Enhancing Web-Based Diabetes Prediction Using Random Forest Optimization and SMOTE

Durotun Nafisah Amalia Ahli[1], Salamun Rohman Nudin[2*], Dodik Arwin Dermawan[3], Hafizhuddin Zul Fahmi[4], and I Gde Agung Sri Sidhimantra[5]

[1,2*,3,4,5]Universitas Negeri Surabaya, Surabaya, Indonesia

Keywords:
diabetes, information system, random forest, smote, flask framework

ABSTRACT

This research develops a website-based diabetes early detection information system using random forest method and synthetic minority oversampling (SMOTE) technique. This system is designed to predict the risk of diabetes based on user-inputted symptoms. Flask framework is used to optimize web application development. System testing was conducted using black box testing method and validation by medical experts, showing accurate prediction results. The use of the Flask framework facilitates the integration of modeling and user interface development. The data used in this research was balanced using SMOTE, resulting in a prediction accuracy of 96%. The results show that this information system is effective in providing early prediction of diabetes risk and can be a tool for the community to increase awareness of the importance of periodic health checks. The system also provides information that can be used to take preventive measures against diabetes mellitus, supporting government efforts to improve public health.

INTRODUCTION

Indonesia is a country with a population of 271.9 million according to WHO data in 2020. With a birth percentage of 77%. With the number 4 population in the world, the mortality rate in Indonesia is still relatively high with the main cause of chronic diseases suffered. According to WHO data, the highest causes of death in Indonesia are stroke, ischemic heart disease, and followed by diabetes mellitus in third place with a range of 43 per 100,000 population in women and 39 per 100,000 population in men (data.who.int, 2020). According to the Indonesian Ministry of Health, it is predicted that the level of vulnerability to diabetes in Indonesia could reach 30 million people by 2030. It cannot be denied that diabetes mellitus, which is a chronic condition due to problems with insulin production, can affect individuals in all age ranges, from children to the elderly, supported by analysis of diet and lifestyle (Lestari et al., 2021).

There are two types of diabetes in the world, namely type 1 diabetes which is caused by gene abnormalities resulting in a condition where the body cannot produce insulin at all, then type 2 diabetes which is caused by the body's failure to respond to insulin as a whole so that insulin cannot work properly. Looking at the report released by the International Diabetes Federation (IDF) in 2022, the number of people with type 1 diabetes mellitus in Indonesia reached 41.8 thousand people. This makes Indonesia the country with the most type 1 diabetes mellitus sufferers in ASEAN and ranked 34th on an international scale, this figure is 10% of the total Diabetes Mellitus sufferers in Indonesia because 90% of all diabetes mellitus sufferers are type 2 diabetes mellitus sufferers. From IDF statistics, most patients are in the age range of twenty to fifty-nine years, with a graph at a young age that looks quite high (Ahdiat, 2023). Diabetes mellitus is a disease categorized as a chronic disease characterized by changes in the body's performance in the metabolic process of carbohydrates, fats, and proteins, causing blood sugar levels to rise. This can be detected through the sugar content in the patient's urine which is not controlled (Mohajan & Mohajan, 2023).

This disease is feared by the community so that it has the nickname The Great Imitator, a disease that can damage all organs of the body after becoming a patient suffering from it. The damage in the body comes slowly so that patients with diabetes mellitus cannot feel and realize the various changes that occur in their bodies (Ahlin & Billhult, 2012). In the current era, lifestyle is a secondary need that depends on the era. People's lifestyle today tends to ignore the condition of their body's health by having a diet that is high in fat, salt and sugar. These ingredients are found in fast food, coffee and sugary drinks, and excessive cigarette use. Meanwhile, to control blood sugar to remain stable, people are encouraged to reduce food portions, maintain a diet, and set a stable eating schedule (Medina-Remón et al., 2018).

Organizing a healthier diet and lifestyle by exercising regularly, reducing cigarette use, and maintaining a good diet can reduce the development of diabetes mellitus in the community. Seeing the increase in the graph of diabetes mellitus patients per year, there are still many people who complain about the health services provided by the government through the BPJS program. What should be the government's media in monitoring public health and community service media for periodic health checks is currently experiencing a lot of maladministration where there is an imbalance between BPJS user patients and independent patients or insurance users. Maladministration in the provision of health services is due to the absence of standards or regulations from the government so that many health service people take advantage of this by committing maladministration (ombudsman.go.id, 2023). This action has caused the community to lose interest in checking their health conditions regularly with the health service so that the process of diagnosing diabetes in the community is delayed.

Therefore, a website-based detection information system was developed that aims to be an early detection of diabetics by utilizing the dataset that has been collected as user comparison data. This system is designed to provide diabetes risk prediction to users based on the symptoms inputted through the form so that it can help the community in detecting the risk of diabetes early and increase awareness of the importance of regular health checks. The information system was developed using the random forest method and synthetic minority oversampling technique as a detection method by utilizing the development of an expert system to diagnose diabetes mellitus based on the symptoms that arise in the user through a form that will be filled in by the user. The dataset used has 18 features that are used as a data analysis model so that the system can predict diabetes diagnoses in users. The development of the random forest method in diabetes prediction information systems is based on reference to previous research journals with the title "Machine Learning Based Diabetes Prediction and Development of Smart Web Application". The journal has several machine learning methods studied and one of them is random forest. The accuracy value obtained from each method does not have much difference, but the highest accuracy value is obtained from the random forest method. There are two datasets used in the research and from these two datasets the four highest accuracy values appear, namely SVM and Random Forest on dataset 1 and Decision Tree and Random Forest on dataset 2 (Ahmed et al., 2021).

RESEARCH METHOD

This research is designed with several stages of methodology adapted from the IBM analysis method (Rollins, 2015) and the Waterfall development method (Bassil, 2012), namely the problem analysis stage, model development, website development, and reporting. The research flow is organized as in Figure 1.

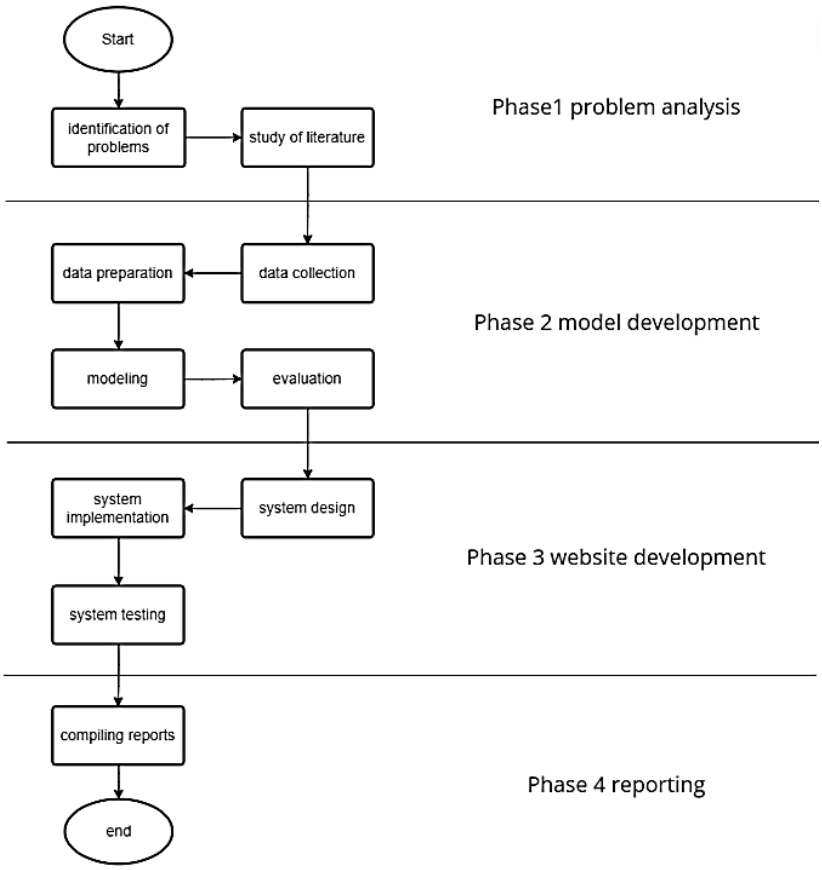


Figure 1. Research Flow

2.1 Phase I Problem Analysis

The initial stage of this research is problem analysis, in problem analysis the researcher will choose a topic, determine the formulation of the problem, objectives, and conduct a literature study for the topic of diabetes, random forest algorithms (Nudin et al., 2022), and synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002). This stage has two processes, namely identification of problems and study literature.

2.2 Phase II Model Development

Stage two is model development. This stage is adapted from IBM Data Science Methodology and there are four processes for its development. At this stage the data will be processed into a data model that is used for data analysis using the random forest algorithm and SMOTE technique. The four development processes are data collection, data preparation, modelling, and evaluation. The authors collect various dataset provided by the official international dataset website, Kaggle.com. The dataset used is Neha Prerna Tigga's dataset with the title "Diabetes Dataset 2019". This dataset has 18 features with 954 respondents (Tigga1, 2019). After obtaining the data, the authors will

examine the data to identify problems in the data obtained with three preparatory processes, namely cleaning data, combining data, and transforming data. Random forest algorithm as a modeling technique, the algorithm will be implemented in the python programming language and integrated into a web-based diabetes prediction information system. The selection of the algorithm is obtained from the reference journal literature with the title "Machine Learning Based Diabetes Prediction and Development of Smart Web Application". The literature compares various machine learning algorithms with the highest result being the random forest algorithm on the two datasets tested. The modeling stage begins with training data on the random forest algorithm to build a model. After carrying out the modeling stage, an evaluation stage is needed to determine whether the model that has been used can run in accordance with the original research objectives. At this stage, testing the model used with the confusion matrix method is carried out to produce several values (Santra & Christy, 2012).

1. True Positive (TP): how much data is actually labeled positive and how much the model predicts the label is positive.
2. True Negative (TN): how much data has an actual negative label and how much the model predicts an actual negative model.
3. False Negative (FN): how much data is actually labeled negative and how much the model predicts is actually positive.
4. False Positive (FP): how much data has an actual positive label and how much the model predicts an actual negative model.

Through the above data, it can allow researchers to calculate the probability of the quality of the model used through the calculation of the model's confusion matrix value with the formula:

1. Accuracy: The overall total of the model classifies the data accurately. The following is the formula for calculating the accuracy value:

$$Accuracy = \frac{TP+TN}{Total} \quad (1)$$

2. Precision: The accuracy value between the requested data and the prediction results given by the model. The following is the calculation formula:

$$Precision = \frac{TP}{FP+TP} \quad (2)$$

3. Recall: The value of the model's success in retrieving information. The following is the calculation formula:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-Score: The average value of the results of the precision and recall calculations. The following is the calculation formula for f1-score:

$$F1\text{-}Score = 2\ x\ \frac{precision\ x\ recall}{precision+recall} \quad (4)$$

2.3 Phase III Website Development

The process of developing this diabetes prediction information system uses a waterfall model. This model has an initial stage before entering system design, namely system requirements analysis, this stage aims to narrow down existing problems and analyze

system development needs. Before conducting research, the authors must identify the problems that will be raised in the topic. After getting the identification of the problem above, the authors will create a website-based diabetes prediction information system. The information system is to predict diabetes in users so that users can more easily diagnose diabetes. The system analysis will be used as follows in Table 1.

Table 1. System analysis

Hardware	HP 14s laptop, AMD Ryzen 5 5625 with Radeon Graphics, Windows 11 64 Bit, 8.00 GB RAM
Software	Microsoft Office Word 2021, Microsoft Excel 2019, Visual Studio Code, Figma, google colab
Input Data	Diabetes diagnosis indicator form
Output data	Diabetes diagnosis prediction results
Result	No diabetes, Diabetes

After understanding and analyzing system requirements, researchers proceed to the next stage, namely system design to system testing. The flow of how this diabetes prediction information system works is available at Figure 2. The first stage user will fill out the form according to the user's health condition through the website, the second stage information that has been provided by the user will be sent to the back-end website, the third stage Flask server will perform analysis and provide prediction results from data analysis through the random forest method, fourth stage the website displays the prediction results to the user.

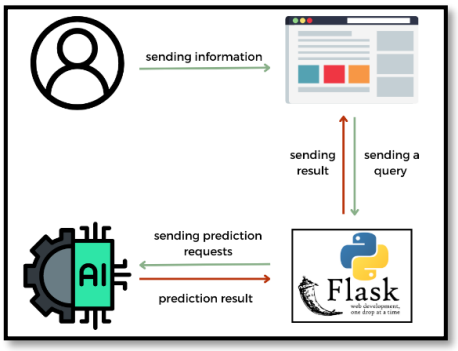


Figure 2. Working Flow Website

There are two stages of testing carried out by the authors are interface testing and validation testing.

1. Interface Testing

Testing the interface / user interface aims to find out the performance of the tools on the website can be used as expected by the authors. There is a test plan for the system interface, namely:

- The home feature is expected to bring users to the description of diabetes on the main page.
- The data input form is expected to be filled in according to the predetermined data type
- The prediction button is expected to display the results of predictions made by the system

2. Validation Testing

Validation testing aims to determine the match between the test results carried out by experts and the results of the analysis carried out by the system. The validation will be carried out by two expert experts as a parameter for the match between the doctor's analysis and the results of data analysis. Testing is done by entering analysis parameters according to the form on the prediction information system to determine the accuracy of the expert system and a questionnaire with yes / no answers to test the feasibility of the system.

2.4 Phase IV Reporting

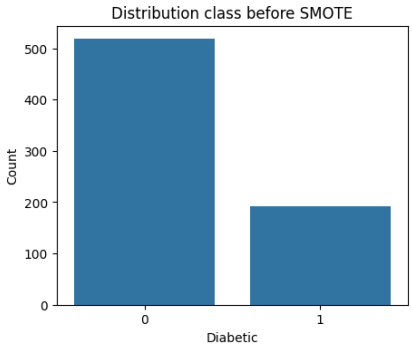
The reporting stage is the final stage of the research at this stage in the form of the process of preparing the final project report. The process of preparing this report includes several important steps that must be followed carefully to ensure the report is comprehensive, accurate, and in accordance with established guidelines. The research results and discussion will be presented in the form of tables, graphs, or clear and informative narrative descriptions, analyzing the research results, relating them to existing theories, and explaining the implications of the findings. In the conclusion section, the authors summarize the main findings of the research and provides relevant recommendations based on the results obtained.

RESULTS AND DISCUSSION

Model development is made through the google collab application with data collection from Kaggle. Website development is made through the vscode application with the flask framework for website display and integration with data analysis modeling (Musse Bekabil, 2014). The following is an explanation of both points 3.1 until 3.3.

3.1 Model development

Based on the images in Figure 3 and Figure 4 before implementing the smote technique, the amount of data for diabetics and non-diabetics is drift about 300 data and after implementing the smote technique, the data for diabetics and non-diabetics is equal in each class with total of 500 data.



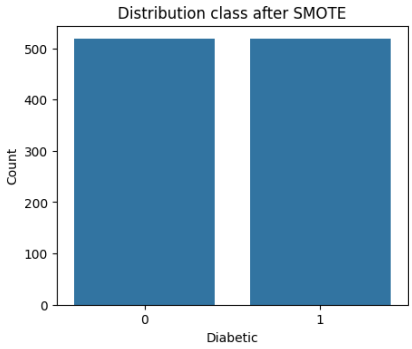


Figure 3. Distribution class before smote **Figure 4.** Distribution class after smote

From the visualization of Figure 6, it can be seen that according to the mapping of the confusion matrix method obtained by the test data as much as 237 data has an accuracy of 0.96 or 96%. This fig.6 diagram visualization determines the loss value of the test data as much as 5 in non-diabetes data and 5 in diabetes data, the value with red and beige

color blocks means that the value of valid data with a total of 159 non-diabetes data and 68 diabetes data. From the visualization of Figure 5, it can be seen that according to the mapping of the confusion matrix method obtained by train data as much as 1021 data has an accuracy of 0.99 or 99%. Visualization diagram Figure 5 determines the loss value of the test data as much as 9 in non-diabetic data and 6 in diabetic data, the value with the beige color block means that the value of valid data with a total of 512 non-diabetic data and 509 diabetic data. the data has undergone oversampling made with the smote technique.

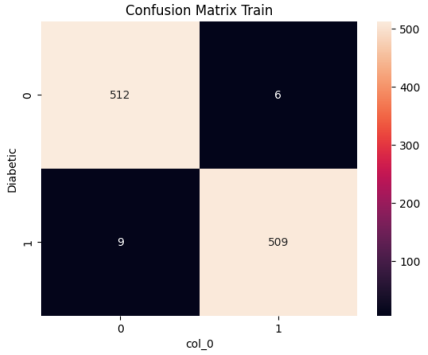


Figure 5. Confusion matrix of train

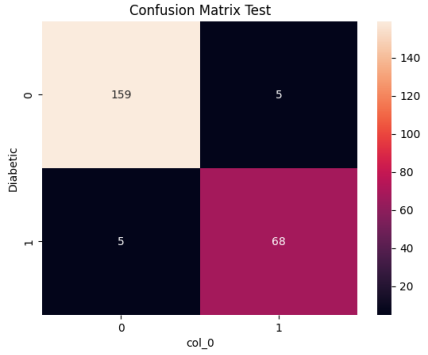


Figure 6. Confusion matrix of test

3.2 System implementation

The development of the diabetes prediction information system is adapted from the waterfall method. The initial stages of system analysis, system design, and wireframe of the system have been discussed in research methodology. The next stage is system implementation by creating the interface of the prediction information system and integrating it into the data analysis modeling that has been made in the previous stage. The following is the display of the diabetes prediction information system website.

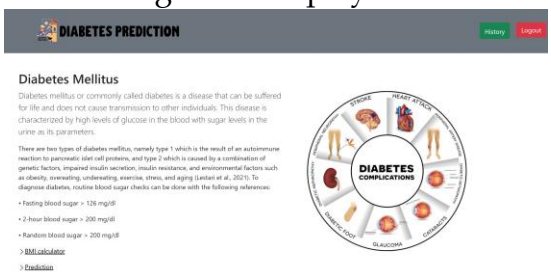


Figure 7. Basic summary

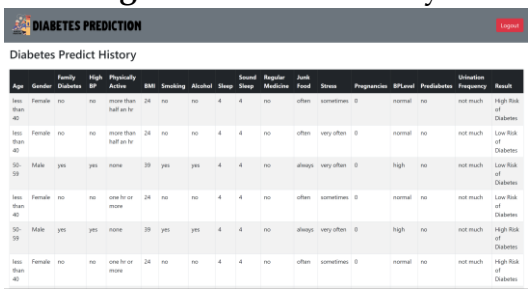


Figure 9. Result history

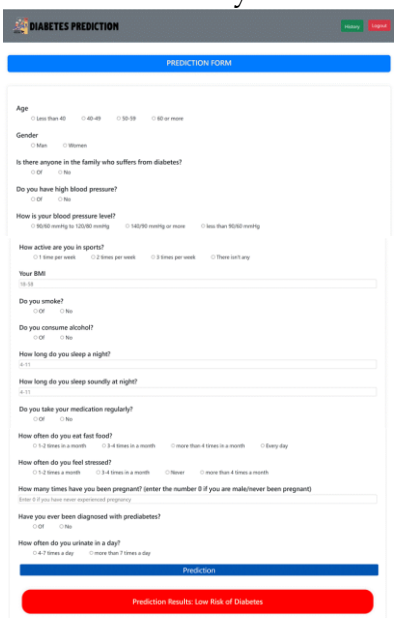


Figure 8 Form predict and result

3.3 System testing

In validation testing, there are two test forms, namely the expert system accuracy feasibility test and the system feasibility test. The following are the test results from both

stages. The testing through experts by comparing the results of system predictions with the results of expert diagnoses. Comparison of diagnostic results can have results that are not the same as each other. This depends on the user filling process on the prediction form and modeling performance.

Table 2. Results of expert system accuracy test

No	Testing	Parameters	Answer	Result System	Diagnoses Doctor
1	Uji 1	Age	40-49	High Risk of Diabetes/ ~~Low Risk of Diabetes~~	High Risk of Diabetes/ ~~Low Risk of Diabetes~~
		Gender	Female		
		family diabetes	yes		
		high blood preasure	no		
		physically active	3-4 times per week		
		BMI	25		
		Smoking	No		
		Alcohol	No		
		Sleep	6		
		sound sleep	6		
		regular medicine	No		
		junkfood	3-4 times a month		
		stress	1-2 times a month		
		blood preasure level	90/60 - 120/80		
		pregnancies	3		
		pdiabetes	yes		
		urination frequency	4-7 times a day		
2	Uji 2	Age	40-49	High Risk of Diabetes/ ~~Low Risk of Diabetes~~	~~High Risk of Diabetes~~/ Low Risk of Diabetes
		Gender	Female		
		family diabetes	not		
		high blood preasure	No		
		physically active	3-4 times per week		
		BMI	23		
		Smoking	No		
		Alcohol	No		
		Sleep	6		
		sound sleep	6		
		regular medicine	No		
		junkfood	1-2 times a month		
		stress	Never		
		blood preasure level	90/60 - 120/80		
		pregnancies	6		
		pdiabetes	yes		
		urination frequency	4-7 times a day		

Table 2 above is an expert system feasibility test table by filling in the expert diagnosis column and will be compared with the system prediction results. Of the two tests, there is one invalid test because the test results are not the same as each other. This is because

filling in the pdiabetes column is not in accordance with the dataset used so that the model cannot provide good analysis. Table 3 below is an information system feasibility test table, the table is filled in by giving a check mark in the column provided with four assessment categories, namely NS (not suitable), LS (less suitable), S (suitable), and VS (very suitable). The test table is filled in by expert who also validate the results of system predictions. The following is a table of information system feasibility test results based on expert.

Table 3. System feasibility testing results

Testing	Question	Scoring Category			
		NS	LS	S	VS
Test 1	System Purpose			√	
	Features of the system			√	
	Order of presentation			√	
	Suitability of the symptoms asked			√	
	The suitability of the prediction results given with the results of expert diagnosis		√		
	Alignment of questions with prediction results			√	
Test 2	System Purpose			√	
	Features of the system			√	
	Order of presentation			√	
	Suitability of the symptoms asked		√		
	The suitability of the prediction results given with the results of expert diagnosis			√	
	Alignment of questions with prediction results		√		

CONCLUSION

In this final project, the authors develop a diabetes prediction information system using Random Forest and SMOTE methods to overcome data imbalance between "diabetes" and "non-diabetes" results. SMOTE is used to equalize data values to improve accuracy, recall, and f1-score. Random Forest algorithm was chosen because it is more complex and optimal than Decision Tree, as described in the journal "Machine Learning Based Diabetes Prediction and Development of Smart Web Application". Model evaluation using confusion matrix showed an accuracy of 96%. This system is proven to be accurate in predicting new data entered through the form on the website. Some of the supporting modules installed include Flask, virtual environment, and others. The file `__init__.py` is required to initialize Flask and ensure the data structure and folders are in accordance with the Flask framework. Based on the results, although the accuracy reached 96%, there is still room for improvement. It is recommended to expand and increase the dataset and implement better optimization. This is expected to improve the accuracy and quality of information system predictions. Improving methods and optimization is also important to improve the prediction results in this diabetes prediction information system.

REFERENCES

Ahdiat, A. (2023, February 10). *Indonesia Punya Penderita Diabetes Tipe 1 Terbanyak di ASEAN*. Databooks.Katadat.Co.Id.

https://databoks.katadata.co.id/datapublish/2023/02/10/indonesia-punya-penderita-diabetes-tipe-1-terbanyak-di-asean

Ahlin, K., & Billhult, A. (2012). Lifestyle changes - A continuous, inner struggle for women with type 2 diabetes: A qualitative study. *Scandinavian Journal of Primary Health Care*, *30*(1), 41–47. https://doi.org/10.3109/02813432.2011.654193

Ahmed, N., Rayhan, A., Md. Manowarul, I., Md. Ashraf, U., Arnisha, A., Md. Alamin, T., & Bikash Kumar. Paul. (2021). Machine Learning Based Diabetes Prediction and Devlopment of Smart Web Application. *International Journal of Cognitive Computing in Engineering*, *2*, 229–241.

Bassil, Y. (2012). A Simulation Model for the Waterfall Software Development Life Cycle. In *International Journal of Engineering & Technology (iJET)* (Vol. 2, Issue 5). http://iet-journals.org/archive/2012/may_vol_2_no_5/255895133318216.pdf

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

data.who.int. (2020). *Indonesia Health Data Overview for the Republic of Indonesia*. Data.Who.Int. https://data.who.int/countries/360

Lestari, Zulkarnain, & ST. Aisyiyah Sijid. (2021). *Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan*. http://journal.uin-alauddin.ac.id/index.php/psb

Medina-Remón, A., Kirwan, R., Lamuela-Raventós, R. M., & Estruch, R. (2018). Dietary patterns and the risk of obesity, type 2 diabetes mellitus, cardiovascular diseases, asthma, and neurodegenerative diseases. In *Critical Reviews in Food Science and Nutrition* (Vol. 58, Issue 2, pp. 262–296). Taylor and Francis Inc. https://doi.org/10.1080/10408398.2016.1158690

Mohajan, D., & Mohajan, H. K. (2023). Basic Concepts of Diabetics Mellitus for the Welfare of General Patients. *Studies in Social Science & Humanities*, *2*(6), 23–31. https://doi.org/10.56397/sssh.2023.06.03

Musse Bekabil, A. (2014). *REST API Implementation with Flask-Python*.

Nudin, S. R., Budi Warsito, & Adi Wibowo. (2022). *Impact of Soft Skills Competencies to predict Graduates getting Jobs Using Random Forest Algorithm*. https://doi.org/10.1109/ICISIT54091.2022.9872669

ombudsman.go.id. (2023, March 1). *Pembatasan Layanan Pasien BPJS Kesehatan Diskriminatif*. Ombudsman.Go.Id. https://ombudsman.go.id/news/r/pembatasan-layanan-pasien-bpjs-kesehatan-diskriminatif

Rollins, J. B. (2015). *Foundational Methodology for Data Science*.

Santra, A. K., & Christy, C. J. (2012). *Genetic Algorithm and Confusion Matrix for Document Clustering*. www.IJCSI.org

Tigga, N. P. (2019). *Diabetes Dataset 2019*. Kaggle.Com. https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019/data