Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Performance Accuration Method of Machine Learning for Diabetes Prediction

Dwi Harini Sulistyawati[1,a)], Ali Murtadho[2,b)],
[1,2] Informatics Engineering , University 17 August 1945 Surabaya
n
[a)] Dwiharini@untag-sby.ac.id
[b)] Id.alimurtadho@gmail.com

ARTICLE INFO

Article history:
Received: 02/04/2020
Revised: 10/04/2020
Accepted: 01/05/2020

Keywords:
Machine Learning Prediction Diabetes,
Performa Accuration Method ,
Supervised Learning,
AI(artificial intelligence)

A B S T R A C T

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning (ML) techniques allow us to obtain predictively, the dataset we are testing is pima-indian-diabetes with a dataset of 765 raw data with 8 data features and 1 data label we developed a method to achieve the best accuracy from the 5 methods we use with the stages of separation training and testing the dataset, scaling features, parameters evaluation, confusion matrix and we get the accuracy of each method, and the results of the accuracy we get with these 5 methods Gradient-boosting is best with an accuracy score of 0.8, Decision Tree 0.72, Random Forest 0.72, next is Logistic Regression 0.7, and then followed by K-NN method with a score of 0.65.

1. Introduction

Diabetes is a long-lasting or chronic disease and is characterized by high blood sugar (glucose) levels or above normal values. Glucose which accumulates in the blood due to not being absorbed by body cells properly can cause various disorders of the body's organs. If diabetes is not well controlled, various complications that can endanger the lives of patients can arise [1].Machine Learning (ML) is one branch of the discipline of Artificial Intelligence (AI) which discusses the development of systems based on data. Many things are learned, but basically there are 4 main things learned in machine learning [2]. 1. Supervised Learning, 2. Unsupervised Learning, 3. Semi-Supervised Learning, 4. Reinforcement Learning[3].

Classification techniques in this research produces more accurately such predictive models are one of the most common applying research Machine Learning (ML) techniques train the data and make the function inferred, which is can be used to map new or invisible examples. The main purpose of classification techniques is to accurately estimate the target class for each case in data. Classification algorithms generally require that class is defined based on data attribute value.

The dataset we tested this time we took from pima indian diabetes specification data features 8 raw and data labels, the amount of data we tested amounted to 785 datasets, for learning process in this time we apply several stages to see the performance of the accuracy of the method we will test the steps we use.

This trial first 1. Raw data, 2. Featureextraction, 3. Hyperparameter and Tunning, 4. Confusion matrix and result acurration. For more details can be seen in Figure 1. In this research, we have studied the performance of five different models to compare the accuracy of the model, model 1. Gradient Boosting, 2. K-Nearest Neighbor (K-NN), 3. Decision Tree, 4. Logistic Regression, 5. Random Forest [3].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

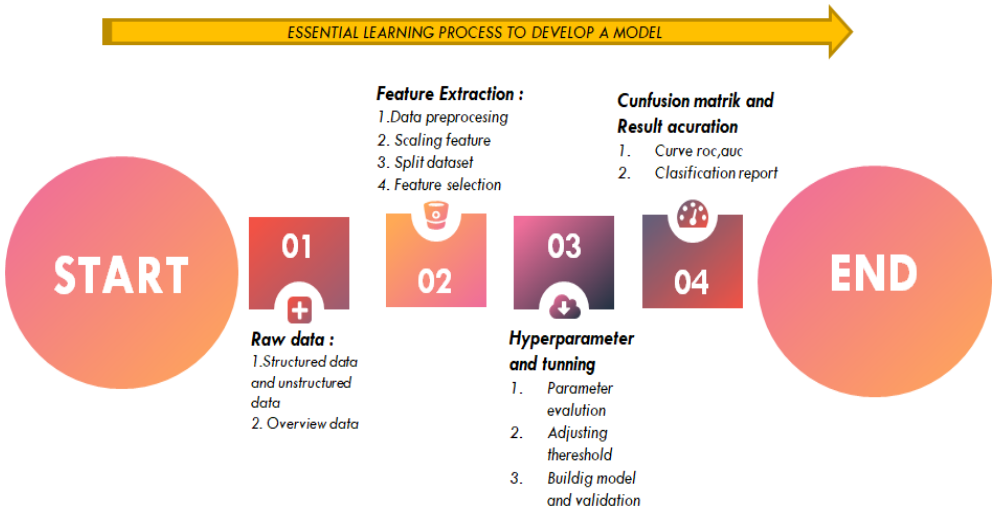


Fig 1. Essential Learning Process

2. Material And Method

2.1 Proposed Methodology

In this model, we have used 5 methods; 1. GradientBoosting, 2. Random Forest, 3. Decision Tree, 4. Logistic Regression, 5. K-NN for comparison of accuracy, the main focus in this research is knowing how accurate the method we will use so we can find out which methods will produce the best accuracy and which is the worst accuracy among the 5 methods. For the dataset there are a total of 768 rows and are divided into 2 classes: diabetics and non-diabetics with eight feature data, eight features are 1. Pregnancy, 2. Glucose, 3. Blood Pressure, 4. Blood Pressure, 5. Insulin, 6. BMI (Body Mass Index), 7.Diabetes Pedigree Function, 8.Age, and we include the flow of the algorithm performance process in the learning process as shown in Figure 2.

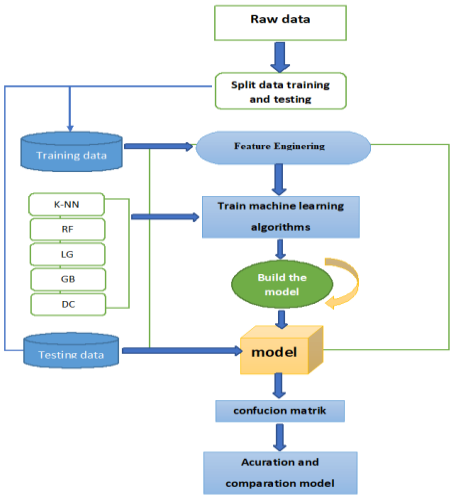


Fig. 2 Flow Proces

This research we made a split test with parameter constraints like the following train_test_split (X, y, test_size= 0.25, random_state=42),where we made a test data of 0.25 from testing data and supplying data with parameters 42.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

2.2 Gradient Boosting

In the gradient boosting algorithm this time we use the adjusting development thresholdGradientBoostingClassifier (learning_rate= 0.05, max_depth= 3, max_features= 0.5, random_state= 42), result this accuracy on training set: 0.882, accuracy on test set: 0.750 and obtain the matrix confusion results as in Figure 3[4].

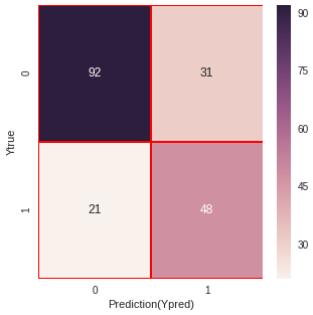


Fig 3. ConfusionMatrikGradientBoosting

2.3 Random Forest

Random Forest algorithm we use the customize development thresholdRandomForestClassifier (n_estimators= 100, criterion= 'gini', max_depth= 6, max_features= 'auto', random_state= 0), produce accuracy on this training set: 0.917, accuracy on the test set: 0.745 and get the results of the matrix confusion as shown in Figure 4[7].

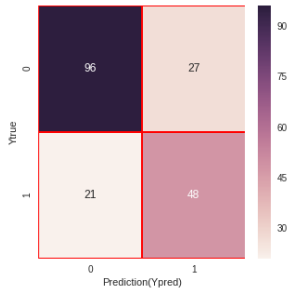


Fig 4. Confusion Matrik Random Forest

2.4 Decision Tree

The Decision Tree algorithm we use the customize development threshold DecisionTreeClassifier (max_depth= 6, max_features= 4, min_samples_split= 4, random_state= 42), produce accuracy on this training set: 0.852, accuracy on the test set: 0.729 and get the results from the confusion matrix as shown in Figure 5.

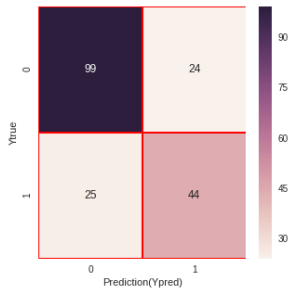


Fig 5. ConfusionMatrikDecision Tree

2.5 Logistic Regresion

Logistic Regression algorithm we use the Customize development threshold logreg_classifier= LogisticRegression(C = 1, penalty = 'l1'), produce accuracy on this training set: 0.783, accuracy on the test set: 0.724 and get the results from the confusion matrix as shown in Figure 6 [6].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

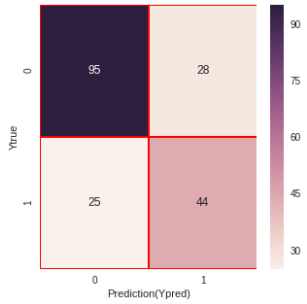


Fig 6. ConfusionMatrik Logistic Regresion

2.6 K-NN

K-NN algorithm we use the customize development threshold Kneighbors Classifier (algorithm='auto',leaf_size=30,metric='minkowski',metric_params=None, n_jobs=1,n_neighbors=5,p=2,weights='uniform'), produce accuracy on this training set: 0.77, accuracy on the test set: 0.71 and get the results from the confusion matrix as shown in Figure 7.

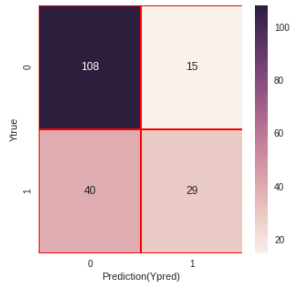


Fig 7. ConfusionMatrik K-NN

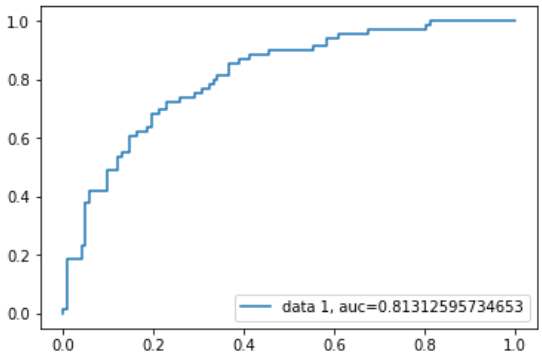
3. Result And Discusion

The final results of the experiment of the confusion matrix of each algorithm will be continued with the classification report calculation which includes [6].

1. Precision = (TP / (TP + FP)) * 100%.
2. Recall = (TP / (TP + FN)) * 100%.
3. F1 Score = 2 * (Recall * Precission) / (Recall + Precission)
4. Accuracy = (TP + TN) / (TP + TN + FP + FN)) * 100%
5. Macro avg = calculates metrics independently for each class and then takes the average (hence treating all classes equally).
6. Weighted avg = returns the average by considering the proportions for each label in the dataset.

From the results of the classification report, each method will get the best accuracy result and will be explained in the curcoca in the image below.

3.1 Gradient Boosting



Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Fig 8. Curvroc_Aucalgoritm Gradient Boosting

Figure 9 is a preview of the accuracy of the Gradient Boosting method. The graph described in Figure 9 is a linear graph with blue for horizontal variables is the range of the smallest accuracy to the largest number and for the same vertical variable means the smallest range of accuracy to the smallest number with the following accuracy calculation:

Auc = (TP + TN) / (TP + TN + FP + FN)) * 100%
= (48 +96) / (48 + 96 + 27 + 21) * 100%
= 0.813

on the variable in the graph there is an auc data information which means this is the result of the accuracy of the Gradient Boosting method having a value (0.813).

3.2 Decision Tree

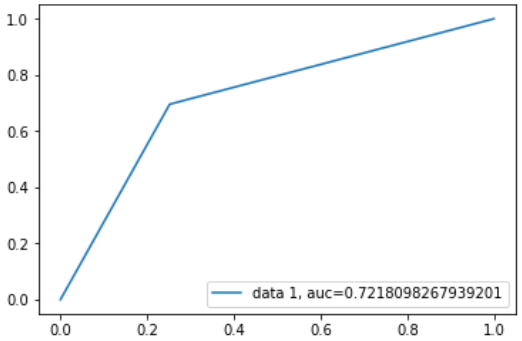


Fig 9. Curvroc_aucalgoritm Decision Tree

Figure 10 is a preview of the accuracy of the Decision Tree method. The graph described in Figure 10 is a linear graph with blue for horizontal variables is the range of the smallest accuracy to the largest number and for the same vertical variable means the smallest range of accuracy to the smallest number with the following accuracy calculation:

Auc = (TP + TN) / (TP + TN + FP + FN)) * 100%
= (48 +92) / (48 + 92 + 32 + 21) * 100%
= 0.72

on the variable in the graph there is an auc data information which means this is the result of the accuracy of the Decision Tree method has a value (0.72).

3.3 Random Forest

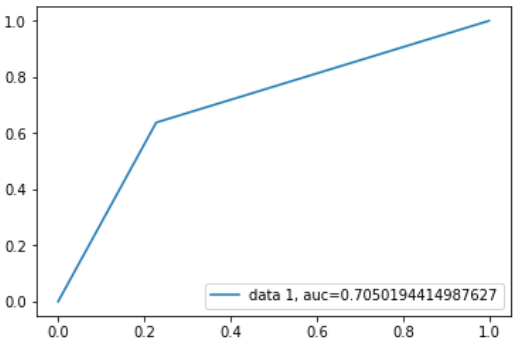


Fig 10. Curvroc_aucalgoritm Random Forest

Figure 11 is a preview of the accuracy of the Random Forest method. The graph described in Figure 11 is a linear graph with blue for horizontal variables is the range of the smallest accuracy to the largest number and for the same vertical variable means the smallest range of accuracy to the smallest number with the accuracy calculation as follows:

Auc = (TP + TN) / (TP + TN + FP + FN)) * 100%
= (44 +99) / (44 + 99 + 24 + 25) * 100%

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

= 0.72

on the variable in the graph there is an auc data information which means this is the result of the accuracy of the Decision Tree method has a value (0.72).

3.4 Logistic Regresion

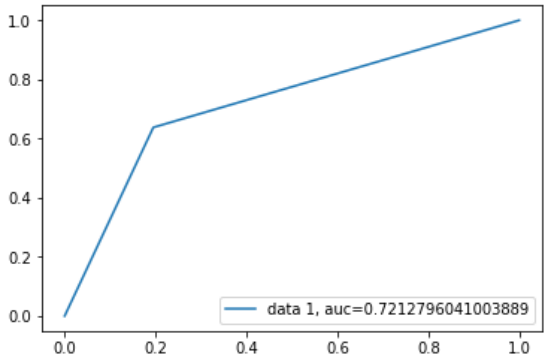


Fig 11. Curvroc_aucalgoritmLogistic Regresion

Figure 12 is a preview of the accuracy of the Logistic Regression method. The graph described in Figure 12 is a linear graph with blue for horizontal variables is the range of the smallest accuracy to the largest number and for the same vertical variable means the smallest range of accuracy to the smallest number with the accuracy calculation as follows:

Auc = (TP + TN) / (TP + TN + FP + FN)) * 100%
= (44 +95) / (44 + 95 + 28 + 25) * 100%
= 0.7

on the variable in the graph there is an auc data information which means that this is the result of the accuracy of the Logistic Regression method which has a value (0.7).

3.5 K-NN

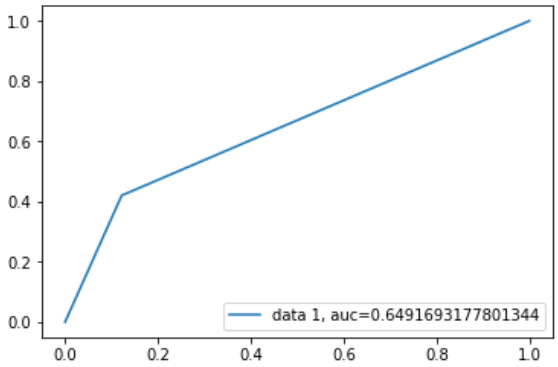


Fig 12. Curvroc_aucalgritm K-NN

Figure 13 is a preview of the accuracy of the K-Nearest Neighbors method. The graph described in Figure 13 is a linear graph with blue for horizontal variables is the range of the smallest accuracy to the largest number and for the same vertical variable means the smallest range of accuracy to the smallest number with the accuracy calculation as follows:

Auc = (TP + TN) / (TP + TN + FP + FN)) * 100%
= (29 + 108) / (29 + 108 + 15 + 40) * 100%
= 0.64

on the variable in the graph there is an auc data information which means this is the result of the accuracy of the K-Nearest Neighbors method has a value (0.64).

3.6 Comparison of five algorithms

The results of the comparison of the five algorithms that have been explained in the above stages are explained in this 3.6 chart with the best accuracy rating obtained by the Gredient boosting algorithm with

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

an accuracy value of 0.81 and for the lowest accuracy, the K-NN method with an accuracy value of 0.64, this information detail can see on figure 14.

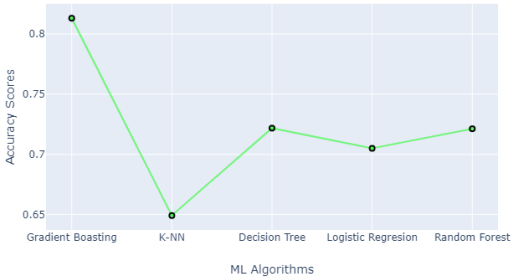


Fig 13. Five Method Comparison Chart

The final results of the trial have been done as the table above, there you can see thatthe best level of accuracy is the Gradient Boosting method with accuracy (0.81) and for the method with the worst accuracy is K-NN (0.64), followed by the results of the Logistic Regression method with accuracy (0.705), Random Forest (0.72) and Decision Tree (0.72).

From the results we can analyze why the BoostingGradient method is more accurate because there are several things that are affected as follows:

The effect of cross-validation techniques that are determined by the same data split and data testing. Here we can know the difference when entering the cross-validation stage the results of the confusion matrix show the number of results from the gradient boosting method with the TP result "true positive" highest than the other methods "48" although the Decision Tree method also produces the same number on the TP results, However, in the Decision Tree FN method "false negative" or miss more accuracy than the Gradient Boosting method[11].

4. Conclusion

Based on the results of trials of supervised learning techniques with a comparison of 5 methods namely K-NN, Logistic Regression, Random Forest, Decision Tree, Gradient Boosting can be seen the results of the accuracy of the 5 methods, that the most accurate method for prediction of diabetes with supervised learning techniques is pima -indian is the Gradient Boosting method and for the worst accuracy the accuracy is the K-NN method, and for the prediction results of the decision tree method, logistic regression and random forest results, it is almost the same. Henceforth, we should use more data to train the model because it is in the machine learning the more data used in training the model, the better the model will be.

From these results produce an analysis of why the gradient boosting method is more accurate because there are several things that are affected as follows.

1) The effect of cross validation techniques that are determined by the same data split and data testing. Here we can know the difference when entering the cross validation stage the results of the confusion matrix show the number of results from the gardient bosting method. With the TP result "true positive" highest than the other methods "48" although the Decision Tree method also produces the same number on the TP results , but in the Decision Ttree FN method "false negative" or miss more accuracy than the Gradient Boosting method.
2) In one of the journal works by Jordan Frery, Amaury Habrard, Marc Sebban, Olivier Caelen, and Liyun He-Guelton, entitled "Efficient top-ranking optimization with improved gradients for controlled anomaly detection explainsthe journal also proves that for searching the method with the most efficient and accurate supervised learning technique is the boosting technique and from the method used above for comparison of the accuracy of the boosting technique there is the Gradient Boosting method.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 164-171 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

3) Articles written by AlbolfazlRavanshad "Data Scientist, Ph.D. from the university of florida and he graduated from Udacity's nano degre machine learning program. Explain about the performance of Gradient Boosting performance with the Random Forest method that:
   a) Gradient Boosting: GBT creates trees one by one, where each new tree helps correct the mistakes made by trees that were previously trained
   b) Strength of the model: Because the enhanced tree is derived by optimizing objective functions, GBM can basically be used to solve almost all objective functions that can be written to gradients. This includes things like ranking and poetic regression, which RF is more difficult to achieve.of learning, and each tree that is built is generally shallow.

5. Acknowledgement

This research presents his sincere appreciation goes to works have provided motivation, advice, and support. On this occasion, we would like to thank everyone, especially the 17 August 1945 Surabaya University, Faculty of Informatics, supervisors who have helped him patiently complete this undergraduate thesis by giving advice, guidance, and correction until the completion of this thesis. this undergraduate thesis is far from perfect, but is expected to be useful in the future, not only for researchers but also for readers and junior at the 17 August 1945 Surabaya University.

6. References

[1] B. S. D. Soumya, "J. Diabetes Metab.,," Late stage complications of diabetes and insulin resistance, vol. 2 (167), pp. pp. 2-7, 2011.

[2] towardsdatascience.com, "type-of-machine-learning," https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861.

[3] U. s. K. C. o. T. S.Saru, "International Journal of Emerging Technology and Innovative Engineering," ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING, vol. Volume 5, no. Issue 4, p. (ISSN: 2394 – 6598), April 2019.

[4] D. Chaturvedi, "Mathematical Models, Methods and Applications," Soft computing techniques and their applications, p. 31–40. Springer Singapore, 2015.

[5] https://www.python-course.eu/Boosting.php, "Boosting," python-course.eu.

[6] databricks.com, "random-forest," https://databricks.com/blog/2015/01/21/random-forests-and-boosting-in-mllib.html.

[7] towardsdatascience.com, "Logistic-Regresion," https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102.

[8] T. Y. e.-m. t. Kamer Kayaer e-mail: kayaer@yildiz.edu.tr, "Medical Diagnosis on Pima Indian Diabetes," Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks,Yildiz Technical University , Department of Electronics and Comm. Eng. Besiktas, Istanbul 34349 TURKEY .

[9] A. H. M. S. O. C. a. L. H.-G. Jordan Frery, "Efficient top rank optimization with gradient boosting," Efficient top rank optimization with gradient boosting for supervised anomaly detection, pp. 1 Univ. Lyon, Univ. St-Etienne F-42000, UMR CNRS 5516, Laboratoire Hubert-Curien, France .

[10] www.codepolitan.com, "python," https://www.codepolitan.com/memulai-pemrograman-python.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.