ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research, volume 440
International Conference on Online and Blended Learning 2019 (ICOBL 2019)

Prediction Analysis Student Graduate Using Multilayer Perceptron

Mariana Windarti[1]
[1]*Faculty of Computer Science, Universitas Widya Dharma*
Klaten, Indonesia
marianawindarti@gmail.com

Putri Taqwa Prasetyaninrum[2]
[2]*Faculty of Information Technology, Universitas Mercubuana Yogyakarta*,
Yogyakarta, Indonesia
putri@mercubuana-yogya.ac.id

***Abstract*—Student graduation data is a data that is important to the College, especially for the Faculty as well as the courses in question. Acquisition of knowledge in a database (a number of large data) commonly referred to as data mining. This research aims to analyze the student's graduation predictions that can be done on a fourth semester using Multilayer Perceptron (MLP) classifier which available in WEKA software implementations. Then do the testing and performance comparisons of MLP against Naïve Bayes classification, IBk and Tree J48. Cross Validation and Percentage Split are used as the testing procedure in this research. The parameters in the process of testing using correctly classified instances and Root Mean Squared Error (RMSE). On the mode of Cross Validation, MLP has better performance compared to all contender methods with accuracy of J48 81.82% and the value of the smallest RSME i.e. 0.273. On a Percentage Split MLP mode has the same accuracy value with Naïve Bayes i.e. 92.31%, and the value of the RMSE on the MLP of 0.182.**

Keywords: Multilayer Perceptron (MLP), data mining, correctly classified instances, Root Mean Squared Error (RMSE)

I. INTRODUCTION

World Education is currently cannot be separated from technology advancement [1]. Department of information systems of Universitas Mercu Buana Yogyakarta store data in computerized form. It contains data about students, lecturers, staffs, and financial stuffs. Meanwhile, this huge amount of data is actually can be used as a source of strategic information for the department to perform classification of the study period of the student graduates using data mining techniques. Aside from providing information that is strategic for the Faculty and course of study, it can also enhance the efforts to encourage and expedite their graduation. So, in addition to can be beneficial to the students themselves, it can also increase the value of accreditation for the department [2]. Student graduation data is one of the elements that are present in High College accreditation standard BAN-PT [3]. The number of students graduating on time can be improved by improving the quality of learning and academic services to students. In addition, if a student can study completion time predicted then the handling of students would be more effective. One of the prediction techniques that can be used is with the technique of data mining or data mining. Data mining based on data in a college education can improve the quality of student learning in College [4]. Therefore, colleges need to detect the behavior of students who have a status of active in order to be known the cause of the failure of the student who does not comply with the study period. Some of the causes of student failure include low academic ability, financial factor, domicile when sitting for study, and other factors [5]. There are some researcher suggesting the application of backpropagation algorithm to predict, such as in research conducted by Rudy. Backpropagation algorithm with initial setting of 12 parameters input, 9 hidden node (neuron) and 3 output used to predict student graduation requirements in STMIK Indonesia Banjarmasin and produce accuracy of 92.49% [6]. Other data mining techniques also used by Sri Redjeki, using the Backpropagation algorithm and K-Nearest Neighbor (KNN) to identify a disease with 82 patients and divided into 72 training data and the remaining 10 data as testing data. The algorithm performance results show that KNN has accuracy better than backpropagation with the accuracy of 100% [7].

Research using Multilayer Perceptron (MLP) for the prediction of new Admissions done and produce accuracy rate of 89,59% [8]. Method of Multilayer Perceptron (MLP) Neural Network and Logistic Regression to predict the behavior of the population implemented in a junior high school in Saudi Arabia. The results of the research have MLP better performance than in the case of Logistics Regression prediction [9]. Research by using the classifier method or the same with MLP is done to determine the nutritional status of the toddler and the required food menu recommendations. The sample data used as many as 166, retrieved the value of the precision of 82.609% [10]. Other research to analyze performance comparison against 2 or more classifier done on the case of predictions. Comparative analysis of Multilayer Perceptron (MLP) and Autoregressive Integrated Moving Average (ARIMA) using performance measurement parameters evaluation of Root Mean Squared Error (RMSE) and coefficient of determination (R2). Performance results show that MLP has a lower prediction error of on Autoregressive Integrated Moving Average (ARIMA) [11]. Another case study for prediction of death from a heart attack is done using some of the techniques of classification such as the Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN) the Support Vector Machine (SVM), and the Mixture of Expert (ME). The combination of these four classifiers done by experts and produce performance superior specificity and sensitivity, i.e. the accuracy value of 84.24%, 85.71% and 82.85% [12].

Student's graduation best prediction analysis is performed using data mining classification MLP, and compare the performance of MLP with the other classifier such as Naïve Bayes, KNN, and Decision Tree C 4.5. By analyzing the performance of each party, the classifier are expected to lower the number of students who drop out in on the course of information systems Universitas Mercu Buana Yogyakarta and aims to improve academic quality. The author uses the classification method of Multilayer Perceptron MLP method (MLP). It is a simple method of statistical probability yet produces accurate results.

II. MATERIAL AND METHOD

A. Material

Data collection is done directly by requesting data on graduates to the academic part. The dataset used is a data information system study Program graduates during the 2014-2018. A dataset has been collected and has gone through the stages of data preprocessing totaled 66 record data. The dataset will be used in the process of training and testing in the classification process. The number of graduates is still quite a bit because of the information system department is still a new course at UMBY. The input variables used consisted of parent's income, 1st Semester's performance index, 2nd semester performance index, 3rd semester's performance index, school majors, type of student and age at the fourth semester. While the output or target variable is the study period. Prediction of graduates can be done on a fourth semester student.

A classification or grouping variable or attribute values that are used can be seen in table 1 below. While the display visualizes the inputs and outputs attributes of 8 attributes can be seen in figure 1 below.

Table 1. The Classification Variable Value

Attribute	Classification
Parent's income	< 1.5 million
	> = 1.5 – 2 million
	> = 2 – 2.5 million
	> = 2.5 – 3 million
	> = 3 million
Grade Point (GP)	>=2.5–3
	>=3–3.5
	>=3.5
Department	IPA
	IPS
Student Type	New
	Transfer
Age	19 year old
	20 years old
	21 years old
	22 years old
Study period	< 4 Years
	> = 4 – 4.5 years
	> = 4.5 – 5 years
	> = 5 years

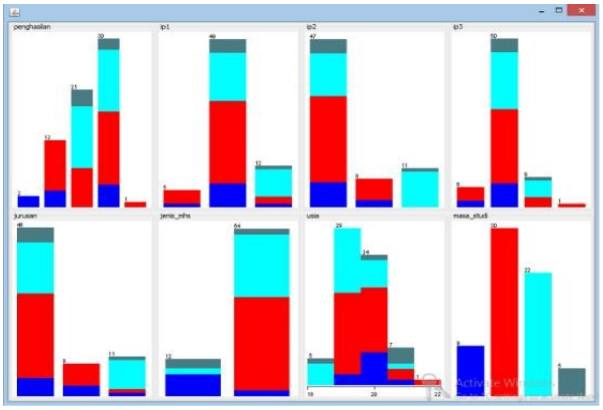


Figure 1. Visualization Display Attribute Input and Output on the WEKA

B. Method

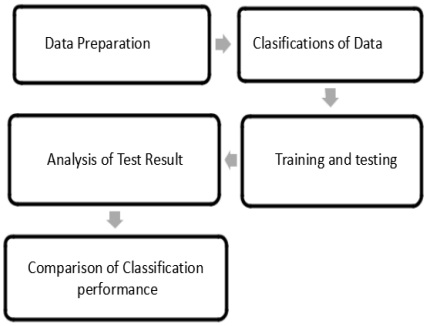


Figure 2. Research Flow

1. The preparation of the Data.

Raw data retrieved in a format .xls or .xlsx (Microsoft Excel). The data is stored in the format. CSV (Comma Separated Value), then converted to data with the format extensions .ARFF. Data that can be used by the WEKA is a data format .ARFF (Attribute Relation File Format) and have gone through preprocessing stage.

2. Classification of Data

A classification or grouping of the value of each attribute or variable can be done based on an existing value.

3. Training and testing

The process of training and testing on the WEKA is done via module Classify using classifier Multilayer Perceptron (MLP), Naïve Bayes, IBk (K-NN) and Tree J48 (implementation of Decision Tree C 4.5). The testing method used is the percentage split and cross validation [13]. For percentage split is determined in advance that is 80%, that is to say the amount of training data that are used as much as 80% of the entire dataset i.e. 53 record and the remaining 20% as the training data that is as much as 13 record. Cross validation Testing mode uses 10-fold meaning data training and test data is shared as much as 10 section 10 of the data or a subset of the same size. The process of training and testing will be done by as much as 10 times repeatedly.

4. The analysis of the test results

Measurement of performance evaluation of each classifier using the parameters Correctly Classified Instances and RMSE (Root Mean Squared Error). Correctly Classified data is the number of instances that are predicted correctly or referred also to the value of accuracy, while for the data predicted incorrectly included into Incorrectly classified instances. Evaluation with the Root Mean Squared Error (RMSE) was conducted to find out how big the error resulting from the classifier used.

5. Comparison of classification performance

The performance results of each classification are MLP, Naïve Bayes, KNN and Decision Tree C 4.5 then compared to using parameter Correctly Classified Instances and RMSE.

III. RESULTS AND DISCUSSION

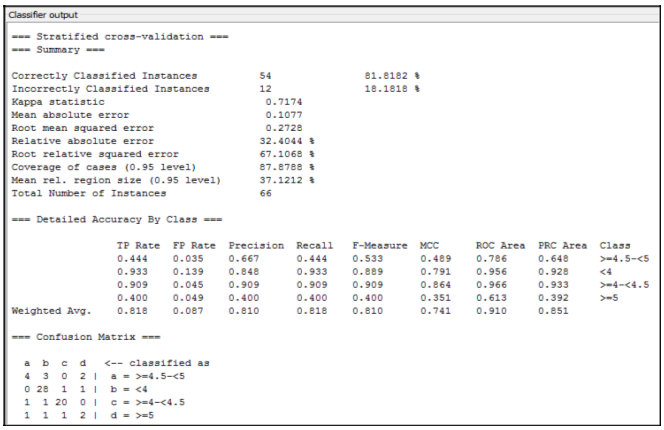
The process of training and testing on this research use the dataset as much as 66 record. The set of data has gone through stages so that there is no longer a preprocessing of data that the attribute value is empty or is missing and data is ready to be processed. Sample data ready to be used in the process of training and testing of WEKA can be seen in table 2 below.

Table 2. Sample Training Data & Testing

Parents Income	GP1	GP2	GP3	Department	Student type	Age	Study period
>=1.5-2	>=2.5 - 3	>=3 - 3.5	>=2.5 - 3	IPA	transfer	19	>=4.5-<5
<1.5	>=3 - 3.5	>=2.5 - 3	>=2.5 - 3	IPA	fresh	21	<4
>=2.5-3	>=3.5	>=3 - 3.5	>=3 - 3.5	IPA	fresh	19	>=4-<4.5
>=1.5-2	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPA	fresh	19	>=4.5-<5
>=2.5-3	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPA	transfer	20	>=5
>=3	>=3.5	>=3 - 3.5	>=3 - 3.5	IPS	transfer	20	>=4.5-<5
<1.5	>=3 - 3.5	>=2.5 - 3	>=3 - 3.5	IPS	fresh	19	<4
>=2.5-3	>=3 - 3.5	>=3.5	>=3 - 3.5	IPA	fresh	19	>=4-<4.5
>=3	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPA	transfer	20	>=4.5-<5
>=2.5-3	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPA	transfer	21	>=5
>=3	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPA	fresh	19	<4
>=3	>=3.5	>=3.5	>=3.5	IPA	fresh	19	>=4-<4.5
>=3	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPA	fresh	20	>=4.5-<5
<1.5	>=2.5 - 3	>=2.5 - 3	<2.5	IPS	fresh	20	<4
<1.5	>=3 - 3.5	>=3 - 3.5	>=3 - 3.5	IPS	fresh	19	<4

The initial dataset is saved in the Microsoft Excel format is .xls or .xlsx. Later datasets are stored in format CSV (*.csv), after that data is converted into the format of ARFF (*. arff). Data that can be used or recognized by WEKA is a data format extension ".arff". The process of training and testing the classifier using Multilayer Perceptron (MLP) on the WEKA. Multilayer Perceptron is one of the methods in the Neural Network. The results of the prediction accuracy of graduation a student or correctly classified instances with MLP uses a test mode 10-fold cross validation of 81.8182%. 66 record from data use, data record 54 predicted correctly or produce output in accordance with the actual data and the predicted 12 record is not correct or does not match the actual data. This shows that the performance of MLP very well, evidenced by the percentage of accuracy above 80%.

In addition to using classifier MLP, the author uses other classifier to test whether the MLP have better performance or not. Other classifiers used i.e. Naïve Bayes, IBk (K-Nearest Neighbor) and J48 which is implementation of Decision Tree C 4.5. Evaluation of measurement results of testing each classifier uses the parameters of the accuracy (Correctly Classified Instance) and RMSE. RMSE value shown through data error after iteration of the data. RMSE values retrieved from the error test results test data compared to target or output [14]. Implementation process of testing on WEKA can be seen in figure 3 below. While the comparison of test results based on parameter accuracy (correctly classified instance) can be seen in table 3 and use the RMSE parameter can be seen in table 4.



```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               81.8182 %
Incorrectly Classified Instances        12               18.1818 %
Kappa statistic                          0.7174
Mean absolute error                      0.1077
Root mean squared error                  0.2728
Relative absolute error                 32.4044 %
Root relative squared error             67.1068 %
Coverage of cases (0.95 level)          87.8788 %
Mean rel. region size (0.95 level)      37.1212 %
Total Number of Instances               66

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.444    0.035    0.667      0.444    0.533      0.489    0.786     0.648     >=4.5-<5
                 0.933    0.139    0.848      0.933    0.889      0.791    0.956     0.928     <4
                 0.909    0.045    0.909      0.909    0.909      0.864    0.966     0.933     >=4-<4.5
                 0.400    0.049    0.400      0.400    0.400      0.351    0.613     0.392     >=5
Weighted Avg.    0.818    0.087    0.810      0.818    0.810      0.741    0.910     0.851

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  4  3  0  2 |  a = >=4.5-<5
  0 28  1  1 |  b = <4
  1  1 20  0 |  c = >=4-<4.5
  1  1  1  2 |  d = >=5
```

Figure 3. The test results with the MLP method of Cross Validation

Table 3. Comparisons of the Result Test Correctly Classified Instances.

Classifier	Cross Validation (%)	Percentage Split (%)
MLP	81.82	92.31
Naïve Bayes	75.76	92.31
Lazy.IBk	77.27	86.62
Trees.J48	75.76	84.62

Based on the test results, the overall classifier MLP has better performance compared with other Naïve Bayes classifier i.e., IBk and J48. Whereas the classifier with the less good performance in General is owned by J48 (Decision Tree). Figure 4 and 5 below shows the graph of the results of training and performance comparison testing using correctly classified and instance of the RSME.

Table 4. Comparison Of The Results Of The Test Correctly Classified Instances

Classifier	Cross Validation	Percentage Split
MLP	0.273	0.182
Naïve Bayes	0.308	0.29
Lazy.IBk	0.329	0.282
Trees.J48	0.324	0.284

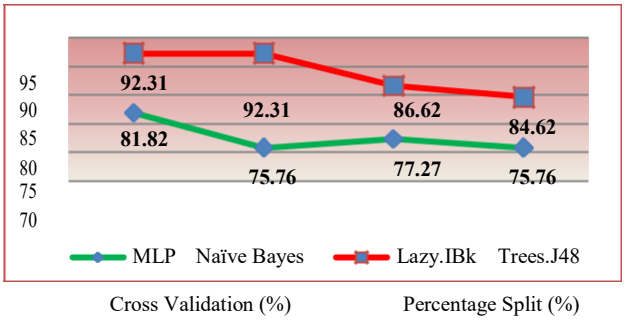


Figure 4. Comparisons Chart Correctly Classified Instance

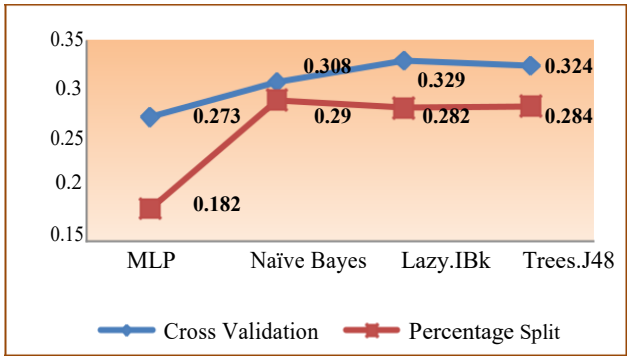


Figure 5. RMSE Comparison Chart

On mode testing with cross validation, i.e. 10-fold cross validation, MLP has accuracy classifier than most other classifier 81.82%. To the average value of error RMSE value MLP small compared to most other classifier. The smaller RMSE means that the methods do better. While in test mode percentage split, MLP and Naïve Bayes has most improved accuracy compared to IBk and J48 with equal accuracy value of 92.31%, then followed by the classifier IBk and last J48. RMSE value for MLP has a value other than the most minor of 0182.

In the table 4, there is a column of Study period which is the target of the data to be analyzed or actual data, and the column Study period is the predicted output or result predictions on the test process. If the prediction is equal to actual data (target) then the data calculated in the parameters correctly classified instances. For data with the results of the prediction does not correspond to real or actual data is marked in yellow on the record.

Table 4. The Test Result on the WEKA

Income	GP	department	age	'predicted study period'	Study period
>2.5-3	'>=3 - 3.5'	IPA	19	<4	
>=3	'>=3 - 3.5'	IPA	20	<4	
>=3	>=3.5	IPA	19	>=4-<4.5	
>=3	'>=3 - 3.5'	TKJ	21	>=5	>=4.5-<5
>=2.5-3	'>=3 - 3.5'	IPA	20	<4	
>=3	'>=3 - 3.5'	IPA	20	<4	
>=2.5-3	'>=3 - 3.5'	IPS	19	>=4-<4.5	<4
>=3	'>=3 - 3.5'	TKJ	20	>=4-<4.5	
>=3	'>=3 - 3.5'	IPA	19	>=4-<4.5	
>=2.5-3	'>=3 - 3.5'	IPA	20	>=5	
>=3	'>=3 - 3.5'	IPA	20	>=5	>=4.5-<5

IV. CONCLUSION

The graduation of students can be seen upon graduating on time or not. The students are said to pass on time or are still in the reasonable limits if the period of study is less than 4.5 years. If the study period had already passed the limit is then said to be passed not on time. Student graduation prediction can be done on a fourth semester with variable input and output among other parents income, Grade Point (GP) 1-3, department, student types, age and study period. The value of each variable or attribute is grouped first. Data for training and testing that is used on the WEKA software derived from data information system study Program graduates during the year of graduation 2014 until 2018 totaling 66 record data. The number of graduates belongs a little because this course is a study of a new stand at the University of Mercu Buana Yogyakarta. The data are ready to be used on the WEKA is a data format extension. ARFF. Based on actual data on students who have study period more than 4.5 years less than half the overall total graduates i.e. amounted to 13 people. Of the 13 people mostly from student transfer as many as 10 people and three new students.

Prediction analysis of graduation at WEKA using fad testing 10-fold cross validation and the percentage split of 80%. Performance measurement evaluation using parameter RMSE (Root Mean Squared Error) and correctly classified instances or accuracy. Based on the test results, in General classifier Multilayer Perceptron (MLP) has the most improved performance compared with other classifier third (IBk, Naïve Bayes, J48). This is shown on the test cross validation mode, that the value of accuracy highest IE MLP 81.82% RMSE values and very small i.e. 0.273. Being on a percentage split mode, accuracy of MLP 92.31% and the value of the RMSE 0.182.

REFERENCES

[1] R. P. S. Putri and I. Waspada 2018 Penerapan Algoritma C4.5 pada Aplikasi Prediksi

[2] M. A. Nurrohmat 2017 Aplikasi Pemrediksi Masa Studi dan Predikat Kelulusan Mahasiswa Informatika Universitas Muhammadiyah Surakarta Menggunakan Metode Naive Bayes J. Khazanah Inform. 1 29

[3] .M. D. Yalidhan 2018 Implementasi Algoritma Backpropagation Untuk Memprediksi Kelulusan Mahasiswa Klik - Kumpul. J. Ilmu Komputer 5 169

[4] M. M. Abu Tair and A. M. El-Halees 2012 International Journal of Information and Communication Technology Research Mining Educational Data to Improve Students' Performance: A Case Study 2 140–146

[5] Khafiizh Hastuti 2012 Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa Non aktif Seminar Nasional Teknologi Inf. Komunikasi Terapana (Semantik 2012) 14 241–249

[6] R. Ansari 2012 Prediksi Kelulusan Mahasiswa Dengan Jaringan Syaraf Tiruan jtiulm 1–6

[7] S. Redjeki 2013 Perbandingan Algoritma Backpropagation dan K-Nearest Neighbor (KNN) untuk Identifikasi Penyakit," Seminar Nasional Aplikasi Teknologi Informasi (SNATI) 1–5

[8] R. E. Putra, A. I. Nurhidayat, and A. Y. Wicaksono 2018 Implementation of Neural Network to determine the New College Students IOP Conf. Ser. Mater. Sci. Eng 288

[9] K. J. Assi, K. M. Nahiduzzaman, N. T. Ratrout, and A. S. Aldosary 2018 Mode choice behavior of high school goers: Evaluating logistic regression and MLP neural networks Case Studies Transport Policy 6 225–230

[10] Fitri, O. Setyawati, and D. Rahadi S 2013 Aplikasi jaringan syaraf tiruan untuk penentuan status gizi balita dan rekomendasi menu makanan yang dibutuhkan J. EECCIS 7 119–124

[11] A. Olawoyin and Y. Chen 2018 Predicting the Future with Artificial Neural Network Procedia Computer Science 140 383–392

[12] E. Ebrahimzadeh et al. 2019 An optimal strategy for prediction of sudden cardiac death through a pioneering feature-selection approach from HRV signal Computer Methods and Programs Biomedicine 169 19–36

[13] H. Jiawei, M. Kamber and J.Pei 2001 Data Mining: Concepts and Techniques

[14] A. S. Budiman and X. A. Parandani 2017 Uji Metode Klasifikasi Data Dalam Proses Seleksi Penerima Beasiswa SMK PGRI Ploso IJCIT (Indonesian J. Comput. Inf. Technol) 2 68–76