Quest Journals
Journal of Software Engineering and Simulation
Volume 7 ~ Issue 12 (2021) pp: 20-25
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Application of Logistic Regression to Analyze Student Performance in Elective Courses

Felix Andreas Sutanto, Heribertus Yulianton, Budi Hartono
[1,2,3](Teknik Informatika, Stikubank University, Indonesia)

ABSTRACT: *Learning achievement is one indicator that can be used to assess how much a student has succeeded in his education. In the informatics engineering study program curriculum, there are 2 concentrations of specialization that need to be analyzed for their effectiveness. This study will analyze other factors outside of academic ability that affect the success of a student in pursuing his studies, especially in specialization courses that are the flagship of the study program. Other factors include gender, parents' education, and city. The research objects are students of the Informatics Engineering, Stikubank University. With this research, it is hoped that the researcher will get an idea of the predicted student performance from the supporting factors. This research will use machine learning approach and Logistic Regression method to process data. The data on gender, city and parents' education are of ordinal type, and the final destination is of binary type. For its implementation using the Python programming language. The results showed that gender had a fairly high difference in learning success, but the city factor and parents' education were not too high.*

KEYWORDS: *Logistic Regression, student performance, machine learning, gender, ordinal.*

R*eceived 08 Dec, 2021;* ***R****evised 21 Dec, 2021;* ***A****ccepted 23 Dec, 2021 © The author(s) 2021.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

In general, the curriculum in higher education is structured to achieve a graduate standard in which there are learning outcomes. Learning achievement is one indicator that can be used to assess the success of a student in his education. Every year, Stikubank University welcomes students from various regions and different backgrounds. Because they come from different environments, student profiles are interesting things to analyze. In the informatics engineering study program, there are elective courses that are superior and have characteristics that distinguish them from others. These courses are usually taken after students complete basic courses. Based on the university's academic manual, there are 2 choices of specialization, each consisting of 4 courses.

This study will analyze student learning outcomes from the point of view of the student's background or profile. The purpose of this study, the researchers wanted to know whether the factors of gender, city and parents' education of students will affect student learning achievement, especially in the elective courses of the study program. In addition, the results of the analysis can be used for management in mapping student profiles and making learning strategies suitable for students. The results of the study can be used as a model for student grouping to improve learning success in the future.

II. LITERATURE REVIEW

Research in the field of categorical data clustering has begun to develop, although the development is still far less than clustering on numeric data types. Logistics Regression is a non-linear regression, used to explain the non-linear relationship between X and Y, the non-normality of the distribution of Y, the variability of the response is not constant which cannot be explained by the ordinary linear regression model [1]. Logistics regression is a statistical analysis method to describe the relationship between the dependent variable which has two or more categories and one or more independent variables on a categorical or interval scale [2].

Research on logistic regression analysis for accreditation cases aims to determine the accreditation model for senior high school in Ambon city based on the factors contained in the school profile. In this study, the predictor variables used for logistic regression analysis were school status (there are two categories, namely public and private), length of time in school, number of students, number of teachers, status of building land (1

= self-owned; 0 = rented/ride), and total average value of the national examination. The results showed that a significant variable affecting the accreditation of senior high school in Ambon was the number of teachers [3]. To detect heart disease, a logistic regression algorithm is used by using a logistic function to generate binary zeros and ones as classification determinations. After the experiment was carried out with the logistic regression algorithm, it gave results that have different advantages over other methods based on the confusion matrix analysis model [4]. Cardiovascular disease identification techniques are complicated to do. The diagnosis and treatment of cardiovascular disease are very complex. For this reason, a support system is implemented to predict cardiovascular disease through a machine learning model. By using the Heart Disease UCI dataset consisting of fourteen variables, including age, sex, cp, fbs, restecg, thalac, exang, oldpeak, slope, ca, thal, and target, it was found that the use of the logistic regression algorithm is effective and efficient in predicting cardiovascular disease where based on the results of data validation it is found that the accuracy of the prediction results with the algorithm reaches 85% with an error rate that tends to be small at 0.1406565. [5]

III. STUDENT DATA

In this study, the student data used as the sample were students of the 2018 and 2019 Informatics Engineering study program. The data was taken from the Smart Campus information system at Stikubank University. Students in that year were selected because they used the curriculum set for the flagship subjects in this study. In 2021 the student data that can be studied are 2 batches in the informatics engineering study program. The data needed in this study are only gender, city and education of parents (father) as well as the values of elective courses. The attributes of gender, city and education are categorical variables, not numeric values. The gender attribute has the value of male and female. City attributes consist of 2 groups, namely Semarang and outside Semarang groups. As for the education of parents consists of high school, diploma and undergraduate. In this curriculum, elective courses are divided into 2 concentrations. Students usually choose one of these concentrations. Concentration 1 courses consist of Social Networking, Geographic Information Systems, Big Data, and Cloud Computing Technology. Concentration 2 courses consist of Robotics, Artificial Neural Networks, Computer Vision, Intelligent Computing for concentration 2.

Before carrying out the data analysis process, data cleaning is done first by removing the data of students who have not taken the 4 elective courses and ensuring that none of the data to be processed has a null value. After the data selection and cleaning process, 117 student data were obtained to be used in further data processing. If the top 10 data are taken, sample data can be seen in Figure 1.

	Gender	City	ParentEdu	Preference	Course1	Course2	Course3	Course4
0	L	semarang	sma	2	2.75	3.75	3.75	4.00
1	L	semarang	sma	1	4.00	3.25	3.25	3.75
2	L	luar semarang	sma	2	3.75	4.00	3.75	3.75
3	L	semarang	sma	1	4.00	3.75	3.75	3.25
4	L	semarang	d3	2	4.00	3.75	3.00	4.00
5	L	semarang	s1	1	1.75	2.75	0.00	3.75
6	P	semarang	s1	1	3.25	3.75	3.75	4.00
7	P	semarang	sma	1	3.25	3.75	3.00	4.00
8	L	semarang	sma	2	2.75	3.00	4.00	3.00
9	L	semarang	sma	2	4.00	2.75	3.75	4.00

Figure1: Student Initial Data

After getting the initial student data (figure1), the researcher carried out the process of calculating the average score and determining the pass/fail. To be considered successful, the average value of the course must be above 3.25. By using the python programming language, the program code to perform the average course value and determination of success is carried out as in Appendix a.

Appendix a

```
student['AvgScore']=(student['Course1']+student['Course2']+student['Course3']+student['Course4'])/4

def result(RU):
  if (RU > 3.25):
    return '1'
  else:
```

```
    return '0'

student['Pass/Fail']=student.apply(lambda x: result(x['AvgScore']),axis = 1 )
```

For the purposes of the logistic regression process, the ordinal data is converted to numeric (0 and 1 for gender and city, and 1,2,3 for parents' education). So that the data on Gender (X1), City (X2) and ParentEdu (X3) becomes as shown in Figure 2.

	X1	X2	X3	Pass/Fail
0	1	1	1	1
1	1	1	1	1
2	1	0	1	1
3	1	1	1	1
4	1	1	2	1
5	1	1	3	0
6	0	1	3	1
7	0	1	1	1
8	1	1	1	0
9	1	1	1	1

Figure2: Data For Logistic Regression

IV. DATA VISUALIZATION

The purpose of this study was to get an overview of student performance in terms of gender, city, and parents' education. But first it is necessary to look at the results of student performance without paying attention to these factors. With a total of 117 students, 90 students with an average score above 3.25 (Pass), while 27 students with a score below 3.25 (Fail). The success rate is achieved with a percentage of 76.9% as shown in Figure 3.

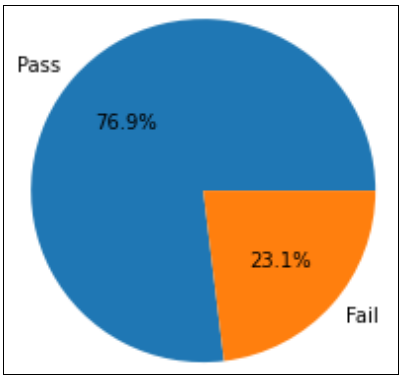

Figure3: Student Success Percentage

5.1 Student Performance Based On Gender

Based on gender, the analysis in this study shows that the female gender has a higher success rate than the male. In this study, there were 22 female (P) students and 95 male (L) students. The success rate for female was 90.91% and the success rate for male was 73.68%. The success diagram can be seen in Figure 4.

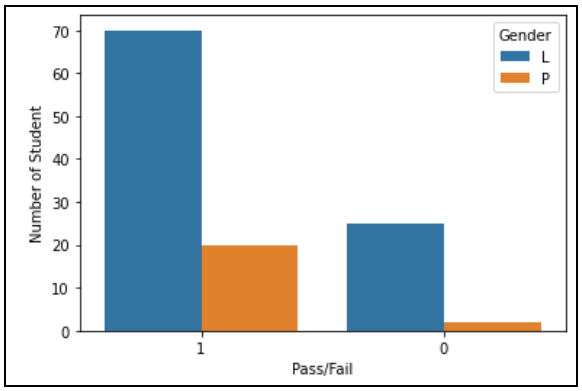

Figure4: Student Performance Based On Gender

5.2 Student Performance Based on City

Based on the city, this research is separated into two categories, namely Semarang and Outside Semarang. There are 99 students from Semarang and the remaining 18 are from outside Semarang. The percentage of success of students from Semarang is 76.77%. While the percentage of success outside Semarang is 77.78%. This shows that the city has no effect on the success rate of students. The success diagram can be seen in Figure 5.

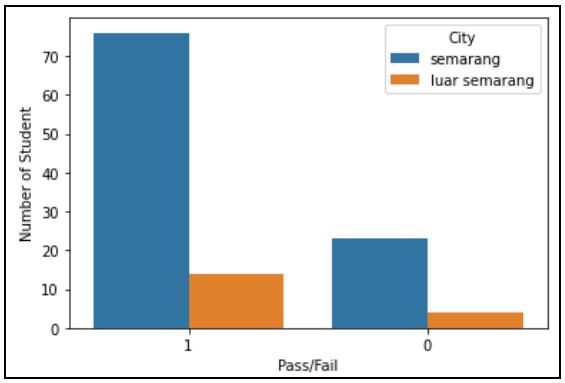

Figure5: Student Performance Based On City

5.3 Student Performance Based On Parents' Education

Based on parents' education, there are 75 High School levels (sma), 5 Diploma levels (d3), and 37 Undergraduate levels (s1). The success rate of students based on parents' education is 81.33% (High School), 80.00% (Diploma), and 67.57% (Undergraduate). This shows that the education of parents does not affect the success of students in the elective courses of the study program. The success diagram can be seen in Figure 6.

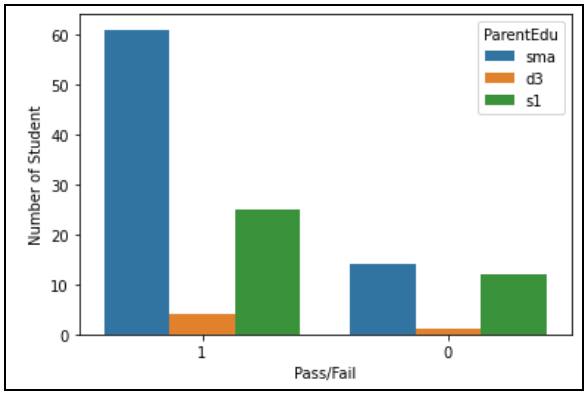

Figure6: Student Performance Based On Parents' Education

V. LOGISTIC REGRESSION

In Machine Learning, classification is one of the most frequently used techniques by researchers. One technique is logistic regression. Logistic regression is a model used to predict whether something is true or false (0 or 1). Logistic regression will measure the relationship between the target variable (which you want to predict) and the input variable (the features used) with the logistic function. In this study, the target variable is Pass/Fail, while the input variables are Gender (X1), City (X2) and ParentEdu (X3). This is in accordance with the research objective, which is to find out whether there is a relationship between student success in completing the elective course with gender, city and level of education of their parents.

For the classification process, train data and test data are needed. In this study, both data were taken from the same source. Split the data using the train_test_split() function. The function refers to sklearn.model_selection. Researchers used 75% of the data for train data and 25% for test data. After separating the data, the next step is to conduct training on the selected data using the logistic regression algorithm. To predict from the test data, the logistic_regression.predict() function is used.

Confusion matrix is often used to measure the performance of a classification model where the output can be in the form of two or more classes. The results can be seen in figure 7. From the classification model, the accuracy is 0.76. Prediction process with logistic regression and accuracy calculation can be explained with program code as in appendix b.

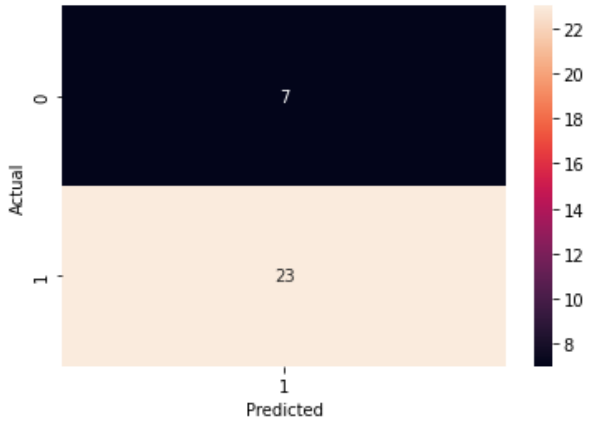

Figure7: Confusion Matrix

Appendix b

```
#set x and y
X = student[['X1', 'X2','X3']]
y = student['Pass/Fail']

#set data test and data train
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.25,random_state=0)

#train dan test
logistic_regression= LogisticRegression()
logistic_regression.fit(X_train,y_train)
y_pred=logistic_regression.predict(X_test)

#count accuracy
print('Accuracy: ',metrics.accuracy_score(y_test, y_pred))
plt.show()
```

VI. CONCLUSION

From the results of the study, it can be concluded that the success of IT students in elective course is good. This can be seen from the success rate of 76.9%. Female students have a higher success rate than male. Students from Semarang and outside Semarang city have almost the same value. Parents' education has no effect on student success. So that the only factor that may have an effect is gender. Logistic regression model provides an accuracy of 76%.

REFERENCES

[1]. Agresti, A. 1996. An Introduction to Categorical Data Analysis. Toronto: John Wiley and Sons Inc
[2]. Hosmer, D.W., Lemeshow, S. 2000. Applied Logistic Regression. 2nd Edition. New Yor: John Willey and Sons.
[3]. Pentury, T., Aulele, S.N., and Wattimena, R. ANALISIS REGRESI LOGISTIK ORDINAL (Studi kasus: Akreditasi SMA di Kota Ambon), Jurnal Ilmu Matematika dan Terapan Maret 2016 Volume 10 Nomor 1.
[4]. Pangaribuan, J.J., Tanjaya, H., Kenichi. Mendeteksi Penyakit Jantung Menggunakan Machine Learning Dengan Algoritma Logistic Regression. Information System Development Volume 6 No. 2 Juli 2021.
[5]. Ciu, T., Oetama, R.S., Logistic Regression Prediction Model for Cardiovascular Disease. IJNMT, Vol. VII, No. 1 June 2020.