ISSN 2354-0082

Analysis Of UMN Student Graduation Timeliness Using Supervised Learning Method

Christian Pangestu Kuncoro
Information Systems, Universitas Multimedia Nusantara, Tangerang, Indonesia
christian.pangestu@student.umn.ac.id

Accepted 15 November 2021
Approved 07 February 2022

***Abstract*—— Education is one of the most important things in human life, and in the world of education. However, there are still many students who graduate not on time. The purpose of this study is to find out an overview of what factors influence, then data analysis, and visualization so that students can graduate on time or not on time for UMN student graduates in 2018-2020. The method or approach used to solve the problem is data collection, independent variable, dependent variable, CRISP-DM, with SQLYog tools, to store data, rapid miner for data cleaning, then calculate prediction accuracy with rapid miner using nave Bayes algorithm, and regression logistics, using the included 10-fold validation method, and visualizing the data with Tableau. The conclusion of the final result that is done from this research is for the project to be able to process simple mysql pentaho storage, For data mining, suggesting using the model with the greatest accuracy in each semester in the Information Systems study program is for Semester 1 to use the IPS model - Cross Validation Logistic Regression, then Semester 2 to Semester 7 using the GPA-NaiveBayes (Normal) or GPA-NaiveBayes (Traning With CrossValidation) model). For Data Visualization, there are also insights that will be discussed further in this thesis.**

***Index Terms*—Data Analysis; Data Cleaning; Data Mining; Data Visualization; Education**

I. INTRODUCTION

Education is one of the most important things for human life, education itself has the meaning of knowledge, skills, to a group of people that are passed on from generation to generation through teaching, or research, education is important for generational transfer, because the future is determined by a new generation, current work many are replaced by the new generation [1]. Universities in Indonesia, can take the form of institute, polytechnic, Academy,University and high School. Universities can organizeeducation, vocation, profession, academic, with educational programs Diploma, or Bachelor, or Master, or Doctoral, and Specialist [2].

Multimedia Nusantara University is a university that was established in 2005 with 4 faculties with 12 study programs at the undergraduate level (S1) in 2020, and 1 study program, namely hospitality at the Diploma level (D3) UMN is located in Kelapa Dua Summarecon Serpong, Tangerang Regency. Students graduate in the Bachelor (S1) program with a minimum credit that must be completed is 144 semester credit units (credits), and the maximum length of learning is 7 years, the length of study for undergraduate students (S1) normally according to the curriculum is 8 semesters or for 4 years. However, many students who complete their studies pass the general standard of graduation or can be said to be in the category of not graduating on time. In education, especially Indonesia, the quality of education must be improved, so that it can be useful in the world of work, especially for service to the country[3]. The number and percentage of graduate study programs that are not punctual in 2018 to 2020, where in 3 years the percentage of study programs that increase in punctuality is Accounting, Film, and Television, Communication Studies, and finally Information Systems, data obtained from the Academic Information Bureau (BIA) of Multimedia Nusantara University. [4]. However, there is lack of research that analyzes the category of UMN graduates that has been traced on the UMN knowledge center website [4]. From the existing background, the formulation of the problem emerges, where the questions that will discuss what are the factors that affect the punctuality of graduation for students at Multimedia Nusantara University in the 2018-2020 graduation year.

II. METHODOLOGY

A. Object of research

The object of research in this thesis proposal is UMN students, many UMN students want to graduate quickly or want to improve grades in the intermediate semester, for the Bachelor program is an academic education level that has a study load of between 144 semester credit units (credits) to 160 credits, with a curriculum of 8 semesters. and the length of the program is between 7 to 14 semesters [5], Diploma Three Program (D-3) is an academic education level that has a study load between 108 to 120 semester credit

units (credits), with a curriculum of 6 semesters and program duration between 6 to 10 semesters [6].

The research method is a quantitative research, because it measures a data problem through numbers, and also measures as descriptive words pass on time, or not. Data can be converted in statistical form and taken into account in making a solution and the method used is not from questionnaires, surveys, polls, or interviews that are questions, the number of participants in quantitative methods tends to be more than qualitative[7].

B. *Research Method*

Fig. 1 explaining about CRISP-DM, CRISP-DM stands for Cross-Industry Standard Process for Data Mining which was developed in 1996 by several analysts from the industry, namely NCR, SPSS, Daimler Chrysler. CRISP-DM is a standardization process in data mining for general problem solving for business or for research.

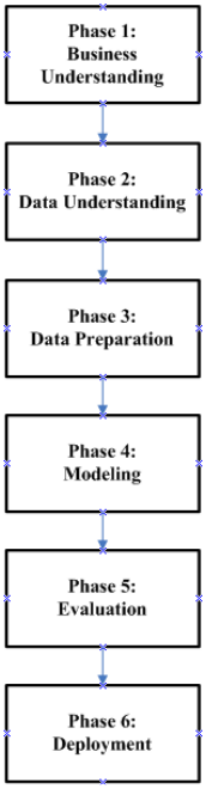

Fig. 1 CRISP-DM Graphics [8]

1) *Business Understanding*

Business Understanding in this thesis is an analysis of UMN student graduation per study program, and per semester, the rate of non-graduation on time is still large and increasing in the 2018-2020 period at UMN, chapter 4 will discuss graduation in the UMN academic guide. Pass is when you have completed all the credits of lessons in each study program including internship and thesis sessions, and have taken the IELTS English exam, pass on time for Bachelor (S1) if you pass 3.5 years or 4 years, for not being on time pass above this number, including students who drop out of lectures[5].

2) *Data Understanding*

The data used is graduate student data (S1) who have graduated from 2018 to 2020 at Multimedia Nusantara University. data was obtained from the Academic Information Bureau (BIA) of Multimedia Nusantara University. The data rows were 3625 rows of data. The dependent variable is a variable that is influenced by the independent variable, the data obtained by students who have graduated from 2018 to 2020 only, in the dataset, the dependent variable is as follows: Category Of Graduates. Whilst independent variables are variables that can affect other variables. In the dataset, the independent variables are as follows: Force, Origin School, Study Program, Faculty, Sex, Ips Semester 1, Ips Semester 2, Semester Between 1, Ips Semester 3, Ips Semester 4, Ips Semester Between 2, Ips Semester 5, Ips Semester 6, Ips Semester Between 3, Ips Semester 7, Gpa Semester 1, Gpa Semester 2, Semester Between 1, Gpa Semester 3, Gpa Semester 4, Gpa Semester Between 2, Gpa Semester 5, Gpa Semester 6, Gpa Semester Between 3 , Semester Gpa 7, Place Of Birth.

3) *Data Preparation*

The data must be clean and free from missing values, therefore by eliminating the missing row labels and missing row attributes, and removing duplicates so that the resulting data is more valid, connect the data to Pentaho MySQL. This project saves MySQL pentaho data using SQLYog software to input data into a table, or create a table, which has a data type, and the length of each column.

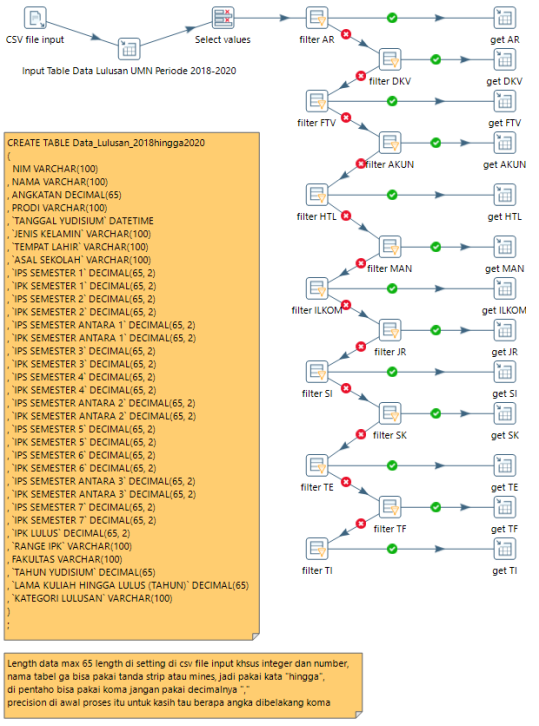

Fig. 2. Image Database Schema

Fig. 2 explain about making data tables getting from csv format a total of 3625 data, inputting MySQL pentaho data from each study program so that it can store data optimally, the name of the thesis database, with the data_graduate_2018 to 2020 table, for a small part of the data, divided into 13 study programs namely the account table, architecture, DKV, FTV, Hospitality, Communication Studies, Journalism, Management, SI, SK, TE, TF, IT.

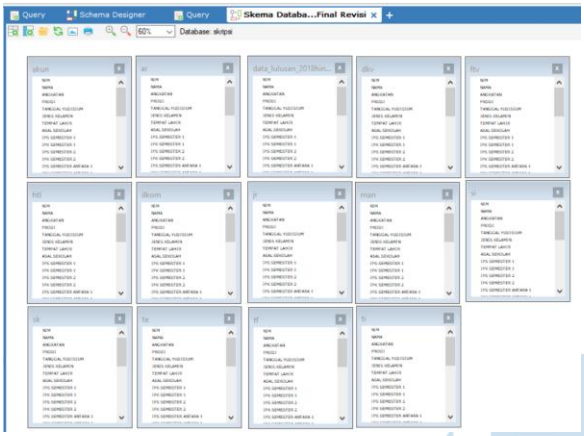

Fig. 3 Image Database Schema

Fig. 3 explain about relational tables, or a collection of tables.

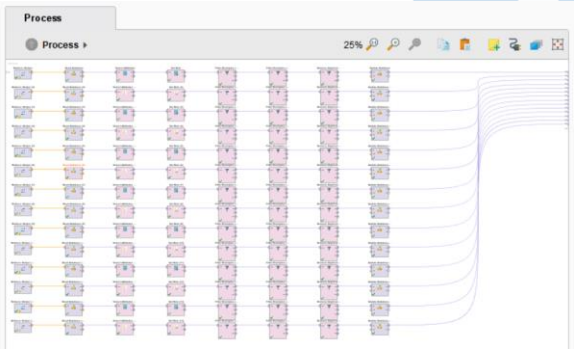

Fig. 4 Data Cleaning

Fig. 4 explain about data cleaning on all tables, by connecting with mysql, then select attributes, and set labels, namely the graduate category, with filters for no missing labels, and no missing attributes so that the data is clean, as well as remove duplicates.

4) Modeling

There are two pupular algorithms that have been applied for this topic. The first algorithm is Naïve Bayes with 96.67% accuracy [9] and and 80% accuracy [10]. Secondly, reseachers uses a logistic regression algorithm with 90.2 % accuracy [11]. The reasons for choosing Naïve Bayes because, it does not have to be numerical for all predicted variables such as neural networks, can be used for quantitative and qualitative data, does not require a large amount of data, if used in programming languages, the code is simple, can be used for problem classification. binary or multiclass, compared to logistic regression where the dependent variable must be binary yes or no, the more variables the more precise. The reason for choosing Logistics Regression is because the independent variables or attributes in logistic regression do not have to be all numeric to predict the dependent variable, using logarithmic or logarithmic logistics, suitable for 2-choice logistic regression or true, false.

The modeling used is classification, using Naïve Bayes data mining algorithm, and logistic regression, for data mining using rapidminer tools, in this study by making a combination of naive bayes and logistic regression with the normal model, Cross Validation, and reducing the IPS attribute per semester with 70% training data, and 30% testing data, after getting a conclusion, data visualization is formed using the Tableau tools.

TABLE I. DATA MINING SOFTWARE COMPARISON

Software	**Advantages**	**Deficiency**
Rapid Miner	*Open source* using 10,000 Rows of data, usage is simpler because it uses the drag and drop method for data mining and data cleansing processes	The data mining process is not described in detail, usually it's just simple, for example, how much accuracy is directly in the result
Python	*Open source*, Multiplatforms there are operating systems windows, Linux, mac, neural network in rapid miner if you want to get accuracy the label must be text, so use python	*coding* more complex than rapid miner and r studio
R Studio	*Open source*, Multiplatforms there are windows, Linux, mac operating systems, relatively good graphics facilities	Coding is more complex than rapid miner and 91amper is the same as complex with r studio

Based on Table I. explains about the selection of data mining software for research from the advantages and disadvantages that exist, then the rapid miner is used because Open-source uses 10,000 data rows, usage is simpler because it uses the drag and drop method for data mining and data cleansing processes.

TABLE II. COMPARISON OF DATA VISUALIZATION SOFTWARE

Software	Advantages	Deficiency
tableau	Interactive visual options, there are moving graphics, User friendly , does not require a lot of hard coding, Dashboard is mobile friendly , can process data on mobile, can connect with many types of databases, there are story features, and dashboards	Paid software, the cost is quite expensive
Power BI	*Open source* software , Can create dashboards, User friendly , doesn't need a lot of hard coding	The amount of data has reached 2GB, must upgrade to the paid version of power BI

Table II. explained about the selection of data visualization software for research from the advantages and disadvantages that exist, by choosing the Tableau software because of the interactive visual options, there are moving graphics, User friendly, does not need a lot of hard coding, Mobile friendly dashboard, can process data on mobile, can connected to a database of many types, there is a story feature, and a dashboard.

5) Evaluation

The evaluation in this study is to compare the accuracy of the Naive Bayes model and logistic regression with the normal model, cross validation, and reduce the Social Studies attribute per semester with 70% training data and 30% testing data.

6) Deployment

In this study, it was used only for learning, the deployment stage was not used because it did not use a system model to UMN students

III. RESULT AND DISCUSSIONS

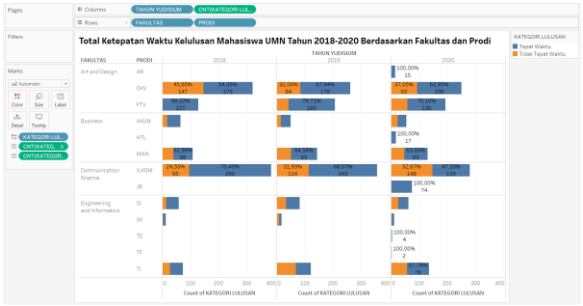

Fig. 5 Graph of Total Timeliness of Graduation of UMN Students in 2018-2020 Based on Faculties and Study Programs

Fig. 5 Explaining the Total Timeliness of Graduation of UMN Students in 2018-2020 Based on Faculties and Study Programs or within a period of 3 years, the blue color indicates the number of students on time, and the orange color indicates the number of students who are not on time, in the graph there is also a number label for see how many students graduate on time or not, including the percentage per pane, from the conclusion that in 2020 graduates, architecture, hospitality, journalism, electrical engineering, engineering physics get graduates 100% on time.

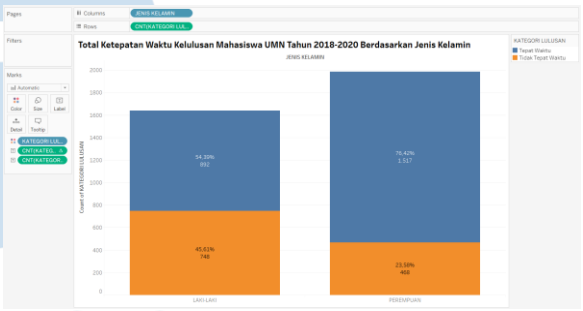

Fig. 6 Graph of Total Timeliness of Graduation of UMN Students in 2018-2020 by Gender

Fig. 6 Explaining the Total Timeliness of Graduation of UMN Students in 2018-2020 Based on Gender or in a 3 year period, the blue color indicates the number of students on time, and the orange color indicates the number of students who are not on time, in the graph there is also a number label to see how many students graduate on time or not, including the percentage, it can be seen that the percentage of women is 76.42% graduating on time compared to 54.39% for men.

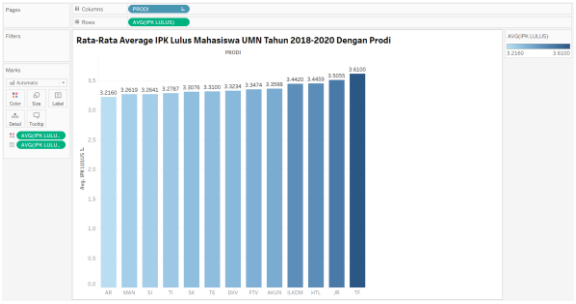

Fig. 7 Graph of Average GPA of Graduated UMN Students in 2018-2020 Based on Study Program

Fig. 7 explained about the average GPA of UMN students from 2018 to 2020 based on study programs, with the winner being engineering physics with an average GPA of 3.61.

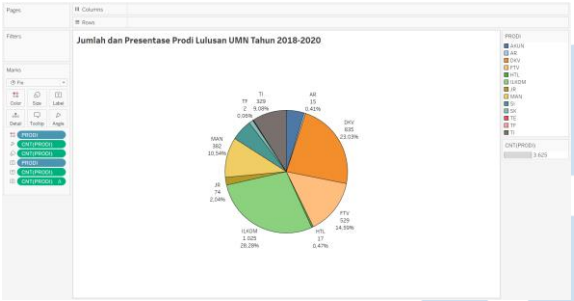

Fig. 8 Number and Percentage of 2018-2020 UMN Graduate Study Programs Based on Study Program

Fig. 8 explains the number, and percentage of study programs at UMN graduates in 2018-2020 based on study programs, with the largest percentage of ILKOM being 1025 students with a percentage of 28.28%, and for diplomas only 17 people, with a percentage of 0.47%.

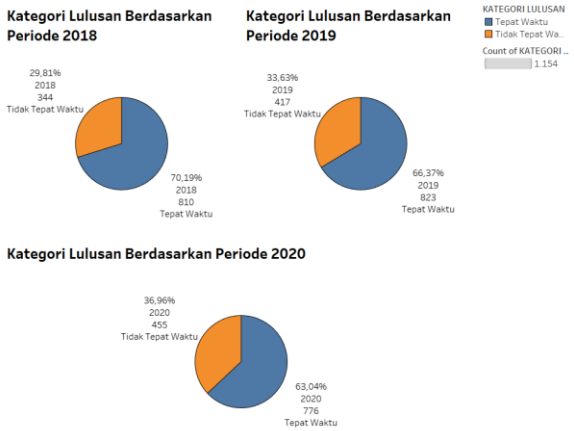

Fig. 9 Graph of Graduate Categories by Period 2018-2020

Fig. 9 Explaining the dashboard of graduates of UMN S1 students over a 3 year period with the percentage, year, number, and category of graduates, from this dashboard it can be said that the percentage of inaccuracy of UMN student graduates increased from a period of 3 years, the blue color shows the number of students on time, and the orange color indicates the number of students is not on time, the data is obtained from the Academic Information Bureau (BIA) Universitas Multimedia Nusantara which was approved by the head of the Information Systems study program, Mrs. Ririn Ikana Desanti.

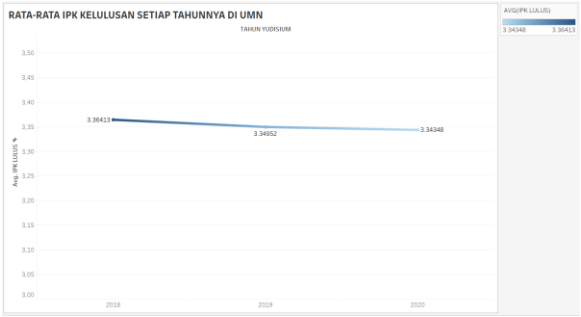

Fig. 10 Average Graduation GPA every year at UMN for the 2018-2020 period

Seen from Fig. 10 that there is also a problem in this research, namely the Average GPA of Graduation Annually at UMN for the 2018-2020 period decreases every year in a special 3-year period for S1, data obtained from the Academic Information Bureau (BIA) Universitas Multimedia Nusantara which was approved by the chairman Information Systems study program, namely Mrs. Ririn Ikana Desanti.

A. Evaluation

TABLE III. ACCURACY RESULTS

Study Program and Semester	Model Algorithm	Accuracy
Semester 1 Information System	IPS - CrossValidationLogisticRegression	75.31
Semester 2 Information System	GPA - NaiveBayes (Normal)	73.77
	GPA - NaiveBayes (Traning With CrossValidation)	73.77
Intermediate Semester Information System 1	GPA - NaiveBayes (Normal)	86.89
	GPA - NaiveBayes (Traning With CrossValidation)	86.89
Semester 3 Information Systems	IPS - NaiveBayes (Normal)	83.61
	IPS - NaiveBayes (Traning With CrossValidation)	83.61
Information System Semester 4	IPS - NaiveBayes (Normal)	83.61
	IPS - NaiveBayes (Traning With CrossValidation)	83.61
Intermediate Semester Information System 2	IPS - NaiveBayes (Normal)	83.61
	IPS - NaiveBayes (Traning With CrossValidation)	83.61
Semester 5 Informasi Information System	IPS - NaiveBayes (Normal)	81.97
	IPS - NaiveBayes (Traning With CrossValidation)	81.97
Information	IPS - NaiveBayes (Normal)	81.97

Study Program and Semester	Model Algorithm	Accuracy
Systems Semester 6 Sistem	IPS - NaiveBayes (Traning With CrossValidation)	81.97
SI Semester Between 3	IPS - NaiveBayes (Normal)	81.97
	IPS - NaiveBayes (Traning With CrossValidation)	81.97
SI Semester 7	IPS - NaiveBayes (Normal)	80.33
	IPS - NaiveBayes (Traning With CrossValidation)	80.33

Table III. describes the comparison of the output accuracy of the winning model in each semester in each study program which for 3 years from 2018 to 2020 experienced an increase in untimely graduation, in this thesis data mining all using the most influencing variable with the greatest accuracy is class, school origin, IPS Semester 1, IPS Semester 2, GPA Semester 2, IPS Semester 1, IPS Semester 3, IPS Semester 4, IPS Semester 2, IPS Semester 5, IPS Semester 6, IPS Semester 3, IPS Semester 7.

In conclusion, suggesting using the model with the greatest accuracy in each semester in the Information Systems study program is for Semester 1 to use the IPS model - Cross Validation Logistic Regression, then Semester 2 to Semester 7 using the GPA-NaiveBayes (Normal) or GPA-NaiveBayes (Traning With CrossValidation) model).

TABLE IV. RESULTS OF COMPARISON OF DATA ANALYSIS

Category	Before Data Analysis	With Data Analysis
Data storage on Pentaho MySQL	Previously there was no data that wanted to be analyzed on the data stored in Pentaho MySQL	Now there is data then it can be stored in Pentaho MySQL
Data Mining	Previously there was no data that you wanted to mine	Now that there is data, it can be mined to get new knowledge
Data Visualization	Previously there was no data, which was used for visualization	Now it helps to see visualizations, to get new knowledge

Table IV. describes the results of the comparison of data analysis from 3 categories, namely data storage in Pentaho MySQL, data mining, and data visualization.

IV. CONCLUSIONS

A. *Conclusions*

From the results of the practicum in this thesis report, the goal was achieved. By knowing the factors that affect the timeliness of graduation for Bachelor (S1) students at Multimedia Nusantara University in the 2018-2020 graduation year with the variable of the algorithm winner from each semester in the Information Systems study program, namely the generation, origin school, IPS Semester 1, IPS Semester 2, GPA Semester 2, IPS Semester 1, IPS Semester 3, IPS Semester 4, IPS Semester 2, IPS Semester 5, IPS Semester 6, IPS Semester 3, IPS Semester 7, with processing where MySQL data input was successful using Pentaho software, and SQLYog, for visualization data, the percentage of inaccuracy level of UMN student graduates increased from a 3-year period, Average Graduation GPA Annually at UMN 2018-2020 period decreased every year in a 3-year periodit can be seen that there are more percentages of women graduating on time than men, and in 2020 graduates, architecture, hospitality, journalism, electrical engineering, physics engineering obtained graduates 100% on time, then the average GPA graduated from UMN students in 2018 to 2020 based on study programs, the winner is engineering physics with an average GPA of 3.61, the number, and percentage of study programs at UMN graduates in 2018-2020 based on study programs, with the largest percentage of ILKOM being 1025 students with a percentage of 28.28%, and for diplomas only 17 people, with a percentage of 0.47%.then the average GPA graduated from UMN students in 2018 to 2020 based on study programs, the winner was physics engineering with an average GPA of 3.61, the number, and percentage of study programs for UMN graduates in 2018-2020 based on study programs, with the largest percentage of ILKOM being 1025 students with a percentage of 28.28%, and for diploma only 17 people, with a percentage of 0.47%.then the average GPA graduated from UMN students in 2018 to 2020 based on study programs, the winner was physics engineering with an average GPA of 3.61, the number, and percentage of study programs for UMN graduates in 2018-2020 based on study programs, with the largest percentage of ILKOM being 1025 students with a percentage of 28.28%, and for diploma only 17 people, with a percentage of 0.47%.

B. *Suggestions*

The findings of this model can be used as input to create a database, or create a timely graduation rate analysis system for UMN students, so that with a directly connected database they can issue reports and complete the analysis.

REFERENCES

[1] A. A. Zafi, "Transformasi Budaya Melalui Lembaga Pendidikan (Pembudayaan dalam Pembentukan Karakter)," vol. I, no. 1, pp. 1–16, 2018, [Online]. Available: https://ejournal.stainupwr.ac.id/index.php/al_ghzali/article/download/5/1.

[2] A. Abdullah, "Pengembangan Pembelajaran Karakter Berbasis Soft Skill di Perguruan Tinggi," *Ishraqi*, vol. 1, no. 1, pp. 18–30, 2017, [Online]. Available: http://journals.ums.ac.id/index.php/ishraqi/article/view/2926/2300.

[3] RISTEK DIKTI, "PERATURAN MENTERI RISTEK DAN DIKTI NO 44 TAHUN 2015," 2016. http://kopertis3.or.id/v2/wp-content/uploads/Bu-Illah-SN-DIKTI-44-2015-SOSIALISASI-APTISI.pdf (accessed Feb. 19, 2021).

[4] UMN, "UMN Knowledge Center," 2021.

https://kc.umn.ac.id/ (accessed Jun. 16, 2021).
[5] UMN, *Buku Panduan Akademik Sistem Informasi*. 2020.
[6] UMN, *Buku Panduan Akademik Perhotelan*. 2020.
[7] H. H. Elkatawneh, "Comparing Qualitative and Quantitative Approaches," *SSRN Electron. J.*, vol. 3, no. 2, 2016, doi: 10.2139/ssrn.2742779.
[8] D. Feblian and D. U. Daihani, "Implementasi Model Crisp-Dm Untuk Menentukan Sales Pipeline Pada Pt X," *J. Tek. Ind.*, vol. 6, no. 1, pp. 1–12, 2017, doi: 10.25105/jti.v6i1.1526.
[9] N. Frastian, "Implementasi Komparasi Algoritma Klasifikasi Menentukan Kelulusan Mata Kuliah Algoritma Universitas Budi Luhur," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 3, no. 1, p. 1, 2018, doi: 10.30998/string.v3i1.2334.
[10] E. Purnamasari, D. P. Rini, and Sukemi, "Seleksi Fitur menggunakan Algoritma Particle Swarm Optimization pada," vol. 1, no. 10, pp. 469–475, 2020.
[11] I. Ketut, P. Suniantara, and M. Rusli, "Klasifikasi Waktu Kelulusan Mahasiswa Stikom Bali Menggunakan Chaid Regression – Trees dan Regresi Logistik Biner," *Statistika*, vol. 5, no. 1, pp. 27–32, 2017.
[12] F. Chollet, *Deep Learning with Python*. Manning Publications Co, 2020.
[13] N. Ketkar and E. S., *Deep Learning with Python : A Hands-on Introdcution*. Banglore: Karnataka: Apress, 2017.
[14] F Provost, *Data Science for Business: What you need to know about data mining and data analytic thinking*. O'Reilly Media, Inc, 2013.
[15] M. A. K, *Business Intelligence And Data Mining*. NY: Business Expert Press, 2015.
[16] A. Fauzan, "Implementasi Software R pada Mata Kuliah Kalkulus I untuk Menunjang Kemampuan Pemrograman Mahasiswa," vol. 2, no. 1, pp. 279–286, 2020.
[17] "RStudio Pricing - RStudio." https://rstudio.com/pricing/ (accessed Nov. 20, 2020).
[18] "RapidMiner Pricing | RapidMiner." https://rapidminer.com/pricing/ (accessed Nov. 20, 2020).
[19] J. D. Miller, *Big Data Visualization*. UK: Packt Publishing Ltd, 2017.
[20] D. Santos, *Tableau 10 Business Intelligence Cookbook*. UK: Packt Publishing, 2016.
[21] "Tableau Pricing for Individuals and Personal Use." https://www.tableau.com/pricing/individual (accessed Nov. 20, 2020).
[22] "Pricing & Product Comparison | Microsoft Power BI." https://powerbi.microsoft.com/en-us/pricing/ (accessed Nov. 20, 2020).