JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

ANALYSIS OF ACCURACY IMPROVEMENT IN RANDOM FOREST USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Hanna Willa Dhany, Muhammad Iqbal
Sains dan Teknologi Universitas Pembangunan Pancabudi Medan
E-mail: hdhany@dosen.pancabudi.ac.id

Abstract

Decision tree is used to classify a data that still does not know its class to existing classes. The data testing path is the first step that the root node goes through and finally the leaf node will predict the class for the data that has been concluded. Random Forest cannot be relied on for data types that have different categorical variables and therefore needs to be improved in the classification process, this is influenced by differences in the value of the variable. Therefore a method is needed to reduce features that are less relevant to the process of determining accuracy in the classification of the Random Forest method. In research conducted on the PCA + Random Forest classification model, using the Water Quality Status Dataset that has been simplified into 5 attributes, 4 classes and 117 instances with an accuracy rate of 91.43% with a classification error rate of 8.57%. Based on the test results from the four classification models, it can be concluded that the success of the PCA can be used as a reference to improve the accuracy performance of the Random Forest classification model.

Keywords: PCA, Random Forest, Decision Tree, Classification.

INTRODUCTION

Information Technology is very influential in every life management, from its use in education to industry to make a management decision. With the existence of information technology that has developed this has a positive impact on human life. Information technology is very valuable, because it provides benefits both directly and indirectly. So humans become more productive in making changes and enhancing knowledge[1]. Random Forest is a combination model of a tree that uses random vectors taken separately from input vectors, and each tree provides its popular class for classifying input vectors (Breiman, 1999). In other words, this model uses randomly selected features or a combination of features in each node to generate a tree [2]. Random Forest used to solve problems. The Random Forest method is one of the methods in the Decision Tree. Decision Tree is a flowchart shaped like a tree that has a root node that is used to collect data, an inner node that is at the root node that contains questions about data and a leaf node that is used to solve problems and make decisions[3]. Decision tree classifies a sample of data that is not yet known its class into existing classes. The use of decision trees in order to avoid overfitting a data set when achieving maximum accuracy[4]. PCA gives good results when applied to correlated attributes. In this study, PCA was applied in training and testing factors that greatly influenced the dataset. PCA will identify patterns in the data set, find similarities and differences between each factor. Because PCA serves as a powerful model

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

ISSN : 1979-9292

E-ISSN : 2460-5611

for analyzing data[5]. Hussainet used the Principal Component Analysis (PCA) approach as a method of feature selection to reduce indicators related to prediction of survival rates of patients infected with breast cancer. The data used were from the SEER dataset of 684,394 patient medical records. With the proposed approach obtaining an accuracy of 92%[5]. Principal Component Analysis (PCA) approach is expected to be able to simplify and eliminate factors that are less dominant or relevant to affect the data tested but have a large correlation to the formation of the tested data factors with a total proportion of expected variance of covariance of 60%. So this makes it easier for educational institutions to further improve the accuracy of the data being tested[6]. Data classification is data categorization into different categories according to the rules. In this classification aims to change the structure of the object instance. Classification algorithms are made from training sets and build models and models used to classify new objects. The decision tree evaluates the power of classification by analyzing the performance and results of the analysis[6]. Weaknesses in the Random Forest method can affect performance in grouping data. Performance can be interpreted as the level of achievement seen from the percentage of accuracy and accuracy in classifying. Therefore, based on previous studies, this research is proposed with the aim of increasing the performance of the Random Forest method by reducing irrelevant features and classifying the classes in the dataset using Principal Components Anaysis (PCA) so that the process of determining the root-node will more optimal[7]. It is hoped that this will be able to overcome weaknesses in Random Forest and result in increased accuracy in classifying the data used[8].

METHOD

Decision tree is used to classify a data that still does not know its class to existing classes. The data testing path is the first step that the root node goes through and finally the leaf node will predict the class for the data that has been deduced. *Random Forest* Data representation in the form of trees has advantages over other approaches that are meaningful and easily interpreted. The goal is to create a classification model that predicts target attribute values (often called classes or labels) based on multiple input attributes from sample data sets. Each tree interior node corresponds to one of the input attributes. The number of sides of a nominal interior node is equal to the number of possible values of the corresponding input attribute. The exit side of the numeric attribute is labeled. Each leaf node represents the value of the label attribute given by the input attribute value which is represented by the path from the root to the leaf. Pruning is a technique in which leaf nodes that do not add to the discriminatory power of the tree are removed. This is done to convert trees that are too specific or too many to be placed into more general shapes to increase their predictive power in invisible datasets. Pre-pruning is a type of pruning done parallel to the process of making trees. Random forest produces a set of a number of random trees which are the results of a random forest (forest / collection of trees). Selected sound

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

models from all trees produced. Development of trees in the random forest to achieve the highest size of the data tree. however, the formation of a random forest tree is not done by pruning which means a method to reduce the contents of the space. The formation is done by applying the randomization method to reduce errors. Formation of trees with sample data uses variables that are taken at random and carry out classification on all trees that have been formed. In Random Forest a solution is used to divide data based on the type of attribute used. Random Forest is one way of applying the stochastic discrimination approach to classification. In the classification process that will run when all trees have been formed and when the classification process is finished, the initialization is done with as much data based on its accuracy value. Random forest has its own advantage that is able to classify data that has incomplete attributes, and can also be used for classification and regression but not too good for regression, because it is more suitable for classifying data and also serves to process quite a lot of data. In the process after the process is formed, a selection is made for each class of existing data. After that the selection of each class is then taken the most elections, classification of data using random forest will produce the best selection. Advantages of Random Forest One of the most accurate algorithms available. On a large dataset, it produces a very accurate classification. Large databases run efficiently. Random forest can handle thousands of input variables without removing variables. Random forest provides an estimate of what is important in classification a. Losses from Random Forest has been over-observed for some datasets by random classification. In variable data with a number of different levels, supports these attributes with many levels. Therefore, the most important value of the random forest variable is not reliable for this type of data.

RESULT

Random Forest cannot be relied on for data types that have different categorical variables and therefore need to be improved in the classification process, this is influenced by differences in the value of the variable. Therefore a method is needed to reduce features that are less relevant in the process of determining accuracy in the classification of the Random Forest method. This research seeks to improve the accuracy of the classification of Random Forest using the Principal Component Analysis (PCA) method so that the two methods are compared to measure the performance of the method based on the resulting accuracy.

Table 1 Normalization Results of Water Quality Status Data

No	TSS (mg/L)	DO (mg/L)	COD (mg/L)	*BOD (mg/L)*	*Total phospat (mg/L)*	…	*Pij*
1	-0.90	-0.45	-0.44	-0.35	-0.42	…	-1.04
2	-0.89	-0.08	-0.20	-0.31	-0.24	…	-1.01
3	-0.89	-0.16	-0.27	-0.33	-0.33	…	-1.00
4	-0.87	-0.38	-0.51	-0.43	-0.06	…	-1.01

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

5	-0.87	-0.31	-0.44	-0.35	-0.37	…	-1.03
6	-0.85	-0.38	-0.44	-0.35	-0.42	…	-1.03
7	-0.85	-0.45	-0.44	-0.32	0.07	…	-1.01
8	-0.83	-0.16	0.07	-0.38	-0.33	…	-0.99
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
117	0.77	1.25	0.57	0.14	-0.78	…	1.04

Normalization calculation uses equation 3.1, which is Z-Score for TSS attribute (mg / L) from data to -1 as follows:

$$z = \frac{x-\mu}{\sigma} ; z = \frac{2-53.23}{56.39} = -0.90$$

z: standard score, x: observation data, μ: mean per variable and σ: standard deviation per variable. The results of the Z-score are data with mean = 0 and standard deviation = 1.

Correlation Calculation Results

The next process is to calculate the correlation value between attributes using the covariance equation. Covariance is used to measure the magnitude of the relationship between two attributes.

Next is the calculation of the Water Quality Status rating value between attribute X1 and attribute X2:

Table 2 Calculation of Correlation (Covariance) of the Water Quality Status dataset

No	X_1	X_2	X_1-X_{avg}	X_2-X_{avg}	Product
1	-0.90	-0.45	-0.90	-0.45	0.41
2	-0.89	-0.08	-0.89	-0.08	0.07
3	-0.89	-0.16	-0.89	-0.16	0.14
4	-0.87	-0.38	-0.87	-0.38	0.33
5	-0.87	-0.31	-0.87	-0.31	0.27
⋮	⋮	⋮	⋮	⋮	⋮
117	0.77	1.25	0.77	1.25	0.96

$$\text{Cov}(X_1, X_2) = \frac{\sum(-0.9-0)(-0.45-0)+(-0.89-0)(-0.08-0)\ldots(0.77-0)(1.25-0)}{117-1} = \mathbf{-0.011}$$

Next is to enter the results of the calculation of the covariance value for each pair of attributes into the form of a covariance matrix (Covariance Matrix) measuring 8x8 where the value of Cov (X1, X2) is equal to Cov (X2, X1), the value of Cov (X1, X3) is equal to the value of Cov (X3, X1) and so on in the same way applies to each attribute pair.
The following are the results of obtaining Covariance values from the Water Quality Status dataset using Rapidminer

Attributes	TSS (mg/L)	DO (mg/L)	COD (mg/L)	BOD (mg/L)	Total phosp...	Fecal Colifo...	Total Colifor...	Pij
TSS (mg/L)	1	-0.011	0.095	0.091	0.082	-0.068	-0.064	-0.004
DO (mg/L)	-0.011	1	-0.212	-0.221	0.156	-0.083	0.050	0.080
COD (mg/L)	0.095	-0.212	1	0.897	0.068	-0.045	0.001	0.167
BOD (mg/L)	0.091	-0.221	0.897	1	0.032	-0.031	0.033	0.213
Total phospat (mg/L)	0.082	0.156	0.068	0.032	1	0.092	0.158	0.195
Fecal Coliform (mg/L)	-0.068	-0.083	-0.045	-0.031	0.092	1	0.430	0.283
Total Coliform (mg/L)	-0.064	0.050	0.001	0.033	0.158	0.430	1	0.685
Pij	-0.004	0.080	0.167	0.213	0.195	0.283	0.685	1

Figure 1 Covariance Matrix Water Quality Status (Rapidminer) Calculation Results

In the Water Quality Status dataset, the TSS attribute (mg / L) has a variance covariance correlation of 1, while the TSS attribute correlation (mg / L) with

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

DO (mg / L) has a correlation (variance covariance) of -0.011 and so on in the same way also applies to each attribute pair. The image is the result of calculating the correlation between attribute pairs using the covariance equation.

Eigenvalue Decomposition Results from Covariance Matrix

The covariance matrix that is formed has a square matrix of size m x m, then for the eigenvalue (λ) that corresponds to the covariance matrix it has a scalar (λ1, λ2 …… ..λm) that is used to calculate the attribute weights using the eigenvector.

The eigenvalue for the Water Quality Status dataset is formed from the diagonal of the covariance matrix (Cov (X1, X1), Cov (X2, X2), Cov (X3, X3), Cov (X4, X4), ..., Cov (X8, X8)). The diagonal value is the same as the result of calculating the covariance variance value of each attribute as follows:

Varians kovarians = $Cov(X_1,X_1)+Cov(X_2,X_2)+Cov(X_3,X_3)+\ldots.+Cov(X_8,X_8) = 1+1+1+1+1+1+1+1 = 8$

Then proceed with calculating the percentage value of the proportion of covariance variance for the Water Quality Statistics dataset as follows:

1. Proportion Value PC 1 (%) = $\frac{\text{Nilai Eigen PC 1}}{\text{Varians kovarians}}$ x100% = $\frac{216.80}{8}$ x100% = 27.10%
2. Proportion Value PC 2 (%) = $\frac{\text{Nilai Eigen PC 2}}{\text{Varians kovarians}}$ x100% = $\frac{191.20}{8}$ x100% = 23.90%
3. Proportion Value PC 3 (%) = $\frac{\text{Nilai Eigen PC 3}}{\text{Varians kovarians}}$ x100% = $\frac{116.80}{8}$ x100% = 14.60%
4. Proportion Value PC 4 (%) = $\frac{\text{Nilai Eigen PC 4}}{\text{Varians kovarians}}$ x100% = $\frac{94.40}{8}$ x100% = 11.80%
5. Proportion Value PC 7 (%) = $\frac{\text{Nilai Eigen PC 7}}{\text{Varians kovarians}}$ x100% = $\frac{28}{8}$ x100% = 3.50%
6. Proportion Value PC 8 (%) = $\frac{\text{Nilai Eigen PC 8}}{\text{Varians kovarians}}$ x100% = $\frac{10.40}{8}$ x100% = 1.30%

Table 3: Results of EigenWater Quality Status Decomposition Results

PC	Eigen Value	Proportion Varian (%)	Cumulative
1	216.80	27.10	27.10
2	191.20	23.90	50.90
3	116.80	14.60	65.50
4	94.40	11.80	77.30
5	80.80	10.10	87.40
6	62.40	7.80	95.30
7	28.00	3.50	98.70
8	10.40	1.30	100.00

Eigen value decomposition results from the Water Quality Status dataset use the Eigen Value Decomposition (EVD) equation. The eigenvalue is obtained from the product of the variance proportion value with the total variance covariance of the attributes. Can be seen as follows:

Eigenvalue = Proportion of Variants (%) x Covariance Variance

= 27.10 *x* 8

= 216.80

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

and so on in the same way also applies to each PC

Orthogonal Principal Component (PC) Results

In this study, the number of principal components chosen is the maximum proportion of variance covariance that is able to explain the variance of covariance from the original attribute. Based on the above table, the proportion of covariance variance taken for the Water Quality Status dataset is the cumulative proportion of 50.90% obtained from the sum of the variance proportion values of the 1st principal component to the 2nd principal component so that a number of 2 principal components is obtained as in the following table:

Table 4: Orthogonal Principal Component (PC) Results for Water Quality Dataset Status

PC	Nilai Eigen	Proporsi Varian (%)	Cumulative
1	216.80	27.10	27.10
2	191.20	23.90	50.90

***Variance Threshold* = 50.90%**

The percentage of variance of PC 1 variance can only meet 27.10% of the total variance of original data covariance, if coupled with the proportion of variance of PC 2, it has been able to fulfill 50.90% of the total variance of original data covariance so that it can already represent the maximum proportion of variance in original data of 50% of the data original whole.

Factor Rotation Results using Varimax Rotation (Eigenvector)

The factor rotation process aims to find factors that are able to optimize the correlation between the observed indicators. In this study, the rotational factor used is the varimax rotation with a loading factor greater than 0.3. To choose which original attributes are included in a number of principal components (attributes that affect the variance of the original data covariance), the factor rotation process is performed using an eigenvector, where each x-vector value formed will correspond to one eigenvalue (λn).The results of the factor rotation use the eigenvector (varimax rotation) from the Water Quality Statistics dataset as follows:

Attribute	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
TSS (mg/L)	0.052	-0.143	-0.465	-0.815	0.302	0.065	-0.027	-0.001
DO (mg/L)	-0.123	0.270	-0.569	0.440	0.343	0.525	-0.044	0.008
COD (mg/L)	0.477	-0.463	-0.056	0.155	-0.018	0.173	-0.104	-0.700
BOD (mg/L)	0.493	-0.448	-0.025	0.150	0.037	0.155	-0.038	0.711
Total phospat (mg/L)	0.187	0.173	-0.576	0.017	-0.750	-0.195	-0.018	0.037
Fecal Coliform (mg/L)	0.260	0.366	0.335	-0.299	-0.300	0.695	0.154	-0.001
Total Coliform (mg/L)	0.414	0.464	0.097	-0.027	0.207	-0.210	-0.719	0.011
Pij	0.487	0.335	-0.054	0.070	0.306	-0.318	0.667	-0.055

Figure 2: Factor Rotation Results (eigenvector) Dataset Water Quality Status

Based on Figure 4.3, the eigenyang value symbolized as (λ) is a scalar number. The Water Quality Status dataset measures 8 x 8 to obtain values of oxygen values (λ1, λ2 …… ..λn). Eigenvalues and Eigenvectors can both define the Water Quality Status dataset matrix. The equation for calculating the Eigenvector is as follows:

$Ax = \lambda x$
$Ax - \lambda x = 0$
$(A - \lambda)\, x = 0$
$(A - \lambda I)\, x = 0,\ x \neq 0$

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

Where :
A = nxn matrix which has n eigenvalue (λ_n)
λ = Eigenvalue Value
x = matrik non-zero
I = matrik identities
In order to obtain a linear combination, namely:

a. $\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_n$ is *eigenvalue* matrix dataset
b. $x_1, x_2, x_3 \ldots x_n$ is *eigenvector* same with *eigenvalue* (λ_n)

In order to obtain a linear combination, namely:
$AX = XD$
$A = X D X^{-1}$
A = nxn matrix which has n eigenvalue (λn)
D = eigenvalue from the eigenvector
X = eigenvector of matrix A
X-1 = inverse of eigenvector X
The calculation result of the eigenvector equation is the value of Loading Factor from PC 1 to PC 8 is the value of the magnitude of the correlation between a number of principal components with the indicators that exist in the Dataset Water Quality Status. Interpretation of the results is done by looking at the value of factor loading contained in the results of factor rotation with a loading factor value greater than 0.3 . The process of determining which variables will be included in the factor is done by looking at the ratio of the correlation in each row in each factor matrix table as follows: the loading factor value is greater than 0.3

Table 5 Dominant Factors of Water Quality Status

Factor	Variable	Label	Eigenvalue	Loading Value	Variance %
PC 1	X_3	COD (mg/L)	216.8	0.477	27.10%
PC 1	X_4	BOD (mg/L)	216.8	0.493	27.10%
PC 1	X_8	Pij	216.8	0.487	27.10%
PC 2	X_6	Fecal Coliform (mg/L)	191.2	0.366	23.90%
PC 2	X_7	Total Coliform (mg/L)	191.2	0.464	23.90%

Based on table 4.10, the dominant factors of Water Quality Status are obtained based on the highest eigenvector value (factor loading) generated from PC 1 and PC 2 that meet the proportion of covariance variance of 50.90%. After analyzing the factors using the principal component analysis method, 2 factors are obtained: The first factor (PC 1) is the most dominant factor having an eigenvalue of 216.80 and able to explain a total diversity of 27%. This factor consists of variables X3 = COD (mg / L), X4 = BOD (mg / L) and X8 = Pij, called the main factor. The second factor (PC 2) consists of X6 = Fecal Coliform (mg / L) and X7 = Total Coliform (mg / L) with an eigenvalue of 191.20 and is able to explain a total diversity of 24%. This factor is said to be a supporting factor.

Acquisition of Water Quality Status Dataset Accuracy Results

The Water Quality Status dataset has 8 attributes, 4 classes and 120 instances, class distribution in the form of good conditions (30 instances), lightly polluted (30 instances), medium polluted (30 instances) and heavily polluted (30 instances). The data is divided by 70% from the data will be used as training data and as much as 30% of the data will be used as test data conducted randomly.

Table 6 Water Quality Status Dataset Attribute Information

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

Atributte	Value
TSS (mg/L)	[2-266]
DO (mg/L)	[0.02-8.43]
COD (mg/L)	[1.7-416]
BOD (mg/L)	[0.6-150]
Total phospat (mg/L)	[0.0016-1.23]
Fecal Coliform (mg/L)	[27-2800000]
Total Coliform (mg/L)	[74-53000000]
Pij	[0.54-15.31]
Quality Status	{*good condition, lightly polluted, medium polluted, heavily polluted*}

Next is a simulation model of the Random Forest classification using Rapidminer:

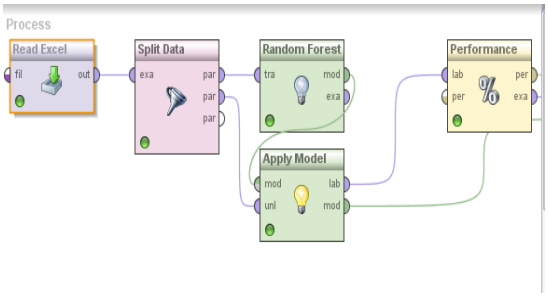

Figure 3. Simulation of the Random Forest Classification Model

Water Quality Status (PCA + Random Forest) Dataset Accuracy Test Results

The results obtained from the PCA analysis in the form of dominant factors Water Quality Status were selected based on the highest eigenvector value (factor loading) generated from PC 1 and PC 2 that met the proportion of covariance variance of 50.90%.

Table 6 Details of Water Quality Status Data (PCA)

COD (mg/L)	BOD (mg/L)	Fecal Coliform (mg/L)	Total Coliform (mg/L)	Pij	Quality Status
8	2.6	92	150	0.76	Good Condition
19.2	3.1	92	150	0.88	Good Condition
16	2.9	930	2400	0.91	Good Condition
4.793	1.32	1100	1400	0.87	Good Condition
8	2.5	230	750	0.81	Good Condition
8	2.6	150	210	0.78	Good Condition
8	3	750	2100	0.87	Good Condition
32	2.1	36	740	0.99	Good Condition
16	2.3	92	740	0.88	Good Condition
⋮	⋮	⋮	⋮	⋮	⋮
55.4	10	1500	2800000	10.74	Heavily Polluted

In testing the accuracy of the PCA + Random Forest classification model the data still uses the proportion of data sharing 70% of training data and 30% of test data. So that the accuracy of PCA + Random Forest classification results obtained in the Water Quality Status dataset can be seen in the following confusion matrix table:

Table 7 Confusion Matrix for PCA + Random Forest classification

Kinerja Klasifi	**Predicted Class**

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (113-122)

ISSN : 1979-9292

E-ISSN : 2460-5611

kasi				
Actual Class	Predicted. Good Condition	Predicted. Lightly Polluted	Predicted. Medium Polluted	Predicted. Heavily Polluted
Actual. Good Condition	9	1	0	0
Actual. Lightly Polluted	0	8	1	0
Actual. Medium Polluted	0	0	6	0
Actual. Heavily Polluted	0	0	1	9

Then proceed with calculating the Accuracy value and the classification error level (Classification_error) of the PCA + Random Forest (Water Quality Status) classification model. Following the calculation results:

a. *Accuracy* = $\frac{9+8+6+9}{9+8+6+9+1+1+1} = \frac{32}{35}$ = 0.914285*100% = **91.43%**
b. *Classification_error* = $\frac{1+1+1}{9+8+6+9+1+1+1} = \frac{3}{35}$ = 0.857*100% = **8.57%**

CONCLUTION

In the research that has been done, the authors produce several conclusions as follows: 1. In research conducted on the PCA + Random Forest classification model, using the Water Quality Status Dataset which has been simplified into 5 attributes, 4 classes and 117 instances with an accuracy rate of 91.43% with a classification error rate of 8.57%. 2. Based on the test results of the four classification models, it can be concluded that the success of PCA can be used as a reference to improve the accuracy performance of the Random Forest classification model.

REFERENCE

[1] Agjee Na'eem Hoosen., Mutanga Onisimo., et al. 2018. *The Impact of Simulated Spectral Noise on Random Forest and Oblique Random Forest Classification Performance*. Journal of Spectroscopy.

[2] Chang, C., Wu, Y., Hou, S. 2009. *Preparation and Characterization of Superparamagnetic Nanocomposites of Aluminosilicate/Silica/Magnetite*, Coll. Surf. A336: 159,166.

[3] Dai Qin-yun,. Zang Chun-Ping., Wu Hao. 2016. *Research of Decision tree Classification Algorithm in Data Mining*. Dept. of Electric and Electronic Engineering, Shijiazhuang Vocational and Technology Institute. China

[4] Hussain, H., Quazilbash. N.Z., Bai. S. &Khoja, S. 2015. Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis.*International Conference on Computer-Based Medical Systems*, pp. 131-134.

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

ISSN : 1979-9292

E-ISSN : 2460-5611

[5] Manasi M. Phadatare, Sushma S. Nandgaonkar. 2014. *Uncertain Data Mining usig Decision Tree and Bagging Technique*. Department of Computer Engineering, India.

[6] Pal. M. 2007. *Random Forest Classifier for Remote Sensing Classification***.** National Institute of Technology, Department of Civil. Haryana

[7] Patel, B. N., Prajapati G. Satish., Lakhtaria I. Kamaljit. 2012. *Efficient Classification of Data Using Decision Tree*. Bonfring International Journal of Data Mining, Vol. 2, No.1.

[8] Paul Angshuman, Mukherjee Dipti Prasad, et.al. 2018. *Improved Random Forest for Classification*. IEEE Transaction on Image Processing Volume: 27, Issue:8

[9] Seema., Rathi Monika., Mamta. 2012. *Decision Tree: Data Mining Techniques*. Department of Computer Science Engineering. India.

[10] Yang Bo-Suk., Di Xiao., and Han Tian. 2008. *Random Forests Classifier for Machine Fault Diagnosis*. Journal of Mechanical Science and Technology 22.

Submitted : 04 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020