Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Identification of Diabetes Mellitus Risk in Women using Random Forest

Eka Wijaya

Department of Informatics Engineering, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia
email: eka.wijaya@student.pelitaindonesia.ac.id

ARTICLE INFO

Article history:
Received: 14 December 2024
Revised: 2 January 2025
Accepted: 29 January 2025
Available online: 15 February 2025

Keywords:
Diabetes
Algorithm
Classification
Women
Random Forest

Please cite this articlein IEEE style as:
E. Wijaya, "Identification of Diabetes Mellitus Risk in Women Using Random Forest", Data Science Insights.

ABSTRAK

Diabetes Mellitus (DM) is one of the chronic diseases that can cause various serious complications, especially in women. Early risk identification is an important step in preventing the progression of this disease. This study aims to identify the factors influencing the risk of diabetes in women by analyzing data from several parameters, namely the number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. A quantitative approach was used in this study with descriptive and inferential statistical analysis methods. The research results show that glucose levels and BMI are the most significant factors in increasing the risk of diabetes, followed by family history of diabetes and age. In addition, the number of pregnancies also has an impact on the risk of diabetes, especially in women with a history of gestational diabetes. This research concludes that the combination of several parameters can be used to predict the risk of diabetes more accurately, especially in women. These results are expected to support early prevention efforts and better clinical decision-making in the management of diabetes.

Correspondence:
Eka Wijaya
Department of Informatics Engineering, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia

1. Introduction

Diabetes Mellitus (DM) is one of the major health issues worldwide. This chronic disease not only affects the quality of life of individuals but also imposes a significant economic burden on the global healthcare system. Among the population vulnerable to diabetes, women show a higher risk level due to certain biological factors and lifestyle choices, such as hormonal changes during pregnancy and menopause, as well as suboptimal physical activity patterns.

Early detection of diabetes risk in women is an important step to prevent serious complications and reduce the incidence of this disease. Factors such as the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), family history of diabetes (diabetes pedigree function), and age have been proven to play a significant role in determining the risk of diabetes. This research aims to identify women at high risk of diabetes based on these parameters using a machine learning-based approach. By utilizing an accurate predictive model, it is hoped that intervention measures can be taken earlier to prevent the development of diabetes.

In this study, machine learning-based classification methods are used to analyze the data. Four machine learning algorithms—Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest—will be tested to compare their accuracy and effectiveness in predicting diabetes risk. The analysis is conducted with the support of software such as Tableau and Rapid Miner to facilitate the research process. The results of this research are expected to contribute to the development of a diabetes risk prediction model that can be implemented in the healthcare system to improve the quality of life for women.

2. Metodologi Penelitian

Figure 1. Data Science Process

In the field of Data Science, there is a series of procedures that must be followed to ensure that the processed data has a high level of accuracy. Systematic and meticulous procedures are necessary to produce relevant and credible information to support research objectives. As shown in Figure 1, this process involves several important stages. Here are the steps taken in this research.

2.1 Data Collection

The first stage of this research is data collection. Data is obtained from various sources, including direct surveys and online data repositories. In the context of this research, the data was obtained from the Kaggle website (www.kaggle.com), which is a community platform for data science and machine learning. Kaggle provides a variety of high-quality datasets covering various research topics. The selection of datasets is carried out with attention to accuracy, completeness, and relevance to the research objectives..

2.2 Data cleaning

The collected data usually needs to be selected so that unimportant data can be removed from the dataset and values can be added to empty fields. At this stage, cleaning of errors, redundancy, and data inconsistencies is carried out. This process is important to ensure that the data to be researched is accurate and reliable. This data cleaning process is carried out using Google Colab. At this stage, problematic data is selected and the data format is standardized to ensure that the processed data yields accurate results.

2.3 Exploratory Data Analysis

The next stage is exploratory data analysis, which involves the application of various statistical and visualization techniques to understand patterns, trends, and relationships within the dataset. This analysis aims to uncover initial insights from the data without jumping to conclusions. The visualization process is carried out using Tableau, a software that offers high flexibility in visualizing data. With Tableau, data can be thoroughly analyzed using various graphs and diagrams, making it easier to identify important patterns.

2.4 Model Development

After initial insights are obtained from exploratory analysis, the next step is to build a predictive model. At this stage, various machine learning algorithms are applied to predict, classify, or cluster new data. This process is carried out using RapidMiner, a data analysis platform that supports the application of various machine learning algorithms. RapidMiner simplifies the model development process and helps researchers choose the algorithm that best fits the characteristics of the dataset and the research objectives.

2.5 Model Implementation

The model that has been built is then applied according to the research objectives. In this context, the model is used to identify the risk of diabetes in women based on various parameters. The implementation of this model is expected to provide real benefits, such as helping the healthcare system in early detection of diabetes risk and improving the effectiveness of medical interventions. The process of implementing the model is carried out carefully to ensure optimal results that are relevant to the research needs.

This systematic procedure allows for thorough research, so the results obtained can make a significant contribution to supporting data-driven decision-making.

3 Results and Discussion

In this section, the results will be presented according to the stages carried out:

3.1 Data Collection

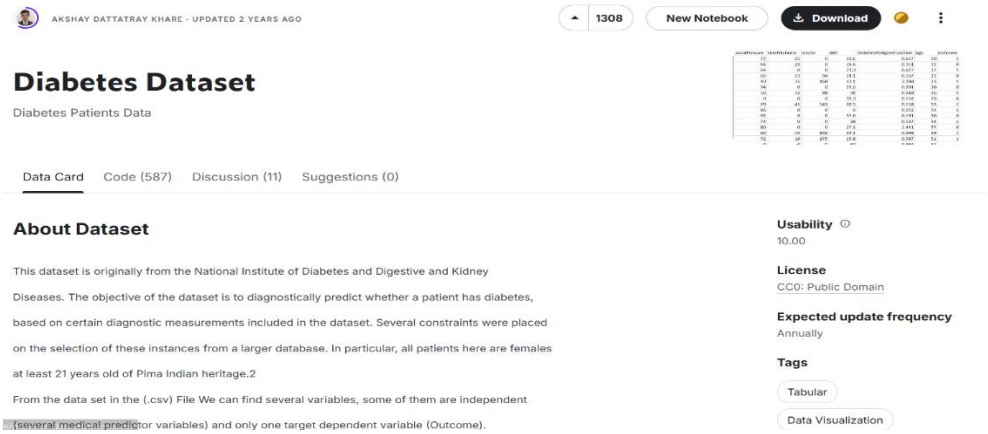

Figure 2. Diabetes Dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	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
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

Figure 3. Initial Data

This dataset (Figure 2) contains women's health data used to predict diabetes risk based on a number of medical and demographic parameters. The purpose of collecting this dataset is to analyze and identify patterns or characteristics related to the risk of diabetes. Each row in the dataset represents an individual, while each column describes medical or personal attributes, such as the number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), family history of diabetes, and age. This dataset is highly relevant for research, as it includes information that can assist in the development of diabetes risk prediction models. The results of the analysis of this dataset are expected to contribute to clinical decision-making and more effective prevention strategies.

3.2 Data Cleaning

In Figure 3, it can be seen that the data has many 0 values in certain rows and columns, possibly due to missing data. Therefore, I performed data cleaning using Python in Google Colab by replacing the 0 values with randomly distributed values for each column to make it look more natural.

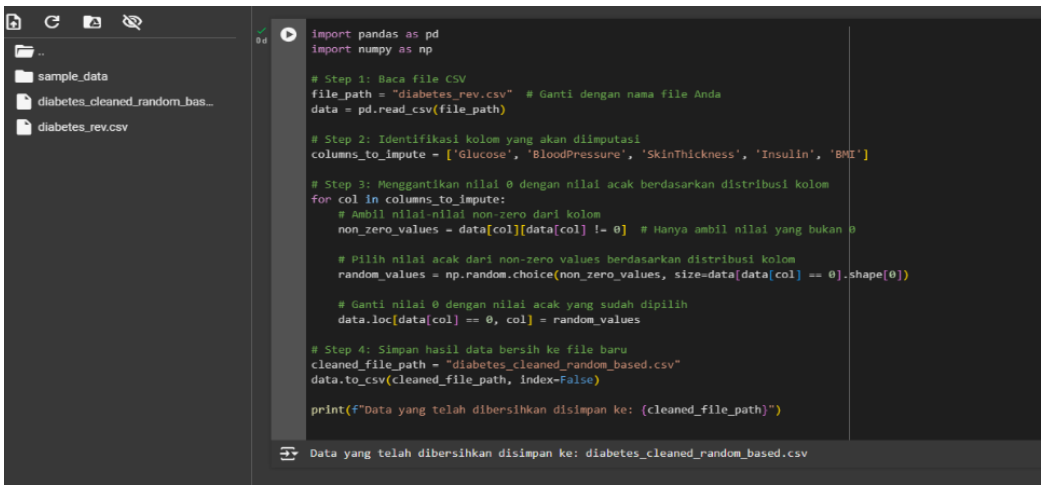

Figure 4. Cleaning Process using Google Colab

After the cleaning stage shown in Figure 4 is completed, the resulting data looks like the following image,

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	148	33.6	0.627	50	Iya
1	85	66	29	215	26.6	0.351	31	Tidak
8	183	64	35	744	23.3	0.672	32	Iya
1	89	66	23	94	28.1	0.167	21	Tidak
0	137	40	35	168	43.1	2.288	33	Iya
5	116	74	36	127	25.6	0.201	30	Tidak
3	78	50	32	88	31	0.248	26	Iya
10	115	72	33	155	35.3	0.134	29	Tidak
2	197	70	45	543	30.5	0.158	53	Iya
8	125	96	18	168	28.7	0.232	54	Iya
4	110	92	18	182	37.6	0.191	30	Tidak
10	168	74	36	115	38	0.537	34	Iya
10	139	80	19	274	27.1	1.441	57	Tidak
1	189	60	23	846	30.1	0.398	59	Iya
5	166	72	19	175	25.8	0.587	51	Iya
7	100	66	23	48	30	0.484	32	Iya
0	118	84	47	230	45.8	0.551	31	Iya
7	107	74	26	96	29.6	0.254	31	Iya
1	103	30	38	83	43.3	0.183	33	Tidak
1	115	70	30	96	34.6	0.529	32	Iya
3	126	88	41	235	39.3	0.704	27	Tidak
8	99	84	30	56	35.4	0.388	50	Tidak
7	196	90	15	126	39.8	0.451	41	Iya
9	119	80	35	144	29	0.263	29	Iya

Figure 5. Data that is cleaned using Google Colab

As can be seen from Figure 5, the value 0 has been replaced with random values for each column. Thus, there are no more missing or empty values in the dataset.

3.3 Exploratory Data Analysis

Data analysis using Tableau software for data visualization to facilitate the analysis process. The following are some of the analyses obtained:

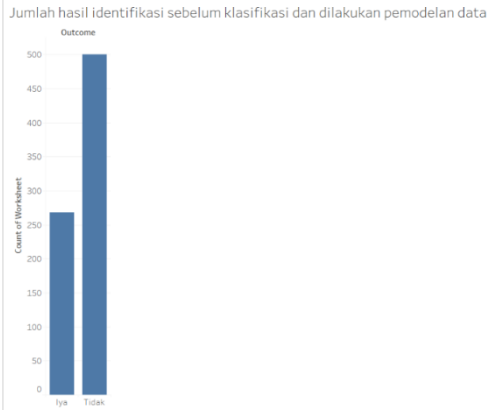

Figure 6. First Exploratory Data Analysis

Figure 6 is a tableau visualization showing how many women are affected by diabetes and how many are not, based on the existing aspects. It can be seen here that the number of women diagnosed as not having diabetes is quite significant compared to the number who have diabetes.

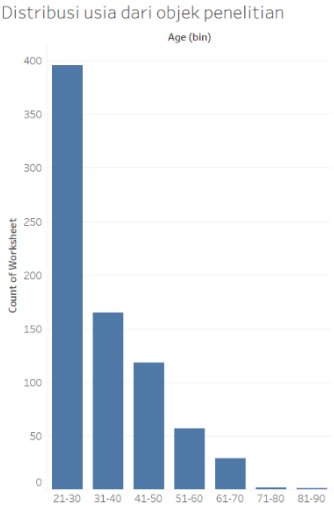

Figure 7. Second Exploratory Data Analysis

Figure 7 is a tableau visualization displaying the age range distribution of the research subjects. It can be concluded from the data visualization in Figure 7 above that the mode of the age range is in the 21-30 year class, which means the majority of the research subjects are still relatively young, specifically 21-30 years old.

3.4 Model Development

The model development was carried out by testing five different algorithms and comparing them to find the most suitable one.

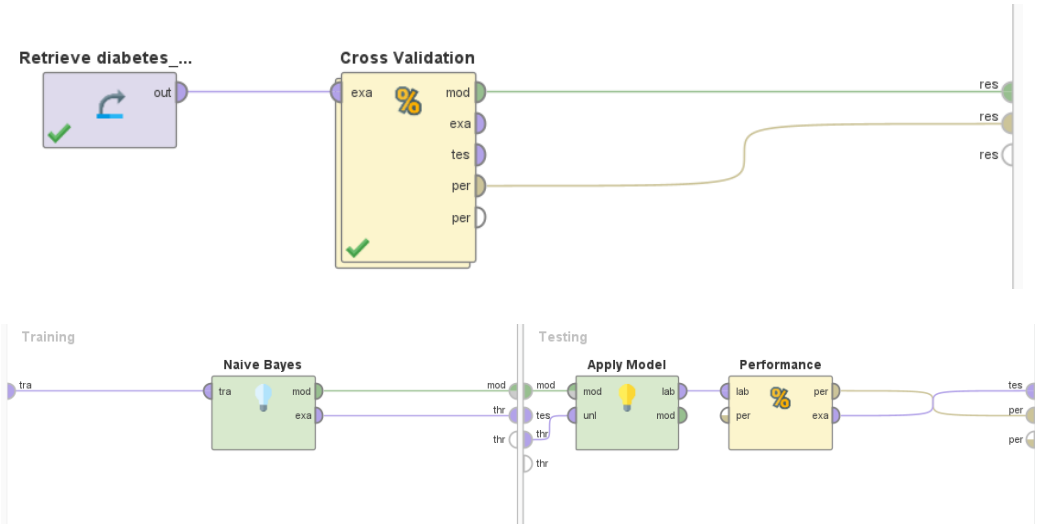

Figure 8. Naïve Bayes Algorithm using Rapidminer

The first algorithm tested using RapidMiner was the Naïve Bayes algorithm. The structure of RapidMiner, as shown in Figure 8, begins with reading the dataset, then dividing the dataset into training and testing sets 10 times with the help of cross-validation. After that, classification has been performed using the first stage of the Naive Bayes algorithm. Then, a second stage of classification was carried out until the accuracy, precision, and recall results were obtained for the data.

accuracy: 76.27% +/- 0.28% (micro average: 76.27%)

	true Iya	true Tidak	class precision
pred. Iya	1471	699	67.79%
pred. Tidak	941	3801	80.16%
class recall	60.99%	84.47%	

Figure 9. Naïve Bayes Algorithm Results

The results of the Naïve Bayes algorithm can be seen in Figure 9. The algorithm has an accuracy rate of 76.27%. This level of accuracy is already quite high, but there is still room for further improvement.

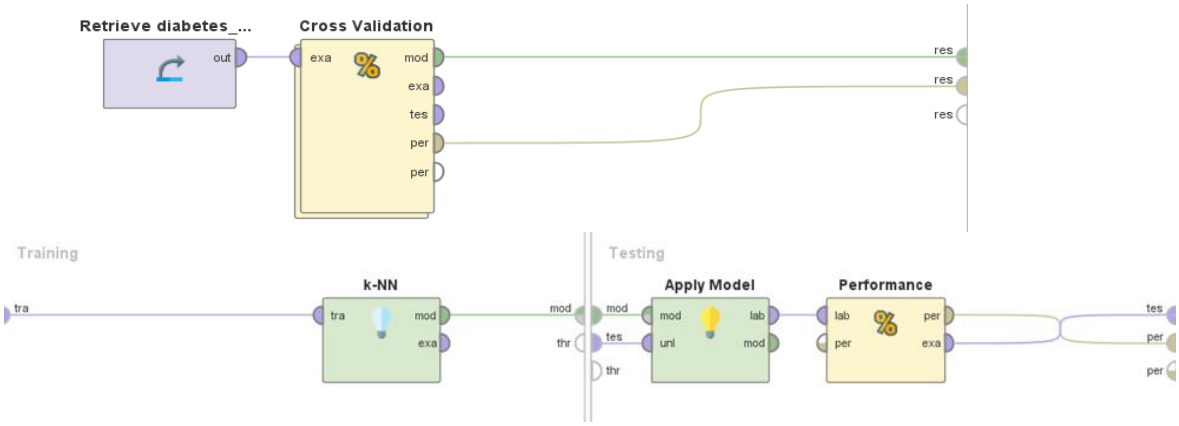

Figure 10. K-Nearest Neighbors Algorithm using Rapidminer

The second algorithm tested using RapidMiner was the K-Nearest Neighbors algorithm. The structure of RapidMiner, as shown in Figure 10, begins with reading the dataset and then splitting the dataset into training and testing sets 10 times with the help of cross-validation. After that, classification has been performed using the K-Nearest Neighbors algorithm in the first stage. Then, a second stage of classification was performed until accuracy, precision, and recall results were obtained for the data.

accuracy: 71.63% +/- 4.56% (micro average: 71.61%)

	true Iya	true Tidak	class precision
pred. Iya	139	89	60.96%
pred. Tidak	129	411	76.11%
class recall	51.87%	82.20%	

Figure 11. K-Nearest Neighbors Algorithm Results

The results of the K-Nearest Neighbors algorithm can be seen in Figure 11. The algorithm has an accuracy rate of 71.63%. This accuracy level is still quite low and falls below Naïve Bayes in terms of accuracy.

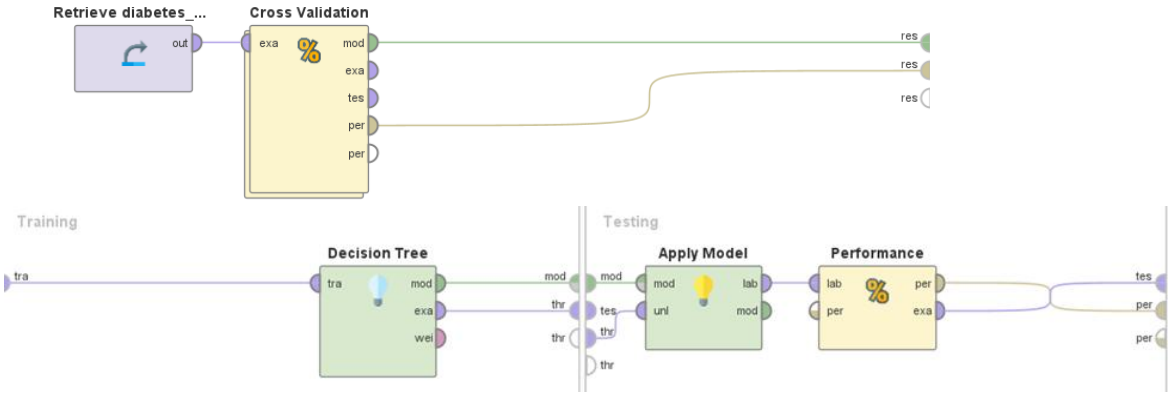

Figure 12. Decision Tree Algorithm using Rapidminer

The third algorithm tested using RapidMiner was the Decision Tree algorithm. The structure of RapidMiner, as shown in Figure 12, begins with reading the dataset and then splitting the dataset into training and testing sets 10 times with the help of cross-validation. After that, classification was performed using the Decision Tree algorithm in the first stage. Then, a second stage of classification was performed until the accuracy, precision, and recall results were obtained for the data.

accuracy: 78.18% +/- 1.56% (micro average: 78.18%)

	true Iya	true Tidak	class precision
pred. Iya	1091	187	85.37%
pred. Tidak	1321	4313	76.55%
class recall	45.23%	95.84%	

Figure 13. Decision Tree Algorithm Results

The results of the Decision Tree algorithm can be seen in Figure 13. The algorithm has an accuracy rate of 78.18%. This accuracy level is already quite high and surpasses K-Nearest Neighbors and Naïve Bayes.

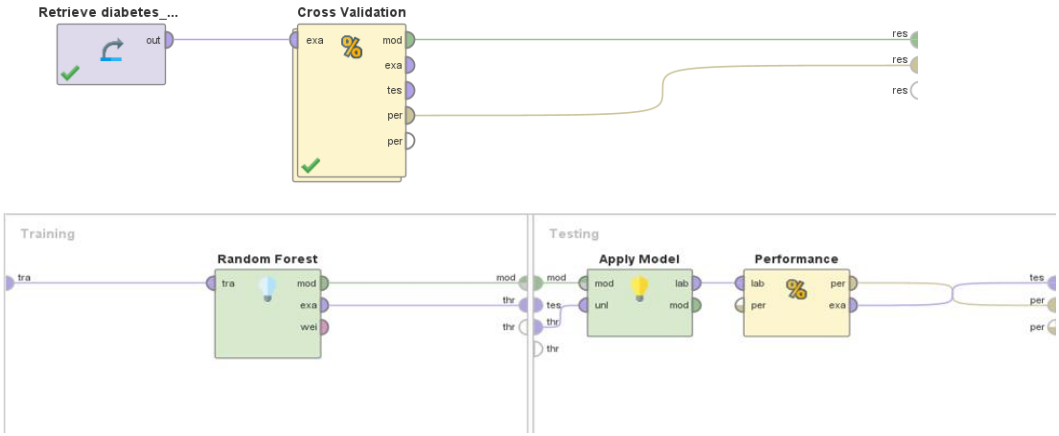

Figure 14. Random Forest Algorithm using Rapidminer

The fourth algorithm tested using RapidMiner was the Random Forest algorithm. The structure of RapidMiner, as shown in Figure 14, begins with reading the dataset, then splitting the dataset into training and testing sets 10 times with the help of cross-validation. After that, classification was performed using the Random Forest algorithm in the first stage. Then, a second stage of classification was carried out until accuracy, precision, and recall results were obtained for the data.

accuracy: 80.15% +/- 1.22% (micro average: 80.15%)

	true Iya	true Tidak	class precision
pred. Iya	1099	59	94.91%
pred. Tidak	1313	4441	77.18%
class recall	45.56%	98.69%	

Figure 15. Random Forest Algorithm Results

The results of the random forest algorithm can be seen in Figure 15. The algorithm has an accuracy rate of 80.15%. This accuracy level is currently the highest compared to Naïve Bayes, Decision Tree, and Random Forest.

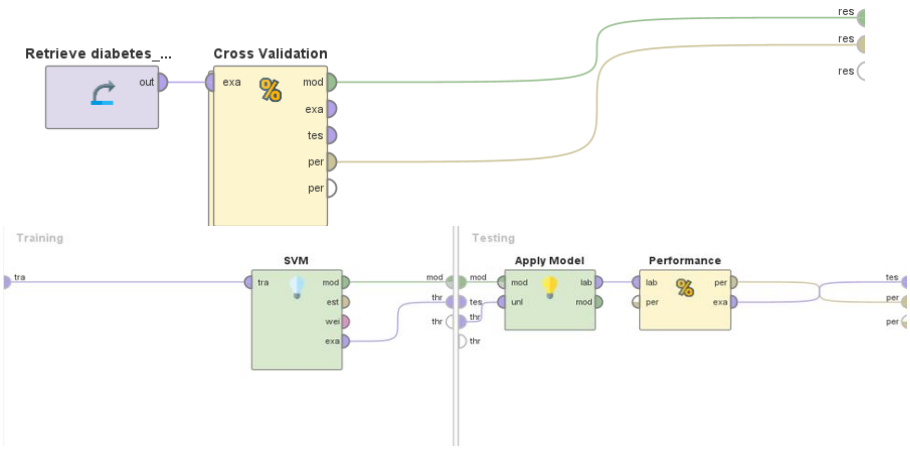

Figure 16. Support Vector Machine Algorithm using Rapidminer

The last algorithm tested using RapidMiner was the Support Vector Machine algorithm. The structure of RapidMiner, as shown in Figure 16, begins with reading the dataset and then splitting the dataset into training and testing sets 10 times with the help of cross-validation. After that, classification has been performed using the first stage of the Support Vector Machine algorithm. Then, a second stage of classification was performed until the accuracy, precision, and recall results on the data were obtained.

accuracy: 77.65% +/- 0.82% (micro average: 77.65%)

	true Iya	true Tidak	class precision
pred. Iya	1387	520	72.73%
pred. Tidak	1025	3980	79.52%
class recall	57.50%	88.44%	

Figure 17. Support Vector Machine Algorithm Results

The results of the random forest algorithm can be seen in Figure 17. The algorithm has an accuracy rate of 77.65%. This accuracy level is quite high and surpasses the accuracy of the Naïve Bayes and K-Nearest Neighbors algorithms, but is still below that of Decision Tree and Random Forest.

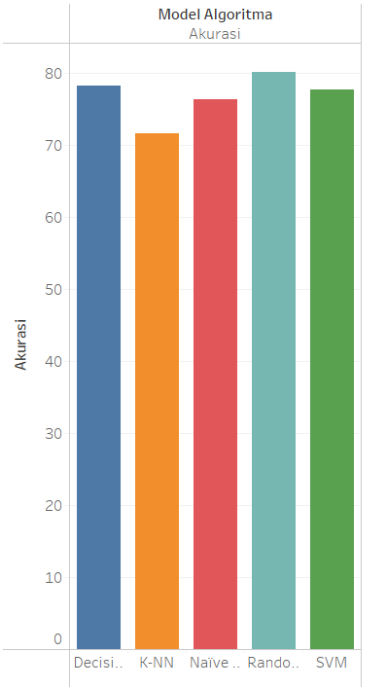

Figure 18. Accuracy based on Algorithm Comparison

After testing the five algorithms, it was found that the Random Forest algorithm is the most suitable for the purpose of this research. As seen in Figure 18, the Random Forest algorithm has the highest accuracy compared to the other four algorithms tested in this study, with an accuracy level of 80.15%.

4 Conclusion

This study shows that the risk of Diabetes Mellitus (DM) in women can be effectively identified through the analysis of several health parameters, such as blood glucose levels, BMI, family history, age, and number of pregnancies. Based on the testing of five machine learning algorithms (Naïve Bayes, KNN, Decision Tree, Random Forest, and SVM), the Random Forest algorithm produced the highest accuracy of 80.15%, making it the best choice for predicting diabetes risk. This research emphasizes the importance of early detection through predictive models to support more effective medical interventions and disease prevention. For the performance of the algorithm, additional evaluation with a larger and more diverse dataset will provide stronger validation of its performance, and collecting real-time data that includes additional parameters, such as diet and physical activity, can be conducted for a more comprehensive analysis.

References

[1] E. Dritsas and M. Trigka, “Data-Driven Machine-Learning Methods for Diabetes Risk Prediction,” *Sensors*, vol. 22, no. 14, 2022, doi: 10.3390/s22145304.

[2] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” *Front. Genet.*, vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.

[3] A. Tuppad and S. D. Patil, “Machine learning for diabetes clinical decision support: a review,” *Adv. Comput. Intell.*, vol. 2, no. 2, pp. 1–24, 2022, doi: 10.1007/s43674-022-00034-y.
[4] S. Yadu, R. Chandra, and V. K. Sinha, “Comparing Different Machine Learning Techniques in Predicting Diabetes on Early Stage,” p. 20, 2024, doi: 10.3390/engproc2024062020.
[5] Y. Zhao *et al.*, “Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy Among Patients With Type 2 Diabetes Mellitus: A Cohort Study,” *Front. Endocrinol. (Lausanne).*, vol. 13, no. May, pp. 1–8, 2022, doi: 10.3389/fendo.2022.876559.
[6] P. D. Petridis, A. S. Kristo, and A. K. Sikalidis, “A Review on Trending Machine Learning Techniques for Type 2 Diabetes Mellitus Management,” pp. 1–24, 2024.
[7] N. Ghaffar Nia, E. Kaplanoglu, and A. Nasab, “Evaluation of artificial intelligence techniques in disease diagnosis and prediction,” *Discov. Artif. Intell.*, vol. 3, no. 1, 2023, doi: 10.1007/s44163-023-00049-5.
[8] C. N. Noviyanti and A. Alamsyah, “Early Detection of Diabetes Using Random Forest Algorithm,” *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, 2024, doi: 10.52465/joiser.v2i1.245.
[9] R. Qureshi *et al.*, “Artificial Intelligence and Biosensors in Healthcare and Its Clinical Relevance: A Review,” *IEEE Access*, vol. 11, pp. 61600–61620, 2023, doi: 10.1109/ACCESS.2023.3285596.
[10] K. Wu, “Optimizing Diabetes Prediction with Machine Learning: Model Comparisons and Insights,” *J. Sci. Technol.*, vol. 5, no. 4, pp. 41–51, 2024, doi: 10.55662/jst.2024.5403.
[11] D. Lamba, W. H. Hsu, and M. Alsadhan, *Predictive analytics and machine learning for medical informatics: A survey of tasks and techniques*, no. February. 2021. doi: 10.1016/B978-0-12-821777-1.00023-9.
[12] M. S. Reza, A. R. S. Fakir, M. R. Islam, A. M. T. Alom, T. Sen, and M. S. Islam, “Early Stage Diabetes Prediction Using Machine Learning Techniques,” *2023 26th Int. Conf. Comput. Inf. Technol. ICCIT 2023*, pp. 1–21, 2023, doi: 10.1109/ICCIT60459.2023.10441427.
[13] S. Ucha Putri, E. Irawan, F. Rizky, S. Tunas Bangsa, P. A. -Indonesia Jln Sudirman Blok No, and S. Utara, “Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5,” *Januari*, vol. 2, no. 1, pp. 39–46, 2021.
[14] A. Veronica Agustin and A. Voutama, “Implementasi Data Mining Klasifikasi Penyakit Diabetes Pada Perempuan Menggunakan Naïve Bayes,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1002–1007, 2023, doi: 10.36040/jati.v7i2.6808.
[15] C. Y. Chou, D. Y. Hsu, and C. H. Chou, “Predicting the Onset of Diabetes with Machine Learning Methods,” *J. Pers. Med.*, vol. 13, no. 3, 2023, doi: 10.3390/jpm13030406.