,02	81,86	81,73	79,435	78,006	76,99333	76,20143	75,56125

Based on the table above, the prediction result that is closest to the actual value is k=1. The actual HDI value in the first data is 66.82. However, after being predicted to be close to the actual value, which is 66.76. This is also proven through the MAPE value using the following equation.

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100\%$$

Where A_t is the actual data, F_t is the prediction value, n is the amount of data and t is the value scale from 1 to n. The following are the results of the MAPE calculation from a single K-NN prediction

Table 4. Evaluation of Single K-NN Prediction Results

K-NN	MAPE (%)
$k = 1$	1,041
$k = 2$	1,218
$k = 3$	1,347
$k = 4$	1,527
$k = 5$	1,739
$k = 6$	1,89

$k = 7$	2,012
$k = 8$	2,118

Based on the table above, it can be seen that the smallest MAPE value is found at k=1 with a value of 1.041%. The smaller the MAPE value, the better the result.

5. K-Nearest Neighbor Ensemble Prediction

The ensemble technique used in calculating the final result of the prediction is weighted mean. In the first step, the correlation value between the actual data and the results of the single K-NN prediction will be carried out using the following equation.

$$r_h = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

Where r_h is the correlation value between the actual data and the data of the single K-NN prediction, x_i is the data on the x variable and y_i is the data on the y variable and n is a lot of data. Here are the results of the calculation.

Table 5. Correlation Results

r_h	Value
$r_{k=1}$	176320,942
$r_{k=2}$	172924,338
$r_{k=3}$	169179,5
$r_{k=4}$	113655,505
$r_{k=5}$	86364,9412
$r_{k=6}$	70636,8815
$r_{k=7}$	60446,987
$r_{k=8}$	52238,8227
Jumlah	901767,9174

Next, calculations are carried out on the ensemble weights using the following equation.

$$w_i = \frac{r_h}{\sum_{h=1}^{n} r_h}$$

Where W_i is the ensemble weighting value and h is the value scale 1 to (n). Here are the results of the calculation.

Table 6. Ensemble Weights

w_i	Weight
w_1	0,195528
w_2	0,191761
w_3	0,187609
w_4	0,126036
w_5	0,095773
w_6	0,078332
w_7	0,067032
w_8	0,057929
Jumlah	1

After the calculation of correlation and ensemble weighting, the final result will be predicted using the weighted mean technique through the following equation.

$$\mu_i = \frac{\sum_{i=1}^{n} w_i \bar{y}_i}{\sum_{i=1}^{n} w_i}$$

Where μ_i is the final result of the ensemble prediction, w_i is the ensemble weight, and y¯_i is the single K-NN prediction value. Here are the results of the prediction.

Table 7. Ensemble Prediction Final Result

No	Ensemble Prediction	*Actual Data*
1.	66,15511	66,82
2.	70,91772	71,7

3.	72,12117	73,02
4.	71,48375	72,48
5.	66,06054	66,95
6.	66,39346	67,66
7.	68,66956	69,57
8.	65,90274	67,11
9.	65,56821	66,39
10.	68,94527	69,93
11.	66,26893	66,93
12.	68,80832	69,73
13.	80,06675	82,52

In the table above, it can be seen that the final result of the prediction using the ensemble method with the weighted mean technique (weighted mean) obtained a result, which is an average of 69.03. Here are the results of the calculation:

$$\text{Average} = \frac{\text{Amount of data}}{\text{a lot of data}} = \frac{897{,}3615}{13} = 69{,}02781 \approx 69{,}03$$

Based on the final results of the prediction, the results were evaluated using MAPE to find out how much error level was in predicting the results.

$$\text{MAPE} = \frac{0{,}189137}{13} \times 100\% = 1{,}45\%$$

3.2. The Relationship between HDI and the Unemployment Rate

In finding the relationship between two variables, namely HDI and the unemployment rate, a correlation test was carried out using SPSS software with data taken through the Central Sulawesi Province BPS website, the following data.

Table 8. HDI Data and Unemployment Rate

Year	IPM (x)	Unemployment Rate (y)
2010	63,29	4,61
2011	64,27	3,93

2012	65,00	3,93
2013	65,79	4,27
2014	66,43	3,68
2015	66,76	4,10
2016	67,47	3,29
2017	68,11	3,81
2018	68,88	3,37
2019	69,50	3,11
2020	69,55	3,77
2021	69,79	3,75
2022	70,28	3,00
2023	70,95	2,93
2024	69,03	-

Based on the data in the table above, HDI and the unemployment rate will be illustrated in the form of the following graph:

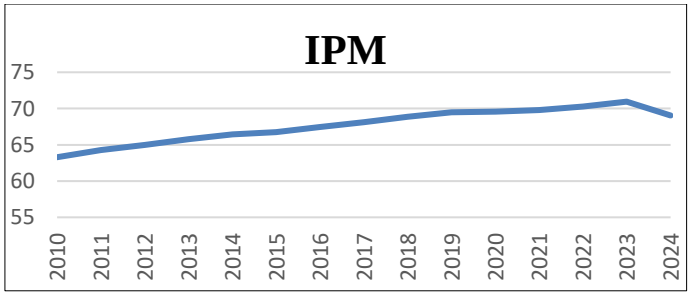


Figure 1. Human Development Index (HDI) graph

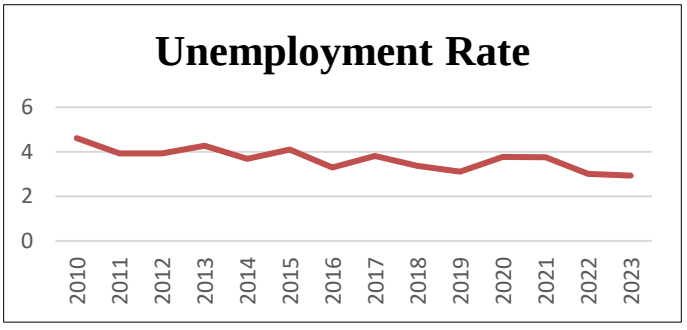
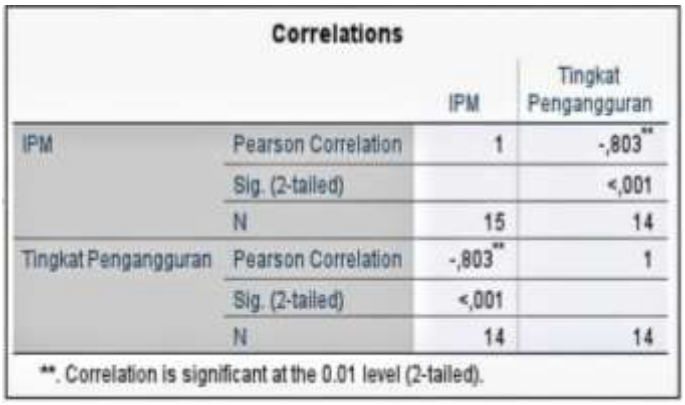


Figure 2. Unemployment Rate Graph

The correlation test was used to find the relationship between two quantitative variables, namely HDI as the x variable and the unemployment rate as the y variable. The correlation test was carried out by inputting data in SPSS software, here are the results:



Correlations

		IPM	Tingkat Pengangguran
IPM	Pearson Correlation	1	-,803**
	Sig. (2-tailed)		<,001
	N	15	14
Tingkat Pengangguran	Pearson Correlation	-,803**	1
	Sig. (2-tailed)	<,001	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 3. Correlation Test Results

From the results of the correlation test, a significant value of less than 0.05 was obtained, namely 0.001, which means that there is a relationship between HDI and the unemployment rate with a correlation value of -0.803. The resulting correlation is negative, meaning that if variable x decreases, then variable y increases, and vice versa. Meanwhile, the correlation is very strong with the result of the correlation being 0.803.

4. Conclusion

Based on the formulation of the problem and research objectives, several key conclusions can be drawn from this study. First, the HDI prediction results for Central Sulawesi Province in 2024, using the Ensemble K-Nearest Neighbor (K-NN) method, show an average value of 69.03. This result indicates that the model provides a reasonable estimate for future development based on the available data.

Second, the Ensemble K-Nearest Neighbor method can be considered effective for this research. This is evidenced by the evaluation of the final results, where the accuracy of the data testing process, measured using the Mean Absolute Percentage Error (MAPE), yielded a value of 1.45%. Since the MAPE is below 10%, it indicates a high level of prediction accuracy, confirming the reliability of the model used in this study.

Lastly, the correlation test conducted between HDI and the unemployment rate reveals a significant relationship between the two variables. The results from the SPSS software output showed a p-value of 0.001, which is below the threshold of 0.05, indicating statistical significance. Additionally, the correlation coefficient of -0.803 suggests a strong negative correlation. This means that when the HDI decreases in a particular year, the unemployment rate tends to increase, and conversely, when the HDI increases, the unemployment rate tends to decrease. The negative sign of the correlation coefficient further supports this inverse relationship, highlighting the strength of the connection between these two variables.

References

Bhirawa, W. T. (2020). The data processing process from the regression equation model using the Statistical Product and Service Solution (SPSS). Statistics, 71–83. http://journal.universitassuryadarma.ac.id/index.php/jmm/article/download/528/494

Devita Rosmadayanti, Niniek Imaningsih, R. S. W. (2021). The Influence of Economic Growth, Regional Original Income, Special Allocation Funds and Regional Expenditure on the Human Development Index in East Java. Paper Knowledge . Toward a Media History of Documents, 3(2), 6.

Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). The application of data mining to predict the sales of best-selling products using the K-Nearest Neighbor method. Building of Informatics, Technology and Science (BITS), 3(4), 639–648. https://doi.org/10.47065/bits.v3i4.1408

Franita, R., & Fuady, A. (2019). Analysis of Unemployment in Indonesia. Journal of Social Sciences, 2, 88–93. http://jurnal.um-tapsel.ac.id/index.php/nusantara/article/view/97/97

Jusman, M., Nur'eni, N., & Handayani, L. (2022). Ensemble K-Nearest Neighbors Method to Predict Composite Stock Price Index (CSPI) in Indonesia. Journal of Mathematics, Statistics and Computing, 18(3), 423–433. https://doi.org/10.20956/j.v18i3.19641

Mahroji, D., & Nurkhasanah, I. (2019). The Effect of the Human Development Index on the Unemployment Rate in Banten Province. Journal

of Economics-Qu, 9(1). https://doi.org/10.35448/jequ.v9i1.5436

Nasution, R. S. (2020). Analysis of the Influence of the Human Development Index (HDI), Number of Creative Economy Workers, and Wages of Creative Economy Workers on Creative Economy Growth in Indonesia. FEB Student Scientific Journal.

Pratiwi, I. A. A. S., & Wijayanto, A. W. (2019). Classification of Human Development Index with K-Nearest Neighbor Method and Support Vector Machine in Java. Journal of Computer Science, 15(1), 8–21. https://ojs.unud.ac.id/index.php/jik/article/download/68565/44248

Satriya, R. H. D., Santoso, E., & Sutrisno, S. (2018). Implementation of the K-Nearest Neighbor Ensemble Method for Prediction of the Rupiah Exchange Rate Against the US Dollar. Journal of Information Technology and Computer Science Development (JPTIIK) Universitas Brawijaya, 2(4), 1718–1725

V. Wiratna Sujarweni, S.E., M,M., M.T. & Lila Retnani Utami, S.E., S.Pd., M.Si., C. (2021). The Master Book of SPSS.

Widaningsih, S. (2019). Comparison of Data Mining Methods for Predicting the Grades and Graduation Time of Informatics Engineering Study Program Students with C4.5 Algorithms, Naïve Bayes, Knn and SVM. Journal of Incentive Technology, 13(1), 16–25. https://doi.org/10.36787/jti.v13i1.78.