Classification of Type 2 Diabetes using Decission Tree Algorithm

Ivandari [1], Much. Rifqi Maulana[2], M. Adib Al Karomi[3]
[1-2] *Computer Science, STMIK Widya Pratama Pekalongan, Indonesia*

Abstract— Diabetes is a disease that causes many deaths. According to data from WHO, in 2019 there were 2 million deaths due to diabetes. The recording of the patient's condition has been carried out for medical purposes. The large number of records that are only used as stored data will only later become digital waste. Data mining offers a classification process to process data into new knowledge. The recognition of new patterns from existing data results from algorithmic calculation processes as well as statistics. One of the best and widely used classification algorithms for high-dimensional datasets is the decision tree. This study uses the type 2 diabetes dataset from the uci repository which was released in 2020. The results showed that the accuracy of the decision tree algorithm for type 2 diabetes data classification was 95.96%. Another output of this study is a decision tree from the early stage diabetes risk prediction dataset.

Keywords— Data mining, Decision tree, Diabetes type 2.

1. Introduction

Diabetes is classified as a deadly disease in the world [1]. From exploration in 2019, 2 million deaths were recorded due to diabetes [2]. Type 2 diabetes is not a contagious disease or hereditary disease. Most cases of diabetes are caused by unhealthy eating habits of patients. Patients who consume a lot of glucose while the body is unable to control excess glucose in the blood [3]. Processing of glucose in the blood becomes slow due to lack of exercise or the patient is often in a passive state. Early treatment of patients is proven to reduce the risk of patient death.

Data mining is a field of science that studies old datasets to find new patterns or knowledge from the data [4]. The pattern that is formed is the result of mathematical calculations arranged in such a way as to become an algorithm [5]. This calculation model is similar and can be said to be a development of statistics. From these calculations, algorithmic validation and evaluation can be carried out. Evaluation is carried out to find out the best algorithm model that can be used for certain datasets [6].

Data mining can handle estimation, classification, prediction, association and clustering. This process is carried out in accordance with the existing dataset and the desired output. Classification is a process of adjusting new data (training data) based on existing old data (testing data). In the classification process the data used must have a label attribute. The label attribute is the destination attribute to be searched for. The classification process can be used in various fields with large data. In the health sector, classification is widely used. One of the classifications is done for the detection of kidney disease [7]. In addition to detecting kidney disease, the classification process is also carried out for breast cancer detection [8] and early detection of diabetes [9]. In 2021 a comparison of classification algorithms for diabetes will be carried out. This research uses the PIMA dataset. The results of this study the best performance of the classification algorithm is the soft voting classifier with an accuracy rate of 79.08% [10]. In 2023, a similar comparison will also be carried out using the Gradient Boost Machine (GBM). The result is that GBM obtains the best accuracy rate of 80.4% [11].

Several classification models have proven to be good for handling certain data. The data type can affect the performance of a classification algorithm. One of the best and most widely used classification algorithms is the decision tree [12]. Decission tree is an algorithm that can handle datasets with nominal and numeric types. Another advantage of the decision tree is that the resulting output can be easily understood by human language.

This study uses a diabetes dataset from the uci repository. Uci repository is a public dataset provider web that is widely used for algorithm testing. The decision tree algorithm is used to classify the early stage diabetes risk prediction dataset. The level of accuracy of the decision tree algorithm in this study reached 95.96%. Another result of this study is that there is a decision tree that can be easily understood by human language.

2. Literature Review

2.1. Related Research

Research related to the classification of diabetes was conducted in 2021 by Saloni Kumari et al [10]. In this study, a comparison was made of several classification algorithms. The result is that the soft voting classifier has the highest accuracy rate with 79.80%. The dataset used is the diabetes PIMA dataset from the uci repository.

In early 2023 another study regarding the classification of diabetes was conducted by Cesar Carpinteiro [11]. This research, conducted by academics from the University of Minho, Portugal, made a comparison of the classification algorithms for diabetes. In this study, the Gradient Boost Machine (GBM) got the highest accuracy rate of 80.4%.

2.2. Classification of Data Mining

Data mining is a field of science that processes data to seek new knowledge [6]. Large and large data can be useless and become digital waste. Data mining can create new rules, patterns or models from a large dataset. This model can be used to analyze new records in previously unstored data [13]. The process carried out in data mining is actually a mathematical calculation. This calculation uses an algorithm as well as statistical calculations. The main function of data mining is divided into two. Namely supervised learning and unsupervised learning. Classification is part of data mining which belongs to supervised learning. The dataset used in classification is a dataset with label attributes or destination attributes. This label attribute is often referred to as a class. In the classification process learning data is needed to calculate or determine the class of new data. Many algorithms can be used in the classification process. Some algorithms are powerful for certain datasets. However, it is weak when used for other data models. One of the best algorithms and widely used by researchers is the decision tree.

2.3. Decission tree

Decision tree is one of the most widely used classification algorithms. This algorithm is widely used because it is known to be good at handling high-dimensional data. Decision trees can be used for classification or prediction and are proven to be strong [14]. The calculation process in this algorithm is to calculate the gain values for all existing attributes. The attribute with the highest gain value is used as the first node. The node has several branches according to the number of variations of the attribute. This model proved to be very robust for nominal or discrete datasets [15]. Furthermore, the same loop is performed to find other nodes and branches. This process stops when each branch already has a class. Figure 1 is an example of a decision tree from golf data.

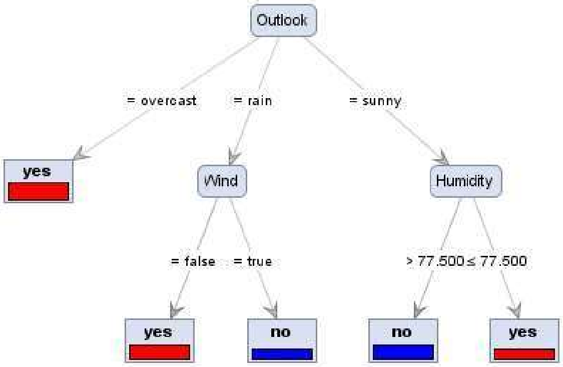
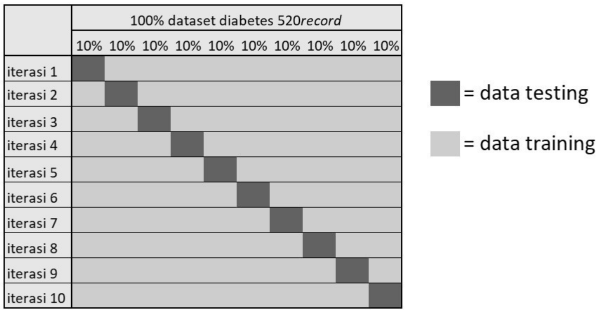


Figure 1. Golf Data Decision Tree

2.4. 10 folds cross validation

Cross validation is a widely used validation process and is proven to have good capabilities for the classification process [16]. The process in this cross validation is to divide the data into several parts and then one part is used as data testing, the rest is used as training data. The distribution of this data is adjusted according to existing needs. The most widely used process in classification research is 10 folds cross validation [17]. Figure 1 is an overview of the process of 10 folds cross validation.



	100% dataset diabetes 520*record*									
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
iterasi 1	■	□	□	□	□	□	□	□	□	□
iterasi 2	□	■	□	□	□	□	□	□	□	□
iterasi 3	□	□	■	□	□	□	□	□	□	□
iterasi 4	□	□	□	■	□	□	□	□	□	□
iterasi 5	□	□	□	□	■	□	□	□	□	□
iterasi 6	□	□	□	□	□	■	□	□	□	□
iterasi 7	□	□	□	□	□	□	■	□	□	□
iterasi 8	□	□	□	□	□	□	□	■	□	□
iterasi 9	□	□	□	□	□	□	□	□	■	□
iterasi 10	□	□	□	□	□	□	□	□	□	■

■ = data testing

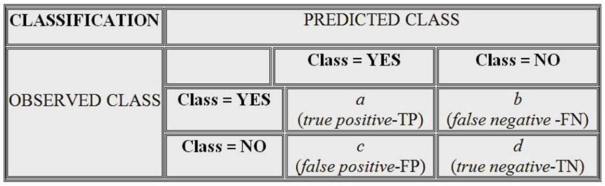
□ = data training

Figure 1. An overview of 10 folds cross validation

From Figure 1 it is shown that there are 10 iterations or repetitions. Iteration 1 uses the first 10% data or the first 52 records out of a total of 520 records for testing and the remaining 468 records are used as training data. The next iteration is carried out until the entire data gets the 1x portion used as testing data.

2.5. Confusion matrix

The confusion matrix is the percentage of the classification process that matches the actual conditions or labels [18]. The final result of the confusion matrix is in the form of a percentage which is also used as the value or level of accuracy of an algorithm. Figure 2 is a representation of the confusion matrix [19]. The confusion matrix evaluation process can be calculated using formula (1) below.



CLASSIFICATION	PREDICTED CLASS		
OBSERVED CLASS		Class = YES	Class = NO
	Class = YES	*a* (*true positive*-TP)	*b* (*false negative* -FN)
	Class = NO	*c* (*false positive*-FP)	*d* (*true negative*-TN)

Figure 2. Representation of the confusion matrix [19]

$$\boldsymbol{accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

a = (true positive) = true classification
b = (false negative) = wrong classification
c = (false positive) = wrong classification
d = (true negative) = true classification

3. Research Methods

The best research method that can be used in this type of research is experimental. The calculation process is carried out using the rapid miner assist application. In more detail, the research method carried out is divided into several stages as follows:

3.1. Data collection

The first step in this research is data collection. The data used is the diabetes dataset which is public data from the uci repository. The complete dataset can be accessed at https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset. This dataset has 17 attributes and 520 data records.

3.2. Validation

Validation is a mandatory process in a classification. Several ways to carry out validation adjust to the problems and the classification process carried out. This study uses 10 folds cross validation and is run with the rapid miner application.

3.2. Decission tree calculation

The decision tree calculation in this study uses the rapid miner assist application. The calculation process uses drag and drop. This process combines 10 folds cross validation using a decision tree and then evaluates it using a confusion matrix. The output of the decision tree calculation is a tree and the level of accuracy of the algorithm.

3.2. Algorithm Evaluation

Evaluation is a process to assess the performance of an algorithm. The evaluation process is carried out using various methods. One of the best evaluation methods and is widely used in classification research is the confusion matrix. The process of calculating this matrix in detail uses the rapid miner application.

4. Results and Discussion

The dataset used in this study is the early stage diabetes risk prediction dataset. This data can be downloaded from https://archive.ics.uci.edu/static/public/529/early+stage+diabetes+risk+prediction+dataset.zip. This dataset is a recap of the records of 520 patients at Sylhet Diabetes Hospital, Bangladesh. This dataset was approved by the relevant physician and donated in November 2020 [20]. To make it easier for researchers to manage this dataset, it has been in the form of a csv file. Table 1 is the metadata of the early stage diabetes risk prediction dataset.

Table 1. Metadata early stage diabetes risk prediction dataset

Role	Name	Type	Statistics	Range	Missings
label	class	binominal	mode = Positive (320), least = Negative (200)	Positive (320), Negative (200)	0.0
regular	Age	integer	avg = 48.029 +/- 12.151	[16.000 ; 90.000]	0.0
regular	Gender	binominal	mode = Male (328), least = Female (192)	Male (328), Female (192)	0.0
regular	Polyuria	binominal	mode = No (262), least = Yes (258)	No (262), Yes (258)	0.0
regular	Polydipsia	binominal	mode = No (287), least = Yes (233)	Yes (233), No (287)	0.0
regular	sudden weight loss	binominal	mode = No (303), least = Yes (217)	No (303), Yes (217)	0.0
regular	weakness	binominal	mode = Yes (305), least = No (215)	Yes (305), No (215)	0.0
regular	Polyphagia	binominal	mode = No (283), least = Yes (237)	No (283), Yes (237)	0.0

regular	Genital thrush	binominal	mode = No (404), least = Yes (116)	No (404), Yes (116)	0.0
regular	visual blurring	binominal	mode = No (287), least = Yes (233)	No (287), Yes (233)	0.0
regular	Itching	binominal	mode = No (267), least = Yes (253)	Yes (253), No (267)	0.0
regular	Irritability	binominal	mode = No (394), least = Yes (126)	No (394), Yes (126)	0.0
regular	delayed healing	binominal	mode = No (281), least = Yes (239)	Yes (239), No (281)	0.0
regular	partial paresis	binominal	mode = No (296), least = Yes (224)	No (296), Yes (224)	0.0
regular	muscle stiffness	binominal	mode = No (325), least = Yes (195)	Yes (195), No (325)	0.0
regular	Alopecia	binominal	mode = No (341), least = Yes (179)	Yes (179), No (341)	0.0
regular	Obesity	binominal	mode = No (432), least = Yes (88)	Yes (88), No (432)	0.0

From the early stage diabetes risk prediction dataset metadata, there are 16 regular attributes and 1 class attribute. This data is classified as balanced data, as evidenced by variations for the positive class of 320 patients and for the negative class of 200 patients. Most of the attribute types in this dataset are binomial. Only one attribute with integer type is the age attribute.

4.1 Validation and Evaluation of Results

The cross validation validation process using the rapid miner application can be seen in Figure 3. Meanwhile, Figure 4 is the process of evaluating the decision tree algorithm that is in the cross validation process. Using the rapid miner application is very easy by dragging and dropping datasets, then using the tools that have been prepared. In the validation process, 10 folds cross validation is used, so the cross validation tools in rapid miner must be set according to 10 iterations.

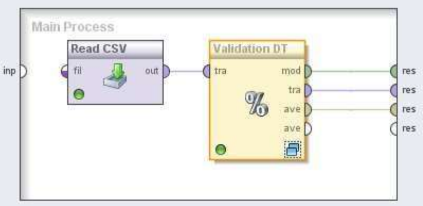


Figure 3. Rapid miner's cross validation process

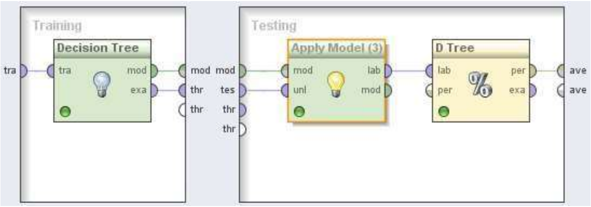


Figure 4. The evaluation process for the rapid miner

Figure 4 shows the training data used by the decision tree algorithm. Then in data testing evaluation is used to measure the level of accuracy of the decision tree algorithm. Table 2 is the result of the confusion matrix of diabetes classification using the decision tree algorithm.

Table 2. Results of the confusion matrix

	True positive	True negative	Class precission
Pred. positive	306	7	97,76%
Pred. negative	14	193	93,24%
Class recall	95,62%	96,50%	

From table 2 it can be seen that there are 520 data records that are classified. The 520 records are divided into 4 parts as the confusion matrix table. A total of 306 positive classifications predicted positive, 193 negative classifications predicted negative. In addition, there are 14 positive classifications which are predicted to be negative, and 7 negative classifications which are predicted to be positive. The level of accuracy of the decision tree can be obtained by calculating (306+193) divided by (306+193+14+7). A value of 95.96% was obtained which is the accuracy level of the decision tree for type 2 diabetes classification.

4.2 Representation of Decision Trees

Figure 5 is the decision tree output created from the calculation results. In the decision tree the polyuria attribute becomes the first node. This means that the polyuria attribute is the attribute with the highest gain value of all other attributes. This calculation continues by separating the yes and no variants of the polyuria attribute. Obtained from all the yes variants in the polyuria attribute it is recalculated that the highest gain value is age. Then for the no variant in the attribute polyuria the attribute with the highest gain value is polydipsia. This process continues until a positive or negative class appears in the classification.

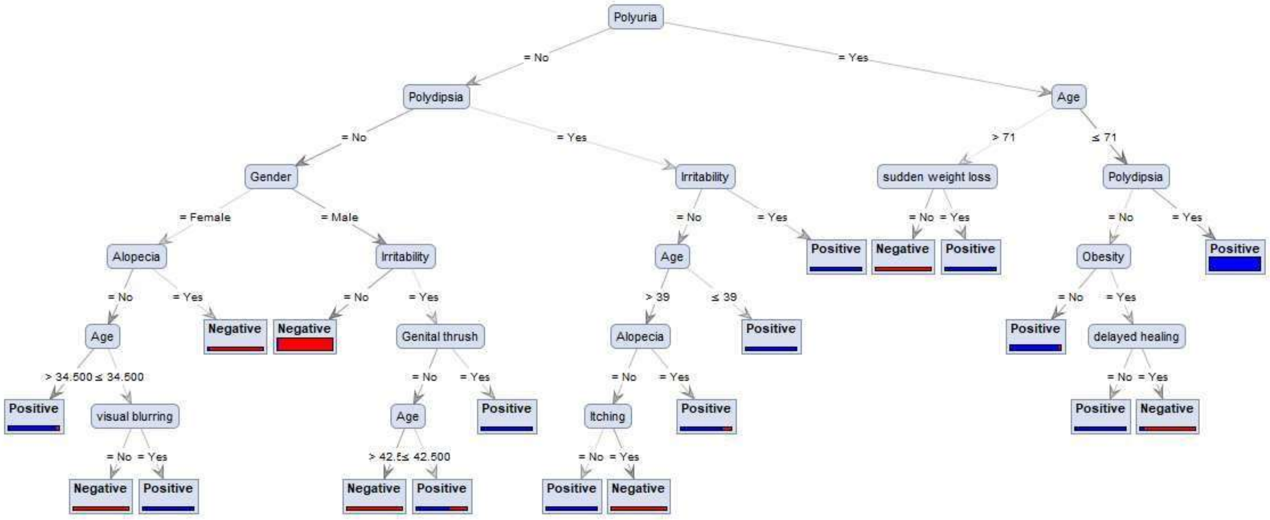


Figure 5. Decision tree formed

4.3 Discussion

Decission tree is an algorithm that can handle numeric and nominal data types. Besides being able to handle various types of data, decision trees are also proven to be able to handle high-dimensional data. In this study, the early stage diabetes risk prediction dataset was used. This dataset has 17 attributes with 16 of them being binomial. With only 2 data variants for some of the attributes, the resulting output is very easy for human language to understand. The performance of this decision tree algorithm is very good with an accuracy value of 95.96%. This level of accuracy is very good for a classification process.

5. Conclusion

Classification of type 2 diabetes data using a decision tree algorithm produces an accuracy rate of 95.96%. The dataset used is public data from the uci repository which was released at the end of 2020. The dataset from the uci repository is widely used for algorithm testing around the world. For further research, pre-processing of data can be carried out. The pre-processing stage can be carried out by selecting features to trim attributes that are less relevant in the classification process.

References

[1] B. J. G. Rozo, J. Crook, and G. Andreeva, "The role of web browsing in credit risk prediction," *Decis. Support Syst.*, p. 113879, 2022, doi: 10.1016/j.dss.2022.113879.

[2] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, "Generating customer's credit behavior with deep generative models," *Knowledge-Based Syst.*, vol. 245, p. 108568, 2022, doi: 10.1016/j.knosys.2022.108568.

[3] M. R. Maulana and M. A. Al Karomi, "Sistem Pendukung Keputusan Persetujuan Kredit Menggunakan Algoritma C4.5," *J. IC-Tech*, vol. Vol. XI No, no. 1, pp. 29–38, 2016, [Online]. Available: http://jurnal.stmik-wp.ac.id/gdl.php?mod=browse&op=read&id=ictech--muchrifqim-80.

[4] Ivandari and M. A. Al Karomi, "Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance," *IC Tech*, vol. 16, no. 1, pp. 12–15, 2021, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137.

[5] M. A. Al Karomi, M. R. Maulana, S. J. Prasetiyono, Ivandari, and Arochman, "Strengthening campus finance by analyzing attribute attributes for student registration classifications." p. 1, 2019, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/1431.

[6] V. K. Xindong Wu, *The Top Ten Algorithm in Data Mining*. 2009.

[7] Ivandari and M. A. Al Karomi, "Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm," *Jaict*, vol. 6, no. 1, pp. 44–49, 2021, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/2896.

[8] Ivandari, T. T. Chasanah, S. W. Binabar, and M. A. Al Karomi, "Data Attribute Selection with Information Gain to Improve Credit Approval Classification Performance using K-Nearest Neighbor Algorithm," *IJIBEC*, vol. I, pp. 15–24, 2017.

[9] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute

Selection using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[10] I. Indrayanti, S. Devi, and M. A. Al Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-TECH*, vol. XIII, no. 2, pp. 1–6, 2017, [Online]. Available: ejournal.stmik-wp.ac.id.

[11] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2007.

[12] O. Maimoon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, vol. 40, no. 6. Springer, 2010.