Plagiarism Checker X Originality Report

Similarity Found: 7%

Date: Tuesday, December 05, 2023
Statistics: 195 words Plagiarized / 2742 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

Comparison of Multi Layer Perceptron, Random Forest & Logistic Regression on Students Performance Test 1st Musthofa Galih Pradana Computer Faculty UPN Veteran Jakarta Jakarta, Indonesia musthofagalihpradana@upnvj.ac.id 4th Dhina Puspasari Wijaya Computer Engineering Faculty Alma Ata University Yogyakarta, Indonesia dhina.puspa@almaata.ac.id 2nd Kenneth Palilingan Computer Engineering Faculty Sam Ratulangi University Yogyakarta, Indonesia kennethpalilingan@unsrat.ac.id 5th Pujo Hari Saputro Engineering Faculty Sam Ratulangi University Manado, Indonesia pujoharisaputro@unsrat.ac.id 3rd Yuri Vanli Akay Engineering Faculty Sam Ratulangi University Manado, Indonesia yuriakay@unsrat.ac.id Abstract—The test is one thing that can be taken to measure a person's ability to understand a material or a competency.

In general, there is a final test taken by students at the school level, before reaching the final test, usually students will take a series of preparatory tests. In reality, of course, not all students can take the test preparation well. therefore of course the school has data related to test preparation. From this test preparation data, a classification technique can be used to classify the data of students who have completed the preparatory test and students who have not completed the preparatory test, so that schools can prepare the best strategy.

To assist in classifying data, data classification techniques are needed, in this study the Multi- Layer Perceptron, Random Forest and Simple Logistics algorithms were used. These three methods produce different accuracy when used for the data classification process. For testing the data, scenarios are used using cross-validation. The results of this test scenario show that the Logistic Regression method is superior to the Random Forest and Multi-Layer Perceptron methods with an accuracy of 73.9%.

The best Root Mean Square Error results are in the Multi-Layer Perceptron method with the smallest value of 0.363. Keywords—Classification, Multi-Layer Perceptron, Random Forest, Logistic Regression I. INTRODUCTION The test is one thing that can be taken to measure a person's ability to understand the material or a competency. The test produces a numerical value that represents how well a person's understanding is described by the score. In schools, various types of tests are usually carried out to measure students' abilities. With this test, it is hoped that it can describe the abilities and results of understanding related to the material and students' knowledge.

Of the many types of tests, schools will usually have a final test that is used as the main measure of student success in learning. Before arriving at the final test, the school will hold a series of tests to prepare students more mature in completing their final test. A series of tests are held within a certain period and are also carried out in stages and continuously. Of course in preparation, not all students complete the test preparation well.

The data of students who take the preparatory test can be done with a classification technique to be able to classify the data of students who have completed the preparatory test and students who have not completed the preparatory test. This test grouping data can be used to prepare the best strategy for the school to achieve better results related to students final test results. To classify data, it is necessary to apply data mining, namely classification. The classification technique performs data grouping, the data is grouped based on the relationship or data related to the sample data [1].

In this study, the dataset used is a student performance dataset based on several test scores and test preparation scores. The data will be processed using classification techniques, or in other words, the data will be classified using the Multi- Layer Perceptron, Random Forest, and Simple Logistic Classifier algorithms. These three classification methods will be compared with each other and look for the one with the highest accuracy in classifying the data. The technique used in testing the data in this study is to use cross-validation.

Cross-validation is the process of randomly dividing data into several parts. From this test, it will be found how much truth the algorithm has in classifying the data appropriately according to its class. The testing parameters of the three algorithms are the value of precision, recall, and F-Measure The reason for using these three methods is that there has been no research on the detailed comparison of the three methods that will be tested and compared the results, namely Multi-Layer Perceptron, Random Forest, and Logistic Regression. This is what distinguishes it from previous relevant reference research, in reference research, there is no detailed research that compares these three methods. II.

STATE OF THE ART The relevant research referred to is that of Kanish Shah and the team who made comparisons with the classification of texts. The results show that the logistic regression classifier using the TF-IDF vectorizer function achieves the highest accuracy of 97% on the data set. This algorithm has been proven to be the most stable classifier on small data sets [2]. More research from Jaime Lynn Speiser and the team who wrote a comparison of random forest variable selection methods for classi?cation prediction modeling.

Comparison of results of available methods for randomly selecting forest variables in the context of classification using 311 data sets freely available online. Preference was given to the method with the lowest out-of-bag error rate, computation time, and several variables [3]. Mohammad Reza Sheykhmousa wrote a systematic review Support Vector Machine vs. Random Forest for 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) 978-1-6654-7327-9/22/$31.00 ©2022 IEEE 462 Remote Sensing Image Classification.

The result challenges, recommendations, and Possible directions for future research are also discussed in detail. Additionally, a summary of the results is provided Researchers can tailor their efforts to achieve maximum results Accurate results based on theme application [4]. Random Forest vs Logistic Regression Kasichtten by Kaitlinof Kirasich. The result of each case study 1000 simulations and model performance consistently showed this Statistical False Positive Rate for Random Forest with 100 Trees It is different from logistic regression.

In all four cases, logistic regression and Random forest achieved different relative classification scores among different Simulated recording conditions [5]. Random forest for big data classification in the internet of things using optimal features writing by S. K. Lakshmanaprabu. The result of research observed by the implementation of the maximum accuracy of the proposed method was 94.2%. Validate the effectiveness of the proposal The method analyzes various key performance indicators and compares them with existing methods [6].

Shalin Savalia published research on Cardiac Arrhythmia Classi?cation by Multi-Layer Perceptron and Convolution Neural Networks. As a result, the accuracy of MLP was 88 the and 83.5% accuracy of CNN. The expected method performance is Decent, but the problem of arrhythmia diagnosis is still not resolved. has many complications The algorithm can efficiently diagnose various cardiovascular diseases with an accuracy of 88.7 [7]. Weibiao Qiao experiments with his research classification using local wavelet acoustic patterns and Multi-Layer Perceptron neural networks. Based on the results obtained, more It classified the sonar data as 1.2891 better than GMDH.

Score along with other classifications. Overall, the use of MLP-more A study of the classification of passive sonar targets shows this. This algorithm can be used to classify different high grades A dimensional underwater data sets [8]. Image classification is written by Zhifei Lai using the method Multi-Layer Perceptron. We base our model on neural networks and merge the different feature groups obtained in the first and second steps. we rate them The approach proposed for his two benchmark medical image datasets of HIS2828 and ISIC2017. achieve global classification with 90.1% and 90.2% higher accuracy than currently successful methods [9].

Method Multi-Layer Perceptron was also used by M. Khishe with object sonar dataset classification. Used a sonar dataset and compared the results obtained using PSO, ACO, ES, and GA benchmark algorithms. The result is, SFS with a simple structure and powerful search ability In the utilization phase, it can bring better results. Convergence speed, confinement to local minima, and classification accuracy Compared with benchmark algorithms [10]. Adhien Kenya published research also about comparison logistic regression. The result shows that the k-means clustering model results (22%) are much lower than the logistic regression model results (91%) [11].

Method logistic regression for prediction customer churn written by Arno De Caigny. LLM provides more accurate results, as evidenced by this customer benchmark study. Models using LR and DT building blocks as standalone [12]. Implementation logistic regression in papers Fadi Thabtah machine learning autism classification. The results obtained show that the machine learning technique was able to generate an acceptable classification system. Above all performance in terms of sensitivity, specificity, and accuracy [13]. Abrar Ahmed research about logistic regression for Scene Classi?cation. Using the proposed system object segmentation.

The problem was investigated using two robust algorithms, MFCS and MSS. Also, object similarity was studied by multiple kernel learning. We used logistic regression to classify complex scenes. Experimental evaluation shows that the proposed system consistently outperforms other state-of-the-art systems A system of calculations, divisions, and precision [14]. III. METODOLOGY A. Research Stage The stage of this research start from collecting data, the data has been classified using 3 different methods, and the result the is accuracy of each method, the accuracy results will be compared and the best accuracy from the three methods.

The testing parameters of the three algorithms are the value of precision, recall, and

F-Measure. Details are shown in Error! Reference source not found. Fig. 1. Research Stage B. Random Forest A random forest is a combination of all good trees combis ined into a model. A random forest relies on random vector values that have the same distribution in all trees, and each decision tree has maximum depth. A random forest consists of trees {h(x, k), k = 1.[15]. C. Multi-Layer Perceptron Multilayer Neural Networks, also known as Multilayer Perceptron. It is a further development of the single-layer perceptron. to learn.

Using a delta algorithm called Error Backpropagation Training Algorithm, input arguments are compared forward during the process. Uses learning in addition to forward propagation Backpropagation. If the result does not match the target, the weight. It is updated during the learning cycle process until the error value is reached. Expected minimum or output equals target. [1] D. Logistic Regression Logistic regression is a technique used to describe the relationship between input and output variables. Input 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) 463 variables are considered independent variables and output variables are called dependent variables.

A dependent variable can only take on a fixed set of values. These values correspond to problem classification classes. The goal is to identify the relationship between independent and dependent variables by estimating probabilities using the logistic function. A logistic function is a sigmoid curve used to construct functions with various parameters. This is closely related to general linear model analysis, which tries to fit a line to as many points as possible to minimize error. [15] IV. RESULT A.

Data Preparation The data used in this study such as test preparation data which contains completed and none completed means that you have taken the preparatory exam, if none means that you have not taken the preparation test, as well as test scores for each subject such as math score, writing score and The detailed reading score is shown in Error! Reference source not found.. Fig. 2. Datasets B. Multi-Layer Perceptron Classification using the Multi-Layer perceptron method, testing using cross-validation technique. The results of the classification test using a multi-layer perceptron are shown in Error! Reference source not found.. TABLE I.

CLASSIFICATION RESULT MLP No K-Fold Accurate Precision Recall F-Measure 1 10 66,9 0,788 0,788 0,754 2 15 64,7 0,718 0,741 0,730 3 20 65,2 0,716 0,759 0,737 4 25 66,2 0,721 0,773 0,746 Fig. 3. Chart Accuracy MLP The best accuracy in the Multi-Layer Perceptron method is at k-fold 10, with an accuracy of 66.9. Apart from the accuracy data, the next data from the Root Mean Square Error is shown in Error! Reference source not found. TABLE II. RMSE RESULT MLP No K-Fold Root Mean Square Error 1 10 0,496 2

15 0,363 3 20 0,514 4 25 0,506 Fig. 4. Chart RMSE MLP Root Mean Square Error value, the best value is at k-fold 15 with a value of 0.363. C.

Random Forest Classification using the random forest method, testing using a cross validation technique. The results of the classification test using a random forest are shown in Error! Reference source not found. TABLE III. CLASSIFICATION RESULT RANDOM FOREST No K-Fold Accurate Precision Recall F-Measure 1 10 65,3 0,696 0,815 0,751 2 15 65,8 0,701 0,816 0,754 3 20 64,8 0,640 0,808 0,747 4 25 65,8 0,699 0,819 0,755 Fig. 5. Chart Accuracy Random Forest The best accuracy in the Random Forest method is at k- fold 15 and k-fold 25, with an accuracy of 65.8. Apart from the accuracy data, the next data from the Root Mean Square Error is shown in Error! Reference source not found.

66.9 64.7 65.2 66.2 K=10ÿK=15ÿK=20ÿK=25 ACCURATE Accurate 0.496 0.363 0.514 0.506 K=10ÿK=15ÿK=20ÿK=25 RMSE RMSE 4.3 2.5 3.5 4.5 K=10ÿK=15ÿK=20ÿK=25 ACCURATE Accurate 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) 464 TABLE IV. RMSE RESULT RANDOM FOREST No K-Fold Root Mean Square Error 1 10 0,470 2 15 0,470 3 20 0,472 4 25 0,467 Fig. 6. Chart RMSE Random Forest Root Mean Square Error value, the best value is at k-fold 25 with a value of 0.467. D.

Logistic Regression Classification using logistic regression method, testing using cross validation technique. The results of the classification test using a random forest are shown in Error! Reference source not found. TABLE V. CLASSIFICATION RESULT LOGISTIC REGRESSION No K-Fold Accurate Precision Recall F-Measure 1 10 73,6 0,757 0,868 0,808 2 15 73,9 0,758 0,872 0,811 3 20 72,9 0,753 0,860 0,803 4 25 72,8 0,752 0,860 0,802 Fig. 7. Chart Accuracy Logistic Regression The best accuracy in the Logistic Regression method is at k-fold 15, with an accuracy of 73.9. Apart from the accuracy data, the next data from the Root Mean Square Error is shown in Error! Reference source not found.

TABLE VI. RMSE RESULT LOGISTIC REGRESSION No K-Fold Root Mean Square Error 1 10 0,417 2 15 0,418 3 20 0,417 4 25 0,417 Fig. 8. Chart RMSE Logistic Regression Root Mean Square Error value, the best value is k-fold 10, 20 and 25 with a value of 0.418. E. Comparison Based on the testing of each algorithm, be compared between the three methods used in classifying the algorithm that has the highest accuracy is logistic regression with an accuracy of K=15 of 73.9%. This shows that with not too many datasets with a data range of 1000 data, the classification method with logistic regression is more effective, because many at least can affect the accuracy of results.

While the smallest RMSE value is an anomaly that occurred in the study. The details of the comparison of each accuracy result along with the best RMSE result or the smallest value in the multilayer perceptron method with a value of 0.363. This anomaly could have happened in the research conducted by Candra Dewi writing the comparison of neural network and ANFIS, the result is that the smaller RMSE value does not necessarily indicate a higher level of accuracy [16]. The result of the three methods is shown in Error! Reference source not found. Fig. 9.

Result The results obtained are that the logistic regression value with better accuracy is due to the longer classification process compared to the other 2 methods, with a longer iteration process. Logistic Regression takes 10 ms, Multilayer perceptron 5 ms, and random forest 2 ms. V. CONCLUSION The conclusion that can be drawn from the research that has been done is, the best accuracy between the three methods is the logistic regression method with an accuracy of 73.9% followed by the multi-layer perceptron method with an accuracy of 66.9% and finally the random forest method with 0.47 0.47 0.472 0.467 K=10ÿK=15ÿK=20ÿK=25 RMSE RMSE 73.6 73.9 72.9 72.8 K=10ÿK=15ÿK=20ÿK=25 ACCURATE Accurate 0.417 0.418 0.417 0.417 K=10ÿK=15ÿK=20ÿK=25 RMSE RMSE 66.9 65.8 73.9 0.363 0.467 0.418 MLPÿRF LR RECAP Accurate RMSE 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) 465 an accuracy of 65. 8%.

While the root means square error value states that the best value is found in the multi-layer perceptron method with the smallest error value compared to the others of 0.363, the result of RMSE is an anomaly.

INTERNET SOURCES:

1% -
https://www.semanticscholar.org/paper/Comparison-of-Multi-Layer-Perceptron%2C-Random-Forest-Pradana-Palilingan/014db57ec0a36d6f9d7769969e0ef327aa29d3e5/figure/1
<1% - https://www.thoughtco.com/the-purpose-of-tests-7688
2% - https://ieeexplore.ieee.org/document/10017501/
<1% -
https://www.wallacefoundation.org/knowledge-center/Documents/using-data-strategically-group-students-tip-sheet.pdf
<1% -
https://academia.stackexchange.com/questions/117314/how-to-write-represent-the-state-of-the-art-analysis-in-a-research-paper
<1% - https://www.sciencedirect.com/science/article/pii/S0957417419303574

<1% - https://www.researchgate.net/publication/367319922_A_Preliminary_Study_of_Decision_Support_Model_of_Photovoltaic_for_Village_Area_in_South_of_Sumatera
<1% - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4170322/
1% - https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview
<1% - https://www.researchgate.net/publication/348901338_Medical_Image_Classification_Techniques_and_Analysis_Using_Deep_Learning_Networks_A_Review
1% - http://ijiis.org/index.php/IJIIS/article/view/74
<1% - https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262
<1% - https://corporatefinanceinstitute.com/resources/data-science/random-forest/
<1% - https://lib.ugent.be/en/catalog/ebk01:26036112200041
<1% - https://www.sagepub.com/sites/default/files/upm-binaries/45679_1.pdf
<1% - https://stats.stackexchange.com/questions/164412/fail-to-improve-recall-in-classification
<1% - https://statisticsbyjim.com/regression/root-mean-square-error-rmse/
<1% - https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
<1% - https://www.thewindowsclub.com/error-reference-source-not-found-microsoft-office