DIABETES MELLITUS ATTRIBUTE CLASSIFICATION USING THE NAIVE BAYES ALGORITHM BASED ON FORWARD SELECTION

Dwi Puji Prabowo[1], Rama Aria Megantara[2)], Ricardus Anggi Pramunendar[3)], and Yuslena Sari[4)]

[1)] Department of Visual Communication Design, Universitas Dian Nuswantoro, Semarang, Indonesia
[2,3)] Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
[4)]Department of Information Technology, Faculty of Engineering, Universitas Lambung Mangkurat, Banjarmasin, Indonesia
e-mail: dwi.puji.prabowo@dsn.dinus.ac.id[1)], aria@dsn.dinus.ac.id[2)], ricardus.anggi@dsn.dinus.ac.id[3)], yuzlena@ulm.ac.id[4)]

Corresponding Author: Dwi Puji Prabowo

ABSTRACT

Diabetes Mellitus is a chronic condition that frequently results in death. Almost every nation has experienced and contributed to this rise in mortality. Consequently, several researchers are motivated to determine this disease's source and prevent the increase in mortality rates. The research was conducted in the field of informatics in partnership with health professionals to determine the causes of this condition. Many informatics researchers employ machine learning techniques to aid in analyzing existing data. This study suggests feature selection based on forward selection and the naive Bayes classification approach to determine this disease's primary aetiology. The results demonstrate that our proposed strategy can increase the classification accuracy of patients. The performance outcomes improved by 169%. According to this theory, it is also known that the primary cause of this disease is its dependence on body mass index and age. Therefore, additional research must explore these two variables' impact on various other disorders.

***Keywords**: Diabetes Mellitus, Classification, naïve Bayes, Forward Selection*

I. INTRODUCTION

DIABETES Mellitus (DM) is one of the most persistent diseases [1]. Diabetes Mellitus is closely associated with the pancreas and occurs when the pancreas produces the natural hormone insulin, but its production is disrupted, resulting in an increase in blood sugar. A high blood sugar level in the human body will impair the function of organs such as the kidneys, heart, and brain [1][2]. According to a WHO report, diabetes mellitus is a major cause of death worldwide, accounting for 82% of all deaths. Four hundred fifty million people worldwide suffer from this disease, according to WHO data. In 2014, 8.5% of the world's adult population was affected by diabetes [3]. In 2012, diabetes caused the deaths of 1,6 million people worldwide. In 2016, high blood sugar caused 2.2 million deaths worldwide. Premature mortality due to diabetes mellitus increased by 5% between 2000 and 2016 [3]. In developed countries, the premature death rate due to diabetes mellitus decreased from 2000 to 2010[4], whereas it increased in lower-middle countries. This is due to a rise in population, age, obesity, irregular eating habits, and lack of physical activity [5]. Identifying potential diabetes risk factors is the initial step in preventing diabetes. Among the influential factors are sociodemographic and behavioural variables [6][7]. The number of collected factors necessitates a discipline able to extract useful information from enormous data sets [8][9]. Typically, information derived from diabetes mellitus risk factors is divided into two categories, diabetes mellitus and not. Current diabetes mellitus research focuses on the risk factor-based classification of diabetes mellitus patients [10-12]. Logistic Regression (LR), Nave Bayes (NB), k-Nearest Neighbor (k-NN), Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) are well-known classification algorithms for predicting whether or not a person has diabetes mellitus [13-18]. Several algorithms have been selected for optimal performance with the appropriate dataset [18]. No single optimal classification algorithm applies to all datasets [19] [20]. The majority utilize the Pima Indian Diabetes Dataset. Popular in developing classification algorithms, this dataset is a metric data collection [13][14] [21-24].

Generally, metric data quality is irrelevant, noisy, and redundant. These characteristics reduce the efficiency of the learning algorithm [25]. Consequently, when presented with irrelevant, noisy, and redundant datasets, this learning algorithm failed to distribute the data and produced inaccurate results. Generally, feature selection is used to improve noisy data, irrelevant data, and redundant data [18]. The three methods of feature selection are embedded, wrapper, and filter. Current research demonstrates that embedded feature selection improves learning performance [18]. Learning algorithms determine whether a feature should be selected or deleted when selecting wrapper features. The feature selection filter is dataset-dependent. P. D. Sheth et al. [26] mentioned that the filter feature selection method is very fast in computation but does not consider the classification algorithm when

selecting attributes. Some classification algorithms have sensitivity in attribute selection. Therefore, a wrapper feature selection method utilizing a classification algorithm to select attributes is proposed.

To contribute to the problem of identifying the attributes that influence the Diabetes Pima Indian dataset, specifically by employing the wrapper feature selection method and forward selection for attribute selection, also classification algorithm as naive Bayes. It is hoped that the most valuable contribution can be obtained to solve the problem of identifying the attributes that influence the Pima Indian Diabetes dataset.

II. RELATED RESEARCH

Multiple researchers have conducted a classification to identify individuals with DM diabetes mellitus. Some researchers, such as Iyer et al. [16], who proposed NB and DT J48 using the WEKA software calculator, have proposed NB and DT J48—using the Diabetes Data Set for Pima Indians dataset. Using ten-fold cross-validation, the results of DT J48 yield an accuracy of 76.96%. In the meantime, NB displays an accuracy of 79.57%.

In their study, Kumari et al. [22] proposed optimization of the SVM algorithm based on the RBF Kernel and utilizing the 768-record Pima Indians Diabetes Data Set. The entire data set is purged of missing data, resulting in 460 data records. They used 10-fold cross-validation to separate training data from test data. This study employs MatLAB 2010a and achieves 78% precision, 80% sensitivity, and 76% specificity.

By D. Sisodia et al. [13], the NB, SVM, and DT algorithms were compared utilizing the Pima Indians Diabetes Data Set. The experiment was conducted using a 10-fold cross-validation design. The three algorithms were compared, and the resulting NB was superior in every investigation aspect, with an accuracy of 76.30%, precision of 75%, recall of 76%, f-measure of 76%, and ROC of 0.81. At the same time, the SVM method yields an accuracy of 65.10%, precision of 42.4%, recall of 65.1%, f-measure of 51.3%, and ROC of 0.5. The DT method yields an accuracy of 73.82%, a precision of 73.7%, a recall of 73.7%, an f-measure of 73.7%, and a ROC of 0.751.

P. Berchilla et al. [17] proposed NB using feature selection based on their research. This study compares NB using the feature selection method with the filter and wrapper approaches. Diabetes among Pima Indians is the dataset used. The results indicated that the NB method was 85.5% accurate, the NB and filter approach was 88.09% valid, and the NB and wrapper approach was 92.7% accurate.

P. D. Sheth et al. [26] also used feature selection with 11 public datasets, including PID, Immunotherapy, WBCD, Fertility, Lung Cancer, Sonar, Wine, Zoo, Ionosphere, Musk 1, and Lymphography. The dataset was obtained from the Machine Learning Repository at UCI. Wrapper Selection is used to select the dataset's attributes, and k-NN is used as the classification algorithm. Each dataset is chosen using the Wrapper Selection attribute, and the k-NN classification algorithm is then applied to the results of PID with six attributes, Immunotherapy with two attributes, WBCD with 12 attributes, Fertility with five attributes, and Lung Cancer with 26 attributes. Experiments with initial processing followed by k-NN classification were also conducted. The outcomes of these experiments demonstrate performance improvements in several datasets (e.g., the WBCD dataset increased by 3.8% to 98%, Musk 1 by 4% to 90.8%, Ionosphere by 6.5% to 93.4%, Lymphography by 11.1% to 97.9%, Sonar by 11.5% to 95.8%, Wine by 5.1% to 98.9%, and Zoo by 2.3% to 99.3%).

Based on the studies described previously, the classification of diabetes mellitus disease data is anticipated to be completed using the NB algorithm and forward selection due to its excellent performance.

III. RESEARCH METHOD

The Pima Indians Diabetes Database was used to gather the data for this study on diabetes mellitus illness. This study seeks to increase the precision of the classification outcomes of diabetes mellitus by using the NB classification algorithm and forward selection as a feature selection technique. This proposal enhances classification performance by removing pointless features from the classification process and identifying the qualities that significantly impact the dataset for diabetes mellitus illness. The suggested research is depicted in Figure 1.

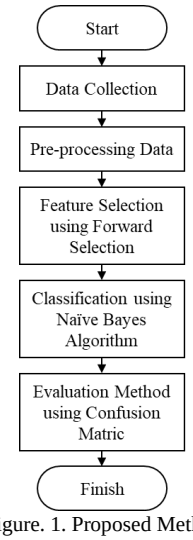

Figure. 1. Proposed Method

A. Data Collection

This study focuses on diabetes mellitus research that can be downloaded from the Pima Indians Diabetes dataset in the UCI Machine Repository Dataset. The Diabetes in Pima Indians dataset contains metric data. Contains one binominal class and eight number properties (Pregnant, Glucose, Diastolic, Triceps, Insulin, BMI, Diabetes History, and Age). The Pima Indians Diabetes database contains 768 records.

TABLE I
PIMA INDIANS DIABETES DATASET

No.	Pregnant	Glucose	Diastolic	Triceps	Insulin	BMI	Diabetes	Age	Result
1	6	148	72	35	0	33.6	0.627	50	**1**
2	1	85	66	29	0	26.6	0.351	31	**0**
3	8	183	64	0	0	23.3	0.672	32	**1**
4	1	89	66	23	94	28.1	0.167	21	**0**
5	0	137	40	35	168	43.1	2.288	33	**1**
6	5	116	74	0	0	25.6	0.201	30	**0**
7	3	78	50	32	88	31	0.248	26	**1**
8	10	115	0	0	0	35.3	0.134	29	**0**
….	….	….	….	….	….	….	….	….	**….**
….	….	….	….	….	….	….	….	….	**….**
….	….	….	….	….	….	….	….	….	**….**
766	0	126	84	29	215	30.7	0.52	24	**0**
767	14	100	78	25	184	36.6	0.412	46	**1**
768	**8**	**112**	**72**	**0**	**0**	**23.6**	**0.84**	**58**	**0**

This study utilized a 768-record diabetes mellitus disease dataset from the Pima Indians Diabetes Database. There are eight (integer data type) characteristics and one class in the dataset (binominal data type). From a total of 768 entries, it was noted that 268 people had diabetes mellitus and 500 did not.

TABLE II
PIMA INDIANS DIABETES DATASET

No	Attribute	Data Type	Description
1.	Pregnant	Integer	Number amount of pregnancies
2.	Glucose	Integer	According to the World Health Organization, one of the diagnostic criteria for diabetes is a plasma glucose concentration of 200 mg/dL or more after two hours of an oral glucose tolerance test.
3.	Diastolic	Integer	Diastolic blood pressure/blood pressure in the diastole (mm Hg)
4.	Triceps	Integer	Triceps skin fold depth (mm)
5.	Insulin	Integer	The early stages of type 2 diabetes are characterized by an insulin excess, as well as type 1 diabetes, also known as insulin-dependent diabetes. Insulin affects the symptoms of diabetes from a medical standpoint and is responsible for many of the physiological and behavioural aspects of the disease.
6.	BMI	Integer	Maintaining body weight so as not to be obese is one way to reduce the risk of diabetes, according to the American Diabetes Association (ADA).
7.	Diabetes	Integer	Histories of diabetes mellitus in the subject's genetically related, inherited relatives
8.	Age	Integer	Age (years)
9.	Result	Binominal	Diabetes test results (positive, negative)

According to [31], eight integer attributes and one binominal class dataset of Diabetes in Pima Indians were converted into an example format so that the data could be identified and utilized as a model for the techniques and methods that have been determined.

B. Feature Selection

Feature selection is a fundamental data preprocessing technique used in data mining to enhance performance and accelerate algorithm processing. The method can select a subset of a sufficient number of features, thereby reducing or eliminating irrelevant or less influential features for classification. The primary concept of feature selection is selecting a subset of existing features, as not all features are relevant or significant to the problem. On the other hand, some of the existing features can cause interference and reduce the level of precision, so features with no effect must be eliminated to increase the precision value [27]. Forward selection is an element of the wrapper approach method for selecting features. Wrapper feature selection is frequently used because its performance outcomes are superior to those of the filter, embedded, and hybrid techniques. This is because the wrapper approach can generate a subset of better-suited features for classification, resulting in greater accuracy.

Based on the search strategy, the wrapper approach is divided into two categories: sequential search and random search. In the category of sequential search, forward selection and backward elimination are included with the concept of adding or removing features sequentially. Forward selection is a modelling technique that begins with an empty model (zero variables) and adds variables incrementally until a set of criteria is satisfied. Forward selection outperforms backward elimination because the process is more efficient and scalable for large datasets. Listed below are the stages of the forward selection method:

1) *Using regression, the attributes of the dataset will be calculated. Where the first attribute enters the model, select the attribute with the highest correlation to the target. If the obtained model is not significant, the calculation will stop, report that there are no significant attributes, and proceed to the next step.*

$$r = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{\{n \sum X_i^2 - (\sum X_i)^2\}\{n \sum Y_i^2 - (\sum Y_i)^2\}}} \quad (2)$$

Description

n = copious amounts of data
X = independent variable
Y = dependent variable
$\sum X_i Y_i$ = number of multiples of the independent variable by the dependent variable
$\sum X_i$ = sum of all independent variables
$\sum Y_i$ = sum of all dependent variables
$\sum X_i^2$ = square sum of the independent variable
$\sum Y_i^2$ = square sum of the dependent variable

The following criteria determine the correlation relationship's strength:

0 : There is no association between two variables
$0 < r < 0.25$: The observed correlation is very weak
$0.25 < r < 0.5$: The correlation is adequate
$0.50 < r < 0.75$: Significant correlation
$0.75 < r < 0.99$: Extremely robust correlation
1 : Absolute correlation

2) *After calculating with the aforementioned formula, the highest correlation value can be determined, and attributes with a high (significant) correlation value can be used as targets for further calculations. The following calculation is performed using the multiple correlation formula:*

$$r_{y.x1.x2} = \sqrt{\frac{r_{yx1}^2 + r_{yx2}^2 - 2\, r_{yx1}\, r_{yx2}\, r_{x1x2}}{1 - r_{x1x2}^2}} \quad (3)$$

3) *The calculation above determines the correlation between variables or attributes until the highest value of the multiple correlation coefficient is determined. This calculation will be repeated numerous times. The NB classifier algorithm is then used to classify the selected properties.*

C. Naive Bayes Classification

Naive Bayes (NB) is a straightforward probabilistic classifier that computes a set of probabilities by summing the frequencies and combinations of values from a given dataset. The algorithm employs Bayes' theorem and assumes all attributes are independent or interdependent based on the class variable's value. NB assumes that attribute values are conditionally independent if an output value is given. In other words, the probability of observing collectively, given the output value, is the product of the individual probabilities. The advantage of using NB is that it only requires a small amount of training data to determine the classification process's required parameter estimates. NB frequently outperforms expectations in the majority of complex real-world situations. Several advantages of NB over other theories include Interpolation, Languages, and instincts. NB can be used as a decision-making tool to update the information's confidence level. The NB theory is one of the branches of a mathematical-statistical approach that permits us to create a model of an event's uncertainty by combining general knowledge with observational data.

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

The following is an explanation of each variable:
X = Unknown class of data
H = Hypothesis data is a particular category
P(H|X) = Probability of H given X (posteriori probability)
P(H) = Probability of the hypotheses H (prior probability)
P(X|H) = The probability of X under the conditions of the hypothesis. H
P(X) = Probability X

D. Accuracy Evaluation

The objective of the evaluation is to describe the outcomes of the examined data. The evaluation phase is used to assess the method's precision. The assessment is conducted by dividing the total number of test data within the dataset by the number of recognized test data. The equation is used to compute accuracy (5).

$$accuracy = \frac{known\ data}{total\ data} \times 100\% \tag{5}$$

IV. Result and Discussion

The dataset's conversion results are displayed in Table III. The acquired results are consistent with [28]. The objective is to discover the point with the highest R-value by calculating all attributes using the simple regression correlation algorithm. Table IV displays the results of the computation of the correlation between the quality of X1 through X8 and the dependent variable Y.

Table III
Dataset after preprocessing

No.	Pregnant	Glucose	Diastolic	Triceps	Insulin	BMI	Diabetes	Age	Result
1	Grande Multipara	Medium	Low	High	Low	Fat	Medium	Old	1
2	Primpara	Good	Low	High	Low	Fat	Low	Young	0
3	Grande Multipara	Bad	Low	Low	Low	Normal	Medium	Young	1
4	Primpara	Good	Low	High	Normal	Fat	Very Low	Young	0
5	Primpara	Good	Low	High	High	Very Fat	Very High	Young	1
6	Multigravida	Good	Low	Low	Low	Fat	Very Low	Young	0
7	Multigravida	Good	Low	High	Normal	Fat	Low	Young	1
8	Grande Multipara	Good	Low	Low	Low	Very Fat	Very Low	Young	0
….	….	….	….	….	….	….	….	….	….
….	….	….	….	….	….	….	….	….	….
….	….	….	….	….	….	….	….	….	….
766	Primpara	Good	Normal	High	High	Fat	Low	Young	0
767	Grande Multipara	Good	Low	High	High	Very Fat	Low	Old	1
768	Grande Multipara	Good	Low	Low	Low	Normal	High	Old	0

According to the results of the correlation coefficient table presented previously, the correlation connection with the highest value is X6, which corresponds to the BMI characteristic. Multiple linear regression is then used to calculate the correlation coefficient between X6, Y, and qualities other than X6 itself, after obtaining the first correlation coefficient value, X6 using (3).

The method for calculating the table of correlation between X6, Y, and X1 to X5, X7, and X8 is same to that for simple regression correlation. X6 was identified as an attribute having a relationship and effect throughout the classification process, resulting in a substantial correlation between the findings of X6, Y, and other characteristics,

for which the multiple linear regression correlation formulas are not utilized. Table IV display correlation coefficient results from simple regression x1-x8 with y.

TABLE IV
CORRELATION COEFFICIENT RESULTS FROM SIMPLE REGRESSION X1-X8 WITH Y

No	Combination	Result
1	X1 dan Y	2.0105
2	X2 dan Y	2.0383
3	X3 dan Y	1.8069
4	X4 dan Y	1.7552
5	X5 dan Y	1.7047
6	X6 dan Y	3.2642
7	X7 dan Y	1.6849
8	X8 dan Y	2.3787

There is a significant association between X6, Y, and X1-X5, X7, and X8, with X8, age, having the greatest correlation coefficient value (as in table V). The calculation is then completed using the method for multiple regression correlation to identify attributes with a strong and significant correlation. The final result of the correlation coefficient calculation indicates that four traits are highly associated with the dependent variable Y, notably X6 (BMI) and X8 (Age).

TABLE V
RESULTS OF MULTIPLE REGRESSION CORRELATION COEFFICIENTS BETWEEN X6 AND Y AGAINST X1-X5, X7, AND X8

No	Combination	Result
1	X6, Y dan X1	Ry.x6.x1 = 2.1912
2	X6, Y dan X2	Ry.x6.x2 = 1.7051
3	X6, Y dan X3	Ry.x6.x3 = 1.8061
4	X6, Y dan X4	Ry.x6.x4 = 1.3995
5	X6, Y dan X5	Ry.x6.x5 = 1.7860
6	X6, Y dan X7	Ry.x6.x7 = 1.9418
7	X6, Y dan X8	Ry.x6.x8 = 2.3705

Based on the selected quality, classification will be performed using the Naive Bayes Classifier method to determine the diabetes mellitus classification, with the decision outcome being the probability of classification. The effect of feature selection is displayed in table VI.

TABLE VI
DATASET AFTER FEATURE SELECTION USING FORWARD SELECTION

No.	BMI	Age	Result
1	Fat	Old	1
2	Fat	Young	0
3	Normal	Young	1
4	Fat	Young	0
5	Very Fat	Young	1
6	Fat	Young	0
7	Fat	Young	1
8	Very Fat	Young	0
….	....	....	....
….	....	....	....
….	....	....	....
766	Fat	Young	0
767	Very Fat	Old	1
768	Normal	Old	0

Table VI displays the findings of diabetes mellitus illness classification with a negative classification (0), as the posterior probability value of diabetes mellitus with a negative result (0) is greater than the value of diabetes mellitus with a positive result (1). From the calculation utilizing all test data, then testing using a confusion matrix diagram, the NB algorithm's forward selection accuracy value is determined. Using the confusion matrix displayed in Table VII, the accuracy of the forward selection-based NB technique.

TABLE VII
CONFUSION MATRIC FOR NB USING FORWARD SELECTION

	TRUE 1	**TRUE 0**
PRED. 1	136	49
PRED. 0	132	451

The 768-dataset categorization results include true positive results in 136 data, false positive results in 49 data, false negative outcomes in 132 data, and true negative effects in 451 data. Based on this, the accuracy value for Table VII of the confusion matrix is 76,43%.

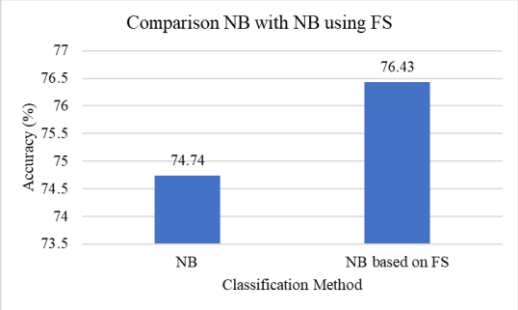

Figure.2. Proposed Method

This research demonstrates that our proposed strategy can enhance accuracy performance. The increase was reached by employing a forward selection-based strategy for feature selection. When the forward selection approach is compared to the original method without the selection feature, a gain of 1.69% is observed (as shown in Figure 2). The NB method's accuracy is 74.74% without the selection feature and climbs to 76.43% with the selected feature. Since feature selection eliminates unnecessary aspects and retains influential ones, it is possible to attain enhancements. However, the usage of this feature selection has ignored some additional information. Therefore, it is feasible that this feature selection approach can still be enhanced to increase classification performance.

V. CONCLUSION

In this study, the Diabetes dataset for Pima Indians was classified using Naive Bayes with forward selection. This study's findings demonstrate that the dataset's accuracy improves when the recommended idea is implemented. The implementation of this study suggestion resulted in a 1.69% improvement in accuracy performance, for a total performance of 76.43%. In addition, this research proposal presents significant conclusions concerning the influence of body mass index and age on the incidence of diabetes mellitus. The forward selection strategy yielded these two factors and rejected others, including pregnancy, glucose, diastolic, triceps, insulin, and diabetes. So that future research should be able to analyze these two features in conjunction with others..

REFERENCES

[1] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," Clin. Epidemiol. Glob. Heal., vol. 7, no. 4, pp. 530–535, 2019, doi: 10.1016/j.cegh.2018.12.004.
[2] S. A. Paschou, G. I. Sydney, K. J. Ioakim, K. Kotsa, and D. G. Goulis, "Comment on the systematic review and meta-analysis titled 'Gestational diabetes and the risk of cardiovascular disease in women,'" Hormones, vol. 19, no. 3, pp. 447–448, 2020, doi: 10.1007/s42000-019-00158-w.
[3] N. Sarwar et al., "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies," Lancet, vol. 375, no. 9733, pp. 2215–2222, 2010, doi: 10.1016/S0140-6736(10)60484-9.
[4] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digit. Signal Process. A Rev. J., vol. 17, no. 4, pp. 702–710, 2007, doi: 10.1016/j.dsp.2006.09.005.
[5] F. Milita, S. Handayani, and B. Setiaji, "Kejadian Diabetes Mellitus Tipe II pada Lanjut Usia di Indonesia (Analisis Riskesdas 2018)," 2018.
[6] Perkumpulan Endokrinologi Indonesia, "Konsensus Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2," Diss. Abstr. Int. Sect. A Humanit. Soc. Sci., vol. 71, no. 2-A, p. 730, 2015, [Online]. Available: http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3393923%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc7&NEWS=N&AN=2010-99150-140.
[7] J. J. Pangaribuan, "Mendiagnosis Penyakit Diabetes Melitus Dengan Menggunakan Metode Extreme Learning Machine," J. ISD, vol. 2, no. 2, pp. 2528–5114, 2016.
[8] M. North, Data Mining for the Masses. 2012.
[9] I. H. Witten, E. Frank, and M. A. Hall, Data mining. 2011.
[10] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," J. Intell. Learn. Syst. Appl., vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/jilsa.2017.91001.
[11] T. A. ASFAW, "Prediction of Diabetes Mellitus Using Machine Learning Techniques ," Int. J. Comput. Eng. Technol., vol. 10, no. 4, pp. 145–148, 2019, doi: 10.34218/ijcet.10.4.2019.004.
[12] M. M. Islam, M. J. Rahman, D. Chandra Roy, and M. Maniruzzaman, "Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach," Diabetes Metab. Syndr. Clin. Res. Rev., vol. 14, no. 3, pp. 217–219, 2020, doi: 10.1016/j.dsx.2020.03.004.
[13] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.
[14] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Comput. Sci., vol. 167, no. 2019, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.
[15] G. Naveen Kishore, V. Rajesh, A. Vamsi Akki Reddy, K. Sumedh, and T. Rajesh Sai Reddy, "Prediction of diabetes using machine learning classification algorithms," Int. J. Sci. Technol. Res., vol. 9, no. 1, 2020.
[16] A. Iyer, J. S, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 1, pp. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.

[17] P. Berchialla, F. Foltran, and D. Gregori, "Naïve Bayes classifiers with feature selection to predict hospitalization and complications due to objects swallowing and ingestion among European children," Saf. Sci., vol. 51, no. 1, pp. 1–5, 2013, doi: 10.1016/j.ssci.2012.05.021.
[18] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction : Research Trends , Datasets , Methods and Frameworks," vol. 1, no. 1, 2015.
[19] V. U. B. Challagulla, F. B. Bastani, and R. a. Paul, "Empirical Assessment of Machine Learning based Software Defect Prediction Techniques," 10th IEEE Int. Work. Object-Oriented Real-Time Dependable Syst., pp. 263–270, 2005, doi: 10.1109/WORDS.2005.32.
[20] Q. Song, Z. Jia, M. Shepperd, S. Ying, and J. Liu, "A general software defect-proneness prediction framework," IEEE Trans. Softw. Eng., vol. 37, no. 3, pp. 356–370, 2011, doi: 10.1109/TSE.2010.90.
[21] C. Hapter, "G Enetic a Lgorithms for D Esigning," vol. 2, no. 1, pp. 83–102, 2011.
[22] K. and Chitra, "Classification Of Diabetes Disease Using Support Vector Machine," vol. 3, no. 2, pp. 1797–1801, 2018, [Online]. Available: https://www.researchgate.net/publication/320395340.
[23] D. R. Edla and R. Cheruku, "Diabetes-Finder: A Bat Optimized Classification System for Type-2 Diabetes," Procedia Comput. Sci., vol. 115, pp. 235–242, 2017, doi: 10.1016/j.procs.2017.09.130.
[24] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics Med. Unlocked, vol. 10, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.
[25] H. He and E. A. Garcia, "Learning from Imbalanced Data," {IEEE} Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/tkde.2008.239.
[26] P. D. Sheth, S. T. Patil, and M. L. Dhore, "Evolutionary computing for clinical dataset classification using a novel feature selection algorithm," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2021, doi: 10.1016/j.jksuci.2020.12.012.
[27] M. F. Nugroho and S. Wibowo, "Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes," *J. Inform. Upgris*, vol. 3, no. 1, pp. 63–70, 2017, doi: 10.26877/jiu.v3i1.1669.
[28] Y. Sinatrya and L. A. Wulandhari, "Deteksi Diabetes Melitus Untuk Wanita Dan Penyusunan Menu Sehat Dengan Pendekatan Adaptive Neuro Fuzzy Inference System (Anfis) Dan Algoritma Genetika (Ga)," *J. Tek. Inform.*, vol. 12, no. 1, pp. 39–58, 2019, doi: 10.15408/jti.v12i1.9578.