13th SCBTII
Sustainable Collaboration in Business, Technology, Information and Innovation

Telkom University

Recover Together, Recover Stronger: Challenges and Opportunities Toward Sustainable Digital Economy

PROCEEDING BOOK 2022

e-ISSN: 2621-3192

Organized By

Fakultas Ekonomi & Bisnis
School of Economic & Business
Telkom University

University Partner

Proceedings of International Conference on Sustainable Collaboration in Business, Technology, Inforation and Innovation (SCBTII)

Third Edition, December 2022

Vol 3. Year 2022

e-ISSN: 2621-3192

Chief Editor
Yuhana Astuti, S.Si.,S.E., M.T., M.Agr., Ph.D

Editor Team
Dr. Putri Fariska S,Si., M.Si
Dr. Nidya Dudija, S.Psi., M.A.
Hani Gita Ayungningtias, S.Psi., M.M
Dian Putri Rahmdhani, S.M., M.M

Design By
Indra Gunawan, S.Kom

Publisher
School of Economics and Business, Telkom University

Editorial Address

Jln. Telekomunikasi, Terusan Buah Batu, Bandung 40257, Indonesia Telp: +62 22 7503621
Website: seb.telkomuniversity.ac.id
Email : seb@telkomuniversity.ac.id

ORGANIZED BY

Universitas Telkom

SUPPORTED BY

UNIVERSITY PARTNER

Neural Network Algorithm For Classification Of Student Graduation In Faculty of Economics, University of Garut

Fikri Fahru Roji[1], Agna Hilyah[2], Ridwan Setiawan[3], and Diqy Fakhrun Shiddieq[4]
[1,2,4] Faculty of Economics, University of Garut, Garut, Indonesia
[3]Information Technology, Garut Institute of Technology, Garut, Indonesia

Abstract
The academic performance is one of aspect which has remained the benchmark of the success in learning activities at the 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, Garut University. In this research explore the model that able to classify the graduation of student through the data mining classification technique using Neural Network Algorithm. The classification conducted by evaluating the academic performance based on Semester Performance Index (IPS) during first years, Semester Credit Unit (SK S) at second years and use the demographics of students as attributes that will be used in the dataset. Based on the examinations that conducted by using k-fold cross validation, there are 12 attributes that influence the graduation of students. Moreover, based on the evaluation of accuracy and recall score, with some modified number of hidden layers, and number of hidden nodes and removing some features will indicated the high value. The model that represented able to applied for classify the active students from 2012 to 2017 that indicate accuracy value of 82.96% and recall value was 66.54%.

Keywords— Data Mining; Student Graduation; Classifications; Neural Network; K-Fold Cross Validation; Confusion Matrix.

I. INTRODUCTION

Academic performance is an aspect to measure the success of universities in carrying out learning activities. A good indicator of academic performance for tertiary institutions is that students can complete the study period with a predetermined minimum time limit. Problems regarding student academic performance in terms of accomplishing the study period on time occur in the Faculty of Economics, University of Garut. Table 1 shows the percentage of graduate students from the economics faculty who accomplish their study period on time from the 2012 – 2017 class. This case makes the ratio between lecturers and students not ideal. In other words, the value of accreditation of study programs regarding student affairs is not optimal, and too many active students lead teaching and learning activities to be less conducive. The obstacles occur in determining the indicated students will not graduate on time because each lecturer guides too many students, and the number of human resources in the study program does not match the number of students.

Table 1. Percentage of Graduates of the Faculty of Economics on time in 2015-2021

Period	Accounting Department	Management Department
2012	5,7%	6,2%
2013	14,5%	7,7%
2014	13,5%	4%
2015	26,9%	27,7%
2016	8%	5,7%
2017	10,3%	23,9%

The rapid growth of information technology creates the accumulation of data but lacks information. In the process of browsing for that information, the term Data Mining appears. Data mining is regarded as the process of finding interesting patterns and knowledge from data collections of databases, data warehouses, data marts, and repositories (Han et al., 2012) that focus on extracting hidden knowledge (Hussain et al., 2018). Educational Data Mining (EDM) is widely used in education to explore and analyze student academic performance, predict dropouts, feedback analysis, data visualization, and assessment in the learning process. The classification method is the method in Data Mining that is often operated in EDM to predict student performance (Shahiri et al., 2015).

This study aims to find a model for classifying student graduation patterns using the Data Mining classification technique by the Neural Network algorithm. Classification is used by evaluating the output of student academic performance in the form of social studies in the first year and the number of credits in the second year and using student demographics as a feature to be used in the dataset. The methodology for designing Data Mining uses the CRISP-DM with the model performance test method through Cross Validation. The model from the classification implemented on a graph and a list of students classified by graduation time is a data visualization. The classifications of student graduation support decision-makers in taking the necessary actions if the student's study period can be known earlier.

II. LITERATURE REVIEW

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 (Baker, 2010). EDM is a Data Mining concept that is used to explore educational data, identify patterns and predict descriptively what characterizes student behavior and achievement, domain content knowledge, assessment, educational functions, and educational applications (Peña-Ayala, 2014). EDM analyzes data generated by various types of information systems that support learning or education. Data in education have distinctive characteristics such as multiple hierarchical levels (subjects, assignments, question levels), context (certain students in a particular class encountering certain questions at specific times and on specific dates), and longitudinal (lots of data recorded over many sessions in a while) (Romero & Ventura, 2013).

(Shahiri et al., 2015) conducted a review of studies related to the implementation of Data Mining to predict the study period and produced a review of the attributes in the dataset used in predicting graduation using Data Mining and the algorithm used. The results of this study suggest that the Neural Network algorithm is the most popular and has a good level of accuracy for predicting student graduation. In addition, the Neural Network algorithm is a suitable algorithm for classifying student graduation after a comparison experiment has been carried out between Neural Network, Support Vector Regression, and Linear Regression (Obsie & Adem, 2018)

This study uses the CRISP-DM methodology, referring to (Shafique & Qaiser, 2014) who compared several methodologies in the development of Data Mining and concluded that the Data Mining project mostly uses CRISP-DM and is considered quite complete. Classification of data that has accuracy or similarity to measurement data with actual numbers or data, testing is carried out based on accuracy with test data parameters and training data determined by Cross-Validation. Therefore, data accuracy affects the final results of data classification because when data accuracy is not accurate it will affect the percentage of grouping test data and training data (Hulu et al., 2020) Cross validation is a standard procedure for evaluating performance classification algorithms (Cao et al., 2007) and has been extensively used in data mining for the sake of model selection or modeling procedure selection (Hastie et al., 2009) given a priori or developed by modeling procedures. Based on data splitting, part of the data is used to fit each competing model and the rest is used to measure the predictive performance of the model with validation errors, and the model with the best overall performance is selected (Zhang & Yang, 2015).

III. RESEARCH METHODOLOGY

The methodology used in this study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) method. As depicted in Fig.1 CRISP-DM consists of six phases, namely (Chapman et al., 2000): Business Understanding; Data Understanding; Data Preparation; Modeling; Evaluation and Deployment.

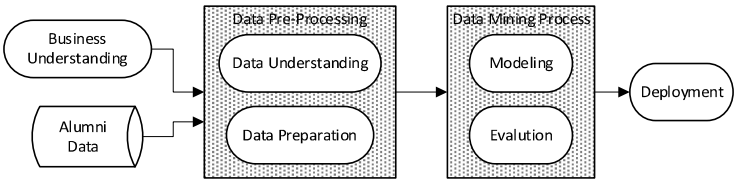

Fig. 1. Methodology Process

A. Business Understanding

The purpose of implementing this data mining is to classify students who graduate on time and graduate not on time. This information is used by guardian lecturers (mentors) and study programs to understand the characteristics and provide special treatment to their students such as suggesting taking a short semester and

several other treatments based on the policy of the study program. Institutions expect graduation classification data to increase the number of students who complete their studies on time.

B. Data Understanding and Preparation

This study uses data from graduates from an academic information system that runs within the economics faculty from the class of 2012 to 2017. The data taken is in the form of a worksheet file with 20 features for 2022 data. The data is raw data with data conditions that do not meet the standards to enter the mining process. In the next stage, the researcher selects features by deleting several features that do not influence the Data Mining process. Many factors influence the data mining process, namely accuracy, completeness, consistency, timeliness, trustworthiness, and interpretability (Han et al., 2012). Next, check the data to see the mistakes, inconsistencies, and noise that affect the modeling results. In addition, cleaning data by handling missing values in the dataset, then assigning a constant value, middle value, or average of the feature or giving a lot of values for the feature (Han et al., 2012). Next, look for the information gain value for each feature in the dataset. Researchers use information gained as a reference for modeling scenarios. Table 2 describes the features that will be used after going through the data preparation stage.

Table 2. Description of the features of the data used

Attribute	Description	Values	Information Gain
PROG	Program reg	Reg, Non Reg	0.004
GND	Gender	Male, Female	0.011
AGE	Age when	16 - 38	0.002
MRG	Marriage status	Marriage, Not Marriage	0.002
LOC	Local city	Local, Out Of Town	0.023
FTSCH	Father last school	SD, SMP, SMA, S1, S2, S3	0.042
MTSCH	Mother last school	SD, SMP, SMA, S1, S2, S3	0.049
FTJOB	Father job	gov emp, soldier, priv emp, entpr, laborer, not, etc	0.048
MTJOB	Mother job	gov emp, soldier, priv emp, entpr, laborer, not, etc	0.001
IP1	Semester 1 Performance Index	0,0 – 4,0	0.098
IP2	Semester 2 Performance Index	0,0 – 4,0	0.112
IP3	Semester 3 Performance Index	0,0 – 4,0	0.118
IP4	Semester 4 Performance Index	0,0 – 4,0	0.083
SKS3	Semester 3 Credit Unit	0 – 24	0.159
SKS4	Semester 4 Credit Unit	0 – 24	0.095
LULUS		On time / Not on time	-

C. Modeling

The modeling in this research is using Neural Network. Determination of train data and testing data in this modeling process uses the Cross Validation method. The dataset has a total of 1304 records. In finding the comparison value,the researher look for the resulting accuracy and recall values. Features are sorted by information gained. The test scenario of this model uses all the features in the first scenario with a different number of hidden layers and hidden nodes, then deletes the feature with the lowest information gain value in the next experiment. In doing the modeling, this research was assisted by using Rapidminner software.

IV. RESULT/FINDING

A. Evaluation

Evaluation of models at the modelling stage to find out which model is the best in classifying student graduation. The comparison appears in the number of hidden layers, the size of each hidden layer, and features. The way to evaluate the model is to analyse the accuracy and recall values generated by each model. Table 3-6 is a recapitulation of the performance of all models.

Table 3. Classification results using all attributes

Table 4. Classification results by removing 1 attribute with the lowest information gain value

Table 5. Classification results by removing 2 attributes with the lowest information gain value

Table 6. Classification results by removing 3 attributes with the lowest information gain value

Test	*Accuracy*	*Recall*	Test	*Accuracy*	*Recall*	Test	*Accuracy*	*Recall*	Test	*Accuracy*	*Recall*
M1	81,11	61,52	**M31**	79,71	57,62	**M61**	80,33	59,85	**M91**	80,95	57,81
M2	80,64	55,76	**M32**	80,69	52,23	**M62**	81,31	55,39	**M92**	80,79	57,43
M3	80,9	53,9	**M33**	80,69	54,65	**M63**	80,59	61,34	**M93**	81,11	56,69
M4	80,44	59,29	**M34**	82,03	56,32	**M64**	81,98	55,58	**M94**	80,8	65,24
M5	81,26	60,59	**M35**	82,29	63,57	**M65**	82,08	60,59	**M95**	82,96	66,54
M6	81,16	57,43	**M36**	80,44	55,02	**M66**	81,16	58,92	**M96**	79,97	59,85
M7	80,85	57,81	**M37**	80,95	57,43	**M67**	80,59	54,28	**M97**	81,16	57,81
M8	80,95	52,04	**M38**	81,52	64,31	**M68**	81,47	60,04	**M98**	82,03	55,39
M9	81,78	58,74	**M39**	81,37	58,55	**M69**	80,54	59,29	**M99**	81,16	56,13
M10	81,93	59,48	**M40**	82,34	57,06	**M70**	81,36	58,55	**M100**	82,34	60,59
M11	80,8	57,81	**M41**	80,18	53,16	**M71**	80,85	54,09	**M101**	79,19	54,65
M12	80,79	61,71	**M42**	80,9	58,18	**M72**	80,64	60,41	**M102**	79,14	54,46
M13	80,43	62,64	**M43**	81	61,71	**M73**	82,6	58,92	**M103**	81,52	65,8
M14	82,09	61,9	**M44**	81,73	56,69	**M74**	81,36	65,61	**M104**	81,1	58,18
M15	82,39	61,9	**M45**	81,88	59,67	**M75**	80,79	61,52	**M105**	81,57	61,52
M16	81,05	59,11	**M46**	79,51	48,14	**M76**	80,7	59,11	**M106**	80,07	61,52
M17	80,85	60,04	**M47**	79,71	55,2	**M77**	80,85	60,04	**M107**	81,05	60,59
M18	80,84	57,62	**M48**	80,69	55,39	**M78**	81,72	61,71	**M108**	80,95	62,08
M19	81,88	53,9	**M49**	80,02	58,55	**M79**	81,52	59,67	**M109**	81,78	63,38
M20	81,46	58,92	**M50**	80,95	59,67	**M80**	81,62	60,59	**M110**	82,8	60,04
M21	81,57	64,31	**M51**	80,02	58,92	**M81**	80,54	61,34	**M111**	80,59	56,69
M22	80,43	57,06	**M52**	80,38	56,32	**M82**	79,35	55,95	**M112**	80,9	54,46
M23	81,52	63,2	**M53**	80,64	58,55	**M83**	81,62	62,08	**M113**	81,72	60,41
M24	81,32	56,88	**M54**	81,42	56,32	**M84**	81,21	62,08	**M114**	81,26	60,22
M25	82,08	60,97	**M55**	80,79	63,94	**M85**	81,67	62,64	**M115**	82,6	62,64
M26	80,95	64,68	**M56**	79,15	49,44	**M86**	80,02	52,97	**M116**	80,64	57,43
M27	80,74	63,75	**M57**	80,85	59,11	**M87**	80,85	59,85	**M117**	80,28	53,35
M28	80,84	59,85	**M58**	81,21	54,28	**M88**	82,09	61,71	**M118**	80,69	61,9
M29	80,74	56,69	**M59**	81,88	60,59	**M89**	81,31	62,45	**M119**	82,24	62,08
M30	81,93	57,99	**M60**	80,79	61,15	**M90**	80,85	61,34	**M120**	82,24	58,74

Accuracy shows the ratio of True positives (true positives and true negatives) of the overall data. Accuracy answers the question about the percentage of students who are predicted to pass on time and graduate on time from all students. Recall indicates the ratio of true positive predictions compared to all data that is true. Recall answers the question about the percentage of students is predicted to graduate on time compared to the total number of students who graduate on time.

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 are 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. 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. Based on the test results, the highest recall value is in the M95 model, where the model has 12 features (PROG, GND, LOC, FTSCH, MTSCH, FTJOB, IP1, IP2, IP3, IP4, SKS3, SKS4) with a recall value of 66.54 %, besides that the M95 model has a fairly high accuracy value, which is 82.96%.

B. Deployment

Data collection for students of 2019-2020 who have not passed and meet the feature criteria used in the recommendation model. The number of data is 999, consisting of 423 2019 students and 576 students from the 2020 batch. Furthermore, the data preprocessing process is carried out and leaves as many as 809 data (436 data from the 2019 batch and 373 from the 2020 batch). Table 7 is a recapitulation of the graduation classification of 2019 and 2020 students of the Faculty of Economics. Fig 2 shows the distribution of graduation classification data by batch.

Table 7. Recapitulation of the graduation classification of 2019 and 2020

Graduation Prediction	Amount	Fraction	Percentation
On Time	205	0,756	24%
Not On Time	636	0,244	76%

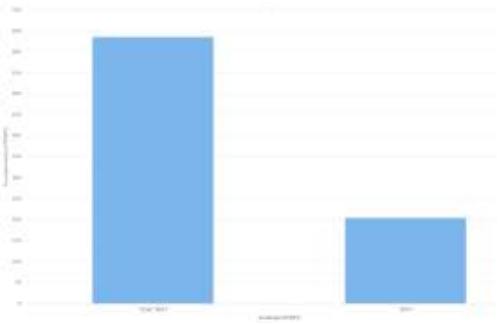

Fig 2. Distribution of graduation classification data by batch

V. DISCUSSION

From the results of research conducted at the Faculty of Economic Garut University, it can be concluded that:

a. The predicted results show results that illustrate that student graduation tends to be late because the dominant training data used has a late trend
b. The number of hidden layers and the number of hidden nodes greatly affect the accuracy and recall of the neural network algorithm.
c. From the results of the research conducted and the results of the application of the neural network algorithm, the most influential parameter is the student's SKS3 (Third Semester Credit Unit).
d. The results of the prediction of the application of the neural network algorithm to the prediction of graduation of management students at the Faculty of Economic University of Garut with a sample from 2012 to 2017 batches were 82.96% for accuracy and 66.54% for recall.

With the results obtained from this study, the accuracy can be improved by reducing the number of parameters and increasing the number of input nodes and layers in the next study. The discussion and application of a comprehensive feature model are expected to increase the accuracy value which is getting closer to 100% in subsequent studies.

VI. CONCLUSION AND RECOMMENDATION

The finding concluded that student graduation tends to be late because the dominant training data used has a late trend. In addition, the number of inputs such as input nodes, hidden layers, and hidden nodes greatly affect the accuracy of neural network algorithms. The results of the neural network algorithms illustrated the most influential parameter is the SKS3 (Third Semester Credit Unit) of students. The calculation of predictions of the accuracy of the application of neural network algorithms to the graduation predictions of the Faculty of Economic University of Garut with batch samples from 2011 to 2017 was 98.27%. The students who are predicted to not graduate on time, FEKON UNIGA, in this case, the study program and the student guardian lecturer, perform several treatments such as suggesting taking a short semester, more routine counseling, and other treatments based on the policy of the study program.

Furthermore, the accuracy and recall improve by reducing the number of features and increasing the number of hidden layer layers and nodes in subsequent studies. Discussion and implementation of comprehensive feature models are expected to increase the accuracy value closer to 100% in future studies.

REFERENCES

Baker, R. S. J. (2010). Data Mining for Education Advantages Relative to Traditional Educational Research Paradigms. *International Encyclopedia of Education*, *7*(3).

Cao, L., Yu, P. S., Zhang, C., & Zhang, H. (2007). Data Mining for Business applications. In *The Multimedia Handbook*. Springer.

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

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Editio). Springer.

Hulu, S., Sihombing, P., & Sutarman. (2020). Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data. *International Journal of Research and Review*, *VII*.

Hussain, S., Dahan, N. A., 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.

Obsie, E. Y., & Adem, S. A. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications*, *180*(40), 39–47.

Peña-Ayala, A. (2014). *Educational Data Mining Applications and Trends*. Springer.

Romero, C., & Ventura, S. (2013). Data Mining in Education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222.

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

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112.