Jurnal Sistem Informasi ISSN P : 2598-599X; E: 2599-0330

# DATA MINING CLASSIFICATION ALGORITHMS FOR DIABETES DATASET USING WEKA TOOL

Rahma Fitria[1], Desvina Yulisda[2] , Mutammimul Ula [3]
Sistem Informasi Universitas Malikussaleh Lhokseumawe
Jl. Cot Tgk Nie-Reulet, Aceh Utara, 141 Indonesia
email: rahmafitria@unimal.ac.id, desvina.yulisda@unimal.ac.id,
mutammimul@unimal.ac.id

**Abstrak**

*Data mining explores a huge amount of data to extract the information to be meaningful. In the field of public health, data mining hold a crucial contribution in predicting disease in early stage. In order to detect diseases, the patients need to conduct various tests. In the context of disease predicion, Data mining techniques aims to reduce the test that patients need to accomplish. Also the techniques is used to increase the accuracy rate of detection. Nowadays, diabetes attacks many adults in the world. Moreover, in order to reduce the number of adult having diabetes, an effective and efficient diabetes detection mechanism should be found. This report will apply some data mining techniques on diabetes dataset that has been downloaded at UCI Machine Learning Repository.Three kind of classification algorithm such as Naïve Bayes Classifier, Multilayer Perceptrons (MLP's) and Desicion Tree (J.48) have been performed on this dataset. Obtained outcomes indicated that Naïve Bayes Classifier achieved the highest accuracy with 76,30%. As the result, this algorithm is a good method to classify and diagnose diabetes diseases on studying dataset.*

***Keywords:*** *Data mining, Classification, Diabetes, Naïve Bayes, Multilayer Perceptrons, Desicion Tree, WEKA*

## 1. INTRODUCTION

Diabetes mellitus is a chronic condition that arises when the pancreas fails to produce sufficient insulin or when the body fails to adequately use the insulin that is produced. Insulin is a hormone that transfers glucose from the bloodstream to the body's cells for use as energy (Chaves & Marques, 2021). As a result, the person is unable to get the glucose that enters the bloodstream from consumed food, and blood glucose levels rise. Diabetes Mellitus is a disease that contributes significantly to the development of a number of severe ailments, including kidney failure and blindness. (American Diabetes Association, 2005).

Diabetes is a deasease that is commmonly faced by U.S adults among aged 45 years or older. According to U.S fact sheet data 2017, 88 million population in U.S have diabetes over the total number of population. In addition, 1 of 3 from the U.S people have prediabetes and almost 90% of

them do not realize having it (Centers for Disease Control and Prevention, 2017). According to the above data facts, diabetes becomes the most common disease that treated adults that can cause any other health complications. Applying data mining methods for diabetes will utilize huge volumes of diabetes dataset in order to gather knowledge. The classification algorithm will be used to process the dataset. There are three algorithms to be used in this report such as Naive Bayes classifier, Multilayer Perceptron and Decision Tree in order to predict the best methods for diabetes dataset.

The proposed of this study is to analyse how the different classification algorithm can be performed to a training dataset. This data mining tool can be utulized in diagnosing diabetes in early stage.

## 2. RELATED WORK

In 2021, Alpan and Ilgi conducted data mining classification techniques in order to classify diabetes diseases by utilizing WEKA tools. They performed 7 different data mining strategies including Random Tree, Random Forest, Bayes Network, Decision Tree, Naïve Bayes, SVM and k-NN. The used data set have been implemented to predict and compare in order to acquire the best accuracy technique (Hasdyna et. al.,2020) (Ula et al., 2021). The studied exhibited that k-NN performed well with the maximum accuracy which is 98.07%.

In 2020, Shuja, Mittal, & Zaman performed 5 kind of data mining models in order to have early prediction of diabetes diseases. These models are Desicion Tree, MLP'S, Bagging, Simple Logistic and SVM in order to find the most accuracy techniques for classification of diabetes dataset by utilizing WEKA tool as well. The result inferred that decision tree is achieving the highest accuracy models with 94%.

In the other hand, other researchers studied 12 of classification techniques such as Naïve Bayes, MLP'S, Lazy IBK, Ada Boost-M1, Lazy K-Star, Logistic Regression, Random Tree, J48, Random Forest, Decision Table, Multiclass Classifier and J-Rip by using WEKA tool. The research indicates the Lazy K-Star, Lazy IBK, Random Forest and Random Tree are having outstandingly well with nearly 100% accuracy (Kumar, Mishra, Mazzara, Thanhx, & Verma, 2019).

## 3. METHODOLOGY

The Classification is the process to identify a new observation category of training set of data. This report will compare three distinct classifiers such Naive bayes, Multilayer Perceptron, and Decision Tree. This dataset will be implemented by using WEKA tool.

### 3.1 Naïve Bayes Classifier

Naive Bayes classifier is such a well known type of classifiers. The Bayes theorem is used to create a probabilistic classifier called Naive Bayes. It is also a collection of programs that, based on the descriptive attributes, assign a class from a preset set to an object under creation. This is accomplished through the use of a probabilistic technique that computes class probabilities and forecasts the most likely classes. Furthermore, this technique suggests that predictors are independent, despite the fact that the stated independence assumptions may not hold true in practice (Alpan & Ilgi, 2020). This method is also useful when dealing with excessively large datasets (Flach, 2004).

### 3.2 Multilayer Perceptrons (MLP'S)

The multilayer perceptron (MLP'S) is one of the most extensively used neural network classification technique (Rahman & Afroz, 2013). During simulations with the PIDD dataset, the MLP was built with a three-layer feed-forward neural network: one hidden, one input, and one output layer. The following parameters have been chosen for the model: learningRate = 0.3/0.15; momentum = 0.2; randomSeed = 0; validationThreshold = 20; Number of Epochs = 500. In addition, Multilayer Perceptrons (MLPs) are universal in that they may arbitrarily well approximate any continuous nonlinear function on a compact interval. As a result, MLPs became popular for parametrizing nonlinear models and classifiers, with often better results than traditional approaches.

### 3.3 Desicion Tree (J.48)

Decision Tree (J.48) is a tree that have the shape of a flow chart. Most decision tree induction algorithms use a top-down approach, beginning with a training set of tuples and their associated class labels (Han, Kamber, & Pei, 2012). It is used as a classification and prediction method with nodes and internodes as representation. The test cases that are used to separate the instances with different features are the root and internal nodes. Internal nodes are the output of attribute test cases. The class variable is represented by the leaf nodes (Iyer, Jeyalatha, & Sumbaly, 2015).

During categorization, a decision tree is generated using Decision Tree (J.48). One of the most common tools in data mining is a system that

creates classifiers. This divides the data collected during the inspection into branches, which can then be used to form a tree for better classification accuracy. The inputs are assigned to one of a small number of classes with a set of fixed attributes, and the output is a classifier that reliably predicts the class to which the case belongs.

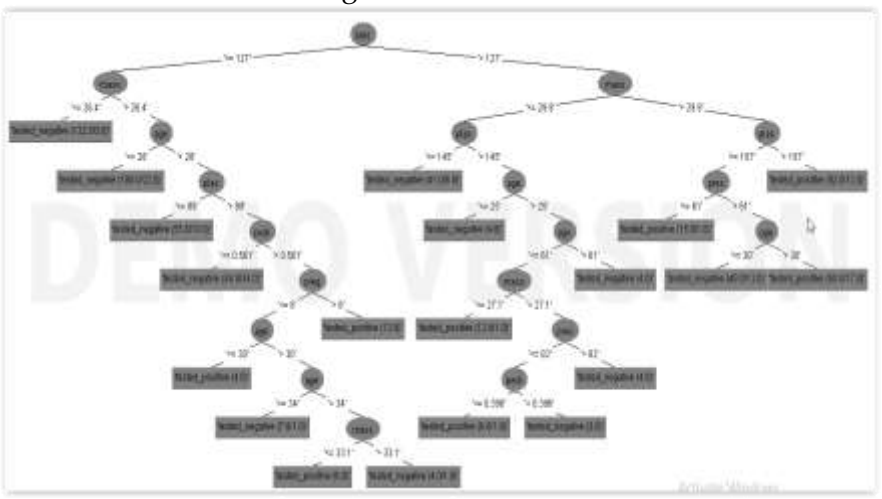

Fig.1. Decision tree for J48 algorithm for diabetes dataset after experiment with WEKA

## 4. EVALUATION AND TESTING

### 4.1 Cross Validation

Cross validation is a data mining approach for evaluating classification algorithm performance. It is used to assess the rate of error in learning procedures (Thirumal & Nagarajan, 2015). The dataset is divided into n folds, with each fold serving as a testing and training platform. In testing and training, the procedure is repeated n times. The data is separated into ten parts, each of which is roughly the same, to produce the whole dataset in a 10 fold cross validation. Each term is held out, and the error rate in the holdout set is determined during the learning scheme that focused on the remaining nine-tenths. The learning procedure is repeated ten times on training sets, and the error rates for the ten sets are averaged to yield an overall error rate.

### 4.2 Confusion Matrix

The confusion matrix is used to display the accuracy of classifiers derived through classification. It is used to demonstrate the connection between results and expected classes.

Table 1 Confusion matrix

| Confusion Matrix | | Targeted Values | |
|---|---|---|---|
| | | Positive | Negative |
| Model | Positive | a | b |
| | Negative | c | d |

## 5. EXPERIMENTAL RESULT

Before using a learning system, its accuracy must be checked. Because of the scarcity of data, measuring accuracy is a challenging process. Choosing a good assessment technique is critical in the context of a machine learning system. There are various approaches for separating data into training and testing sets.

The accuracy comparison of these three classification algorithm can be descibed as the table below:

Table 2 The classification result

| Algorithm | Accuracy (%) |
|---|---|
| Naïve Bayes | 76.3021 |
| Multilayer Perceptron | 75.3906 |
| Decision Tree (J.48) | 73.8281 |

From the above result, Naive Bayes performs the best classifying process than Multilayer Perceptron and Decision Tree (J.48).

### 5.1 Classification Accuracy

#### 5.1.1 Naïve Bayes

Confusion Matrix

a b ← classified as
422 78 | a = tested_negative
104 164 |b = tested_positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

= 422 + 164/ 422+164+78+104
= 586/768
= 0.7630208
= 0.7630208 x 100%

= 76.3021%

### 5.1.2 Multilayer Perceptron (MLP'S)

Confusion Matrix

a b ← classified as
416 84 | a = tested_negative
105 163 |b = tested_positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

= 416 + 163/ 416+163+84+105
= 579/768
= 0.753906
= 0.753906 x 100%
= 75.3906%

### 5.1.3 Decision Tree (J.48)

Confusion Matrix

a b ← classified as
407 93 | a = tested_negative
108 160 |b = tested_positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

= 407 + 160/ 407+160+93+108
= 567/768
= 0.738281
= 0.738281 x 100%
= 73.8281%

## 6. CONCLUSION

As a researcher, it's very interesting applying data mining models to medical datasets because there are numerous problems with health and cases to research. Diabetes Mellitus is one of the challenging diseases effected to adult in the world. Classification is extremely beneficial method for knowledge exploration as It has the ability to precisely and efficiently classify data. In this study, the classification of diabetes dataset applied three classification algorithms which are Naives Bayes, Multilayer Perceptron and Decicion Tree (J.48) taken from the University of California, Irvine's machine learning repository. These three classification algorithms is done based on the performance factors classification accuracy in order to

discover the most effective classification strategy in diabetes diagnosis. For diabetes dataset, it can be concluded that the best classification algorithms is Naives Bayes technique with 76,30% accuracy. Based on the attributes, Naives Bayes technique is a reliable and an effective technique in investigating diabetes disease compared to MLP'S and J.48. As a result, It can help medical workers to make a quick desision in handling this disease.

## REFERENCES

Alpan, K., & Ilgi, G. S. (2020). Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. *4th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2020 - Proceedings*. https://doi.org/10.1109/ISMSIT50672.2020.9254720

American Diabetes Association. (2005). Diabetes Mellitus and Other Categories of Description of Diabetes. *World Health*, *28*(Suppl 1), 224102. https://doi.org/10.2337/diacare.27.2007.S5

Centers for Disease Control and Prevention. (2017). *Diabetes and Prediabetes and improve the health of all people with diabetes* . Retrieved from www.cdc.gov/chronicdisease

Chaves, L., & Marques, G. (2021). applied sciences Data Mining Techniques for Early Diagnosis of Diabetes. *Appl. Sci.*, *11*(2218), 1–12. https://doi.org/doi.org/10.3390/app11052218

Flach, P. A. (2004). Naive Bayesian Classification of Structured Data. *Machine Learning*, *57*(1), 233–269.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Third Edit). Waltham: Morgan Kaufmann.

Hasdyna, N., & Dinata, R. K. (2020). Analisis Matthew Correlation Coefficient pada K-Nearest Neighbor dalam Klasifikasi Ikan Hias. INFORMAL: Informatics Journal, 5(2), 57-64.

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(1), 1–14.

Kumar, V., Mishra, B. K., Mazzara, M., Thanhx, D. N. H., & Verma, A. (2019). Prediction of malignant & benign breast cancer: A data mining approach in healthcare applications. *ArXiv*, 1–8.

Rahman, R. M., & Afroz, F. (2013). Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. *Journal of Software Engineering and Applications*, *2013*(March), 85–97.

Shuja, M., Mittal, S., & Zaman, M. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. In *Advances in Computing and Intelligent Systems* (pp. 195–211). https://doi.org/10.1007/978-981-15-0222-4_17

Thirumal, P. C., & Nagarajan, N. (2015). Utilization Of Data Mining Techniques For Diagnosis Of Diabetes Mellitus - A Case Study. *ARPN Journal of Engineering and Applied Sciences*, *10*(1), 8–13.

Ula, M., Ulva, A. F., & Mauliza, M. (2021). Implementasi Machine Learning Dengan Model Case Based Reasoning Dalam Mendiagnosa Gizi Buruk Pada Anak". Jurnal Informatika Kaputama (JIK), 5(2), 333-339.