A proposed model using Naïve Bayes and generalized linear models for early detection of heart attack risk

Oman Somantri, Linda Perdana Wanti
Cyber Security Engineering, Department of Computer and Business, State Polytechnic of Cilacap, Cilacap Regency, Indonesia

Article Info

Article history:

Received Oct 24, 2023
Revised Dec 9, 2023
Accepted Dec 16, 2023

Keywords:

Early detection
Generalized linear model
Genetic algorithm
Heart attack
Naïve Bayes

ABSTRACT

Timely identification of diseases, particularly heart attacks is crucial for individuals, particularly the elderly, to accurately anticipate the onset of the disease based on symtoms. The objective of this study is to develop a highly accurate model for early detection of heart disease using the Naïve Bayes (NB) and generalized linear model (GLM) techniques. In addition, another concern is the model's subfar accuracy levels, promting the implementation of measures to optimize it. The suggested approach fot optimization involves the utilization of a genetic algorithm (GA). The research findings indicate that the NB and GLM approaches achive a reasonably high level of accuracy. Specifically, the NB model achieves an accuracy of 82.53%, while the GLM achieves an accuracy of 84.50%. Following optimization, the accuracy levels notably improved, with the NB_M-GA model reaching 85.83% and the GLM_M-GA model achieving 86,48%.

Corresponding Author:

Oman Somantri
Cyber Security Engineering, Department of Computer and Business, State Polytechnic of Cilacap
Dr. Soetomo street No. 01, Sidakaya, Cilacap, Indonesia
Email: oman_mantri@yahoo.com

1. INTRODUCTION

According to data published by the WHO, heart disease is a prominent cause of mortality on a global scale. The data for 2021 revealed a total of 17.8 million deaths attributed to heart disease, accounting for one-third of all global deaths. Myocardial infarctions are a medical condition that necessitates specific care, and the enhancement of preventive measures can be further optimized. Efforts must be undertaken to ensure the importance of disease prevention and early detection. Due to technological advancements ini data mining and machine learning, it is now feasible to provide utilize this technology for early detection of heart disease and heart failure [1], [2].

Data mining is extensively utilized not only for detecting heart attacks but also for indentifying diseases such as diabetes [3], [4] and Parkinson's disease [5], among others. Utilizing machine learning for early detection of heart attacks risk is a technique than can facilitate the prompt recognition of heart attack and implementation of preventive measures. Machine learning algorithms have offered numerous solutions for users, particular focus on hearth disease. Commonly employed algorithms encompass neural networks [6], support vector machines [7], Naïve Bayes (NB), linear regression, deep learning [8], convolutional neural networks [9], deep learning [10], and diverse other techniques. Data mining, wich utilizes machine learning, offers a solution due to its inherent advantages. Nevertheless, the accuracy of the proposed models differs depending on the algorithm employed. Hence, further endeavors are required to enhance the precision of theses models, enabling their application and andvancement into intelligent systems capable of accurately forecasting the targeted ailment.

The NB algorithm is a probabilistic method that utulizes Bayes' theorem to classify data. Bayesian inference possesses the benefit of forcasting future probabilities by leveraging prior experiences. Researchers have found that NB demonstrates superior performance in comparison to other classification models. Futhermore, one of the benefits is that NB only necessitates a limitated quantity of training data in order to ascertain the essential parameter estimated during the process of data classification. Generalized linear models (GLM) are algorithms developed from linear regression models, offering improved advantages [11].

Researchers have conducted multiple studies on the categorization and forecasting of heart disease using diverse methodologies and technologies. Ozcan and Peker [12] conducted a study where they used classification and decision tree (CART) method to predic heart disease in 1190 patients. The model they developed had an accuracy rate of 87%. Sheeba *et al.* [13] conucted a study that aimed to enhance the performance of ensemble classification techniques for predicting heart disease. They employed metaheuristic training and utilized methods such as support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), and principal component analysis (PCA), the resulting model achieved an accuracy rate of 57.8%. In their study, Ali *et al.* [14] investigated the prediction of the heart disease using supervised machine learning algorithms, including k-NN, decision tree (DT), and RF. The most accurate model, with a 100% accuracy rate, was found to be the RF algorithm. Deepika and Balaji conducted additional research [15] and successfully predicted heart disease using the multi-layer-perceptron technique, achieving an impressive model accuracy rate of 94.28%. A separate investigation conducted by Rajendran and Karthi [16] employed linear regression and NB to forecast heart disease, achieving a remarkable accuracy rate of 92,4%. In contrast to these previous studies, Ali *et al.* [17] devised a sophisticated system for heart disease monitoring utilizing ensemble deep learning, resulting in an impressive accuracy rate of 98.5%. A study conducted by George and Gaikwad [18] utilized simulation modeling to map the cholesterol levels of heart attack patients. Additionally, other researchers employed a genetic algorithm (GA) based K-Means method to classify heart attacks classification [19].

This research proposes a model for early detection of heart disease using data mining techniques based on machine learning. Specifically, the NB and GLM algorithms are employed to achieve the highest level of accuracy. Prior studies have demonstrated that employing distinct methodologies has yielded divergent findings, leadning to disparate research outcomes and distinct accuracy models. Unlike prior studies that primarily concentrated on develoving Naive Bayes models and refining them for diabetes detection [20], alternative research employed fuzzy logic mamdani and NB for the purpose of detecting dental diseases [21]. Nevertheless, the models derived from prior research necessitate endeavors to enhance precision and explore optimal models for the classification of heart disease. Another proposed endeavor in this article involves optimizating NB and GLM models using GA in order to enhance accuracy when compared to non-optimized approaches. This study is anticipated to be advantageous for individuals requiring timely identification for the purpose of hearth disease prevention.

2. METHOD

2.1. Dataset and tools

The data in this study was obtained from the repository https://www.kaggle.com/. The dataset used is the heart attack dataset collected in the year 2021 [22]. It consists of 13 variables and 1 label, with a total of 303 records. In this research, the data was divided into two parts: the training data and the testing data, with a composition of 90% for training and 10% for testing. An example of the dataset used in this study is presented in Table 1. The tools used for analysis and experiments in this research were performed using RapidMiner Studio software.

As shown in Table 1, the variable descriptions include Age: Age of the patient, Sex: Gender of the patient. CP: Chest pain type (0: typical angina; 1: atypical angina; 2: non-anginal pain; 3: asymptomatic). TRTBPS: Resting blood pressure (mm Hg). Chol: Cholesterol (mg/dl) fetched via BMI Sensor. FBS: Fasting blood sugar >120 mg/dl (1: true; 0: false). Rest_ECG: Resting electrocardiographic results (0: normal; 1: having ST-T wave abnormality; 2: showing probable or definite left ventricular hypertrophy). Thalach: Maximum heart rate achieved. Exang: Exercise-induced angina (1: yes; 0: no). Old_Peak: ST depression induced by exercise relative to rest. SLP: The slope of the peak exercise ST segment (0: unsloping; 1: flat; 2: downsloping). CAA: Number of major vessels (0-3). Thall: Thalassemia (0: null; 1: fixed defect; 2: normal; 3: reversible defect). Output: Diagnosis of heart disease (0: <50% diameter narrowing, less chance of heart disease; 1: >50% diameter narrowing, more chance of heart disease).

Figure 1 illustrates the stages of the research process carried out in this study, which involved several activities. The initial stage involves data research, followed by data preprocessing before feeding it into the model. In the data preprocessing stage, the data is divided into two parts: training data and testing data, with a 90:10 ratio. The next step in the process is the application of the NB and GLM algorithms for modeling. Subsequently, model optimization is carried out as an effort to improve the performance accuracy of the model by optimizing weights using the evolutionary GA. The data validation stage involves the use of the k-fold cross-validation method. Furthermore, the process of determining the model's performance values in the next stage is carried out using the confusion matrix method.

Table 1. Example of a research dataset

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
57	0	1	130	236	0	0	174	0	0	1	1	2	0

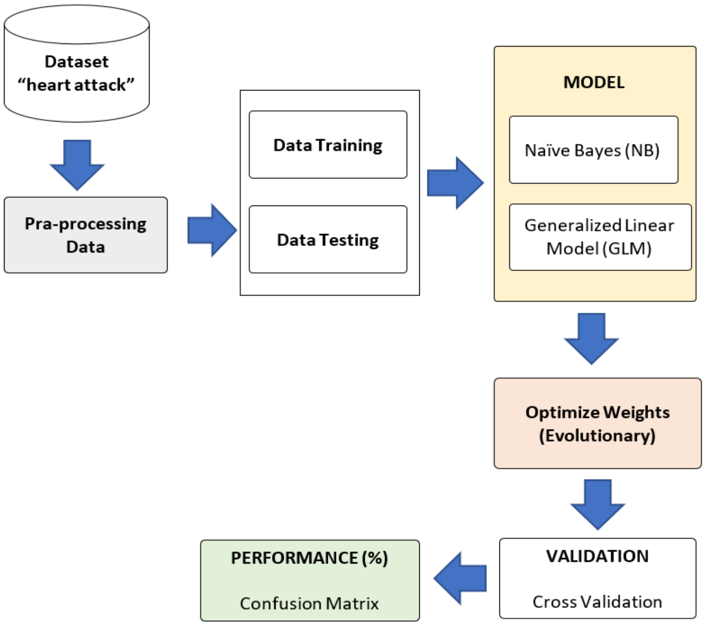

Figure 1. A proposed framework models

2.2. NB and GLM

NB is an algorithm that utilizes the concept of predicting future probabilities based on past experiences. The NB method technically applies supervised learning techniques to classify future objects by assigning class labels to instances using conditional probabilities or the likelihood of an occurring event based on other previously observed events [23]. The NB equation is depicted in (1).

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{1}$$

Where: X represents data with an unknown class, H is the hypothesis that data X belongs to a specific class, *P(H|X)* is the probability of hypothesis H given condition X (posterior probability), *P(H)* is the probability of hypothesis H, $P(X|H)$ is the probability of X based on hypothesis H, and *P(X)* is the probability of X.

The GLM is an extension of regression models used to analyze both discrete and continuous response variables. The GLM does not require a normal distribution of data but falls within the exponential family of distributions [24], [25]. Furthermore, GLM assumes that observations are independent and do not consider correlations between the outcomes of n observations. In general, the equation used in the GLM is shown in (2).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \tag{2}$$

Y is a dependent variable, β is beta weight (parameter estimates), X is regressor, and ε is residual.

2.3. Evaluation and validation model

K-fold cross-validation is applied [26] using (3) to validate the experimental results of the model using the proposed method in the research. Additionally, to measure the performance values of the generated model, the confusion matrix method is used [27], as shown in (4).

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i \tag{3}$$

Where *E* is error, *k* is the total of *k*, E_i is *error* of-*i*.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

where TP is *True Positive*, TN is true negative, FP is false positive, and FN is false negative.

3. RESULTS AND DISCUSSION

3.1. Experiments using GLM

The first experiment conducted was the application of the GLM method to the acquired dataset to obtain the best model. In this experiment, the analysis results are presented in Table 2. Based on the trial results, the highest accuracy values were achieved for each sampling category. For linear sampling, an accuracy rate of 77.88% was obtained with fold=8, while for shuffled sampling, an accuracy rate of 84.50% was achieved with fold=6. Lastly, using stratified sampling resulted in an accuracy performance of 82.85%. The best trial results using the confusion matrix are shown in Table 3, revealing that the highest accuracy rate is 84.50%.

In some of the experimental trials, the best model which presents a comparison of the accuracy performance achieved and is visualized using the ROC curve was obtained as shown in Table 4. The receiver operating characteristic (ROC) curve is presented to illustrate the performance. For Model 4, which is the best model, the parameters were set using the shuffled sampling method with fold=6, resulting in the highest accuracy performance of 84.50%, precision of 88.64%, and recall of 75.63%. The difference in model accuracy achieved in these trials is due to variations in parameter settings, especially the fold value and the sampling method. A comparison of model performance using GLM is presented in Table 4.

Table 2. The experimental results using the GLM

Sampling	Fold	Accuracy	Precision	Recall	AUC
linear	10	77.49%	48.12%	71.74%	0.087
linear	8	77.88%	48.21%	72.46%	0.106
linear	6	74.27%	48.21%	69.57%	0.139
linear	4	71.95%	48.44%	64.49%	0.211
Shuffled	10	83.51%	85.56%	74.05%	0.903
Shuffled	**8**	**84.50%**	**87.76%**	**75.31%**	**0.905**
Shuffled	6	84.50%	88.62%	75.63%	0.909
Shuffled	4	82.18%	85.10%	73.71%	0.905
stratified	10	82.48%	85.58%	74.73%	0.896
stratified	8	82.85%	85.10%	76.10%	0.898
stratified	6	82.83%	84.26%	76.81%	0.912
stratified	4	81.84%	76.81%	76.81%	0.897

Table 3. The best model performance using the confusion matrix

Accuracy: 84,50%	True 1	True 0	Class precision
Pred. 1	151	33	82,07%
Pred. 0	14	105	88,24%
Class recall	91.52%	76.09%	

Table 4. Comparison of the best models using GLM

Sampling	Fold	Accuracy	Precision	Recall
linear	8	77.88%	48.21%	72.46%
shuffled	6	**84.50%**	88.62%	75.63%
stratified	8	82.85%	85.10%	76.10%

3.2. Optimizing the GLM using GA

The best GA method proposed in this study is a GA that has had its parameter values modified, thus it is named ‘modify genetic algorithm (M-GA)’. To achieve a better-performing model in the GLM, this research applied the GA as an optimization method to enhance accuracy. In this test, model optimization was carried out by optimizing the weight values in GLM model using GA. The experimental results with the best performance values are presented in Table 5, and the visual comparison is shown in Figure 2. In this model trial, the GA was configured with a parameter population ranging from 5 to 10, resulting in multiple models with slightly varying accuracy values. The model with the highest accuracy was achieved using a population parameter of 10, with an accuracy performance of 86.64%, precision of 88.89%, and recall of 80.01%.

Table 5. Comparison of the optimized GLM model results with M-GA

Population	Accuracy	Precision	Recall
5	85.77%	87.25%	80.58%
10	**86.48%**	88.89%	80.01%
15	86.14%	88.55%	80.64%
20	86.15%	87.50%	80.86%

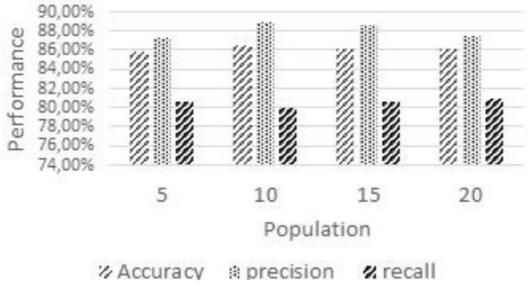

Figure 2. Comparison of the optimized GLM model results with M-GA

3.3. Experiments using the NB method

The NB method was applied in the experiments to obtain the best model. Furthermore, in the NB experiments, three sampling methods were used: linear, shuffle, and stratified. The use of various sampling methods was done to find the model with the best performance, rather than relying on just one sampling method. Additionally, different fold values were set for validation to observe the changes in each obtained model. The first experiment was conducted using the linear sampling method, resulting in models with varying performance, as shown in Table 6. The next step was to perform experiments using the shuffle and stratified sampling methods, with the results presented in Tables 7 and 8.

Table 6. The results of the NB experiments using with linear sampling

Fold	Accuracy	Precision	Recall
10	**79.15%**	48.24%	75.36%
9	78.27%	48.15%	74.64%
8	77.22%	48.21%	75.36%
7	77.83%	50.00%	74.64%
6	77.25%	48.44%	74.64%
5	76.56%	53.33%	73.91%
4	75.59%	48.58%	73.19%
3	73.93%	57.50%	73.19%
2	38.19%	0.00%	0.00%

Table 7. The results of the NB experiments with shuffle sampling

Fold	Accuracy	Precision	Recall
10	82.48%	81.58%	78.63%
9	81.50%	81.15%	77.74%
8	**82.53%**	82.02%	77.84%
7	80.85%	79.56%	77.21%
6	82.84%	82.68%	78.85%
5	81.16%	79.61%	78.91%
4	81.19%	80.92%	76.93%
3	81.85%	81.23%	78.17%
2	80.86%	79.81%	77.50%

Table 8. The results of the NB experiments with stratified sampling

Fold	Accuracy	Precision	Recall
10	81.51%	81.18%	77.64%
9	80.87%	80.53%	77.50%
8	81.85%	82.01%	77.49%
7	**82.51%**	82.73%	78.23%
6	82.17%	81.99%	78.26%
5	82.50%	82.94%	77.54%
4	81.19%	80.46%	77.52%
3	80.86%	80.33%	76.81%
2	82.18%	81.41%	78.99%

In the experiments conducted using NB, when the linear sampling technique was used for the best model, as shown in Table 6, it resulted in an accuracy performance of 79.15%. This accuracy level is not quite as high as expected and suggests the need for efforts to improve accuracy. In contrast, the NB experiments using the shuffle sampling technique and a fold value of 8 achieved a higher accuracy of 82.53%, as shown in Table 8. Slightly different results were obtained using the stratified sampling technique, where the highest accuracy was achieved with a fold value of 7, which is 82.51%. This value is not significantly different from the shuffle sampling results but is slightly lower. Based on the experimental results, when comparing accuracy values, fold values of 7 and above have a higher likelihood of achieving high accuracy.

3.4. Model optimization of NB using optimize weights (evolutionary)

To enhance accuracy, model optimization was performed by conducting experiments using GA as the optimization algorithm, resulting in several best models based on predefined parameter values. In this optimization experiment, two sampling techniques were used: linear sampling and shuffle sampling. For the GA method, the parameter values set for the linear sampling technique were fold=10 and crossover=uniform, as shown in Table 9. Additionally, Table 10 presents the analysis results using shuffle sampling with fold=8 and crossover=uniform. In Table 9, the results of the experiments using the NB+GA model show the highest accuracy level, 81.82%. For this model, the GA parameter was set to use the tournament technique with a population value of 15. In contrast, the results obtained using shuffle sampling had a higher accuracy of 85.83%, as shown in Table 10. Furthermore, this top model was achieved after specifying a population value of 15 and using the tournament schema.

Table 9. Experiment with the NB and GA model using linear sampling

Population	Schema	Accuracy	Precision
5	tournament	81.81%	83.20%
10	tournament	81.81%	83.20%
15	tournament	**81.82%**	83.20%
20	tournament	81.82%	83.20%
5	Uniform	80.82%	48.24%
10	Uniform	81.81%	83.20%
15	Uniform	81.14%	48.24%
20	Uniform	80.83%	49.38%
5	roulette wheel	79.16%	48.24%
10	roulette wheel	81.47%	48.24%
15	roulette wheel	81.14%	48.24%
20	roulette wheel	81.15%	82.40%

Table 10. Experiment with the NB and GA model using shuffle sampling

Population	Schema	Accuracy	Precisin
5	tournament	84.84%	86.00%
10	tournament	85.48%	86.81%
15	tournament	**85.83%**	87.25%
20	tournament	85.81%	87.19%
5	Uniform	83.50%	83.71%
10	Uniform	85.19%	86.71%
15	Uniform	83.49%	84.96%
20	Uniform	84.49%	85.24%
5	roulette wheel	84.14%	83.64%
10	roulette wheel	83.83%	84.05%
15	roulette wheel	84.82%	85.90%
20	roulette wheel	84.48%	85.91%

3.5. Evaluation of the best NB and GLM model

Based on the results obtained in the NB_M-GA model experiments, it is clear that the best NB method model is superior to the NB model before optimization process. In this case, the increase in performance is quite significant, as indicated by the higher accuracy achieved compared to the pre-optimized NB model. The comparison of accuracy performance values between the two models can be seen in Table 11. Based on the results obtained, it is evident that the GA has an impact on increasing the accuracy of the NB model. There is a noticeable increase in accuracy, where the initial best NB model, which had an accuracy of 82.53%, improved to 85.83%. This increase in accuracy is also highly influenced by the parameter values set for each applied method, making these values crucial to consider.

Apart of that, based on the results obtained in the experiments, two models were obtained using the GLM method: the model before optimization and the model after optimization. Efforts to improve model accuracy have been quite successful. However, this improvement is not very significant because the difference is not very large, but the accuracy values have increased. The accuracy level obtained in the experiments for the GLM+GA model showed an accuracy performance of 86.48%, which is higher than the regular GLM model, which achieved 84.50%.

The GLM base on the M-GA exhibits a higher level of accuracy compared to other models. The M-GA shown in Figure 3 is a modified method where the most influential parameter values have been adjusted, including maximal fitness = infinity, selection scheme=tournament, cross over type=shuffle, mutation variance =1.0, dan population=10. It is important to note that the proposed M-GA method's setting, using the rapidminer studio application for model analysis, may differ from those of other researchers. A comparison of the two models, GLM and GLM_M-GA, can be seen in Table 12. Based on the results obtained, it is evident that the application of the GA algorithm as a model optimization method significantly impacts the performance accuracy of the GLM model, as indicated by the change in accuracy values. In the experimental model, GLM_M-GA achieved an accuracy of 86.48%, which is higher than the classic GLM model's accuracy of 84.50%.

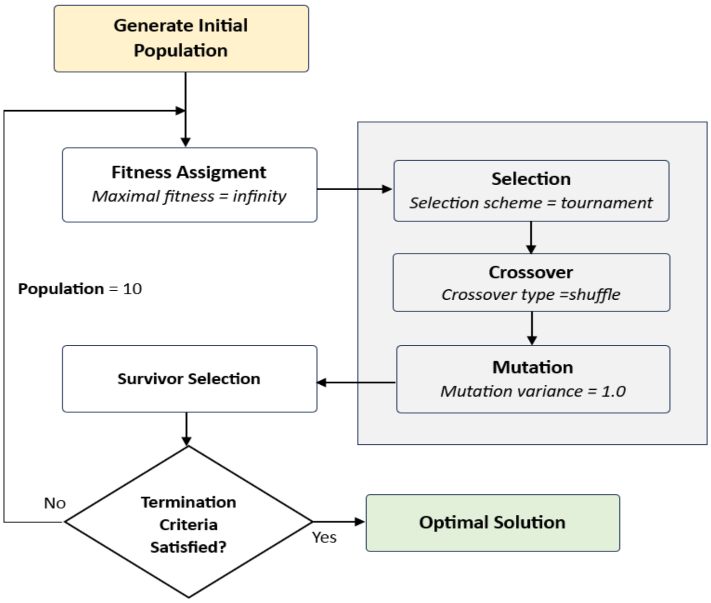

Figure 3. A proposed M-GA method

Table 11. Comparison of the NB and NB_M-GA models

Model	Accuracy (%)	Precision (%)	Recall (%)
NB	82.53%	82.02%	77.84%
NB_M-GA	**85.83%**	87.25%	79.50%

Table 12. Comparison of the GLM and GLM_M-GA models

Model	Accuracy (%)	Precision (%)	Recall (%)
GLM	84.50	88.62	75.63
GLM_M-GA	**86.48**	88.89	80.01

4. CONCLUSION

Early detection of heart disease is highly beneficial in identifying the condition at an early stage, allowing for the determination of the appropriate treatment process for those affected by the disease. The classification model for early detection of heart disease using the NB and GLM algorithms, based on the experimental results, shows a relatively good level of accuracy. In this article, the GA has a significant impact on optimizing the model and achieving a notable increase in accuracy for both proposed models. Efforts in optimization and the use of other machine learning models for further research may lead to the possibility of applying these models, aiming to provide the best model comparison for early detection of heart disease.

ACKNOWLEDGEMENTS

The author would like to express gratitude to the Research and Development Center (P3M) of Politeknik Negeri Cilacap for providing research funding support through the research funding scheme from the Politeknik Negeri Cilacap's 2023 Budget Allocation (DIPA).

REFERENCES

[1] M. M. Ali *et al.*, "A machine learning approach for risk factors analysis and survival prediction of heart failure patients," *Healthcare Analytics*, vol. 3, p. 100182, Nov. 2023, doi: 10.1016/j.health.2023.100182.

[2] H. Hasanova, M. Tufail, U. J. Baek, J. T. Park, and M. S. Kim, "A novel blockchain-enabled heart disease prediction mechanism using machine learning," *Computers and Electrical Engineering*, vol. 101, p. 108086, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108086.

[3] J. K. Alwan, D. S. Jaafar, and I. R. Ali, "Diabetes diagnosis system using modified Naive Bayes classifier," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 28, no. 3, pp. 1766–1774, 2022, doi: 10.11591/ijeecs.v28.i3.pp1766-1774.

[4] Rajni and Amandeep, "RB-Bayes algorithm for the prediction of diabetic in 'PIMA Indian dataset," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 4866–4872, Dec. 2019, doi: 10.11591/ijece.v9i6.pp4866-4872.

[5] N. Prasath, V. Pandi, S. Manickavasagam, and P. Ramadoss, "A comparative and comprehensive study of prediction of Parkinson's disease," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 23, no. 3, pp. 1748–1760, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1748-1760.

[6] S. P. Patro, G. S. Nayak, and N. Padhy, “Heart disease prediction by using novel optimization algorithm: A supervised learning prospective,” *Informatics Medicine Unlocked*, vol. 26, p. 100696, 2021, doi: 10.1016/j.imu.2021.100696.

[7] G. Sugendran and S. Sujatha, “Earlier identification of heart disease using enhanced genetic algorithm and fuzzy weight based support vector machine algorithm,” *Measurement Sensors*, vol. 28, p. 100814, Jun. 2023, doi: 10.1016/j.measen.2023.100814.

[8] C. V. Aravinda, M. Lin, K. R. U. K. Reddy, and G. A. Prabhu, “A deep learning approach for the prediction of heart attacks based on data analysis,” in *Deep Learning for Medical Applications with Unique Data*, Elsevier, 2022, pp. 1–18.

[9] A. Jain, A. C. S. Rao, P. K. Jain, and Y. C. Hu, “Optimized levy flight model for heart disease prediction using CNN framework in big data application,” *Expert Systems with Applications*, vol. 223, p. 119859, Aug. 2023, doi: 10.1016/j.eswa.2023.119859.

[10] G. Rajkumar, T. Gayathri Devi, and A. Srinivasan, “Heart disease prediction using IoT based framework and improved deep learning approach: Medical application,” *Expert Systems with Applications*, vol. 111, p. 103937, Jan. 2023, doi: 10.1016/j.medengphy.2022.103937.

[11] L. P. Fávero, P. Belfiore, and R. de Freitas Souza, “Generalized linear mixed models,” in *Data Science, Analytics and Machine Learning with R*, Elsevier, 2023, pp. 303–319.

[12] M. Ozcan and S. Peker, “A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthcare Analytics*, vol. 3, p. 100130, Nov. 2023, doi: 10.1016/j.health.2022.100130.

[13] P. T. Sheeba, D. Roy, and M. H. Syed, “A metaheuristic-enabled training system for ensemble classification technique for heart disease prediction,” *Advances in Engineering Software*, vol. 174, p. 103297, Dec. 2022, doi: 10.1016/j.advengsoft.2022.103297.

[14] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,” *Computers in Biology and Medicine*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[15] D. Deepika and N. Balaji, “Effective heart disease prediction using novel MLP-EBMDA approach,” *iomedical Signal Processing and Control*, vol. 72, p. 103318, Feb. 2022, doi: 10.1016/j.bspc.2021.103318.

[16] R. Rajendran and A. Karthi, “Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers,” *Expert Systems with Applications*, vol. 207, p. 117882, Nov. 2022, doi: 10.1016/j.eswa.2022.117882.

[17] F. Ali *et al.*, “A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion,” *Information Fusion*, vol. 63, pp. 208–222, Nov. 2020, doi: 10.1016/j.inffus.2020.06.008.

[18] J. P. George and S. M. Gaikwad, “Simulation modeling for heart attack patient by mapping cholesterol level,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 1, p. 16, Apr. 2020, doi: 10.11591/ijeecs.v18.i1.pp16-23.

[19] A. A. Hussein, “Improve the performance of k-means by using genetic algorithm for classification heart attack,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 2, p. 1256, Apr. 2018, doi: 10.11591/ijece.v8i2.pp1256-1261.

[20] O. Somantri, “An optimize weights naïve bayes model for early detection of diabetes,” *Telematika*, vol. 15, no. 1, Feb. 2022, doi: 10.35671/telematika.v15i1.1307.

[21] L. P. Wanti and O. Somantri, “Comparing fuzzy logic mamdani and naïve bayes for dental disease detection,” *Journal of Information Systems Engineering & Business Intelligence*, vol. 8, no. 2, pp. 182–195, Oct. 2022, doi: 10.20473/jisebi.8.2.182-195.

[22] R. Rahman, “Heart attack analysis and prediction dataset,” *Kaggle*, 2021, https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

[23] P. Cichosz, “Naïve Bayes classifier,” in *Data Mining Algorithms*, Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 118–133.

[24] J. Neuhaus and C. Mcculloch, “Generalized linear models,” *Wiley Interdiscip Reviews: Computational Statistics*, vol. 3, no. 5, pp. 407–413, Sep. 2011, doi: 10.1002/wics.175.

[25] F. Zuniga, T. J. Kozubowski, and A. K. Panorska, “A generalized linear model for multivariate events,” *Journal of Computational and Applied Mathematics*, vol. 398, p. 113655, Dec. 2021, doi: 10.1016/j.cam.2021.113655.

[26] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: A review with examples from ecology,” *Ecological Monographs*, vol. 93, no. 1, Feb. 2023, doi: 10.1002/ecm.1557.

[27] M. Heydarian, T. E. Doyle, and R. Samavi, “MLCM: Multi-label confusion matrix,” *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

BIOGRAPHIES OF AUTHORS

Oman Somantri is scientist and lecturer at college of the Department of Cyber Security Engineering, State Polytechnic of Cilacap, Indonesia. His research areas are data mining, text mining, and sentiment analysis. He has received the S.Kom degree in informatics engineering from the STMIK Sumedang, Indonesia, and the M.Kom degree in informatics from the Universitas Dian Nuswantoro, Semarang, Indonesia. He can be contacted at email: oman_mantri@yahoo.com.

Linda Perdana Wanti received the S.Kom degree in informatics from the University of AMIKOM Purwokerto, Indonesia, and the M.Kom degree in informatics from the University of AMIKOM Yogjakarta, Indonesia. His research areas are data mining, text mining, and sentiment analysis decision support systems, expert systems, and databases. She can be contacted at email: linda_perdana@pnc.ac.id.