JSE Journal of Science and Engineering, Vol. 1, No. 1 (page: 30-34)
e-ISSN: 2723-3944 September 2023
DOI: https://doi.org/10.30650/jse.v1i1.3435

Prediction of Late Tuition Fees at Muhammadiyah University Kalimantan Timur Using the Logistic Regression Method

Taufiqurrahman[1], Taghfirul Azhima Yoga Siswa[2*]
[1,2] Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
* Corresponding Email: tay758@umkt.ac.id

Abstract –One of the funding to finance the university's activities is tuition fees. However, some activities or agendas or procurement of facilities need to be postponed when many students cannot pay the tuition fee in time. This situation will affect the quality of education. Therefore, the purpose of this study is to use student economical data to predict the lateness of students who pay the tuition fee using Logistic Regression. The data is sourced from the Academic Administration and the Financial Administration. There are 12,408 data with several attributes such as the faculty, study program, class, gender, father's income, mother's income, father's education, mother's education, and label (late or not late). Based on the experiment, the accuracy of Logistic Regression to classify the lateness of students to pay tuition fees is 55.89%.

Keywords: Prediction, Logistics Regression, Confusion Matrix, Accuracy

Submitted: 29 Juli 2023 - Revised: 15 September 2022 - Accepted: 26 September 2022

1. Introduction

Tuition fees are one of the things that play an important role in the sustainability of operational activities in a private university. This is the concern of the University Muhammadiyah Kalimantan Timur towards a campus that will continue to grow and create students and students to become the best graduates from all fields so that they can be applied in society or the world of work. University in carrying out its operational activities, of course, also relies on funds from students, one of which is tuition fees. This creates problems if students are late in making tuition payments, because of the tuition financing, the majority of private universities can improve the quality of education and facilities such as computer labs, lecture halls, internet, prayer rooms, libraries and so on.

Some phenomena of student delays in paying tuition fees are very diverse, ranging from the income of parents who is not sufficient for their daily lives even the lack of a sense of responsibility from students towards the accuracy of paying tuition. This is exacerbated by the outbreak of Covid 19, which has made the entire world economy, including Indonesia. Based on the above phenomenon, it is necessary to analyze the prediction of late payment. One of the techniques used is classification with data mining. Classification is the task of assessing data objects to have a choice so that they can be included in a particular class from various existing classes[1]. Data mining is defined as an interaction to discover new relationships, data sets, and patterns with cycles that use static methods, mathematics, artificial intelligence, and

machine learning to obtain and recognize useful data from information linked to large database sets[2].

Several previous studies include research conducted by Ginting entitled "Implementation of the C4.5 Algorithm to Predict Late Payment of School Tuition Fees Using Python". The research was conducted using several variables such as the amount of income, family dependents, parents' educational background and parents' age. With the object of research conducted at SMK Al-Islam Surakarta produces an accuracy rate of (90%)[3]. Then the next research conducted by Muqorobin entitled 'Optimization of the Naïve Bayes Method with Feature Selection Information Gain for Predicting Late Payments' The results of the study using the Naïve Bayes method showed that the results obtained were accuracy of (90%)[4]. As for the research conducted by Rohmayani about Analysis Of Student Tuition Fee Pay Delay Prediction Using Naïve Bayes Algorithm With Particle Swarm Optimation Optimization (Case Study: Politeknik Tedc Bandung), The results of testing the classification model using the highest accuracy Naive Bayes algorithm were tested based on Particle Swarm Optimization (PSO), with the results obtained accuracy of (73.94%)[5].

Based on previous research, the author tries to make an analysis of the Logistic Regression method to predict the accuracy of the prediction of late payment of tuition fees at the Muhammadiyah University Kalimantan Timur, to find out which students are right or late in paying fees with attributes of the name, gender, number, faculty, study program, class, gender, father's education, father's income, mother's education, mother's income, and information (late or not late).

2. Data Mining

Data mining involves collecting important information from extensive data sets using certain techniques. The information obtained from data mining can be used to improve decision-making[6].

Data mining is a process that uses statistical techniques, artificial intelligence, mathematics, and machine learning to identify which information is useful and which knowledge is obtained from various large databases. From the explanation of data mining above, data mining is knowledge stored in a database that has been processed to find forms and machine learning techniques to identify knowledge information obtained through the database[7].

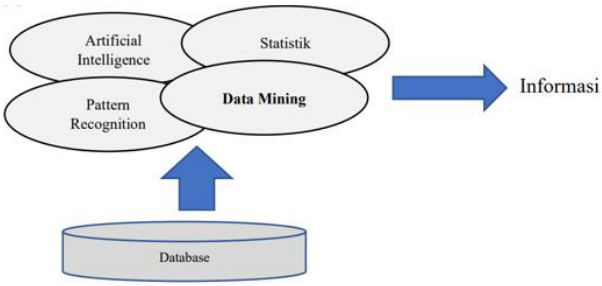

Figure 1. Root of Data Mining

In data mining, there is another term that has the same meaning as data mining, namely Knowledge Discovery in Database (KDD). Data mining and KDD have the same goal: using data that already exists in the database and then processing the data to get new useful information. Knowledge Discovery in Database (KDD) is a knowledge extraction process, pattern/data analysis, archaeological data, and data dredging[8].

3. Resampling

Resampling is a technique used when processing data that is a balance or not. An imbalance in data that occurs in a particular class or category has more data than other classes or categories. The imbalance in the data greatly affects the accuracy of the data classification process[9].

There are two techniques for resampling data: oversampling and undersampling. Oversampling works by introducing a certain amount of data into the minority class. In the undersampling technique, the data in the majority class is reduced[10]. The visualisasion the resampling can be seen in Figure 2.

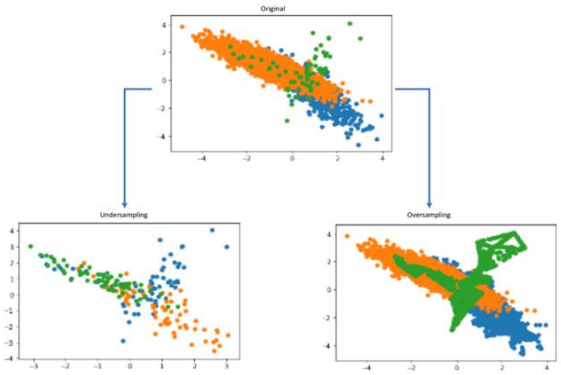

Figure 2. Ilustration Oversampling and Undersampling

4. Logistic Regression

Logistic Regression is a relationship between variables X (Free), and Y (Bound) that does not have a non-linear relationship. The dependent variable is a scale with two categories, for example, yes and no, good and bad, or high and low, which is meant by binary classification using the probability prediction method[11].

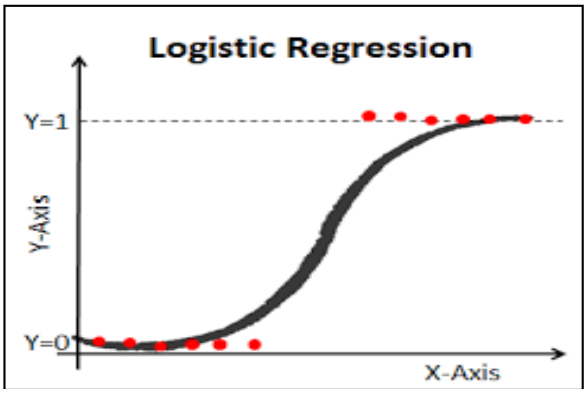

Figure 3. Curve Pattern

Figure 3 shows that if the curve goes positive, the y (output) value will be predicted to be 1. If the curve goes negative, the y (output) value is predicted to be 0. If formulated :

$$p \geq 0.5, class = 1 \quad (1)$$

$$p < 0.5, class = 0 \quad (2)$$

If the number of variables is not limited, the Logistic Regression Equation is expressed by the following formula :

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \quad (3)$$

Or

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (4)$$

e-ISSN: 2723-3944

Change the logarithm (Ln) to exponential (e) or probability Logistic Regression as follows :

$$p(Y = 1) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (5)$$

$$\text{Ln}\left(\frac{P}{1\text{-}P}\right) = \beta_0 + \beta_1 x \quad (6)$$

Description :

Ln : Natural Logarithm

β0 : Constant

β1 : The coefficient of each variable

P : Logistics probability

5. Label Encoder

Label encoder is a transformation method to convert categorical data into numeric form. In the case of Regression, if it contains categorical variables and their values cannot be factored in the form of levels, a dummy process is carried out, and each value in that variable becomes another variable[12]. Examples of changes in the form of this data include the categories of Morning, Afternoon, Afternoon and Evening into numeric data in the form of 1, 2, 3, 4.

6. Evaluation

Evaluation serves to determine the accuracy of the algorithm model created. The evaluation criteria considered are accuracy, standard deviation, f1 score, recall, precision and specificity[14]. In this study, the evaluation was carried out by calculating the accuracy. To get the value of the accuracy results, this study uses a confusion matrix as an evaluation test[15].

Confusion Matrix is a measurement tool used to evaluate performance on a classification model to compare the actual value with the predicted value. When classifying and having true data, the True-Positive and True-Negative values provide that information. If the classifier has an error when classifying the data, then the values of False-Positive and False-Negative will provide that information[16]

Table 1
Confusion Matrix

Class		Prediction	
		TRUE	*FALSE*
Actual	*TRUE*	TP	FP
	FALSE	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad 7)$$

The accuracy value shows how accurate the model is in doing classification.

Description :

TP = The number of positive classes classified as positive.

TN = The number of positive classes classified as negative.

FP = The number of negative classes classified as positive.

FN = The number of negative classes classified as negative.

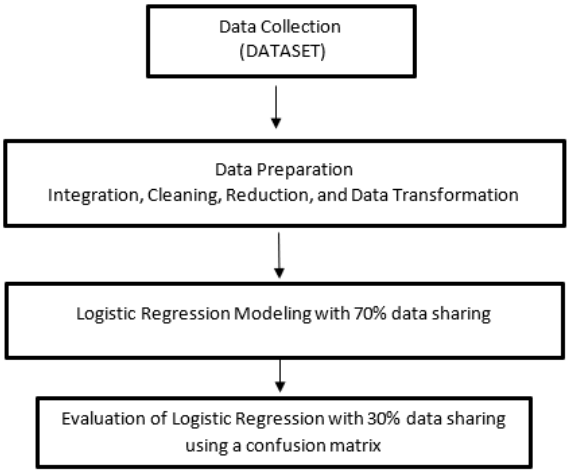

Figure 4. Research Stages

7. Result and Discussion

To solve the problems, the experiment in this research has four stages, as shown in Figure 4. As in Figure 4, this research starts with collecting data from Academic Administration and the Financial Administration. Furthermore, data preparation must be conducted to ensure Logistic Regression can process the data. The Logistic Regression acts as a classifier model to classify the lateness of the student who pays the tuition fees. Then, for evaluation, this research employs a confusion matrix as a tool to calculate the accuracy.

7.1. Data Colection (Dataset)

The data includes student tuition payment data obtained from the Sub of Academic Administration and Financial Administration in the form of student tuition payment data

e-ISSN: 2723-3944

for 2019–2021. The data obtained from the Academic Administration Section amounted to 10,959 with attributes of nim, name, faculty, study program, class, gender, father's income, mother's income, father's education, and mother's education. Then the data obtained from the Financial Administration Section with NIM attributes, names, and labels amounted to 8,833 student data late and 30,811 student data on time. The data used in this study was 39,644 student data.

7.2. Data Preparation

The data that has been obtained from the Sub of Academic Administration and Financial Administration will go through a preparation process to clean the data from previously unstructured data into more structured data to facilitate the classification process. The stages of preparation for the data that has been obtained include data integration, data cleaning, data reduction, and data transformation. Here are the steps :

1) Data Integration

At this step, the data is combined by matching each of the same attributes from the data obtained from the sub of the Academic Section and the sub of the Finance Section. Student data obtained from the Finance Department has 3 attributes: nim, name, and label. After the integration process, it becomes name, nim, faculty, study program, class, gender, father's education, father's income, mother's education, mother's income, and label (late or not late).

2) Data Cleaning

The next step is to clean the data. The cleaned data is data that does not have a complete attribute value (missing value), there is an error in data entry (noise). The previous data amounted to 39644 rows of data, then after being cleaned it became 29,545 rows of data with the exact number of 23,341 data and 6204 data late.

3) Data Reduction

The next step is to balance the existing data. In the initial data, there are 23,341 correct data, while the late data is 6,204, so that the data is not balanced. So it is necessary to balance the data by randomly taking the data of students who make proper tuition payments as much as 6,204 and 6,204 late data, for a total of 12,408 data.

4) Data Transformation

At this step, the label encoder technique is used. The way the label encoder works is to change the category data type to be in the form of numeric data. This is necessary because logistic regression modeling can only work on numerical data. An example of changing the data is changing each value in a column into consecutive numbers such as 1,2,3,4, and so on.

7.3. Modeling and Evaluation

In modeling Logistic Regression, the results are obtained after going through several stages using python programming with a confusion matrix to find the accuracy results. Here is the process:

1) Import Data

```
import pandas as pd

dataset = pd.read_csv('Data_Name.csv')
dataset
```

Figure 5. The process of importing the data

Figure 5 shows the process of importing the data. The data used is 12,408 data. The file to be read is then displayed from index 0 to index 12,407. There are several attributes such as faculty, study program, class, gender, father's income, mother's income, father's education, mother's education, and information (late or not late).

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size = 0.3, random_state=0)

len(X_train), len(X_test)
```

Figure 6. The process of dividing the data

2) Split Data 70:30

Before conducting the evaluation, it is necessary to divide the data by dividing the dataset into 70% training data with a total of 8685 and 30% of testing data with a total of 3723. Figure 6 show the process of dividing the data to data training and data testing.

3) Modeling Logistic Regression

The modelling using Scikitlearn and Logistic Regression approach. The purpose of the value of parameter C is to show the inverse of the regulatory power, which must have a positive float value of 0.01. A smaller value determines a stronger regularization of the modelling be performed. Figure 7 shows the creation of regression model in Python.

```
from sklearn.linear_model import
LogisticRegression

model = LogisticRegression(C=0.01,
solver='liblinear')
model.fit(X_train, y_train)

print(model.coef_)
```

Figure 7. Creating the Model

4) Evaluation

The model's accuracy is a measurement to evaluate how good the model's performance is in classification.

e-ISSN: 2723-3944

Calculate the accuracy at the first stage by constructing the confusion matrix (Table 3). Furthermore, equation 7 calculated the model's accuracy and was found to be 0.5589578297072254, or 55.89%. All the processes in this stage are shown in Figure 8.

```
from sklearn.metrics import accuracy_score,
confusion_matrix

y_pred = model.predict(X_test)

print(confusion_matrix(y_test, y_pred))
print("Hasil Akurasi = ",accuracy_score(y_test,
y_pred))
```

Figure 8. Creating evaluation matrix and Calculating the accuracy.

Table 3 Confusion Matrix

Actual	***Predict***	
	TRUE	*FALSE*
TRUE	997	827
FALSE	815	1084

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{997 + 1084}{997 + 1084 + 827 + 815}$$

$$= \frac{2081}{3723} \times 100\% = 55\%$$

8. Conclusion and Suggestion

This research implements the Logistic Regression algorithm in modelling the prediction of late tuition fees at the Muhammadiyah University of East Kalimantan, resulting accuracy of 55.89%. The accuracy is still relatively low. Optimization methods such as resample and grid search need to be explored for further research.

Acknowledgments

I am enormously grateful to My supervisor for his continuous encouragement and advice throughout my study, and I am thankful to my friends for supporting me.

References

[1] Robi Wariyanto Abdullah, & Kusrini, Prediksi Keterlambatan Pembayaran SPP Sekolah Dengan Metode K-NEAREST NEIGHBOR (Studi Kasus SMK AL-ISLAM Surakarta), Jurnal Informatika Vol. 4/ No. 3, 1-18, Sept. 2019

[2] Agus Bahtiar, Mulyawan, Suryani, & Dindin Firmansyah, Klasifikasi Ketepatan Waktu Pembayaran SPP Di Pondok Pensantren Al-Arifah Menggunakan Algoritma Naive Bayes, Ilmiah Manajemen Informatika dan Komputer, 2549, 1-8, 2017.

[3] V.Ginting, Kusrini, & E.Taufiq, Implementasi Algoritma C4.5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python, Inspiration: Jurnal Teknologi Informasi dan Komunikasi, 10(1), 36-44, Juny. 2020.

[4] Muqorobin, Kusrini, E.Taufiq, Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah, Jurnal Ilmiah SINUS, 17(1), 1693-1173, Jan. 2019.

[5] Dini, Rohmayani Analysis of Student Tuition Fee Pay Delay Prediction Using Naive Bayes Algorithm With Particle Swarm Optimation Optimazation, Jurnal Teknologi Informasi dan Pendidikan, 13(2), 1-8, Sept. 2020.

[6] Santosa, B., & Umam, A. Data Mining dan Big Data Analytics. Media Pustaka.Yogyakarta, 2018

[7] tomo, D. P., & Mesran, M. Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. Jurnal Media Informatika Budidarma, Jurnal Media Informatika Budidarma, 4(2), 437-444, April 2020.

[8] Prasetyo, E. Data Mining : Mengolah data Menjadi Informasi Menggunakan Matlab. Yogyakarta. 2014.

[9] Amelia, R., Fitrianto, KOMPARASI TEKNIK UNDERSAMPLING DAN OVERSAMPLING PADA REGRESI. X(2), 1–11, Dec 2021.

[10] Indrawati, A, Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset. JIKO (Jurnal Informatika Dan Komputer), 4(1), 38–43, April 2021.

[11] Primartha, Rifkie, Algoritma Machine Learning, ed.1, Informatika Bandung, 2021.

[12] R. Tyasnurita, A. Pamungkas, Deteksi Diabetik Retinopati menggunakan Regresi Logistik, ILKOM Jurnal Ilmiah, 12(2), 130-135, August. 2020.

[13] N.R. Sandrianus, & Latipah, Sistem Rekomendasi Tujuan Poli Pada Rumah Sakit Umum Daerah Bajawa Berdasarkan Metode Decision Tree, Jurnal Ilmiah Teknik Informatika, 15(1), 62-74, Mei. 2021.

[14] H.Briliant Argario, N.Hidayat, & R.Kartika Dewi, Implementasi Metode Naive Bayes Untuk Diagnosis Penyakit Kambing, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(8), 2719-2723, August 2018.

[15] Utami, L. A, Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization. Jurnal Pilar Nusa Mandiri, 13(1), 103–112, Maret. 2017.

[16] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. Physical human activity recognition using wearable sensors. Sensors (Switzerland), 15(12), 31314–31338, Dec. 2015.

[17] E. Hary Candana, L A, & Gunadi, Perbandingan Fuzzy Tsukamoto, Mamdini Dan Sugeno Dalam Penentuan Hari Baik Pernikahan Berdasarkan Wariga Menggunakan Confusion Matrix, Jurnal Ilmu Komputer Indonesia, 6(2), 2615-2711, Nov. 2021.

e-ISSN: 2723-3944