Comparative Analysis of Classification Algorithms for Citizens Welfare Status using PCA as Feature Selection

Erfin Nur Rohma Khakim
Master of Information Technology, Postgraduate Program, Universitas Teknologi Yogyakarta
Sleman, Indonesia

Erik Iman Heri Ujianto
Master of Information Technology, Postgraduate Program, Universitas Teknologi Yogyakarta
Sleman, Indonesia

ABSTRACT

The government has launched various programs to improve the welfare of citizens in order to solve the problem of poverty. The problem in poverty alleviation is on its databases. Classification of the level of welfare conventionally with the estimation method causes the classification results to be invalid. In addition, many poor people who should be the target recipients of poverty alleviation programs have yet to be recorded. This study proposes a machine learning data mining method to classify the welfare of citizens so that the results of the category of welfare levels are more computable and valid. The proposed algorithms are Naïve Bayes, Decision Tree and K-Nearest Neighbor (K-NN) and using Principal Component Analysis (PCA) as feature selection and normalization method on the preprocessing. The data that used in this research is Data Indikator Kesejahteraan Sosial (IKS). IKS data is data collected from residents of Bantul Regency in 2022. The IKS data currently consists of 95,347 rows and uses 27 attributes. There are 4 (four) class or label in this dataset include very poor, poor, nearly poor and not poor. The results of the test show that generally the best algorithm performance is K-NN with accuracy, precision and recall values respectively 96.71%, 95.16% and 88.79%. In this study, using PCA and the normalization method also had a significant effect on improving the performance of the classification algorithm. For further research, it is expected to be able to use deep learning algorithms in classifying because it has large data dimensions.

Keywords

Classification; feature selection; welfare; poverty

1. INTRODUCTION

The availability of data that abundant resulting from the use of information technology in almost all areas of life raises the need to be able to utilize the information and knowledge contained in the overflow of data, which named to data mining [1]. There are various techniques in data mining that are used to extract information from one or more data. These techniques include classification, clustering, association and prediction [2]. Economic inequality in each region triggers the existence of pre-prosperous and prosperous citizens. In general, welfare can be measured in terms of demography, food adequacy, education, health, employment, and environmental conditions [3]. The impact of this low welfare, among other things, can result in many children not experiencing quality education, not being able to pay for health expenses, the amount of savings that is still very minimal, access to public services cannot be fulfilled, the lack of social security and protection for families, and increasingly high degree of urbanization to the city [4]. This low level of welfare also results in the emergence of criminality from a conflict perspective [5].

Bantul Regency is also a district that cannot escape the problem of poverty. According to the Central Statistics Agency for Bantul Regency, the poverty rate for 2021 is 14.05%, has increased 0.5% from 2020 [6].

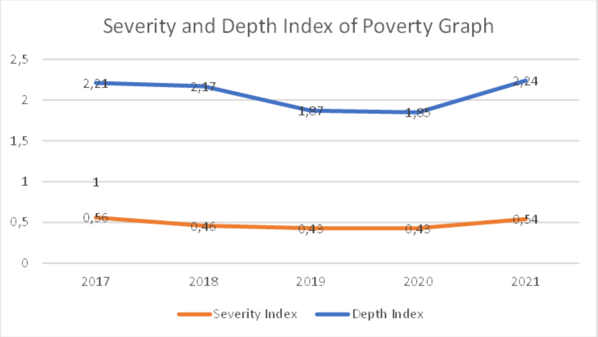

Fig 1: Severity and Depth Index of Poverty Graph [7]

From Figure 1, it can be seen from year to year that there is no significant change in the index of severity and depth of poverty in Bantul Regency. The graph in figure 1 above also shows that in recent years the level of poverty inequality in Bantul Regency is still high.

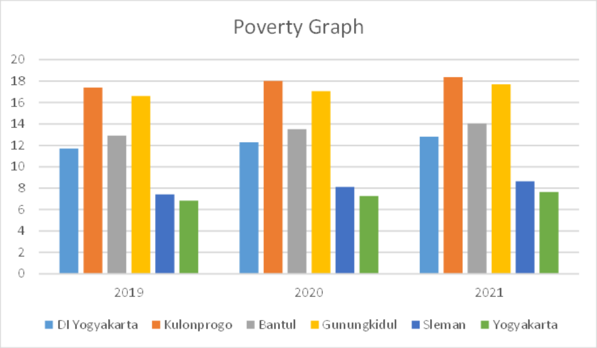

Fig 2: Poverty Graph [7]

Figure 2 shows that the poverty rate in Bantul Regency is still above 10%. It is very different from the city of Yogyakarta and Sleman Regency, which have quite good poverty alleviation programs. From this reality, both the central government and regional governments have made efforts to reduce the poverty rate with various poverty alleviation programs through social assistance. The Indonesian government has taken various ways to reduce poverty or increase welfare. New problems emerged after the social assistance beneficiary data was determined by the Ministry of Social Affairs and the Regional Government.

The problem in poverty alleviation is on its database. Classification of the level of welfare that is done conventionally with the estimation method causes the classification results to be invalid. In addition, many poor people in Bantul Regency who should be the target recipients of the poverty alleviation program have yet to be recorded. This is due to the welfare database which cannot cover all residents in Bantul Regency. The data collection and classification process is long and requires large costs, so the welfare database is still incomplete. Poverty alleviation programs that have been launched by the government are unable to target the poor who have not been included in the data.

Previous research that has been done is research that classifies based on data on the poor population obtained from the Tibawa District using data mining techniques, namely Naïve Bayes. Attributes that will be used in classifying the population include Age, Education, Occupation, Income, Dependents, and Marital/Never Married Status [8]. The weakness of that research is that the dataset used still has a few row data with a small number of attributes. In addition, the accuracy value is below 80% because it has not used feature selection.

The aim of this research is to classify residents in Bantul Regency who have not been included in welfare data. In this study, classification data mining techniques will be used to model existing welfare data, then classify based on the model that has been made and test the classification results. To get the best results, it will compares three classification methods, namely Naïve Bayes, Decision Tree, and k-Nearest Neighbor. Tests are carried out to measure the highest level of performance of several data mining methods that are compared.

2. RESEARCH METHODS

2.1 Research Authenticity

Prior to this research, there have been many other studies that have been done. The first is a research that presents two supervised multi classified machine learning models to predict the poverty status of households in Costa Rica as a way to support government and business sectors in making decisions in a rapidly evolving social and economic environment [9]. The second research is research that focuses on the problem of multidimensional poverty in Jordan where there is a survey of household expenditure and income which can provide data used to identify and measure the status of household poverty over time [10]. The third research is research that identifies poor households as potential beneficiaries of poverty alleviation programs. This identification is carried out using a machine learning algorithm, especially classification [11]. The fourth research revealed that the Bayesian Network Classifiers model offers an adequate alternative to address policy challenges in measuring vulnerability to multidimensional poverty [12]. The fifth research also aims to classify the poverty line according to district/city which was conducted in North Sumatra Province to determine the poverty line [13]. Welfare is closely related to poverty alleviation programs. One of the most popular poverty alleviation programs is the provision of social assistance. As in research which states that the provision of precise and focused poverty data is an important component of a poverty reduction strategy [14].

Subsequent research is the process of classifying citizens who are entitled to receive assistance is still carried out manually and is considered to be less accurate in obtaining the results of social assistance recipients [15]. Apart from being in the form of social assistance, welfare improvement programs can also take the form of financial assistance for children's schools as discussed in research which aims to create an application program that is capable of conducting data analysis in a school to classify students who are eligible to receive an Indonesian Smart Card [16]. Another way of welfare improvement program is related to assistance for the construction of uninhabitable houses as revealed in research with the aim of helping the problem with the difficulty of determining recipients of housing repair assistance based on predetermined criteria [17]. In addition to the classification of welfare data, there is also research on the classification of risks arising from a low level of welfare. This research’s aims for the most efficient classification [18].

Subsequent research is the classification of the poor using the naïve Bayes algorithm in the people of Tibawa District [8]. Next research related to the prediction of home improvement recipients using the naïve Bayes method on data on home improvement recipients in Bali [17]. Another research that discusses the classification of poor families is research that uses naïve Bayes validations 2 and 3 in cases of families in Banjar Regency [19]. In addition, research discussing the eligibility of PKH recipients uses C4.5 and naïve Bayes in Banjar District [20]. Another research on poverty classification using the QUEST algorithm on households in Semarang City [21]. Finally, there is research to classify the poor with three classification algorithms with the result that the decision tree is the best algorithm for classifying the poor [22].

2.2 Theoritical Basic

Welfare is a cycle that include changing in several basic aspects of human life that do not increase towards a better condition in society, lifestyle and social relations [23]. Welfare is inseparable from the problem of poverty. Poverty itself is a situation where there is a shortage of things that are usually owned, such as food, clothing, shelter and drinking water [24]. Types of poverty are divided into six namely absolute, subjective, relative, natural, cultural and structural poverty [25].

Data mining is a process of analyzing hidden data patterns according to various perspectives for categorization into useful information, collected in common areas, data warehouses for efficient analysis, data mining algorithms, facilitating business decision making, and other information [26]. There are several techniques in data mining that can be used to extract information from a set of data. These techniques include classification, clustering, association and prediction [27]. Classification is a data mining model that tests a number of records, and each record contains a target variable and a set of input or predictive variables [28]. The classification process has two stages, the first is learning: training data is analyzed using a classification algorithm and the second is classification, where test data is used to estimate the accuracy of the classification rules [29]. The classification itself has several algorithms, including Naïve Bayes classification, K-Nearest Neighbor, decision trees and Support Vector Machines [30].

The Naïve Bayes algorithm is an algorithm for predicting the probability of membership in a class [29]. This algorithm is very simple and each attribute is independent, which allows each attribute to contribute to the final result [31]. The Naïve Bayes probability formula [32] is written in equation 1 as follows.

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (1)$$

Another algorithm used in this research is Decision Tree C4.5. The advantage of this method is that it is able to eliminate unnecessary calculations or data because existing samples are usually only tested based on certain criteria or classes [33]. The gain and entropy formulas in the Decision Tree [32] are written in equations 2 and 3 as follows.

$$Gaint(S,A) = Entrophy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \qquad (2)$$

$$Entrophy(S) = \sum_{i=1}^{n} -pi * log2pi \qquad (3)$$

The next algorithm used in this research is K-NN. The K-NN algorithm is an algorithm that can be used to predict or classify data depending on the type of data in the existing data set [34]. The K-NN distance formula in general [34] is written in equation 4 as follows.

$$d(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{i1} - x_{i2})^2} \qquad (4)$$

Feature selection functions to determine a class at target values by reducing the number of irrelevant features and reducing data dimensions to improve system performance, efficiency and improve accuracy [35]. The feature selection used in this research is Principal Component Analysis (PCA). PCA aims to reduce dimensions or information. PCA works by calculating the variance of each attribute [36]. This research chose to use PCA because PCA can reduce data dimensions that are too large.

The main objectives in preprocessing this data are as follows [37]: Data cleaning that is filling in missing values, smoothing data noise, identifying and removing outliers and resolving inconsistencies. Normalization is one of the preprocessing techniques to remove outliers. Z-score normalization, also known as standardization, is a technique in which the values on the attributes are normalized based on the mean and standard deviation [38]. Equation 5 is the Z-score normalization formula [39].

$$PZ = \frac{x - \bar{x}}{\sigma} \qquad (5)$$

Meanwhile, Min-max normalization is a method that transforms a data set into a scale ranging from 0 (min) to 1 (max) [38]. Equation 6 is the Min-max normalization formula [40].

$$X = \frac{MinRange + (X - MinValue)(MaxRange - MinRange)}{MaxValue - MinValue}$$

(6)

The Confusion Matrix is a method that can be used to measure the performance of a classification method [41]. Confusion Matrix is a table consisting of accuracy, precision and recall. The formula for accuracy, precision and recall is shown in equation 7 as follows [26].

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN} \qquad (7)$$

$$Recall = \frac{TP}{TP + FN} \qquad (8)$$

$$Precision = \frac{TP}{TP + FP} \qquad (9)$$

2.3 Research Steps

The steps in this research are shown in figure 5 as follows.

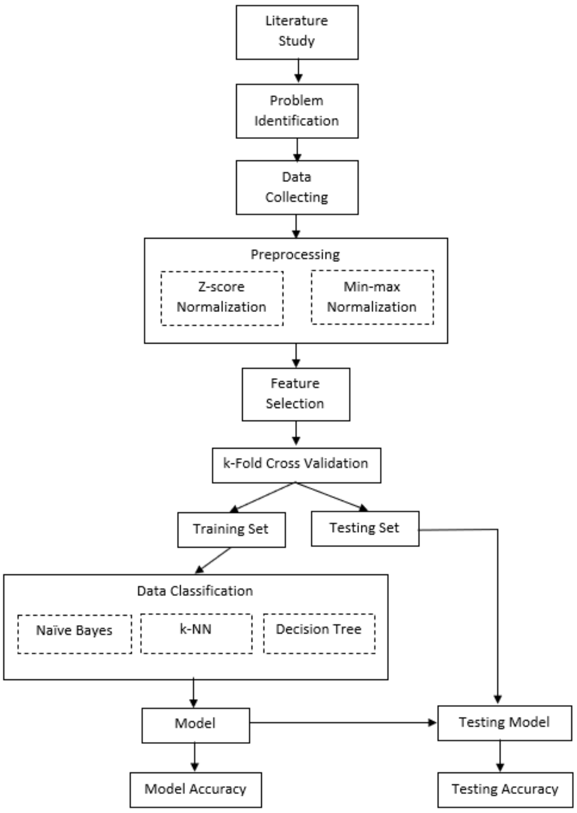

Fig 3: Research Stages Diagram

Figure 5 shows a process diagram or stages of the research conducted. It starts from literature study, problem identification and data collecting. After that, data preprocessing will be carried out clean data. After data preprocessing, normalization will also be carried out. Next, the feature selection with Principal Component Analysis (PCA) and Cross Validation. Then the data will be split into training and testing set. The training set data will processed by classification with naïve bayes, k-Nearest Neighbor and Decision Tree algorithm. From classification, when the model has been obtained, it will try this model into testing set data. The last, model accuracy and testing accuracy will be obtained. It will analysis the accuracy of all research to know what the best classification.

3. RESULT AND DISCUSSIONS

3.1 Result

The data that used in this research is Data Indikator Kesejahteraan Sosial (IKS). IKS data is data collected from residents of Bantul Regency in 2022. The IKS data currently consists of 95,347 rows and uses 27 attributes as shown in table 1 below. It will classified in 4 labels or classes in accordance with the regulations announced by BPS in a presentation about welfare classification, that very poor, poor, nearly poor and not poor.

Table 1. List of Attributes

Num	Features	Num	
1	Building State	15	Healthy Service Access
2	Floor Area	16	Disability
3	Floor Type	17	Chronical Disease
4	Wall Type	18	Education Burden
5	Roof Type	19	Highest Degree
6	Water Source	20	Occupation

7	Lighting Source	21	Income Family Members
8	Cooking Fuel	22	Income per Capita
9	Defecate Facility	23	Electronics Asset
10	Closet Type	24	Number of Motorcycle
11	Landfills	25	Number of Car
12	Eating Capability	26	Immovable Assets
13	Protein Consumption	27	Farm and Pet Animals
14	Ability To Buy Clothes		

Table 1 is a list of the attributes of the dataset used in the research. This attribute is listed in the data collection on Indikator Kesejahteraan Sosial in Bantul Regency which will be carried out in 2022.

a. Naïve Bayes

The results of the Confusion Matrix for the Naïve Bayes method with Z-score normalization before PCA feature selection are performed are shown in table 2 as follows.

Table 2. Confusion Matrix Naive Bayes, Z-score, without PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	306	1499	82	0
Poor	15	29226	6371	0
Nearly Poor	0	4284	48624	103
Not Poor	0	0	2889	1888

Table 2 shows the Confusion Matrix table from the Naïve Bayes method with Z-score normalization before PCA feature selection is performed. From the table it can be calculated the value of accuracy, precision and recall. The Confusion Matrix results for the Naïve Bayes method with Z-score normalization after PCA feature selection are shown in table 3 as follows.

Table 3. Confusion Matrix Naive Bayes, Z-score, with PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	305	1601	175	4
Poor	16	23698	4127	0
Nearly Poor	0	9710	51933	321
Not Poor	0	0	1791	1666

Table 3 shows the Confusion Matrix table from the Naïve Bayes method with Z-score normalization after PCA feature selection. From the table it can be calculated the value of accuracy, precision and recall.

The Confusion Matrix results for the Naïve Bayes method with Min-max normalization before PCA feature selection are performed are shown in table 4 as follows.

Table 4. Confusion Matrix Naive Bayes, Min-max, without PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	306	1499	82	0
Poor	15	29226	6371	0
Nearly Poor	0	4284	48684	103
Not Poor	0	0	2889	1888

Table 4 shows the Confusion Matrix table from the Naïve Bayes method with Min-max normalization before PCA feature selection is performed. From the table it can be calculated the value of accuracy, precision and recall. The Confusion Matrix results for the Naïve Bayes method with Min-max normalization after PCA feature selection are shown in table 5 as follows.

Table 5. Confusion Matrix Naive Bayes, Min-max, with PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	285	887	14	0
Poor	36	26583	3692	8
Nearly Poor	0	7539	53363	175
Not Poor	0	0	957	1808

Table 5 shows the Confusion Matrix table from the Naïve Bayes method with Min-max normalization after PCA feature selection. From the table it can be calculated the value of accuracy, precision and recall.

b. Decision Tree

The Confusion Matrix results for the Decision Tree method with Z-score normalization before PCA feature selection is performed are shown in table 6 as follows.

Table 6. Confusion Matrix Decision Tree, Z-score, without PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	92	18	1	0
Poor	215	25253	6109	1
Nearly Poor	14	9738	51169	640
Not Poor	0	0	747	1350

Table 6 shows the Confusion Matrix table from the Decision Tree method with Z-score normalization before PCA feature selection is carried out. From the table it can be calculated the value of accuracy, precision and recall. The results of the Confusion Matrix for the Decision Tree method with Z-score normalization after PCA feature selection are performed are shown in table 7 as follows.

Table 7. Confusion Matrix Decision Tree, Z-score, with PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	107	10	0	0
Poor	215	29771	6017	0
Nearly Poor	0	522	51861	685
Not Poor	0	0	148	1306

Table 7 shows the Confusion Matrix table from the Decision Tree method with Z-score normalization after PCA feature selection is performed. From the table it can be calculated the value of accuracy, precision and recall.

The Confusion Matrix results for the Decision Tree method with Min-max normalization before PCA feature selection is performed are shown in table 8 as follows.

Table 8. Confusion Matrix Decision Tree, Min-max, without PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	92	18	1	0
Poor	215	25253	6109	1
Nearly Poor	14	9738	51169	640
Not Poor	0	0	747	1350

Table 8 shows the Confusion Matrix table from the Decision Tree method with Min-max normalization before PCA feature selection is carried out. From the table it can be calculated the value of accuracy, precision and recall. The Confusion Matrix results for the Decision Tree method with Min-max normalization after PCA feature selection are shown in table 9 as follows.

Table 9. Confusion Matrix Decision Tree, Min-max, with PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	88	6	0	0
Poor	233	27971	4901	0
Nearly Poor	0	7032	52936	451
Not Poor	0	0	189	1540

Table 9 shows the Confusion Matrix table from the Decision Tree method with Min-max normalization after PCA feature selection is performed. From the table it can be calculated the value of accuracy, precision and recall.

c. K-Nearest Neighbor

The results of the Confusion Matrix for the K-NN method with Z-score normalization before PCA feature selection are performed are shown in table 10 as follows.

Table 10. Confusion Matrix K-NN, Z-score, without PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	236	22	0	0
Poor	85	33111	970	0
Nearly Poor	0	1876	56948	213
Not Poor	0	0	108	1778

Table 10 shows the Confusion Matrix table from the K-NN method with Z-score normalization before PCA feature selection is performed. From the table it can be calculated the value of accuracy, precision and recall. The Confusion Matrix results for the K-NN method with Z-score normalization after PCA feature selection are shown in table 11 as follows.

Table 11. Confusion Matrix K-NN, Z-score, with PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	236	22	0	0
Poor	85	33140	955	0
Nearly Poor	0	1847	56976	219
Not Poor	0	0	95	1772

Table 11 shows the Confusion Matrix table from the K-NN method with Z-score normalization after PCA feature selection. From the table it can be calculated the value of accuracy, precision and recall.

The results of the Confusion Matrix for the K-NN method with Min-max normalization before PCA feature selection are performed are shown in table 12 as follows.

Table 12. Confusion Matrix K-NN, Min-max, without PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	227	21	0	0
Poor	94	33154	817	0
Nearly Poor	0	1834	57140	181
Not Poor	0	0	69	1810

Table 12 shows the Confusion Matrix table from the K-NN method with Min-max normalization before PCA feature selection is performed. From the table it can be calculated the value of accuracy, precision and recall. The results of the Confusion Matrix for the K-NN method with normalized Min-max after PCA feature selection are shown in table 13 as follows.

Table 13. Confusion Matrix K-NN, Min-max, with PCA

Real Prediction	Very Poor	Poor	Nearly Poor	Not Poor
Very Poor	230	18	0	0
Poor	91	33149	914	0
Nearly Poor	0	1842	57029	169
Not Poor	0	0	83	1822

Table 13 shows the Confusion Matrix table of the K-NN method with Min-max normalization after PCA feature selection. From the table it can be calculated the value of accuracy, precision and recall.

3.2 Discussions

Based on the results of the research that has been done, a comparison of the accuracy, precision and recall values is obtained as follows.

Comparison of accuracy values for the Naïve Bayes algorithm is shown in table 14 as follows.

Table 14. Comparison Table for Accuracy of Naive Bayes Algorithm

	Z-score	Min-max
Without PCA	84,01	84,01
With PCA	84,39	86,04

Table 14 shows a comparison of the accuracy values of the Naïve Bayes algorithm when tested with and without PCA and with the Z-score and Min-max normalization methods. The accuracy value before using PCA in the Z-score normalization method obtained a value of 84.01%, while after using PCA it became 84.39%. The accuracy value before using PCA in the Min-max normalization method obtained a value of 84.01%, while after using PCA it became 86.04%.

Comparison of accuracy values for the Decision Tree algorithm is shown in table 15 as follows.

Table 15. Comparison Table for Accuracy of Decision Tree Algorithm

	Z-score	Min-max
Without PCA	81,66	81,66
With PCA	87,10	86,56

Table 15 shows a comparison of the accuracy values of the Decision Tree algorithm when tested with and without PCA and with the Z-score and Min-max normalization methods. The accuracy value before using PCA in the Z-score normalization method obtained a value of 81.66%, while after using PCA it became 87.10%. The accuracy value before using PCA in the Min-max normalization method obtained a value of 81.66%, while after using PCA it became 86.56%.

Comparison of accuracy values for the K-NN algorithm is shown in table 16 as follows.

Table 16. Comparison Table for Accuracy of K-NN Algorithm

	Z-score	Min-max
Without PCA	96,57	96,84
With PCA	96,62	96,93

Table 16 shows a comparison of the accuracy values of the K-NN algorithm when trials were carried out with and without PCA and with the Z-score and Min-max normalization methods. The accuracy value before using PCA in the Z-score normalization method obtained a value of 96.57%, while after using PCA it became 96.62%. The accuracy value before using PCA in the Min-max normalization method obtained a value of 96.84%, while after using PCA it became 96.93%.

Comparison of precision values for the Naïve Bayes algorithm is shown in table 17 as follows.

Table 17. Comparison Table for Precision of Naive Bayes Algorithm

	Z-score	Min-max
Without PCA	57,41	57,41
With PCA	57,96	66,15

Table 17 shows a comparison of the precision values of the Naïve Bayes algorithm when tested with and without PCA and with the Z-score and Min-max normalization methods. The precision value before using PCA in the Z-score normalization method obtained a value of 57.41%, while after using PCA it became 57.96%. The precision value before using PCA in the Min-max normalization method obtained a value of 57.41%, while after using PCA it became 66.15%.

Comparison of precision values for the Decision Tree algorithm is shown in table 18 as follows.

Table 18. Comparison Table for Precision of Decision Tree Algorithm

	Z-score	Min-max
Without PCA	78,15	78,15
With PCA	88,77	88,93

Table 18 shows a comparison of the precision values of the Decision Tree algorithm when tested with and without PCA and with the Z-score and Min-max normalization methods. The precision value before using PCA in the Z-score normalization method obtained a value of 78.15%, while after using PCA it became 88.77%. The precision value before using PCA in the Min-max normalization method obtained a value of 78.15%, while after using PCA it became 88.93%.

Comparison of precision values for the K-NN algorithm is shown in table 19 as follows.

Table 19. Comparison Table for Precision of K-NN Algorithm

	Z-score	Min-max
Without PCA	94,78	95,53
With PCA	94,95	95,47

Table 19 shows a comparison of the precision values of the K-NN algorithm when trials were carried out with and without PCA and with the Z-score and Min-max normalization methods. The precision value before using PCA in the Z-score normalization method obtained a value of 94.78%, while after using PCA it became 94.95%. The precision value before using PCA in the Min-max normalization method obtained a value of 95.53%, while after using PCA it became 95.47%.

Comparison of recall values for the Naïve Bayes algorithm is shown in table 20 as follows.

Table 20. Comparison Table for Recall of Naive Bayes Algorithm

	Z-score	Min-max
Without PCA	83,98	86,87
With PCA	89,39	89,39

Table 20 shows a comparison of the recall values in the Naïve Bayes algorithm when trials were carried out with and without PCA and with the Z-score and Min-max normalization methods. The recall value before using PCA in the Z-score normalization method obtained a value of 83.98%, while after using PCA it became 89.39%. The recall value before using PCA in the Min-max normalization method obtained a value of 86.87%, while after using PCA it became 89.39%.

Comparison of recall values for the Decision Tree algorithm is shown in table 21 as follows.

Table 21. Comparison Table for Recall of Decision Tree Algorithm

	Z-score	Min-max
Without PCA	64,17	64,17
With PCA	68,32	68,96

Table 21 shows a comparison of the recall values in the Decision Tree algorithm when trials were carried out with and without PCA and with the Z-score and Min-max normalization methods. The recall value before using PCA in the Z-score normalization method obtained a value of 64.17%, while after using PCA it became 68.32%. The recall value before using PCA in the Min-max normalization method obtained a value of 64.17%, while after using PCA it became 68.96%.

Comparison of recall values for the K-NN algorithm is shown in table 22 as follows.

Table 22. Comparison Table for Recall of K-NN Algorithm

	Z-score	Min-max
Without PCA	88,88	88,69
With PCA	88,84	89,03

Table 22 shows a comparison of the recall values in the K-NN algorithm when trials were carried out with and without PCA and with the Z-score and Min-max normalization methods. The recall value before using PCA in the Z-score normalization method obtained a value of 88.88%, while after using PCA it became 88.84%. The recall value before using PCA in the Min-max normalization method obtained a value of 88.69%, while after using PCA it became 89.03%.

4. CONCLUSION

Based on the three classification algorithms that have been carried out in this research, it is concluded that the best welfare classification performance is using the K-NN algorithm. This can be seen from the value of accuracy, precision and recall of the classification with the K-NN algorithm. The K-NN algorithm has the highest performance compared to other algorithms where the average values for accuracy, precision and recall are 96.71%, 95.16% and 88.79% respectively.

In several experiments without using feature selection, the Z-score and Min-max normalization methods have the same value, in the experiments with the Naïve Bayes and Decision Tree classification algorithms where the average performance produces the same value as the Z-score normalization and Min-max. So it can be concluded that the normalization method does not have a big effect on the classification if feature selection is not used. However, in all experiments that have used feature selection, there is a difference in performance between the experiments using the Z-score and Min-max normalization methods. So it can be concluded that the normalization method greatly influences the classification when feature selection is used. The normalization method can further improve accuracy in experiments with feature selection compared to without feature selection because normalization helps in overcoming the problem of scale and variation between features selected after feature selection. With normalization, these features are brought to a uniform scale, making it easier for the model to understand and extract information from each feature, which can ultimately improve the performance of the model.

In almost all experiments, the Principal Component Analysis (PCA) feature selection method is very influential in improving the performance of the welfare level classification. Although in experiments using the K-NN algorithm there was a decrease in performance in precision and recall values, namely from 95.53% to 95.47% in Min-max normalization and from 88.88% to 88.84% in Z-score normalization. So it can be concluded that the PCA feature selection method is very influential in improving the welfare level classification performance.

5. ACKNOWLEDGMENTS

Our thanks to Dinas Sosial Kabupaten Bantul.

6. REFERENCES

[1] R. Manurung, P. S. Ramadhan, and M. I. Perangin-angin, “Perbandingan Akurasi Klasifikasi Tingkat Kemisinan Antara Algoritma C 4.5 Dan Naive Bayes,” *J. Ilm. NERO*, vol. 2, no. 4, pp. 37–43, 2019, [Online]. Available: http://nero.trunojoyo.ac.id/index.php/nero/article/view/42

[2] L. Afifah, “Apa itu Regresi, Klasifikasi, dan Clustering (Klasterisasi)?,” Ilmu Data Py. [Online]. Available: https://ilmudatapy.com/apa-itu-regresi-klasifikasi-dan-clustering-klasterisasi/

[3] Y. Hastuti and M. Muzaini, “Algoritma Chaid Pada Klasifikasi Rumah Tangga Miskin Kota Palopo,” *J. Mat. dan Apl.*, vol. 1, no. 2, pp. 22–30, 2021.

[4] N. P. N. Hendayati and M. Nurhidayati, “Regresi Logistik Biner dalam Penentuan Ketepatan Klasifikasi Tingkat Kedalaman Kemiskinan Provinsi-Provinsi di Indonesia,” *J. Sains dan Teknol.*, vol. 12, no. 2, pp. 63–70, 2020.

[5] A. M. Wahyu, P. G. Anugrah, A. M. Danyalin, and R. D. Noorrizki, “Ketimpangan Ekonomi Berdampak pada Tingkat Kriminalitas? Telaah dalam Perspektif Psikologi Problematika Sosial,” *J. Ilm. Ilmu Sos.*, vol. 7, no. 2, p. 170, 2021, doi: 10.23887/jiis.v7i2.35361.

[6] B. BPS, “Tabel Kemiskinan,” BPS Kab Bantul. [Online]. Available: https://bantulkab.bps.go.id/subject/23/kemiskinan.html#subjekViewTab3

[7] K. DKB Ditjen Dukcapil, “Statistik Penduduk DIY,” Biro Tata Pemerintahan Setda DIY. [Online]. Available: https://kependudukan.jogjaprov.go.id/statistik/penduduk/jumlahpenduduk/14/0/12/04/.clear

[8] H. Annur, “Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes,” *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[9] J. Y. Kim, “Using Machine Learning to Predict Poverty Status in Costa Rican Households,” *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.3971979.

[10] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman, “Poverty classification using machine learning: The case of Jordan,” *Sustain.*, vol. 13, no. 3, pp. 1–16, 2021, doi: 10.3390/su13031412.

[11] J. A. Talingdan, “Performance comparison of different classification algorithms for household poverty classification,” *Proc. - 2019 4th Int. Conf. Inf. Syst. Eng. ICISE 2019*, no. 4, pp. 11–15, 2019, doi: 10.1109/ICISE.2019.00010.

[12] M. Gallardo, “Measuring vulnerability to multidimensional poverty with Bayesian network classifiers,” *Econ. Anal. Policy*, vol. 73, pp. 492–512, 2022, doi: 10.1016/j.eap.2021.11.018.

[13] M. A. Hanafiah and A. Wanto, “Implementation of Data Mining Algorithms for Grouping Poverty Lines by District/City in North Sumatra,” *(International J. Inf. Syst. …*, vol. 3, no. 36, pp. 315–322, 2020.

[14] Y. Shino, Y. Durachman, and N. Sutisna, “Implementation of Data Mining with Naive Bayes Algorithm for Eligibility Classification of Basic Food Aid Recipients,” *Int. J. Cyber IT Serv. Manag.*, vol. 2, no. 2, pp. 154–162, 2022, doi: 10.34306/ijcitsm.v2i2.114.

[15] E. Firasari, N. Khasanah, U. Khultsum, D. N. Kholifah, R. Komarudin, and W. Widyastuty, “Comparation of K-Nearest Neighboor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance,” *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012077.

[16] E. Afrianto, J. E. Suseno, and B. Warsito, “Decision Tree Method with C4.5 Algorithm for Students Classification Who is Entitled to Receive Indonesian Smart Card (KIP),” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, 2020, doi: 10.1088/1757-899X/879/1/012072.

[17] L. G. P. Suardani, I. M. A. Bhaskara, and M. Sudarma, “Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients,” *Int. J. Eng. Emerg. Technol.*, vol. 3, no. 1, pp. 66–70, 2018.

[18] J. Drábeková, “Classification model of poverty risk in the European Union,” *Math. Educ. Res. Appl.*, vol. 7, no. 2, pp. 73–80, 2021, doi: 10.15414/meraa.2021.07.02.73-80.

[19] Fitria, “Perbandingan Algoritma Naive Bayes Validasi 2

dan 3 Pada Klasifikasi Keluarga Miskin di Kabupaten Banjar," *J. Phasti*, vol. 05, no. April, pp. 8–14, 2019.

[20] E. Fitriani, "Perbandingan Algoritma C4.5 dan Naive Bayes untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan," *J. Sist. Inf.*, vol. 9, no. 1, pp. 103–115, 2019.

[21] D. Ispriyanti, A. Prahutama, and Mustafid, "Analisis Klasifikasi Kemiskinan di Kota Semarang Menggunakan Algoritma Quest," *J. Stat.*, vol. 7, no. 1, 2019.

[22] K. S. Utomo, "Perbandingan Algoritma Machine Learning untuk Penentuan Klasifikasi Kemiskinan Multidimensi di Provinsi Nusa Tenggara Timur," *J. Stat. Terap.*, vol. 2, no. April, pp. 36–46, 2022.

[23] N. Zaman *et al.*, *Sumber Daya dan Kesejahteraan Masyarakat*. Medan: Yayasan Kita Menulis, 2021. [Online]. Available: https://books.google.co.id/books?id=bKIjEAAAQBAJ&hl=id&source=gbs_navlinks_s

[24] D. Arfiani, *Berantas Kemiskinan*. Semarang: Alprin, 2019. [Online]. Available: https://books.google.co.id/books?id=xnn7DwAAQBAJ&hl=id

[25] H. Samsudin, Sadiman, and I. Pachrozi, *Kajian Sosial : Menuju Kemiskinan Satu Digit*. Banyuasin: Bappeda Litbang Banyuasin, 2019. [Online]. Available: https://books.google.co.id/books?id=dKndDwAAQBAJ&hl=id

[26] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*, 1st ed. Yogyakarta: Andi, 2020. [Online]. Available: https://books.google.co.id/books?id=AtcCEAAAQBAJ&hl=id

[27] L. Afifah, "Algoritma K-Nearest Neighbor (KNN) untuk Klasifikasi," Ilmu Data Py. [Online]. Available: https://ilmudatapy.com/algoritma-k-nearest-neighbor-knn-untuk-klasifikasi/

[28] L. Muflikhah, D. E. Ratnawati, and R. R. M. Putri, *Data Mining*. Malang: Tim UB Press, 2018. [Online]. Available: https://books.google.co.id/books?id=V_NqDwAAQBAJ&hl=id

[29] A. Wanto *et al.*, *Data Mining : Algoritma dan Implementasi*. Medan: Yayasan Kita Menulis, 2020. [Online]. Available: https://books.google.co.id/books?id=gAnfDwAAQBAJ&hl=id

[30] U. Sa'adah, M. Y. Rochayani, D. W. Lestari, and D. A. Lusia, *Kupas Tuntas Algoritma Data Mining dan Implementasinya menggunakan R*. Malang: Tim UB Press, 2021. [Online]. Available: https://books.google.co.id/books?id=SI1TEAAAQBAJ&hl=id

[31] S. Marpaung, Solikhun, and Irawan, "Penerapan Metode Naïve Bayes Dalam Memprediksi Prestasi Siswa Di SMA Negeri 1 Panombeian Panei," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 8–13, 2021, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.1522.

[32] E. N. R. Khakim, "Perbandingan Algoritma Klasifikasi Data Kesejahteraan Sosial Kabupaten Bantul," *Process. J. Ilm. Sist. Informasi, Teknol. Inf. dan Sist. Komput.*, vol. 17, no. 2, pp. 91–100, 2022.

[33] Binus, "Decision Tree Algoritma Beserta Contohnya Pada Data Mining," Binus. [Online]. Available: https://sis.binus.ac.id/2022/01/21/decision-tree-algoritma-beserta-contohnya-pada-data-mining/

[34] A. Khairi, A. F. Ghozali, and A. D. N. Hidayah, "Implementasi K-Nearest Neighbor (KNN) untuk Mengklasifikasi Masyarakat Pra-Sejahtera Desa Sapikerep Kecamatan Sukapura," *TRILOGI J. Ilmu Teknol. Kesehatan, dan Hum.*, vol. 2, no. 3, pp. 319–323, 2021, doi: 10.33650/trilogi.v2i3.2878.

[35] Sulandri, A. Basuki, and F. A. Bachtiar, "Metode Deteksi Intrusi Menggunakan Algoritme Extreme Learning Machine dengan Correlation-based Feature Selection," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, pp. 103–110, 2021, doi: 10.25126/jtiik.0813358.

[36] D. Dahman, "Dimensionality Reduction : LDA, PCA, t-SNE," Medium. [Online]. Available: https://medium.com/sysinfo/dimensionality-reduction-lda-pca-t-sne-b85254e04348#:~:text=Perbedaan mendasar lain yang membedakan,data dapat dipisahkan dengan baik.

[37] Alfarisi, "Data Preprocessing - Konsep Pembelajaran Data Mining," Steemit. [Online]. Available: https://steemit.com/education/@alfarisi/data-preprocessing-konsep-pembelajaran-data-mining

[38] Trivusi, "Normalisasi Data : Pengertian, Tujuan dan Metodenya," Trivusi. [Online]. Available: https://www.trivusi.web.id/2022/09/normalisasi-data.html#:~:text=Normalisasi min-max biasanya memungkinkan,tidak memperlakukan outlier dengan baik.

[39] R. G. Whendasmoro and Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 4, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i4.4526.

[40] H. E. Wahanani, M. H. P. Swari, and F. A. Akbar, "Case based Reasoning Prediksi Waktu Studi Mahasiswa Menggunakan Metode Euclidean Distance dan Normalisasi Min-Max," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, pp. 1279–1288, 2020, doi: 10.25126/jtiik.2020763880.

[41] W. Nengsih, "Analisa Akurasi Permodelan Supervised Dan Unsupervised Learning Menggunakan Data Mining," *Sebatik*, vol. 23, no. 2, pp. 285–291, 2019, doi: 10.46984/sebatik.v23i2.771.