Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

An Optimize Weights Naïve Bayes Model for Early Detection of Diabetes

Oman Somantri[1], Ratih Hafsarah Maharrani[2] and Linda Perdana Wanti[3]

[1.2,3] Department of Informatics, Politeknik Negeri Cilacap, Cilacap, Central Java, Indonesia
E-mail: oman_mantri@yahoo.com[1], ratih.hafsarah@pnc.ac.id[2], lindaperdana16@gmail.com[3]

ARTICLE INFO

History of the article:
Received Juny 06, 2021
Revised January 25, 2022
Accepted February 22, 2022

Keywords:
optimize weights, naïve bayes, diabetes, genetic algorithm

Correspondece:
E-mail:
oman_mantri@yahoo.com

ABSTRACT

This research proposes a method to optimize the accuracy of the Naïve Bayes (NB) model by optimizing weight using a genetic algorithm (GA). The process of giving optimal weight is carried out when the data will be input into the analysis process using NB. The research stages were conducted by preprocessing the data, searching for the classic naïve Bayes model, optimizing the weight, applying the hybrid model, and as the final stage, evaluating the model. The results showed an increase in the accuracy of the proposed model, where the naïve Bayes classical model produced accuracy rate of 87.69% and increased to 88.65% after optimization using GA. The results of the study conclude that the proposed optimization model can increase the accuracy of the classification of early detection of diabetes.

INTRODUCTION

Diabetes is a disease that can increase the risk of blood vessel damage, kidney disease, heart disease, blindness, stroke, nerve damage and birth defects. One of the causes of this diabetes is when the pancreas does not release enough fluid. Diabetes is a health problem that is currently experienced by most people; this is caused by various things, one of which is the person's diet and lifestyle. (Mok et al., 2021).

One of the efforts to prevent this disease is by early detection of diabetes. This early detection is a solution that can be used as a reference for anyone who is indicated to have these symptoms; therefore it can be treated early. The level of knowledge and education in the assessment and prevention of diabetes is an important point; this will have an impact on the ways that each person experiencing these symptoms takes preventive measures. Currently, along with the development of information technology, various technologies are being developed in the application of diabetes detection, one of which is data mining (Shivakumar & Alby, 2014) .

Data mining provides a workable solution currently, in which the implementation of a model is gained by using certain machine learning algorithms; therefore the model obtained can be a decision support. Currently there are various machine learning applications used in predicting or detecting diabetes according to existing data, such as decision tree (DT), neural network (NN), support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF), naïve bayes (NB), deep learning and other methods (Anwar et al., 2020; Mujumdar & Vaidehi, 2019; G. Tripathi & Kumar, 2020).

Naïve bayes is one of the best machine learning tools that currently are widely used for classifying data. One of the advantages of this NB is that the data does not require a long analysis process so it is

efficient. Moreover, NB is quite good to be used for classification data analysis with insufficient amount of data. Through its advantages, NB is currently widely used for sediment analysis (Somantri & Apriliani, 2019), education (A. Tripathi et al., 2019), data security (Shrivastava et al., 2019), in addition to prediction and detection of diabetes.

There are several previous studies that apply Naïve Bayes for the prediction and detection of diabetes. The research was conducted by (Permana & Patwari, 2021). Their research applying NB and Decision Tree resulted in the best level of accuracy was using NB with the best level of accuracy. Subsequent research was carried out by (Ridwan, 2020), by applying NB for the classification of Diabetes Mellitus (DM).

Another study tried to compare a model using the ID3 algorithm with Naïve Bayes for the classification of diabetes mellitus. In this research, the best model was produced by using NB with an accuracy rate of 76% (Nurdiana & Algifari, 2020). Slightly different from the research conducted in machine learning tree classifiers (Vigneswari et al., 2019).

Based on the advantages it has, NB still has shortcomings, namely it still has a level of accuracy that is not quite significant from the model applied. Even so, the level of accuracy is still adequate but still requires an increase in accuracy and optimization of the model. Model optimization that can be done is by optimizing the weight value, thus resulted in a different accuracy. As one of the algorithms in the optimization method, the Genetic algorithm (GA) can provide a reliable solution so that the proposed Naïve Bayes model is better and optimal. The current genetic algorithm can be relied on to perform the best weight search in NB models. However, GA is a model that has been widely used by previous researchers as a model that can provide increased accuracy in the fields of health, electricity, technology, and others. (Wang & Sobey, 2020).

The research contribution in this article is an effort to improve the performance of a different model from previous studies. The difference in the research conducted in this article is to improve the performance of the model's accuracy level by applying an optimization algorithm to get the classification accuracy level for early detection of diabetes, namely using GA. Unlike previous studies, the focus of GA in this study is proposed to do the best weighting for each attribute to support the increase in the accuracy of the naïve Bayes model by using a dataset which consist of different attributes. The research objective is to apply a genetic algorithm to optimize the Naïve Bayes model therefore lead to an increase in accuracy in the classification of early detection of diabetes, in the hope that this model can be used as a decision support for those with an interest in detecting and assessing diabetes symptoms.

RESEARCH METHODS

1. Dataset and Tools

The experiment for this study use a windows 8 operating system device, an Intel core i7 processor, and 8 GB of memory. The analysis of the research data in the search for the proposed model use Rapidminer Studio version 9 software. This research dataset was taken from the UCI machine learning repository, which is in the form of early-stage diabetes risk prediction dataset (M M Faniqul Islam; Rahatara Ferdousi, 2020). Furthermore, this data was first published in 2020. Research data is data taken from Sylhet Diabetes Hospital Bangladesh (Islam et al., 2020). The dataset is a multivariate data type consisting of 17 attributes with 520 instances of data.

The attributes used in this study were age, sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle

stiffness, alopecia, and obesity. There is one attribute in a label form used as an answer to early detection of diabetes, namely a class filled with “positive” and “negative” answers.

2. Proposed Method

In this study, we propose a model for the classification of the early assessment of diabetes. A model based on the optimized Naïve Bayes algorithm is used. Optimization is carried out as an effort to improve classification accuracy; this gives a significant contribution to the resulting model.

The first stage of this experiment is to carry out the data preprocessing namely the process of labeling attribute where the target set role of the attributes is determined. At this stage, the "class" attribute is used as a label. The next process is to perform data analysis by applying the Naïve Bayes algorithm as the proposed model. The process of model optimization is carried out by applying a genetic algorithm to optimize the weight of each weight attribute used.

The data validation stage was carried out to see how far the level of accuracy can be obtained by the proposed model. This study used the cross-validation method for data validation, by first dividing the dataset used into training and testing data. Furthermore, data sharing was carried out with a percentage of 90% training data and 10% testing data.

In the last stage of this research, a comparison of the accuracy level resulted from the proposed model was carried out after optimization using a model that had not been optimize to be able to get an overview of the success rate. The framework for the research stages of the proposed model is shown in Figure 1.

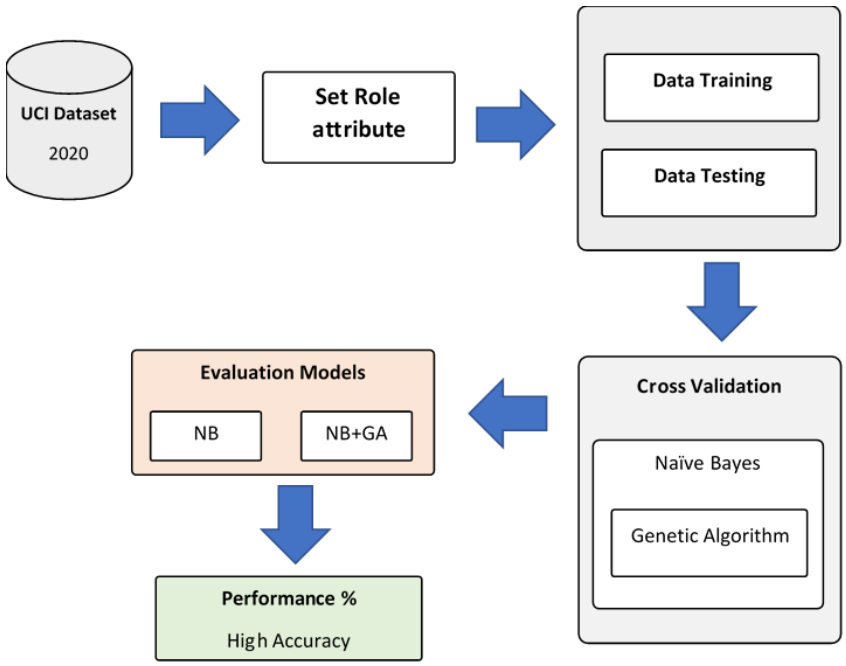

Figure 1. Proposed model Framework

In the process of finding a model with the best performance accuracy level using confusion matrix (Kotu & Deshpande, 2019), equations (1), (2), and (3) are used. In the equation for TP is *True Positive*, FP is *False Positive*, FN is *False Negative*, and TN is *True Negative*.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

3. Naïve Bayes & Genetic Algorithm

Naïve Bayes (NB) is a machine learning algorithm developed from the development of the Bayes theorem by utilizing a statistical and probabilistic calculation. The concept of NB itself is predicting a probability in the future based on previous experiences. For the NB equation, it can be seen in the equation (4) (Friedman et al., 1997). For P(*H*) is the prior probability of class *H*, P(*x*) is the prior probability of predictor x, P(*H* | *x*) is the posterior probability of class *H* given predictor x, and P(*x*|*H*) is the likelihood probability of predictor *x* given class *H*.

$$P\ (H|x) = \frac{P(x|H)\ P(H)}{P(X)} \quad (4)$$

Genetic Algorithm (GA) is an optimization algorithm developed to enable the search for the optimal value of the optimization process against existing models. (Melanie, 1996).

RESULTS AND DISCUSSION

This research has resulted in a model that has a level of accuracy in accordance with the parameter values that have been determined previously. It should be noted that the parameter values of the naïve Bayes model and the genetic algorithm are set manually; this gives an advantage during the experiment to be able to know more about every change in the level of accuracy of the model being sought.

1. Results of the Naïve Bayes Model Experiment

Experiments on each model were carried out by providing comparisons of each validation stage. The experimental stages were conducted using Naïve Bayes and each stage was done using cross validation with different parameters of the number of folds for each experiment, namely using 5 and 10.

In the experimental classification of diabetes, early detection done by using Naïve Bayes with Fold = 5 produces results as in Table 1. Experimental results using Fold = 10 can be seen in Table 2, and the graphical representation between the two folds is depicted in Figure 2. The highest level of accuracy obtained amounted to 87.69%.

Table 1. Naïve Bayes with k-Fold=5

Sample Type	*Accuracy*	*Precision*	*Recall*
linear	83.85%	60.75%	82.50%
shuffled	87.12%	79.92%	88.71%
stratifiied	**87.69%**	80.91%	89.50%

Table 2. Naïve Bayes with k-Fold=5

Sample Type	*Accuracy*	*Precision*	*Recall*
linear	86.92%	58.41%	89.00%
shuffled	**87.69%**	80.13%	90.24%
stratifiied	87.69%	81.27%	90.00%

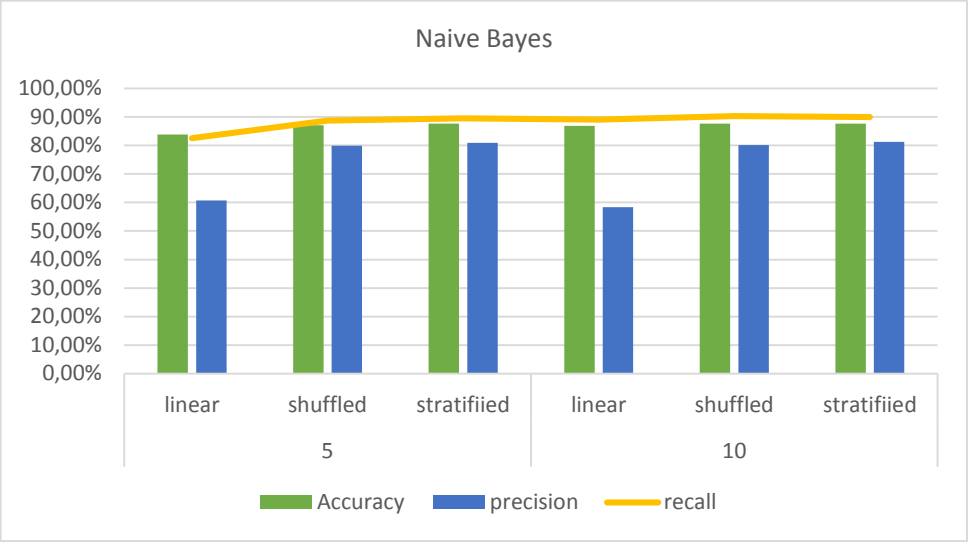

Figure 2. An accuracy result of Naïve Bayes model

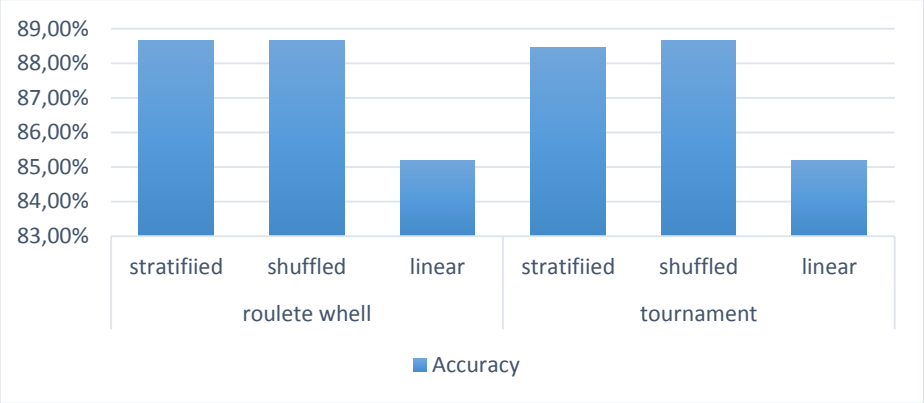

Figure 3. Result of NB+ GA models with k-Fold=5

2. Naïve Bayes with Optimization

Another experimental stage is carried out is to optimize the model. The optimization stage is conducted by optimizing the attribute weights of the existing model, namely Naïve Bayes. The method used in this optimization is a genetic algorithm.

The results of the experiment using GA optimization previously determined the k-Fold parameter = 5, then the number of populations was determined at GA = 5. The parameter selection scheme settings are then set into two schemes, namely roulette wheel and tournament, resulting in an accuracy level as shown in Table 3, Table 4, and Figure 3. Experiments in this model show the highest level of accuracy as high as 88.65%.

Table 3. NB + GA using roulette wheel with k-Fold=5

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	**88.65%**	82.06%	90.50%
shuffled	88.65%	82.25%	89.77%
linear	85.19%	61.12%	85.00%

Table 5. NB + GA using tournament with k-Fold=5

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	88.46%	81.55%	90.50%
shuffled	**88.65%**	81.97%	90.24%
linear	85.19%	61.12%	85.00%

In the next experimental stage with k-Fold = 5, optimization is carried out with GA settings for the population = 10 parameter, the results are shown in Table 6, Table 7, and Figure 4. The highest level of accuracy in this model is 88.65%.

Table 6. NB+GA using roulette wheel with population=10

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	88.65%	82.21%	90.50%
shuffled	88.46%	81.56%	89.69%
linear	85.19%	61.12%	85.00%

Table 7. NB+GA using tournament with population=10

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	**88.65%**	81.92%	90.50%
shuffled	88.65%	82.19%	90.01%
linear	85.19%	61.12%	85.00%

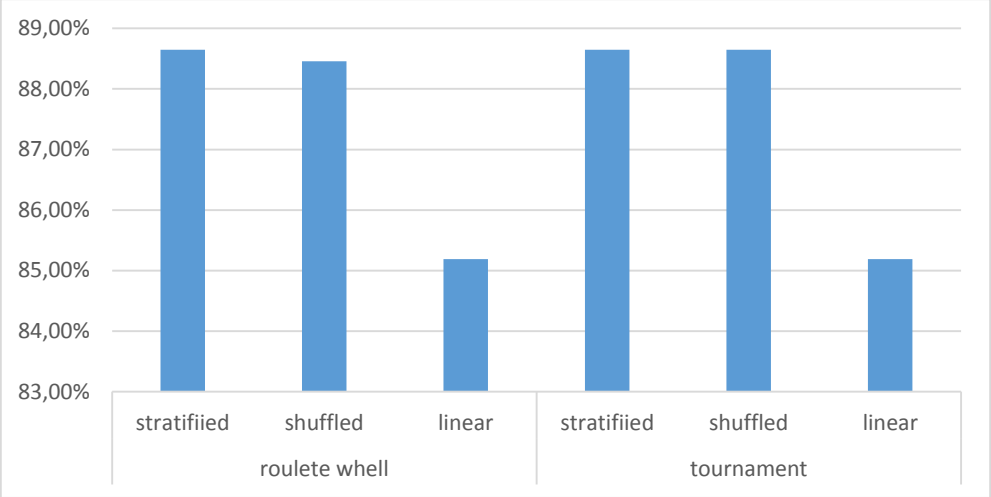

Figure 4. Result of NB+ GA models with k-Fold=5, and population=10

To get a model with a better level of accuracy, it is redone by setting other parameters, namely the k-Fold parameter = 10. In addition, the population parameter in GA is still set with values of 5 and 10. The results of model optimization using the roulette wheel and tournament selection scheme show the highest level of education at 88.27% and it can be seen in Table 8 and Table 9. For the experimental results using population = 10 in GA, the highest accuracy value is 88.46% as shown in Table 10 and Table 11.

Table 8. NB + GA using roulette wheel with population=5, k-fold=10

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	88.08%	81.81%	90.00%
shuffled	88.27%	81.73%	90.33%
linear	87.31%	58.41%	89.50%

Table 9. NB + GA using tournament with population=5, k-fold=10

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	88.27%	82.04%	90.00%
shuffled	88.08%	81.28%	89.91%
linear	87.31%	58.41%	89.50%

Table 10. NB + GA using roulette wheel with population=10, k-fold=10

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	88.27%	81.80%	89.50%
shuffled	88.27%	81.41%	89.93%
linear	87.31%	58.41%	89.50%

Table 11. NB + GA using tournament with population=10, k-fold=10

Sampling	*Accuracy*	*Precision*	*Recall*
stratifiied	88.46%	82.82%	90.00%
shuffled	88.46%	81.25%	89.70%
linear	87.31%	58.41%	89.50%

The result of the research can be presented in the form of tables, graphs or figures. They can be compiled with written text to build a discussion of the findings, that is about the new, the modification or the established theory.

3. Evaluation Models

The experimental results show a value with a different level of accuracy, this happens because of the changes made in parameter values. Model optimization is an effort to increase the level of accuracy in the classification of early detection of diabetes. Based on the experimental results, a hybrid model with different accuracy values has been obtained, as shown in Table 12 and Figure 5.

Table 12. NB and NB+GA Model Comparison

Model	*Accuracy*	*Precision*	*Recall*
Naïve Bayes	87.69%	80.91%	89.50%
NB+GA (proposed)	**88.65%**	81.92%	90.50%

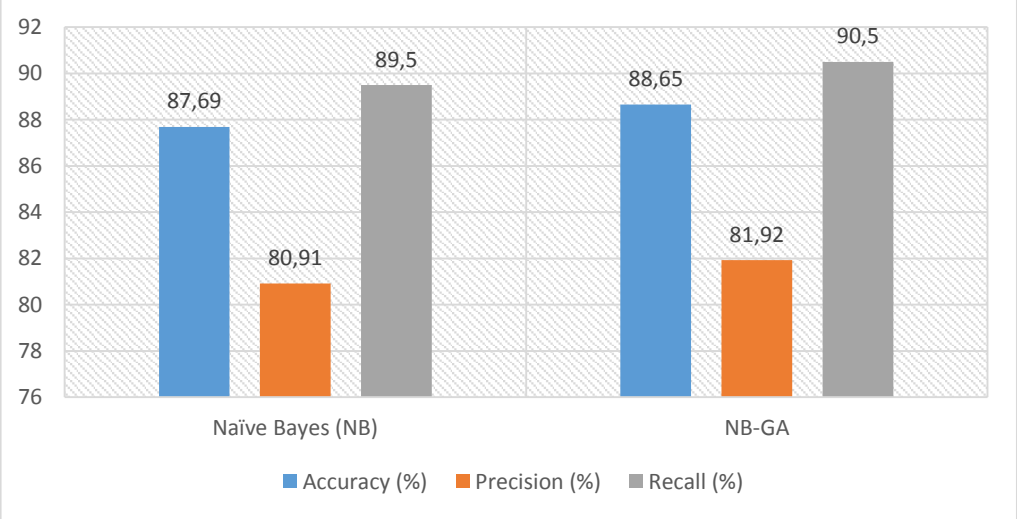

Figure. 5 Evaluation of a proposed models

The optimization of the model carried out in this study provides a change in the level of accuracy. The highest level of accuracy produced using Naïve Bayes is 87.69%, while the NB + GA hybrid model produces the highest accuracy rate of 88.65%, experiencing an increase in accuracy of 0.96%.

Based on the research, there are deficiencies in the proposed model which is the level of accuracy that requires more significant improvement. This increase can be done if the experiment is carried out using other methods with appropriate and precise parameter values.

CONCLUSIONS AND RECOMMENDATIONS

Optimized weights applied to Naïve Bayes using genetic algorithms have succeeded in increasing the accuracy of the proposed model. The best accuracy rate for the NB-GA hybrid model as the classification of early detection of diabetes that was proposed was 88.65%. For future research, it is recommended to apply the model using other algorithms, thus different levels of accuracy will be obtained. The search for parameter values in this study plays an important role, therefore in addition to the estimated weight; an optimization that focuses on optimizing the parameter values of the Naïve Bayes model and other models is needed.

REFERENCES

Anwar, F., Qurat-Ul-Ain, Ejaz, M. Y., & Mosavi, A. (2020). A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey. *Informatics in Medicine Unlocked*, *21*, 100482. doi: 10.1016/j.imu.2020.100482

Candra Permana, B. A., & Dewi Patwari, I. K. (2021). Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes. *Infotek : Jurnal Informatika Dan Teknologi*, *4*(1), 63–69. doi: 10.29408/jit.v4i1.2994

Friedman, N., Geiger, D., Goldszmidt, M., Provan, G., Langley, P., & Smyth, P. (1997). Bayesian Network Classifiers *. *Machine Learning*, *29*, 131–163. https://doi.org/10.1023/A:1007465528199

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Advances in Intelligent Systems and Computing*, *992*, 113–125. doi: 10.1007/978-981-13-8798-2_12

Kotu, V., & Deshpande, B. (2019). Model Evaluation. In *Data Science* (pp. 263–279). Elsevier. doi: 10.1016/b978-0-12-814761-0.00008-3

M. M. Faniqul Islam; Rahatara Ferdousi. (2020). *UCI Machine Learning Repository: Early stage diabetes risk prediction dataset. Data Set*. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Melanie, M. (1996). An introduction to genetic algorithms. *Cambridge, Massachusetts London, England.* doi: 10.1016/S0898-1221

Mok, C. H., Kwok, H. H. Y., Ng, C. S., Leung, G. M., & Quan, J. (2021). Health State Utility Values for Type 2 Diabetes and Related Complications in East and Southeast Asia: A Systematic Review and Meta-Analysis. *Value in Health*. doi: 10.1016/j.jval.2020.12.019

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. doi: 10.1016/j.procs.2020.01.047

Nurdiana, N., & Algifari, A. (2020). Studi Komparasi Algoritma Id3 Dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *INFOTECH Journal*, *6*(2), 18–23. https://ejournal.unma.ac.id/index.php/infotech/article/view/816

Ridwan, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *4*(1), 15–21. doi: 10.47970/siskom-kb.v4i1.169

Shivakumar, B. L., & Alby, S. (2014). A survey on data-mining technologies for prediction and diagnosis of diabetes. *Proceedings - 2014 International Conference on Intelligent Computing Applications, ICICA 2014*, 167–173. doi: 10.1109/ICICA.2014.44

Shrivastava, R. K., Ramakrishna, S., & Hota, C. (2019). Game theory based modified naïve-bayes algorithm to detect DoS attacks using Honeypot. *2019 IEEE 16th India Council International Conference, INDICON 2019 - Symposium Proceedings*, 1–4. doi: 10.1109/INDICON47234.2019.9030355

Somantri, O., & Apriliani, D. (2019). Opinion mining on culinary food customer satisfaction using naïve bayes based-on hybrid feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, *15*(1), 468–475. doi: 10.11591/ijeecs.v15.i1

Tripathi, A., Yadav, S., & Rajan, R. (2019). Naive Bayes Classification Model for the Student Performance Prediction. *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2019*, 1548–1553. doi: 10.1109/ICICICT46008.2019.8993237

Tripathi, G., & Kumar, R. (2020). Early Prediction of Diabetes Mellitus Using Machine Learning. *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, 1009–1014. doi: 10.1109/ICRITO48877.2020.9197832

Vigneswari, D., Kumar, N. K., Ganesh Raj, V., Gugan, A., & Vikash, S. R. (2019). Machine Learning Tree Classifiers in Predicting Diabetes Mellitus. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, 84–87. doi: 10.1109/ICACCS.2019.8728388

Wang, Z. Z., & Sobey, A. (2020). A comparative review between Genetic Algorithm use in composite optimisation and the state-of-the-art in evolutionary computation. *Composite Structures*, *233*, 111739. doi: 10.1016/j.compstruct.2019.111739