International Journal of Computer Science and Information Technology (IJISIT)
Volume 1, No. 1 (2024) pp. 1-7
e-ISSN 3063-4156 | p-ISSN 3063-4172
DOI: 10.55123/ijisit

Design and Development of Artificial Intelligence Based Applications Using Classification Algorithms in Datamining to Predict Early Diabetes

Rano Agustino[1,a*], Ratna Mutu Manikam[2,b]
[1,2]Jakarta 13550, Universitas Mohammad Husni Thamrin, Indonesia
[a]rano.agustino@gmail.com[*], [b]ratnamutu2811@gmail.com

Article Info

Article history:

Received January 24, 2024
Revised March 05, 2024
Accepted June 15, 2024

Keywords:

Diabetes Prediction Application
Prediction Naive Bayes
RAD Method
Classification Comparation

ABSTRACT

Diabetes is a chronic disease characterized by high blood sugar levels. High blood sugar levels can cause various serious complications, such as heart disease, stroke and kidney failure. Early detection of diabetes is essential to prevent this complication. This research is using an experimental method. The experimental method is a research method carried out by deliberately manipulating the independent variable to see its effect on the dependent variable. In this case, the independent variable being manipulated is the machine learning algorithm used to predict diabetes. The dependent variable observed was the accuracy of the diabetes prediction model. From the research results, it was found that the Naive Bayes model had the highest accuracy, namely 81.5%. This model is better than the Decision Tree C.45 (72.7%), Random Forest (73.8%), and K-Nearest Neighbor (71.5%) models. Based on the research results, it can be concluded that the Naive Bayes model is the best model for predicting diabetes. The appropriate system development method for this research is the Rapid Application Development (RAD) method. Further research can be carried out to improve the accuracy of diabetes prediction models by using optimization models, such as Particle Swarm Optimization or Colony Optimization. In addition, further research can be carried out to develop diabetes prediction applications that can predict diseases that may be caused by diabetes, such as heart disease, stroke, kidney disease and blindness.

Corresponding Author:

Rano Agustino
Information Systems, Fakultas of Computer
Universitas Mohammad Husni Thamrin
13550, Jakarta, Indonesia
Email: rano.agustino@gmail.com

1. INTRODUCTION

The Glucose is a simple sugar which is the main source of energy for the human body [1]. Glucose is absorbed from the food we eat and stored in the liver and muscles. When the body needs energy, glucose is released into the bloodstream. In healthy people, blood glucose levels will remain stable within a narrow range. However, in people with diabetes, blood glucose levels can become too high. This is caused by the body's inability to produce or use insulin effectively. Insulin is a hormone that helps the body to absorb glucose from the bloodstream.

Glucose that is not absorbed by the body will accumulate in the bloodstream [2]. This can cause various serious complications, such as heart disease, stroke and kidney failure. Glucose is one of the most

important factors in predicting diabetes. High blood glucose levels are an early sign of diabetes. Therefore, *Artificial Intelligence* (AI)-based diabetes prediction applications often use diabetes dataset to predict diabetes risk [3]. Blood glucose data can be collected from a variety of sources, such as blood tests, portable blood glucose meters, and health apps. AI-based diabetes prediction applications will use blood glucose data to train classification models. This classification model will then be used to predict the risk of diabetes in individuals.

Diabetes is a chronic disease characterized by high blood sugar levels. High blood sugar levels can cause various serious complications, such as heart disease, stroke and kidney failure. Early detection of diabetes is essential to prevent this complication.

Potential Applications of Diabetes Prediction Based on *Artificial Intelligence* (AI) offers new and promising solutions for diabetes prediction [4]-[5]. AI-based applications can analyze an individual's demographic, medical and lifestyle data to predict their risk of diabetes. AI-based diabetes prediction applications have the potential to improve early detection of diabetes in several ways. AI-based applications can use complex classification algorithms to predict diabetes risk with high accuracy. This can help detect diabetes early, so the risk of complications can be reduced.

AI-based applications can process large amounts of data quickly and efficiently. This can help doctors diagnose diabetes more quickly and accurately. AI-based applications can be accessed at relatively affordable costs. This can help increase the accessibility of diabetes diagnosis to the wider community. AI-based diabetes prediction applications have the potential to improve early diabetes detection. Further research is needed to evaluate the potential for this application in a clinical context.

AI-based diabetes prediction applications have the potential to improve early diabetes detection in several ways, namely:

- AI-based applications can use classification algorithms consisting of Naive Bayes, C45, Random Forest and K-Nearest Neighbor algorithms to compare predicting diabetes datasets from hospital medical records in Jakarta and see which accuracy is higher. After obtaining an algorithm with higher results, that algorithm will be used
- This AI-based application will be web and mobile based. The programming languages used is PHP and JavaScript. It is hoped that this website can help people to diagnose diabetes more quickly and accurately.

Further research is needed to evaluate the potential application of AI-based diabetes prediction in a clinical context.

In 2021, Ahmad Shaker Abdalrada, et al. in his article entitled "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study", concluded that Diabetes Mellitus (DM) and cardiovascular diseases (CVD) showed that participants with HbA1c > 6.45 and TC\HDL ratio > 5.5 risks developing both diseases (probability 97.9%) [7]. In contrast, participants with HbA1c >6.45 and TC\HDL ratio ≤ 5.5 were more likely to suffer from DM only (probability 84.5%) and participants with HbA1c ≤5.45 and HDL >1.45 were more likely to be healthy (probability 82, 4%). Furthermore, participants with HbA1c ≤5.45 and HDL <1.45 were only at risk of CVD (100% probability). The prediction accuracy of the ML model for detecting co-occurrence of DM and CVD was 94.09%, sensitivity 93.5%, and specificity 95.8%.

In 2022, Edeh, MO, et al. in his article concludes, Four Machine Learning classification algorithms, namely the Supervised learning algorithm (Random Forest, SVM and Naïve Bayes, Decision Tree DT) and the unsupervised learning algorithm (K-means), have been the techniques used in this investigation to identify diabetes at this stage. beginning [8]. The experiment was carried out on two databases, one taken from Frankfurt Hospital in Germany and the other from the database. PIMA Indian Diabetes (PIDD) is provided by the UCI machine learning repository. The results obtained from the database extracted from Frankfurt Hospital, Germany show that the random forest algorithm outperforms with the highest accuracy of 97.6%, and the results obtained from the Pima Indian database show that the SVM algorithm outperforms with the highest accuracy of 83.1%. compared to other algorithms. The validity of these results is ensured through the process of splitting the data set into two parts: the training set and the test set, which is explained below. The training set is used to develop the model's capabilities. The test set is used to test the model and determine its correctness.

In 2023 Rashi Rastogi conducted research in which the article was entitled "Diabetes prediction model using data mining techniques" where the article concluded that diabetes could cause various health problems, such as heart disease, kidney problems and eye problems [9]. Therefore, it is very important to prevent, monitor and increase awareness about diabetes. Data collection techniques can contribute to healthcare decisions for accurate diagnosis and treatment of diseases, thereby reducing the workload of experts. In this research, researchers used a diabetes prediction model using data mining techniques, namely Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. The proposed mechanism is trained using

Python and analyzed with real datasets, collected from Kaggle. The research results show that logistic regression has high accuracy, namely 82.46%, compared to other data mining techniques. This shows that logistic regression can be used to predict diabetes with high accuracy.

Still in 2023, Muhammad Exell Febrian, et al. in his article entitled "Diabetes prediction using supervised machine learning" concluded that by comparing the two k-Nearest Neighbor algorithms and the Naive Bayes algorithm to predict diabetes based on several health attributes in a data set using supervised machine learning [10]. According to the results from our experiments and evaluating the algorithm using the Confusion Matrix, the Naive Bayes algorithm outperforms KNN, with an average value of 76.07 percent accuracy, 73.37 percent precision, and 71.37 percent recall on Naive Bayes and average accuracy values 73.33 percent, precision 70.25 percent, and recall 69.37 percent on KNN. As a result, it can be concluded that the Naive Bayes algorithm is preferable to the KNN algorithm in predicting diabetes using the Pima Indian data set.

1.1 Research Problems

Based on the previous background, there are problems in this research are as follows:

1. What classification algorithm is suitable for diabetes predictions using diabetes datasets from hospitals in nearby Jakarta?
2. What method of system development is suitable for *Artificial Intelligence* -based applications that can detect diabetes?

1.2 Objectives and benefits Study

Based on the research problems above, the aims and benefits of the research are as follows

1. Assess and compare the performance of classification algorithms, such as Naive Bayes, Decision Tree C.45, Random Forest, and K-Nearest Neighbor, with the aim of improving diabetes prediction accuracy.
2. Determine the classification algorithm that provides diabetes prediction results with the highest accuracy, precision, recall and AUC based on the comparisons made.
3. Design and develop an efficient AI-based diabetes prediction application, accessible via web and mobile, using PHP and JavaScript programming languages.
4. Evaluating the potential of AI-based diabetes prediction applications in supporting early diabetes detection, especially in helping Health Institutions and the Community to diagnose diabetes more quickly and efficiently.
5. Identify the benefits of using AI-based diabetes prediction applications, such as increased accessibility and efficiency of diagnosis, and identify challenges that may arise in their implementation.

2. METHOD

The type of research used in this research is applied research. Applied research aims to produce products or services that can be applied in real life. In this case, the product or service produced is an Artificial Intelligence Application to predict AI-based diabetes. The research method that can be used is the experimental method. The experimental method is a research method carried out by deliberately manipulating the independent variable to see its effect on the dependent variable. In this case, the independent variable being manipulated is the machine learning algorithm used to predict diabetes. The dependent variable observed was the accuracy of the diabetes prediction model).

3. RESULTS AND DISCUSSION

The steps in this research are based on CRISP-DM [11]-[16]. The following are the stages:

3.1. Business understanding

This step aims to understand the business needs of developing an AI-based diabetes prediction application. Based on the previous background, the aim of developing an AI-based diabetes prediction application is to improve early detection of diabetes. Users of this application are health workers and the public.

3.2. Data understanding

This step aims to understand the data that will be used to develop an AI-based diabetes prediction model. The dataset used is a medical record dataset about diabetes originating from hospitals around Jakarta. Before this diabetes dataset, the data will be prepared so that the dataset has supporting data quality so that the data can be easily processed.

3.3. Data preparation

This step aims to prepare data for the modeling process. Some of the activities carried out in this step are:

- Cleaning diabetes data sets from duplicate and damaged data
- Divide the data contents in the dataset into 2, namely training data *and* testing *data*. The dataset consists of 768 examples or records, whose attributes consist of; Jumlah Kehamilan, Gula Darah Sewatu, Tekanan Darah Diastolik, Ketebatalan Kulit, Insulin, IMT, Riwayat Keturunan Diabetes, Usia and the table label are the Results. The following is an example of a diabetes table:

Table 1. Example of Diabetes Table

Jumlah Kehamilan	Gula Darah Sewatu	Tekanan Darah Diastolik	Ketebatalan Kulit	Insulin	IMT	Riwayat Keturunan Diabetes	Usia	Hasil
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2,288	33	1

Data cleaning is performed to remove errors and anomalies in the data. Data standardization is carried out to standardize the data format. Data sharing was carried out to test the performance of the model developed.

3.4. Modeling

This step aims to develop an AI-based diabetes prediction model. Several algorithms that can be used to develop this model are Naive Bayes [17], Random Forest, Decision Tree, K-Nearest Neighbor [18]-[20]. The following is the picture:

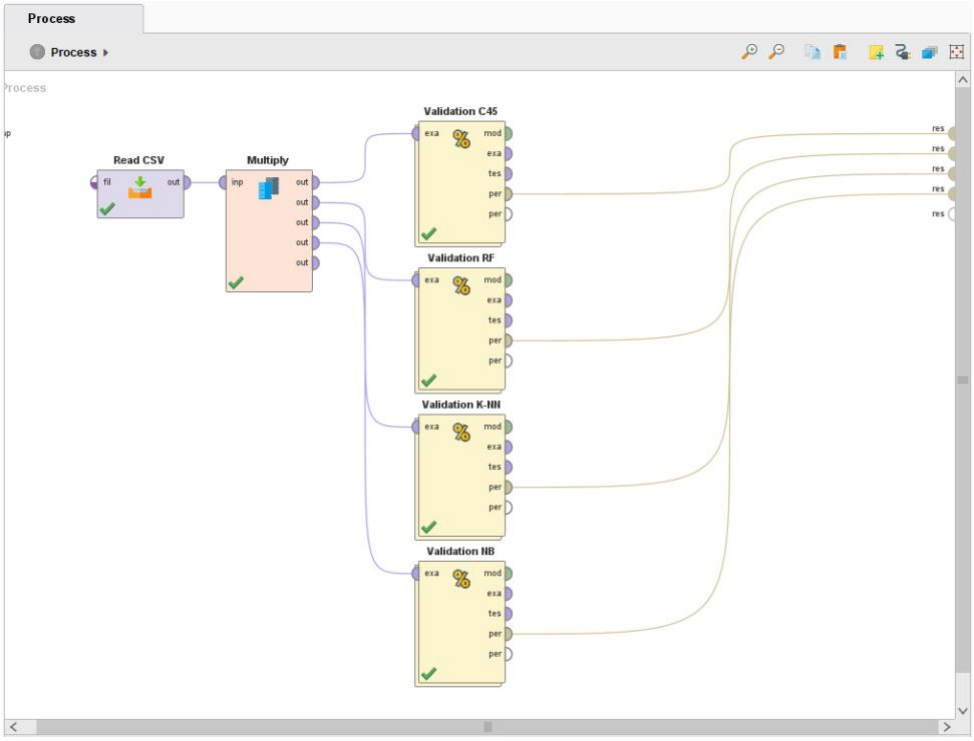

Figure 1. Modeling Process

3.5. Evaluation

This step aims to evaluate the performance of the AI-based diabetes prediction model. Some metrics that can be used to evaluate model performance are measuring Accuracy, AUC and Recall [21][22]. The following is a table evaluating the results of the four classification algorithm models, namely.

Table 2. Evaluation results table

No	Model	Accuracy	AUC	Precision
1	Decision Tree C.45	72.7	77.6	78.9
2	Random Forest	73.8	80	80.6
3	K-NN	71.5	74	81.8
4	Naive Bayes	81.5	83	83.9

From the results of table 2. From the table above, it can be concluded that the Naive Bayes model is the highest with an accuracy value of 81.5, an AUC value of 83 and a precision value of 83.9. Therefore, the model will be used to predict diabetes. And the next step will be to build an application using the PHP and JavaScript programming languages with the Naive Bayes pattern [23]-[24].

3.6. Evaluation

This step aims to apply the AI-based diabetes prediction model in a real context. Some of the activities carried out in this step are:

- Developing a system using the Rapid Application Development (RAD) method. With the RAD method, it is hoped that application creation can be done quickly and efficiently [25]-[26]. The following is a picture of the stages of RAD.

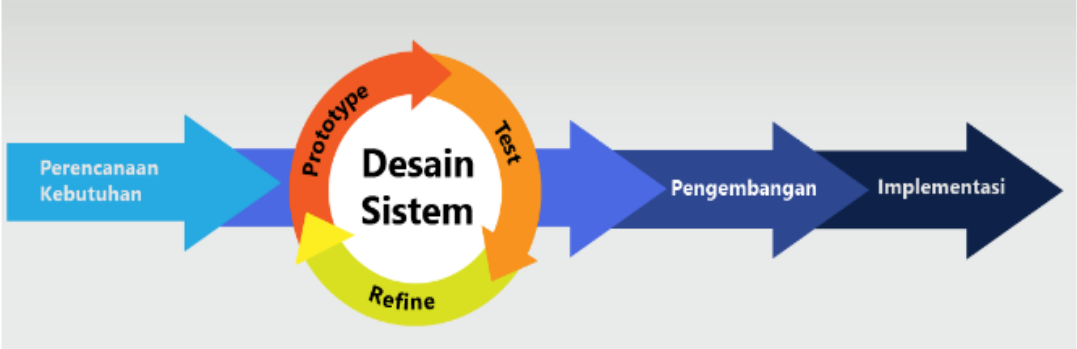

Figure 2. RAD stages

1) Planning.
   This stage is the initial stage of RAD. In this research, the problem has been identified and the flow or pattern of making this application is also clear, namely making an artificial intelligence application to predict early diabetes. The algorithm used is Naive Bayes, which produces quite high accuracy, precision, recall and AUC.
2) System Design.
   At this system design stage, researchers carry out designs according to their research background, namely about public health. For the process flow, the user wants to use this application, the user must log in first, if not, then they must register first. After that, you will be asked several questions so that you can conclude whether you are predicting diabetes or not. If at this design stage users still ask for it, improvements will continue to be made
3) Development
   After designing the system according to the needs of the user, if it is appropriate then develop the application. At this stage, testing is always carried out, whether using Black Box or White Box.
4) Implementation
   After testing or testing has been carried out, this application is ready to be implemented directly for users to use

- Setting up an interface or display for users. The following is a display of the AI-based diabetes prediction application

Figure 3. Display of the AI-based diabetes prediction application

4. CONCLUSION

Based on the results of research conducted by researchers, several conclusions were found, including:

1) Based on the results of the evaluation stage, it shows that the Naive Bayes model has a higher accuracy of 81.5, a recall of 83, and an AUC of 83.9, while others such as C.45, RF and K-NN have values below the Naive Bayes value. So, it can be concluded that in this problem to predict diabetes according to the diabetes dataset from a hospital in Jakarta is to use Naive Bayes.
2) method that suits the needs of creating an artificial intelligence application to predict diabetes is the Rapid Application Development (RAD) method, which is used if the application or system needs to be implemented quickly.

Meanwhile, for suggestions for this research, the researcher has several suggestions for further research, namely:

1) It is hoped that in further research, optimization models such as Particle Swarm Optimization or Colony Optimization or other optimizations can be added so that the Accuracy, AUC and Recall values can be further increased.
2) Researchers hope that future applications can predict diseases that may be caused by diabetes, such as applications predicting heart disease, stroke, kidney disease and blindness.

ACKNOWLEDGEMENTS

The author would like to thank the Mohammad Husni Thamrin University who have played a role in this research both in terms of financial and moral support so that this publication can be achieved.

REFERENCES

[1] Luhov yy, B.L., & Kathirvel, P. (2022). Food proteins in the regulation of blood glucose control. In *Advances in Food and Nutrition Research* (Vol. 102, pp. 181-231). Academic Press.
[2] Jiwakanon, S., & Mehrotra, R. (2013). Nutritional management of end-stage renal disease patients treated with peritoneal dialysis. In *Nutritional Management of Renal Disease* (pp. 539-561). Academic Press.
[3] Nomura, A., Noguchi, M., Kometani, M., Furukawa, K., & Yoneda, T. (2021). Artificial intelligence in current diabetes management and prediction. *Current Diabetes Reports , 21* (12),
[4] Kaul, S., & Kumar, Y. (2020). Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. *SN Computer Science , 1* (6), 322.

[5] Li, J., Huang, J., Zheng, L., & Li, X. (2020). Application of artificial intelligence in diabetes education and management: present status and promising prospect. *Frontiers in public health* , *8* , 173.

[6] Abdalrada, AS, Abawajy, J., Al-Quraishi, T., & Islam, SMS (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. *Journal of Diabetes & Metabolic Disorders* , *21* (1), 251-261.

[7] Edeh, MO, Khalaf, OI, Tavera, CA, Tayeb, S., Ghouali, S., Abdulsahib, GM, Richard-Nnabu, NE and Louni, A., 2022. A classification algorithm-based hybrid diabetes prediction model. *Frontiers in Public Health* , *10* , p. 829519.

[8] Rastogi, R. and Bansal, M., 2023. Diabetes prediction model using data mining techniques. *Measurement: Sensors* , *25* , p.100605.

[9] Febrian, ME, Ferdinan, FX, Sendani, GP, Suryanigrum, KM, & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. Procedia Computer Science, 216, 21-30.

[10] Emmanuel, G., Hungilo, G.G., & Emanuel, AWR (2021, March). Performance evaluation of machine learning classification techniques for Diabetes disease. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1098, No. 5, p. 052082). IOP Publishing.

[11] Agustino, R. (2019). Comparison of Classification Algorithms Using Anaconda to Predict Crowds of Moviegoers in Cinemas. Journal of Information and Computer Technology, 5(1), 24-28.

[12] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534.

[13] Grady, N. W., Payne, J. A., & Parker, H. (2017, December). Agile big data analytics: AnalyticsOps for data science. In *2017 IEEE international conference on big data (big data)* (pp. 2331-2339). IEEE.

[14] Solano, J. A., Cuesta, D. J. L., Ibáñez, S. F. U., & Coronado-Hernández, J. R. (2022). Predictive models assessment based on CRISP-DM methodology for students performance in Colombia-Saber 11 Test. *Procedia Computer Science*, *198*, 512-517.

[15] Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp, 79, 403-408.

[16] Hariyanto, D., Sastra, R., & Putri, FEPEP (2021). Implementation of the Rapid Application Development Method in Library Information Systems. JUPITER (Journal of Computer Science and Engineering Research), 13(1), 110-117.

[17] Libnao, M., Misula, M., Andres, C., Mariñas, J., & Fabregas, A. (2023). Traffic incident prediction and classification system using naïve bayes algorithm. *Procedia Computer Science*, *227*, 316-325.

[18] Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

[19] Chisholm, A. (2013). *Exploring data with Rapidminer* (pp. 306-7). Packt publishing.

[20] Ramjan, S., & Sunkpho, J. (Eds.). (2023). *Principles and Theories of Data Mining with RapidMiner*. IGI Global.

[21] Delfani, P., Carlsson, A., King, T., Ney, A., Pereira, S. P., & Mellby, L. D. (2019). Differentiating Pancreatic Ductal Adenocarcinoma (PDAC) from individuals with symptoms suggestive of PDAC, including type II diabetes, with ROC AUC values above 0.95. *Pancreatology*, *19*, S191.

[22] Zhang, X., Li, X., Feng, Y., & Liu, Z. (2015). The use of ROC and AUC in the validation of objective image fusion evaluation metrics. *Signal processing*, *115*, 38-48.

[23] Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361.

[24] Beynon-Davies, P., Carne, C., Mackay, H., & Tudhope, D. (1999). Rapid application development (RAD): an empirical review. *European Journal of Information Systems*, *8*(3), 211-223.

[25] Beynon-Davies, P., Carne, C., Mackay, H., & Tudhope, D. (1999). Rapid application development (RAD): an empirical review. *European Journal of Information Systems*, *8*(3), 211-223.

[26] Mohan, V. (2022). System Development Life Cycle. In *Clinical Informatics Study Guide: Text and Review* (pp. 177-183). Cham: Springer International Publishing.