MATHunesa
Jurnal Ilmiah Matematika
e-ISSN : 2716-506X | p-ISSN : 2301-9115

Volume 13 No 02
Tahun 2025

COMPARISON OF THE PERFORMANCE *OF NAÏVE BAYES* AND *CORRELATED NAÏVE BAYES METHODS* WITH THE APPLICATION OF *SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE*

Radia Sultan
Statistics Department, Faculty of Mathematics and Natural Sciences, Hasanuddin University
e-mail: radiasultan28@gmail.com*

Siswanto Siswanto
Statistics Department, Faculty of Mathematics and Natural Sciences, Hasanuddin University
e-mail: siswanto@unhas.ac.id

Andi Isna Yunita
Statistics Department, Faculty of Mathematics and Natural Sciences, Hasanuddin University
e-mail: andiisnayunita176@gmail.com

Abstrak

Klasifikasi merupakan proses membuat model untuk mengenali pola dengan tujuan memetakannya ke dalam kelas tertentu dan memprediksi kelas. Naive bayes adalah metode klasifikasi yang populer sederhana dan efektif dengan pendekatan probabilistik berdasarkan Teorema Bayes. Asumsi independensi dalam metode ini terkadang membuat kinerja klasifikasi menurun. Correlated naive bayes memperbaiki asumsi ini dengan mempertimbangkan korelasi atribut, sementara SMOTE digunakan untuk mengatasi ketidakseimbangan data. Pendekatan ini penting dalam analisis data medis salah satunya memprediksi penyakit jantung iskemik. Penelitian ini bertujuan untuk membandingkan kinerja metode Naïve Bayes dan Correlated Naïve Bayes dalam klasifikasi penyakit jantung iskemik, dengan penerapan SMOTE untuk mengatasi ketidakseimbangan data. Analisis dilakukan menggunakan data penyakit jantung iskemik di Pusat Jantung Terpadu RSUP Dr. Wahidin Sudirohusodo Kota Makassar, periode Juli 2021 hingga Juli 2022. Naïve Bayes berhasil mengklasifikasikan 66 data dengan akurasi 75%, presisi 94%, dan sensitivitas 62%. Sementara itu, Correlated Naïve Bayes menunjukkan kinerja yang lebih baik dengan mengklasifikasikan 77 data secara benar, menghasilkan akurasi 87,5%, presisi 86%, dan sensitivitas 94%. Hasil ini menunjukkan bahwa Correlated Naïve Bayes memiliki kinerja yang lebih unggul dalam mengklasifikasikan penyakit jantung iskemik.
Kata Kunci: *Correlated Naïve Bayes*, Klasifikasi, *Naive Bayes*, Penyakit Jantung Iskemik, SMOTE..

Abstract

Classification is the process of creating a model to recognize patterns with the aim of mapping them into specific classes and predicting classes. Naive Bayes is a popular, simple and effective classification method with a probabilistic approach based on Bayes' Theorem. The assumption of independence in this method sometimes makes the classification performance decrease. Correlated naïve bayes corrects this assumption by considering attribute correlations, while SMOTE is used to overcome data imbalances. This approach is important in medical data analysis, one of which is predicting ischemic heart disease. This study aims to compare the performance of Naïve Bayes and Correlated Naïve Bayes methods in the classification of ischemic heart disease, with the application of SMOTE to overcome data imbalance. The analysis was carried out using ischemic heart disease data at the Integrated Heart Center of Dr. Wahidin Sudirohusodo Hospital, Makassar City, for the period of July 2021 to July 2022. Naïve Bayes managed to classify 66 data with 75% accuracy, 94% precision, and 62% sensitivity. Meanwhile, Correlated Naïve Bayes showed better performance by correctly classifying 77 data, resulting in 87.5% accuracy, 86% precision, and 94% sensitivity. These results show that Correlated Naïve Bayes has a superior performance in classifying ischemic heart disease.
Keywords: *Correlated Naïve Bayes, Classification, Naïve Bayes, Ischemic Heart Diseases, SMOTE.*

INTRODUCTION

Classification is creating a model or function to identify patterns from historical data that has been studied and labeled (Jiawei & Micheline, 2006). The data classification process aims to map data into classes based on information learned from previous data to predict classes for new data (Aggarwal, 2015). This process groups data based on similar characteristics or attributes (Hart et al., 2000). Several types of classification algorithms are used such as C4.5, k-nearest neighbor classifier, naive bayes, SVM, and others (Nikam, 2015).

One of the popular, simple, and effective classification methods is machine learning techniques naïve bayes, which uses a simple probabilistic classification method based on Bayes' Theorem (Kumar et al., 2019). Naïve bayes assumes the value of an input attribute in a certain class is independent or independent of the values of other attributes (Tan et al., 2021). This independence assumption allows probability calculations to be simpler and more efficient, which makes this method fast and easy to implement even in large data. Naïve bayes also have a variant called gaussian naïve bayes, which is designed to work with numerical data, so this method is very effective for classifying continuous data (Berrar, 2019). However, this assumption also means that each attribute is considered to not influence other attributes, which in reality, is often inappropriate because many attributes in real datasets are interconnected (Arar & Ayan, 2017).

Method naïve bayes has proven effective in a variety of classification applications, including disease classification. Research conducted by Moreno-Ibarra and friends (2021) compared various machine learning algorithms for disease classification and showed that naïve bayes outperforms other methods such as support vector machine, logistic regression, C4.5, deep learning, and neural network in terms of accuracy performance (Moreno-Ibarra et al., 2021). With the simple assumption of attribute independence, naïve bayes can quickly build accurate models, even when the data used has many attributes.

The assumption that each attribute is mutually independent sometimes impairs the performance of naïve bayes in carrying out descending classification. Therefore, there are modifications and developments in the classification naïve bayes to overcome this assumption by adding correlation parameters between attributes to classes known as methods correlated naïve bayes (Mansour et al., 2022). Correlation calculation (R-Square) is carried out to show the relationship between attributes and their classes (Muktamar et al., 2015). By adding correlation parameters, classifying with correlated naïve bayes can learn more complex relationships between attributes and their classes, to improve classification performance and produce more accurate predictions. Research by Muktamar and friends (2015) shows the addition of correlation in the algorithm naïve bayes can significantly increase the level of accuracy on some datasets such as Balance-scale, Iris, Haberman, and Servo (Muktamar et al., 2015). Yulhendri and friends (2023) apply the method correlated naïve bayes in their research to evaluate the cure rate for Hepatitis patients, achieving an accuracy of 86.04% (Yulhendri et al., 2023).

The main problem of machine learning with imbalanced data can significantly harm the performance of most classification algorithms. Most of these algorithms assume a balanced class distribution (He & Garcia, 2009). Synthetic Minority Over-Sampling Technique (SMOTE) is one approach to oversampling that aims to correct class imbalance in the dataset. Data for the minority class will be synthesized so that the data is balanced between data for the majority and minority classes (Sugiyono, 2014). Combination of SMOTE with classification methods naïve bayes can produce better accuracy values, this was shown in research by Sharfina & Ramadhan (2023) which compared the combination of SMOTE with the classification method naïve bayes and random forest For Hepatitis C classification, the results showed that the classification performance with the SMOTE combination was better with accuracy results of 89% and 97% respectively (Sharfina & Ramadhan, 2023).

Data classification plays an important role in data analysis and decision making. One area that can have a big impact on the development of classification techniques is the health sector. Analysis of medical data sets can help in analyzing large amounts of patient data to identify patterns and make predictions about disease progression, including ischemic heart disease (Nayar et al., 2019).

Ischemic heart disease is one of the largest causes of death in the world, causing 16% of total global deaths. This disease occurs due to a lack of oxygen and decreased or absent blood flow to the myocardium due to the narrowing of the coronary arteries (Sukandar et al., 2008). Since 2000, the number of deaths due to ischemic heart disease has increased significantly from 2.7 million to 9.1 million in 2021 (World Health Organization, 2024). South Sulawesi, especially the city of Makassar, established an Integrated Heart Center at RSUP Dr. Wahidin Sudirohusodo to deal with the increase and complexity of heart disease cases, as well as to anticipate a spike in heart and blood vessel disease in eastern Indonesia.

The significant increase in the value of this disease makes classifying patient data important, as early detection and accurate grouping can help in implementing more precise and effective medical interventions. Based on this description, this research will focus on carrying out classification by comparing methods naive bayes and correlated naive bayes. By applying data mining methods, especially classification, patterns and relationships in large data sets can be identified, making it possible to make accurate predictions regarding ischemic heart disease.

THEORITICAL REVIEW

Naïve Bayes Classification

Classification naïve bayes is a machine learning algorithm that uses the principles of Bayes' theorem to carry out classification (Darwis et al., 2021). The algorithm uses naïve bayes, which assumes that all independent attributes are not interconnected given the values of the class variables (Maulana & Yahya, 2019). The way it works is by carrying out statistical calculations by calculating the probability of similarities between previous experiences on a case basis with new cases (Sartika & Sensuse, 2017). In general, Bayes' theorem can be written as in equation (1) (Berrar, 2019).

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (1)$$

where:

X = Attribute

C = Class

$P(C|X)$ = Probability of C occurring based on condition X (*posterior probability*)

$P(C)$ = Probability of C occuring (*prior probability*)

$P(X|C)$ = The probability of X occurring based on the conditions in hypothesis C

$P(X)$ = Probability of X occuring with $P(X) \neq 0$

Maximizing the probability value of each class is carried out in the classification by the naïve bayes method, as shown in equation (2).

$$P(C_i|x_1, x_2, \dots, x_n) = argmax \frac{P(C_i) \prod_{k=1}^{n} P(x_k|C_i)}{P(X)} \quad (2)$$

So that a class that maximize $P(C_i|X)$ is found. Maximum value $P(C_i|X)$ for the class C_i referred to as the maximum posterior hypothesis (Ehsani-Moghaddam et al., 2018). The value of $P(C_i)$ can be calculated using equation (3).

$$P(C_i) = \frac{N_{C_i}}{N} \quad (3)$$

where:

N_{C_i} = Number of training data for the class C_i

N = Number of training data

The value of $P(x_k|C_i)$ can be calculated using equation (4).

$$P(x_k|C_i) = \frac{N(x_k, C_i)}{N(C_i)} \quad (4)$$

where:

$N(x_k, C_i)$ = The number of occurrences of the value x_k in the training data for the class C_i

$N(C_i)$ = Total number of training data for the class C_i

If the data used is continuous, the calculation of the value of $P(x_k|C_i)$ can be calculated as follows in equation (5) (Berrar, 2019).

$$P(x_k|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2} \quad (5)$$

where:

x_k = The value of the k data

μ = Average

σ = Standard deviation

Correlated Naïve Bayes Classification

Classification method using correlated naïve bayes is one of the classification methods from the development of classification methods naïve bayes. Parameters added to the algorithm correlated naïve bayes is the correlation value between attributes and their classes and numbers laplacian. On classification methods correlated naïve bayes This will consider the correlation between the independent variables or attributes (X) on the dependent variable (Y) namely by calculating the correlation value to show the

relationship between attributes and their classes (Muktamar et al., 2015). Number laplacian used to prevent this from happening zero probability or it is called a zero probability value, which usually occurs if there is a data set/sample that does not exist in the training data. Algorithm formula correlated naïve bayes for classification can be written as shown in equation (6) (Hairani et al., 2018).

$$P(C_i|X) = \frac{P(C_i)\prod_{k=1}^{n} P(x_k|C_i)^{\ell} . R(x_k|C_i)}{P(X)} \quad (6)$$

$P(x_k|C_i)^{\ell}$ can be calculated using equation (7):

$$P(x_k|C_i)^{\ell} = \frac{N(x_k, C_i) + \ell}{N(C_i)} \quad (7)$$

where:

$R(x_k|C_i)$ = correlation x_k based on hypothesis C_i

ℓ = *laplacian* number

The correlation coefficient is the number of strong relationships between two or more variables, while the formula for calculating the correlation value can be written as shown in equation (8) (Sugiyono, 2014).

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2)}\sqrt{(n\Sigma Y^2 - (\Sigma Y)^2)}} \quad (8)$$

where:

r = attribute correlation value between classes

n = total data on the dataset

Synthetic Minority Over-Sampling Technique

Synthetic Minority Over-Sampling Technique (SMOTE) is an approach oversampling which aims to correct class imbalance in the dataset (Siringoringo, 2018). Replication of this data is known as synthetic data. The minority class data will be synthesized until the numbers are the same as the majority class data (Muqiit WS et al., 2020). Synthetic data is created based on k-nearest neighbor. Adding minority class data is done by creating synthetic data along the lines connecting one or all k-nearest neighbor minority class data with euclidean distance (Sulistiyono et al., 2021). Euclidean distance use to measure data similarity, it can be written as shown in equation (9) (Habibi & Santika, 2020).

$$d(x_k, y_k) = \sqrt{\sum_{i=1}^{n} (x_{ki} - y_{ki})^2} \quad (9)$$

where:

$d(x_k, y_k)$ = Euclidean distance of the *k* attribute data

x_{ki} = The *k* train data point on the *i* attribute

y_{ki} = The *k* test data point on the *i* attribute

n = Number of attributes

Synthetic data is new data that is generated based on k nearest neighbor. Chosen k the data class closest to the data to be duplicated. Formation of synthetic data through random interpolation using two sample data that can be written as shown in equation (10) (Douzas et al., 2018).

$$x_{syn(ij)} = x_{ij} + \left(x_{knn(ij)} - x_{ij}\right) \times w \quad (10)$$

where:

$x_{syn(ij)}$ = The *i* synthetic data of attribute *j* to be created

x_{ij} = The *i* data of attribute *j* to be replicated

$x_{knn(ij)}$ = The *i* data of attribute *j* that has the closest distance from x_{ij}

w = $0 \leq w \leq 1$

METHODE

Pada The type of data used is secondary data obtained from the Integrated Cardiovascular Center of RSUP Dr. Wahidin Sudirohusodo, Makassar. This data is sample data from patient medical records for one year starting from July 2021 to July 2022. The amount of data used was 230 data. The variables used in this study consisted of a response variable, namely ischemic heart disease status. Consist of eight predictor variables namely leukocytes, history of hypertension, age, hemoglobin, hematocrit, platelets, body weight, and body mass index.

The stages of data analysis carried out in this research are as follows:

1. Data collection from samples of medical records from the Integrated Heart Center of RSUP Dr. Wahidin Sudirohusodo Makassar City in July 2021 to July 2022.
2. Transforming data on categorical variables into numerical form.
3. Scaling data is carried out for all numerical variables using min-max normalization as shown in equation (11):

$$X_* = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (11)$$

where:

X_* = Scaling data

X_i = Preliminary data, $i = 1,2,3 \ldots, n$

X_{min} = Minimum data

X_{max} = Maksimum data

4. Overcoming class imbalance in data with the SMOTE method algorithm (Synthetic Minority Over-sampling Technique). The synthesized

minority data will be added so that the amount of data for both classes is balanced.

5. Divide the data into training data and test data with a ratio of 80:20.
6. Carry out classification using methods naïve bayes.
7. Carry out classification using methods correlated naïve bayes
8. Conduct performance evaluations on classification methods naïve bayes and correlated naïve bayes in classifying test data using confussion matrix. Performance is evaluated based on accuracy, precision and sensitivity values as shown in Equations (12), (13), and (14):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Sensitivity = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (14)$$

where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

9. Comparing the performance between classification methods with naïve bayes and classification methods with correlated naïve bayes. Classification is compared to being better than other classification methods if the performance measurement results obtained are greater.

RESULT AND DISCUSSION

Data Transformation

Data transformation is carried out to simplify the classification process, due to algorithms machine learning generally can only process numeric data. Categorical data cannot be directly used in this algorithm because it does not allow mathematical calculations to be carried out on the text. Therefore, categorical variables need to be transformed into numerical form. In the dataset used, there are two categorical variables, namely the ischemic heart disease status variable and the hypertension history variable. These two variables have two data categories, namely 'Yes' and 'No', so the transformation can be carried out as follows:

No : 0
Yes : 1

Scaling Data

Scaling is an important step before carrying out classification, which aims to normalize the range of values of each variable so that they have the same scale. Process scaling data helps prevent variables with a larger range of values from dominating variables with a smaller range of values. Scaling applied to numeric variables to change their values to a range between 0 to 1. In this process, the method used is min-max normalization. The attributes to be scaled are the variables Leukocytes, Age, Hemoglobin, Hematocrit, Platelets, Body weight, Body mass index. The following example shows the calculation process scaling data.

Variable X_1 (Leukocytes)
Data ke-1

$X_{1*} = \frac{6790-15.7}{37850-15.7} = 0.179$

Balancing Data

Data on the status of ischemic heart disease is divided into two categories, namely data with the status Yes for 220 patients and data with the status No for 10 patients, making it minority data while data with the status Yes is the majority data. The difference in data reached 91.3 percent, which shows a very high figure, so it can be said that the data is unbalanced. Synthetic Minority Over-sampling Technique (SMOTE) is a popular method used to overcome class imbalance in data. By adding synthetic data to minority classes, SMOTE helps improve minority class representation, which in turn improves the performance of machine learning models in recognizing patterns and making more accurate predictions of minority classes.

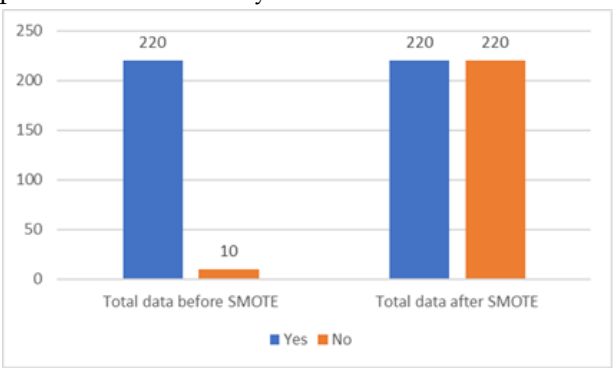

Figure 1. Proportion of data before and after applying the SMOTE method

Based on Figure 1, you can see the proportion of data classes before and after the application of the SMOTE method, where 210 synthetic data were added to the minority class or No status category. Obtained class proportions that are balanced are 50% and 50% with a total of 440 data. The SMOTE method is not only effective in correcting class imbalance but is also flexible in setting oversampling proportions. Can customize a lot of synthetic data added in minority classes according to the specific needs of the data set and data analysis goals. In Figure 3, 100% oversampling is used, namely adding synthetic data so that the amount of minority class data is balanced with majority class data. SMOTE has this flexibility that allows fine tuning to achieve the desired class balance and improve machine learning model performance.

Naïve Bayes Classification

Classification performance is used to show how good the model built during the training process is at predicting data classes. Three metrics, namely accuracy, precision and sensitivity, are used to measure this performance. The results of the performance in the form of a confusion matrix for the 88 test data that have been used are shown in Table 1.

Table 1. Confussion matrix naïve bayes classification

Actual Class	Prediction Class	
	Yes	**No**
Yes	33	20
No	2	33

Based on Table 1, of the 88 test data, 66 test data were correctly predicted into the correct class and 35 test data were predicted incorrectly. The following are the results of measuring performance metrics.

1. Accuracy

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{33+33}{33+33+2+20} = 75\%$$

2. Precision

$$Precision = \frac{TP}{TP+FP} = \frac{33}{33+2} = 94\%$$

3. Sensitivity

$$Sensitivity = \frac{TP}{TP+FN} = \frac{33}{33+20} = 62\%$$

Based on the performance measurement of the classification using the Naïve Bayes method, the accuracy result of 75% represents the percentage of patients with ischemic heart disease status that was correctly predicted by the model, whether classified as "Yes" or "No." The precision of 94% represents the percentage of patients predicted as "Yes" for ischemic heart disease who actually have the disease, from all patients predicted as "Yes." The sensitivity of 62% represents the percentage of patients who actually have a "Yes" status for ischemic heart disease and were correctly predicted by the model, from all patients whose actual data shows a "Yes" status.

Correlated Naïve Bayes Classification

Classification performance is used to show how good the model built during the training process is at predicting data classes. Performance results for 88 test data are shown in the form confussion matrix in Table 2.

Table 2. Confussion matrix correlated naïve bayes classification

Actual Class	Prediction Class	
	Yes	**No**
Yes	50	3
No	8	27

Based on Table 2, of the 88 test data, 77 test data were correctly predicted into the correct class and 11 test data were predicted incorrectly. Performance measurement is carried out in the classification method correlated naïve bayes with three main metrics, namely accuracy, precision and sensitivity.

1. Accuracy

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{50+27}{50+27+8+3} = 87.5\%$$

2. Precision

$$Precision = \frac{TP}{TP+FP} = \frac{50}{50+8} = 86\%$$

3. Sensitivity

$$Sensitivity = \frac{TP}{TP+FN} = \frac{50}{50+3} = 94\%$$

Based on the performance measurement of classification with accuracy, precision, and sensitivity using the Correlated Naïve Bayes method, it was found that the developed model performs well as the performance metrics exceed 85%. The accuracy of 87.5% represents the percentage of all patients whose ischemic heart disease status was correctly predicted by the model, whether classified as "Yes" or "No." The precision of 86% represents the percentage of patients predicted as "Yes" for ischemic heart disease who actually have the disease, from all patients predicted as "Yes." The sensitivity of 94% represents the

percentage of patients who actually have a "Yes" status for ischemic heart disease and were correctly predicted by the model, from all patients whose actual data shows a "Yes" status.

Comparison Of Classification Performance Of Naïve Bayes Method And Correlated Naïve Bayes Method

Based Comparison of classification performance of the two classification methods, namely naïve bayes and correlated naïve bayes, was carried out by comparing the results of the evaluation performance of the two methods, namely accuracy, precision and sensitivity. If a classification has higher performance measurement values, the classification is considered better than other classifications. The results of the comparison of the classification performance of the naïve bayes and correlated naïve bayes methods can be seen in Table 3.

Table 3. Comparison of classification performance

Method	Correct Prediction	Accuracy	Precision	Sensitivity
Naïve Bayes	66	75%	94%	62%
Correlated Naïve Bayes	77	87.5%	86%	94%

Based on Table 3, the classification method with correlated naïve bayes is better than the naïve bayes method in classifying ischemic heart disease at the Integrated Heart Center of Dr. Wahidin Sudirohusodo Hospital, Makassar. The performance measurement results of the correlated naïve bayes method showed a value greater than 85%, which means that the method is very good at classifying. Of the 88 patient test data, the correlated naïve bayes method correctly predicted 77 data while the naïve bayes method managed to correctly predict only 66 data. In the measurement of accuracy, the correlated naïve bayes method is better than the naïve bayes method with a ratio of 87.5% : 75%. In precision measurement, the comparison of the correlated naïve bayes method and the naïve bayes method is 86%: 94%. Meanwhile, in the measurement of sensitivity, the correlated naïve bayes method is better than the naïve bayes method with a ratio of 94% : 62%. From this explanation, the correlated naïve bayes method is proven to be more suitable and better than the naïve bayes method to classify ischemic heart disease patients at the Integrated Heart Center of Dr. Wahidin Sudirohusodo Hospital Makassar. This is in accordance with previous research by Yulhendri et al., who compared the two methods in disease classification, and found that correlated naïve bayes gave more accurate results than naïve bayes with a ratio of 84%:78%. Similar research by Subarkah et al. also showed that correlated naïve bayes with an accuracy of 80.6% were superior to naïve bayes with an accuracy value of 76.5%.

CLOSING

Conclusion

Based on the analysis and discussion results, the following conclusions can be drawn.

1. From 88 data on ischemic heart disease at the Integrated Heart Center of Dr. RSUP. Wahidin Sudirohusodo for the period July 2021 to July 2022 which is used as test data, method naïve bayes can correctly predict 66 data and incorrectly predict 22 data. Mark true positive and true negative each amounting to 33 data, for value false positive and false negative respectively 2 and 20 data. Method correlated naïve bayes can correctly predict 77 data and incorrectly predict 11 data. Mark true positive and true negative 50 and 27 data, respectively, for values false positive and false negative respectively 8 and 3 data.
2. Method correlated naïve bayes better than method naïve bayes in classifying ischemic heart disease data at the Integrated Heart Center of RSUP Dr. Wahidin Sudirohusodo in the period July 2021 to July 2022 this can be seen in the results that correlated naïve bayes predicted more data correctly, namely 77 data, whereas naïve bayes only 66 data. Classification performance results correlated naïve bayes namely 87.5% for accuracy, 86% for precision, and 94% for sensitivity.

Suggestion

In this study, the naïve bayes and correlated naïve bayes methods were carried out to classify, so in the next study, the author suggested to classify using other methods such as support vector machine, k-nearest neighbor, neural network, decision tree and can add a selection of features or attributes before classifying. The researcher also suggested using other data balancing methods such as combine sampling.

REFERENCES

Aggarwal, C. C. (2015). *Data classification*. Springer.

Arar, Ö. F., & Ayan, K. (2017). A feature dependent Naive Bayes approach and its application to the software defect prediction problem. *Applied Soft Computing*, *59*, 197–209.

Berrar, D. (2019). *Bayes' theorem and naive Bayes classifier*.

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131–145.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1–20.

Ehsani-Moghaddam, B., Queenan, J. A., MacKenzie, J., & Birtwhistle, R. V. (2018). Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLoS One*, *13*(12), e0209018.

Habibi, A. M., & Santika, R. R. (2020). Implementasi Algoritma K-Nearest Neighbor dalam Menentukan Jurusan Menggunakan Metode Euclidean Distance Berbasis Web Pada SMP Setia Gama. *SKANIKA: Sistem Komputer Dan Teknik Informatika*, *3*(4), 7–14.

Hairani, H., Nugraha, G. S., Abdillah, M. N., & Innuddin, M. (2018). Komparasi akurasi metode correlated naive Bayes classifier dan naive Bayes classifier untuk diagnosis penyakit diabetes. *InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan*, *3*(1), 6–11.

Hart, P. E., Stork, D. G., & Duda, R. O. (2000). *Pattern classification*. Wiley Hoboken.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Jiawei, H., & Micheline, K. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

Kumar, D. P., Amgoth, T., & Annavarapu, C. S. R. (2019). Machine learning algorithms for wireless sensor networks: A survey. *Information Fusion*, *49*, 1–25.

Mansour, N. A., Saleh, A. I., Badawy, M., & Ali, H. A. (2022). Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy. *Journal of Ambient Intelligence and Humanized Computing*, 1–33.

Maulana, D., & Yahya, R. (2019). Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung Di Indonesia Menggunakan Rapid Miner. *Jurnal SIGMA*, *10*(2), 191–197.

Moreno-Ibarra, M.-A., Villuendas-Rey, Y., Lytras, M. D., Yáñez-Márquez, C., & Salgado-Ramírez, J.-C. (2021). Classification of diseases using machine learning algorithms: A comparative study. *Mathematics*, *9*(15), 1817.

Muktamar, B. A., Setiawan, N. A., & Adji, T. B. (2015). Analisis perbandingan tingkat akurasi algoritma naïve bayes classifier dengan correlated-naïve bayes classifier. *Semnasteknomedia Online*, *3*(1), 1–2.

Muqiit WS, A., Nooraeni, R., Ananda, I. P., Rizki, M. A., & Hapsari, Z. D. (2020). Penerapan Metode Resampling Dalam Mengatasi Imbalanced Data Pada Determinan Kasus Diare Pada Balita Di Indonesia (Analisis Data Sdki 2017). *Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya)*, *8*(1), 19. https://doi.org/10.24252/msa.v8i1.13452

Nayar, N., Ahuja, S., & Jain, S. (2019). Swarm intelligence and data mining: a review of literature and applications in healthcare. *Proceedings of the Third International Conference on Advanced Informatics for Computing Research*, 1–7.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology*, *8*(1), 13–19.

Sartika, D., & Sensuse, D. I. (2017). Perbandingan algoritma klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada studi kasus pengambilan keputusan pemilihan pola pakaian. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *3*(2), 151–161.

Sharfina, N., & Ramadhan, N. G. (2023). Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes. *JOINTECS (Journal of Information Technology and Computer Science)*, *8*(1), 33–40.

Siringoringo, R. (2018). Klasifikasi data tidak Seimbang menggunakan algoritma SMOTE dan k-nearest neighbor. *Journal Information System Development (ISD)*, *3*(1).

Sugiyono. (2014). *Metode Penelitian kuantitatif, kualitatif dan R & D*. Alfabeta.

Sukandar, E. Y., Andrajati, R., Sigit, J. I., Adnyana, I. K., & Setiadi, A. A. P. (2008). Iso Farmakoterapi. *PT. ISFI Penerbitan: Jakarta*.

Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi algoritma synthetic minority over-sampling technique untuk menangani ketidakseimbangan kelas pada dataset klasifikasi. *SISTEMASI: Jurnal Sistem Informasi*, *10*(2), 445–459.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2021). *Introduction to Data Mining, second edition*.

World Health Organization. (2024, August 7). *The top 10 causes of death*. World Health Organization.

Yulhendri, Y., Malabay, M., & Kartini, K. (2023). Correlated Naïve Bayes Algorithm to Determine Healing Rate of Hepatitis Patients. *International Journal of Science, Technology & Management*, *4*(2), 401–410.