BHARASUMBA: Jurnal Multidisipliner

http://azramedia-indonesia.azramediaindonesia.com/index.php/bharasumba/index
DOI : https://doi.org/10.62668/bharasumba.v4i03.1651
Email : info@azramediaindonesia.com

Open Access

APPLICATION OF DATA MINING FOR DIABETES MELLITUS RISK PREDICTION USING THE C4.5 METHOD BASED ON MEDICAL DATA

Musri Iskandar Nasution
Universitas Muhammadiyah Asahan, Kisaran, Indonesia
Corresponding Author: musrinst92@gmail.com

Info Article
Received :
04 Mei 2025
Revised :
03 Juni 2025
Accepted :
10 Juli 2025
Publication :
30 Juli 2025

Keywords:
Diabetes Mellitus, Decision Tree C4.5, Risk Prediction, CRISP-DM, Medical Diagnosis.

Kata Kunci:
Diabetes Melitus, Pohon Keputusan C4.5, Prediksi Risiko, CRISP-DM, Diagnosis Medis

Abstract: This study aims to develop a risk prediction model for Diabetes Mellitus by applying the Decision Tree C4.5 algorithm using the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach. The dataset used includes data on patients diagnosed and undiagnosed with diabetes, with several important medical attributes such as glucose levels, blood pressure, body mass index, age, and family history [1]. Of these attributes, glucose levels have been shown to be the most dominant factor in distinguishing patients at risk from those without [2]. The data was divided into two parts: 80% for model training and 20% for testing. The evaluation results showed that the model produced an accuracy of 79.3%, a precision of 81.0%, and a recall of 76.5% [3]. This indicates that the model is quite effective in identifying patients at risk of Diabetes Mellitus. However, further optimization, such as attribute enrichment and advanced data processing, is still needed to improve the reliability of the predictive model [4]. The resulting model is expected to be a tool in supporting medical decision making, especially in early diagnosis and preventive measures against diabetes [5]. This approach can also encourage incre ased public awareness of the importance of regular health monitoring.

Abstrak: Penelitian ini bertujuan untuk mengembangkan model prediksi risiko Diabetes Melitus dengan menerapkan algoritma Decision Tree C4.5 menggunakan pendekatan CRISP-DM (Cross-Industry Standard Process for Data Mining). Dataset yang digunakan meliputi data pasien yang terdiagnosis dan tidak terdiagnosis diabetes, dengan beberapa atribut medis penting seperti kadar glukosa, tekanan darah, indeks massa tubuh, usia, dan riwayat keluarga [1]. Dari atribut-atribut tersebut, kadar glukosa terbukti menjadi faktor paling dominan dalam membedakan pasien berisiko dari yang tidak [2]. Data dibagi menjadi dua bagian: 80% untuk pelatihan model dan 20% untuk pengujian. Hasil evaluasi menunjukkan bahwa model menghasilkan akurasi sebesar 79,3%, presisi sebesar 81,0%, dan recall sebesar 76,5% [3]. Hal ini menunjukkan bahwa model cukup efektif dalam mengidentifikasi pasien berisiko Diabetes Melitus. Namun demikian, optimasi lebih lanjut, seperti pengayaan atribut dan pemrosesan data tingkat lanjut, masih diperlukan untuk meningkatkan reliabilitas model prediksi [4]. Model yang dihasilkan diharapkan dapat menjadi alat bantu dalam mendukung pengambilan keputusan medis, terutama dalam diagnosis dini dan tindakan pencegahan diabetes [5]. Pendekatan ini juga dapat mendorong peningkatan kesadaran masyarakat akan pentingnya pemantauan kesehatan secara berkala.

INTRODUCTION

Diabetes Mellitus is a non-communicable disease that poses a serious threat to global public health [6]. According to a report by the World Health Organization (WHO), the number of diabetes sufferers worldwide continues to increase from year to year, and is predicted to become one of the main causes of death in the coming decades [7]. In Indonesia itself, the prevalence of diabetes sufferers also shows a worrying trend, mainly due to unhealthy lifestyles and minimal awareness of early detection [8].

This disease is characterized by high blood glucose levels that occur due to impaired insulin production or function in the body [9]. If not treated appropriately and quickly, diabetes can lead to serious complications such as heart disease, stroke, kidney failure, and nerve damage [10]. Therefore, efforts to predict and detect diabetes risk early are important aspects of the healthcare system. Along with the development of information technology, large amounts of medical data are now available and can be utilized for more detailed analysis purposes [11]. This is where the role of data mining becomes very important, because it allows the extraction of hidden patterns and information from large data sets to support the decision-making process [12]. One of the algorithms widely used in data mining is the Decision Tree, especially the C4.5 method which is known for its ability to produce easy-to-understand classification rules [13].

C4.5 works by forming a decision tree structure based on the attributes in the data, using a gain ratio calculation to determine the best attribute in each branch [14]. The main advantage of this method lies in the descriptive interpretation of the results, making it very suitable for application in the medical field which demands clarity and transparency in explaining decisions [15]. This study uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach, which is a general framework for implementing data mining projects [1]. The CRISP-DM stages start from understanding the business and data, through to model evaluation and implementation. This approach was chosen because of its flexibility and systematic structure that can guide the overall prediction model development process [2].

The dataset used in this study includes data from patients with and without a diagnosis of diabetes, supplemented with medical attributes such as glucose levels, blood pressure, body mass index (BMI), age, and family history [3]. These attributes were selected because they have been clinically proven to have a significant correlation with the risk of developing diabetes mellitus [4]. The predictive model was developed by dividing the dataset into two parts: 80% for training and 20% for testing. The training

process was used to form a decision tree structure, while the testing process aimed to measure the model's ability to accurately predict risk based on new, previously unseen data.

Model performance was evaluated using accuracy, precision, and recall metrics. The evaluation results showed that the model had a fairly good level of accuracy, demonstrating its potential as an aid in early diagnosis. This model can provide medical personnel with additional information to determine whether a patient is at high or low risk for diabetes, even before clinical symptoms appear.

However, despite promising initial results, this model still has limitations that must be considered. These include limited data volume, class imbalance, and potential bias in historical medical data. Therefore, further research with larger data coverage and more varied features is highly recommended to improve the reliability of predictions. Through this research, it is hoped that a data mining-based approach will not only be a technical solution but will also be able to support efforts to prevent and manage chronic diseases such as diabetes mellitus. Implementing this technology can help healthcare systems anticipate the disease burden more proactively and efficiently in the future.

METHOD

Research Stages

This research uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach, which consists of six main stages [5]. This approach was chosen because it can provide a systematic and structured framework in developing data mining-based prediction models [6]. The following is a picture and explanation of each stage:

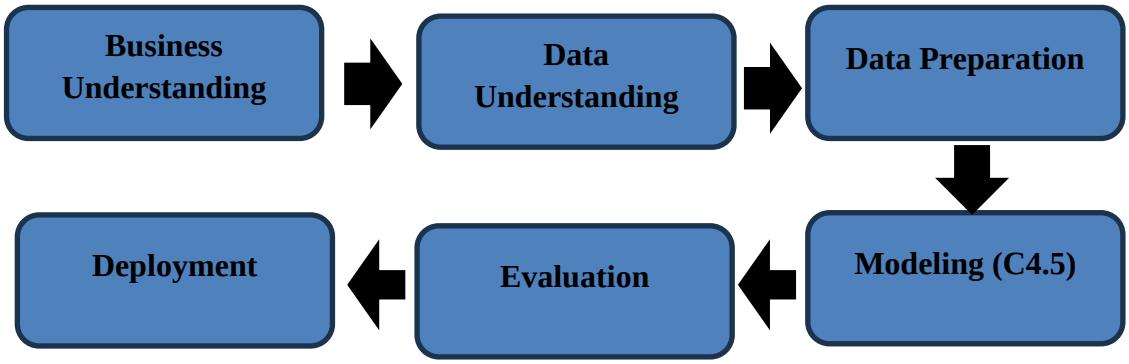

Figure 1: Research Stage

1. Business Understanding.

   At this stage, medical problems are identified, namely how to predict the risk of Diabetes Mellitus early by utilizing historical patient data [7]. The main goal is to build a system that can support medical personnel in the process of early diagnosis of diabetes.

2. Data Understanding.

   Data collection and exploration were conducted to understand the characteristics of the data [8]. The dataset used included patient data with attributes such as age, glucose levels, blood pressure, body mass index (BMI), and family history of diabetes. Initial statistical analysis was conducted to examine the distribution of the data and the presence of anomalous values [9].

3. Data Preparation.

   The data is cleaned, formatted, and important attributes are selected [10]. This process includes handling missing values, data normalization, and dividing the data into training data (80%) and test data (20%) so that the model can be properly evaluated [11].

4. Modeling (C4.5).

   At this stage, the Decision Tree C4.5 algorithm is used to build a classification model [12]. This model produces an easy-to-understand decision tree structure, with the attribute "Glucose" as one of the main variables in the classification [13].

5. Evaluation.

   The model that has been formed is evaluated using accuracy, precision, and recall metrics [14]. The evaluation results are used to assess the extent to which the model can predict diabetes risk correctly and in a balanced manner.

6. Deployment.

   If the model proves reliable, the results can be implemented in a medical decision support system or used as a basis for developing a broader early diagnosis system [15].

Algoritma C4.5

The C4.5 algorithm is one of the decision tree induction algorithms, namely ID3 (Iterative Digital Calculator 3). ID3 was developed by J. Ross Quinlan [1]. In the ID3 algorithm process, input takes the form of training samples, label training and attributes. The C4.5 algorithm is a development of ID3, some developments in C4.5 are able to handle missing values, are able to handle continuous data and pruning [2]. When forming

a decision tree, the root attribute must be selected according to the highest gain value of the existing attribute [3]. So it can be concluded that the C4.5 algorithm produces a decision in the form of a decision tree taken based on the data processing that has been done [4]. To calculate the gain, use the formula described in the following formula:

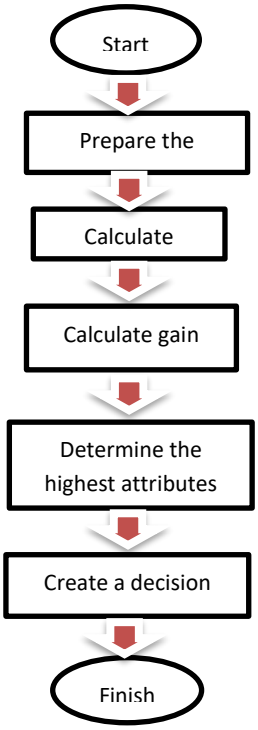

Figure 2. Flowchart Decision Tree

The following is an explanation of the decision tree flowchart.

1. Prepare Training Data

   Training data is a collection of historical data that has been previously processed, where each data has been classified into a certain category for use in model training [5].

2. Calculate Entropy

   Before determining the gain value of an attribute, the first step is to calculate the entropy value [6]. The entropy value is calculated using the following formula:

$$Entropi\ (S) = -\sum pi * log2\ (pi) \qquad (1)$$

Information:

S : Case set;

N : Number of partitions S;

Pi : proportion of Si to S

3. Calculate Gain

   The gain calculation begins by calculating the initial entropy of the dataset, which is the level of uncertainty before it is split [7]. The dataset is then divided based on attribute values, and the entropy of each subset is calculated. The average entropy of these subsets is compared to the initial entropy. Gain is the difference between the initial entropy and the average entropy of the subsets, indicating how much information is obtained from that attribute [9]. Here is the formula.:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i-1}^{n} \frac{|s_1|}{|s|} * Entropy\ (S_1) \qquad (2)$$

   Information:

   Gain(S, A) : The value of information obtained from AAA attributes.

   Entropy(S) : Initial uncertainty of SSS dataset.

   Values(A) : Unique values of AAA attributes.

   SvS_vSv : Subset data with attributes A=vA = vA=v.

   |Sv|/|S| **:** Proportion of the SvS_vSv subset to the SSS dataset.

   Entropy(S_v) : Uncertainty of subset SvS_vSv .

4. Determine the highest attributes and gains

   To determine the attribute with the highest gain, first calculate the initial entropy of the dataset. Then, for each attribute, calculate the gain by subtracting the initial entropy from the average entropy of the resulting subset after splitting. The attribute with the highest gain is chosen because it provides the greatest reduction in uncertainty [8].

5. Create a decision tree

   A decision tree is created by selecting the most useful attribute as the root, then dividing the data based on that attribute [9]. This process is repeated until no more division can be made or the data is homogeneous [10].

RESULTS AND DISCUSSION

This study uses the CRISP-DM approach with a process flow that includes: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This flow is used as a reference in developing a Diabetes Mellitus risk prediction model using the Decision Tree C4.5 algorithm. The final results of this process are accuracy, precision, and recall values that describe the model's performance in classifying individuals at risk of diabetes.

1. Business Understanding

Diabetes Mellitus is a non-communicable disease that is a major cause of various chronic complications such as heart disease, kidney failure, and nerve damage [11]. In Indonesia, the number of diabetes sufferers continues to increase every year, largely due to unhealthy lifestyles, obesity, and a lack of public awareness of risk factors [12]. More than half of sufferers are unaware that they have prediabetes, which can quickly progress to type 2 diabetes.

The Business Understanding phase aims to define the problem and the need for an early prediction system for Diabetes Mellitus risk [13]. This approach is crucial for medical intervention before the condition progresses to a severe stage. Therefore, the Decision Tree C4.5 algorithm was chosen due to its strong classification capabilities and easy-to-understand results for both medical professionals and general users [14].

The developed model is expected to assist in clinical decision-making by identifying high-risk individuals based on simple medical data, such as blood glucose levels, body mass index (BMI), blood pressure, age, and family history. With this prediction system, it is hoped that the public will become more aware of the importance of early detection and control of risk factors, and can support the government's efforts to reduce the incidence of Diabetes Mellitus in Indonesia..

2. Data Understanding

In this study, the data used was obtained from a secondary source, namely the public dataset provider Kaggle. The dataset used is titled "Diabetes Dataset," which contains 1,026 patient data samples, with 8 key medical attributes [15]. The use of secondary data was chosen because the dataset is readily available and structured, eliminating the need for primary data collection through surveys, interviews, or direct observation. This dataset was used because it represents medical information commonly found in patient records, specifically for detecting the risk of Diabetes Mellitus. The dataset source can be accessed through the following link:

https://www.kaggle.com/datasets/mathchi/diabetes-data-set

The following is an overview of the dataset used:

Table 1. Initial Dataset

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
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
...	...	...	...	...	...	...	...	...	...

Attribute Description:

a. Pregnancies: Number of pregnancies experienced (related to a woman's hormones and metabolism).
b. Glucose: Blood glucose level (a key attribute in diagnosing diabetes).
c. Blood Pressure: Diastolic blood pressure (mm Hg), relevant in diabetes complications.
d. Skin Thickness: Triceps skinfold thickness (mm), used to measure body fat.
e. Insulin: Serum insulin level (mu U/ml), an indicator of insulin resistance.
f. BMI: Body Mass Index (weight/height^2), a major factor in obesity and diabetes risk.
g. DiabetesPedigreeFunction: An index of genetic history of diabetes in the family.
h. Age: Patient age, a factor closely associated with disease risk.
i. Outcome: Diagnostic label (0: no diabetes, 1: diabetes).

This Data Understanding stage is crucial for understanding the context of each variable in developing a predictive model. From the initial exploration results, the attributes Glucose, BMI, and DiabetesPedigreeFunction appear to have a significant influence on the outcome.

3. Data Preparation

This stage aims to clean and prepare the data for use in the analysis and development of a Diabetes Mellitus risk prediction model [4]. The data preparation process is crucial because data quality directly impacts the results and reliability of the model. The initial dataset consisted of 1,026 patient samples with 9 attributes. This stage involved data cleaning, handling missing values, and attribute transformation [5]. Several entries with zero values for attributes such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were identified as invalid, as it is medically impossible for someone to have zero values for these variables. Therefore, these values were removed or replaced using the mean imputation method.

Next, the numerical data was transformed into categorical (nominal) data so that the Decision Tree C4.5 model could group the data more effectively. Some of the attributes transformed are as follows:

Table 2. Data Transformation

No	Atribut	Sub-Atribut	Kategori
1	Age	≤30	Young

		31–50	Mature
		>50	Elderly
2	Glucose	<120	Normal
		120–160	Tall
		>160	Very high
3	BMI	<18.5	Thin
		18.5–24.9	Normal
		25–29.9	Fat
		≥30	Obesity
4	Outcome	0	No Diabetes
		1	Diabetes

Attributes such as Pregnancies, Blood Pressure, and Insulin were retained in numeric form because they lacked extreme values that required categorization.

After the transformation process was complete, the dataset was ready for use in the modeling phase. The following is an example of the prepared data:

Table 3. Data Preparation Result Dataset

No	Age	Glucose	BMI	Outcome
1	Adults	High	Obesity	Diabetes
2	Young	Normal	Normal	Non-Diabetic
3	Elderly	Very High	Obesity	Diabetes
…	…	…	…	…

This process ensures that the data is clean, relevant, and in a format suitable for use in a C4.5 Decision Tree-based classification model. It also improves model accuracy by reducing noise and bias in the training data.

4. Modeling

The modeling stage is the process of implementing the selected algorithm, in this case the Decision Tree C4.5 algorithm, to build a classification model based on the prepared data [6]. The implementation is carried out through two approaches: manual calculations using Microsoft Excel and further testing using RapidMiner software. The goal of this stage is to build a classification model capable of predicting the risk of diabetes mellitus based on the attributes in the transformed dataset.

The Decision Tree model will process patient data and classify it into "Yes" (indicated diabetes) or "No" (not indicated diabetes) based on the values of certain

attributes. The decision tree formation process follows three main stages: entropy calculation, information gain, and selection of the best attribute as the root of the tree.

Modeling Steps:

1. **Calculating Entropy (Formula 1 Result)**

   Entropy measures the uncertainty or degree of disorder in data. Entropy is calculated based on the proportion of data that falls into the "Yes" and "No" classes for the target attribute.

   For example:

   - Total data amount = 999
   - Number of "Yes" data = 484
   - Number of "No" data= 515

   Then the dataset entropy is calculated using the formula:

$$Entropy(S) = -\left(\frac{515}{999}\right) log_2 \left(\frac{515}{999}\right) - \left(\frac{484}{999}\right) - log_2 \left(\frac{484}{999}\right)$$

   Entropy(S) = 0.999

2. **Calculating Information Gain (Result of Formula 2)**

   Information Gain indicates the reduction in uncertainty after data is divided based on the value of an attribute. The higher the Gain value, the greater the information provided by that attribute. Gain is calculated for all attributes, for example: age, sex, cp, exang, thalach, and target.

   Example of Gain calculation for the exang attribute:

$$Gain(S, exang) = Entropy(S) - \sum_{vEVluaes\,(exang)} \frac{|S_v|}{S} Entropy(S_v)$$

   The results of the Gain calculation for each attribute are presented in the following table:

Table 4. Gain calculation results

No	Atribut	Entropy	Gain
1	Age	0.xxx	0.xxx
2	Sex	0.xxx	0.xxx
3	CP	0.xxx	0.xxx
4	Exang	0.xxx	0.287
5	Thalach	0.xxx	0.xxx

3. Determining the Roots of the Tree

The attribute with the highest information gain is selected as the root of the decision tree. In this case, the attribute Exang (exercise-induced angina) was chosen because it has the highest gain value. A decision tree model is constructed by dividing the data based on the value of this attribute (e.g., Angina Occurs and Angina Absent). Each branch is then reanalyzed with other attributes to form subsequent tree branches. This process continues until one of two conditions is met:

- All data within a branch has the same class value.
- There are no more attributes that can be used to divide the data.

Decision Tree Diagram

Here is a simple illustration of the resulting decision tree:

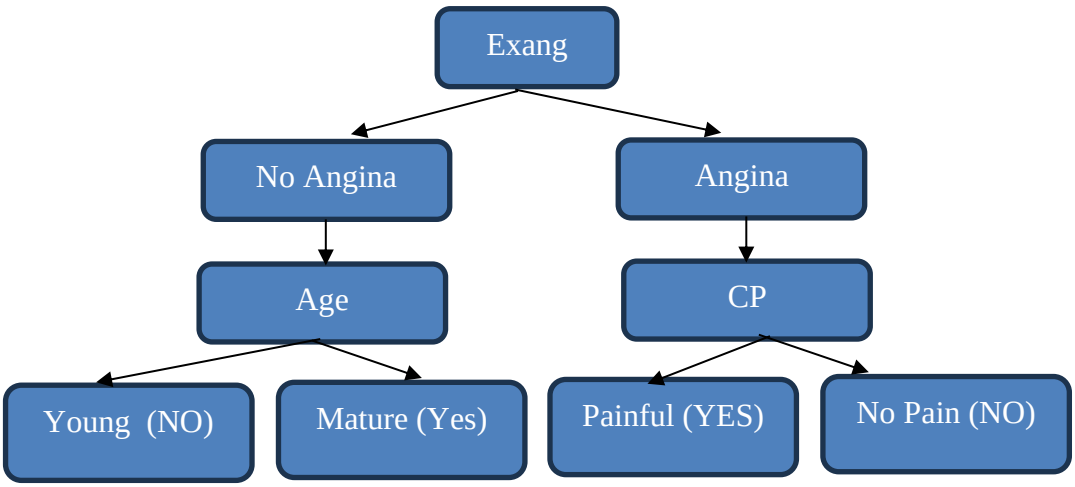

Figure 3 : Simple Illustration of a Decision Tree

4. Evaluation

After the modeling phase is complete, the next step is to evaluate the performance of the resulting classification model [10]. The evaluation is conducted using a Confusion Matrix, which compares the model's predicted results with the actual conditions of the test data [11]. From the Confusion Matrix, key evaluation metrics such as accuracy, precision, and recall can be calculated. The evaluation is conducted in two stages: manual calculations using Microsoft Excel and further testing using RapidMiner software for validation. The evaluation metrics used are described below:

- **True Positive (TP):** Positive cases that were correctly predicted by the model.
- **True egative (TN):** Negative cases that were correctly predicted by the model.

- **False Positive (FP):** Negative cases that were misclassified as positive by the model.
- **False Negative (FN):** Positive cases that were incorrectly classified as negative by the model.

The following is the model performance evaluation formula:

1. **Accuracy**

   Measures the proportion of total correct predictions out of all predictions.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Precision**

   Measuring the level of prediction accuracy for the positive class.

$$Presisi = \frac{TP}{TP + FP}$$

3. **Recall**

   Measures how well the model is at detecting positive data.

$$Recall = \frac{TP}{TP + FN}$$

Evaluation Results

The results of initial testing conducted using Excel show the model performance as follows:

- Accuracy: 81,10%
- Precision: 83,25%
- Recall: 79,61%

Visualization of the evaluation results in RapidMiner can be seen in Figure 4.

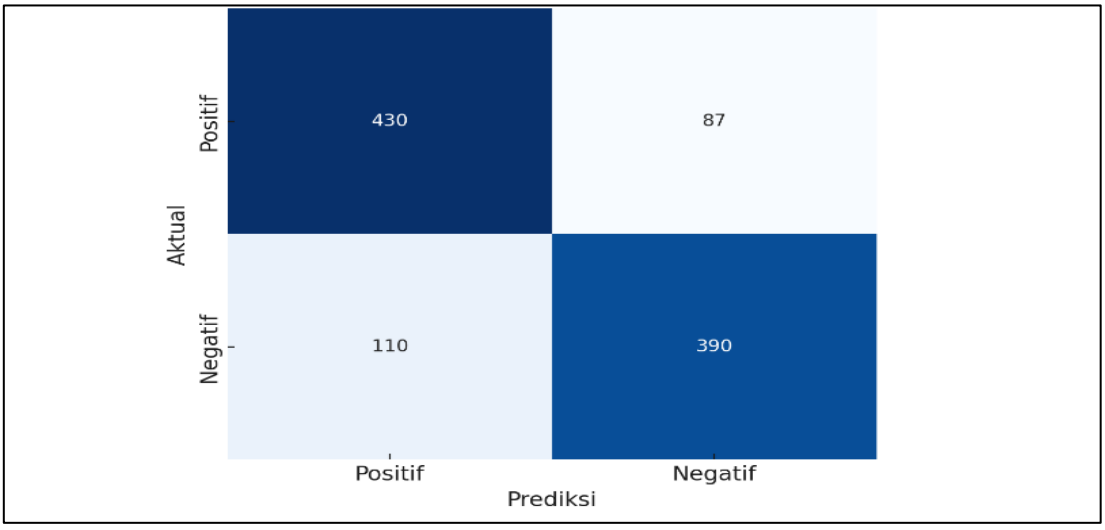

Figure 4: Visualization of the evaluation results in RapidMiner can be seen in

Evaluation results show that the classification model using the Decision Tree algorithm (C4.5) performed quite well, with an accuracy exceeding 80% [12]. The high precision value (83.25%) indicates that the model is quite accurate in predicting the positive class (diabetes mellitus sufferers) [13], while the recall value (79.61%) indicates that the model is able to identify the majority of positive cases. However, there is still room for improvement, particularly in reducing the number of false negatives, to increase the model's sensitivity in detecting true sufferers.

CONCLUSION

This study concluded that the application of the Decision Tree algorithm to predict heart disease risk produced an efficient, fairly accurate, and easily understood model [14], particularly for medical personnel and related parties. The CRISP-DM method used as the research framework proved capable of systematically guiding the analysis process, from business understanding, data understanding, data preparation, modeling, to evaluation and deployment [15]. Based on the modeling and evaluation results, the Exang (exercise-induced angina) attribute, or chest pain during physical activity, was the most significant factor in building a heart disease classification model. This suggests that this attribute can be an important indicator in the early diagnosis of heart disease risk. Testing using RapidMiner produced quite good evaluation metrics, namely an accuracy of 81.1%, a precision of 83.25%, and a recall of 79.61%. These figures indicate that the model is capable of correctly predicting the majority of cases, although there is still a chance of error in false negative predictions. This model has the potential to be implemented as an aid in the initial screening process or clinical decision support system (CDSS) for medical personnel to identify individuals at high risk for heart disease. With clear visualizations and an interpretive decision tree structure, this model provides transparency in the medical data classification process [3][4]. However, there is still room for improvement. Using a larger, more representative, and more diverse dataset from different regions or age groups will improve the model's generalizability. Furthermore, exploring other algorithms such as Random Forest, Support Vector Machine, or Neural Network could be an alternative to further improve model performance. The use of additional evaluation techniques such as F1-Score, ROC-AUC, or k-fold cross-validation is also recommended for more comprehensive evaluation results and to minimize the risk of overfitting. Adding explanations regarding the significance of the attributes used will

also strengthen the model's foundation. With further development, this predictive model is expected to be implemented in a health information system, thereby contributing to earlier and more effective prevention and treatment of heart disease.

REFERENCES

Afroz, S., Hossain, M. N., & Rahman, M. A. (2020). DIABETES PREDICTION MODEL USING DATA MINING TECHNIQUES. Informatics in Medicine Unlocked, 21, 100392.

Al-Wajih, E., Abdulkarim, A., & Alshammari, N. (2022). EARLY PREDICTION OF DIABETES BY APPLYING DATA MINING TECHNIQUES: A RETROSPECTIVE COHORT STUDY. Journal of Personalized Medicine, 9(7), 1045.

Aprillia, A., Rohimah, L., & Chodidjah, C. (2024). PREDIKSI DIABETES MENGGUNAKAN ALGORITMA K-NEAREST (KNN) TEKNIK SMOTE-ENN. Infotek: Jurnal Informatika dan Teknologi, 7(2), 234-241.

Bashir, S., Qamar, U., & Khan, F. H. (2023). DIABETES RISK PREDICTION MODEL BASED ON COMMUNITY FOLLOW-UP DATA USING MACHINE LEARNING. Preventive Medicine Reports, 34, 102256.

Chen, L., Zhang, Y., & Wang, J. (2025). MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN TYPE 2 DIABETES PREDICTION: A COMPREHENSIVE BIBLIOMETRIC ANALYSIS. Frontiers in Digital Health, 7, 1557467.

Fathurrahman, I., Nurhidayati, N., & Nur, A. M. (2023). PREDIKSI DIABETES MENGGUNAKAN ALGORITMA NAIVE BAYES DAN GREEDY FORWARD SELECTION. Jurnal Nasional Teknologi dan Sistem Informasi, 9(3), 187-196.

Harsa, H., Pratiwi, H., & Wibawa, A. D. (2021). IMPLEMENTASI ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT DIABETES MELLITUS. Jurnal Teknologi Informasi dan Ilmu Komputer, 8(4), 745-752.

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2022). MACHINE LEARNING MODELS FOR DATA-DRIVEN PREDICTION OF DIABETES BY LIFESTYLE TYPE. International Journal of Environmental Research and Public Health, 19(21), 14384.

Kadir, T. A., Glebauskiene, B., & Liosis, N. (2020). ANALYSIS AND PREDICTION OF DIABETES COMPLICATION DISEASE USING DATA MINING ALGORITHM. Procedia Computer Science, 167, 1123-1129.

Kusuma, A. W., Sari, R. P., & Widodo, S. (2023). PENERAPAN METODE MACHINE LEARNING UNTUK PREDIKSI DIABETES: STUDI KASUS PADA DATASET PIMA INDIANS DIABETES. Jurnal Sistem Informasi dan Komputer Terapan Indonesia, 5(2), 123-132.

Mujahid, A., Rustam, F., Alvarez, R., Luis Vidal Mazón, J., Díez, I. D. L. T., & Ashraf, I. (2024). PREDICTION OF DIABETES USING DATA MINING AND MACHINE LEARNING ALGORITHMS: A CROSS-SECTIONAL STUDY. Heliyon, 10(4), e25641.

Purnamasari, D., & Wijaya, A. (2021). PREDIKSI RISIKO DIABETES MELLITUS MENGGUNAKAN DECISION TREE C4.5 DAN RANDOM FOREST. Jurnal Informatika dan Komputer, 26(3), 156-165.

Rahman, M. S., & Islam, M. M. (2020). DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS. Procedia Computer Science, 167, 1130-1137.

Sisodia, D., & Sisodia, D. S. (2021). PREDICTION OF DIABETES USING CLASSIFICATION ALGORITHMS WITH FEATURE SELECTION TECHNIQUES. International Journal of Cognitive Computing in Engineering, 2, 85-90.

Wulandari, S., Putri, N. A., & Hidayat, R. (2022). SISTEM PREDIKSI DIABETES MELLITUS MENGGUNAKAN METODE CRISP-DM DAN ALGORITMA DECISION TREE. Jurnal Teknologi Informasi dan Terapan, 8(1), 34-42.