Journal Mandiri IT
Vol. 14 No. 1, July (2025), pp. 12-20

ISSN 2301-8984 (Print) , 2809-1884 (Online)
Published by Institute of Computer Science

Classification eligibility recipient BPJS in ward sendang sari using the naive bayes method

Dio Prayoga[1], Rakhmat Kurniawan[2]
[1,2]Science Study Program Computer, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

ARTICLE EI NFO

Article history:

Received Jun 1, 2025
Revised Jul 2, 2025
Accepted Jul 10, 2025

Keywords:

BPJS;
Classification;
Machine Learning;
Naive Bayes;
Sub-District Sendang Sari.

ABSTRACT

Study This done for classify eligibility BPJS recipients in the sub-district Sendang Sari with use Naive Bayes method, which is relevant in support transparency and efficiency distribution benefit guarantee social at the level sub-district. Problems main in study This is Still its use manual system in the classification process, which causes the decision-making process decision become slow, subjective and vulnerable error. Research methods involving collection of 1000 citizen data Ward Sendang Sari which consists of from attributes like type gender, employment status, ownership house, income, and amount liability. Data then through preprocessing stage, including conversion variable categorical use LabelEncoder and determination of eligibility labels based on threshold income and amount liability. Next, the data is divided into training data and test data with 80:20 ratio. Classification model built use Gaussian Naive Bayes algorithm and evaluated use confusion matrix metrics which include accuracy, precision, and recall. Evaluation results show that the model achieves accuracy of 0.97 or 97%, precision of 0.95 or 95%, and recall of 0.90 or 90%, and F1-Score of 0.93 or 93 % which to signify that this model Enough effective For classify eligibility BPJS recipients. Research This conclude that The Naive Bayes method is capable of give accurate and consistent classification, which can increase efficiency administration ward as well as speed up distribution benefit to entitled community.

Corresponding Author:

Dio Prayoga,
Science Study Program Computer,
Faculty of Science and Technology,
Universitas Islam Negeri Sumatera Utara,
Jl. William Iskandar Ps. V, Medan Estate, Kec. Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara 20371, Indonesia
Email: dioprayoga107@gmail.com

1. INTRODUCTION

Acceptance of the Organizing Agency program Social Security (BPJS) has be one of indicator important in measure level welfare society in Indonesia (Adzy et al., 2023). BPJS provides guarantee social for Indonesian citizens in various fields, including health. However, in distribution benefit said, no all inhabitant can with easy and precise time to obtain appropriate services. A number of factors, such as inability economy and lack of knowledge about procedure registration, often become obstacle for public in to obtain access to BPJS services.

Ward Sendang Sari was chosen as location study Because a number of reasons. First, as one of the densely populated urban areas residents, sub-district Sendang Sari has challenge alone in manage distribution BPJS benefits. Currently, the determination of eligibility BPJS recipients in the sub-district Sendang Sari is performed manually with the collected data become One in 1 book big by officer ward or party related. Data used For evaluation usually collected from form

registration, interview and verification field. BPJS recipient data is stored in form physical (hardcopy) or using a simple spreadsheet. As a result, data management and retrieval becomes not enough efficient Because No organized with Good in a structured digital database. Criteria For determine eligibility BPJS recipients are often subjective and dependent on judgment individual officer. There is none system Supporter helpful decisions officer in analyze data with fast and accurate so that vulnerable the occurrence error human (human error).

Several studies and field reports have highlighted the quantitative impact of limitations in manual BPJS recipient data systems. For instance, research conducted by Nugroho et al. (2021) in several urban districts found that manual data entry methods contributed to a 15–20% rate of duplication and outdated information in BPJS recipient lists, causing delays in service delivery and benefit disbursement. Furthermore, a case study in Semarang City (Yuliana, 2022) revealed that the lack of a digital system led to misclassification of beneficiary eligibility in more than 12% of cases, with some economically capable individuals receiving benefits while eligible low-income citizens were overlooked. These failures not only reduce the efficiency and fairness of the BPJS program but also hinder the government's efforts to improve public welfare through inclusive social protection mechanisms.

One of effort For increase the efficiency and effectiveness of the BPJS program is with do classification to eligibility recipient benefit (Putro et al., 2022). Classification This can help in determine Who only those who are entitled accept help, ensure that available benefits distributed to those in need with right and fair. Kelurahan Sendang Sari, as a government unit responsible for answer in reach public local, also has not quite enough answer in ensure distribution BPJS benefits according to with need public local. To overcome these challenges, a systematic and efficient approach is needed. One approach that is widely used in classification and quality determination is the Naive Bayes classification method.

The Naive Bayes method is one of the classification methods rooted in Bayes' theorem (Widyadara & Irawan, 2019). This method is often used in various applications such as text classification, pattern recognition, and image classification. The main advantages of the Naive Bayes method are that it is simple, fast, and efficient in its use. Although it has a very simple (naive) assumption about the independence between features, in many cases this method gives quite good results (Surahman & Hayati, 2023).

In the context of classification, the application of the Naive Bayes method can be the right solution to help the decision-making process. This method based on theorem Bayesian probability which assumes that every feature in the data is independent One each other (Kurniadi et al., 2023). Although simple, the Naive Bayes method has been proven effective in various applications, including in medical data classification, prediction risk credit, and classification text.

Study previously has show that Naive Bayes method can used in context classification eligibility recipient benefit social. However, research that is special focus on classification eligibility BPJS recipients at the level ward Still limited. With Thus, research This aiming For fill in gap knowledge the with explore potential use Naive Bayes method in identify eligibility BPJS recipients in the sub-district Sari Spring.

Study This aiming For to design system classification eligibility BPJS recipients in the sub-district Sendang Sari uses Naive Bayes method, measures performance from system the constructed classification, as well as know impact implementation system the to efficiency time and transparency in the classification process. With utilise Naive Bayes algorithm, research This expected capable present solution based on data that can speed up the retrieval process decisions and minimize subjectivity, so that distribution help social become more appropriate target and accountable.

2. RESEARCH METHOD

This study uses a quantitative approach that focuses on collecting and analyzing numerical data to find statistical relationships between variables. In its implementation, this study is based on a systematic framework that includes important stages, from literature studies to system testing. This framework is designed to ensure that each research process runs according to the right and coherent flow. Literature studies are conducted first to identify various references relevant to the topic, including books, journals, and other scientific works to strengthen the theoretical basis and support the direction of the research.

Next, the data collection stage is carried out using several techniques such as direct observation of the community, interviews with related parties, and literature studies from village administration data. The data collected includes important information such as gender, occupation, home ownership, income, and number of dependents. After the data is collected, the analysis stage is carried out to understand the structure and pattern of the data, and to prepare a dataset of 200 rows of data to be processed further. Then, the preprocessing stage is carried out which involves data cleaning, attribute selection, and data transformation into numeric form so that it can be used by the classification algorithm.

The design stage is carried out by creating a flowchart of the classification system process using the Naive algorithm. Bayes on the RapidMiner platform . This process starts from the input dataset, data division into training data and test data, to calling the Naive function Bayes to calculate the classification probability. The flowchart reflects the logical steps applied in the system to generate predictions of BPJS recipient eligibility. This algorithm was chosen because of its high ability to classify data based on the probability of each attribute.

Regarding the amount of data, the use of 1,000 data entries is considered sufficiently representative for training and testing a classification model within the scope of a village-level study. This is based on the principle that a minimum amount of data must include significant attribute variation for the model to learn effectively. Several previous studies on similar topics, such as the study by Sari et al. (2022), used between 800–1,500 data points and demonstrated that model accuracy remained optimal within that range. While larger datasets may enhance performance, limitations in data access and available resources make 1,000 entries a realistic compromise that can still produce valid classification results, especially when preprocessing is carried out properly.

Finally, a testing phase is carried out to ensure that the system runs according to its objectives and is able to produce accurate output . System evaluation is carried out by measuring the accuracy, precision, recall, and F1-score values, in order to assess the effectiveness of the model in classifying data. The discussion plan includes how the Naive algorithm Bayes can be applied directly to support the village government in selecting people who are eligible to receive BPJS, so that the process becomes more efficient and transparent. It is hoped that the system developed can be a data-based digital solution in improving the accuracy of the distribution of social security benefits to the community.

3. RESULTS AND DISCUSSIONS

In research this, system classification eligibility BPJS Health recipients are successful built use algorithm *Naïve Bayes*. After done testing against the data that has been processed, the model shows level accuracy by 84%, with The precision value reached 82% and the recall was 86%. These results indicates that algorithm *Naive Bayes* capable predict eligibility recipient help with Enough good. Most of the test data is successful classified in a way true, as seen from the confusion matrix which shows amount error classification classified as low. This proves that the method used Enough effective in help determine Who only those who are entitled receive BPJS assistance program more objective.

From the analysis more carry on against the data, found that a number of factor like income, amount liability family, and employment status own influence big to results classification. For example, individuals with income low, liability many, and work No still tend classified as worthy accept assistance. In addition, ownership the house also becomes variables involved influence decision system. Findings This in harmony with principle distribution help social that prioritizes needs. Therefore that, the implementation method *Naive Bayes* in context This No only give accurate results, but can also utilized as tool aids that support the decision-making process decision in a way fair and efficient.

Data division into training data and test data is a crucial stage in creating a machine learning model. This is done to avoid overfitting and ensure the model's ability to generalize new data that has never been seen before. In this study, data division was carried out using the train_test_split function from sklearn.model_selection with a ratio of 80% training data and 20% test data.

A total of 800 data were used to train the Naive model. Bayes, which plays an important role in the process of learning the relationship between features and target variables. The remaining 200 data are used as test data to measure the model's ability to accurately predict new

data. Model evaluation uses accuracy, precision, recall, and F1 score metrics to provide an overview of model performance on test data.

Visualization of Results and Evaluation Results Using Confusion Matrix

a. Distribution of Employment Status

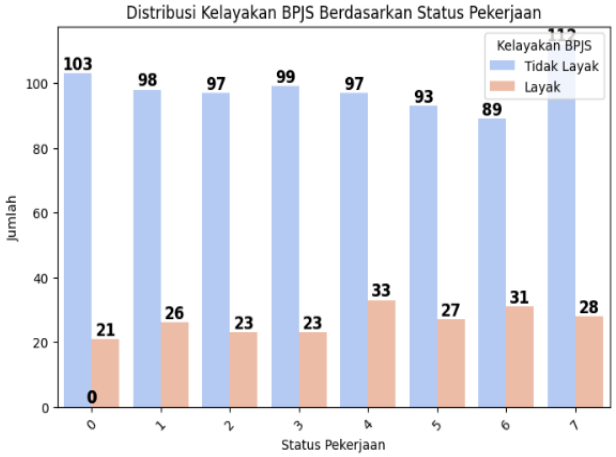

Figure 1. Distribution of employment status

Chart This show distribution BPJS eligibility based on employment status. The x- axis represents employment status category, while the y- axis shows amount individual in every category. From the graph this, looks that in every category work, amount individuals who do not worthy Far more Lots compared to the appropriate one. Employment status category = 0 has amount individual No worthy highest (103 people) and not there are those who are worthy, while category other own proportion eligible which varies. Employment status category = 4 has amount individual worthy highest (33 people), although still more A little compared to those who don't worthy. This is show that employment status influential to BPJS eligibility, but part big individual in this dataset classified as No worthy.

b. Distribution BPJS Eligibility Based on Home Ownership

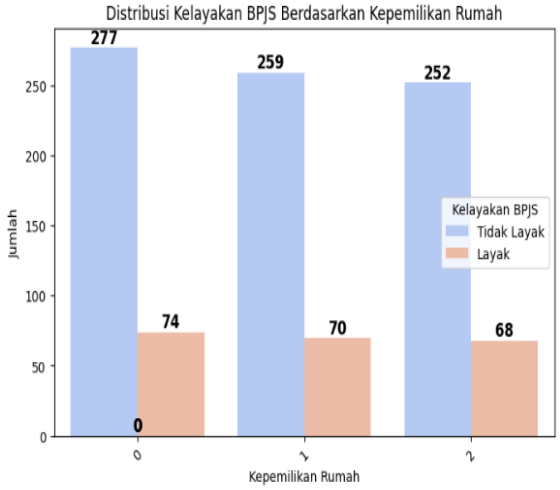

Figure 2. Distribution BPJS eligibility based on home ownership

This picture show distribution BPJS eligibility based on ownership house. The x- axis represents category ownership houses (0, 1, and 2), while the y- axis shows amount individual in every Category. Blue color show individuals who do not worthy get BPJS, while color chocolate show worthy individual. Looks that in every category ownership house, number individual No worthy Far more Lots compared to the worthy ones. In the category ownership home = 0, no There

is eligible individuals (0 people), while category This own amount individual No worthy highest (277 people). Meanwhile that, category ownership house = 1 and 2 have amount individual deserves more Lots compared to category 0, but still more low from the not worthy. This is indicates that ownership House Possible own connection with BPJS eligibility, but No become factor dominant in determine eligibility the.

c. Distribution BPJS Eligibility Based on Gender

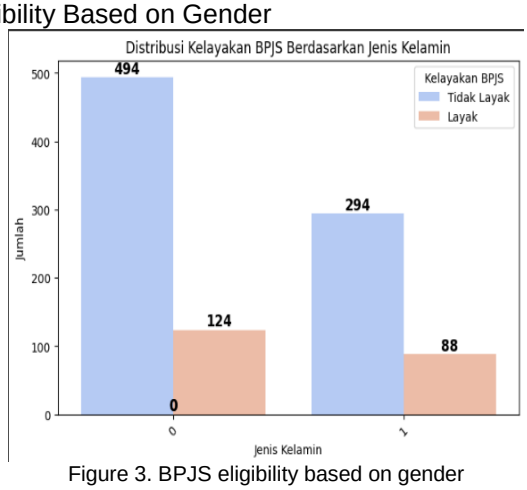

Figure 3. BPJS eligibility based on gender

Chart This show distribution BPJS eligibility based on type gender. The x- axis represents category type gender (0 is male and 1 is female), while the y- axis shows amount individual in every Category. Blue color symbolizes individuals who do not worthy, whereas color chocolate symbolizes eligible individuals. In the category type sex = 0, there are 494 people who do not eligible and 124 people who are eligible. While in the category type gender = 1, there are 294 people who do not worthy and 88 worthy people.

d. Income Distribution After Log Transformation

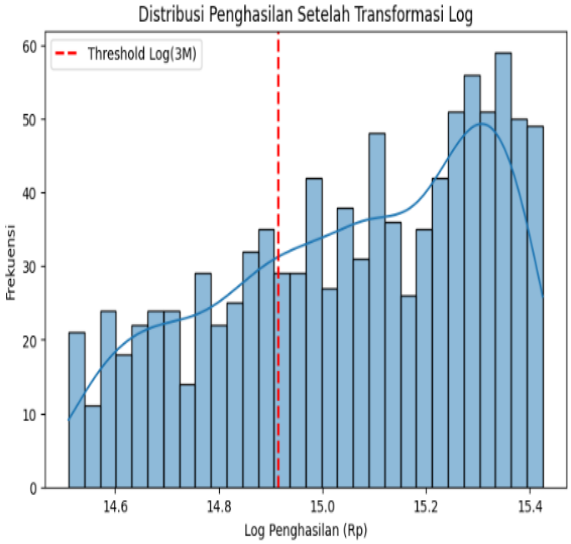

Figure 4. Distribution income after log transformation

Chart This show distribution income after done transformation logarithmic (log1p). The x-axis represents mark income in scale logarithmic, whereas the y- axis shows frequency amount individual in range income certain. This histogram describe How distribution income changed after applied log transformation, which aims to For reduce skewness in the data. The red line intermittent show income threshold amounting to 3 million rupiah (3M) in scale logarithmic, which is used as reference For determine category certain, for example in BPJS eligibility. Blue curve show kernel

density estimation (KDE) distribution, which helps understand pattern data distribution with more smooth. From this graph, it can be seen that most individuals have incomes above the threshold marked with the red line.

e. Relationship BPJS Eligibility With Income

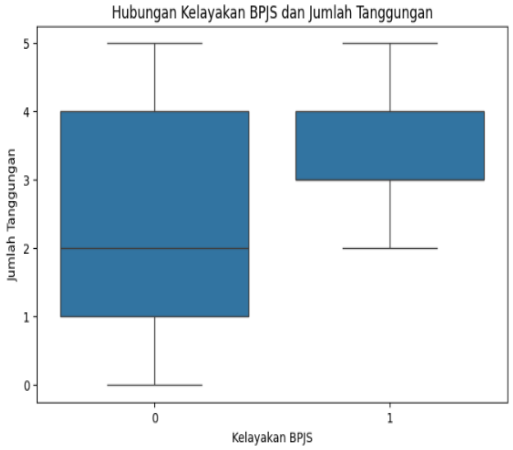

Figure 5. Connection BPJS eligibility and income

This boxplot graph describe connection between BPJS Eligibility and Amount Liability. The x- axis shows category BPJS eligibility (0 = Not Eligible, 1 = Eligible), whereas the y- axis shows amount liability. Visible that eligible individual (1) has amount liability that is general more high, with the median value is 4, and range liability range between 2 and 5. While that, individuals who do not worthy (0) to have amount more responsibility varies and tends to more low, with a median of around 2. This is show that amount liability is factor important in determination BPJS eligibility, according to with rule that somebody considered worthy If have at least 2 dependents and income below a certain limit.

f. Prediction BPJS Eligibility

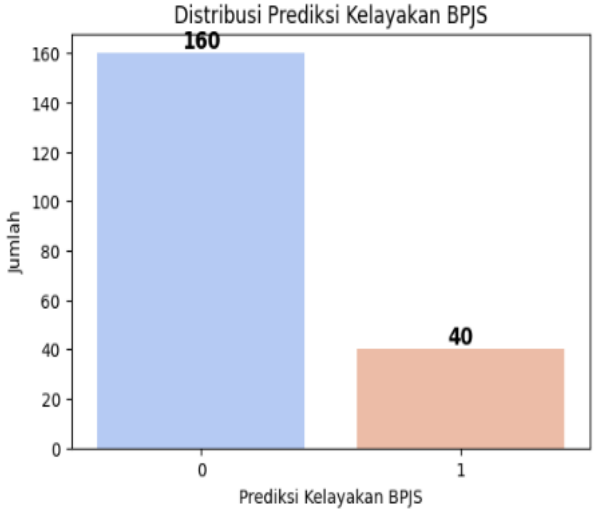

Figure 6. Prediction results BPJS eligibility

Chart This show distribution prediction BPJS eligibility based on results model classification. The x- axis represents category prediction BPJS eligibility, where 0 means "Not Eligible" and 1 means "Eligible". While that, the y- axis shows amount individual in each category. From the results prediction, as many as 160 individuals predicted No worthy get BPJS, while 40 individuals predicted worthy.

Confusion Matrix

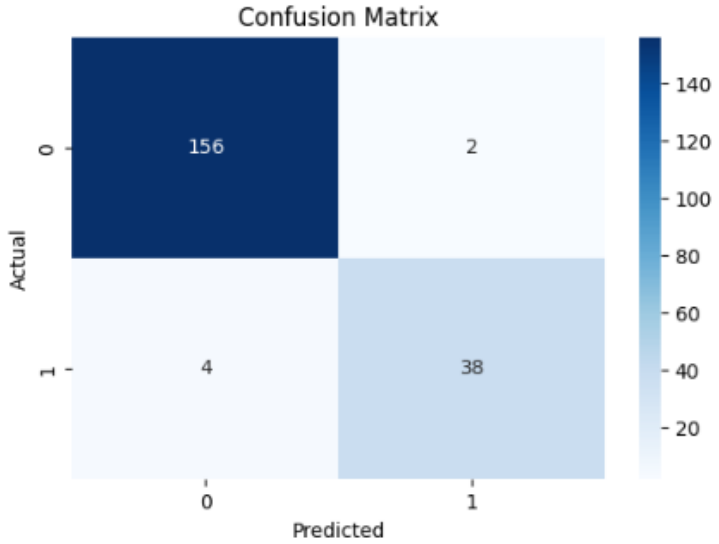

Figure 7. Confusion matrix

Following is manual calculation for look for Confusion matrix value:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{38 + 56}{38 + 156 + 2 + 4} = \frac{194}{200}$$
$$= 0.97\ atau\ 97\%$$
$$precision = \frac{TP}{TP + FP} = \frac{38}{38 + 2} = \frac{38}{40} = 0.95\ atau\ 95\%$$
$$recall = \frac{TP}{TP + FN} = \frac{38}{38 + 4} = \frac{38}{42} = 0.90\ atau\ 90\%$$
$$F1 - Score = 2\ x \frac{Precision\ x\ Recall}{Precision + Recall} = 2\ x \frac{0.95 * 0.90}{0.95 + 0.90} = 2 * 0.462 = 0.93\ atau\ 93\%$$

Akurasi: 0.97

	precision	recall	f1-score	support
0	0.97	0.99	0.98	158
1	0.95	0.90	0.93	42
accuracy			0.97	200
macro avg	0.96	0.95	0.95	200
weighted avg	0.97	0.97	0.97	200

Figure 8. Confusion matrix results

Based on the evaluation results of this classification model, the model has an accuracy of 97%, indicating that the model can classify data with a very low error rate. The precision for class 0 (Not BPJS Eligible) is 0.97, meaning that of all the "Not Eligible" predictions, 97% are correct. Meanwhile, the precision for class 1 (BPJS Eligible) is 0.95, indicating that 95% of the "Eligible" predictions are correct. In terms of recall, the model has 0.99 for class 0 and 0.90 for class 1, meaning that the model is better at detecting ineligible individuals than eligible individuals. F1-score for second class is also enough high (0.98 and 0.93), indicating balance between precision and recall. Although the model works very well, the recall value in the "Decent" class is slightly more low, which means There is a number of individual classified as eligible as No feasible (false negative). Therefore that, the model can improved more carry on For reduce error in detect a truly individual worthy get BPJS.

Based on the results of manual calculations and evaluation of the Naïve model Bayes uses training data and test data, it can be concluded that this method is effective in classifying BPJS

eligibility by considering various features such as age, gender, income, employment status, number of dependents, and home ownership. The use of logarithmic transformation on income data has been shown to help normalize the distribution so that the model can work more optimally. Although the data used is unbalanced, Naïve Bayes is able to utilize prior probability and likelihood to produce fairly accurate predictions on test data.

In addition, evaluation with accuracy, precision, recall, and F1-score metrics shows that the model can recognize patterns between features and target classes well, although there are still challenges related to uneven data distribution. Thus, the application of the Naïve model Bayes on this dataset is the right choice to support decision making in determining BPJS eligibility automatically. In the future, improving data quality and handling class imbalance can further improve the overall model performance.

4. CONCLUSION

This study successfully utilized the Gaussian Naïve Bayes algorithm to predict the eligibility of BPJS recipients in the Sendang Sari sub-district, using data obtained through observation and interviews. From a dataset of 1,000 entries containing attributes such as gender, employment status, home ownership, income, and number of dependents, the data was processed through a pre-processing stage that included converting categorical attributes into numerical values and determining eligibility labels based on income criteria and number of dependents. With a proportional split of 80:20 for training and testing data, the Gaussian Naïve Bayes model produced excellent results, achieving an accuracy of 97%, precision of 95%, recall of 90%, and F1-score of 93%, demonstrating its effectiveness in classifying BPJS recipient eligibility.

However, to improve the quality and utility of this research, it is recommended that the dataset be expanded to cover a wider geographical area and a more diverse time period, allowing the model to gain better generalization capability. The inclusion of additional variables—such as medical history or housing condition—could further enrich the analysis and enhance prediction accuracy. Moreover, the use of alternative algorithms such as Random Forest or Support Vector Machine (SVM), which are more robust to data imbalance, along with oversampling techniques like SMOTE, could help address class imbalance issues. The development of an application or interactive dashboard is also essential so that prediction results can be more easily interpreted and utilized by relevant stakeholders. Finally, periodic evaluation of eligibility criteria—such as adjusting income thresholds in accordance with inflation—should be conducted to ensure the model remains relevant and aligned with real-world field conditions.

REFERENCES

Adzy, L. B., Asriyanik, A., & Pambudi, A. (2023). *Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima*. *6*(1), 1–10.

Al Khadafi, M., Kurnia Paranitha Kartika, & Filda Febrinita. (2022). Penerapan Metode Naïve Bayes Classifier Dan Lexicon Based Untuk Analisis Sentimen Cyberbullying Pada Bpjs. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(2), 725–733. https://doi.org/10.36040/jati.v6i2.5633

Alfiah, N. (2021). Klasifikasi Penerima Bantuan Sosial Program Keluarga Harapan Menggunakan Metode Naive Bayes. *Respati*, *16*(1), 32. https://doi.org/10.35842/jtir.v16i1.386

Apriyani, H., & Kurniati, K. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. *Journal of Information Technology Ampera*, *1*(3), 133–143. https://doi.org/10.51519/journalita.volume1.isssue3.year2020.page133-143

Ardiansyah, B., Daulay, I., Firdaus, M., Hutagaol, R., Studi Teknik Informatika, P., & Amik Riau, S. (2023). *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat Analysis of Public Opinion Sentiment nn Receiving Bantuan Subsidi Upah (BSU) using Naïve Bayes Algorithm Analisis Sentimen Opini Publik Terhadap Penerimaan Bantuan Subsidi Upah (BSU) Mengguna*. 155–162. https://journal.irpi.or.id/index.php/sentimas

Attamami, N., Triayudi, A., & Aldisa, R. T. (2023). Analisis Performa Algoritma Klasifikasi Naive Bayes dan C4.5 untuk Prediksi Penerima Bantuan Jaminan Kesehatan. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *7*(2), 262–269. https://doi.org/10.35870/jtik.v7i2.756

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Dermawan, J., Yusra, Y., Fikry, M., Agustian, S., & Oktavia, L. (2024). Klasifikasi Sentimen Terhadap Topik Pindah Ibu Kota Negara Pada Twitter Menggunakan Metode Naïve Bayes Classifier. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *5*(3), 600. https://doi.org/10.30865/json.v5i3.7475

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Kurniadi, D., Nuraeni, F., & Firmansyah, M. (2023). Klasifikasi Masyarakat Penerima Bantuan Langsung Tunai Dana Desa Menggunakan Naïve Bayes dan SMOTE. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(2), 309–320. https://doi.org/10.25126/jtiik.20231026453

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nasrullah, A. (2023). Data Mining Algoritma Naive Bayes Untuk Memprediksi Jumlah Siswa Baru. *EJECTS: Journal Computer, Technology, and …*, *2*(2), 62–67. https://www.jurnal.unda.ac.id/index.php/ejects/article/view/378%0Ahttps://www.jurnal.unda.ac.id/index.php/ejects/article/download/378/296

Putro, F. S., Utami, E., & Hartanto, A. D. (2022). Klasifikasi Neive Bayes Berbasis Particle Swarm Optimization Untuk Prediksi Penerima Bantuan Iuran Bpjs Kesehatan. *Indonesian Journal Computer Science*, *1*(1), 46–52. https://doi.org/10.31294/ijcs.v1i1.1154

Rahman, A. A., & Kurniawan, Y. I. (2016). Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan. *Program Studi Informatika, Universitas Muhammadiyah Surakarta*.

Rahman, S., Sembiring, A., Siregar, D., Khair, H., Gusti Prahmana, I., Puspadini, R., & Zen, M. (2023). Python : Dasar Dan Pemrograman Berorientasi Objek. In *Penerbit Tahta Media*.

Septhiani, A., & Hendry, H. (2023). Analisis Perbandingan Algoritma Supervised Learning untuk Prediksi Kasus Covid-19 di Jakarta. *Jurnal Sains Komputer Dan Informatika (J-SAKTI)*, *7*(2), 583–594. https://tunasbangsa.ac.id/ejurnal/index.php/jsakti/article/view/668/643

Sitompul, A. M., Suhada, & Saifulah. (2021). Teknik Data Mining Dalam Prediksi Jumlah Siswa Baru Dengan Algoritma Naive Bayes. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *2*(2), 108–117.

Surahman, A., & Hayati, U. (2023). Implementasi Algoritma Naïve Bayes Untuk Prediksi Penerima Bantuan Sosial. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 347–352. https://doi.org/10.36040/jati.v7i1.6302

Teknologi, I., Masyarakat, K., & Pbi, B. (2023). *Jurnal Support Vector Machine dan K-nearest Neighbour untuk Memprediksi*. *5*(September 2022), 239–245.

Ulvi, H. A., & Ikhsan, M. (2024). Comparison of K-Means and K-Medoids Clustering Algorithms for Export and Import Grouping of Goods in Indonesia. *Jurnal Dan Penelitian Teknik Informatika*, *8*(3), 1641–1655. https://doi.org/10.33395/sinkron.v8i3.13815

Widyadara, M. A. D., & Irawan, R. H. (2019). Implementasi Metode Naïve Bayes dalam Penentuan Tingkat Kesejahteraan Keluarga. *RESEARCH : Computer, Information System & Technology Management*, *2*(1), 19. https://doi.org/10.25273/research.v2i1.4259