Comparation of K-Nearest Neighbor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance

Elly Firasari¹*, Nurul Khasanah², Umi Khultsum³, Desiana Nur Kholifah⁴, Rachman Komarudin⁵ and, Wiwiek Widyastuty⁶
¹,²,³,⁴,⁶Universitas Bina Sarana Informatika, Indonesia
³Program Studi Sistem Informasi, STMIK Nusa Mandiri, Indonesia
E-mail: elly.efa@bsi.ac.id

Abstract. Poverty is a problem faced by developing countries, as well as Indonesia. According to data from the Central Statistics Agency in 2018, more than half the distribution of the poor population in Indonesia is in Java, which is 13,340.15 million people. Somokerto village is one of the villages in the district of Magelang, Central Java, which receives government assistance in an effort to reduce poverty. But in the process of classifying citizens who are entitled to receive assistance is still done manually. Manual classification is considered inaccurate in obtaining the results of social assistance recipients. In overcoming this problem, we need a systematic calculation to get accurate results. In this case, the researcher uses data mining classification calculation by comparing 2 calculation methods, namely K-NN and Naive Bayes. The researchers use Rapidminer tools. The research stages are identification of problems, data collection, implementation K-NN, Implementation Naive bayes, data testing process to produce accuracy and compare the result. The results obtained are the accuracy of Naive Bayes higher than K-NN, namely Naive Bayes 89.04% and K-NN 87.67%. This figure is classified in the category of good classification. From the results of the study it can be concluded the Naive Bayes algorithm is suitable to be applied in the calculation of recipients of social assistance.

1. Introduction
Poverty is a problem faced by developing countries, as well as Indonesia. The problem of poverty is very closely related to social, economic, educational, and other aspects. This means that the problem of poverty has indeed become an important point to be overcome. When viewed in terms of economic growth, Indonesia’s condition is getting better. The poverty percentage in September 2018 was 9.66 percent, down 0.16 points compared to March 2018 and decreased 0.46 percentage points compared to September 2017 [1]. Even so, this figure has not met the target set out in the RPJM of 7-8% at the end of 2019, while the reality on the ground the poverty rate in Indonesia in 2018 is still at 9.66%.

According to data from the Central Statistics Agency in 2018, more than half the distribution of poor population in Indonesia is in Java, which is 13,340.15 million people, while the number of poor people in urban areas is 6,573.80 million people. The number of poor people in Magelang Regency from 2012 to 2018 has decreased. From 2012-2018, the number of poor people decreased by 25 960 people from 169.4 thousand people to 143.44 thousand people in 2018. Relatively the
percentage of poor people also decreased from 13.97 percent in 2012 to 11.23 percent in 2018\cite{2}. Compared to the previous year, the number of poor people in 2018 decreased by 13,710 people (8.72%). The decrease in poor population in 2018 shows the biggest decline in the last 6 years.

To reduce poverty in Indonesia, the government already has several programs that have been implemented, including Poor Rice, Program Keluarga Harapan (PKH), Bantuan Langsung Tunai (BLT) and others. Indicators that are considered for the provision of social assistance include: Name of family head, latest education KK, house area, house floor, house wall, availability of electricity, fuel, water source, type of building ownership, assets owned, monthly income, and several other supporting indicators. BPKP Review (Financial and Development Supervisory Agency) of 2014 Social Expenditures Spending at 11 K/L, from the social assistance value of Rp. 18.6 T found: 45.2 percent was not on target at 9 K/L, 54.8 percent was right on target\cite{3}. The distribution of aid that is not on target will certainly have an impact on economic problems that are difficult to overcome.

The Somokerto village government still uses the classification process manually in determining the recipients of social assistance even though there are a lot of data being processed. Implementation of a system that is still manual is considered inaccurate in obtaining the results of recipients of social assistance.

In overcoming this problem, a systematic calculation is needed in determining who is eligible to receive social assistance \cite{4}. Data mining is the process of analyzing the extraction of information and a very large set of data through the use of algorithms and withdrawal techniques in the field of machine learning statistics and data management systems. Classification is a process of grouping, which means collecting objects or entities that are the same and separating objects or entities that are not the same \cite{5}. In the classification process, there are many methods that can be applied. The author compares several previous studies related to the use of the method in this case. comparing 4 PSO-Based Algorithms in predicting student graduation, the K-NN algorithm gave the best results than the other algorithms used\cite{6}. The results obtained are an accuracy value of 74.08% and an AUC value of 0.788. Naive Bayes Classification in Document Classification for Identification of E-Government Content states the results that the classification of documents using the Naive Bayes Classifier in this study with training data of 260 political documents and 222 economic documents using 40 Testing data shows good accuracy in the overall classification, with an overall classification accuracy of 85%\cite{7}.

From several studies that have been done before, there are 2 methods that provide the best results in calculations, namely Naïve Bayes and K-NN. For this reason, in this study the writer will compare the two methods to be applied in the calculation of recipients of social assistance in Somokerto Village. The results of the study are expected to be able to be applied in the process of classifying the recipients of social assistance in Somokerto Village so that the assistance will be on target.

In the study of K-Nearest Neighbor algorithm for the classification of the Indonesia Healthy Card giving a precision value of 97.66% accuracy 98.5% and recall 96.5%\cite{8}, that compared 4 PSO-Based Algorithms in predicting student graduation, the K-Nearest Neighbor algorithm gave the best results than the other algorithms used\cite{6}.

In this study, the method used for the calculation has never been applied to similar cases. We do a comparison of two Algorithms, namely K-NN and Naive Bayes to compare the best Algorithm with high accuracy.

2. Literature Review

2.1. Data Mining
Data Mining is the process of discovering useful new correlations, patterns and trends by mining a large number of data repositories, which use pattern recognition technologies such as statistics and mathematical techniques\cite{9}.
2.2. Classification
Classification is a process of grouping, which means collecting objects or entities that are the same and separating objects or entities that are not the same[5]. In the data mining classification process, there are several algorithms that can be used, including: K-NN, Naïve Bayes, C.45, SVM, Apriori, and others.

Evaluation and validation of the results of the classification rule is done by confusion matrix and ROC (Receiver Operating Characteristic) Curve. For data mining classification, AUC values can be divided into several groups [10]:

- 0.90 - 1.00 Excellent classification
- 0.80 - 0.90 Good classification
- 0.70 - 0.80 Fair classification
- 0.60 - 0.70 Poor classification
- 0.50 - 0.60 Failure

2.3. Naïve Bayes Classifier
Naïve Bayes is a simple probability-based prediction technique based on application of the Bayes theorem with a strong assumption independence[11].

2.4. K-Nearest Neighbor
K-Nearest Neighbor is a method for classifying objects based on learning data that is the closest distance or has the most characteristic similarities with the object[5]. Near or far neighbors are usually calculated by Euclidean distance.

Steps - steps to calculate the K-NN algorithm [5]:

- Determine the value of k.
- Calculate the square of euclid distance (query instances) of each object against the training data.
- Sort objects - these objects into groups that have the smallest euclid distance.
- Collecting class Y labels (Nearest Neighborhood classification)

2.5. Data Analysis Methods
Data analysis techniques use quantitative data in the form of mathematical calculations of numbers or numerical and nominal. In this study, data analysis was carried out using population data in one of the Magelang district villages with an average of poor citizens receiving Social assistance and non-poor citizens receiving social assistance. Data was processed and tested using the K-NN and Naïve Bayes algorithm, then will be tested using a confusion matrix and Receiver Operating Characteristic (ROC) Curve to obtain the level of accuracy generated by the two algorithms in order to get better results from both algorithms.

3. Research Methods
3.1. Research Stages
The following is the research method flow:
4. Result and Discussion

4.1. Calculation with K-Nearest Neighbor

The following are the results of calculation with K-Nearest Neighbor method:

|            | True Poor | True Not Poor | Class Precision |
|------------|-----------|---------------|-----------------|
| Pred. Poor | 41        | 9             | 82.00%          |
| Pred. Not Poor | 0        | 23            | 100%            |
| Class Recall | 100%     | 71.88%        |                 |

The number of True Positive (TP) is 41 which is classified as Poor Population and False Negative (FN) as many as 0 which is classified as Poor apparently not Poor. then 9 for True Negative (TN) which is classified as Non-Poor, and 23 False Positive (FP) turned out to be classified as Not Poor. Based on the table above, it shows that the level of accuracy using the Naïve Bayes Classifier algorithm is 87.67%.

4.2. Calculation with Naïve Bayes

The following are the results of calculation with Naive Bayes method:

|            | True Poor | True Not Poor | Class Precision |
|------------|-----------|---------------|-----------------|
| Pred. Poor | 41        | 8             | 82.00%          |
| Pred. Not Poor | 1        | 23            | 100%            |
| Class Recall | 100%     | 75.00%        |                 |

The number of True Positive (TP) is 41 which is classified as Poor Population and False Negative (FN) as many as 0 which is classified as Poor apparently not Poor. then 8 for True
Negative (TN) which is classified as Non-Poor. and 23 False Positive (FP) turned out to be classified as Not Poor. Based on the table above, it shows that the level of accuracy using the Naïve Bayes Classifier algorithm is 89.04%.

4.3. Comparasion K-NN with Naïve Bayes
Comparison of the percentage accuracy of the K-NN and Naive Bayes methods is as follows:

| No | Method                  | Accuracy | Class Precision | Class Recall |
|----|-------------------------|----------|-----------------|--------------|
| 1  | K-Nearest Neighbor      | 87.67%   | 100%            | 71.88%       |
| 2  | 89.04%                  | 100%     | 75.00%          |              |

4.4. Evaluate Model ROC with K-NN Method
ROC graphics are used to express confusion matrix data. Horizontal lines represent false positive values (FP) and vertical lines represent true negative (TP). With an AUC (Area Under Cover) value of 0.875 included in True Not Poor.

Figure 2. AUC K-NN Method

4.5. Evaluate Model ROC with Naive Bayes Method
ROC graphics are used to express confusion matrix data. Horizontal lines represent false positive values (FP) and vertical lines represent true negative (TP). With an AUC (Area Under Cover) value of 0.953 included in True Not Poor.
5. Closing

5.1. Conclusion

The conclusions of this study include:
1. Naïve Bayes Classifier algorithm has greater accuracy than the K-NN algorithm to be applied in classifying the poor population in receiving social assistance in Somokerto Village.
2. Research using Naïve Bayes Classifier algorithm produces a classification with an accuracy rate of 89.04% while K-NN produces a classification with an accuracy rate of 87.67%.
3. Both calculations are included in the category of good classification.

Suggestions that can be given for the development of this research are:
1. The Research Ligkup Room should be expanded to become a sub-district or district level
2. This research can be developed using other algorithms to produce a better research
3. Adding attributes in order to produce better accuracy.

References

[1] BPS, 2018 Presentase penduduk miskin Sptember.
[2] BPS, 2018 Presentase Penduduk Miskin dan Jumlah Penduduk Miskin Kabupaten Magelang
[3] Kementrian Keuangan, 2015 Tata Kelola Bantuan Sosial Kementerian / Lembaga.
[4] A. Saleh, 2015 Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga Creat. Inf. Technol. J. 2, 3 p. 207–217.
[5] D. Syahid, J. Jumadi, and D. Nursantika, 2016 Sistem Klasifikasi Jenis Tanaman Hias Daun Philodendron Menggunakan Metode K-Nearest Neighbor (KNN) Berdasarkan Nilai Hue, Saturation, Value (HSV) J. Online Inform. 1, 1 p. 20.
[6] M. Zainuddin, 2018 Metode Klasifikasi Berbasis Particle Swarm Optimization (Pso) Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa J. Ilm. Teknol. Inf. Asia 13, 1 p. 1.
[7] H. A. S. Santoso, and A. K. Wijaya, 2016 Naïve Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government J. Appl. Intell. Syst. 1, 1 p. 48–55.
[8] A. A. Rahman and Y.I. Kurniawan, 2018 Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan Algoritma Naïve Bayes Classifier J. Teknol. dan Manaj. Inform. 4, 1 p. 151–161.
[9] K. Fatmawati, and A. P. Windarto, 2018 Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi Comput. Eng. Sci. Syst. J. 3, 2 p. 173.
[10] R. Sulaimani, 2016 Prediksi Keputusan Klien Telemarketing Untuk Deposito Pada Bank Menggunakan Algoritma Naïve Bayes Berbasis Backward Elimination Ilk. J. Ilm. 8, 3 p. 182–189.
[11] C. Fadlan, S. Ningsih, and A. P. Windarto, 2018 Penerapan Metode Naïve Bayes Dalam Klasifikasi Kelayakan Keluarga Penerima Beras Rastra J. Tek. Inform. Musirawas 3, 1 p. 1.

Figure 3. AUC Naïve Bayes Method