Original Research Article

Classification of Java Cities/Regencies Based on Human Development Index Using Discriminant Analysis and Naïve Bayes Classifier

.

ABSTRACT

Aims: This research aims at grouping of cities/regencies on the island of Java, where the central government as well as the most densely populated island in Indonesia, using linear discriminant analysis (LDA) and Naïve Bayes Classifier (NBC).
Study design: Quantitative design.
Place and Duration of Study: Sample: The data used in this study is secondary data from the Indonesian Central Statistics Agency (Badan Pusat Statistik, BPS) regarding the 2022 Human Development Index (HDI) from 119 cities/regencies on the island of Java. The data used are four HDI indicators as independent variables (long and healthy living, knowledge, and the dimensions of decent living standards) and the HDI value as the dependent variable.
Methodology: The grouping was carried out using LDA and NBC. LDA is a type of multivariate analysis used in the dependency method where the relationship between variables can be distinguished between the independent variable and the dependent variable. It aims at obtaining discriminant function equations to group cases into certain groups and to determine differences between groups based on independent variables. Meanwhile, the NBC method is a simple probability-based prediction technique based on the application of Bayes' theorem (Bayes' rule) with a strong assumption of independence.
Results: Both LDA and NBC can be used for prediction and classification. Based on the results of the discriminant analysis, three discriminant functions were formed to group cities/regencies on the island of Java into three HDI groups. In the NBC analysis, the prior probability value for the very high category HDI group was 0.211, the high category HDI group was 0.606, and the medium category HDI group was 0.183. The research results show that LDA is better than the NBC for grouping cities/regencies based on the 2022 HDI indicators with an accuracy rate of 72.92%. Meanwhile, the NBC analysis only provides an accuracy of 64.58%. Three discriminant functions have been obtained to group cities/regencies on the island of Java based on the largest discriminant score where life expectancy makes the largest contribution in distinguishing each group.
Conclusion: As a result, in this case LDA is a better classification method than the NBC. It is also of important to note medium class regions for further actions from stakeholders.

Keywords: Linear Discriminant Analysis, Naïve Bayes Classifier, Human Development Index

1. INTRODUCTION

Development is an effort to optimize existing resources or potential in a planned and sustainable manner with the principle of just and equitable use [3]. The development efforts

carried out to provide welfare to the community equally without exception. Development is considered perfectly successful if all communities can experience the same services and benefits. Achieving the quality of development is not enough if it is only measured based on economic growth to describe the overall level of welfare [9]. For this purpose, development requires other indicators which also play an important role in growth, both from a material perspective and those related to human welfare. Regional development is the process of developing all aspects of life in an area by utilizing its potential and resources to improve community welfare. In general, regional development is related to economic, social and political growth in the area. Regional development requires human resources as the main capital in development, such that the government continues to make efforts for improving the quality of human resources to achieve development success. The measuring tool for regional development success can be seen from the Human Development Index (HDI) value in each region. Indeed, HDI cannot be used as the only measurement of regional development success without other factors that also play a role in regional development.

The United Nations Development Program (UNDP) introduced the concept of human development for the first time in 1990. The Human Development Index (HDI) is a number to measure human development achievements based on the basic components of quality of life which can influence the level of productivity by a person [10]. UNDP determines that there are three dimensions in the formation of HDI, namely: 1) the dimension of long life and healthy living with indicators of life expectancy, 2) the dimension of knowledge with indicators of expected length of schooling and average length of schooling, and 3) the dimension of decent living standards with indicators real expenditure per capita. HDI is used to measure how far human development has been achieved both physically and mentally in an area, in other words humans not only play a role as input but also as output in the development. It is also used to see the efforts and performance results of development programs as a whole in an area. Improving the quality of the population in an area can increase opportunities to participate in sustainable development.

HDI growth in 2020 experienced a slowdown due to the COVID-19 pandemic which occurred over the last two years [1] as per 2023. HDI growth in Indonesia in 2020 slowed down much compared to the previous year, only 0.03 percent. This condition is due to slowing growth in life expectancy and education and real expenditure per capita which has also decreased. The Indonesian government is making every effort to increase the human development index again as an effort to realize regional development success. Indonesia consists of thousands of islands with the central government located on Java Island, of course this presents its own challenges for the government in realizing regional development. Java Island, as the center of government, has the most populous population, 56.10% of the total population of Indonesia, showing that the population distribution in Indonesia is unequal [2]. Therefore, it is necessary to group cities/regencies on the island of Java to support the success of regional development programs so that they are right on target. Previous research on HDI [20] using discriminant analysis explained that life expectancy has a negative correlation with HDI, while expected length of schooling, average length of schooling, and purchasing power parity have a positive correlation with HDI. Similar research was conducted by Hasan [7], where there were four discriminant functions used to group cities/regencies in Indonesia based on ten HDI indicators in 2015 with a success rate of more than 85%. However, research using HDI indicators in 2015 is considered no longer relevant for use in calculating HDI at the moment [2].

Research regarding the comparison between discriminant analysis and Logistic Regression analysis was carried out by Zufa *et al* [21] in predicting the classification of bank health conditions showed that discriminant analysis is better than Logistic Regression analysis with a prediction accuracy of 80%. Furthermore, sensitivity comparison of both discriminant

analysis function and logistic Regression in classifying live birth and stillbirth under varying training and test samples was carried out by Asosega *et al* [22] where the discriminant analysis is preferred. A stepwise discriminant analysis was used for discriminating students in vocational and technical education in Thailand [13]. Meanwhile, customer perceptions of online banking in Indonesia were analyzed using Naïve Bayes algorithm [18]. Furthermore, Sutitis *et. al*. [19] conducted research by comparing the Ordinal Logistic Regression and Naïve Bayes methods for classifying customers' fluency in paying premiums. Based on the accuracy of the classification, the results showed that the best classification is the one that used Naïve Bayes method. Based on previous research, in this study a comparison was made of the classification of cities/regencies on the island of Java based on HDI indicators in 2022 using discriminant analysis and Naïve Bayes Classifier. This research is expected to be able to provide information on a more accurate classification method in grouping cities/regencies on the island of Java based on HDI indicators in 2022. The island of Java was chosen because of the availability of complete data and is the center of economic and human resource development in Indonesia.

2. LITERATURE REVIEW

2.1 Discriminant Analysis

Discriminant analysis is a type of multivariate analysis used in the dependency method where the relationship between variables can be distinguished between the independent variable and the dependent variable. In discriminant analysis, the independent variable has a numerical data scale while the dependent variable has a categorical data scale [16]. Based on the many categories of dependent variables, discriminant analysis is divided into two types: 1) Two-Group Discriminant Analysis, used when the dependent variable consists of two categories, while 2) Multiple Discriminant Analysis is used when the dependent variable has more than two categories [6]. Discriminant analysis aims to obtain discriminant function equations to group cases into certain groups and to determine differences between groups based on independent variables [4]. Apart from that, discriminant analysis also aims at determining which independent variables have the most influence in differentiating between groups and based on the value of these independent variable, respondents will be grouped into certain categories or groups [17]. Discriminant analysis can be carried out after several assumptions are met. Several discriminant analysis assumptions that must be met include: data with a multivariate normal distribution, non-multicollinearity and homogeneity of variance.

The multivariate normality test is carried out by calculating measures of skewness ($b_{1,p}$) and kurtosis ($b_{2,p}$) with the test statistics shown in Equations 2.1 and 2.2 [11].

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{p} \sum_{j=1}^{p} [(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \boldsymbol{S^{-1}} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})]^3 \quad (2.1)$$

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{p} [(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \boldsymbol{S^{-1}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})]^2 \quad (2.2)$$

where n: sample size; $\boldsymbol{x}_i$: vector of i^{th} observation; $\boldsymbol{x}_j$: vector of j^{th}; $\bar{\boldsymbol{x}}$: vector of average; $\boldsymbol{S^{-1}}$: inverse of variance-covariance matrix. The decision making criteria is that null hypothesis (H0) is accepted (multivariate normal distribution data) if $\frac{n}{6} b_{1,p} \leq \chi^2_{\alpha;p(p+1)(p+2)/6}$ or $\frac{b_{2,p} - \boldsymbol{p(p+2)}}{\sqrt{8p(p+2)/n}} \leq Z_\alpha$ or significance value larger than 0.05. The correlation matrix is used to detect multicollinearity problems. Multicollinearity occurs when the correlation coefficient value is between 0.8 to 1 [5]. Furthermore, according to Afifi et al. [14], the Box's M is a test statistic used in testing the homogeneity of the variance-covariance matrix.

The linear discriminant function is used if the assumptions of multivariate normality and homogeneity of the variance-covariance matrix between groups are met. The linear discriminant function aims to separate the population into groups and is used for classification [11]. The linear discriminant function is shown as follows

$$D = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p \quad (2.3)$$

where *D*: discriminant score; *X*: independent variable; *b*: coefficient or discriminant weight. The linear discriminant function to obtain discriminant scores for each group can be formed from Equation 2.4 [8].

$$D_k = \bar{\boldsymbol{x}}_{\boldsymbol{k}}' \boldsymbol{S}_{comb}^{-1} \boldsymbol{x} - \frac{1}{2} \bar{\boldsymbol{x}}_{\boldsymbol{k}}' \boldsymbol{S}_{comb}^{-1} \bar{\boldsymbol{x}}_{\boldsymbol{k}} + \boldsymbol{ln}\, p_k \quad (2.4)$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$: vector of group means at k; $\boldsymbol{S}_{comb}^{-1}$:inverse of combination of variance-covariance matrix for each group; p_k: prior probability at group k. In classification using linear discriminant analysis, an object can be grouped into a category based on the largest discriminant score.
In several independent variables that form the discriminant function, there will be a variable that is the strongest differentiator for classifying objects into certain categories. The strongest differentiating variables contribute to differences between groups. Determining the strongest differentiating variable can be seen from the smallest Wilk's Lambda value [11] which is obtained from Equation 2.5,

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}. \quad (2.5)$$

The value of $\boldsymbol{W}$ and **B** are obtained from $\boldsymbol{W} = \sum_{i=1}^{n_k} \sum_{k=1}^{g} (\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k)'$, and $\boldsymbol{B} = \sum_{k=1}^{g} (\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})'$, where $\boldsymbol{W}$: matrix of sums of squares and products in groups; $\boldsymbol{B}$: matrix of sum of squares and total product; $\boldsymbol{x}_{ik}$: observation vector at i group k; $\bar{\boldsymbol{x}}_k$: vector of group means at k; $\bar{\boldsymbol{x}}$: total mean vector; n_k: the number of group observations, group *k*. By using the *F* test statistical approach, the Wilk's Lambda test statistical value is $F = \left(\frac{\sum n_k - p - 1}{p}\right)\left(\frac{1-\Lambda}{\Lambda}\right)$, with the decision making criteria, namely H_0 is rejected if the independent variable being tested has the smallest Wilk's Lambda value, the test statistic $F > F_{\alpha,2(\sum n_k - p - 1)}$ or significance value < 0.05.

2.2 Naïve Bayes Classifier

The Naïve Bayes Classifier (NBC) method is a simple probability-based prediction technique based on the application of Bayes' theorem (Bayes' rule) with a strong assumption of independence. In Bayes' theorem, let $B_i, i = 1,2,\ldots,p$ be part of the sample space *S* with $P(B_i) \neq 0$ and are mutually exclusive events. Hence, for event *A* where $P(A) \neq 0$, the probability of B_i with condition *A* is as follows $P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{p} P(B_i)P(A|B_i)}$. If $P(A) = \sum_{i=1}^{p} P(B_i)P(A|B_i)$ then we obtain $P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)}$. Naïve Bayes Classifier emphasizes probability estimation with the benefit that the classification will get a smaller error value when the data is large. Naïve Bayes Classifier is based on the simplifying assumption that attribute values are conditionally independent of each other when given values [15]. According to Prasetyo [12], the Naïve Bayes Classifier for classification has the following formula

$$P(Y|X) = \frac{P(Y)\Pi_{i=1}^{p} P(X_i|Y)}{P(X)} \quad (2.6)$$

where: $P(Y|X)$: probability of data with vector *X* for class *Y*; $P(Y)$: initial probability of class *Y* (prior probability), with $Y = Y_k$, k = 1,2, …, *g*; $\Pi_{i=1}^{p} P(X_i|Y)$: independent probability of class *Y* of all observations in vector *X*; $P(X)$: probability of *X*. The probability *P*(*X*) is always constant so that in determining predictions only the maximum value of $P(Y)\Pi_{i=1}^{p} P(X_i|Y)$ is needed by selecting the maximum value as the prediction result class. Meanwhile, the independent probability $\Pi_{i=1}^{p} P(X_i|Y)$ is the influence of all observations from the data on each group *Y* which is denoted in the following equation

$$P(X_i|Y = y) = \Pi_{i=1}^{p} P(X_i|Y = y) \; . \quad (2.7)$$

For observed values of independent variables with a numerical (non-categorical) type, there is certain treatment before carrying out Naïve Bayes Classifier analysis by assuming a certain form of probability distribution for continuous observed values (usually using a Gaussian distribution) and estimating distribution parameters with training data. The Gaussian distribution is characterized by two parameters, namely mean (μ) and variance (σ^2) for each group Y_k, the conditional probability of group Y_k for the observation value X_i in the following equation

$$P(X_i = x_j|Y = y_k) = g(x_j, \mu_{ik}, \sigma_{ik}), \text{ where } g(x_j, \mu_{ik}, \sigma_{ik}) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{(x_j-\mu_{ik})^2}{2\sigma_{ik}^2}},$$

$i = 1,2,\ldots,p$, $j = 1,2,\ldots n,$ and $k = 1,2,\ldots,g$. (2.8)

2.3 Evaluation of Classification Function

Evaluation of the classification function can be done by calculating the accuracy, sensitivity and specificity values. Confusion matrix is a table that can help in evaluating classification functions. The confusion matrix has a main diagonal consisting of actual data that is classified correctly, while on the other diagonal is data that is classified incorrectly. Accuracy is one of the values that can be used to see the accuracy of a classification model. The greater the accuracy value, the better the accuracy of the classification function. Sensitivity is obtained by measuring the proportion of correctly classified data from the entire true classes. Meanwhile, the specificity value is used to measure the model's ability to correctly identify data from non-true categories. The confusion matrix is presented in Table 1.

Table 1 Confusion Matrix

		Prediction class			Total
		I	II	III	
Actual class	I	n_{11}	n_{12}	n_{13}	$n_{1.}$
	II	n_{21}	n_{22}	n_{23}	$n_{2.}$
	III	n_{31}	n_{32}	n_{33}	$n_{3.}$

Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Based on Table 1 Accuracy, sensitivity, specificity calculations can be calculated using the following equation. Accuracy $= \frac{n_{11}+n_{22}+n_{33}}{\text{n}}$; Sensitivity $= \frac{n_{kk}}{n_{k.}}$, for each group *k.* Calculation of specificity values for each group is as follows: Specificity for group 1 $= \frac{n_{22}+n_{23}+n_{32}+n_{33}}{(n_{22}+n_{23}+n_{32}+n_{33})+(n_{21}+n_{31})}$, specificity for group 2 $= \frac{n_{11}+n_{13}+n_{31}+n_{33}}{(n_{11}+n_{13}+n_{31}+n_{33})+(n_{12}+n_{32})}$, and specificity for group 3 $= \frac{n_{11}+n_{12}+n_{21}+n_{22}}{(n_{11}+n_{12}+n_{21}+n_{22})+(n_{13}+n_{23})}$.

3. DATA AND METHODOLOGY

3.1 Data and Research Variables

The data used in this research is secondary data from the Indonesian Central Statistics Agency (Badan Pusat Statistik, BPS) regarding the 2022 Human Development Index (HDI) from 119 cities/regencies on the island of Java. The data used are four HDI indicators as independent variables (X_1, X_2, X_3, X_4) and the HDI value as the dependent variable (Y). The Human Development Index (HDI) is a regional development process that involves humans as both actors and goals in the development process itself. In other words, the goal of development is prosperity that can be felt by residents in the area. HDI is an important indicator to observe development from the human side which can show changes in people's choices in living a decent life [2]. Human development is a process of expanding choices including political freedom, participation in social life, educational choices, survival and health, enjoying a decent standard of living, as well as participating in the community and making decisions that will have an impact on human life itself. A country is categorized developed country if it has a population with good health, intelligent thinking and good purchasing power. However, HDI is not a measure of overall human development and hence other indicators are needed that also play a role in human development. United Nations Development Program (UNDP) determines three dimensions that form HDI, including: Dimensions of Longevity and Healthy Life, Dimensions of Knowledge, and Dimensions of Decent Living Standards.

Table 2 Data Structures for Classification Analysis

City/Regency	**Variable**				
	$\mathbf{X_{1,j}}$	$\mathbf{X_{2,j}}$	$\mathbf{X_{3,j}}$	$\mathbf{X_{4,j}}$	$\mathbf{Y_j}$
City/Regency 1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	Y_1
City/Regency 2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	Y_2
City/Regency 3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	Y_3
⋮	⋮	⋮	⋮	⋮	⋮
City/Regency 119	$X_{1,119}$	$X_{2,119}$	$X_{3,119}$	$X_{4,119}$	Y_{119}

1. Dimensions of Longevity and Healthy Life, measured using the Life Expectancy (LE) indicator at birth, which is the average estimated length of time (in years) that a person can live during their life. In other words, LE is an estimate of the age that someone born at a certain time might reach. Life expectancy in an area describes the health, nutrition and environmental conditions of the area. UNDP states that the standard life expectancy is a minimum of 20 years and a maximum of 85 years.
2. Knowledge Dimension, measured by two indicators in the education sector, namely the Expected Years of Schooling (EYS) indicator and the Average Years of Schooling (AYS) indicator. EYS is the length of school that a 7-year-old child is expected to take in the future. EYS is used to describe people's opportunities to obtain formal education as well as a measure of the success of educational development in the short term. The standard expected length of schooling is a minimum of 0 years and a maximum of 18 years. EYS is obtained from the following calculation [2].

$$EYS_a^t = CF \times \sum_{i=a}^{n} \frac{E_i^t}{P_i^t} \quad (3.1)$$

where EYS_a^t: expected years of schooling at age *a* year *t*; *CF*: correction factor; E_i^t: number of people aged *i* attending school in year *t*; P_i^t: total population age *i* in year *t*. AYS is the average length of school that people aged 25 years and over have taken to undergo formal education [2]. AYS is used to determine the quality of education and measure the success of development in the education sector in the long term. The standard average length of schooling is a minimum of 0 years and a maximum of 15 years. AYS is obtained from the following calculation

$$AYS = \frac{1}{n} \times \sum_{i=1}^{n} x_i \quad (3.2)$$

where AYS: average years of schooling for the population aged 25 years and over; *n*: number of residents aged 25 years and over; x_i: length of school for the *i*-th resident who is 25 years old. Hence, the knowledge dimension value can be obtained with the following calculation

$$I_{knowledge} = \frac{EYS + AYS}{2}. \quad (3.3)$$

3. Dimensions of Decent Living Standards, measured by indicators of real expenditure per capita in a region. UNDP uses Gross National Income (GNI) data to measure people's decent living standards, but these indicators do not reach the regional level. Therefore, an alternative indicator is used, namely adjusted real per capita expenditure or called Adjusted per Capita Expenditure (ACE) which can reach up to the city/regency. This indicator can be used to describe the income and level of welfare of the people in an area [2]. Meanwhile, in calculating the expenditure index, a limit of minimum value is used IDR 1,007,436 and a maximum of IDR 26,572,352 (IDR, Indonesian Rupiah). The HDI is one of the development indicators used to measure the success of developing the quality of life of society. Apart from that, HDI is also used to determine the General Allocation Fund (GAF) in calculating Regional Incentive Funds (RIF). RIF is a funding sourced from the state budget to certain regions based on certain categories which aims to provide awards for improvements and/or achievements of certain performance in the areas of regional financial governance, public government services, basic public services, and community welfare. Based on the achievement status, human development in an area can be grouped into four categories, as the following [2]: 1) Very high: HDI ≥ 80; 2) High: 70 ≤ HDI < 80; 3) Moderate: 60 ≤ HDI < 70; and Low: HDI < 60. Moreover, the variables used in this research are in Table 3.

Table 3 Research Variables

Variable	Variable Names	Unit
X_1	Life Expectancy (LE)	Year
X_2	Expected Years of Schooling (EYS)	Year
X_3	Average Years of Schooling (AYS)	Year
X_4	Adjusted Real Expenditure per Capita (ARCE)	1000 IDR /person /year
Y	Human Development Index (HDI)	-

3.2 Data Analysis Method

Evaluation of the classification function can be done by calculating the accuracy, sensitivity and specificity values. Confusion matrix is a table that can help in evaluating classification.

The research steps are as follows:

- Determining research variables
- Entering standardized data
- Describing research data using descriptive statistical methods
- Carrying out Linear Discriminant classification analysis and Naïve Bayes Classifier
- Comparing the accuracy, sensitivity and specificity values of the two analyses
- Interpretation of the results of the comparison of the Linear Discriminant and Naïve Bayes Classifier methods.

The steps for Linear Discriminant analysis are as follows:

1. Entering standardized data.
2. Dividing the data into training and testing data with a ratio of 60:40.
3. Checking the assumptions of discriminant analysis.
    a. Testing the multivariate normal assumption using skewness and kurtosis test statistics. If the data does not meet the multivariate normal assumptions, data transformation can be carried out.
    b. Testing the non-multicollinearity assumption using Pearson correlation values. If there is an independent variable where multicollinearity is detected then one of the variables will be deleted.
    c. Testing the assumption of homogeneity of the variance-covariance matrix with Box's M test statistics. If the data meets the assumption of homogeneity of the variance-covariance matrix then proceeding using linear discriminant analysis. Furthermore, if the data does not meet the assumption of homogeneity of the variance-covariance matrix then continue using quadratic discriminant analysis.

4. Forming a linear discriminant function and determining the discriminant score.
5. Determining the strongest differentiating variable using Wilk's Lambda test statistics.
6. Forming a confusion matrix table.
7. Calculating accuracy, sensitivity and specificity values.

The steps for classification analysis with the Naïve Bayes Classifier are as follows:

1. Dividing the data into testing and training data with a ratio of 60:40.
2. Determining the prior probability ($P(Y)$) value for each group.
3. Calculating the posterior probability value.
4. Forming a confusion matrix table.
5. Calculating accuracy, sensitivity and specificity values.

The analysis steps were carried out using R and Microsoft Excel software.

4. RESULTS AND DISCUSSION

4.1 Descriptive Statistics

Descriptive statistics are used to describe the characteristics of the data used in this research. Based on the achievement status, HDI in an area can be grouped into four categories, namely very high, high, medium and low HDI. The HDI categories of cities/regencies on the island of Java are presented in Table 4. Based on Table 4, it can be seen that the majority of cities/regencies on the island of Java or as many as 75 cities/ regencies have HDI values in the high category, 70 to 80. The percentage of cities/regencies included in the very high HDI category and the medium HDI category is relatively the same, which is equal to 18.5%. In other words, 22 cities/regencies have HDI values in the very high category as well as the same number 22 cities/regencies have HDI values in the medium category. Meanwhile, in the low HDI category, not a single city/regency on Java Island has an HDI value of less than 60 (low). The distribution of data for each independent variable, namely life expectancy (LE, X_1), expected years of schooling (EYS, X_2), average years of schooling (AYS, X_3), and adjusted real per capita expenditure (ARCE, X4) for each HDI group are presented in Figure 1 and Figure 2

Table. 4 HDI Level of City/Regency in Java Island

No	Count	Precent	HDI Label
1	22	18.49%	Very High
2	75	63.03%	High
3	22	18.49%	Medium
4	0	0.00%	Low

Based on Figure 1, it can be seen that there are two cities/ regencies which are outliers in the life expectancy variable in the very high HDI category, namely Salatiga City and

Semarang City. In the high category HDI group there is one outlier, namely Yogyakarta City, while in the medium HDI category there is no outlier data and the data is symmetrically distributed. Figure 1 regarding the expected length of schooling shows that there is one outlier, namely Cilegon City in the very high HDI category. Meanwhile, in the high and medium HDI categories, no outlier data was found. In Figure 2, the boxplot of average years of schooling and the adjusted real per capita expenditure boxplot shows that in the three HDI groups there were no outlier data found even though the data distribution was not symmetrical.

4.2 Discriminant Analysis Assumption Tests

Evaluation of the classification function can be done by calculating the accuracy, sensitivity and specificity values. Discriminant analysis has three assumptions, the following are the results of testing the assumptions in the case of HDI data analysis:

1. Multivariate Normally Distributed Data, where testing the assumption was carried out by calculating measures of skewness and kurtosis. Based on the results of testing the multivariate normal assumption, it was found that the skewness measure had a *p*-value of 0.3872 and the kurtosis measure had a *p*-value of 0.9868. Both the skewness measure and the kurtosis measure produce a *p*-value greater than 0.05, so the decision to accept the initial hypothesis is obtained. It can be said that with an error rate of 5% the used independent variables fulfill the assumptions of multivariate normal distribution data.

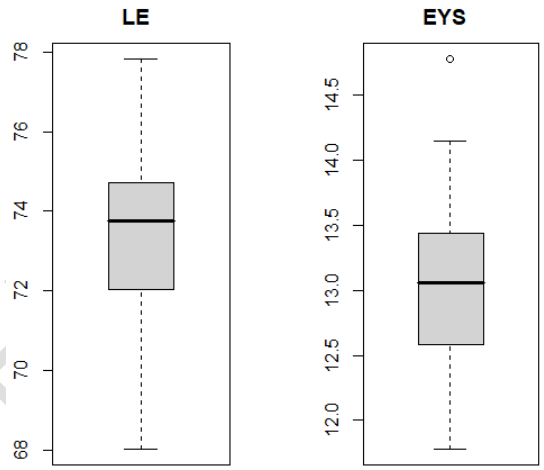

Figure 1 Boxplot of LE and EYS

2. Non-Multicollinearity, where testing the assumption was carried out by looking at the Pearson correlation coefficient value. The correlation matrix resulting for testing the non-multicollinearity assumption is presented in Table 5. Based on the results in this table, the correlation matrix formed does not have multicollinearity problems (the correlation value is no more than 0.8).

3. Homogeneity of Variance-Covariance, where testing the assumption in this study uses the Box's M test statistic with the χ^2 approach. Based on the results of testing the assumption of homogeneity of the variance-covariance matrix with the Box's M test statistic using the χ^2 approach, a test statistical value of 25.907 and a p-value of 0.1689

were obtained, so the decision to accept the initial hypothesis (homogeneous matrix) was obtained. It can be said that with an error rate of 5% the data has a relatively equal variance-covariance matrix, or meets the assumption of homogeneity of the variance-covariance matrix.

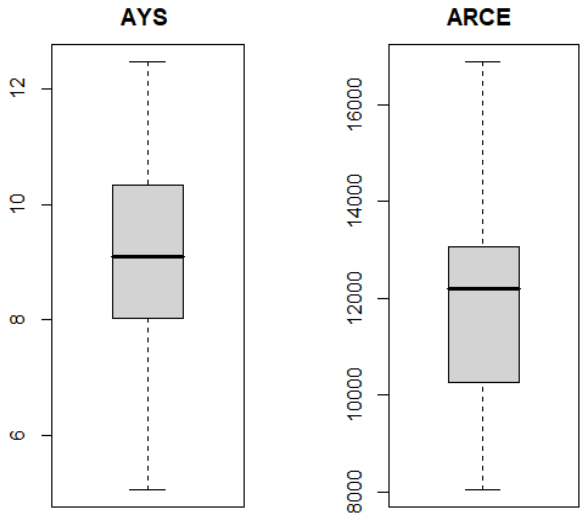

Figure 2 Boxplot of AYS and ARCE

Table 5 Correlation Matrix

	X_1	X_2	X_3	X_4
X_1	1	0.3759	0.2820	0.3429
X_2	0.3759	1	0.5637	0.4704
X_3	0.2820	0.5637	1	0.7276
X_4	0.3429	0.4704	0.7276	1

4.3 Linear Discriminant Analysis

Testing the assumptions of discriminant analysis shows that the data has a multivariate normal distribution and has a homogeneous variance-covariance matrix, so the formation of the discriminant function can be continued using linear discriminant analysis. The linear discriminant function used to classify cities/regencies on the island of Java, three discriminant functions are obtained for getting discriminant scores, namely:

1. The very high category HDI; $D_1 = -1.030 + 0.743X_1 + 0.489X_2 - 0.050X_3 - 0.117X_4$

2. The high category HDI; $D_2 = -0.239 + 0.259X_1 - 0.024X_2 - 0.043X_3 + 0.040X_4$

3. The medium category HDI; $D_3 = -1.793 - 1.858X_1 + 0.324X_2 + 0.394X_3 - 0.378X_4$

The first discriminant function is used to obtain a discriminant score for the very high HDI category group. The formed function shows that the variables of life expectancy (X_1) and expected years of schooling (X_2) make a positive contribution to the discriminant score. The contribution of the two variables to the discriminant score is that every one-year increase in life expectancy can increase the discriminant score by 0.743 if the other variables are constant and a one-year increase in expected length of school can increase the discriminant score by 0.489 if the other variables are constant. Meanwhile, every one-year increase in the average length of school (X_3) can reduce the discriminant score by 0.050 if the other variables are constant and every one thousand rupiah/person/year increase in the adjusted real per capita expenditure variable (X_4) can reduce the discriminant score by 0.117 if the other variables are constant in the first group, namely the very high HDI category. In the second discriminant function, if other variables are constant, every one-year increase in life expectancy can increase the discriminant score by 0.259 and every increase of one thousand rupiah/person/year in adjusted real per capita expenditure can increase the discriminant score by 0.040 in the high category HDI group. Meanwhile, if the other variables are constant, every one-year increase in the expected length of school can reduce the discriminant score by 0.024 and every one-year increase in the average length of school can reduce the discriminant score by 0.043 in the high category HDI group.

In the third group, namely the medium HDI category, the discriminant function formed is inversely proportional to the discriminant function for the second group of high categories of HDI. In the third discriminant function, the expected length of school and average length of school variables has a positive contribution, namely that every one-year increase in the expected length of school can increase the discriminant score by 0.324 and every one-year increase in the average length of school can increase the discriminant score by 0.394 if the variable others are constant values. Meanwhile, if other variables are constant, every one-year increase in life expectancy can reduce the discriminant score by 1.858 and every increase of one thousand rupiah/person/year adjusted real per capita expenditure can reduce the discriminant score by 0.378 in the medium HDI group.
Based on the three discriminant functions formed, a discriminant score can be obtained for each group. The following is an example of manual calculation to obtain the discriminant score for Pacitan Regency with standardized data: Life expectancy (X_1): -0.3236; Expected years of schooling(X_2): -1.9027; Average length of school (X_3): -0.4789; Adjusted real expenditure per capita (X_4): -1.4188.

1. Calculation of discriminant scores for the very high HDI category group

$D_1 = –1.030 +0.743(-0.3236) +0.489(-1.9027) –0.050(-0.4789) –0.117(-1.4188) = – 2.011$

2. Calculation of discriminant scores for the high HDI category group

$D_2 = –0.239 +0.259(-0.3236) –0.024(-1.9027) –0.043(-0.4789) +0.040(-1.4188) = – 0.313$

3. Calculation of discriminant scores for the medium HDI category group

$D_3 = –1.793 –1.858(-0.3236) +0.324(-1.9027) +0.394(-0.4789) –0.378(-1.4188) = – 1.460$

After carrying out calculations, the largest discriminant score was obtained in the second group, which is - 0.313. So Pacitan Regency can be classified in the second group, namely the high HDI category. The results of linear discriminant classification into certain groups based on the largest discriminant score are shown in Table 6. The results of the classification of cities/regencies on the island of Java using Linear Discriminant can be seen in Table 6.

Table 6 Classification Result of Discriminant Analysis

City/ Regency	Y	D_1	…	D_3	Y pred
Pacitan	2	-2.0111	…	-1.4597	2
Ponorogo	2	-1.0317	…	-1.8385	2
Tulungagung	2	-0.6126	…	-2.9066	2
Blitar	2	-0.7934	…	-2.2146	2
Banyuwangi	2	-1.9814	…	-0.1612	2
⋮	⋮	⋮	…	⋮	⋮
Serang City	3	-2.1525	…	3.4330	3
Cilegon City	2	-3.0184	…	2.1319	3

Table 7 Classification of City/Regency with Linear Discriminant

Group	City/Regency
2 (High HDI)	Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Banyuwangi, Nganjuk, Madiun, Magetan, Lamongan, Sumenep, Probolinggo City, Mojokerto City, Surabaya City, Banyumas, Purworejo, Wonosobo, Magelang, Blora, Pati, Jepara, Demak, Temanggung, Pekalongan, Pemalang, Tegal, Magelang City, Salatiga City, Semarang City, Tegal City, Sleman, Bandung, Ciamis, Cirebon, Majalengka, Sumedang, Bekasi, Pengandaran, Cimahi City, Tasikmalaya City, West Jakarta City, North Jakarta City
3 (Medium HDI)	Bondowoso, Situbondo, Pamekasan, Lebak, Serang, Cilegon City

The linear discriminant function formed in this research consists of four independent variables, including life expectancy (X_1), expected length of school (X_2), average length of school (X_3), and adjusted real per capita expenditure (X_4). Among the four independent variables, there is the strongest differentiating variable in classifying cities/regencies on the island of Java based on the HDI value. Determining the strongest differentiating variable can be seen from the smallest Wilk's Lambda value, the largest F value, and a *p*-value of less than 0.05.

Based on Table 8, the results show that the strongest differentiating independent variable in classifying cities/regencies on the island of Java is life expectancy (X_1). The life expectancy variable is the most influential independent variable in determining HDI groups because in society life expectancy is believed to be a valuable factor. Life expectancy is closely related to adequate nutrition and good public health. The higher the community's health, the greater the community's chances of surviving longer. Apart from that, a good diet can influence nutritional intake in a person's body. Regardless of fate, humans have the opportunity to live longer by paying attention to the body's nutritional intake. This shows that the higher the life expectancy in a city/regency can influence the HDI value of that city/ regency. Meanwhile, the variables expected length of schooling (X_2), average length of schooling (X_3), and adjusted real expenditure per capita (X_4) have no greater influence in determining the HDI value in an area. However, the expected length of schooling and the average length of schooling are variables that can describe the availability of human capital in a region. Quality human resources can be seen from the level of education that has been completed. Likewise, the adjusted real per capita expenditure variable also has an influence on the HDI value of a city/regency, although not as much as the life expectancy variable.

Table 8 Strength of Differencing Variable

Variable	Wilk's Lambda	F Test Statistic	*p-value*
X_1	0.6356	33.2512	< 0.0001
X_2	0.9358	3.9797	0.0213
X_3	0.9638	2.1809	0.1176
X_4	0.9583	2.5229	0.0846

4.4 Evaluation of Classification Function of Linear Discriminant

After carrying out classification using linear discriminants, the classification function is then evaluated based on the confusion matrix. Based on the discriminant confusion matrix in Table 9, there are 48 cities/regencies on the island of Java that are correctly classified into certain HDI categories. A total of 30 cities/regencies are appropriately classified in group 2 (high HDI category) and 5 cities/regencies are appropriately classified in group 3 (medium HDI category). However, not a single city/regency on the island of Java is correctly classified in group 1 (very high HDI category). Most of the cities/regencies on the island of Java are included in the group of cities/regencies with high category HDI values, ranging from 70 to 80. Evaluation of the accuracy of the classification function is carried out by calculating the accuracy, sensitivity and specificity values based on Table 9.

a. Accuracy, where the value is used to perceive the accuracy of the model in correctly classifying cities/regencies on the island of Java into certain HDI categories. After forming the discriminant function, the accuracy value was calculated and an accuracy value of 72.92% was obtained. This means that by using linear discriminant analysis, 72.92% of cities/regencies on the island of Java are correctly classified into certain HDI categories.

b. Sensitivity, it can show the ability of the model formed to correctly detect the classification of cities/regencies on the island of Java from all cities/regencies that originate from that

group. In Table 9, the sensitivity value for the very high category of HDI is 0, the high category of HDI is 0.9677, and the medium category of HDI is 0.5556. Based on the three sensitivity values for each group, an average sensitivity value of 0.5078 was obtained. In other words, 50.78% or as many as 60 cities/regencies were able to be classified correctly in a certain category of HDI group compared to all cities/regencies on the island of Java in that group.

Table 9 Confusion Matrix of Linear Discriminant

		Prediction Class			**Total**
		1	**2**	**3**	
Actual Class	**1**	0	8	0	8
	2	0	30	1	31
	3	0	4	5	9
Total		0	42	6	48

Table 10 Sensitivity and Specificity Values of Linear Discriminant Linear

Group	**Sensitivity**	**Specificity**
1 (Very High HDI Category)	0.0000	1.0000
2 (High HDI Category)	0.9677	0.2941
3 (Medium HDI Category)	0.5556	0.9744

c. Specificity, is used to see the goodness of the model in correctly identifying cities/regencies on the island of Java that are not actually included in a certain category of HDI group. So, three specificity values are obtained which are presented in Table 10. Based on Table 10, the specificity value for the very high category HDI group is 1, the high category HDI group is 0.2941, and the medium category HDI group is 0.9744. The average specificity value for the three groups was 0.7562 or 75.62%. So, of the 119 cities/regencies on the island of Java, the model was able to detect 90 cities/regencies of which actually did not fall into a certain category of HDI group.

4.5 Analysis and Classification of Naïve Bayes Classifier

Naïve Bayes Classifier analysis was carried out using RStudio software. The first step in the Naïve Bayes Classifier analysis is to determine the initial probability (prior). Based on the results of the analysis, the prior values obtained for the three groups use the training data presented in Table 11. The prior probability value for the first group is 0.2113. A total of 15 cities/regencies out of 71 cities/regencies on the island of Java have HDI values in the very high category. In the second group, a prior probability value of 0.6056 was obtained or there

were 43 cities/regencies out of 71 cities/regencies in the training data that had high category HDI values. Meanwhile, in the third group, the prior probability value obtained was 0.1831. This shows that of the 71 cities/regencies in the training data, there are 13 cities/regencies that have HDI values in the medium category. Next, the prior probability value for each group will be used to determine the posterior probability value. Classification with the Naïve Bayes Classifier is carried out by looking at the posterior probability value. After carrying out the Naïve Bayes Classifier analysis, the posterior probability values were obtained for each group. Before determining the posterior probability value, information on the average and standard deviation for each variable in each group is required, which is presented in Table 11.

Table 11 Prior Probability Values

Group	Prior Probability
1 (Very High HDI Category)	0.21127
2 (High HDI Category)	0.60563
3 (Medium HDI Category)	0.18310

Table 12 Mean and Standard Deviation for Each Variable

Group	Independent Variable	Mean	Standard Dev.
1	X_1	0.58162	0.48747
	X_2	0.66228	0.93794
	X_3	0.31900	0.98566
	X_4	0.25022	1.19444
2	X_1	0.16038	0.90670
	X_2	0.04201	1.07789
	X_3	0.02891	0.97986
	X_4	0.05739	0.94212
3	X_1	-1.10451	0.71110
	X_2	-0.12564	1.05101
	X_3	-0.09281	1.25605
	X_4	-0.36133	0.97876

The average value and standard deviation of each independent variable for each HDI group are used to determine conditional probabilities based on a Gaussian distribution. Cities/regencies on the island of Java will be included in a certain category of HDI group based on the largest posterior probability value. The posterior probability value for each HDI group is obtained from the following equation.

1. The group for the very high HDI category

$$P(Y_1|X_1, X_2, X_3, X_4) = 0.211 \times \Pi_{i=1}^{4} P(X_i|Y_1)$$

2. The group for the high category of HDI

$$P(Y_2|X_1, X_2, X_3, X_4) = 0.606 \times \Pi_{i=1}^{4} P(X_i|Y_2)$$

3. The group for the medium category of HDI

$$P(Y_3|X_1, X_2, X_3, X_4) = 0.183 \times \Pi_{i=1}^{4} P(X_i|Y_3)$$

The following is an example of a manual calculation to obtain the posterior probability for Pacitan Regency with standardized data: Life expectancy (X_1): -0.3236; Expected years of schooling (X_2): -1.9027; Average length of school (X_3): -0.4789; and adjusted real expenditure per capita (X_4): -1.4188.

1. Calculation of posterior probability for the first group

$P(Y_1|X_1, X_2, X_3, X_4)$ =0.211× $P(X_1 = -0.3236|Y_1) \times P(X_2 = -1.9027|Y_1) \times P(X_3 = -0.4789|Y_1) \times P(X_4 = -1.4188|Y_1)$

= 0.211 × 0.1019 × 0.0098 × 0.289 × 0.1375 = 0.0089

2. Calculation of posterior probability for the second group

$P(Y_2|X_1, X_2, X_3, X_4)$

=0.606× $P(X_1 = -0.3236|Y_2) \times P(X_2 = -1.9027|Y_2) \times P(X_3 = -0.4789|Y_2) \times P(X_4 = -1.4188|Y_2)$=0.606 × 0.3634 × 0.0755 × 0.3525 × 0.1205 = 0.6709

3. Calculation of posterior probability for the third group

$P(Y_3|X_1, X_2, X_3, X_4)$

=0.183× $P(X_1 = -0.3236|Y_3) \times P(X_2 = -1.9027|Y_3) \times P(X_3 = -0.4789|Y_3) \times P(X_4 = -1.4188|Y_3)$=0.183 × 0.2589 × 0.0932 × 0.3396 × 0.2250 = 0.3211

After carrying out the calculations, the largest posterior probability was obtained in the second group, which is 0.6709. Such that Pacitan Regency can be classified in the second group, namely the high HDI category. The results of the posterior calculation as well as the classification prediction for cities/regencies on the island of Java using the Naïve Bayes Classifier are presented in Table 13. The results of the classification of cities/regencies on the island of Java using the Naïve Bayes Classifier can be seen in Table 14.

Table 13 NBC Calculation of Posterior and Classification Prediction

City/Regency	Y	Posterior 1	…	Posterior 3	Y Pred
Pacitan	2	0.0010	…	0.3180	2
Ponorogo	2	0.3190	…	0.0339	2
Tulungagung	2	0.6580	…	0.0036	1
Blitar	2	0.3040	…	0.0247	2
Banyuwangi	2	0.0040	…	0.5610	3
⋮	⋮	⋮	…	⋮	⋮
Serang City	2	<0.001	…	0.8960	3
Cilegon City	1	<0.001	…	0.6780	3

Table 14 Classification Result with Naïve Bayes Classifier

Group	City/Regency
1 (Very hight HDI)	Tulungangung, Mojokerto City, Pemalang, Tegal City
2 (High HDI)	Pacitan, Ponorogo, Trenggalek, Blitar, Nganjuk, Magetan, Lamongan, Wonosobo, Magelang, Blora, Pati, Jepara, Demak, Temanggung, Pekalongan, Tegal, Magelang City, Salatiga City, Semarang City, Sleman, Ciamis, Cirebon, Majalengka, Sumedang, Bekasi, Pangandaran, Cimahi City, Tasikmalaya City, West Jakarta City, North Jakarta City
3 (Medium HDI)	Banyuwangi, Bondowoso, Situbondo, Madiun, Pamekasan, Probolinggo City, Lebak, Serang, Cilegon City

4.6 Classification Evaluation of Naïve Bayes Classifier

Based on the classification results in Table 15, a confusion matrix can be formed which is shown in Table 16. There are 31 cities/regencies on the island of Java that are correctly classified into certain HDI categories. There is 1 city/regency that is correctly classified in group 1 (very high HDI category), 25 cities/regencies that are correctly classified in group 2 (high HDI category), and 5 cities/regencies that are correctly classified in group 3 (medium HDI category). It can be seen that the majority of cities/regencies on the island of Java have

high HDI values, ranging from 70 to 80. Evaluation of the accuracy of the classification function is carried out by calculating the accuracy, sensitivity and specification values.

Table 15 Confusion Matrix Naïve Bayes Classifier

		Prediction Class			**Total**
		1	**2**	**3**	
Actual Class	**1**	1	7	0	8
	2	2	25	4	31
	3	1	3	5	9
Total		4	35	9	48

a. Accuracy. In the Naïve Bayes Classifier analysis, the accuracy value obtained was 64.58%. This shows that by using Naïve Bayes Classifier analysis 64.58% or as many as 77 cities/regencies on the island of Java were correctly classified into certain HDI categories.

b. Sensitivity. The sensitivity value of the Naïve Bayes Classifier analysis is presented in Table 15. In this table, the sensitivity value for the very high category HDI group is 0.1250, the high category HDI is 0.8065, and the medium category HDI group is 0.5556. Based on the three sensitivity values for each group, an average sensitivity value of 0.4975 was obtained. This shows that 49.75% or as many as 59 cities/regencies on the island of Java are classified correctly in a certain category of HDI group compared to all cities/regencies on the island of Java in that group.

c. Specificity. The specificity values obtained in the Naïve Bayes Classifier analysis are presented in Table 16. Based on this table, the specificity value for the very high category HDI group is 0.9250, the high category HDI group is 0.4118, and the medium category HDI group is 0.8974. The average specificity value for the three groups was 0.7447 or 74.47%. Thus, of the 119 cities/regencies on the island of Java, the model was able to detect 88 cities/regencies of which actually did not fall into a certain HDI category.

Table 16 Sensitivity and Specificity Values of Naïve Bayes Classifier

Group	**Sensitivity**	**Specificity**
1 (Very High HDI Category)	0.1250	0.9250
2 (High HDI Category)	0.8065	0.4118
3 (Medium HDI Category)	0.5556	0.8974

4.7 Classification Comparison

After carrying out linear discriminant analysis and Naïve Bayes Classifier and evaluating the classification of each analysis by calculating the accuracy, sensitivity and specificity values, we then compare the classification results between the two analyzes to obtain the better analysis for classifying cities/regencies on the island of Java into certain HDI categories. The better classification analysis can be seen from the highest accuracy, sensitivity and specificity values. A comparison of the classification results is shown in Table 17.

Table 17 Classification Result Comparison

Evaluation Result	Linear Discriminant	Naïve Bayes Classifier
Accuracy	72.92%	64.58%
Sensitivity	50.78%	49.57%
Specificity	75.62%	74.47%

Based on Table 17, it can be seen that linear discriminant analysis has higher accuracy, sensitivity and specificity values than Naïve Bayes Classifier analysis. Linear discriminant has an accuracy value of 72.92%, which shows that linear discriminant analysis is superior in classifying cities/regencies on the island of Java precisely in certain HDI category groups than Naïve Bayes Classifier analysis which has a lower accuracy value of 64.58%. If we look at the sensitivity value, linear discriminant analysis also has a higher value than Naïve Bayes Classifier analysis, which is 50.78%. This means that linear discriminant analysis is able to classify 50.78% of cities/regencies on Java Island into certain HDI categories compared to all cities/regencies in that group. Meanwhile, in the Naïve Bayes Classifier analysis, it was only able to classify 49.57% of cities/regencies on Java Island compared to all cities/regencies in that group. Likewise, the specificity value of linear discriminant analysis has a higher value compared to the Naïve Bayes Classifier analysis, 75.62%. Thus, of all cities/regencies on the island of Java, linear discriminant analysis was able to detect 75.62% of the cities/regencies which actually did not fall into a certain category of HDI group. Meanwhile, the Naïve Bayes Classifier analysis is only able to detect cities/regencies that are not actually included in a certain HDI category, amounting to 74.44% of all cities/regencies on the island of Java. The best analysis in carrying out classification is the analysis that has higher accuracy, sensitivity and specificity values.

Therefore, in this study, linear discriminant analysis is a better analysis than the Naïve Bayes Classifier in classifying cities/regencies on the island of Java based on the 2022 human development index indicators. This is because linear discriminant analysis has higher accuracy, sensitivity and specificity values which are higher than the Naïve Bayes Classifier analysis so that it can classify cities/regencies on the island of Java more precisely and accurately. It is also worth to be noted that LDA is sensitive to outlier while NBC is sensitive to imbalance data [16]. Fortunately, in this study these two phenomena do not exist and the final interpretation can be carried out without caveats.

5. CONCLUSION AND SUGGESTION

Based on the analysis and discussion that has been carried out, the conclusions are as the following: 1) Three discriminant functions have been obtained to group cities/regencies on

the island of Java based on the largest discriminant score where life expectancy makes the largest contribution in distinguishing each group, 2) From the results of the Naïve Bayes Classifier analysis, a prior probability value for each group is provided which is then used to obtain a posterior probability value where cities/regencies will be grouped based on the largest posterior probability value, 3) Linear discriminant analysis is a better classification method than the Naïve Bayes Classifier.

We also provide the suggestion, particularly to policymakers, practitioners, as well as the academicians, as the following: 1) Areas of medium class (Bondowoso and its nearby, as well as Lebak and its nearby) should be paid attention; 2) Other independent variables, such as income, health facility, and education facility should be considered to impact on HDI values; 3) Simulation studies to compare classification methods would be beneficial to have better method for classification on some cases and scenarios.

ETHICAL APPROVAL

All authors hereby declare that all experiments have been examined and approved by the appropriate ethics committee and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

REFERENCES

1. Badan Pusat Statistik. 2020. Indeks Pembangunan Manusia 2020. Badan Pusat Statistik (BPS), Jakarta.
2. Badan Pusat Statistik. 2022. Indeks Pembangunan Manusia 2022. Badan Pusat Statistik (BPS), Jakarta.
3. Effendi, B. 2002. Pembangunan Daerah Otonom Berkeadilan. Yogyakarta: Uhaindo dan Offset.
4. Garson, G.D. 2012. Discriminant Function Analysis. Statistical Associates Publishing. USA.
5. Gujarati, D.N. and Porter, D.C. 2015. Basic Econometrics, fourth edition. The Mcgraw-Hilol, Singapura.
6. Hair, J.F.Jr., Black, W.C., Babin, B.J., and Anderson, R.F. 2010. Multivariate Data Analysis, Seventh Edition. New Jersey: Pearson Prentice Hall.
7. Hasan, A. 2021. Analisis Diskriminan dalam Menentukan Fungsi Pengelompokan Kabupaten/Kota di Indonesia berdasarkan Indikator Indeks Pembangunan Manusia. Jurnal Ekonomi dan Manajemen Teknologi Vol, 5(1).
8. Johnson, R.A. and Wichern, D.W. 2007. Applied Multivariate Statistical Analysis. Pearson Prentice Hall. New Jersey.
9. Koengkan, M., Fuinhas, J. A., and Santiago, R. 2020. The relationship between CO2 emissions, renewable and nonrenewable energy consumption, economic growth, and urbanisation in the Southern Common Market. Journal of Environmental Economics and Policy, 1– 19. doi:10.1080/21606544.2019.1702 902.
10. Mahroji, D., and Nurkhasanah, I. 2019. Pengaruh Indeks Pembangunan Manusia Terhadap Tingkat Pengangguran di Provinsi Banten. Jurnal Ekonomi-Qu, 9(1).
11. Mattjik, A.A. and Sumertajaya, I.M. 2011. Sidik Peubah Ganda. IPB Press. Bogor.
12. Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: ANDI Yogyakarta
13. Yoosomboon, S., Amornkitpinyo, T., Sopapradit, S., Amornkitpinyo, P., and Kinhom, R. 2021. A Discriminant Analysis of Actual Use of Cloud Technology for Vocational and Technical Education. Journal of Theoretical and Applied Information Technology, 99(21), p.5058--5068.

14. Afifi, A., May, S., Donatello, R.A., Clark, V.A. 2020. Practical Multivariate Analysis, Sixth Edition. CRC Press.
15. Santoso, H. 2017. Data Mining Penyusunan Buku Perpustakaan Daerah Lombok Barat Menggunakan Algoritma Apriori. Seminar Nasional TIK dan Ilmu Sosial.
16. Astutik, S., Solimun, S., Darmanto, D. 2018. Analisis Multivariate: Teori dan Aplikasinya dengan SAS. UB Press.
17. Husson, A., Lê, S., Pagès, J. 2017. Exploratory Multivariate Analysis by Example using R. CRC Press.
18. Karmagatri, M., Aziz, C.F.A., Asih, W.R.P, and Jumri, I.A. 2023. Uncovering User Perceptions Toward Digital Bank in Indonesia: A Naïve Bayes Sentiment Analysis of Twitter Data. Journal of Theoretical and Applied Information Technology, 101(12), p.4960--4968.
19. Sutitis, R., Suparti, S., and Ispriyanti, D. 2015. Klasifikasi Tingkat Kelancaran Nasabah Dalam Membayar Premi Dengan Menggunakan Metode Regresi Logistik Ordinal Dan Naïve Bayes (Studi Kasus Pada Asuransi Ajb Bumiputera Tanjung Karang Lampung). Jurnal Gaussian, 4(3), 651-659.
20. Tiro, M. A., and Ahmar, A. S. 2020. Metode Analisis Diskriminan dalam Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Indeks Pembangunan Manusia. VARIANSI: Journal of Statistics and Its application on Teaching and Research, 2(1), 40-45.
21. Zufa, F., Nugroho, S., and Simanihuruk, M. 2017. Perbandingan Analisis Diskriminan dan Analisis Regresi Logistik Ordinal dalam Prediksi Klasifikasi Kondisi Kesehatan Bank. Jurnal Matematika Vol, 7(2), 92-106.
22. Asosega, K.A., Opoku-Ameyaw, K., Otoo, D., Mac-Ocloo, M.K., and Ayinzoya, R. 2020. Comparative Sensitivity Performance of the Discriminant Function and Logistic Regression under Different Training and Test Samples for Predicting Birth Outcomes. Asian Journal of Probability and Statistics, 6(3), 47-60.