JOURNAL OF BIG DATA SCIENCE

P-ISSN 3025-6615

Volume 1 Number 1
September 2023

Literature Review: Data Mining For Student Data Classification

Billi Mahardika[1]*

[1]Laboratory Staff, Universitas Muhammadiyah Palembang
*billymahardika123@gmail.com

Ahmad Yani street No. 13, Palembang, South Sumatra 30263, Indonesia

Received September 25th, 2023; **Revised** September 29th, 2023; **Accepted** September 30th, 2023

Abstract

The abundance of student data and student graduation number data, hidden information can be found by processing student data to be useful to the university. The processing of student data needs to be done to uncover important information in the form of new knowledge (knowledge discovery) such as information on student data classification based on profile and academic data. Therefore, in this research, the researcher plans to conduct a literature review related to data mining for student data classification with the aim of finding out about data mining data processing classification and collecting all designs used in identifying data starting from problems, methodology, equations and results. For this research, researchers used historical data from students from 2007 to 2011 who had graduated. There were 9 research journals that researchers managed to find, each of which used different algorithms or classification techniques. To conduct a literature review, researchers conducted a journal review using PICOT. The results of this research are the success of researchers in classifying student data using data mining techniques.

Keywords: *Classification, Students, Data Processing.*

1. INTRODUCTION

Higher education today must gain a competitive advantage using every available resource. In addition to infrastructure and human resources, information systems are one of the resources that can be used to collect, process and disseminate information to support day-to-day operations as well as outsourcing activities. strategic decisions. Education has a very important role to improve and prepare superior and highly competitive human resources (HR). This is where the role of higher education institutions becomes very important in creating experts who are able to develop knowledge and contribute to development. Higher education as one of the educational institutions, is required to be able to provide quality and quality education to its stakeholders[1]. The abundance of student data and student graduation count data, hidden information can be found by processing student data to be useful to higher education institutions. The processing of student data should be performed to uncover important information in the form of new knowledge (knowledge discovery), for example information on classification of student data by profile and academic data. This new knowledge could help universities rank student graduation rates to determine strategies to increase graduation rates in later years. Data mining is a series of processes that aim to manually extract added value in the form of unknown information from a database by extracting patterns from data with the aim of manipulating the data to obtain More valuable insights by extracting and recognizing important patterns. or dig into the data contained in the database [2].

In relation of Data Mining, there are several previous research that have become references for researchers. The first is research from Cornelia Selvi Diana, Latifah Hanum and Saut Parsaoran Tamba, who in this research implemented data mining using the K-means algorithm to determine the title of the thesis and research journal by making FTIK UNPRI the research object. The results of this application itself certainly provide convenience or solutions for students and their scope to find out ideas for thesis titles and research journals. The second is research conducted by Ni Luh Putu Purnama Dewi, I Nyoman

Purnama and Nengah Widya Utami. In this research, data mining was applied to cluster lecturer performance assessments using the K-means algorithm at STMIK Primakara. The results obtained from this research are research on lecturer performance based on student satisfaction, namely very good cluster 312 (31.74%) student data, good cluster 401 (40.79% student data, quite good cluster 189 (19.23%) student data and the less good cluster was 81 (8.24%). The DBI accuracy level was 0.270 or 27%. The last research was conducted by Hozairi, Anwari and Syarifu Alim, in this research implementing orange data mininh to classify student graduation using a model. K-Nearest Neighbor, Decision Treee and Naive Bayes in the Informatics Engineering Study Program, Madura Islamic University, class of 2016. In this study, a comparison of classification models was carried out with the results of K-Nearest Neighbor having an accuracy level of 77%, Decision Tree with an accuracy level of 74%. % and Naive Bayes is 89%, thus the most recommended model is Naive Bayes.

From these three research, it can certainly be seen that there are many classification models or classification algorithms for data mining. So in this research, researchers will conduct a literature review related to data mining for student data classification with the aim of finding out about data mining classification for data processing and collecting all designs used in identifying data starting from problems, methodology, equations and results. For this research, researchers used historical data from students from 2007 to 2011 who had graduated. Regarding the classification algorithm consists of 6 parts consisting of: Decision Tree Analysis, is a technique that belongs to the machine-learning family, arguably the most popular classification technique in the data mining area. Statistical Analysis [3] Statistical engineering was the source of popular classification algorithms for many years until the emergence of machine learning techniques. Statistical classification techniques include logistic regression and discriminant analysis, both of which assume that the relationship between input and output variables is essentially linear, that the data are normally distributed, and that the data are normally distributed. The variables are neither interdependent nor independent of each other. The nature of these questionable assumptions eventually led to the shift towards machine learning techniques. Neural Networks is one of the most popular techniques in Machine-Learning that can be used for classification problems. Case-Based Reasoning, this approach uses historical cases to recognize similarities to determine a new case into the most probable category. Bayesian Classifiers, this approach uses probability theory to create classification models based on past events that can place a new instance into the most probable class. Genetic Algorithms, using analogies to natural evolution to create a purposeful search-based mechanism for classifying data samples.

2. RESEARCH METHODOLOGY

In this research, the researcher conducted a literature review, in which case the researcher reviewed journals that matched the PICOT and search terms for journals through MESH (Medical Subject Heading), limitations taken by journals and other things. The journal is used in literature reviews obtained through databases that provide Indonesian scientific journals through Google Scholar and websites such as Garuda Kemdikbud.

The researcher wrote down the keywords according to the MESH (Medical Subject Heading) namely "processing", "data", "students" and selected the full text to appear 100 findings, then narrowed down to Dissertations and Theses and found the next 9 findings sorted from the most recent. Regarding the choice of language, it was not carried out because all the journals found used Indonesian. Each of these questions has followed the PICOT where in each of these questions there is P = Problem/Student/Population, I/E= Implementation /Intervention/Exposure, C=Control/Comparative Intervention, O=Results and T=Time. It is relevant that the author used to get journals about the classification of Data Mining processing student data. The author takes all the designs in the research that are used in identifying student data

3. RESULTS AND DISCUSSION

In this research, using history data from students from 2007 to 2011 who have graduated with a total of 377 data with 72 attribute values of the course and 1 target class in the form of study period. This research was conducted by following the stages of data mining work which refers to the process of knowledge discovery in databases. The data mining application was successfully built with experimental results showing that the best study period classification results were obtained by selecting attributes

from all elective courses with accuracy values. Based on data obtained from the Department of Computer Science/Informatics, Diponegoro University, students from the 2007 to 2011 class year with a total of 377 graduates obtained information that the average student study period is still over 4 years. Inna Alvi Nikmatum (2019) and Indra are alert. There are several data mining tools including Rapid Miner, Orange, KNIME, Weka, Keel and R. WEKA is GUI-based so that it minimizes the use of coding which can make it easier for system users. In this study using WEKA tools. The data processed at WEKA has ARFF and CSV formats. Based on the experimental results, it was concluded that using three model criteria, namely the Ratio Gain, Information Gain and Gini Index. The highest accuracy results are found in the Gini Index criterion model, namely 92.18%. The highest result of the three feature selections is the Information Ratio Gain with a value of p = 0.6 and the accuracy results are 92.46. Feature selection is the process of selecting the right features to be used in the classification or clustering process. The independent variable that gives the greatest t value is taken as X1 provided that H0 is rejected. The two taken from the independent variable are taken as X2 provided that H0 is rejected. Wiwit Supriyanti (2018), and Miss Puspitasari in the number of questionnaire results that were successfully collected were 981 samples. Of the 981 data collected, there are 948 valid data that can be used and 33 data that are invalid due to an unbalanced dataset, so the dataset is adjusted randomly to 360 data with an ouas value and 350 data with a dissatisfied value. The validity and readability tests were carried out on 30 randomly selected data samples, meaning that each statement given by the respondent was correlated with the total score and all were declared valid..

From the results of the implementation of the confusion matrix calculation, the accuracy output of the classification model for the three algorithms is obtained as shown in Table 1. The three algorithms, decision tree C4.5, SVM and Naïve Bayes, show quite good accuracy in classifying correctly both for satisfied classes and the dissatisfied class is quite good, namely above 80%, where the accuracy of the C4.5 algorithm and the SVM algorithm is better than the accuracy provided by the Naïve Bayes algorithm. These results are consistent with previous studies which claim that each algorithm has good performance on the dataset used. The results of the questionnaire show that 61% of students answered they were satisfied with the learning facilities and infrastructure and 39% answered they were dissatisfied. C4 Decision Tree Algorithm. 98% compared to the Naïve Bayes and Support Vector Machine algorithms when modeled from training data and tested using test data from the student satisfaction questionnaire dataset for learning facilities and infrastructure Elga Mariati, Ariesta Lestari, and Widiatry (2020).

Table 1 : Table *literature review*

No.	Writer	Research Title	Year	Method	Results
1.	Inna Alvi Nikmatum & Indra Waspada	Implementation of Data Mining for Classification of Student Study Period Using the K-Nearest Neighbor Algorithm	2019	K-Nearest Neighbor	Student data and student graduation data can produce abundant information, hidden information can be found by managing the data. Based on data obtained from the Department of Computer Science/Informatics, Diponegoro University, students from the 2007 to 2011 class year with a total of 377 graduates obtained information that the average student study period is still over 4 years [4]. [5]
2.	Marina Windarti & Agustinus Suradi	Performance Comparison of 6 Data Mining Classification Algorithms for Prediction of	2019	*Decision Tree, Bayesian Network, K-Nearest Neighbors,*	One of the factors affecting the quality of a higher education institution is student learning outcomes, which can be measured over time of study. The acquisition of knowledge in a database is often referred to as data mining or data mining. This study aims to determine the performance of the six classification algorithms used, which are Decision

		Student Study Period		*Naïve Bayes, Neural Network,* Svm	Trees C4.5, Bayesian Networks, Nearest Neighbors, Naïve Bayes, Neural Networks, and SVMs. The acquisition of knowledge in a database is often referred to as data mining. Decision Tree C4. Bayesian Network has the best performance with an accuracy value of 80.615%, precision and recall values of 0.785 and 0.806, for AUC values included in the good category, namely 0.837 [5].
3.	Wiwit Supriyanti & Nona Puspitasari	Implementation of Forward Selection Feature Selection Techniques in the Minig Data Classification Algorithm for Predicting Study Period of Indonusa Surakarta Polytechnic Students	2018	*K-Nearest Neighbor*	The experimental results concluded that using three criteria models namely gain ratio, information gain and Gini index, the highest accuracy results were found in the Gini index criterion model, namely 92.18%. The highest result of the three feature selections is the information gain ratio with a value of p=0.6 and the accuracy results are 92.46. Feature selection is the process of selecting the right features to be used in the classification or clustering process. The independent variable that gives the greatest t value is taken as X1 provided that H0 is rejected. The two taken from the independent variable are taken as X2 provided that H0 is rejected [6]
4.	Elga Mariati, Ariesta Lestari & Widiatry	Engineering Student Satisfaction Classification Model for Learning Facilities Using Data Mining	2020	*Decision Tree*	Of the 16 attributes used, only 8 attributes affect the level of customer satisfaction because they have a higher presentation of answers, the rest have relatively small presentations. The use of the naïve Bayes algorithm in this study resulted in an accuracy of 74.46% which stated that the amount of correct data was greater than the wrong data. Input attributes of customer satisfaction used in this study include price, facilities, service and loyalty. The implementation of the support vector machine algorithm gives a fairly good accuracy value of 80%. Each of the previous studies only implemented one algorithm, then stated that the algorithm used had given good results in terms of accuracy. The results of the questionnaire showed that 61% of students answered that they were satisfied with the learning facilities and infrastructure and 39% answered that they were not satisfied. C4 Decision Tree Algorithm. 98% compared to the Naïve Bayes and Support Vector Machine algorithms when modeled from training

					data and tested using test data from the student satisfaction questionnaire dataset for learning facilities and infrastructure [7] .
5.	Eka Sabna & Yuda Irawan	Data Mining With 2 (Two) Student Performance Prediction Classification Models	2021	*Naïve Bayes*,	The results of implementing data mining using the Rapidminer software were carried out on two classification algorithm models, namely C4.5 and NBC, then entering datasets as test material for the two models, which contained experimental data and test data. By using the Associa on Rule Method. in the Naive Bayes Classifier (NBC) algorithm model is 80% while in the C4.5 algorithm model it is 60%. It can be concluded that the best accuracy value from the results of the comparison of the two algorithm models is obtained by the Naïve Bayes algorithm model with an accuracy value of 80% [8].
6.	Resti, Dodo Zaenal Abidin, & Errissya Rasywir	Application of Outstanding Classification Data Mining at Stikom Dinamika Bangsa Jambi Using the Naïve Bayes Method	2021	Naïve Bayes	This research used 100 student data for 2015 and 2016 respectively, so that the total data used was 200. These attributes will be applied to the Naive Bayes method for students majoring in Informatics Engineering class of 2015-2016 as many as 200 data to predict student potential. achievers, which are categorized into 3 namely Very Potential, Potential, and Potential Enough. The results of this study used the Use Training Set with a Correctly Classified Instances accuracy percentage of 87% and 13% Incorrectly Classified Instances. 5-cross validation Correctly with a percentage accuracy of Correctly Classified Instances of 77% and Incorrectly Classified Instances of 23%. 10-Fold Cross Validation with a Correctly Classified Instances percentage of 78% accuracy, 22% Incorrectly Classified Instances [9].
7.	Anggi Trifani, Agus Perdana Windarto, Hendry Qurniawan	Application of C4.5 Classification of Data Mining in Determining Stress Levels of Final Students	2022	Algoritma C4.5	Algorithm C4.5 Calculation Process Calculation of Algorithm C4.5 to determine the dominant factors causing Stress can be described as follows: Step 1: Count the number of cases, the number of cases for Highly Stressful decisions, the number of cases for Not Stressed decisions. Step 2: Calculating the entropy of all cases and cases divided by attribute class with equation (1) then calculating the gain for each attribute with equation (2). C4.5 Algorithm. Generates 20 rules and the accuracy rate produced by this method is 87.88%. Based on calculations

					using the C4.5 Algorithm, the most dominant factor is Interpersonal with a gain value of 0.328180116 [10].
8.	Arief Jananto,Sulastri, Eko Nur Wahyudi Dan Sunardi	Main Student Data as a Predictor of Timeliness of Graduation Using the CART Algorithm for Data Mining Classification Data Mining	2020	Algoritma CART	Student main data at the preparatory stage was taken from the Academic Administration Bureau of the Faculty who had graduated in the graduation data in the form of student name, gender, city of birth, date of birth, name of school of origin for 5 years distributed in each semester. From this process the data obtained amounted to 1151 records. By implementing categorical data mining techniques and the CART algorithm using repeated binary partitioning, it is expected to obtain a decision tree that can be used to predict class timeliness of graduation from active students. By using graduation data and student main data of 1018 records, a decision tree model is obtained with an accuracy rate of 63% from the data testing [11].
9.	Ni Luh Ratniasih	Optimization of Data Mining Using Naïve Bayes and C4.5 Algorithms for Classification of Student Graduations	2019	Algoritma C4.5 Dan KNN	The results of the classification on the training data are the GPA attribute as the root of the decision tree, while the other attributes are as child nodes. From the training data with a total of 50 data produced 5 rules. The rules that have been obtained from the training data can be used as rules to determine whether graduation is on time or not for STMIK STIKOM Bali students. 4 predictor attributes and 1 target attribute produce 5 rules in the decision tree so that these rules can be used in determining timely graduation for STMIK STIKOM Bali students. The results of the analysis using the Naïve Bayes method obtained an accuracy of 89.27% where the performance results are accurate [12].

Based on the data in Table 1, it can be said that the emphasis on the problems raised in Inna Alvi Nikmatum and Indra Waspada's research has problems with the timeliness of students in completing their studies and the proportion of students who complete their studies within the study period are included in the element of accreditation assessment. In Marina Windarti and Agustinus Suradi's research, the problem that can affect the quality of a tertiary institution is student performance which can be measured through the length of study period. In Wiwit Supriyanti's research, and Miss Puspitasari have problems processing student data needs to be done to find out important information in the form of new knowledge (knowledge discovery), for example information about classifying student data based on profiles and academic data. This new knowledge can help universities to classify student graduation rates in order to determine strategies to increase graduation in the following years. In the research by Elga Mariati, Ariesta Lestari, and Widiatry, assessment of student satisfaction with facilities and

learning at the Faculty of Engineering had been carried out before, but the assessment was still carried out partially and the results of data collection on satisfaction assessment had not been evaluated before. This study uses data mining techniques in classifying.

Next, namely the research steps or methods used, based on Table 1. In the research of Inna Alvi Nikmatum and Indra Waspada using the K-Nearest Neighbor. Marina Windarti and Agustinus Suradi used this study to understand the performance of the six classification algorithms used, namely C4.5 decision tree, Bayesian network, KNearest neighbor, Naïve Bayes, neural network and SVM. There are several data mining tools including Rapid Miner, Orange, KNIME , Weka, Keel and R. WEKA is GUI-based so that it minimizes the use of coding which makes it easier for system users. In this study using WEKA tools. Data processed at WEKA has ARFF and CSV formats. In the research by Wiwit Supriyanti, and Nona Puspitasari using the Information Gain Feature Selection Technique for Predicting Student Academic Performance » argues that the main problem in the process of discovering knowledge from data in the field of education is identifying representative data. J48, RandomForest, MLP, SVM to predict students' academic performance in mathematics. In the research of Elga Mariati, Ariesta Lestari, and Widiatry. In this study, the method from the data mining approach was applied to classify whether students were satisfied or not with the quality of learning facilities at the Faculty of Engineering. This study compares three data mining algorithms, namely Decision Tree C4.5, Support Vector Machine, and Naïve Bayes to get the best algorithm for prediction systems.

4. CONCLUSION

Based on the research conducted by the researcher, it can be concluded that from the 9 studies presented, static classification techniques include logistic regression and discriminant analysis. Both assume that the relationship between input and output variables is basically linear, the data is normally distributed, and the variables are not interrelated and independent of one another. Assessment of student satisfaction with facilities and learning at the Faculty of Engineering has been carried out before, but the assessment is still carried out partially and the results of data collection on satisfaction assessment have never been evaluated before. This research uses data mining techniques in classifying. The research that has been carried out is considered insufficient so it is advisable to classify student data in order to be able to manage the data as a whole consisting of student personal data in the form of name, name, student address, parents' names and have student GPA data so that it can be processed appropriately with the like that, it can be seen how student performance, student accuracy, understanding of lessons and can see graduation in a timely manner or not.

REFERENCES

[1] D. Purwandani, C. Sutarsih, “Pengaruh Mutu Layanan Sarana Dan Prasarana Terhadap Kepuasan Mahasiswa Di Fakultas Pendidikan Teknologi Dan Kejuruan Universitas Pendidikan Indonesia,”.

[2] F. Gorunescu, “Data Mining Concept, Models And Techniques. Berlin Heidelberg: Springer.,” 2011.

[3] M. Alghobiri, “A Comparative Analysis Of Classification Algorithms On Diverse Datasets. Engineering, Technology & Applied Science Research,” 2018.

[4] Inna Alvi Nikmatun, “Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor”.

[5] A. S. Mariana Windarti, “Perbandingan Kinerja 6 Algoritme Klasifikasi Data Mining Untuk Prediksi Masa Studi Mahasiswa”.

[6] N. P. Wiwit Supriyanti1), “Implementasi Teknik Seleksi Fitur Forward Selection Pada Algoritma Klasifikasi Data Mining Untuk Prediksi Masa Studi Mahasiswa Politeknik Indonusa Surakarta,” 2018.

[7] W. C Elga Mariati A,1,*, Ariesta Lestari B,2, “Model Klasifikasi Kepuasan Mahasiswa Teknik Terhadap Sarana Pembelajaran Menggunakan Data Mining”.

[8] Y. I. Eka Sabna1, “Data Mining Dengan 2 (Dua) Model Klasifikasi Untuk Prediksi Kinerja Mahasiswa,” *Http://Doi.Org/10.33060/Jik/2021/*.

[9] E. R. Resti[1], Dodo Zaenal Abidin[2], “Penerapan Data Mining Klasifikasi Untuk Memprediksi

Potensi Mahasiswa Berprestasi Di Stikom Dinamika Bangsa Jambi Dengan Metode Naive Bayes," Vol. Vol.3, 2021.

[10] Anggi Trifani, "Penerapan Data Mining Klasifikasi C4.5 Dalam Menentukan Tingkat Stres Mahasiswa Akhir," 2022.

[11] S. Arief Jananto, Sulastri, Eko Nur Wahyudi, "Data Induk Mahasiswa Sebagai Prediktor Ketepatan Waktu Lulus Menggunakanalgoritma Cart Klasifikasi Data Mining," *Doi 10.32736/Sisfokom.V10i1.991*.

[12] Ni Luh Ratniasih, "Optimasi Data Mining Menggunakan Algoritma Naïve Bayes Dan C4.5 Untuk Klasifikasi Kelulusan Mahasiswa".