JURNAL RISET INFORMATIKA
Vol. 4, No. 3. June 2022

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.164

Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

CLASSIFICATION OF THE POOR IN SUMATERA AND JAVA ISLAND USING NAIVE BAYES ALGORITHM AND NAIVE BAYES ALGORITHM BASED ON PSO

Endang Sri Palupi

Sistem Informasi
Universitas Bina Sarana Informatika
Endang.epl@bsi.ac.id

Abstrak

Angka kemiskinan di Indonesia masih cukup tinggi, dikarenakan jumlah penduduk yang banyak serta pembangunan dan pusat perekonomian belum merata. Dengan jumlah penduduk yang banyak dan negara kepulauan yang terbentang dari barat ke timur, bukan hal mudah bagi pemerintah untuk meratakan perekonomian guna menurunkan angka kemiskinan di Indonesia. Penelitian ini dilakukan untuk mengklasifikasikan angka kemiskinan di kabupaten yang ada di pulau sumatera dan pulau jawa dengan menggunakan algoritma Naïve Bayes dan Naïve Bayes berbasis Particle Swarm Optimization. Dengan demikian diharapkan pemerintah pusat dan pemerintah daerah bisa bekerjasama untuk menjalankan program - program dalam rangka menurunkan angka kemiskinan khususnya pada kabupaten yang angka angka kemiskinannya tinggi. Berdasarkan penelitian yang dilakukan hasil klasifikasi masyarakat miskin pada kabupaten di pulau Sumatera dan pulau Jawa dengan pengujian confussion matrix dan teknik split validasi menggunakan algoritma Naïve Bayes menghasilkan angka akurasi 59,75% dan AUC 0,768 termasuk dalam klasifikasi yang baik. Sedangkan hasil klasifikasi menggunakan algoritma Naïve Bayes berbasis Particle Swarm Optimization menghasilkan angka akurasi 82,93% dan AUC 0,849 termasuk dalam klasifikasi yang baik. Dari hasil penelitian ini maka dapat disimpulkan algoritma Naïve Bayes merupakan teknik yang baik untuk pengklasifikasian dalam data mining, dan untuk hasil yang lebih maksimal bisa menggunakan Particle Swarm Optimization.

Kata kunci: klasifikasi, naïve bayes, particle swarm optimization

Abstract

The poverty rate in Indonesia is still quite high, due to the large population and uneven development and economic center. With a large population and an archipelagic country that stretches from west to east, it is not easy for the government to level the economy in order to reduce poverty in Indonesia. This study was conducted to classify the poverty rate in districts on the island of Sumatra and Java using Nave Bayes and Nave Bayes based on Particle Swarm Optimization. Thus, it is hoped that the central government and local governments can monitor the implementation of programs in order to reduce poverty rates, especially in districts with high poverty rates. Based on research conducted on the classification of the poor in districts on the island of Sumatra and Java with confusion matrix testing and validation validation techniques using the Naïve Bayes algorithm, the accuracy rate is 59.75% and AUC 0.768 is included in a good classification. While the results of the classification using the Naïve Bayes algorithm based on Particle Swarm Optimization produces an accuracy rate of 82.93% and AUC of 0.849 is included in a good classification. From the results of this study, it can be said that Al-Qur'an Naïve Bayes is a good technique for classification in data mining, and for maximum results using Particle Swarm Optimization.

Keywords: classification, naïve bayes, particle swarm optimization

INTRODUCTION

Factors causing poverty in Indonesia include: high population growth rate, increasing unemployment, low education, natural disasters, unequal income distribution, The impact of poverty is increasing crime, increasing mortality rates, access to closed education, rising unemployment, and the emergence of conflict in society. (Ahmad, 2021) It is hoped that the central government can work together with local governments to focus more on implementing programs to alleviate poverty in districts with high poverty rates. The results show that poverty has a significant negative

effect on the Human Development Index, Unemployment has a negative and insignificant effect on the Human Development Index, while Income Inequality has a positive and insignificant effect on the Human Development Index. (Simarnata, 2019)

In the journal entitled Classification of the Poor Using the Naïve Bayes Method written by Haditsah Annur in 2018, the results of the confusion matrix test using the split validation technique, using the nave Bayes classification method for datasets that have been taken on the research object, obtained an accuracy rate of 73% or including in the Good category. While the Precision value is 92% and Recall is 86% (Annur, 2018). The study only used one classification method so that there was no comparison of accuracy values, and the research was only conducted in one district, namely Gorontalo, so the amount of data was not too much. Based on the research that has been done, it can be concluded that the higher the number of records in the dataset, the higher the accuracy value obtained and each dataset has a different algorithm suitability.(Wicaksono & Padilah, 2021)

The journal entitled Application of Data Mining for Clustering of Poor Population Data Using the K-Means Method was written by Parjito and colleagues in 2021, based on the results of testing the training data using Weka, it can be concluded that from 812 population training data, 103 data were classified as poor and 709 being not poor. with 4 iterations. The percentage of 13% in the poor category and 83% in the non-poor category from the results of the application of the Weka tools while the validity of the DBI method resulted in the number of suitable clusters being 2 clusters with a value of 0.1643 which means the cluster is suitable for use (Astuti, 2017). This study only uses one classification algorithm, namely K-Means and the research area is relatively small, only one village, namely Suka Bhakti Village, Gedungaji Baru District, Tulang Bawang Regency, Lampung Province, so the data obtained is not much. The research classification also uses the Weka tools, which are different with RapidMiner which can display the accuracy value. In terms of data processing speed from the classification algorithm, the Rapidminer Data mining Tool has a speed that is superior to the Weka Data Mining Tool, while on the other hand, the superior accuracy level of the Tool is Weka compared to Rapidminer. From the dataset used, the India Liver Patient Dataset (ILPD) dataset which has the fastest processing time of the others. (Faid, Jasri, & Rahmawati, 2019)

In 2020 Yunita Ratna Sari and colleagues conducted a research entitled K-Means Algorithm for Clustering Poverty Data in Banten Province Using RapidMiner. The results of the study were 3 clusters, namely: medium cluster level (C0), high cluster level (C1), and low cluster level (C2). The results from the rapidminer calculations show that Tangerang Regency, Cilegon City, and Serang City are included as members of cluster 0, Pandeglang Regency, Lebak Regency, and Serang Regency are in cluster 1, Tangerang City, and South Tangerang City are in cluster 2. (Sari, Sudewa, Lestari, & Jaya, 2020) This study uses only one algorithm in its calculations and the study was conducted in the district of Banten province and used the clustering technique (Wanto, 2020). The main difference between clustering and classification is that clustering is an unsupervised learning technique which clusters the same instances based on features whereas classification is a supervised learning technique which assigns predefined tags to instances based on features.

Research has also been carried out in 2021 with the title Food Poverty Line Analysis using the K-Means Clustering Algorithm Method by Kamila Aprilia and Falentino Sembiring. This study uses Orange Software and produces 6 provinces for the highest poverty line cluster (Aprilia & Sembiring, 2021). Orange is a software with excellent data visualization advantages, using the Python programming language. Orange is synonymous with interactive and interesting applications, the way of operation is even easier than Rapidminer. (Tandika, 2022)

The poverty classification aims to help the performance of BPS in order to shorten the target classification time and classify all data sets correctly. However, it cannot be denied that the performance of a method cannot work 100% correctly. One method that can be used for the classification of poverty is to use data mining. (Effendy & Purbandini, 2018)

RESEARCH METHODS

This study uses quantitative research methods, namely research that uses data that already exists and is ready to be processed. Quantitative method is a research that is based on positivistic (concrete data), research data in the form of numbers to be measured using statistics as a calculation test tool, related to the problem under study to produce a conclusion. (Sugiyono, 2018), The stages of research are in Figure 1. Research Framework.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.164
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

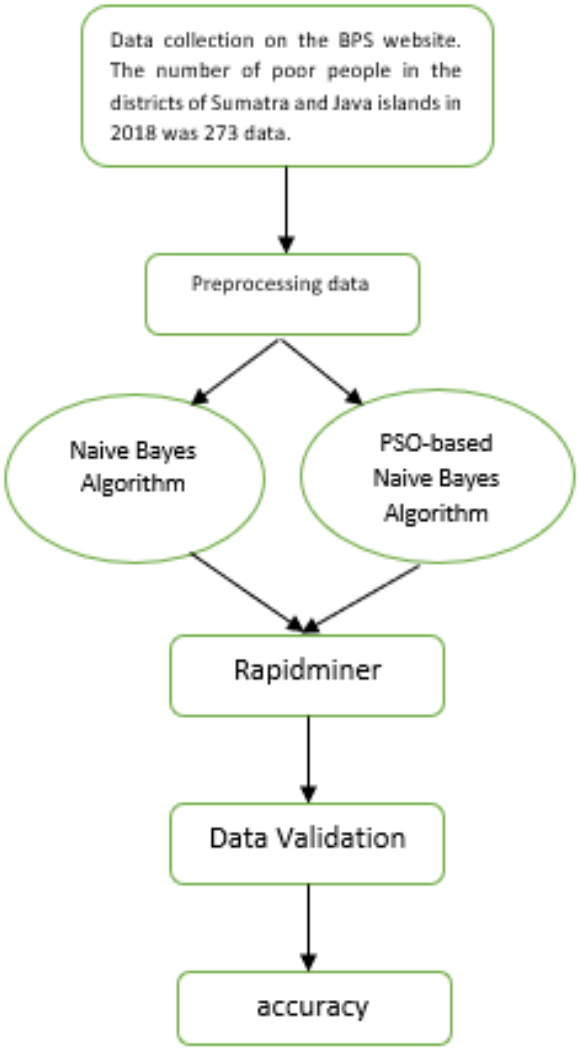

Figure 1. Research Framework

This research was conducted in September 2021. Taking data from the BPS website, namely data on the number of poor people in districts on the island of Sumatra and Java in 2007 to 2015, as many as 273 districts. After taking the dataset on the BPS website, it enters the preprocessing stage, namely: Data Cleaning cleans empty or double data, Data Transformation, which converts data files into excel files, Data Reduction with the aim of increasing storage efficiency and reducing data storage and analysis costs. The following dataset is ready to use Table 1. Data Of The Number of Poor Population In Sumatra and Java District Buku1.xlsx (live.com).

Data Mining

Data mining is the process of discovering something meaningful from new correlations, patterns and trends by sifting through large data stored in repositories, using pattern recognition technology as well as mathematical and statistical techniques. Data mining is the analysis of database observations to find unexpected relationships and to summarize data in a new way or method that is understandable and useful to data owners. Data mining is also an interdisciplinary field that brings together machine learning techniques, pattern recognition, statistics, databases, and visualization to solve the problem of extracting information from large databases. Data mining can be interpreted as a process of extracting useful and potential information from a set of data contained implicitly in a database (Larose & Larose, 2014). Some of the processes carried out by data mining :

- Description (identifies hidden hidden patterns and converts patterns into rules that can be understood by experts)
- Prediction (classifying based on expected behavior in the future)
- Estimation (as prediction except for estimation variables are more numerical)
- Classification (process of finding a function model and describing data to classes)
- Clustering (grouping data without a particular class based on the object)
- Association (find the attribute that appears in time). (Usama M. Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996)

In this study, the author uses a classification process to classify the poor in districts on the islands of Sumatra and Java from 2007 to 2015 using the Naïve Bayes algorithm and the PSO-based Nave Bayes algorithm as a comparison of the accuracy results. The advantages of Naïve Bayes: Can be used for quantitative or qualitative data, do not require a lot of data, calculations are fast and efficient, easy to understand, easy to make, simple programming language, can be used multiclass (Isa, Elfaladonna, & Ariyanti, 2022). Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique (fish, bees, birds, etc.), proposed by Russell C. Eberhart and James Kennedy in 1995 which was inspired by the social behavior of the movement of birds or fish. (Hu, Eberhart, & Shi, 2003)

RESULTS AND DISCUSSION

Naïve Bayes Classification

The following in Figure 2 is modeling using the Naïve Bayes algorithm.

P-ISSN: 2656-1743 | E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.164
JURNAL RISET INFORMATIKA
Vol. 4, No. 3 June 2022
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

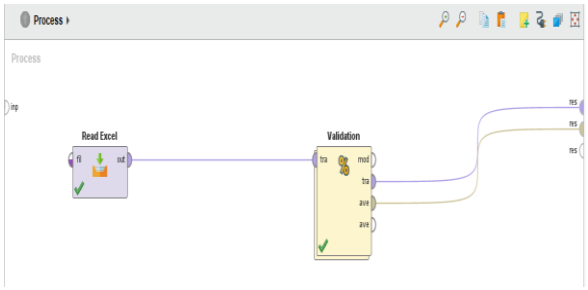

Figure 2. Naïve Bayes Algorithm

In Figure 3 below is the validation process of the Naïve Bayes algorithm.

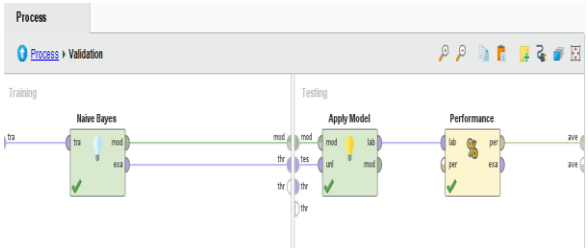

Figure 3. Naïve Bayes Algorithm Validation Process

The following are the results of accuracy using the Naïve Bayes algorithm in Figure 4.

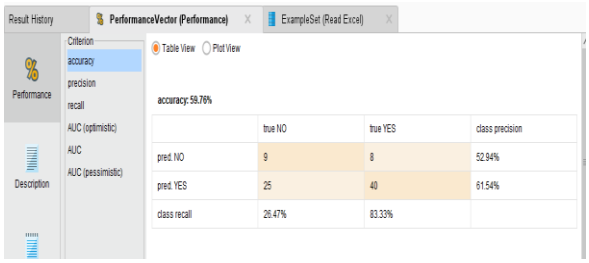

Figure 4. Accuracy of Naïve Bayes Algorithm

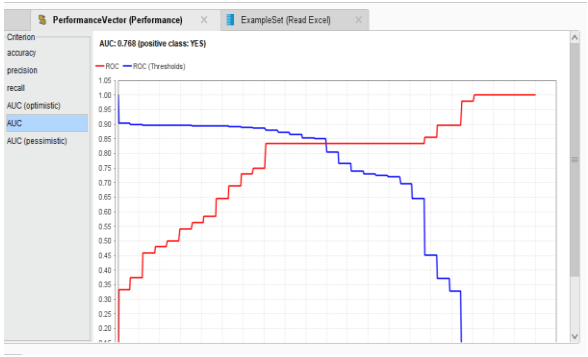

Figure 5. Area Under Curve Naïve Bayes

Figure 5 is the result of an Area Under Curve Naïve Bayes of 0.768 which is included in the good classification category.

Naïve Bayes Classification Based on Particle Swarm Optimization

Figure 6 shows the PSO-based Naïve Bayes Algorithm modeling as a comparison to the previous modeling which only used Naïve Bayes.

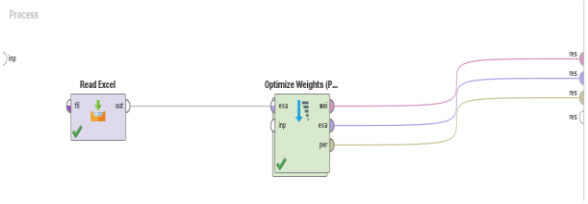

Figure 6. Modeling of Naïve Bayes Algorithm Based on PSO

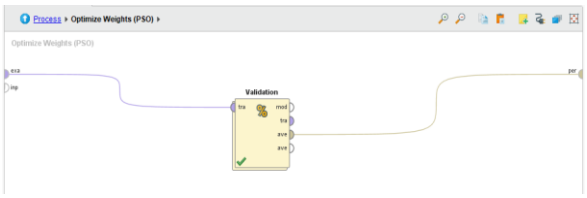

Figure 7. Naïve Bayes Based on PSO

Figure 7 is the PSO-based Naïve Bayes optimization process and then Figure 8 is the PSO-based Naïve Bayes algorithm validation process.

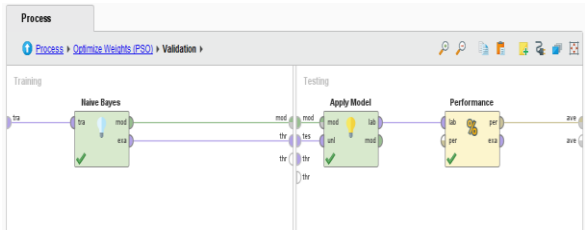

Figure 8. Naïve Bayes Validation Process Based on PSO

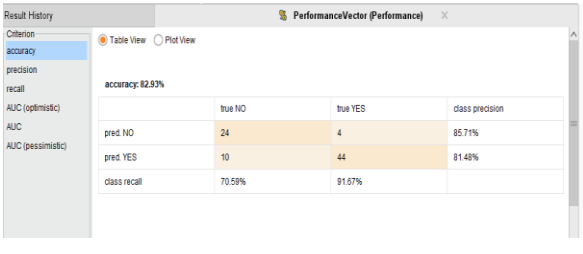

Figure 9 Accuracy of Naïve Bayes Algorithm Based on PSO

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.164
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

Figure 9 shows the results of the PSO-based Naïve Bayes accuracy with a result of 82.93%, the accuracy using PSO is greater.

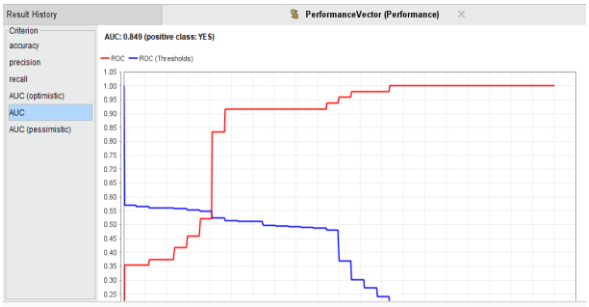

Figure 10. Area Under Curve Naïve Bayes Based on PSO

Figure 10 shows the PSO-based Naïve Bayes AUC of 0.849 which is included in the good classification category. The AUC results using Naïve Bayes based on PSO are greater than using Naïve Bayes alone.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the research conducted, the results of the classification of the poor in districts on the island of Sumatra and Java with the confusion matrix test and the split validation technique using the Naïve Bayes algorithm resulted in an accuracy rate of 59.75% and an AUC of 0.768 included in a good classification. While the results of the classification using the Naïve Bayes algorithm based on Particle Swarm Optimization produces an accuracy rate of 82.93% and AUC of 0.849 is included in a good classification. From the results of this study, it can be concluded that the Naïve Bayes algorithm is a good technique for classification in data mining, and for maximum results, Particle Swarm Optimization can be used.

Suggestion

Further research can classify all districts in Indonesia in full, and can use other algorithm methods so that comparisons can be seen and can produce better accuracy values.

REFERENCES

Ahmad. (2021). Faktor Penyebab Kemiskinan dan Dampaknya. Retrieved July 11, 2022, from Gramedia Blog website: https://www.gramedia.com/literasi/penyebab-kemiskinan/

Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes. *ILKOM Jurnal Ilmiah*, *10*(2), 160–165. https://doi.org/10.33096/ilkom.v10i2.303.160-165

Aprilia, K., & Sembiring, F. (2021). Analisis Garis Kemiskinan Makanan Menggunakan Metode Algoritma K-Means Clustering. *Seminar Nasional Sistem Informasi Dan Manajemen Informatika*, 1–10. Sukabumi: Universitas Nusa Putra. Retrieved from https://sismatik.nusaputra.ac.id/index.php/sismatik/article/view/1

Astuti, F. D. (2017). Penerapan Data Mining Untuk Clustering Data Penduduk Miskin Menggunakan Algoritma Hard C-Means. *Data Manajemen Dan Teknologi Informasi*, *18*(1), 64–69. Retrieved from https://ojs.amikom.ac.id/index.php/dasi/article/view/1836

Effendy, F., & Purbandini, P. (2018). Klasifikasi Rumah Tangga Miskin Menggunakan Ordinal Class Classifier. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *4*(1), 30–36. https://doi.org/10.25077/teknosi.v4i1.2018.30-36

Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi. *Teknika*, *8*(1), 11–16. https://doi.org/10.34148/teknika.v8i1.95

Hu, X., Eberhart, R. C., & Shi, Y. (2003). Particle Swarm With Extended Memory For Multiobjective Optimization. *Proceedings of the 2003 IEEE Swarm Intelligence Symposium. SIS'03 (Cat. No.03EX706)*. Indianapolis: IEEE. https://doi.org/10.1109/SIS.2003.1202267

Isa, I. G. T., Elfaladonna, F., & Ariyanti, I. (2022). *Buku Ajar Sistem Pendukung Keputusan*. Pekalongan: Penerbit NEM.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (4th ed.). New Jersey: John Wiley & Sons.

Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer. *CESS (Journal of Computer Engineering, System and Science)*, *5*(2), 192. https://doi.org/10.24114/cess.v5i2.18519

Simarnata, Y. P. H. (2019). Analisis Pengaruh Kemiskinan, Pengangguran, Dan Ketimpangan Pendapatan Terhadap Indeks Pembangunan Manusia (IPM) Di Indonesia (Universitas Sumatera Utara). Universitas Sumatera Utara. Retrieved from

https://repositori.usu.ac.id/handle/123456789/15100

Sugiyono. (2018). *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Bandung: Alfabeta.

Tandika, B. (2022). 5 Aplikasi Data Mining Favorit Para Spesialis. Retrieved July 7, 2022, from Glints website: https://glints.com/id/lowongan/aplikasi-data-mining/#.YtJQtXZBzIV

Usama M. Fayyad, Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* (MIT Press, Ed.). Massachusetts.

Wanto, A. (2020). *Data Mining : Algoritma dan Implementasi*. Medan: Yayasan Kita Menulis.

Wicaksono, A., & Padilah, T. N. (2021). Pengaruh Jumlah Record Dataset Terhadap Algoritma Klasifikasi Berdasarkan Data Customer Churn. *Jurnal Ilmiah Informatika*, *6*(1), 11–19. Retrieved from https://journal.ibrahimy.ac.id/index.php/JIMI/article/view/1223