Jurnal UJMC, Volume 10, Nomor 1, Hal. 56 - 68
pISSN : 2460-3333 eISSN : 2579-907X

Application of Naïve Bayes Algorithm for Diabetes Prediction

Laili Najla Salsabila[1], Muhamad Riyo Dwi Pangga[2], Syahrul Mauhub Yasser[3], Nabila Arin Riyani[4], Siti Aminah[5], Wahyunengsih[6]

[1]UIN Syarif Hidayatullah Jakarta, lailinajlasalsabilaaa@gmail.com
[2]UIN Syarif Hidayatullah Jakarta, mriyodwipangga@gmail.com
[2]UIN Syarif Hidayatullah Jakarta, syahrulmy27@gmail.com
[3]UIN Syarif Hidayatullah Jakarta, nabilaarin04@gmail.com
[4]UIN Syarif Hidayatullah Jakarta, sitimnh25@gmail.com
[5]UIN Syarif Hidayatullah Jakarta, wahyu.nengsih@uinjkt.ac.id

Abstract. Diabetes is a chronic disease that is considered a significant health problem worldwide. Early detection and prediction of diabetes is a crucial step to enable early intervention and prevent complications. This study aims to apply the Naïve Bayes algorithm in predicting the probability of someone having diabetes. The dataset used in the study was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Attributes such as gender, age, body mass index, glucose level, and others were used as independent variables in the Naïve Bayes algorithm to classify them into two groups: having or not having diabetes. From the research results, it has been shown that the Naïve Bayes algorithm can produce a prediction accuracy of 84.6%, 82.3% precision, and 60.8% recall.

***Keywords*:** *Diabetes, Naïve Bayes, Prediction, Classification*

Abstrak. Diabetes adalah penyakit kronis yang dianggap sebagai masalah kesehatan yang signifikan di seluruh dunia. Deteksi dan prediksi diabetes secara dini merupakan langkah penting untuk memungkinkan intervensi dini dan mencegah komplikasi. Penelitian ini bertujuan untuk menerapkan algoritma Naïve Bayes dalam memprediksi probabilitas seseorang menderita diabetes. Dataset yang digunakan dalam penelitian ini diperoleh dari National Institute of Diabetes and Digestive and Kidney Diseases. Atribut seperti jenis kelamin, usia, indeks massa tubuh, kadar glukosa, dan lainnya digunakan sebagai variabel independen dalam algoritma Naïve Bayes untuk mengklasifikasikan mereka menjadi dua kelompok: memiliki atau tidak memiliki diabetes. Dari hasil penelitian, terbukti bahwa algoritma Naïve Bayes dapat menghasilkan akurasi prediksi sebesar 84,6%, presisi 82,3%, dan recall 60,8%.

***Kata Kunci*:** *Diabetes, Naïve Bayes, Prediksi, Klasifikasi*

1 Introduction

Diabetes is a dangerous illness. It can genuinely impact the body, such as apprehensive framework harm, heart infection, skin contaminations, kidney harm, and hearing misfortune. Unhealthy lifestyles can lead to an increment in the number of individuals with diabetes each year [1]. Hence, early discovery had to be made sometime recently, when diabetes happened, so the analysts conducted several information examinations related to diabetes.

Based on the results of previous studies, diabetes can be analysed using several different methods in each study. First, using six n-gram methods and then evaluated by calculating several validation parameters and found that the model developed using this method has the best performance [2]. Second, using the application of Naive Bayes by processing the original Diabetes dataset into preprocessed data to perform Naive Bayes calculation calculations, this process divides the dataset into training data and test data using the RapidMiner tool. The third is by using Greedy Forward Selection and Naive so it can be concluded that the Naïve Bayes and Greedy Forward Selection algorithms can predict diabetes very well [3].

The gap in this thing is foreseeing diabetes utilizing the Naïve Bayes classification strategy. Naïve Bayes is used to get probabilistic forecasts from information so that the result could be a choice based on past information. Naïve Bayes employs autonomous suspicions with a straightforward likelihood demonstration and Bayes theorem proposed by British researcher Thomas Bayes [4].

The basic theory of this research is the Naïve Bayes Classifier. The credulous Bayes classifier strategy may be a directed strategy that classifies future objects by deciding course names utilizing conditional probabilities. The reason for this method is to classify probabilities based on learning from other probabilities [4]. Gullible Bayes can assess information in real-time and make suggestion frameworks.

Observations that have been made in previous studies have produced some empirical data. Firstly, patient data is obtained from information about each patient, including age, gender, family history of diabetes, body mass index (BMI), blood pressure, blood glucose levels, insulin levels, and others. Laboratory data is obtained from laboratory examinations such as fasting blood glucose levels and oral glucose tolerance [5]. Monitoring data such as patient medical history, laboratory test results, and other relevant clinical measurements are used to develop the prediction model.

Previous research has shown the potential of Naïve Bayes Classifier to predict diabetes. However, this classifier needs to be optimized further to improve its accuracy and reliability in early detection. Diabetes prediction requires analyses from diverse data sources, including patient information, laboratory results, and monitoring data. This research problem aims to investigate integrating these multimodal data sources to improve the prediction capabilities of the Naïve Bayes Classifier [6].

Based on the statement above, the problem formulation is:

1. How does the Naive Bayes algorithm work in predicting diabetes?
2. How effective is the Naive Bayes algorithm in predicting diabetes?

REVIEW OF RELATED LITERATURE

Diabetes is a metabolic disease that occurs when there are high levels of sugar in the body, but it cannot be utilized optimally by the body. These sugar levels should be controlled by the hormone insulin the pancreas produces. However, in people with diabetes, the pancreas cannot produce insulin according to the body's needs [7]. If glucose is not absorbed properly, it will accumulate in the blood, causing various disorders in the body's organs. If uncontrolled, diabetes can lead to life-threatening complications. Naive Bayes Classifier is a Bayes theorem classification method. It predicts future probabilities based on previous experience, assuming strong independence between conditions or events. Examples of naïve Bayes applications include document classification, weather forecasting, spam detection or filtering, and sentiment analysis. This method uses the concepts of probability and statistics proposed by British scientists [8].

The Naive Bayes Classifier is used in a variety of classification tasks. Such as facial and other feature recognition, weather prediction, medical diagnosis for disease risk, and news classification, such as politics or world news [9]. This naive Bayes classifier is easy to understand and implement, making it suitable for beginners. Methods that are often used for data analysis are regression, classification, and clustering. Regression is an analytical technique to identify the relationship between two or more variables. Classification is a technique to classify or categorize some data into a set of discrete classes. Previous research on diabetes data analysis used the logistic Regression method, K-means Clustering for clustering, and Support Vector Machines for classification [10].

2 Method

2.1 Research Methodology

The research methodology used in this research is a quantitative method. Using quantitative methods, this research will collect data from diabetes sufferers regarding variables that can influence diabetes. Next, the data will be analyzed using the Naive Bayes algorithm, assuming that the values between variables are independent of each other in the output value to obtain classification [4]. The classification results will then be evaluated to determine the accuracy and effectiveness of the Naive Bayes algorithm in identifying and predicting diabetes cases [11]. Thus, this research will explain how much the Naive Bayes algorithm can be used as a prediction tool to help diagnose diabetes.

2.2 Data Collection

Participants in this study were a collection of patient data from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). NIDDK is an agency whose mission is to conduct and support medical research and research training and disseminate science-based information about diabetes and other endocrine and metabolic diseases [12]. The study sample population consisted of patients registered with the NIDDK, thus covering a wide range of diabetes data relevant to this study.

The data collection technique in this research is collecting data from the Kaggle website. The data obtained from Kaggle became the basis for the analysis

in this research [13]. From this source, 390 data were collected, with details of 60 positive diabetes patient data and 330 negative diabetes patient data. Each data set includes ten variables relevant to diabetes prediction analysis.

The ethical aspect of this research is the use of sensitive patient data. Ethical aspects related to the use of sensitive health data include data security, transparency, and acknowledgment of data sources [14]. For data security, it seems that it has been protected from the source, namely NIDDK because the data used in this research does not contain sensitive information such as name and place of residence. The bibliography at the end of this research contains the data sources used so that you can verify the information presented.

2.3 Data Analysis

Data analysis in this research uses the Naive Bayes algorithm. The naive Bayes algorithm is an algorithm that uses Bayes' theorem to calculate the probability of a class based on the observed variables. In other words, this algorithm estimates class probability based on the probability of the variables observed in that class. The formula for Bayes' theorem is [15]:

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)} \quad (1)$$

P(A|B) = Opportunity A in condition B
P(B|A) = Opportunity B in condition A
P(A) = Opportunity A
P(B) = Opportunity B

Then, to compare the actual values and predicted values provided by the machine learning model is a confusion matrix. A confusion matrix presents four distinct combinations of predicted and actual values. It includes True Positive, True Negative, False Positive, and False Negative, which characterize the outcomes of the prediction process [16]. Confusion matrices are also helpful in measuring the accuracy, precision, and recall of a machine-learning model.

Tabel 1. Confusion Matrix.

		Actual Value	
		1 (Positive)	**0 (Negative)**
Prediction Value	**1 (Positive)**	True Positive (TP)	False Positive (FP)
	0 (Negative)	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (4)$$

2.4 Limitation of the Study

The research limitations of this study are the limited NIDDK data sources and the limitations of the machine learning model. First, because the data only come from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) with 390 datasets, this may limit the ability to generalize the study results to a broader population [13]. Second, the model has limitations. The predictions used, such as the Naïve Bayes model, have weaknesses, namely that there are assumptions or class conditions independent of each other, so they are less accurate [17]. In contrast, in practice, several conditions usually influence each other [18].

3 Result

3.1 Categories/Theme

In this research, the theme taken is machine learning, especially in the context of the application of the Naive Bayes algorithm in diabetes prediction. Machine learning, as defined by IBM, is a discipline situated within artificial intelligence (AI) and computer science. It focuses on leveraging data and algorithms to replicate human learning processes, with the capacity to refine its accuracy through iterative learning [19]. Within machine learning, there exist two primary categories of learning: supervised learning and unsupervised learning. The Naive Bayes algorithm falls under supervised learning, primarily utilized for classification and probability prediction based on available attributes [20]. Employing statistical methods, machine learning algorithms are trained to categorize or forecast data.

3.2 Interpretation of the data

The dataset used in this research was obtained from NIDDK. The data consists of 390 patient data, 60 diabetic patients, and 330 non-diabetic patients. In addition, this dataset consists of 10 variables that will be used to predict diabetes. Details of the variables are explained in Table 2.

Table 2. Detail Variables.

Variables	Description Variables	State	Description State
Diagnosis (X10)	Disease diagnosis results	1	Diabetes
		2	No Diabetes
Cholesterol (X1)	The amount of cholesterol in the body	1	X1 < 200
		2	200 ≤ X1 < 300
		3	X1 ≥ 300
Gender (X2)	Determines gender	1	Male
		2	Female
Age (X3)	Determines Age	1	X3 < 40
		2	X3 ≥ 40
BMI (X4)	Determines body mass index	1	X4 < 20
		2	20 ≤ X4 < 35
		3	35 ≤ X4 < 50
		4	X4 ≥ 50

Glucose (X5)	Determines the level of sugar in the body	1	X5 < 100
		2	100 ≤ X5 < 150
		3	X5 ≥ 150
HDL (X6)	Determines the amount of good cholesterol in the body	1	X6 < 50
		2	X6 ≥ 50
Pulse Pressure (X7)	Difference between systolic and diastolic blood pressure.	1	X7 < 40
		2	X7 ≥ 40
Mean Arterial Pressure (X8)	Average blood pressure during one cardiac cycle	1	X8 < 85
		2	85 ≤ X8 < 90
		3	X8 ≥ 90
Waist Hip Ratio (X9)	Compare waist circumference with hip circumference	1	X9 < 0,75
		2	X9 ≥ 0,75

The next stage will divide the dataset into training and test data. Training data constitutes a segment of the dataset employed for model training, while test data comprises another segment utilized for model evaluation. This segregation is implemented at an 80:20 ratio, with 80% of the dataset allocated for training the model and the remaining 20% for testing its performance.

The data used at this stage is 80% training data. The training data amounts to 312 data. It is known that the number of patients who do not suffer from diabetes is 275 people, and 37 people who suffer from diabetes. The probability of developing diabetes in 312 data can be seen in Table 3.

Table 3. Probability X10.

Variable	Diabetes	No Diabetes	Chance of Diabetes	Chance of No Diabetes
P(X10)	37	275	0,118589744	0,881410256

Then, probability calculations are performed on the training data. The calculation was conducted on 312 datasets utilizing vectors X1 through X9. The selected attributes for predicting diabetes encompass X1, X2, X3, X4, X5, X6, X7, X8, and X9. Table 4 describes the probability calculations for attributes X1, X2, X3, X4, X5, X6, X7, X8, and X9 which are influenced by X10.

Table 4. Probability Calculation.

Variables	Diabetes	No Diabetes	Chance of Diabetes	Chance of No Diabetes
P(X1 < 200|X10)	14	147	0,378378378	0,534545455
P(200 ≤ X1 < 300|X10)	21	119	0,567567568	0,432727273
P(X1 ≥ 300|X10)	2	9	0,054054054	0,032727273
P(X2=Male|X10)	16	108	0,432432432	0,392727273
P(X2=Female|X10)	21	167	0,567567568	0,607272727
P(X3 < 40|X10)	5	137	0,135135135	0,498181818
P(X3 ≥ 40|X10)	32	138	0,864864865	0,501818182
P(X4 < 40|X10)	0	19	0	0,069090909
P(20 ≤ X4 < 35|X10)	30	213	0,810810811	0,774545455
P(35 ≤ X4 < 50|X10)	6	41	0,162162162	0,149090909
P(X4 ≥ 50|X10)	1	2	0,027027027	0,007272727

P(X5 < 100|X10)	2	219	0,054054054	0,796363636
P(100 ≤ X5 < 150|X10)	9	49	0,243243243	0,178181818
P(X5 ≥ 150|X10)	26	7	0,702702703	0,025454545
P(X6 < 50|X10)	26	155	0,702702703	0,563636364
P(X6 ≥ 50|X10)	11	120	0,297297297	0,436363636
P(X7 < 40|X10)	0	60	0	0,218181818
P(X7 ≥ 40|X10)	37	215	1	0,781818182
P(X8 < 85|X10)	4	41	0,108108108	0,149090909
P(85 ≤ X8 < 90|X10)	3	22	0,081081081	0,08
P(X8 ≥ 90|X10)	30	212	0,810810811	0,770909091
P(X9 < 0,75|X10)	0	10	0	0,036363636
P(X9 ≥ 0,75|X10)	37	265	1	0,963636364

Next, the final stage of creating a prediction model is to use Equation 1 and Table 4. Predictions are made by multiplying the probability values in Table 4 for each attribute to get the predicted probability value. After the probability value is obtained between diabetes and no diabetes, use the highest probability value as the prediction value. The Naïve Bayes model for predicting diabetes can be seen in equations 5 and 6.

P(Diabetes|X) = P(X1=?|Diabetes)...P(X8=?|Diabetes)P(Diabetes) (5)
P(NoDiabetes|X) = P(X1=?|NoDiabetes)...P(X8=?|NoDiabetes)P(NoDiabetes) (6)

The final step is to test the model that has been created. Model testing used 20% of the test data obtained previously. Test data contains 78 data. It is known that the number of patients who do not suffer from diabetes is 55 people, and those who suffer from diabetes are 23 people. The test results of the test data can be seen in Table 5.

Table 5. Data Test Confusion Matrix.

		Actual Value	
		1 (Diabetes)	**0 (No Diabetes)**
Prediction Value	**1 (Diabetes)**	14	3
	0 (No Diabetes)	9	52

3.3 Reliability and validity

An essential thing in predicting diabetes is the reliability and validity of the Naive Bayes algorithm. To ensure the reliability and validity of the Naive Bayes algorithm, accuracy, precision, and recall values are needed from the test data. Accuracy, precision, and recall figures are obtained by entering the values in Table 5 into equations 2, 3, and 4:

$$\text{Accuracy} = \frac{14+52}{14+52+3+9} \times 100\% = 84.6\% \quad (4)$$

$$\text{Precision} = \frac{14}{14+3} \times 100\% = 82.3\% \quad (5)$$

$$\text{Recall} = \frac{14}{14+9} \times 100\% = 60.8\% \quad (6)$$

3.4 Comparison

The comparison of this research with previous research is the difference in variables and algorithms used. In previous research, R. D. Joshi and C. K. Dhakal used Logistic Regression [10], A. Anggrawan and Mayadi used C-means Clustering [22], and N. Mohan and V. Jain used Support Vector Machine [23]. Then there are also previous studies that use different variables, such as Yuni Wardana with eight variables [4] and Ahmad Afif with 17 variables [24]. The current research uses the classification method with Naive Bayes and with ten variables.

3.5 Conclusion of the finding

The conclusion of this research shows that the Naive Bayes algorithm works by calculating a dataset containing relevant variables such as age, BMI, blood sugar levels, blood pressure, and gender, as well as classifying whether the patient suffers from diabetes. This dataset is divided into training data (80%) and test data (20%). The algorithm calculates the prior probability for each class and conditional probability for each variable, then uses Bayes' Theorem to calculate the posterior probability and determine the class with the highest probability. The Naive Bayes algorithm proved quite effective by achieving 84.6% accuracy, 82.3% precision, and 60.8% recall. This means the model can correctly predict 84 to 85 cases out of 100, correctly identify 82 out of 100 positive cases, and detect 61 out of every 100 confirmed cases of diabetes. Thus, the Naive Bayes algorithm is a reliable way to predict diabetes with relatively high accuracy and precision.

4 Discussion

4.1 Conclusion of the finding

The research concludes that the Naive Bayes algorithm is quite effective in predicting diabetes, achieving an accuracy of 84.6%, precision of 82.3%, and recall of 60.8%. This means the model accurately predicts 84 to 85 real cases out of 100, correctly identifies 82 positive cases out of 100, and successfully identifies around 61 out of every 100 diabetes cases [25].

4.2 Comparing the Existing Theories

This research aligns with the theory outlined in the introduction regarding the naive Bayes algorithm. The theory presented by Wardana and Sari states that Naive Bayes is a classification method that employs independent assumptions using a simple probability model with Bayes' theorem proposed by British scientist Thomas Bayes [4]. The advantages of Naïve Bayes include efficiency, fast computation, high accuracy for use with large data containers, and not requiring much data. Another theory by Widianto stated that the main characteristic of the Naive Bayes Algorithm is a very simple (naive) assumption about the independence between each condition or event [8]. The advantage of using Naive Bayes is that this algorithm does not require a lot of training data to determine the necessary parameter estimates during the classification process [27]. Therefore, it can be concluded that this research utilizes the theory outlined in the introduction, namely

applying the naive Bayes algorithm to measure the accuracy of diabetes prediction. However, this is accompanied by modifications to different data variables.

4.3 Theoretical Implication

The implications of the theory obtained in this study are based on applying the theory in previous studies. Research conducted by Ahmad Zaki Arrayyan, Hendra Setiawan, and Karisma Trinanda Putra in the article Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations published in the journal "Semesta Teknika" indicates that Naive Bayes can be effectively used to support diabetes diagnosis through electronic health data analysis. This study shows that the Naive Bayes model can effectively support diabetes diagnosis through the analysis of electronic health data, thereby providing a valuable tool for health professionals in identifying individuals at risk of developing diabetes. [6]. Additionally, research conducted by Yuni Wardana and Devni Primasari, in the article "Prediksi Penyakit Diabetes Dengan Naive Bayes," published in the "Journal of Mathematics UNP," found that the Naïve Bayes Model applied in diabetes prediction research can explain the relationship between diabetes and its symptoms and this model can also effectively identifying individuals at risk of developing diabetes, thereby enabling early intervention and prevention strategies [4]. Similar findings were also revealed in the research conducted by Jwan Kanaan Alwan, Dhulfiqar Saad Jaafar, and Itimad Raheem Ali in the article Diabetes Diagnosis System using modified Naive Bayes Classifier, published in the Indonesian Journal of Electrical Engineering and Computer Science" Research shows that Naive Bayes provides high and optimal accuracy in predicting diabetes. The time required to test predictions is also more efficient [28].

4.4 Generalizability and Transferability

The Naïve Bayes algorithm may unlock insights into diabetes risk across broader populations than the study's sample. The Naïve Bayes algorithm's versatility in application hinges upon its suitability to accommodate new data and its precision across diverse contexts, thereby enabling it to contribute to the prediction of treatment outcomes and patient conditions, leveraging factors like blood pressure, hemoglobin levels, and blood sugar levels [4]. Furthermore, the Naïve Bayes algorithm's simplicity and efficiency make it particularly well-suited for real-time applications in healthcare settings.

4.5 Practical Implication

The practical implication of this research, which highlights the high level of accuracy, precision, and recall in the Naïve Bayes algorithm, is that this model enables early identification of diabetes risk in individuals. This allows for more effective management of diabetes disease, and it facilitates the implementation of accurate and timely prevention and intervention measures. [29]. Furthermore, the early identification of diabetes risk facilitated by the Naïve Bayes algorithm can improve patient outcomes and reduce healthcare costs through targeted preventive measures and lifestyle interventions.

4.6 Alternative Explanation Limitation Study

Research limitations in diabetes prediction research with the Naïve Bayes Algorithm need to be considered. Using an exclusive dataset from the NIDDK Kaggle data may not fully represent the diversity within the overall population, making it difficult to generalize the findings to broader settings [13]. Future research could consider incorporating data from other sources or using more representative sampling methods. Additionally, the Naïve Bayes model may not be sufficient to capture the complexity of diabetes data, so further exploration of more complex models or other machine learning techniques may be warranted. Variables that are unobserved or excluded from the model may also limit understanding of the factors influencing diabetes. Therefore, future research could consider including more relevant variables or conducting more comprehensive analyses.

Furthermore, this study did not consider external variables such as lifestyle and environment, which can significantly impact diabetes risk. Future research could incorporate these external variables into the model to enhance understanding of diabetes risk. Bias in data processing or collection may also influence the accuracy of research findings; therefore, controlling bias through careful study design and meticulous validation is necessary to ensure the reliability of the results.

5 Conclusion and Recommendation

5.1 Conclusion

The results of this research are that the Naive Bayes algorithm can predict diabetes cases with 84.6% accuracy, 82.3% precision, and 60.8% recall. So, we can group and conclude that High Accuracy (84.6%) is a general measure of how well the model performs overall, indicating that the model could classify most of the test data correctly. High Precision (82.3%) is the precision that tells us how many positive predictions are correct, which means the model is good at identifying true positives and avoiding false positives. Low Recall (60.8%) This tells us how many positive cases were determined by the model, which means that the model missed many positive cases (false negatives).

5.2 Recommendation

To overcome the shortcomings of this research and improve the accuracy and applicability of diabetes prediction studies, several improvements should be implemented in future research. First, using data from broader sources, such as populations with diverse backgrounds and geographical locations, can make the study sample more representative. Additionally, exploring more advanced machine learning models, such as ensemble techniques and deep learning, can yield a more specific understanding of the complex interactions between the factors that influence diabetes. Validating the developed models using independent datasets is also crucial to assess their applicability to different populations. By implementing these advancements, future research can create more accurate and generalisable diabetes prediction models, ultimately leading to better patient care and public health strategies.

6 Acknowledgment

Acknowledgments are addressed to our supporting lecturer, Mrs. Dr. Wahyunengsi, M.Pd., who has helped carry out this research, and friends who have supported us in completing this research.

7 Bibliography

[1] Nining Lestari and Burhannudin Ichsan, "Diabetes Melitus sebagai Faktor Risiko Keparahan dan Kematian Pasien Covid-19: Meta-Analisi," *Biomedika*, vol. 13, pp. 83-94, February 2021.

[2] Aisah Mujahidah Rasunah, Erwin Budi Setiawan, and Isman Kurniawan, "Drug Review-based Diabetes Prediction by Using Naïve Bayes Method," in *International Conference Advancement in Data Science, E-learning and Information Systems*, Bali, 2021.

[3] Fitriyani , "Prediksi Diabetes Menggunakan Algoritma Naive Bayes dan Greedy Forward Selection," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 07, pp. 61-69, August 2021.

[4] Yuni Wardana and Devni Prima Sari, "Prediksi Penyakit Diabetes Dengan Naïve Bayes," *Journal Of Mathematics UNP*, vol. 8, pp. 89-97, September 2023.

[5] Fitrokh Nur Ikhromr, Ipin Sugiyarto, Umi Faddillah, and Bibit Sudarsono, "Implementasi Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Naive Bayes dan K-Nearest Neighbor," *Journal of Informatio Technology and Computer Science*, pp. 416-428, May 2023.

[6] Ahmad Zaki Arrayyan, Hendra Setiawan, and Karisma Trinanda Putra, "Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations," *Semesta Teknika*, vol. 27, pp. 28-36, May 2024.

[7] Sherrvy Eva Wijayaningrum. (2023, October) Diabetes - Penyebab, Jenis, Gejala dan Pengobatannya. [Online]. https://www.siloamhospitals.com-/en/informasi-siloam/artikel/diabetes

[8] Mochammad Haldi Widianto. (2019) BINUS University. [Online]. https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/

[9] Trivusi. (2022, September) Pengertian dan Contoh Algoritma Naive Bayes Classifier. [Online]. https://www.trivusi.web.id/2022/07/algoritma-naive-bayes.html? m=1

[10] Ram D. Joshi and Chandra K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *International Journal of Environmental Research and Public Health*, vol. 18, pp. 1-17, July 2021.

[11] Ayuni Fachrunisa Lubis et al., "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Naive Bayes and Decission Tree," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, pp. 44-51, July 2024.

[12] *Mission & Vision - NIDDK* | National Institute of Diabetes and Digestive and Kidney Diseases. [Online]. https://www.niddk.nih.gov/about-niddk/meet-director/mission-vision

[13]"Predict diabetes based on diagnostic measures | Kaggle," https://www.kaggle.com/datasets/houcembenmansour/predict-diabetes-based-on-diagnostic-measures (accessed 02 April 2024).

[14] Arif Rahman. (2020, April) Cyberthreat. Id. [Online]. https://cyberthreat.id/read/6089/Yuk-Simak-Data-data-yang-Termasuk-Sensitif-di-Dunia-Medis

[15] Maulia Indriana Ghani. (2022, January) Zenius. [Online]. https://www.zenius.net/blog/teorema-bayes

[16] Maria Susan Anggreany. (2020, November) School of Computer Science | BINUS University. [Online]. https://socs.binus.ac.id/2020/11/01/-confusion-matrix/

[17] Naomi Chatrina Siregar, Riki Ruli A. Siregar, and M. Yoga Distra Sudirman, "Implementasi Metode Naive Bayes Classifier(NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ)," *Jurnal Teknologia*, vol. 3, pp. 102-110, August 2020.

[18] Alfan Zainal Macfud, Abdi Pandu Kusuma, and Wahyu Dwi Puspitasari, "Analisis Algoritma Naive Bayes Classifier (NBC) Pada Klasifikasi Tingkat Minat Barang Di Toko Violet Cell," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, pp. 87-94, February 2023.

[19] *What Is Machine Learning (ML)* | IBM. [Online]. https://www.ibm.com/topics/machine-learning

[20] Ivan Diryana Sudirman, *Data-Driven Entrepreneur: Bisnis Berdaya Saing dengan Data Science dan Rapid Miner*. Jakarta Selatan: Penerbit Salemba, 2023.

[21] Trivusi. (2022, September) *Data Splitting: Pengertian, Metode, dan Kegunaannya*. . [Online]. https://www.trivusi.web.id/2022/08/data-splitting.html

[22] Anthony Anggrawan and Mayadi, "Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes," *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 22, pp. 405-418, March 2023.

[23] Narendra Mohan and Vinod Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, 2020.

[24] Ahmad Afif, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus di Rumah Sakit Aisyiah," Jurnal Ilmu Komputer dan Matematika, vol. 1, pp. 40-46, 2020.

[25] Sarang Narkhede. (2018, May) Understanding Confusion Matrix | Towards Data Science. [Online]. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[26] Muhammad Ridhwan Galela. (2023, November) Kemenkeu Learning Center. [Online]. https://klc2.kemenkeu.go.id/kms/knowledge/tidak-perlu-confused- dengan-confusion-matrix-728befa0/detail/

[27] Khafifah Munawaroh and Alamsyah, "Performance Comparison of SVM, Naive Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter," *Journal of Advances in Information Systems and Technology*, pp. 113-125, October 2022.

[28] Jwan Kanaan Alwan, Dhulfiqar Saad Jaafar, and Itimad Raheem Ali, "Diabetes Diagnosis Systems Using Modified Naive Bayes Classifier," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, pp. 1766-1774, September 2022.

[29] Turke Althobaiti, Saad Althobaiti, and Mahmoud M. Selim, "An Optimized Diabetes Mellitus Detection Model for Improved Prediction of Accuracy and Clinical Decision-Making," *Alexandria Engineering Journal*, pp. 311-324, March 2024.