ISSN 2347 - 3983
Volume 8. No. 2, February 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter44822020.pdf
https://doi.org/10.30534/ijeter/2020/44822020

The Implementation of Subspace Outlier Detection in K-Nearest Neighbors to Improve Accuracy in Bank Marketing Data

Dimas Aryo Anggoro[1], Putera Islamiyadi Rahmatullah[2]
[1]Informatics Department, Universitas Muhammadiyah Surakarta, Indonesia, dimas.a.anggoro@ums.ac.id
[2]Informatics Department, Universitas Muhammadiyah Surakarta, Indonesia, L200160172@student.ums.ac.id

ABSTRACT

Banks are the most important institution in the economics and social sector. Hence, it provides services or products such as term deposits. Term deposits give banks the most reliable source of profit and credits. Bank uses direct marketing to attract customers to subscribe to the deposit program. By using data mining, a bank can predict the client who will subscribe to the term deposit program. In this research, the dataset to be tested is bank marketing by Portuguese banking institutions. The data mining algorithm used is K-Nearest Neighbors (KNN), but it has a sensitive weakness to outliers which causes the accuracy of this algorithm is not good. This research aims to cover up the weakness of the KNN by detecting outliers in the dataset and improving the accuracy of the KNN. The outlier detection method used in this research is Subspace Outlier Detection (SOD) and Principal Component Analysis (PCA) as dimensional reduction method. The dataset was splitting into 70% training data and 30% testing data as new data. K-fold cross-validation was used to search the K Neighbors value of KNN during the modeling training data. The classification results of testing data based on the confusion matrix show the accuracy of the KNN algorithm which only uses PCA is 91%. While the KNN algorithm accuracy with SOD and PCA is 94%. It can be concluded that the accuracy of KNN using SOD is better than accuracy of KNN without SOD by improving the accuracy by 3% in the bank marketing dataset.

Key words : Direct Marketing, KNN, PCA, SOD

1. INTRODUCTION

A bank is a financial institution that is licensed to raise funds from the public in the form of deposits that can be redirected to the community in the form of credit or other funds [1]. Banks are the most important economic and social aspects, therefore, they are required to provide a variety of banking products or services to customers. One of the services offered by a bank is the deposit. There are different types of deposit accounts, such as demand deposit, savings, term deposit, and money market deposit [2]. Year to year, the term deposit interest rates received by customers are higher when compared to the demand deposit and regular savings rates [3]. Therefore, a term deposit account proffers the most stable and reliable source of profit and credit in the bank sector.

The bankers take advantage of the benefits of the deposit by promoting their deposit program to customers through marketing campaigns. There are two types of marketing campaigns, namely mass marketing and direct marketing [4]. The bank selects the direct marketing strategy since it is more effective compared to mass marketing. One of the banking institutions that uses direct marketing is the Portuguese banking institution. Marketing strategy carried out by this institution utilizes telephone communication. The implementation of direct marketing from time to time generates data and information in the form of reports that need to be analyzed [5]. One of the most effective ways to analyze previous campaign reports is through business intelligence and data mining techniques [5]. Data mining is one such of tool class that is used to recognize and evaluate the hidden pattern of data [6]. In addition, by using data mining in direct marketing, a bank can find out patterns of behavior and choose the right customers for promotion [4]. So a bank can determine which customers will subscribe to the term deposit program.

The classification algorithm used in data mining is K-Nearest Neighbors (KNN), which is a method of classification of objects based on the nearest data in the feature space training [7]. The advantages of this algorithm include fast training, simple and easy to learn, and effective in training big data [8]. However, KNN can be affected by outliers, as a result, its accuracy is not good enough [9]. Another method used in KNN is Subspace Outlier Detection (SOD), which is an outlier detection method that models outliers on high-dimensional data. A study conducted by [10] found that the SOD method is more stable at optimal values to face high-dimensional data than other outlier detection methods such as Angle-Based Outlier Detection (ABOD) and Local Outlier Factor (LOF), which perform better in low-dimensional data. In this current research, the researchers applied the SOD outlier detection method with the KNN classification algorithm to detect outliers in the bank marketing data and provided treatment to the outliers.

Although stored outliers bring out important information in the dataset [11].

This research seeks to detect outliers in bank marketing data using the SOD method and classify it by using the KNN algorithm. It is expected that the results of this research can be used as a reference and consideration for future outlier detection research and ease the bankers in predicting clients who subscribe to deposit programs by applying data mining techniques. Outlier detection using the SOD is also expected to cover the shortcomings of KNN that are sensitive to outliers. The output of this research is the KNN algorithm with SOD which is expected to have better accuracy

2. METHOD

2.1 Data Collection

The dataset used in this research is the Bank Marketing dataset taken from Kaggle. This data relates to direct marketing activities of the Portuguese banking institution from 2008 to 2010 where the agency performed telephone contacts to carry out marketing campaigns for potential customers to subscribe to term deposits. The data consists of 20 input attributes and 1 output attribute. Input attributes used are age, job, marital, education, default, housing, loan, contact, month, day_of_week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, and nr.employed. The output attribute which provides information on the number of clients will subscribe to the term deposit program is indicated by the label y. The number of instances in this dataset is 41188. The detailed description of the attributes used is shown in table 1.

Table 1: Attributes and description of bank marketing data

Attributes	Type of Data	Description of Data
Age	Numeric	age of customers
Job	Categorical	type of customers' occupation
Marital	Categorical	status of customers' marriage
Education	Categorical	levels of customers' education
Default	Categorical	The customers are on default credit or not
Housing	Categorical	The customer has a loan for the property or not
Loan	Categorical	The customers' loans are personal, or not
Contact	Categorical	Type of communication on contact to customers
Month	Categorical	Last months of contact
Day_of_week	Categorical	Last days of contact
Duration	Numeric	Duration of last contact
Campaign	Numeric	The number of contacts that were made during the campaign
Pdays	Numeric	The number of days that have passed since the last customers were contacted from the previous campaign
Previous	Numeric	The number of contacts made during and for this Customer
Poutcome	Categorical	Results of earlier marketing campaign
Emp.var.rate	Numeric	Level of Work Variation
Cons.price.idx	Numeric	Price Index for Consumers
Cons.conf.idx	Numeric	Confidence index for customers
Euribor3m	Numeric	Euribor 3-months rate
Nr.employed	Numeric	The number of people employed
Y	Categorical	The customers subscribed to term deposits or not

2.2 Data Preprocessing

A. Subspace Outlier Detection

Subspace Outlier Detection (SOD) is an outlier detection method that detects and models outliers in high-dimensional dataset. Outliers in the dataset can be found in the subspace in the original data. According to the research by [10], the SOD is modeled by calculating the subspace defining vector $v_i^{R(p)}$ whose results are categorized into high and low (1 and 0), depending on the $VAR^{R(p)}$, variance at the reference point. The subspace defining vector and variance formula can be seen in equations 1 and 2.

$$VAR^{R(p)} = \frac{\sum_{p \in R(p)} dist(p, \mu^{R(p)})^2}{Card(R(p))} \quad (1)$$

$$v_i^{R(p)} = \begin{cases} 1 & if \ var_i^{R(p)} < \alpha \frac{VAR^{R(p)}}{d} \\ 0 & else \end{cases} \quad (2)$$

Subspace hyperplane H(R(p)) is defined as the average value of the reference set $\mu^{R(p)}$ and the subspace defining vector of the reference set $v^{R(p)}$. The hyperplane subspace equation can be seen in equation 3.

$$H(R(p)) = (\mu^{R(p)}, v^{R(p)}) \quad (3)$$

Next, calculate the number of points p deviated from the hyperplane subspace and the reference set *R*(p) using the weight euclidean distance between points p and $\mu^{R(p)}$ with the subspace defining vector as a weight vector. The calculation of distance point p to hyperplane can be seen in equation 4.

$$dist\left(p, H(R(p))\right) = \sqrt{\sum_{i=1}^{d} v_i^{R(p)} \cdot (p_i, \mu_i^{R(p)})^2} \quad (4)$$

If the result of weight euclidean distance approaches 0 where the distance is close to the hyperspace subspace then that point is not categorized as an outlier and vice versa. If the result of the weight euclidean distance is proximate to 1 then that point is categorized as an outlier. From this explanation, the

Subspace Outlier Degree is defined in equation 5.

$$SOD_{R(p)}(p) = \frac{dist(p,H(R(p)))}{||v^{R(p)}||} \quad (5)$$

The SOD which is the distance between point p and hyperplane subspace reference set R(p) is normalized with the number of relevant dimensions $v^{R(p)}$ (equation 5). The results of the SOD in the form of an outlier score is compared to a certain threshold to determine the categorization of the points.

B. Principal Component Analysis

The Principal Component Analysis (PCA) is a method for reducing a dataset dimension without reducing the meaning of the data [12]. PCA extracts dataset features using eigenvector and eigenvalue that have been obtained [13]. The first step in dimensional reduction using PCA is to input X for PCA where X is a training data consisting of n-vectors with data dimensions (m). Next, calculate the average of each dimension using equation 6.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (6)$$

Where n is the number of sample data and X_i is the observation data. The next step is to calculate the covariance matrix (Cx) using equation 7. $\bar{X}$ is the mean data.

$$C_X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \quad (7)$$

The calculation of the eigenvector and eigenvalue from the covariance matrix can be done with equation 8. The eigenvalue is a form of scalar numbers and eigenvector is a matrix of calculations from eigenvalue and a set of initial data [14].

$$C_X v_m = \lambda_m v_m \quad (8)$$

The eigenvalue scores that have been obtained are sorted from the largest to the smallest. Principal Component is a collection of eigenvectors obtained from eigenvalue that has been sorted. The PC dimension is reduced based on the eigenvalue. One way to reduce PC dimensions based on eigenvalues is by using the accumulation of variance values on eigenvectors [13]. Accumulated variance values are taken based on predetermined thresholds. The threshold value is very influential on the eigenvector chosen during data transformation.

C. Data Splitting

Before advancing to the next stage, splitting data must be done. Splitting is done by dividing the dataset into training data and testing data. Training data is employed to form a model of the classification algorithm used. Testing data is useful for validation of the models built [15]. In this research, the dataset was split into 70% of the training data and 30% of the random testing data. 30 percent of the data used as test data were used as new data for prediction purposes.

2.3 Data Processing

A. K-Nearest Neighbors (KNN)

The KNN is a supervised algorithm applied to classify a set of objects or data based on the nearest neighbors whose class is known [16]. In the KNN there are three important elements used for classification, including K values, labeled objects, and distances between objects [17]. The K value in the KNN is used to determine the number of the nearest neighbor. K value is determined based on the optimum K value during the training [18]. In this research, the optimum K value was determined by using K-fold cross-validation. K neighbors value should ≠ 1 and odd [19]. The labeled object was used to create a training data classification model. The calculation of the distance between two objects was calculated based on Euclidean distance. Euclidean distance was used in KNN since it generated the closest distance to each point. The Euclidean distance formula is shown in equation 9 [8].

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(x_{pi} - x_{qi})^2} \quad (9)$$

where:

dist(p,q): the *euclidean distance* from coordinate of p to q

x_{pi}: coordinate of p in iteration i

x_{qi}: coordinate of q in iteration i

n: the number of iteration

From the three aspects above, the steps of KNN are described as follows. First, defining the K value, the optimum K value can be determined by finding the accuracy value in the training data using K-fold cross-validation. Second, calculating the euclidean distance between the testing data and the training data using the formula in equation 10. Third, sorting the results of the euclidean distance calculation from the test data group in ascending order. Fourth, taking k-nearest neighbors from the results that have been sorted for each test data. The test data class was taken from the majority vote on the k-nearest neighbor. After the dataset was modeled by using the KNN algorithm, it was evaluated by using cross-validation and accuracy results.

2.4 Evaluation Model

A. Cross-Validation

Cross-validation is a statistical method which helps to evaluate the performance of algorithm models and the performance of prediction models of unknown datasets [20]. One of the forms of cross-validation is k-fold cross-validation [21]. The k value used for k-fold cross-validation is 10, 10-fold cross-validation. 10-fold cross-validation means that the records in the training data are split into 9 subsets into a training set and 1 subset into a testing set.

B. Confusion Matrix

Confusion Matrix is a method used to evaluate models for classifying true and false objects [22]. Accuracy is used to get percentage of instances which the classification algorithm correctly classifies. Accuracy is defined as the ratio between correct data prediction (TP and TN) and overall data [23]. The accuracy formula is shown in equation 10 [24].

$$accuracy = \frac{TP+TN}{TP+FN+FP+FN} \quad (10)$$

Where TP is true positive; TN is true negative; FP is false positive; FN is false negative.

3. RESULTS AND DISCUSSION

3.1 Data Preprocessing

The dataset used in this research was bank marketing data with 41188 instances and 21 data variables consisting of 20

input variables and 1 target variable. There was a missing value in this dataset, so that, treatment was needed in the form of missing value treatment by replacing the missing value with the mode value for each feature. The data used was multivariate data which required the conversion of categorical data into numerical data by using a label encoder. The label encoder transforms categorical data into numerical data starting from number 0 to the number of data units for each feature.

In this current research, two experiments were carried out on the dataset by implementing Subspace Outlier Detection (SOD) and Principal Component Analysis (PCA). SOD was used to detect outliers in the dataset where outliers could affect the accuracy of the classification model. PCA was used to reduce the dataset features without reducing the meaning of the removed dataset so that it accelerated the performance of the classification algorithm.

A. The Result of Subspace Outlier Detection

The SOD is implemented to detect outliers in the subspace dataset by using Pyod tools. This research detects as much as 15% of observations that are categorized as outliers and searches for the threshold to define the boundary datapoints detected as outliers or not. The results of the SOD are determined by the outlier score on each datapoint. Threshold settings are obtained from the 100th percentile multiplied by outlier contamination from the sample data. The threshold obtained in this experiment is -0.7334582830883063, if the outlier score is above the threshold then it is classified as inliers. Conversely, if the outlier score is below the threshold then it is classified as outliers. Outlier classification results by using SOD, there were 6179 datapoints as outliers and 35009 datapoint as inliers. A total of 6179 outliers from the dataset were removed to create a new dataset with 35009 instances, clean datasets. Comparison of targets on original and clean datasets shown in table 2.

Table 2: Targets on original dataset and clean dataset

Target	Original Dataset	Clean Dataset
No (0)	36548	32802
Yes (1)	4640	2207

According to table 2, it is obvious that the number of targets “no” and “yes” in the original data is 36548 and 4640 instances. After eliminating the outliers, the number of targets “no” and “yes” becomes 32802 and 2207 instances. The two datasets are then reduced for its dimensions by using PCA.

B. The Results of Principal Component Analysis

The original dataset and the clean dataset (the dataset whose outlier has been removed) were processed using PCA. PCA reduces dimensions based on the accumulation of the best variance in the dataset. After the PCA was performed, an accumulation of PCA variance was obtained from each dataset. The number of eigenvectors or components in a PCA depends on the threshold used. The threshold specified in the accumulation of variance is at least 80% [13]. The followings are the results of the accumulation of PCA variance in the original dataset and the clean dataset shown in table 3 and table 4.

Table 3: PCA variance results in the original dataset

Principal Component	Variance	Variance Accumulation
PC1	6.26798747e-01	0.6267987473199792
PC2	**3.30839243e-01**	**0.95763799001592**
PC3	4.08477894e-02	0.9984857794166259
PC4	1.01380124e-03	0.9994995806597874
PC5	1.92031393e-04	0.9996916120525972
PC6	1.17398886e-04	0.9998090109386407
PC7	6.96647226e-05	0.9998786756612444
PC8	4.92985082e-05	0.9999279741694708
PC9	3.64697368e-05	0.9999644439062626
PC10	1.80192123e-05	0.9999824631185513
PC11	7.31553499e-06	0.9999897786535412
PC12	2.81491382e-06	0.9999925935673606
PC13	2.29699535e-06	0.9999948905627138
PC14	1.92586134e-06	0.9999968164240556
PC15	1.34325741e-06	0.9999981596814649
PC16	1.19201638e-06	0.9999993516978456
PC17	2.48691218e-07	0.9999996003890638
PC18	2.17420875e-07	0.9999998178099384
PC19	1.81513212e-07	0.99999999932315
PC20	6.76850027e-10	1.0

Based on table 3, the results of the PCA on the original dataset shows the accumulation of variance on PC1 is 0.61. The results of the accumulation of PC1 and PC2 generate the accumulated variance value of 0.957. From these results, it is possible to reduce the dimensions to 2 features because the best accumulation of variance is found on PC1 and PC2.

Table 4: PCA variance results on clean dataset

Principal Component	Variance	Variance Accumulation
PC1	7.50082833e-01	0.7500828325220159
PC2	**1.71301890e-01**	**0.921384722053264**
PC3	7.55541912e-02	0.9969389132833375
PC4	1.99680881e-03	0.9989357220979856
PC5	3.72798743e-04	0.9993085208410862
PC6	2.74496209e-04	0.999583017049626
PC7	1.53589396e-04	0.9997366064451902
PC8	1.08694409e-04	0.9998453008537498
PC9	8.11770662e-05	0.9999264779199575
PC10	4.07974896e-05	0.9999672754095454
PC11	1.11132665e-05	0.99997838867606
PC12	6.53394511e-06	0.9999849226211727
PC13	5.33033297e-06	0.9999902529541431
PC14	3.23034771e-06	0.9999934833018496
PC15	2.76875572e-06	0.9999962520575669
PC16	2.69847052e-06	0.9999989505280849

PC17	5.04202874e-07	0.9999994547309584
PC18	4.50470067e-07	0.9999999052010259
PC19	9.29438979e-08	0.9999999981449238
PC20	1.85507628e-09	1.0

According to the results in table 4 presents the accumulated variance value of each principal component in the clean dataset. The result of the accumulation of variance on PC1 is 0.750. The result is smaller than the accumulated variance value on PC1 and PC2, which is 0.921. The results of the accumulation of PC1 and PC2 can be determined by reducing the dimensions into 2 features. Before continuing to the data processing stage, the two datasets that have been reduced in dimension are split by dividing the dataset into 70% training data and 30% testing data. The splitting is done randomly.

3.2 Data Processing

The data is modeled using K-Nearest Neighbors (KNN) after going through the preprocessing phase. The first thing to do is to determine the K neighbors value by using k-fold cross-validation. The k value in the k-fold used is 10, 10-fold cross-validation. The K neighbors value used is $1 \leq K \leq 25$, with the condition that the K neighbors value $\neq 1$ and is odd. Calculation of the avarage accuracy using 10-fold cross-validation are shown in table 5.

Table 5: Average accuracy results for each K neighbors

K-neighbors	Average Accuracy	
	Original Dataset	**Clean Dataset**
1	88.20	92.05
3	90.01	93.52
5	90.43	93.85
7	90.67	94.06
9	90.71	93.96
11	90.87	94.04
13	90.80	94.22
15	90.89	94.25
17	90.90	94.23
19	90.97	94.25
21	91.03	94.30
23	91.04	**94.33**
25	**91.06**	94.32

According to table 5, the best K neighbors value for the original dataset is K=25 with an average accuracy of 91.06%. While the best K neighbors value for the clean dataset is K=23 with an average accuracy of 94.33%. The results of K values that have been obtained from each dataset can be used for the KNN training model and a comparison of predictive model evaluations performed on 30% testing data between KNN models with the SOD and PCA and KNN and PCA models.

3.3 Model Evaluation

At the model evaluation phase, the researchers checked the testing data from the two datasets whose labels were removed and then predictions were made by using KNN. Evaluation is done by using the confusion matrix method followed by accuracy calculation. At this stage, the researchers compared the KNN prediction accuracy by using SOD and KNN. The results of the accuracy comparison are shown in table 6.

Table 6: Accuracy results on the KNN

Dataset	K Neighbors	Accuracy
Original dataset (PCA + KNN)	25	91%
Clean dataset (SOD + PCA + KNN)	23	94%

According to the result on table 6, highest accuracy results obtained on the clean dataset using the KNN with the K neighbors value = 23 and SOD of 94%, while the accuracy of the original dataset that only uses the KNN with the K neighbors value = 25 is 91%. It can be concluded that the SOD method combined with KNN and PCA as dimensional reduction has the highest accuracy and SOD can increase the accuracy of KNN in the bank marketing dataset.

4. CONCLUSION

Based on the results of the research, the KNN can be used to predict marketing bank clients using previous client data. From this current research, it is concluded that predictions using KNN and SOD generate better accuracy than predictions using only KNN and PCA. The SOD can increase the accuracy of KNN and PCA as dimensional reduction with an accuracy rate of 94% while the prediction models which only use KNN and PCA have an accuracy rate of 91%. The SOD can increase accuracy by 3% by detecting outliers by 15% from observation. The detected outlier is removed from the dataset.

For further research, other researchers can use other datasets or other methods of outlier detection to improve KNN accuracy. In addition to outlier detection, an algorithm or other methods such as dimensional reduction and normalization of data can be combined. Further researchers can also add time computation analysis so that the resulting computational time is less than what was obtained in this research. The tuning parameters can also be applied to the SOD outlier detection method.

REFERENCES

[1] Margaretha, F., and Letty. **Faktor-faktor yang memengaruhi kinerja keuangan perbankan indonesia**, *Manajemen Keuangan*, vol. 6, no. 2, pp. 84-96, 2017.

[2] Islam, M. A., and Ghosh, P. **A comparative analysis of deposit products in banking industry: an opportunity for eastern bank Ltd**, *Journal of Investment and Management*, vol. 3, no. 1, pp. 7-20, 2014.

[3] Prabowo, A. D. R., and Muljono, M. **Prediksi Nasabah Yang Berpotensi Membuka Simpanan Deposito Menggunakan Naive Bayes Berbasis Particle Swarm**

Optimization, *Techno. Com*, vol. 17, no. 2, pp. 208-219, 2018.

[4] Ling, C. X., and Li, C. **Data mining for direct marketing: Problems and solutions**. In *Kdd*, August 1998, vol. 98, pp. 73-79.

[5] Ispandi, I., and Wahono, R. S. **Penerapan Algoritma Genetika untuk Optimasi Parameter pada Support Vector Machine untuk Meningkatkan Prediksi Pemasaran Langsung**, *Journal of Intelligent Systems*, vol. 1, no. 2, pp. 115-119, 2015.

[6] Mallikarjuna, M. and Rao, R. P. **Application of Data Mining Techniques to Classify World Stock Markets**, *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 1, pp. 46-53, January 2020. https://doi.org/10.30534/ijeter/2020/09812020

[7] Imandoust, S. B., and Bolandraftar, M. **Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background**, *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605-610, 2013.

[8] Mutrofin, S., Izzah, A., Kurniawardhani, A., and Masrur, M. **Optimasi teknik klasifikasi modified k nearest neighbor menggunakan algoritma genetika**, *Jurnal Gamma*, vol. 10, no. 1, 2015.

[9] Bhattacharya, G., Ghosh, K., and Chowdhury, A. S. **kNN classification with an outlier informative distance measure**. In *International Conference on Pattern Recognition and Machine Intelligence*, December 2017, pp. 21-27. https://doi.org/10.1007/978-3-319-69900-4_3

[10] Kriegel, H. P., Kröger, P., Schubert, E., and Zimek, A. **Outlier detection in axis-parallel subspaces of high dimensional data**. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Berlin, Heidelberg, April 2009, pp. 831-838.

[11] Putri, D. A. P., and Sudarmilah, E. **Comparative Study for Outlier Detection In Air Quality Data Set**, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 11, pp. 584-592, November 2019. https://doi.org/10.30534/ijeter/2019/297112019

[12] Noertjahjani, S., Susanto, A., Hidayat, R., and Wibowo, S. **Ictal epilepsy and normal eeg feature extraction based on PCA, KNN and SVM classification**, *Journal of Theoretical and Applied Information Technology*, vol. 83, no. 1, pp. 100, 2016.

[13] Adiwijaya, W. U., Lisnawati, E., Aditsania, A., and Kusumo, D. S. **Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification**, *Journal of Computer Science*, vol. 14, no. 10, 2018.

[14] Yudistira, R., Meha, A. I., and Prasetyo, S. Y. J. **Perubahan Konversi Lahan Menggunakan NDVI, EVI, SAVI dan PCA pada Citra Landsat 8 (Studi Kasus: Kota Salatiga)**, *Indonesian Journal of Computing and Modeling*, vol. 2, no. 1, pp. 25-30, 2019.

[15] Anggoro, D. A., and Supriyanti, W. **Improving Accuracy by Aplplying Z-Score Normalization in Linear Regression and Polynomial Regression Model for Real Estate Data**, *International Journal of Emerging Trends in Engineering Research*. vol. 7, no. 11, pp. 549-555, November 2019. https://doi.org/10.30534/ijeter/2019/247112019

[16] Amra, I. A. A., and Maghari, A. Y., 2017, May. **Students performance prediction using KNN and Naïve Bayesian**. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.

[17] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... and Zhou, Z. H. **Top 10 algorithms in data mining**, *Knowledge and information systems*, vol. 14, no. 1, pp. 1-37, 2008.

[18] Saputra, A. Y., and Primadasa, Y. **Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor**, *Techno. Com*, vol. 17, no. 4, pp. 395-403, 2018.

[19] Rivki, M., and Bachtiar, A. M. **Implementasi Algoritma K-Nearest Neighbor dalam Pengklasifikasian Follower Twitter yang Menggunakan Bahasa Indonesia**, *Jurnal Sistem Informasi*, vol. 13, no. 1, pp. 31-37, 2017.

[20] Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., and Shahmoradi, L. **Accuracy improvement for diabetes disease classification: a case on a public medical dataset**, *Fuzzy Information and Engineering*, vol. 9, no. 3, pp. 345-357, 2017.

[21] Lutfi, M., and Hasyim, M. **Penanganan Data Missing Value pada Kualitas Produksi Jagung dengan Menggunakan Metode K-NN Imputation pada Algoritma C4. 5,** *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, vol. 2, no. 2, pp. 89-104, 2019. https://doi.org/10.31598/jurnalresistor.v2i2.427

[22] Gazalba, I., and Reza, N. G. I., 2017, November. **Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification**. In 2017 2nd *International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 294-298). IEEE.

[23] Devi, R. D. H., and Devi, M. I. **Outlier Detection Algorithm Combined With Decision Tree Classifier for Early Diagnosis of Breast Cancer**, *Int J Adv Engg Tech*, vol. 7, no. 2, pp. 93-98, 2016.

[24] Tharwat, A. **Classification assessment methods**. *Applied Computing and Informatics*, 2018. https://doi.org/10.1016/j.aci.2018.08.003