Volume 6, Nomor 2, Desember 2022
Numerical: Jurnal Matematika dan Pendidikan Matematika
http://journal.iaimnumetrolampung.ac.id/index.php/numerical
DOI:https://doi.org/10.25217/numerical.v6i2.

Increasing Accuracy of Classification in C4.5 Algorithm by Applying Principle Component Analysis for Diabetes Diagnosis

Michael Sitanggang[1], Elmanani S Froilan D. Mobo imamora[2], Froilan D. Mobo[3]

[1, 2] Mathematics Departement, State of University Medan, Indonesia
[3] Department of Research, Devt and Extension Philippines Merchant Marine Academy Philippines
Correspondence: michaelsitanggang18@gmail.com

Article Info
Article History
Received: 02-11-2022
Revised: 20-11-2022
Accepted: 23-11-2022

Keywords:
Classification;
Decision Tree C4.5;
Diabetes Mellitus;
PCA; Machine Learning

Abstract

The data revolution in medical records has increased the automation of medical devices in determining the factors that cause any disease, but it also poses challenges to their analysis. According to WHO, about 6% of the world's population of more than 420 million people live with type 1 or type 2 diabetes and this number has estimated to rise beyond half a billion by 2030, which means that one of the ten adults in the future is suffering from diabetes. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, we used Decision Tree C4.5 to predict diabetes mellitus. This research used a diabetic dataset obtained from UCI machine learning repository with 419 instances and 16 attributes. In this dataset, mostly of attributes are numeric types that are continuous. This research results of the improved C4.5 algorithm by applying PCA. Many algorithms have been proposed to overcome misclassification and overfitting on classifications Decision Tree C4.5. Feature reduction is one option that is intended to eliminate irrelevant data and overcome outliers in the data so as to increase classification accuracy. Based on the results of the experiment, the application of PCA in C4.5 resulted in an increase in accuracy of 6.55% were achieved.

INTRODUCTION

Diabetes is a major health concern affecting all age groups all over the world. Diabetes Mellitus is a disease caused by hyperglycemia or high levels of glucose in the blood coupled with abnormalities in the metabolic process due to lack of insulin which is characterized by frequent drinking, frequent urination, and frequent eating [1]. It is caused mainly due to genetic factors or certain environmental conditions such as the low physical activity of lightworkers at this time, this triggers an increase in blood sugar levels and the risk of developing diabetes [2]. Now it is very important to develop predictive models using risk factors to look at patterns of diabetes development, therefore many studies have suggested machine learning models as predictors [3].

The development of databases in the health sector is growing rapidly. If this data is not utilized, it will only become a pile of useless data. Therefore, a pile of useless data can be used as a very useful data source if they are properly processed using a technique of machine learning. Among machine learning algorithms, decision trees have various advantages including being simple, easy to implement and being able to handle very large data [4].

Print ISSN: 2580-3573, Online ISSN: 2580-2437

Some algorithms used to make decision trees include ID3, ID5, C4.5 and CART. The selection of the C4.5 algorithm in this study is based on the fact that the algorithm contains a split criterion of ID3 which is called the Gain Ratio [5]. The split criteria of the selected node selection, then through the branches, move downwards from the root node until it ends at the leaf node. The C4.5 algorithm is a development of the ID3 Algorithm which can handle missing values, handle data with continuous types and perform pruning trees resulting in a simple and practical composition that affects the formation of a more accurate decision tree [6].

In this research, the diabetic diabetes dataset is taken from the UCI machine learning repository. We propose feature reduction using principal component analysis (PCA) to remove irrelevant features using the assumption of multicollinearity and overcome outliers in the data without affecting the information contained in the original data and then predictors are developed using the C4.5 algorithm to classify the data set used. The aim is to reduce some data features with PCA and then build a classification prediction model with the Decision Tree C4.5 to get the relevant features and to improve the accuracy of the Decision Tree C4.5 [7].

METHOD

1. Decision Tree C4.5

Decision Tree C4.5 is a development of the ID3 algorithm which can handle missing values and the basic concept uses two approaches to test probabilistic rankings: (1) Information gain, which minimizes the total entropy of the subset Si where there is a bias when testing numerical data. (2) Gain ratio, which is the division of information gained by the entropy information of each attribute [5]. The information gain rate is calculated by using the following formula: [8]

Information entropy:

$$Entropi\ (S) = \sum_{i=1}^{n} -p_i Log_2 p_i$$

Information gain:

$$Entropi\ (S, A) = Entropi\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropi\ (S_i)$$

Split Information:

$$Split\ Info\ (S, A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} * Log_2 \frac{|S_i|}{|S|}$$

Information gain rate:

$$GainRatio\ (S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

Then the attribute with the highest gain ratio value is chosen to be the root and the attribute with the gain ratio value lower than the root is selected to be a branch, iterate over the recalculation of the gain ratio value for each attribute until the resulting gain value = 0 for all remaining attributes [9].

Print ISSN: 2580-3573, Online ISSN: 2580-2437

The problem that is focused on the C4.5 algorithm in this study is noise data. Noise data will form branches or nodes that are not needed, causing misclassification and even over-fitting and resulting in a low level of accuracy in classification [10].

2. Principle Component Analysis

Principal Component Analysis (PCA) is a method for constructing some new attributes which are linear combinations of the original attributes [11]. PCA calculation is based on the calculation of eigenvalues and eigenvectors that represent the distribution of data from a dataset [12]. Geometrically this linear combination is a new coordinate system resulting from the iteration of the original system. The new coordinate is the direction with maximum variability and gives a simple covariance so that it can be defined as follows: [13]

$$KU = AY$$

where,

$$KU = \begin{pmatrix} KU_1 \\ KU_2 \\ \vdots \\ KU_k \end{pmatrix} = \text{principle component definition,}$$

$$A = \begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_k^T \end{pmatrix} = \begin{pmatrix} a_1 1 & a_1 2 & \cdots & a_1 p \\ a_2 1 & a_2 2 & \cdots & a_2 p \\ \vdots & \cdots & \ddots & \vdots \\ a_k 1 & a_k 2 & \cdots & a_k p \end{pmatrix} = \text{eigenvector transpose definition,}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \text{j} - \text{th attribute where j} = 1,2,3, \dots, p$$

In this research, for k principal components, it is determined using a scree plot and the cumulative variance can be defined as follows:[13]

$$\frac{\sum_{i=j}^{k} \lambda_j}{\sum_{i=j}^{p} \lambda_j} \times 100\ \%$$

The total main components formed are the same as many of the variables analyzed but the total main components used are if the percentage of total variance is considered sufficient to represent the original data variance of 75% or more [14].

3. Performance Evaluation

The model evaluation is based on the correct and incorrect object prediction results from the test data prediction results to measure the algorithm accuracy of the model built using training data [4]. In this research, the data is divided into training data and test data with each proportion 70% for training data and 30% for test data and the partitioning process is random. Measurement of the accuracy of the classification model is carried out using the confusion matrix, a method to analyze how well the classification model recognizes tuples from different classes [15].

Print ISSN: 2580-3573, Online ISSN: 2580-2437

The results of the confusion matrix are expressed in True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as follows: [4]

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positive (TP)	False Positive (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)

Figure 1. Confusion Matrix

Accuracy is computed by the equation[16].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

4. Previous Research

Feature reduction resulted by PCA and decision tree C4.5 conducted on diabetic data and the other machine learning algorithm resulted, among in: Decision Tree (accuracy 66.89%), Naive Bayes (accuracy 77.24), Extra Trees (accuracy 72.41%), k-Nearest Neighbor (accuracy 71.72%), Radial Basis Function (accuracy 69.27%), Multi-Layer Perceptron (accuracy 80.68%)[17]. The results of this research show that the decision tree algorithm provides a lower level of accuracy than other algorithms. The presence of noise in the form of a large number of outliers affects the structure of the formation of the decision tree and can trigger the formation of unnecessary nodes[18]. This problem will be the focus of this research, namely by applying PCA to be able to trim nodes even branches that are not needed.

RESULTS AND DISCUSSION

The study used a diabetic dataset from the UCI Machine Learning repository and contains 419 observations and 16 attributes. The attributes information is as follows: cholesterol, glucose, low-density lipoprotein, high-density lipoprotein, lipoprotein, age, gender, weight, height, body mass index, systolic bp, diastolic bp, waist, hip, waist-hip ratio and diabetes diagnosis. There are 15 predictor attributes, 14 of which are numeric and 1 is categorical. The, procedure in this research is described as follows:

Print ISSN: 2580-3573, Online ISSN: 2580-2437

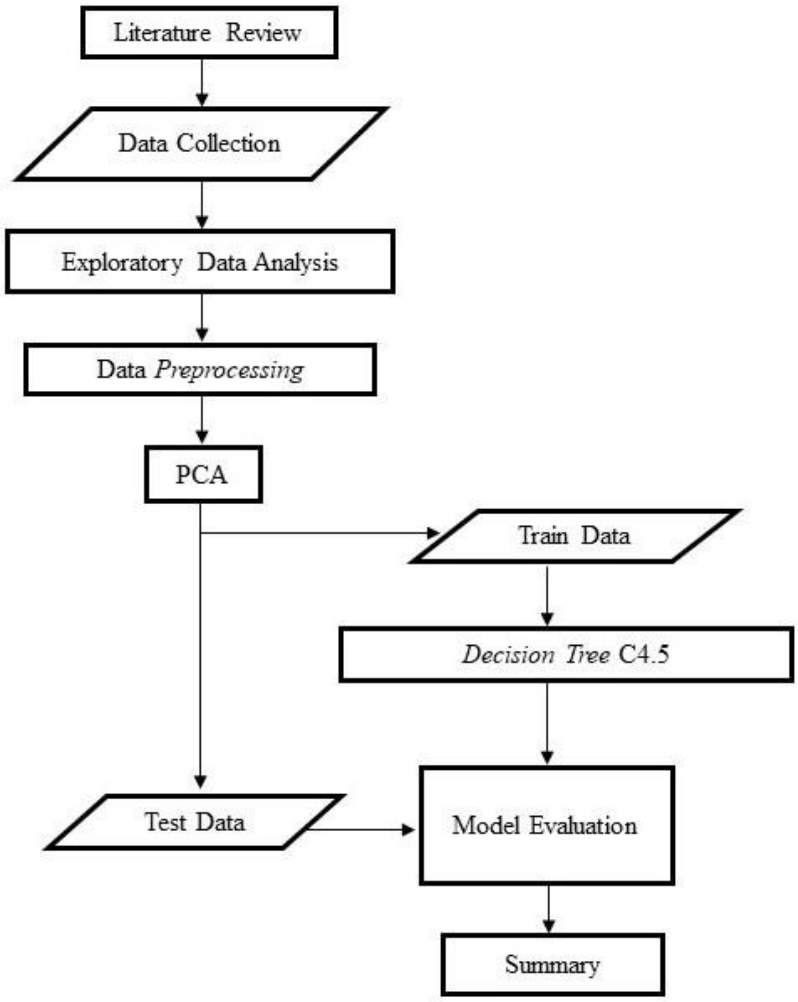

Figure 2. Flowchart of Methodology

Figure 2 is a diagram or Flow chart of Methodology where before doing the classification using the C4.5 algorithm, data must be prepared in advance to minimize errors then approach to analyze the data using visual techniques to check assumptions and patterns with the help of statistical summary[19]. This is useful for optimizing the results of the classifier. The suggested framework is implemented using python programming.

There are some missing values in the lipoprotein attribute of 3%. Therefore, this missing value cannot be ignored and imputation is carried out using the median because in previous studies it was proven that the median value is more robust than the mean value [15]. Based on data after imputation, the analysis of the characteristics of each attribute formed is as follows:

Print ISSN: 2580-3573, Online ISSN: 2580-2437

Table 1. Data Description

Features	Mean	Median	Min	Max	Percentage Outlier
cholesterol	208.96	202	98	331	6.68%
glucose	174.34	168	60	300	4.06%
low-density lp	102.31	103	49	133	5.49%
high-density lp	69.59	72	0	122	5.01%
lipoprotein	1.44	1.40	0.47	3.83	3.58%
Age	37.32	33	21	76	3.58%
weight	93.22	85.14	0	197.28	4.77%
height	174.55	172	105	246	6.44%
body mass indeks	29.96	30.70	0	38.7	1.67%
systolic bp	144.23	136	90	254	7.88%
diastolic bp	91.22	86	60	195	6.44%
waist	41.68	39	27	70	6.92%
hip	44.49	43	30	68	5.73%
waist-hip ratio	0.93	0.90	0.68	1.75	5.97%
gender	Male = 182			Female = 237	
diabetes diagnosis	Yes = 186			No = 233	

Table 1 presents many attributes that are close to a normal distribution (mean close to median), validates the minimum and maximum values of each variable and the outlier data spread out. Outliers in statistics are usually removed or outliers are done with z-score, robust transformation, and log transformation because in general outliers greatly affect the accuracy of the model[20]. Another assumption before applying PCA is to examine the symptoms of multicollinearity in data and the results are as follows: [12]

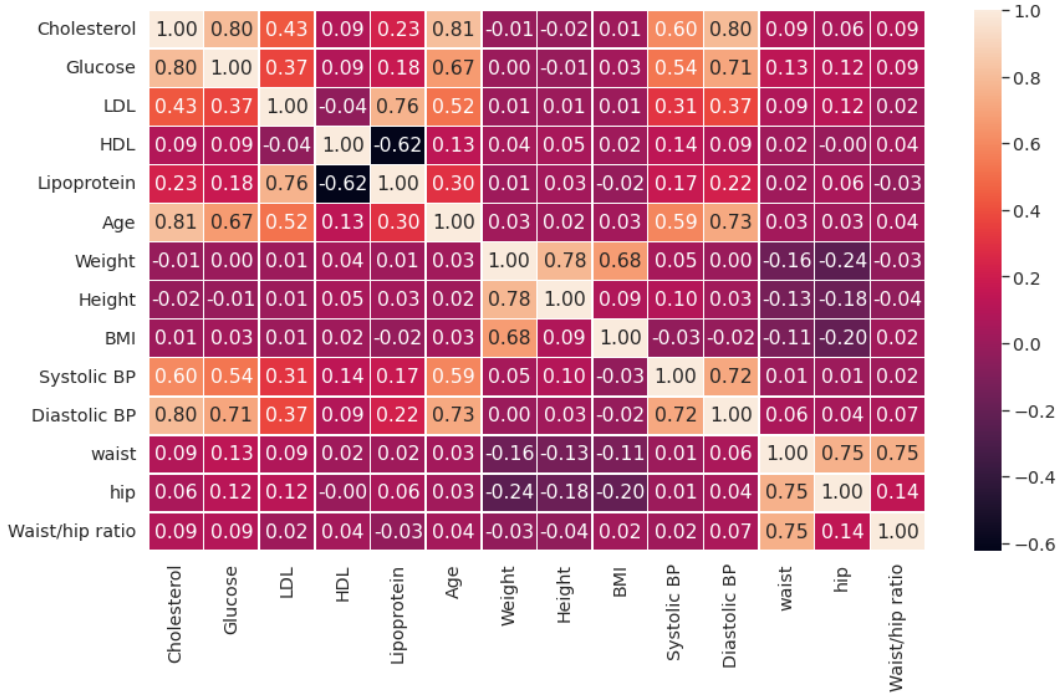

Figure 3. Heatmap Multicollinearity Predictor

Print ISSN: 2580-3573, Online ISSN: 2580-2437

From the results of the assumption test, it was found that the high-density lipoprotein variable did not meet the multicollinearity assumption, then the variable was deleted and the final data has 14 predictor attributes. In this research, attention is focused on outliers on the predictor attribute where the outliers spread in each predictor attribute are not removed but play a role in the formation of the main component attribute. To select relevant attributes, several principal components (PC1 to PC7) are performed, which contribute variances of 90.5% and absolute values of eigenvectors above 0.5. The PCA has reduced 15 attributes to 7 relevant attributes. The results of the calculation of the cumulative proportion of variance of all the principal components can be seen as follows:

Table 2. Cumulative Proportion of Variance

Component (K)	Eigen Value	Variance	Cumulative Proportion
1	4.255	31.5%	31.5%
2	2.488	18.4%	49.9%
3	1.761	13.0%	62.9%
4	1.424	10.5%	73.4%
5	0.960	7.1%	80.5%
6	0.858	6.3%	86.8%
7	0.499	3.7%	90.5%
8	0.406	3.0%	93.5%
9	0.315	2.3%	95.8%
10	0.230	1.7%	97.5%
11	0.136	1.0%	98.5%
12	0.182	1.3%	99.8%
13	0.003	0.1%	99.9%
14	0.000	0.0%	99.9%
15	0.000	0.0%	99.9%

Another technique involving the eigenvalues is using a screeplot by looking at the point before the curve decreases sharply or slopes [12]. In the scree plot, the vertical side defines the eigenvalues, while the horizontal side defines the number of principal components to be selected and the results are as follows:

Print ISSN: 2580-3573, Online ISSN: 2580-2437

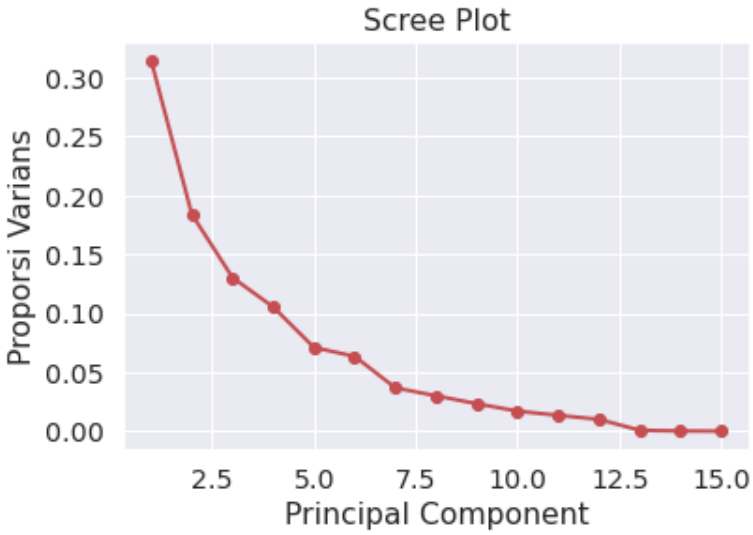

Figure 4. Scree plot Principal Component

In this paper, the diagnosis results are used for leaf nodes. When forming the C4.5 algorithm before and after conducting the PCA, there are differences in the pattern of the two results of the C4.5 algorithm where it is clear that the difference in the number of nodes and branches is that the C4.5 algorithm formed using component data has a simple shape with 9 branches and 105 vertices, where if only decision tree modelling has 10 branches and 89 vertices. After evaluating the model, it can be seen that there is an increase of 6.55% in the C4.5 algorithm model after conducting PCA. The comparison performance of both C4.5 before PCA and C4.5 after PCA is shown in Table 3.

Table 3. The Performance Metrics and Treeplot of the Model

	Data non-PCA	**Data PCA**	**Difference**
Accuracy	**64.29%**	**70.84%**	**6.55%**
Node	**105**	**89**	**16**

Based on the test results contained in Table 3, an increase in accuracy occurs in Decision Tree C4.5 based on principal component data and it can be seen that there are original variables that have a significant weight on the formation of the main component attributes based on the calculation of attribute weights using eigenvectors and the results are as follows:

Print ISSN: 2580-3573, Online ISSN: 2580-2437

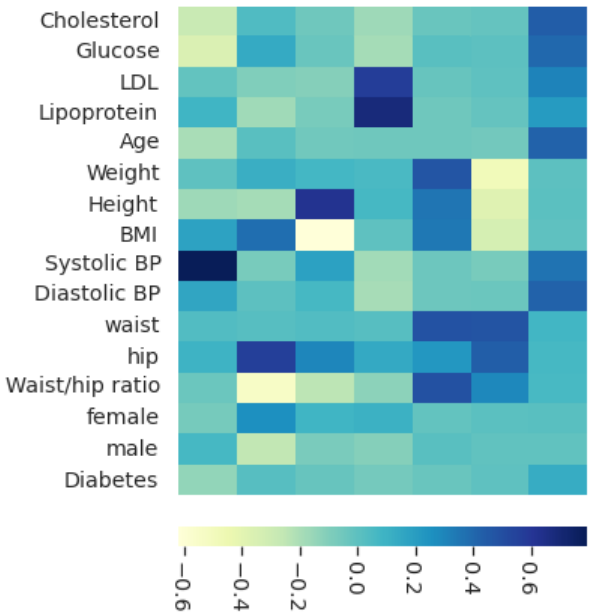

Figure 5. Weight of Principal Component in Variable

If the top 5 sequences are taken from each of the main component attributes, it can be seen that the variables that have a greater weight in the formation of the main component attributes include hip, gender by women, weight, glucose, systolic blood pressure, and body mass index in more detail affects the diagnosis of diabetes mellitus.

CONCLUSIONS

In summary, we have compared two prediction models for predicting diabetes mellitus using 15 predictors and 1 class target. One is before applying PCA to the dataset and the other applies PCA to the dataset. Here the studies conclude that the decision tree C4.5 classifier after applying PCA achieves higher accuracy by 6.55 % and simpler node than before applying the PCA. In another result that is after applying PCA to the dataset, the C4.5 algorithm has a simple tree plot shape with 9 branches and 105 vertices, while C4.5 without PCA modelling has 10 branches and 89 vertices and the C4.5 algorithm after applying PCA have accurate result better than before apply PCA. In this case, we can come to know that PCA can improve the C4.5 algorithm without removing the noisy data from our dataset it will provide good results for our problems. This study can be used to select the best classifier for predicting diabetes. In future, we can use this type of study for any other diseases with their suitable datasets. For the next research, the Decision Tree C4.5 model will be more optimal if the number of data and variables are added to the medical record data for diagnosing diabetes mellitus such as things that support the existence of diabetes mellitus, including hypertension, physical condition, hereditary history and others.

Print ISSN: 2580-3573, Online ISSN: 2580-2437

REFERENCES

[1] Hestiana, D. W. (2017). Faktor-faktor yang berhubungan dengan kepatuhan dalam pengelolaan diet pada pasien rawat jalan diabetes mellitus tipe 2 di Kota Semarang. *JHE (Journal of Health Education)*, *2*(2), 137–145.

[2] Nasution, F., Andilala, A., & Siregar, A. A. (2021). Faktor Risiko Kejadian Diabetes Mellitus. *Jurnal Ilmu Kesehatan*, *9*(2), 94–102.

[3] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, 45–51.

[4] Witten, I., Frank, E., & Hall, M. (2011). *Data Mining. Practical Machine Learning Tools and Techniques, ISBN 978-0123748560*.

[5] Santhosh, K. (2013). Modified C4. 5 algorithm with improved information entropy. *International Journal of Engineering Research & Technology*, *2*(14), 485–512.

[6] Rim, P., & Liu, E. (2020). Optimizing the C4. 5 Decision Tree Algorithm using MSD-Splitting. *International Journal of Advanced Computer Science and Applications*, *11*(10), 41–47.

[7] Liu, J., Ning, B., & Shi, D. (2019). *Application of improved decision tree c4. 5 algorithms in the judgment of diabetes diagnostic effectiveness*. *1237*(2), 022116.

[8] Muslim, M. A., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2017). Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4. 5 dalam Mendiagnosa Diabetes. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, 135–143.

[9] I. S. Damanik, A. P. Windarto, A. Wanto, S. R. Andani, and W. Saputra, "Decision tree optimization in C4. 5 algorithm using genetic algorithm," 2019, vol. 1255, no. 1, p. 012012.

[10] R. S. Wahono, "Penerapan Naive Bayes untuk Mengurangi Data Noise pada Klasifikasi Multi Kelas dengan Decision Tree," *Journal of Intelligent Systems*, vol. 1, no. 2, pp. 136–142, 2015.

[11] S. T. Ikram and A. K. Cherukuri, "Improving accuracy of intrusion detection model using PCA and optimized SVM," *Journal of computing and information technology*, vol. 24, no. 2, pp. 133–148, 2016.

[12] Muhammad, M. U., Jiadong, R., Muhammad, N. S., Hussain, M., & Muhammad, I. (2019). Principal component analysis of categorized polytomous variable-based classification of diabetes and other chronic diseases. *International Journal of Environmental Research and Public Health*, *16*(19), 3593.

[13] Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

[14] Tamonob, A. M., Saefuddin, A., & Wigena, A. H. (2015). *Nonlinear Principal Component Analysis and Principal Component Analysis with Successive Interval in K-Means Cluster Analysis*. *20*(2).

[15] Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. *Morgan Kaufmann*, *340*, 94104–3205.

[16] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.

Print ISSN: 2580-3573, Online ISSN: 2580-2437

[17] Theerthagiri, P., & Vidya, J. (2021). *Diagnosis and Classification of the Diabetes Using Machine Learning Algorithms*.

[18] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28.

[19] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[20] H. J. Escalante, "A comparison of outlier detection algorithms for machine learning," 2005, pp. 228–237.

Print ISSN: 2580-3573, Online ISSN: 2580-2437

Print ISSN: 2580-3573, Online ISSN: 2580-2437