http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.2, Juni 2022 **ISSN : 2302-9706**

APPLICATION OF DATA MINING TO IDENTIFY DIABETES MELLITUS USING THE SUPPORT VECTOR MACHINE (SVM) ALGORITHM AND KNN

Windania Purba, Yessy, Riski Nofarianus Gulo
Program Studi Sistem Informasi, Universitas Prima Indonesia, Medan, Indonesia
Email : Windania@unprimdn.ac.id, yessytan3@gmail.com , riskygulo20@gmail.com

Abstract

Article Info
Received : 10 May 2022
Revised : 30 May 2022
Accepted : 30 June 2022

Damage to the performance of human organs is very detrimental And is the source of the most problems at this time. One of the diseases that is the number one killer in the world is diabetes mellitus. Diabetes mellitus is a metabolic disease characterized by hyperglycemia caused by and obstacle in insulin secretion from insulin action or both. Diabetes mellitus is divided into several types, type 1 diabetes mellitus generally gives rise to indications before the patient is 30 years old. Although in fact the indications of the disease can arise at any time. This study aims to apply the Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) method to identify diabetes mellitus and calculate the comparison value of the accuracy of the two algorithms. From the results of this study. It can be concluded that the Support Vector Machine (SVM) algorithm produces an accuracy value of 76% while the accuracy value of the K-Nearest Neighbor (KNN) algorithm is 75%

Keywords: Diabetes Mellitus, Data Mining, Support Vector Machine, K-Nearest Neighbor

1. INTRODUCTION

Damage to the performance of human organs is very detrimental and is the source of the most problems at this time. One of the diseases that is the number one killer in the world is diabetes mellitus. Diabetes mellitus is a metabolic disease characterized by hyperglycemia caused by an obstacle in insulin secretion from insulin action or both. Chronic hyperglycemia in diabetes mellitus can cause a lot of damage to human organs such as kidneys, eyes, nerves, heart and blood vessels. Diabetes mellitus is divided into several types, type 1 diabetes mellitus generally gives rise to indications before the patient is 30 years old, although in fact the indications of the disease can arise at any time.

People with type 1 diabetes need insulin from outside the body to survive. Type 2 diabetes usually occurs when the patient is over 30 years old and does not depend on insulin from outside the body except in certain circumstances. Another type of diabetes is gestational diabetes, which is diabetes that occurs in pregnant women and is caused by impaired glucose tolerance in these patients. This disease is a hereditary disease that can be passed down from parents to children and it is very disappointing if you have diabetes from an early age. The number of people with diabetes in Indonesia continues to increase every year.

It is very difficult to determine whether a person has diabetes or not. Valid data is made through medical records and laboratory tests. This data will later indicate the presence of other diseases or just diabetes. The lack of manipulation in determining disease, especially with the application of data mining science, encourages the world of information technology and makes it easier for the medical community, especially medical personnel, to determine the identity of patients in diabetes classification.

Based on research conducted by Fida Maisa Hana (2020), Classification of Diabetes Patients Using the Decision Tree C4.5 Algorithm, grouping diabetes using the C4.5 method. In this case, a more accurate accuracy value is needed in detecting diabetes mellitus using the algorithm. different. Therefore, the author wants to make a study entitled "Application of Data Mining to Identify Diabetes Mellitus Using the Support Vector Machine (SVM) and KNN Algorithm".

2. METHOD

This research method is carried out by following the steps shown in the diagram below :

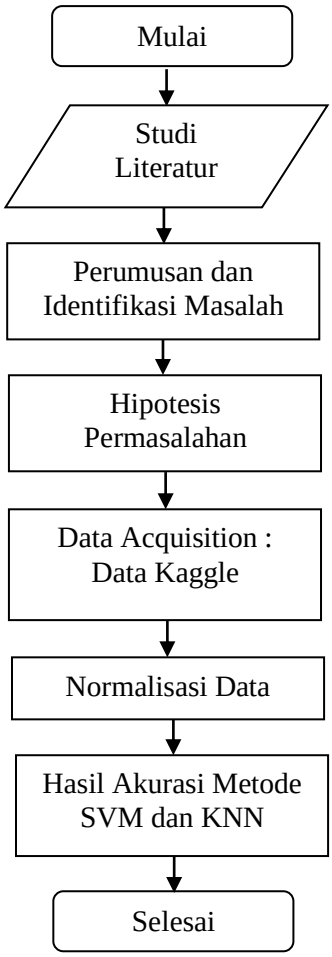

Figure 1. Research Stage Diagram

2.1 Literatur Study

Literatur study is a study conducted by researches by collecting a series of books and journals related to the problem and research objectives. This is used as a reference in the discussion of research results with the aim of clarifying various theories that are in accordance with the problem under study.

2.2 Problem Formulation and Identification

In addition to the background and problem formulation, problem identification is the most important process in conducting research. the formulation of the problem is a summary of the problems described in the study. How to optimize the search for diabetes mellitus information that is directly related to the symptoms experienced by the sufferer and find out the cause of the disease in a person's body.

2.3 Problem Hypothesis

There is some information about diabetes mellitus which can be identified using several methods which can produce a level of accuracy in identifying the disease and seeing what the factors that cause diabetes.

2.4 Data Acquisition

Dataset retrieval is obtained through kaggle data. The dataset is used to identify several factors that cause diabetes mellitus and calculate the accuracy of how much someone is affected by the disease.

2.5 Data Normalization

At this stage, the process of taking several variables which will be used as variables in the study is carried out so that comparative analysis between variables is easier to do.

2.6 Result Accuracy

At the end of the study, the results of the accuracy of data mining processing using the Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree will be a ratio or comparison of the level of accuracy (accuracy) to identify diabetes mellitus

3. RESULT AND DISCUSSION

Diabetes mellitus is a chronic disease characterized by high levels of sugar (glucose) in the blood. This condition is also often referred to as diabetes or diabetes. There are 3 types of diabetes mellitus, namely Type 1 Diabetes (patients are often detected at the age of children to adolescents), Type 2 Diabetes (patients are generally over 30 years old), Gestational Diabetes (only occurs in women who are pregnant). Based on the Global Health Data Exchange, Indonesia's health condition is currently entering the transition from the era of infectious diseases to non-communicable diseases. Indonesia is one of the 10 countries with the most cases of diabetes in the world.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33
...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63
764	2	122	70	27	0	36.8	0.340	27
765	5	121	72	23	112	26.2	0.245	30
766	1	126	60	0	0	30.1	0.349	47
767	1	93	70	31	0	30.4	0.315	23

768 rows × 8 columns

Figure 2. Diabetes Mellitus Dataset

To detect patients with diabetes or not, it can be seen through the outcome variables 0 and 1. 0 means negative for diabetes and 1 means positive for diabetes. The results of these outcomes were then concluded between positive and negative for diabetes.

3.1 Heatmap

A heatmap is a by displaying data in different color representations. Generally, the larger the number of records, the darker the color, usually represented mapping in red.

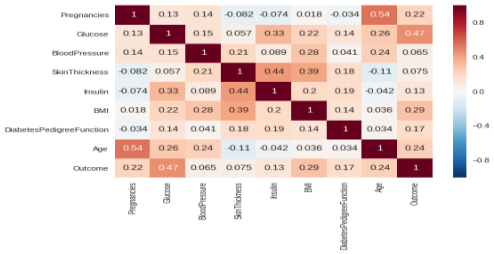

Figure 3. Variable Correlation Heatmap

There are 2 conclusions that can be drawn from the heatmap above, namely:

1. Age and pregnancy have a positive correlation indicating that adults have more children.
2. There is a positive correlation between insulin and glucose variables which can be explained by the fact that it is possible for diabetic patients who generally have high glucose levels to be given insulin injections so that they do not worsen the level of diabetes.

3.2 Positive Examples

From a total of 768 datasets, we divide into 2 examples of positive cases (500 datasets) and negative (268 datasets) according to the results of previous data processing in predictive modeling, to take positive and negative examples of suffering from diabetes, this study uses 3 variables that are used as parameters in determine the pregnancy, Blood Pressure, Age.

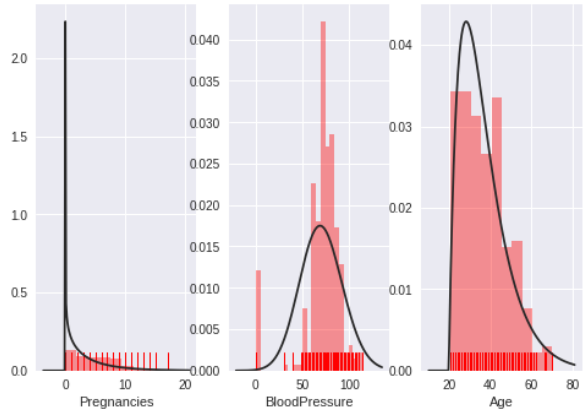

Figure 4. Graphic Examples of Positive Cause of Diabetes

There is a positive graphic example of having diabetes caused by 3 main factors, namely pregnancy, blood pressure and age. At the age of 20-50 years with blood pressure of more than 100mm/Hg are more susceptible to diabetes. This happens because high blood pressure (hypertension) is caused by consuming too many sweet, salty foods or drinks and lack of physical activity.

3.2 Negative Examples

There is an example of a negative graph of having diabetes, it can be seen that people with blood pressure values below 100mm/Hg are more difficult to get diabetes because blood pressure (blood pressure) is normal and stable. good for the body.

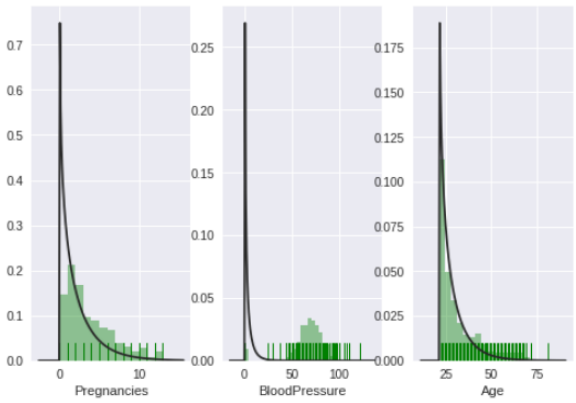

Figure 5. Graphic Examples of Negatives Cases of Diabetes

3.3 Glucose

One of the things that causes diabetes is high levels of sugar (glucose) in the blood. The reason glucose can cause diabetes is that the body is not able to regulate the amount of high sugar levels and often consume high-carbohydrate foods.

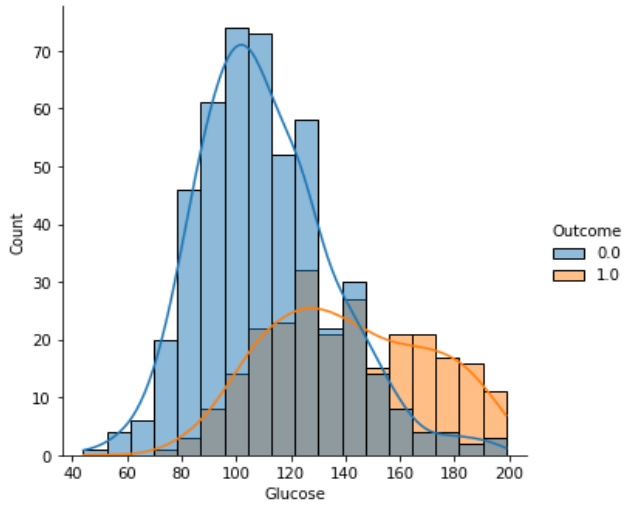

Figure 6. Diabetes Glucose Level Chart

People who do not have diabetes have a normally distributed mean glucose of about 100, whereas those with diabetes have much higher glucose levels ranging between 100 and 200.

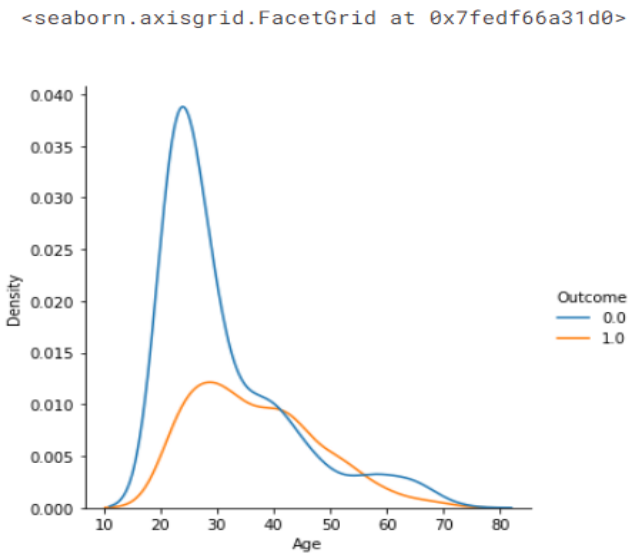

Figure 7. Age of Diabetics

It can be concluded that at the age of 10 - 80 years, the density (density) of glucose varies from 0 to 0.4%. Density (density) is a measurement of the mass of each unit volume of an object. The higher the density of an object, the greater the mass of each volume. The blue graph shows negative for diabetes and the orange graph shows positive for diabetes. At the age of 21-45 years, they are more susceptible to diabetes because of the high amount of plasma glucose levels.

3.4 Accuracy Comparison Resuts

Comparison was made on 2 variables, namely Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The following are the results of the accuracy of the two variables:

Table 1 Comparison of Accuracy Results

No	Model	Accuracy Results
1	K-Nearest Neighbor	75%
2	Linear Kernel (SVM)	76%
3	RBF Kernel (SVM)	65%

4. CONCLUSION

Diabetes can be experienced by anyone and at any time this is caused by an unhealthy lifestyle, rarely doing physical activity and not following a healthy lifestyle. This can be avoided by adopting a healthy lifestyle such as exercising, avoiding sugary foods and drinks. From the results of this study, it can be concluded that the Support Vector Machine (SVM) algorithm produces an accuracy value of 76% while the accuracy value of the K-Nearest Neighbor (KNN) algorithm is 75%.

REFERENCE

[1] Azis, W. A., Muriman, L. Y., & Burhan, S. R. (2020). Hubungan Tingkat Pengetahuan dengan Gaya Hidup Penderita Diabetes Mellitus. Jurnal Penelitian Perawat Profesional, 2(1), 105-114.
[2] Wulansari, D. D., & Wulandari, D. D. (2018). Pengembangan Model Hewan Coba Tikus Diabetes Mellitus Tipe 2 dengan Induksi Diet Tinggi Fruktosa Intragastrik. Media Pharmaceutica Indonesiana, 2(1), 41-47.
[3] Silalahi, L. (2019). Hubungan pengetahuan dan tindakan pencegahan diabetes mellitus tipe 2. Jurnal Promkes: The Indonesian Journal of Health Promotion and Health Education, 7(2), 223-232.
[4] Setiawan, H., Suhanda, S., Sopatilah, E., Rahmat, G., Wijaya, D. D., & Ariyanto, H. (2018). Hubungan tingkat pengetahuan dengan kecemasan penderita diabetes mellitus. Proceeding of The URECOL, 241-248.
[5] Aris, F., & Benyamin, B. (2019). Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi. Router Research, 1(1), 1-6.
[6] Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 2(1), 39-46.
[7] Ridwan, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan), 4(1), 15-21.
[8] Setyawan, D., & Suradi, A. (2017). Implementasi web service dan analisis kinerja algoritma klasifikasi data mining untuk memprediksi diabetes mellitus. Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer, 8(2), 701-710.
[9] Mustafa, M. S., & Simpen, I. W. (2019, August). Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. In SISITI: Seminar Ilmiah Sistem Informasi dan Teknologi Informasi (Vol. 8, No. 1).
[10] Hayuningtyas, R. Y., & Sari, R. (2022). Implementasi Data Mining Dengan Algoritma Multiple Linear Regression Untuk Memprediksi Penyakit Diabetes. Jurnal Teknik Komputer AMIK BSI, 8(1), 40-44.
[11] Afif, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus di Rumah Sakit Aisyiah. JURNAL ILMU KOMPUTER DAN MATEMATIKA, 1(1), 40-46.

[12] Siallagan, R. A. (2021). PREDIKSI PENYAKIT DIABETES MELLITUS MENGGUNAKAN ALGORITMA C4. 5. Jurnal Responsif: Riset Sains dan Informatika, 3(1), 44-52.
[13] Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4. 5. Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan), 4(1), 32-39.
[14] Novianti, N., Zarlis, M., & Sihombing, P. (2022). Penerapan Algoritma Adaboost Untuk Peningkatan Kinerja Klasifikasi Data Mining Pada Imbalance Dataset Diabetes. JURNAL MEDIA INFORMATIKA BUDIDARMA, 6(2), 1200-1206.
[15] Nurdiana, N., Rodiyansyah, S. F., & Algifari, A. (2020). Studi Komparasi Algoritma ID3 dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. INFOTECH journal, 6(2), 18-23.
[16] Pramadhana, D. (2021). Klasifikasi Penyakit Diabetes Menggunakan Metode CFS Dan ROS dengan Algoritma J48 Berbasis Adaboost. Edumatic: Jurnal Pendidikan Informatika, 5(1), 89-98.
[17] Pahlevi, R., Fredlina, K. Q., & Utami, N. W. (2021). Penerapan Algoritma Id3 Dan Svm Pada Klasifikasi Penyakit Diabetes Melitus Tipe 2. Prosiding Snast, 64-75.
[18] Septiani, W. D., & Marlina, M. (2021). Comparison of Decision Tree, Naïve Bayes, and Neural Network Algorithm for Early Detection of Diabetes. Jurnal Pilar Nusa Mandiri, 17(1), 73-78.
[19] Maulidah, N., Supriyadi, R., Utami, D. Y., Hasan, F. N., Fauzi, A., & Christian, A. (2021). Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes. Indonesian Journal on Software Engineering (IJSE), 7(1), 63-68.
[20] Elfaladonna, F., & Rahmadani, A. (2019). Analisa Metode Classification-Decission Tree dan Algoritma C. 45 untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner. SINTECH (Science And Information Technology) Journal, 2(1), 10-17.