ICIC Express Letters
Part B: Applications
Volume **12**, Number **12**, December **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 1177–1184

APPLICATION OF DISCRETIZATION AND ADABOOST METHOD TO IMPROVE ACCURACY OF CLASSIFICATION ALGORITHMS IN PREDICTING DIABETES MELLITUS

Annisa Maulana Majid and Wiranto Herry Utomo

Information Technology Department
Faculty of Computing
President University
Jl. Ki Hajar Dewantara, Kota Jababeka, Cikarang Baru, Bekasi 17550, Indonesia
annisa.majid@student.president.ac.id; wiranto.herry@president.ac.id

Received March 2021; accepted June 2021

Abstract. *The death rate caused by diabetes mellitus can be reduced if there is an accurate diagnosis early on. Previous research in predicting diabetes mellitus with an accuracy level has been carried out but has resulted in little accuracy on the Decision Tree C4.5 algorithm and the K-Nearest Neighbor (KNN) algorithm. For this reason, it is necessary to increase accuracy in order to produce accurate information. The purpose of this study was to improve the accuracy of the Decision Tree C4.5 and K-Nearest Neighbor (KNN) classification algorithms using the Pima Indian Diabetes Dataset data by applying the discretization technique and the ensemble method, namely Adaboost and to handle numerical attributes. The results of this study with a single algorithm resulted in an accuracy of 76.31% for the Decision Tree and 73.21% for the KNN. The Decision Tree used discretization and Adaboost techniques of 83.67% and KNN using discretization and Adaboost techniques of 83.18%. The results showed an increase in classification algorithms using discretization and Adaboost techniques.*

Keywords: Discretization, Diabetes mellitus, Decision Tree C4.5, Ensemble technique, Adaboost

1. **Introduction.** Diabetes mellitus is a disease that can cause complications and even death. Diabetes mellitus is a disease with an increased prevalence rate of diagnosis from 2013-2018 based on the results of the 2018 Basic Health Research (Rikesdas). For this reason, early treatment is needed to prevent complications and premature death by making an early diagnosis of diabetes mellitus. Research on diabetes mellitus diagnosis has been carried out using various algorithms, including Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), K-means, Support Vector Machine (SVM), and others, using the Pima Indian Diabetes Dataset. However, the results of previous research indicate that the level of accuracy of the Decision Tree and KNN has a low level of accuracy with an insignificant difference. Therefore, it is necessary to handle datasets that have continuous attributes using the discrete technique. Discritization technique is one of the most basic data reduction techniques, which focuses on transferring continuous or numeric attributes to discrete or nominal attributes with finite numbers at intervals [1]. The assessment of previous research has been carried out using discretization. Tigga and Garg's research in 2020 used the KNN algorithm with 70.8% accuracy, and a Decision Tree of 69.7% [2]. However, the resulting level of accuracy does not increase significantly, so it is necessary to increase the accuracy to provide the best decision results. Adaptive Boosting stands for Adaboost, which is an ensemble algorithm with the advantage of focusing on misclassified tuples and a higher level of accuracy. Previous research on discretization and Adaboost techniques had been applied independently which resulted in an increase in accuracy; therefore, this

DOI: 10.24507/icicelb.12.12.1177

research was conducted by combining discretization and Adaboost techniques to produce better accuracy. Discretization technique to handle the Pima Indian Diabetes Dataset which is a dataset with numerical attributes and the Adaboost method is used to improve the accuracy of the classification algorithm in predicting diabetes mellitus. The classification algorithm has weaknesses, one of which is overfitting which can cause misclassification because of noisy data and can produce a low level of accuracy, so it is necessary to increase the accuracy of the classification algorithm. This study uses discretization techniques and ensemble methods, namely Adaboost with the Pima Indian Diabetes Dataset which aims to increase the level of accuracy in predicting diabetes mellitus. Research begins with the introduction, including problem identification, second, the literature review section is a brief description of the research, the third part is the methodology that describes the methods used in the study, the fourth part is the results and discussion containing the results of research and comparisons of other studies and finally, the fifth part concludes the paper.

2. **Literature Review.** Research on the diagnosis of diabetes mellitus has been carried out using various algorithms, one of which was developed by Wu et al. to make a prediction model of the K-Means algorithm and logistic regression for high accuracy and adapt to several datasets on diabetes mellitus analysis [3]. Deepti and Sisodia designed a diabetes patient likelihood model with maximum accuracy, using the Decision Tree, SVM and Naïve Bayes [4]. Tigga and Garg predicted the risk value of type 2 diabetes using Logistic Regression, KNN, SVM, Naïve Bayes Classification, Decision Tree and Random Forest [2]. Vigneswari et al. compared the decision method in Machine Learning, and the methods used were Random Forest, C4.5, Random Tree, REPTree, and Logistic Model Tree [5]. Hebbar et al. developed a DRAP method with a hybrid technique to predict diabetes mellitus, using a decision tree and a random forest classifier [6]. Jasim et al. measured and evaluated the performance of the classification method with the spiritual spinning technique used, namely KNN and Artificial Neural Network [7]. Tsai and Chen combined feature selection and discretization in the supervised learning and unsupervised learning methods, and the methods used in the study are the feature selection method, including principal component analysis, Genetic Algorithm (GA), and Decision Tree C4.5, and the classification method is the SVM and the Decision Tree [1]. Mazini et al. evaluated and classified the features of the IDS System for IDS detection accuracy, using the Adaboost method [8]. Maryono et al. implemented numerical discretization attributes for outliers detection of mixed datasets, and the methods used are the Z-Discretization technique and clustering-based discretization using K-means for discretization [9]. The results of existing research show that the comparison of the accuracy of the Decision Tree C4.5 and the KNN has a low level of accuracy. Discritization technique can handle numeric data to nominal with finite numbers with intervals but cannot yet handle the level of accuracy. The classification algorithm has weaknesses of overfitting, and there needs to be an increase in the accuracy of the classification algorithm. This study uses discretization techniques and Adaboost methods to improve the classification algorithm.

3. **Methodology.** This study uses the application of discretization techniques and Adaboost methods to the classification algorithm, namely Dicision Tree C4.5 and KNN to improve accuracy in predicting diabetes mellitus. The following are the stages in the research.

3.1. **Dataset.** The data source used in this study came from public data, namely the Pima Indian Diabetes Dataset from GitHub with a total of 768 data. The Pima Indian Diabetes Dataset consists of 9 attributes, namely as follows:

- Number of times pregnant
- Plasma glucose concentration

- Diastolic blood pressure
- Triceps skin fold thickness
- 2-hour serum insulin
- Body mass index
- Diabetes pedigree function
- Age
- Class

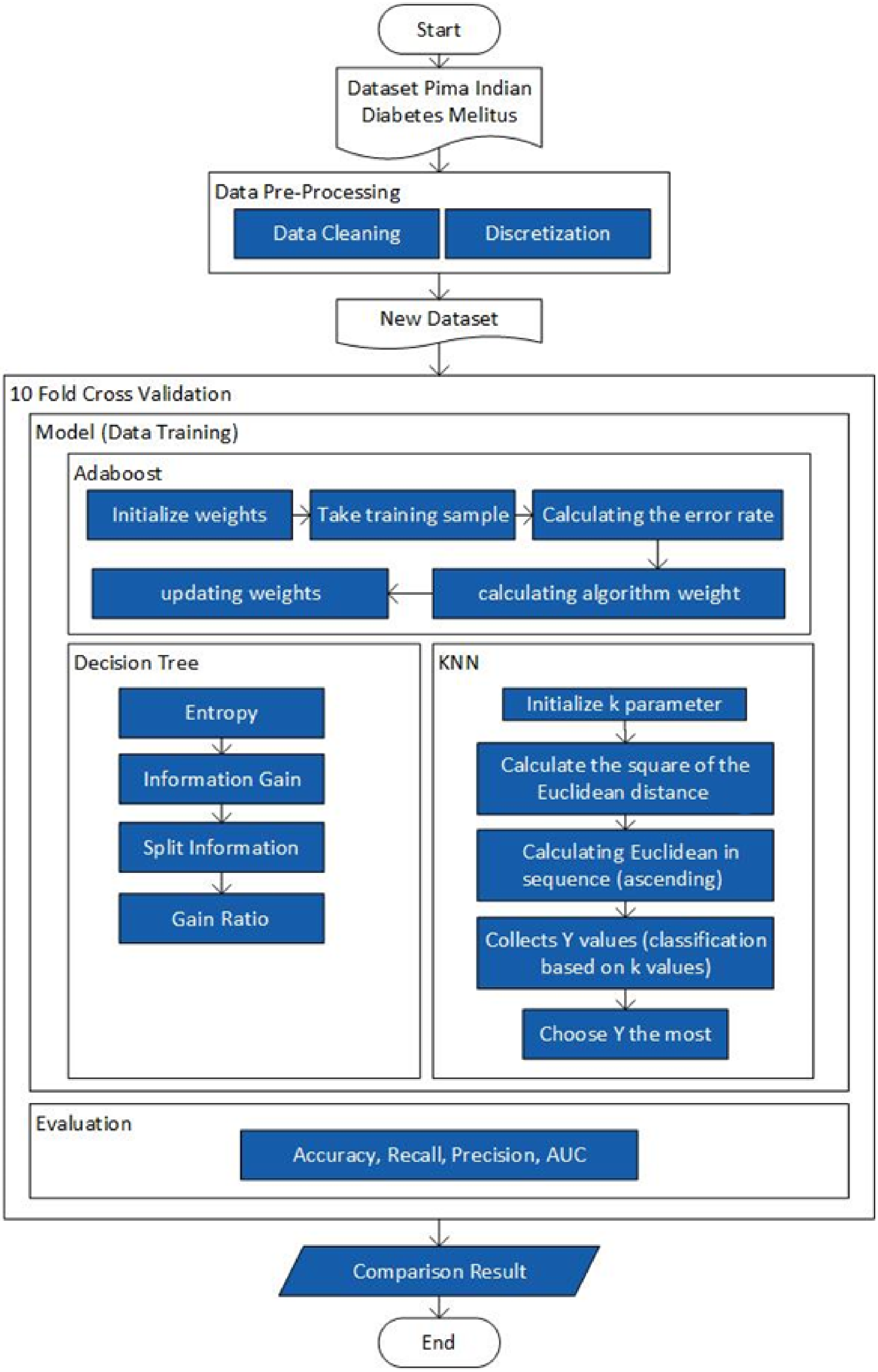

Figure 1. Research stages

In the dataset, the number of positive values is 268 data, while negative values are 500 data.

3.2. **Preprocessing data.** The data preprocessing stage needs to be done to produce quality data in making a decision or determining accuracy. Data preprocessing is used to clean data from missing values, inconsistencies, incomplete data, and noise data. The following are the stages of preprocessing data, namely

1) **Data Cleaning**. Data cleaning is data filtering or data cleaning that is not needed due to data errors to produce quality data. In the Pima Indian Diabetes Dataset, there are no attributes with a missing value, but there are 6 attributes that have a value of 0. The attribute number of times pregnant can be considered the true value because it states how many pregnancies, if it is 0 it means that you have never been pregnant but for the attributes of plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, and body mass index, the value of 0 in the data cannot occur, and therefore these attributes are missing values. Cleaning data from missing values uses a technique of deleting tuples or data records with a value of 0, namely using the filter examples feature in the RapidMiner application.
2) **Discretization Technique**. The diabetes mellitus dataset from Pima Indian Diabetes produces data in the form of numerical attributes for it to be handled using discretization techniques. Discretization can minimize the number of intervals without significant loss of a dataset. Discritization technique is one of the most basic data reduction techniques, which focuses on transferring continuous or numeric attributes to discrete or nominal attributes with finite numbers at intervals [1]. Discretization is a data reduction technique that aims to project a continuous set of values into a discrete and finite space [10].

3.3. **Decision Tree C4.5.** Decision Tree C4.5 is an algorithm developed from the ID3 algorithm. Decision Tree algorithm is included in the classification algorithm category [11]. Algorithm selection is based on information gain [12]. Decision Tree on the ID3 algorithm has been improved and changed to the C4.5 algorithm, and one of the improvements of the algorithm is in terms of pruning. In the Decision Tree, pruning techniques are used to avoid over fitting [13]. Decision Tree C4.5 is a simple algorithm so that users can easily understand the meaning of the rules formed in this algorithm [14].

3.4. **K-Nearest Neighbor (KNN).** The KNN algorithm is a non-parametric method that is widely used for classification in pattern recognition. The main principle of the KNN is that the categories of data points are determined according to the classification of K's closest neighbors [15]. KNN is an algorithm that is widely used in pattern formation in classification algorithms, but it affects the sensitivity of the size k, so it can reduce accuracy [16].

3.5. **Adaboost method.** Adaboost stands for Adaptive Boosting, and Adaboost is a technique of giving weight to weak classifications and aggregating them into strong classifications [17]. Adaboost was successfully applied because the theory in the Adaboost technique was strong, the predictions produced were accurate, and it was implemented simply [18].

3.6. **Data modeling and validation techniques.** In this study using k-folds cross validation as a method of validation with a value of $k = 10$. Validation is carried out to test the algorithm model used. K-folds cross validation is a method to determine the success rate of the algorithm model by retesting random input attributes, in this method the data is divided into k subsets randomly, one subset is used for testing data and the rest is for training data [19]. The k value used is 5 or 10, commonly called 10 folds cross validation, where the data is divided into 10 parts, 90% is for training and the other

10% is used for testing. The process is repeated up to 10 times or 10 iterations until all data records are part of the testing data [20]. How k-folds cross validation works, namely the total data is divided into n parts, iteration or fold 1, namely the 1st part becomes testing, the remaining part becomes the training data, then calculate the accuracy using the following equation:

$$\text{Accuracy} = \frac{\text{the number of classifications is correct}}{\text{amount of test data}} \times 100\% \tag{1}$$

In the 2nd fold, where the 2nd part becomes testing, the rest becomes training, then calculates the accuracy, the process is repeated until it reaches the k-fold. Calculate the average of all k values, the accuracy result is the final accuracy result. In the validation process, the modeling is carried out, in this study using the ensemble method with the boosting technique, namely Adaboost and using the Decision Tree C4.5 and KNN on the training data. After that, the evaluation process is continued with a confusion table and ROC curve. The results of confusion table are used to provide accuracy, recall, and precision in classification algorithms. Accuracy is the percentage between the predicted value and the actual value that exists. Recall is the percentage of the success value of the algorithm used. Precision is an accuracy value with a predicted class. Here is a confusion table.

Table 1. Confusion table

Confusion matrix		Predicted value	
		Positive	Negative
True value	Positive	TP	FN
	Negative	FP	TN

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{4}$$

where TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative.

Receiver Operating Characteristic (ROC) is used to evaluate accuracy results in graphical form. ROC is a curve that will produce the Area Under Curve (AUC) value. AUC is the area accuracy value under the curve generated by the ROC [21]. The accuracy of the AUC value can be classified into 5 groups [22] among others, namely

- 0.90-1.00 = Exellent Classification
- 0.80-0.90 = Good Classification
- 0.70-0.80 = Fair Classification
- 0.60-0.70 = Poor Classification
- 0.50-0.60 = Failure

4. Results and Discussion.

The research conducted is testing I using the Decision Tree C4.5 classification algorithm, testing II using the discretization technique and the Adaboost method with the Decision Tree C4.5 classification algorithm, testing III using the KNN classification algorithm, and testing IV using the discretization technique and the Adaboost method with the algorithm. KNN classification yields accuracy, recall, precision, and AUC values. The preprocessing technique uses the technique of reducing observations on records with a value of 0 and reducing unused attributes. The following is a table of results from 4 tests that have been carried out.

Table 2. Research result

	Testing I	Testing II	Testing III	Testing IV
	Decision Tree	Decision Tree + Discretization + Adaboost	KNN	KNN + Discretization + Adaboost
Accuracy	76.31%	83.67%	73.21%	83.18%
Recall	87.36%	90.81%	84.74%	86.61%
Precision	80.60%	85.79%	77.58%	88.23%
AUC	0.677	0.800	0.733	0.813

Based on the results obtained, it shows that there is an increase in the accuracy, recall, precision, and AUC values. The Decision Tree C4.5 and KNN algorithms using the discretization technique and the Adaboost method increased compared to using only one learning technique. The following are the results of comparisons with other research.

Table 3. Comparison with other experiment

Method	**Accuracy**	**Reference**
DT	76.31%	This paper
DT + Discretization + Adaboost	83.67%	This paper
KNN	73.21%	This paper
KNN + Discretization + Adaboost	83.18%	This paper
KNN	77.3%	Tigga and Garg [2]
DT	73.82%	Deepti and Sisodia [4]
SVM	65.1%	Deepti and Sisodia [4]
NB	76.3%	Deepti and Sisodia [4]
DT	76.25%	Vigneswari et al. [5]
Random Forest	78.54%	Vigneswari et al. [5]
Random Tree	72.41%	Vigneswari et al. [5]
REPTree	75.48%	Vigneswari et al. [5]
Logistic Model Tree	79.31%	Vigneswari et al. [5]
DT	72%	Hebbar et al. [6]
Random Forest	76.5%	Hebbar et al. [6]
KNN	77.24%	Jasim et al. [7]

The comparison results show that there is an increase when applying the discretization technique and the Adaboost method to the classification algorithm. The increase occurs due to the factor of changing the attribute from nominal to interval using the discretization technique. In addition, giving weight to a single algorithm with the Adaboost method can improve the accuracy of the classification algorithm. However, the problem in this study is the reduction of class attributes when implementing data preprocessing with discretization techniques. Further research needs to be done to process data pre-processing using other techniques so that the results of the accuracy are more accurate.

5. **Conclusions.** The results of the tests that have been carried out in this study can be concluded that the discretization technique and the Adaboost ensemble method can improve the accuracy of the classification algorithm, namely Decision Tree C4.5 and KNN in diagnosing diabetes. From the test results, the highest accuracy is produced by the second test, namely applying the discretization technique and the Adaboost ensemble method, which can improve the accuracy of the Decision Tree C4.5 algorithm. The accuracy results obtained were 83.67% and increased by 7.36% from the single Decision Tree results of 76.31%. This research can be used to assist medical personnel in predicting

diabetes early. Further research can be carried out to predict other diseases or by using other methods to increase the level of accuracy.

Acknowledgment. This work is partially supported by President University. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] C.-F. Tsai and Y.-C. Chen, The optimal combination of feature selection and data discretization: An empirical study, *Information Sciences*, vol.505, pp.282-293, DOI: 10.1016/j.ins.2019.07.091, 2019.

[2] N. P. Tiggaa and S. Garg, Prediction of type 2 diabetes using machine learning classification methods, *Procedia Comput. Sci.*, vol.167, pp.706-716, DOI: 10.1016/j.procs.2020.03.336, 2020.

[3] H. Wu, S. Yang, Z. Huang, J. He and X. Wang, Type 2 diabetes mellitus prediction model based on data mining, *Informatics in Medicine Unlocked*, vol.10, pp.100-107, DOI: 10.1016/j.imu.2017.12.006, 2018.

[4] S. Deepti and D. S. Sisodia, Prediction of diabetes using classication algorithms, *Procedia Computer Science*, vol.132, pp.1578-1585, DOI: 10.1016/j.procs.2018.05.122, 2018.

[5] D. Vigneswari, N. K. Kumar, V. G. Raj, A. Gugan and S. R. Vikash, Machine learning tree classifiers in predicting diabetes mellitus, *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS2019)*, Coimbatore, pp.84-87, DOI: 10.1109/ICACCS.2019.8728388, 2019.

[6] P. A. Hebbar, M. V. M. Kumar and H. A. Sanjay, DRAP: Decision tree and random forest based classification model to predict diabetes, *2019 1st International Conference on Advances in Information Technology (ICAIT2019)*, pp.271-276, DOI: 10.1109/ICAIT47043.2019.8987277, 2019.

[7] I. S. Jasim, A. D. Duru, K. Shaker, B. M. Abed, H. M. Saleh and I. Technology, Evaluation and measuring classifiers of diabetes diseases, *The International Conference on Engineering & Technology (ICET2017)*, Antalya, Turkey, DOI: 10.1109/ICEngTechnol.2017.8308165, 2017.

[8] M. Mazini, B. Shirazi and I. Mahdavi, Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms, *Journal of King Saud University – Computer and Information Sciences*, vol.31, no.4, pp.541-553, DOI: 10.1016/j.jksuci.2018.03.011, 2019.

[9] D. Maryono, P. Hatta and R. Ariyuana, Implementation of numerical attribute discretization for outlier detection on mixed attribute dataset, *International Conference on Information and Communications Technology*, pp.715-718, DOI: 10.1109/ICOIACT.2018.8350795, 2018.

[10] S. Ramírez-Gallego, S. García and F. Herrera, Online entropy-based discretization for data streaming classification, *Futur. Gener. Comput. Syst.*, vol.86, pp.59-70, DOI: 10.1016/j.future.2018.03.008, 2018.

[11] S. Manhas, S. Taterh and D. Singh, A novel approach for phishing websites detection using decision tree, *International Journal of Advanced Science and Technology*, vol.29, no.3, pp.943-952, 2020.

[12] S. Guggari, V. Kadappa and V. Umadevi, Non-sequential partitioning approaches to decision tree classifier, *Future Computing and Informatics Journal*, vol.3, no.2, pp.275-285, DOI: 10.1016/j.fcij.2018.06.003, 2018.

[13] A. Al-Qerem, G. Alnaymat and M. Alhasan, Model improvement through comprehensive preprocessing for loan default prediction, *International Journal of Scientific & Technology Research*, vol.9, no.1, pp.1314-1318, 2020.

[14] N. E. I. Karabadji, I. Khelf, H. Seridi, S. Aridhi, D. Remond and W. Dhifli, A data sampling and attribute selection strategy for improving decision tree construction, *Expert Syst. Appl.*, vol.129, pp.84-96, DOI: 10.1016/j.eswa.2019.03.052, 2019.

[15] Y. Guo, S. Han, Y. Li, C. Zhang and Y. Bai, K-nearest neighbor combined with guided filter for hyperspectral image classification, *Procedia Comput. Sci.*, vol.129, pp.159-165, DOI: 10.1016/j.procs.2018.03.066, 2018.

[16] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao and H. Yang, A generalized mean distance-based k-nearest neighbor classifier, *Expert Syst. Appl.*, vol.115, pp.356-372, DOI: 10.1016/j.eswa.2018.08.021, 2019.

[17] S. Cheng, B. Liu, Y. Shi, Y. Jun and B. Li, *Data Mining and Big Data*, Springer, 2015.

[18] E. Listiana and M. A. Muslim, Implementation of adaboost for support vector machine classification to improve accuracy in diagnosing chronic kidney disease, *SNATIFF*, pp.35-40, 2017.

[19] M. A. Banjarsari, I. Budiman and A. Farmadi, Implementation of k-optimal on the KNN algorithm for prediction of timely graduation of students of the Faculty of Mathematics and Natural Sciences,

computer science study program based on grade point semester 4, *KLIK – Kumpul. J. Ilmu Komput.*, vol.2, no.2, pp.159-173, DOI: 10.20527/KLIK.V2I2.26, 2016.

[20] Indrayanti, D. Sugianti and M. A. Al Karomi, K parameter optimization of the k-nearest neighbor algorithm for the classification of diabetes mellitus, *SNATIF*, pp.551-554, 2017.

[21] A. Saifudin and R. S. Wahono, Implementatin of ensemble techniques to contrl of class imbalances in software defect prediction, *J. Softw. Eng.*, vol.1, no.1, 2015.

[22] D. K. Silalahi, H. Mur and Y. Satria, Comparative study of feature selection for support vector machine in the classification of credit risk assessment, *J. EduMatSains*, vol.1, no.2, pp.119-136, 2017.