ICORHESTECH 2024
Vol. 1, No. 1, 2024, pp. 447~453
Journal Homepage: https://journal.ibrahimy.ac.id/index.php/icorhestech

Comparison Support Vector Machine and Random Forest Algorithms in Detect Diabetes

Habib Alrasyid[1], Ahmad Homaidi, M. Kom.[2], Zaehol Fatah, M. Kom.[3*]
[1]Sistem Informasi, Fakultas Sains & Teknologi Universitas Ibrahimy, Indonesia
[2] Teknologi Informasi, Fakultas Sains & Teknologi Universitas Ibrahimy, Indonesia
[3]Sistem Informasi, Fakultas Sains & Teknologi Universitas Ibrahimy, Indonesia

*Corresponding author: *habibalrasyid5@gmail.com*

ABSTRACT

One of the diseases that are very concerning and numerous cases that occur all over the world because the impact is very significant is diabetes. Diabetes sufferers experience disorders in metabolism that identify hyperglycemia caused by no the inability of the pancreas to secrete insulin, which has an impact on death Because No functioning of other body organs. Data states in 2019 that, 433 million people were diagnosed with diabetes, and the Number is predicted to increase until peaking in 2045 will be 700 million people. This matter needs to be anticipated as soon as possible, perhaps by society, with several characteristics that occur in patients. Data on diabetes sufferers can processed with data mining that utilizes machine learning to detect diabetes. Study This will compare the Support Vector Machine and Random Forest algorithms to find accurate results. Researchers use the KDD (Knowledge Discovery in Database) model in several stages, such as data selection, preprocessing, Transformation, and Evaluation. The dataset used was sourced from the kaggel.com website; there were 768, consisting of 500 negative and 268 positive for diabetes. The SVM algorithm with a linear kernel produces a mark accuracy of 77%, Precision of 75%, Recall of 51%, and F1 score of 61%. For the Random Forest algorithm with n_estimators =100, random_state =42, results mark Accuracy 75%, precision 69%, recall 55% and F1 score 61%. The process and results state that more SVM algorithms are suitable for detecting diabetes. Models made using the Python programming language will be implemented with stremlit so you can use Web-based.

Keywords:

Classification; Data Mining; SVM; Random Forest, Diabetes

INTRODUCTION

Because of its significant role in society, diabetes has become a matter of great concern worldwide. Between the effects experienced, Diabetics are disorders in metabolism that identify hyperglycemia caused by disability pancreas for secreting insulin or disturbance in insulin action. In phase chronic hyperglycemia, ongoing damage and disability in the functioning of growing organs such as kidneys, eyes, and blood vessels occur. According to the International Diabetes Federation (IDF), in 2019, 433 million people were suffering from diabetes. With a pattern life that is not healthy, this total estimated will increase to 578 million in 2030 and 700 million in 2045. Indonesia is one of the countries with the highest number of diabetes sufferers in 2019, ranking ten (Bingga, 2021).

Indonesian people know diabetes by its Name. Diabetes is increasing blood sugar levels in the body, especially after eating. Increasing blood pressure above normal 120mg/dl describes hypertension as a symptom of diabetes (Rahayu & Qurrota, 2022). Other influencing factors for Diabetes sufferers include (Ikhromr et al., 2023). With factual data existing in reality, action began fast in handling early diabetes. To handle diabetes, of course, with several existing symptoms displayed. With Diabetes prediction processing, big data from diabetics is later processed with method machine learning. Several research studies have been done on detecting diabetes using the KNN algorithm (Argina) and algorithm and Naïve Bayes (Putry et al., 2022).

Between Lots, the existing algorithms are SVM and Random Forest matching algorithms in the finish problem classification of diabetes. Support Vector Machine Work by looking for a separator from A, which is the most optimal space in a dataset in different classes (Apriyani, 2020). The Random Forest algorithm removes part of the big data, replacing the random sample (Sistem et al., 2021). The second algorithm will process data to produce knowledge in detecting diabetes with statistical, mathematical, and machine learning processes internal processes find This Can called data mining (Maulida, 2020).

Study This aim is to compare the best SVM and random forest algorithms for detecting diabetes. The best algorithm will be assessed based on mark performance, Accuracy, Precision, and Recall. Later, it will be implemented as a web version that uses streamlit so you can use it easily.

RESEARCH METHODS

Research flow This consists of dataset collection, data preprocessing, the classification uses SVM and Random Forest algorithms, model evaluation, and implementation of classification models in a web with stremlit, as for the process in Figure 1

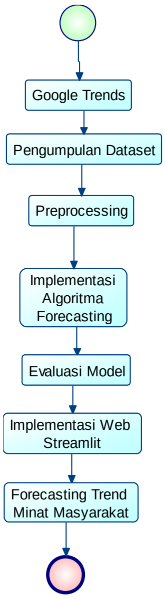

Google Trends
Pengumpulan Dataset
Preprocessing
Implementasi Algoritma Forecasting
Evaluasi Model
Implementasi Web Streamlit
Forecasting Trend Minat Masyarakat

Figure 1. Research Method Flow

Datasets & Data Splitting

The researcher used open-source online data on the kaggle.com website, which had 768 data sets. Splitting data divides a dataset into 80% training and 20% testing data. Training data on duty as a shaper model pattern while testing data as an examiner from model.

SVM algorithm

as system-trained learning through algorithm learning based on optimization and use hypothesis with extensive inner linear function dimensional features (Puspitasari et al., 2018). Support vector machines are method algorithms that classify and predict by looking for room separators from optimal edges on the dataset with various classes (Apriyani, 2020). Classification is done with an SVM algorithm that uses training data for a classification model, and then a model is formed to predict a new data class that does not. This is previously known as data testing (Kurniawan & Falentina, n.d.).

$$f(x_d) = \sum_{i-1}^{ns} a_i y_i \vec{x_i} \vec{x}_d + b$$

Keterangan:
ns = Jumlah support vector
αi = Nilai bobot setiap titik data
yi = Data kelas
$\vec{x}i$ = Variabel *support vector*
$\vec{x}d$ = Daya yang akan diklasifikasi
b = Nilai bias atau error

Random Forest algorithm

Random Forest is an algorithm that reaches the end node in structure tree classification and regression using recursive binary splitting. This algorithm has several advantages, including low error results, good classification performance, and the ability to deal with missing data. Using bootstrap from the sample train and input variables at each node, the algorithm forest random can produce Lots of trees independent of the selected subset randomly (Pamuji & Ramadhan, 2021).

Model Evaluation

Matiks on evaluation are used to evaluate each Accuracy, Precision, and recall model. Ways of working from the evaluation matrix Can Be known with a confusion matrix, namely, processing data to compare results and predict the actual label (Suryati & Aldino, 2023).

Table 1. Confusion Matrix

Actual	**Prediction**	
	Positif	**Negatif**
Positif	True Positif (TP)	True Negatif (TN)
Negatif	False Positif (FP)	False Negatif (FN)

Accuracy

Accuracy merupakan perbandingan pada jumlah prediksi yang benar dengan total jumlah data yang diprediksi. Skor ini akan menggambarkan seberapa baik model melakukan sebuah prediksi secara keseluruhan.

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}} \qquad \textbf{(2)}$$

b. Recall

Recall mengukur banyak dari keseluruhan data positif yang berhasil teridentifikasi dengan benar pada model. Recall menggambarkan sebarapa baik model dapat menemukan kelas tertentu.

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \qquad \textbf{(3)}$$

c. Precission

Precission mengukur banyaknya prediksi positif yang benar dari semua prediksi positif yang dilakukan. Presisi dapat menggambarkan seberapa akurat model dalam mengidentifikasi sebuah kelas.

$$\boldsymbol{Precission = \frac{TP}{TP + FP}} \qquad \textbf{(4)}$$

Application Development with Stremlit

The model that has been created can be developed with a simple and interactive interface using the Streamlit framework (Syafarina, 2023). This framework is designed to be able to build data mining applications that use machine learning and data science systems. Forecasting models that have been integrated can be used on a web basis, by determining the number of months in the selected period.

RESULTS AND DISCUSSION

Dataset Preparation & Preprocessing

The data that has been collected is cross checked to correct data that is null and data that has unclear variables so that errors do not occur in the forecasting process.

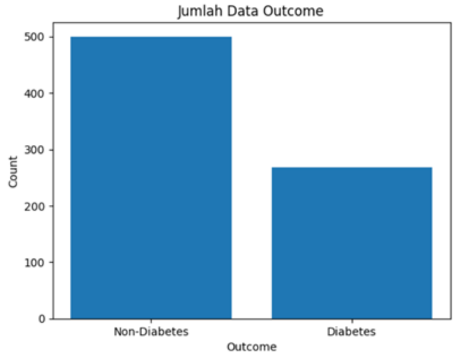


Figure 2. Diabates dataset

After collecting the data, the researcher split it into two: 20% testing data and 80% training data. One hundred fifty-four data records were used for testing, and 614 data were used for training.

Table 2. Data Column Head

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
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

Support Vector Machine Algorithm

Implementation from the results dataset collection is carried out using the programming language Python with a Jupyter notebook for classification using the Support vector machine algorithm. In this algorithm, a value, namely "kernel=linear," is used, which produces a confusion matrix in the image below.

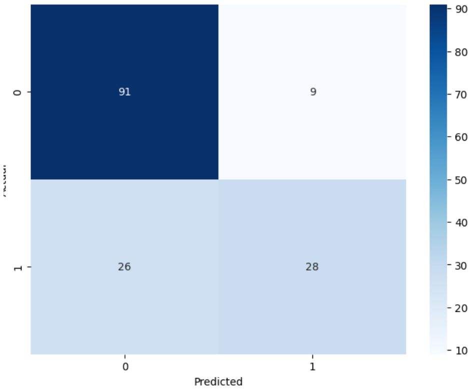


Figure 3. Support Vector Machine confusion matrix result value

After the confusion matrix value of the Support Vector Machine algorithm is obtained, calculation accuracy, Precision, and Recall are calculated. The calculation was done using the Python library " sklearn.metrics," and the results are shown in the table below.

Table 3. Support Vector Machine Algorithm Calculated Value

Accuration = $\frac{91 + 28}{91 + 9 + 26 + 28} = 77\%$

Precission = $\frac{28}{9 + 28} = 75\%$

Recall = $\frac{28}{26 + 28} = 51\%$

Random Forest Algorithm

Implementation from the results dataset collection is carried out using the programming language Python with a Jupyter Notebook for classification using the Random Forest algorithm. In the algorithm, a value, namely " n_estimators =100, random_state =42," produces a confusion matrix in the image below.

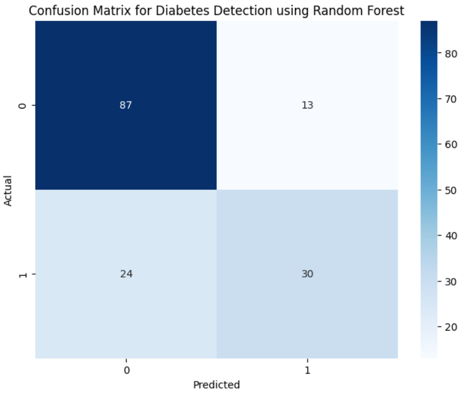


Figure 4. Random Forest confusion matrix result value

After the confusion matrix value of the Random Forest algorithm is obtained, calculation accuracy, Precision, and Recall are calculated. The calculation was done using the Python library " sklearn.metrics," and the results are shown in the table below.

Table 4. Random Forest Algorithm Calculated Value

Accuration = $\frac{30+87}{30+87+24+13} = 75\%$

Precission = $\frac{30}{30 + 13} = 69\%$

Recall = $\frac{30}{24 + 30} = 51\%$

Comparison Algorithms & Model Evaluation

From both results algorithms, the best algorithm can be compared to the third variable, Accuracy, Precision, and Recall, as shown in the table below, to determine performance.

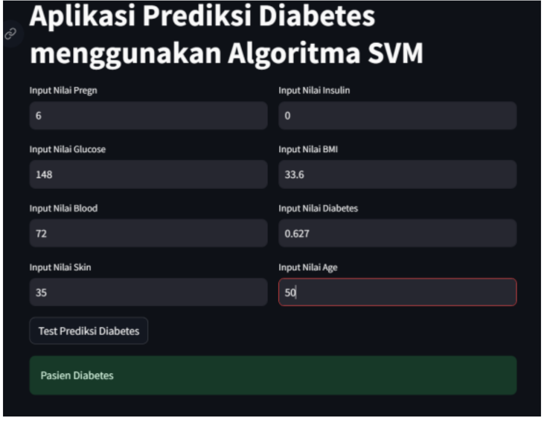
Table 5. Comparison of the two algorithms

Algorithm	Accuration	Precission	Recall
Support Vector Machine	77%	75%	51%
Random Forest	75%	69%	51%

Comparison results between the Support Vector Machine and Random Forest algorithms show that, with notes rounded values, the Support Vector Machine algorithm has more values tall, with an accuracy of 77% and a precision of 76%.

Classification of Diabetes using Streamlit

Results from the Support Vector Machine model that has been created are then carried out through a Pickle import process so that they can be used from the website using the streamlet library. The user application can enter and mark existing criteria listed. Then, the results classification will be obtained.



Aplikasi Prediksi Diabetes menggunakan Algoritma SVM

Input Nilai Pregn: 6

Input Nilai Insulin: 0

Input Nilai Glucose: 148

Input Nilai BMI: 33.6

Input Nilai Blood: 72

Input Nilai Diabetes: 0.627

Input Nilai Skin: 35

Input Nilai Age: 50

Pasien Diabetes

Figure 5. Model implementation using streamlit

According to the results of the diabetes classification using the support vector machine algorithm, if the mark accuracy is 78%, then in 10 trials, 8 data will be classified as true, and 2 data will be classified incorrectly.

CONCLUSION

This study compares Support Vector Machine and Random Forest algorithms for classifying diabetes with the dataset taken from kaggle.com, which is open source. Studies show that the Support Vector Machine algorithm has marked greater Accuracy and Recall, 0.78 and 0.76, respectively. The calculation results were made using Language Python programming, Jupyter Notebook, and available Python libraries. Thus, the Support Vector Machine algorithm is assessed for good performance in classifying this diabetes disease. Next, the model was trained and implemented with an easy web-based framework.

SUGGESTIONS

Study This, of course, has many things that could be improved in research. Furthermore, we can use second algorithms in research. This is for the classification of diabetes. Data can be looked for. More other data references are complex for maximizing results classification, increasing outlook science, and becoming an alternative public in overcoming diabetes

REFERENCES

Amal, I., Pamungkas, E. W., Kom, S., Kom, M., & Ph, D. (n.d.). *APLIKASI PENDETEKSI BERITA PALSU BAHASA INDONESIA MENGGUNAKAN FRAMEWORK FLASK DAN STREAMLIT SERTA ALGORITMA MACHINE LEARNING Teknik Informatika ,(2019….*)1–18.

Apriyani, H. (2020). *Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus*. *1*(3), 133–143.

Bingga, I. A. (2021). *Kaitan kualitas tidur dengan diabetes melitus tipe 2*.

Fikri, M., Syahbani, N., & Ramadhan, N. G. (2023). *Klasifikasi Gerakan Yoga dengan Model Convolutional Neural Network Menggunakan Framework Streamlit*. *7*, 509–519. https://doi.org/10.30865/mib.v7i1.5520

Health, P. (2022). *1 . Improving Care and Promoting Health in Populations : Standards of Medical Care in Diabetes — 2022*. *45*(January), 8–16.

Ikhromr, F. N., Sugiyarto, I., Faddillah, U., Sudarsono, B., Mandiri, U. N., Bina, U., & Informatika, S. (2023). *Implementasi data mining untuk memprediksi penyakit diabetes menggunakan algoritma naives bayes dan k-nearest neighbor implementation of data mining to predict diabetes disease using naives bayes and k-nearest neighbor algorithms*. *6*, 416–428.

Kurniawan, M. A., & Falentina, A. T. (n.d.). *Analisis Big Data dan Official Statistics dalam Melakukan Nowcasting Pertumbuhan Ekonomi Indonesia Sebelum dan Selama Pandemi*. *19*, 521–532.

Maulida, A. (2020). *Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes*. *1*(2), 29–33.

Pamuji, F. Y., & Ramadhan, V. P. (2021). *Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy*. *7*(1), 46–50.

Puspitasari, A. M., Ratnawati, D. E., & Widodo, A. W. (2018). *Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine*. *2*(2), 802–810.

Putry, N. M., Sari, B. N., Kom, M., Informatika, T., & Karawang, U. S. (2022). *KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELITUS*. *10*(1).

Rahayu, P. T., & Qurrota, A. (2022). *Jurnal Smart Teknologi Perbandingan Algoritma K-Nearest Neighbor Dan Gaussian Naïve Bayes Pada Klsifikai Penyakit Diabetes Melitus Comparison Of K-Nears Neighbor And Gaussian Naïve Bayes Algorithm On The Classification Of Diabetes Mellitus Jurnal Smart Teknologi*. *3*(4), 366–373.

Sistem, R., Pasien, D., & Diabetes, P. (2021). *JURNAL RESTI Perbandingan Support Vector Machine dan Modified Balanced Random*. *1*(10), 393–399.

Suryati, E., & Aldino, A. A. (2023). *Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)*. *4*(1), 96–106.

BIOGRAPHIES OF AUTHORS

	Habib Alrasyid Born in Songgon, Banyuwangi, East Java, on February 13 2001, to Sutrisno and Arofah, he is a santri student and one of the media crew at the Sukorejo Salafiyah Syafiiyah Islamic boarding school; he is a student at Ibrahimy Situbondo University with an Information Technology study program. Has an interest in videography and editing using the Premier Pro, Capcut & Aftereffect applications and has produced many works in the form of videos that educate and provide a lot of information to the broader community through the S3tv YouTube channel.