International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET23103145

The Study of Neural Network Algorithm, Random Forest for Classification of Student Graduation

Riyad Sabilul Muminin*[1], Ana Hadiana[2,3], Nila Natalia[4]
[1]Department of Engineering, Garut University, Garut, Jawa Barat, Indonesia
[2]Research Center for Information and Data Science (PRSDI), BRIN, Indonesia
[3]Department of Information System, STIMIK-LIKMI, Bandung, Indonesia
[4]Departement of Computer Engineering, Politeknik Sukabumi, Sukabumi, Indonesia

ARTICLEINFO

Article History:

Accepted: 01 June 2023
Published: 20 June 2023

Publication Issue
Volume 10, Issue 3
May-June-2023

Page Number
517-522

ABSTRACT

The academic performance is one of aspect which has remained the bechmark of the success in learning activities at an university. The indicator of academic performance in the university is the students able to complete their studies on time. Unfortunately, the problem regarding academic performance was associated with the completion time of student studies in Faculty of Economics, University of Garut. In this research explore the model that able to classify the graduation of student through the data mining classification technique by comparing the Neural Network Algorithm dan Random Forest. The classification conducted by evaluating the academic performance based on Semester Performance Index (IPS) first years in the beginning and use the demographics of students as attributes that will be used in the dataset. Based on the results of several model tests from the data train, totaling 1467 data records and 25 attributes. It shows that the 14th Random Forest test model produces a Recall Performance value of 72.70%, 74.70% for Accuracy Performance, 72.80% for Precision Performance and 74.70% for F-measure Performance.

Keywords: Neural Network, Random Forest, Science, Academic performance, Education, and Data Mining

I. INTRODUCTION

A College is a formal educational institution that creates human resources. Graduation rate, self-confidence, ability, passing all courses, graduating on time, or more specific achievements are indicators of student success [1]. The Ministry of Education and Culture of the Directorate General of Higher Education (Kemenristek-Dikti) regulates the education system in universities. The student must

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited

complete a minimum study load of 144 credits (Semester Credit Units) for the Bachelor's Degree and 108 credits for the Diploma. Generally, a bachelor finishes the study for eight semesters or four years and four semesters or three and a half years for diplomas. Meanwhile, the longest time for a bachelor's degree is 14 semesters (seven years) and ten semesters (five years) for diploma programs.

The study period and the timely graduation rate of students are aspects of the assessment of the study program by the National Accreditation Board for Higher Education (BAN-PT). The late graduation of students indicates the institution's performance is not good because the ratio of the number of lecturers to students is not balanced due to the addition of previous students. Various factors cause late student graduation, for example, the number of repeating courses and the limited number of semester credit units (SKS) for students due to poor GPA[2]. The problem is that it is difficult to determine which students will not graduate on time, because of the large number of students handled by one lecturer and the small number of human resources (in this case, study programs and lecturers at the university.

The growth of information technology is fast, especially data mining. Data mining is the process to find interesting patterns from large amounts of data stored in databases[3]. Data mining is the process of finding models and knowledge from large amounts of data. Data mining sources include databases, data warehouses, the web, and other information storage places[4]. Data mining has a wide range of applications in various fields, including marketing, banking, educational research, surveillance, telecommunications fraud detection, and scientific discovery[5]. Educational data mining is used to explore and analyze student academic performance, predict student dropout, feedback analysis, visualize data, assess the learning process, and include predictions of student study duration. Educational data mining assesses the students' performance using several factors such as personal, social, economic, and other environmental information [6][7]. Several classification algorithms have been applied to predict student performances, such as Decision Tree, Neural Network, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machines (Hussain et al., 2018). Meanwhile, this research will search for a model that can classify student graduation patterns using the classification Data Mining technique by comparing algorithms and Neural Networks, Random Forests. Classification is used by evaluating academic outcomes in an achievement index in the first semester of lectures and using student demographics as attributes used in the dataset. The student classification aims to help decision-makers determine the strategy for preventing late graduation.

One of the requirements of the graduate Science, Engineering and Technology courses is that you conduct research and write a research paper on some aspects of software engineering. The paper may present original work, discuss a new technique, provide a survey and evaluation of recent work in a given area, or give comprehensive and taxonomic tutorial information. The paper must emphasize concepts and the underlying principles and should provide authentic contribution to knowledge. If your paper does not represent original work, it should have educational value by presenting a fresh perspective or a synthesis of existing knowledge.

II. EDUCATIONAL DATA MINING

The fundamentals of this reserach are theoretically based on data mining. Firstly, data mining finds patterns in data sets by the process of applying models and group the data [8]. The model is a technique and algorithm applied to the data to determine the similarity of the pattern and group the data. General models of data mining are classification, association rules, and clustering[9]. The process of knowledge discovery from data begins with data cleaning to remove noise and inconsistent or irrelevant data. Second, data integration is the data merging from various databases into one new database. Then, data selection and data transformation into a suitable format, are to be processed in data mining. After that, perform data mining and pattern evaluation to identify interesting patterns in the knowledge-based that are found. Finally, the presentation of knowledge is a visualization and presents knowledge about the methods used to obtain the knowledge obtained by the user. From the description of Knowledge Discovery from Data, it is concluded that the 1st to 4th stages are a form of preprocessing data. Interesting patterns are presented to the user and can be saved as new knowledge in the knowledge base.

Educational Data Mining (EDM) is an emerging discipline with a series of computational and psychological methods as well as a research approach to understanding how students learn [10]. EDM is a data mining concept that is used to explore educational data to find out descriptive patterns and predictions that characterize student behavior and achievement, domain content knowledge, assessment, educational functions, and educational applications[11]. EDM is widely used in education to explore and analyze student academic performance, predict dropouts, and analyze feedback, data visualization, and assessment in the learning process[12]. In other words, EDM is a field of learning in data mining. The availability of educational data is important to find out descriptive patterns and predictions and understand how students learn.

Classification is a process of determining a model or function that describes and distinguishes data classes or concepts[13] to predict categories (discrete, unordered) into the class label of an object whose label is unknown[14]. In classification, a collection of records is in the form of a training dataset, where each record contains a set of attributes, and one of the attributes is a class. Then, look for a model for the class attribute as a function of the values of the other attributes to get a class that is as accurate as possible from the previously invisible records. A training dataset is prepared to determine the accuracy of the model and, at the same time its validation document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

III. RESEARCH METHODOLOGY

The main methodology used in this research is based on Cross-Industry Standard Process for Data Mining (CRISP-DM) model. Figure 1 shows the steps that will be carried out in this research.

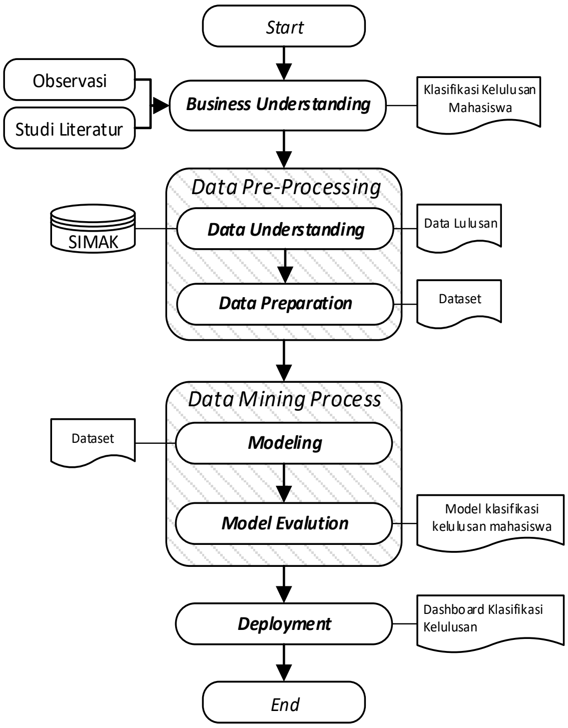

Figure 1: Research Methodology

IV. RESULT AND DISCUSSION

The model used inthis research is CRISP-DM model. The model provides a standard process for conducting data mining as a general problem-solving business or research unit as described in Figure 2.

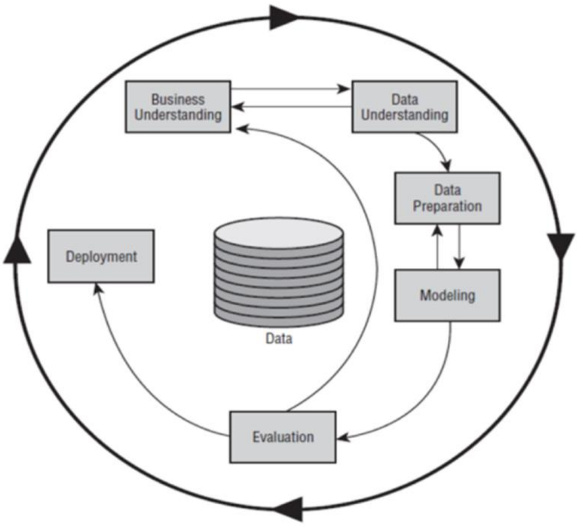

Figure 2: Research Methodology

The CRISP-DM process can be sequenced into the following steps;

1) Business Understanding focuses on understanding project objectives and requirements from a business perspective and transforming business knowledge into data mining problem definitions and initial design plans to achieve business objectives. The project requirement used is a graduation scheme at a faculty at the university, namely compiling a final project.

2) Data understanding begins with collecting data, if the data comes from more than one source, then the data integration process is carried out. Further identifying the quality of the data, developing an investigative analysis of the data to recognize further data, and seeking first insights into thedata or to detect subsets of interest to formhypotheses about hidden information. This study uses demographic data and data on studentacademic results in one data from graduates from the 2013 to 2018 years which consists of the Accountingprogram (diploma), Accounting Study Program (bachelor) and Management Study Program (bachelor). The data is taken by the Faculty Information System Unit in the form of a worksheet file (excel format *.xlxs). In addition, the data that will be used in this study is only data that is under the responsibility of the Faculty Information System Unit. Demographic and academic data are personal data of economy faculty students contains personal data of students who have graduated from the 2013 to 2018 batch with a total of 1467 data records and 25 attributes.

3) Data Preparation, includes all activities to build the final dataset from the initial raw data. Data preparation tasks may be performed multiple times and not in a specified order. Tasks include tables, records and attribute selection as well as transformations and data cleaning for use in the modeling phase. Study Program: this attribute is inputted with the number 021 for Diploma of Accounting Study Program, 022 for Bachelor of Accounting Program and 023 for Bachelor of Management Study Program. To distinguish it from a numeric format, it is adjusted by replacing 021 with the AD3 code, 022 with the AS1 code and 023 with the MS1 code.

4) Modeling through the modeling phase begins with selecting and applying appropriate data modeling techniques. Then, calibrate the model rules to optimize the results. Next, use some of the same techniques for the same problem, and return to the data preparation phase.

5) Evaluation evaluates one or more models used in the modeling phase to determine whether the model or pattern used is relevant to the business objectives in the initial phase.

6) Deployment and data grouping can adjust to user needs and can perform the Data Mining process repeatedly. The form of this phase can be in the form of reporting or implementing the Data Mining process in other departments. information gain value for each attribute. In doing the modeling, this research was assisted by using Orange Software.

Table 1 shows the recapitulaion of this research. Determination of the model refers to the value of accuracy, recall, precision, or f-measure. The high accuracy value indicates the performance of the algorithm has a very close number of False Negative and False Positive data (Symmetric). A high recall value is very good if preferring False Positive is better than False Negative.

A high precision value is very good if we want a True Positive and don't want a False Positive to occur. The f-measure value is considered good. The f-measure value is the average of precision and recall values. In classifying students, it is highly expected that the resulting value will have a False Positive value or pass on time. The results of the classification can be used as a warning to the guardian lecturer or the head of the study program to take action against students classified as graduating on time.

TABLE 1. FONT SIZES FOR PAPERS

Skenario	Sampling (Fold)	Algoritm	Model	Accuracy (%)	Recall (%)	Precision (%)	F-Measure
1	5	NN	1	66.80	66.70	66.50	66.80
		RF	2	73.40	71.30	71.20	73.90
	10	NN	3	68.00	67.80	67.70	68.00
		RF	4	74.00	71.90	71.90	74.00
	20	NN	5	66.70	66.80	66.80	66.70
		RF	6	73.30	71.70	71.40	73.30
2	5	NN	7	66.60	66.60	66.60	66.60
		RF	8	73.10	71.20	71.00	73.10
	10	NN	9	66.40	66.40	66.50	66.40
		RF	10	73.90	72.00	71.90	73.90
	20	NN	11	65.90	66.20	66.50	65.90
		RF	12	74.10	72.10	72.10	74.10
3	5	NN	13	66.10	66.10	66.20	66.10
		RF	14	74.70	72.70	72.80	74.70
	10	NN	15	67.30	67.10	67.00	67.30
		RF	16	74.10	72.10	72.10	74.10
	20	NN	17	66.20	66.20	66.20	66.20
		RF	18	74.10	72.30	72.20	73.10
4	5	NN	19	66.20	66.10	65.90	66.20
		RF	20	73.90	72.00	71.90	73.90
	10	NN	21	65.40	65.40	65.40	65.40
		RF	22	73.80	72.00	71.90	72.70
	20	NN	23	68.00	67.90	67.70	68.00
		RF	24	73.20	71.60	71.20	73.20

Therefore, it can be concluded that the recommended model is the model with the best recall value as the model that chooses False Positive is better than False Negative. The model uses a dataset for two attributes with the lowest Information Gain value are deleted (Dadup's Life and Origin). The sampling using Cross-Validation with the number of Fold 5 with a recall value of 72.70%, even for the value of accuracy, Precision, and F-measurement this model produces superior values compared to others, namely 74.70% for Accuracy Performance, 72.80% for Precision Performance and 74.70% for F-measure Performance.

V. CONCLUSION

In the test scenario, several factors affect student graduation, namely GPA2, GPA1, Study Program, Mother's Occupation, Gender, Program ID, Mother's Education, Father's Education, Father's Occupation, AdmissionAge, CivilStatus, and Mother'sLife. The accuracy value in the Random Forest algorithm dataset is better for classifying student graduation with Recall Performance scores of 72.70%, 74.70%, Accuracy Performance, 72.80%, Precision Performance, and 74.70% F-measure Performance. Therefore, the conclusion is the values of accuracy, recall, precision, and f-measure in all datasets of the Random Forest algorithm have a higher value than the Neural Network. For further research we need more data for calculation in more detail, and it also can be analyzed and explored another method to find better result.

VI. REFERENCES

[1] E. C. Ploutz. 2018. Machine Learning Applications in Graduation Prediction at the University of Nevada, Las Vegas. https://doi.org/10.34917/13568668

[2] N. Azwanti. 2018. Algoritma C4.5 Untuk Memprediksi Mahasiswa Yang Mengulang Mata Kuliah (Studi Kasus Di AMIK Labuhan Batu). Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer, 9(1).

[3] A. A. N. Mostafa. 2015. Review of Data Mining Concept and its Techniques Factors Affecting Acceptance of Mobile Banking in Developing Countries View project Factors Affecting Acceptance of E-commerce by SMEs in Libya View project Review of Data Mining Concept and its Techniques Introduction. https://doi.org/10.13140/RG.2.1.3455.2729.

[4] G. E. Sakr, I. H. Elhajj, I. H., Mitri, G., & Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction. IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM, 1311–1316. https://doi.org/10.1109/AIM.2010.5695809

[5] J. Han, M. Kamber, J. Pei. 2012. "Data mining: Data mining concepts and techniques (3rd ed.)". Morgan Kaufmann Publishers.

[6] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, F. M., & Ribata, N. 2018. Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447–459.

[7] A. M. Shahiri, W. Husain, & N. A. Rashid. 2015. A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422.

[8] J. Leskovec, A. Rajaraman, & J. D. Ullman. 2014. Mining of Massive Datasets. In Mining of Massive Datasets. https://doi.org/10.1017/cbo9781139924801.

[9] R. Boire. 2014. Applying Data Mining Techniques. Data Mining for Managers, March, 95–113.https://doi.org/10.1057/9781137406194_12

[10] R. S. Baker, J. 2010. Data Mining for Education Data Mining for Education Advantages Relative to Traditional Educational Research Paradigms. International Encyclopedia of Education, 7(3), 112–118.

[11] A. Peña-Ayala. 2014*. Educational Data Mining Applications and Trends*. Springer.

[12] D. Sugianti. 2012. Algoritma Bayesian Classification untuk Memprediksi Heregistrasi Mahasiswa Baru di STMIK Widya Pratama. Jurnal Ilmiah ICTech, 10(2), 1–5.

[13] E. N. Wahyudi. 2013. Teknik Klasifikasi untuk Melihat Kecenderungan Calon Mahasiswa Baru dalam Memilih Jenjang Pendidikan Program Studi di Perguruan Tinggi. Jurnal Teknologi Informasi DINAMIK, 18(1), 55–64.

[14] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth. 2000. *CRISP-DM Step-by-step Data Mining Guide*. SPSS Inc.

Cite this article as :

Riyad Sabilul Muminin, Ana Hadiana, Nila Natalia, "The Study of Neural Network Algorithm, Random Forest for Classification of Student Graduation", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 3, pp. 517-522, May-June 2023. Available at doi : https://doi.org/10.32628/IJSRSET23103145
Journal URL : https://ijsrset.com/IJSRSET23103145