Predicting Computer Science Student's Performance using Logistic Regression

1st Noviyanti T M Sagala
Statistics Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
noviyanti.sagala@binus.edu

2nd Syarifah Diana Permai
Statistics Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
spermai@binus.edu

3rd Alexander Agung Santoso Gunawan
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
aagung@binus.edu

4th Rehnianty Octora Barus
Student Advisory & Support Center
Bina Nusantara University
Jakarta, Indonesia, 11480
rbarus@binus.edu

5th Cito Meriko
Student Advisory & Support Center
Bina Nusantara University
Jakarta, Indonesia, 11480
Cito.meriko@binus.edu

2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-6654-5512-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISRITI56927.2022.10052968

***Abstract*— The prediction of student academic performance has drawn noticeable attention in higher education institutions. It may help educational institutions in providing quick assistance to students who are low performed at an early stage. However, it becomes challenging to predict the performance of students because the numerous amount and complexity of the variables such as student's demographics information, student's achievement data, and student's activity data. The purposes of this study are to develop a predictive model using Logistic Regression algorithm on student data coming from a computer science faculty in an Indonesian private university and to identify variables in playing a part in student's performance. The dataset is gathered from the past decade's student records: 2010-2020. It is then explored and preprocessed using various approaches such as merging, aggregating, feature encoding, and SMOTE. About 75% of the dataset is used for training and the remaining is for the testing set. Out of 20 variables used, 10 variables are statistically significant to the response variable. The results show that Logistic Regression achieves a considerably good predictive model with an accuracy rate of 91%, recall rate of 98%, and precision rate of 85%. The model can be used as an early warning system to identify low-performing students and inform both the faculty members and the students. Faculty members can then employ a range of strategies to communicate with low-performing students and give them with opportunities to improve their performance.**

Keywords*— *Data Mining, Higher Education, Logistic Regression, Low-performing Student, Non-Regular Student

I. Introduction

Student performance is important part of education. Students with poor academic performance will encounter numerous problems, including late graduation and even leaving out. As a result, educational institutions should regularly monitor their students' academic performance and provide quick assistance to pupils who are underperforming. One approach is to identify and predict students' academic achievement. With an accurate predictive model, it is also possible to identify the factors that contribute to student performance. This strategy will assist educational institutions in identifying and assisting low-performing students at an early stage.

However, measuring student academic performance becomes difficult since student academic performance is affected by factors such as demographics, personal traits, educational background, psychological, academic record advancement, and other environmental variables [1]. According to a recent comprehensive review, nearly 70% of the evaluated work studied student performance prediction using student grades and GPAs, while less than 1% of the research investigated academic achievement prediction using demographic and student activity data [2]. In addition, researchers are growing more interested in performance categorization as the relationships between many of these parameters remain unknown. The authors in [3, 4] said that it is understandable that it would be easier to make a prediction if there were fewer target classes. According to the authors of [5], many scholars have attempted to characterize student performance using binary categories such as pass-fail, below and above a reference level, good-poor, and so on. Moreover, student's academic performance prediction has been performed at different levels: subject [6-9], semester [10-13], and degree grade level [14-16].

The goal of this study is to predict student success using the most important variables obtained from students' activity data, demographic information, and achievement data in a computer science faculty. It is achieved by developing a model using Logistic Regression. The following are the primary goals of this research:

1. Predict student academic performance into regular or non-regular students using student's demographically data, student's activity's data, and student's achievement data.
2. Finding the variables in contributing student's performance

This study contributes to the literature in several ways. First, student's data used are coming from a faculty of a private university in Indonesia. Second, the variables used are obtained from student's academic achievement, demographic, and activity data. To the best of our

knowledge, none of work investigated on these data and variables has been published.

II. Related Work

Students' performance prediction has been performed at different levels, with the most accurate results at the level of single subject marks, followed by semester grades and then overall grades. At the subject level, the authors have predicted the marks of the Introduction to informatics module of distance learning at Hellenic Open University, Greece using demographic features/variables (age, sex, occupation, etc.). We believe in using assignment marks and face-to-face meetings to get the most out of our students [17]. This study [18] found that cognitive features (such as GPA, pre-requisites course marks, and midterm marks) can predict an undergraduate's performance in an engineering dynamic course at Utah State University, Logan, USA. Previous research [8, 9] has shown that cognitive features (such as a student's progressive academic history and grade point average) and observations of the student's on-campus activities can predict how well they will perform in core courses.

At the semester level (which is also the focus of this study), the authors in [10] used academic information, student activity, and student video interactions to predict whether students would pass or fail at the end of the semester. Another study [11] conducted an experiment to predict Semester GPA (SGPA) using quizzes, discussions, assignments, attendance, and lab work. Using pre-university characteristics and previous academic performance [12], previous four semester grades were used to predict SGPA [13].

A study [17] conducted experiments with 25 characteristics on a sample of 250 students to predict 3rd semester performance (excellent, above average, average, or below average) using a decision tree with an accuracy of 94.40%. Another study [18] examined a sample of 300 students to predict final semester performance and identified factors affecting semester performance using different supervised machine learning algorithms. The results show that Random Forest outperforms other classifiers in terms of accuracy. A study conducted [12] investigated the relationship between social factors and academic performance to predict the performance of third semester students. Parents' education and academic performance in the second semester were good predictors. In [19], Logistic Regression algorithm was trained to predict the performance of 100 students from the four department of federal school of statistics, Ibadan Oyo, Nigeria. The results show that students' monthly salary and student study time were significant predictors. While the gender of the parents and the level of education were insignificant predictors. The suitability of the model was assessed using the Hosmer and Lemeshow test, the split sample approach, and other additional indices to validate the model. Studies [20-22] conducted experiments to predict student performance at degree levels: Electrical Engineering, Computer Science, and Civil Engineering, respectively.

The literature review shows that aspects influencing academic achievement can vary according to courses, semesters and classes, and there is a need to investigate aspects influencing academic achievement at the local level.

III. Methodology

The proposed method for this study consists of four stages as displayed in Fig. 1. It starts with data understanding, followed by exploratory data analysis, then data processing, and modeling & evaluation.

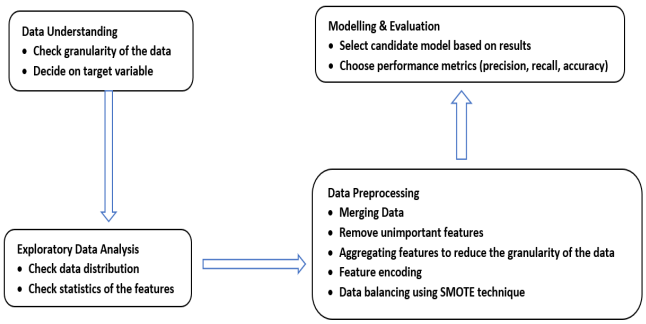

Fig. 1. The Proposed Methodology

A. Data Understanding

The dataset used in this study was collected from different departments. Operation & Resources department, student advisory and support center, IT department. It is required to receive official approval for gathering this dataset for research purposes. The data were provided in three Csv files. The first file contains student activities data with 195,081 rows and 9 columns. The second dataset was student achievement data with 116665 records and 13 columns, and the last dataset consists of students' demographic features with 44169 rows and 26 columns. They were collected from past decades: 2010- 2020, of selected school, Bina Nusantara University, Indonesia.

The granularity/level of the data must be examined. Previous studies often used training data at the achievement level. It is also possible to incorporate student activities data, which comprises information on the students' activities during the term, such as organization or competition. All datasets are generally be contained in a separate data set that would then be transposed and merged with the achievement and activities data, so that each student has specific accomplishment and activities statistics as characteristics in the data set. Including demographic data would have the advantage that we can investigate whether specific student background information is essential for student performance in terms of whether they underperform or well-performed

Most of the previous research approached student performance prediction problem as a dropout rates problem with classification issue (Yes or No). Some considered the issue as a multiclass problem, employing data mining techniques to estimate the CGPA point and then predicting a category based on the expected points ((excellent, very good, good, average, and poor)). The authors eventually discovered that approaching the problem as a classification produced greater results, but that is not to say that this would be accurate for all student performance or all data sets.

Term *Binusians* means those who are engaged with BINUS University (students, alumni, teachers, training participants, and *BINUS* employees). With regards to data privacy, program and program categories are not described specifically. In total, there are 20 variables used in this work. List of variables used are presented in Table I.

TABLE 1. DESCRIPTION OF VARIABLES

No	Category of Attributes	Attributes	Values
1	Achievement	Student Status	(1 = non-regular; 0 =Regular)
		Semester	Numerical (1 to 6)
		CGPA Status	>2; <= 2
		Course Credit Status	Multiples of 15, Less than multiples of 15
		Campus Location	Aggrek, Alsut
		Program	9 different programs
		Program Categories	Categories of Program
2	Demographic	Gender	Female, Male
		Scholarship types	Binusian, Other, Regular
		English Score	Beginner, Intermediate, Advance
		Father's job	Employee, Unemployment, Other
		Mother's Job	Employee, Unemployment, Other
		Mother's status	Alive, Died
		Father's Status	Alive, Died
		Father Education Status	Level 1, Level 2, Level3
		Mother Education Status	Level 1, Level 2, Level3
3	Activity	ExtOrg(External Organization)	Yes, No
		IntOrg(Internal Organization)	Yes, No
		ParInNonAcadCom(Participation in non-academic competition)	Yes, No
		PartInAcadComp (Participation in academic competition)	Yes, No

B. Exploratory Data Analysis

In order to understand the dataset in hand, it must be explored in a statistical and visualization form using plots. Table II includes summary statistics of the dataset (mode and total of missing values).

TABLE II. SUMMARY STATISTICS

Variable	Mode	Missing Values (%)
Father's Education	Level 2	11.25
English Score	Intermediate	10.29
Mother's Education	Level 3	8.46
Father's Job	Other	6.45
Mother's Job	Unemployment	5.14
Father's Status	Alive	5.06
Mother's Status	Alive	4.97
Scholarship Type	Other	1.05
PartInNonAcadCom	N	0.37
PartInAcadCom	N	0.37
IntOrg	N	0.37
ExtOrg	N	0.37

The following analysis is a presentation of data for analysis based on the final status of student. From Fig. 2, majority of the mother's education of students still gathering around level 3 for both categories of student status. In other words, most of mother's education of the students are tertiary level at most. As compared to level 2 and level 3 categories, there is no significant differentiation between student status '0' and '1' in level 1 category.

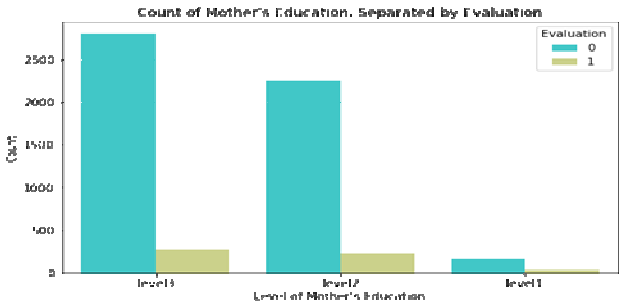

Fig.2. Count of Level of Mother's Education, separated by Evaluation (Student Status)

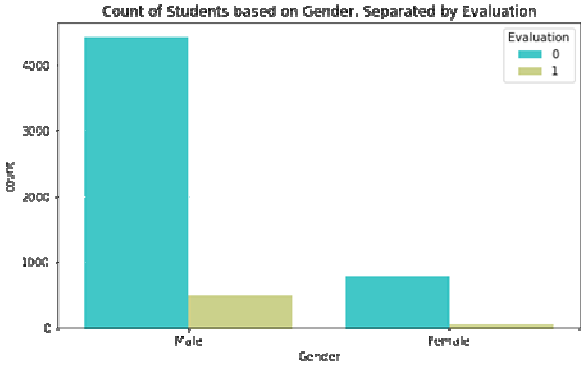

Fig.3. Gender, separated by Evaluation (Student Status)

From Fig. 3, it is shown that the female student has greater proportion of '0' or well-performed status than male student.

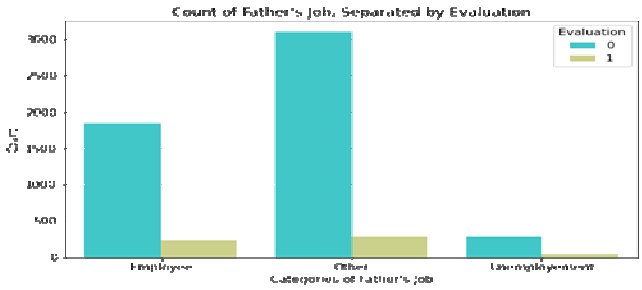

Fig.4. Categories of Father's Job, separated by Evaluation (Student Status)

In terms of father's job, majority of students who are underperformed are unemployed as in Fig. 4. However, student who are well-performed came from family where father's job is entrepreneurs (other).

C. Data Preprocessing

Python's SciKit learn, and Pandas' libraries were used for pre-processing. Since the datasets were collected from different sources, there were a need to merge them as one to simplify data exploration and to avoid data redundancy. Initially, unimportant features were removed from the datasets. We applied eliminating approach for:

1. Three valueless variables from activity dataset (ID, student name, status, and term); 12 variables from demographics dataset (formular number, academic career, student name, student ID, academic group, age, father range salary, mother range salary, mother salary, father salary, tuition level, and address);
2. Activity dataset: Duplicated records (about 157,806); demographic: 1669 rows
3. Achievement dataset: short term records. The short-term record has been combined with even term record.

Three datasets were merged based on term and student ID. The dataset was 28066 rows and 27 columns. Some categorical features possess high cardinalities. It means that too many unique values. It is required to handle this type of data problem which may lead to curse of dimensionality. Aggregating function was employed on the dataset to reduce cardinality. The objective is to keep instances belonging to values with a high frequency and replace the others with a new category named other. The high cardinality variables and its conversion are listed in Table III.

Table III. Variables for Aggregation

Variable	Before Aggregation	After Aggregation
Father's Education; Mother's Education	Doctor, Master, Specialist 2,	Level 1
	Bachelors, Specialist 1, Diploma 4, Diploma 3, Diploma 2, Diploma 1	Level 2
	not completed High School, completed high school, completed junior school, not completed junior school, completed primary school, not completed primary school	Level 3
Scholarship	Binusian Family Scholarship, Sibling Scholarship, Binusian Community Scholarship, Binus ambassador Scholarship, binusian community and early bird Scholarship, Binusian Scholarship	Binusian
	TDP, NDP, BIS, direct admission, AyoKuliah, TeacherChildren, KompasWinner, TPKS, Special TPK, BidikMisi Scholarship, Widia Scholarship, widia partial Scholarship, widia outstanding achievers Scholarship, education expo Scholarship, talent mapping Scholarship	Other
	Regular Admission, Early Batch, Regular, Kalbis	Regular
Father's Job; Mother's job	Indonesian National Armed Forces, Teacher, Employee in National University, Employee in Private University, Farmer	Other
	Unemployed, Pension	Not employee
	Civil Servant, Private Employee	Employee
English Score	>550	Advance
	456 - 550	Intermediate
	<467	Beginner

Some machine learning algorithms is inefficient in handling categorical features, therefore categorical features were transformed into numeric form using one-hot encoding where binary valued dummy variables were added for each category. Furthermore, some features may have a greater impact while training a machine learning model due to differences in the range values of distinct numeric/quantitative data. Quantitative features were converted into a single scale with a zero mean and unit variation to eliminate this type of bias. The dataset was imbalanced: 90.76% of students belonged to the non-regular category 0 (majority class), and only 9.24 belonged to the non-regular category 1 (minority class), which can lead to the non-generalized machine learning model (aka overfitted model). The synthetic minority oversampling technique (SMOTE) was used to correct for the uneven nature of the data set. Based on a random sampling algorithm, new cases of minority groups were generated using synthetic sampling technique to create a more balanced distribution. For the minority class, the SMOTE technique identifies examples close to the feature space by drawing a line between the examples and drawing a new sample at a point along this line. After SMOTE algorithm was implemented, the number of records for regular and non-regular classes are 221 records, respectively.

D. Modelling & Evaluation

Logistic Regression (LR) is a type of regression that works best when the outcome is binary (1). Various academic aspects of the student are involved in predicting the final status of the student whether regular or not. Logistic Regression commonly used to compute the probability of an event with the score ranging from 0 to 1. The cut-off value is a critical feature of logistical regression and is determined by the classification issue itself.

The predictive model in this work was developed using Logistic regression algorithm. Statmodels module and Logit Funtion were applied on the data. The data set was divided into around 75% training data (8216 records) and 25% test data (442 instances).

Out of 20 original variables, 10 variables identified statistically significant to the response variable (p-values = 0.05). These variables were then employed for modelling. They are gender, scholarship_status, father's job, mother's job, father's education, mother's education, mother's status, English score, and campus location. The model was evaluated using classification_report() for accuracy, precision, and recall. The results were presented in Table IV.

Accuracy: the closeness of a measurement to the real value of a quantity

Precision: the proportion of complete TP matches among all TP matches. If the precision is close to one, then the expectations will gradually become accurate.

Recall: the proportion of the TP matches among all the possible positive matches.

Table IV. Logistic Regression Model Performance

	Precision	Recall	F1-Score	Support
0	0.98	0.83	0.90	217
1	0.85	0.98	0.91	217
Accuracy			0.91	434
Macro avg	0.92	0.91	0.90	434
Micro avg	0.92	0.91	0.90	343

The model achieved a classification rate of 91% considered as good accuracy. Out of all the students that the model predicted would be non-regular, 85% actually were. Out of all the students that actually were non-regular, the model predicted this outcome correctly for 98% of those students.The F1-score is close to 1 means that the model dose a good job of predicting whether or not students are regular. Support values simply provide information on how many students belonged to each class in the test dataset. We can see that the number of regular and non-regular students are balance.

IV. CONCLUSION

The goals of this study are to develop a prediction model utilizing the Logistic Regression algorithm on student data from a computer science faculty at an Indonesian private institution, as well as to identify characteristics that influence student performance. The dataset was created using student records from 2010 to 2020. It is then investigated and preprocessed using various methods such as merging, aggregation, feature encoding, and SMOTE. Approximately 75% of the dataset is used for training, with the remainder for testing. Ten of the twenty variables utilized are statistically significant to the response variable. The results reveal that Logistic Regression produces a very good prediction model, with an accuracy rate of 91%, a recall rate of 98%, and a precision rate of 98%.. A limitation of this study is that variables are mainly in categorical form. We intend to expand the study by gathering more additional characteristics such as grade for each course and frequent absence, as well as analyzing psychological aspects that affect student performance. In the future, we aim to use more detailed data sets from other faculties/departments.

REFERENCES

[1] Bilal M, Omar M, Anwar W, Bokhari RH, Choi GS. The role of demographic and academic features in a student performance prediction. Sci Rep. 2022 Jul 22;12(1):12508. doi: 10.1038/s41598-022-15880-6. PMID: 35869103; PMCID: PMC9307570.

[2] Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018; pp. 175–199

[3] Vijayalakshmi (2022) Predicting the academic progression in student's standpoint using machine learning, Automatika, 63:4, 605-617, DOI: 10.1080/00051144.2022.2060652

[4] Dutt A, Ismail MA, Herawan T. A systematic review on educational data mining. IEEE Access. 2017; 5:15991– 16005.

[5] Saqr M, Fors U, Tedre M. How the study of online collaborative learning can guide teachers and predict students' performance in a medical course. BMC Med Educ. 2018;18(1):1–14.

[6] Ahmed NS, Sadiq MH. Clarify of the random forest algorithm in an educational field. In 2018 international conference on advanced science and engineering (ICOASE) (pp. 179-184). IEEE; 2018, October.

[7] Kotsiantis, S. B. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artif. Intell. Rev. https://doi.org/10.1007/s10462-011-9234-x (2012).

[8] Huang, S. & Fang, N. Computers & Education Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Comput. Educ. 61, 133–145 (2013).

[9] Xu, J., Moon, K. H., Member, S. & Van Der, S. M. A machine learning approach for tracking and predicting student performance in degree programs. IEEE J. Sel. Top. Signal Process. 11, 742–753 (2017).

[10] Hasan, R., Palaniappan, S., Rafez, A., Mahmood, S. & Sarker, K. Student academic performance prediction by using decision tree algorithm. In 2018 4th Int. Conf. Comput. Inf. Sci. 1–5 (2018).

[11] Hasan, R. et al. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Appl. Sci. 10(11), 3894 (2020).

[12] Razaque, F. et al. Using naïve bayes algorithm to students' bachelor academic performances analysis. In: 4th IEEE Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2017 1–5 (2018).

[13] Singh, W. & Kaur, P. Comparative analysis of classifcation techniques for predicting computer engineering students' academic performance. Int. J. Adv. Res. Comput. Sci. 7(6), 31–36 (2016).

[14] Mishra, A. & Chaudhary, N. Student performance measure by using diferent classifcation methods of data mining. Turk. J. Comput. Math. Educ. 12, 4063–4069 (2021)

[15] Asif, R., Merceron, A., Ali, S. A. & Haider, N. G. Analyzing undergraduate students' performance using educational data mining. Comput. Educ. https://doi.org/10.1016/j.compedu.2017.05.007 (2017).

[16] Asif, R., Hina, S. & Haque, S. I. Predicting student academic performance using data mining methods. Int. J. Comput. Sci. Netw. Secur. 17(5), 187–191 (2017).

[17] Asif, R., Haider, N. & Ali, A. Prediction of undergraduate student' s performance using data mining methods. Int. J. Comput. Sci. Inf. Secur. 14, 374–380 (2016).

[18] Kotsiantis, S. B. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artif. Intell. Rev. https://doi.org/10.1007/s10462-011-9234-x (2012).

[19] Huang, S. & Fang, N. Computers & Education Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Comput. Educ. 61, 133–145 (2013).

[20] E. K. Akinyemi, O. A. Ogunleye, H. Olaoye, and J. Brakoru, "Binary Logistic Regression Analysis on Predicting Academics Performance", *CJAST*, vol. 40, no. 20, pp. 1-6, Aug. 2021.

[21] Asif, R., Haider, N. & Ali, A. Prediction of undergraduate student ' s performance using data mining methods. Int. J. Comput. Sci. Inf. Secur. 14, 374–380 (2016).

[22] Dutt A, Ismail MA, Herawan T. A systematic review on educational data mining. IEEE Access. 2017; 5:15991– 16005. .