Naïve Bayes for Analysis of Student Learning Achievement

N Pandiangan[1]*, *M* Lintang[2], *BA* Priyudahari[3]

[123] Departement of Computer Education, Universitas Musamus, Merauke, Indonesia

Abstract. Student achievement is measured by the achievement index value obtained every semester, student achievement is measured by several factors, and in this research the author takes several factors including study paths, choice of majors, monthly living expenses, relationships with friends, relationships with family, motivation study, employment, scholarships, transportation, and internet services. Analysis and prediction of student achievement using Naïve Bayes Algorithm classification method, the result is this algorithm works very well using 14 student datasets to determine the grades of the 15th student. Based on theAnalysis, variables that affect student achievement include choice of majors, residence, relationships with friends, relationships with family, job, and scholarships. The accuracy of the naïve bayes algorithm for this student achievement case study model reaches 60%, precision 25%, and recall 100%.

Keywords: Naïve Bayes Algorithm, Classification, Information System, student achievement

1 Introduction

In college education, students are required to be able to achievers and become students competing to get good academic achievement, and the achievement index is a benchmark in that regard [1].

The scope of achievement index (IP) is the average credit score of students or the number of student success rates in participating in the teaching and learning process during each semester. If the student achievement index is high, then it identifies that the student is able to attend lectures properly and correctly. But, if the achievement index is low, it means that the student is not able to attend college well. The benefits of a good achievement index are that these students can get scholarships, students can also take undergraduate courses faster because they can take more courses, and students can easily get jobs when they graduate[2].

Based on the literature, there are several writings that say student success in teaching and learning is gender, place of birth, student residence, socioeconomic level, and aspects of nature including basic abilities, attitudes and appearance[3]. Meanwhile, students have learning motivation influenced by intrinsic factor, quality of lectures, weight of lecture material, lecture methods, condition and atmosphere of lecture halls, and library facilities [4], other articles also state that report cards, National Examination scores, entry paths, choice of majors, place of residence, study methods, monthly living costs, student relationships with friends, student ralationships with family, and motivation to study are important factors in influencing student achievement index [1].

From some of the factors that have been mentioned previously, the authors take several variables are college entrance path, major choice, residence, study method, monthly living expenses, relationship with friends, relationship with family, motivation to learn, job, scholarship, transport, internet service as material for analysing student achievement in this research. This research aims to analyse student achievement using the Naïve Bayes Algorithm.

Naïve bayes is one of the classification algorithms in data mining to predict a class label. Data mining functions to search for knowledge in a database to find valid, useful, and understandable data patterns to be used as knowledge [5]. Usually, knowledge is obtained from experts in a particular field and adapted into a computer program to make decision and provide information from reasoning [6]. The data used must also be of quality data and information obtained from procedures such as collection, maintenance, dissemination and good regulation of data[7]. Data mining consists of several groups, are Description, Estimation, Prediction, Classification, Clustering, and Association [8]. Classification is a method to find a model that describes and distinguishes the class of a data concept. The model will be obtained from the analysis of the traing dataset and will be used to predict the class label of an object whose class is not known[9]. Naïve bayes as a classification algorithm that has high accuracy and speed will be used in the classification of achievement and analysis of factors that affect student achievement.

* Corresponding author: nurlela@unmus.ac.id

2 Research Methods

Naïve bayes is one of the classification algorithms, where the classification method is a method to see the behaviour of the grouped variable attributes. Naïve bayes comes from Bayes Theorem which means that the attributes or variables are independent. Naïve bayes is an algorithm that has accuracy and it fast in managing large databases. Naïve bayes utilizes training data to obtain probabilities of attributes or variables that can be used to predict classes in a classification case. Naïve bayes works by looking at the frequency of each classification in the training data and looking for the greatest opportunity from the possible classifications. The advantage of this algorithm is that naïve bayes can work in classification even though it uses a small amount of training data[10].

Naïve bayes is also considered to have good potential in classifying documents compared to other classification methods in terms of accuracy and computational efficiency. Naïve bayes performs a classification by calculating a simple probabilistic based on the number of frequencies and combinations of values from the dataset. Naïve bayes predicts future opportunities based on past experiences

In the naïve bayes algorithm, the dataset must be equipped with an output value or label, so that Naïve Bayes observes the probability of each attribute to determine the output value or label for the dataset to be classified[11].

Equation of naïve bayes algorithm [12]:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Description:

X : data with unknown class

H : hypothesis data is a specific class

P(H|X) : probability Hypothesis H based on condition X (posteriori probability)

P(H) : hypothesis probability H (prior probability)

P((X|H): probability X based on the conditions on the hypothesis H

P(X) : probability X

In this research, the authors used 14 student data for computer education at the University of Musamus as training and testing data. The training data is 9 data, and the testing data is 5 data and will be analysed using the Rapid Miner Application. Students will be classified based on Bad Class and Good Class. Bad class means a class that contains students who have a IPK below 3, while good class means a class that contains students who have a IPK of 3 and above.

The variables or attributes used in this study are as follows:

Table 1. Variable or attributes

Variable or attributes	Definition
College Entrance Path (X_1)	SNMPTN, SBMPTN, or independent,
Major Choice (X_2)	Computer education as the first choice, computer education as the second choice
Residence (X_3)	Boarding house, private house, or other
Study method (X_4)	Alone or group
Monthly living expenses (X_5)	Less than 500.000, between 500.000 – 1.000.000, and more than 1.000.000
Relationship with friends (X_6)	bad, good enough, or good
Relationship with family (X_7)	bad, good enough, or good
Motivation to learn (X_8)	Low, medium, high
job (X_9)	Study while working, or study not while working.
scholarship (X_{10})	Get a scholarship, or didn't get a scholarship
Transport (X_{11})	Private motorbike, taxi, by friends, or other
Internet service (X_{12})	Personal quota and Indihome, only personal quota, only Indihome, or other.

In addition, as material for predicting student achievement, 14 student achievement datasets will be used to predict the 15th student achievement data that does not yet have a class label.

Table 2. The 15th student achievement data

College Entrance Path (X_1)	Independent
Major Choice (X_2)	he second choice
Residence (X_3)	private house
Study method (X_4)	alone
Monthly living expenses (X_5)	between 500.000 – 1.000.000
Relationship with friends (X_6)	good
Relationship with family (X_7)	Good enough
Motivation to learn (X_8)	medium
job (X_9)	study not while working
scholarship (X_{10})	didn't get a scholarship
Transport (X_{11})	Taxi
Internet service (X_{12})	Personal quota
class	**???**

3 Literature Review

Based on the naïve bayes analysis for prediction of the 14th student achievement data, the probability value is good = 3.47 and the probability value is bad = 0.

P = GOOD

{P(P(College Entrance Path=independent|Y=good). P(major choice =second choice | Y= good) . P(residence = private house|Y=good). P(study method = alone| Y= good). P(monthly living expenses = between 500.000-1.000.000| Y= good). P(relationship with friends = good| Y= good). P(relationship with family = good enough| Y= good). P(motivation to learn = medium| Y= good). P(job = study not while working| Y= good). P(scholarship = didn't get a scholarship| Y= good). P(transport = taxi| Y= good). P(internet service = personal quota| Y= good)}

=(8/12)* (8/12)*(4/12) *(4/12)*(5/12)*(10/12)*(1/12)*(4/12)*(9/12)*(2/12)*(7/12)

= 3.47

Fig.1 Good probability

P = bad

{P(P(College Entrance Path=independent|Y=bad). P(major choice =second choice | Y= bad) . P(residence = private house|Y=bad). P(study method = alone| Y= bad). P(monthly living expenses = between 500.000-1.000.000| Y= bad). P(relationship with friends = bad| Y= bad). P(relationship with family = good enough| Y= bad). P(motivation to learn = medium| Y= bad). P(job = study not while working| Y= bad). P(scholarship = didn't get a scholarship| Y= bad). P(transport = taxi| Y= bad). P(internet service = personal quota| Y= bad)}

=(3/3)* (1/3)*(0/3) *(1/3) *(0/3)*(1/3)*(1/3)*(2/3)*(3/3)*(1/3)*(3/3)

= 0

Fig.2 Bad probability

Because the probability good value is greater than the probability bad, the 15th student is included in the good class label.

Based on the analysis using Rapid Miner for each variable, on the variable of the college Entrance Path, more students are in bad class when their college entrance path is independent.

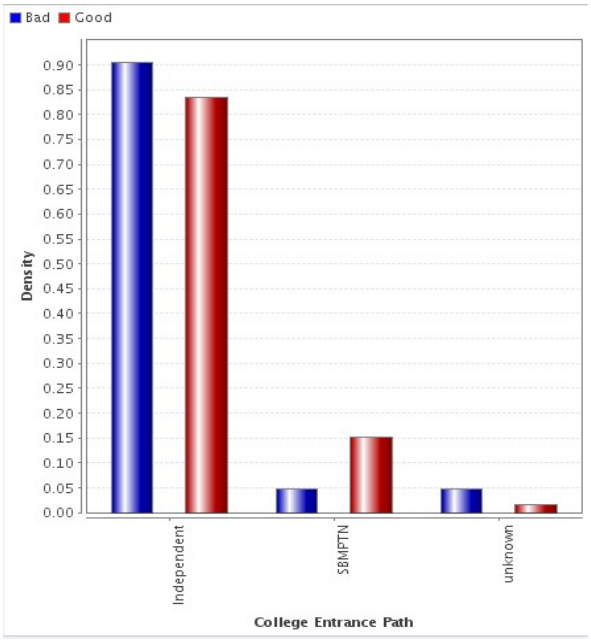

Fig.3 college Entrance Path analysis

While in the Major Choice variable, it can be seen that students who choose computer education as the second choice are students who have good achievements:

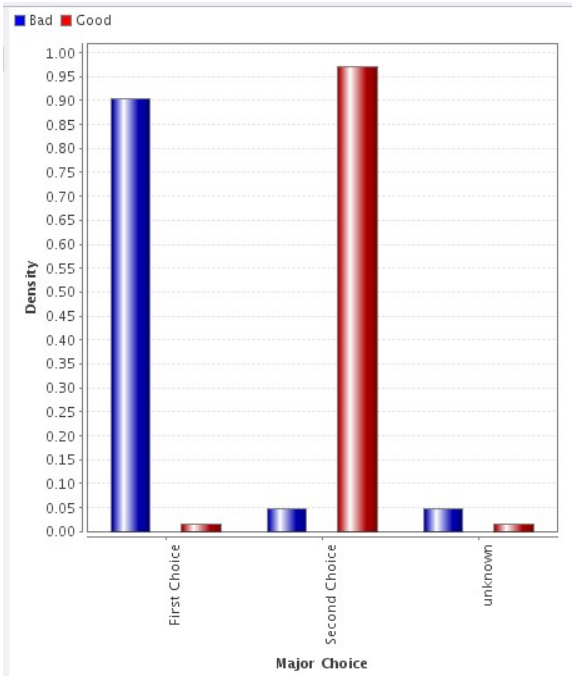

Fig.4 Major Choice Analysis

Based on the place of residence, students who live in boarding houses are more accomplished.

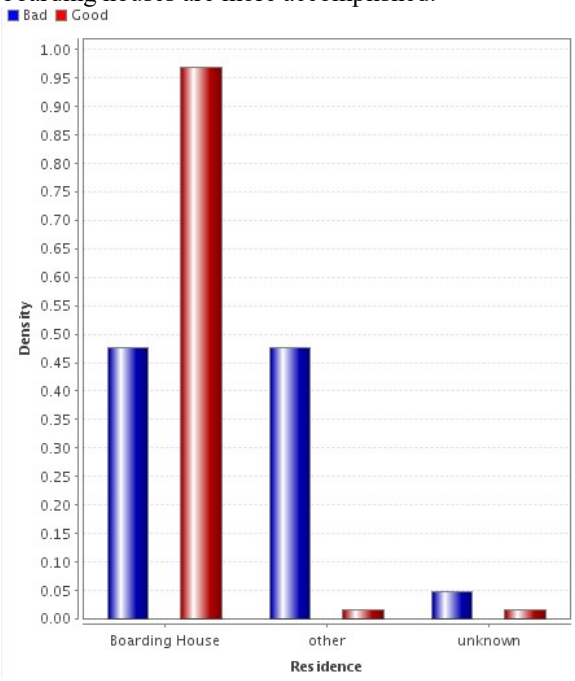

Fig.5 Residence analysis

Based on the study method, students who choose their own learning method are more accomplished.

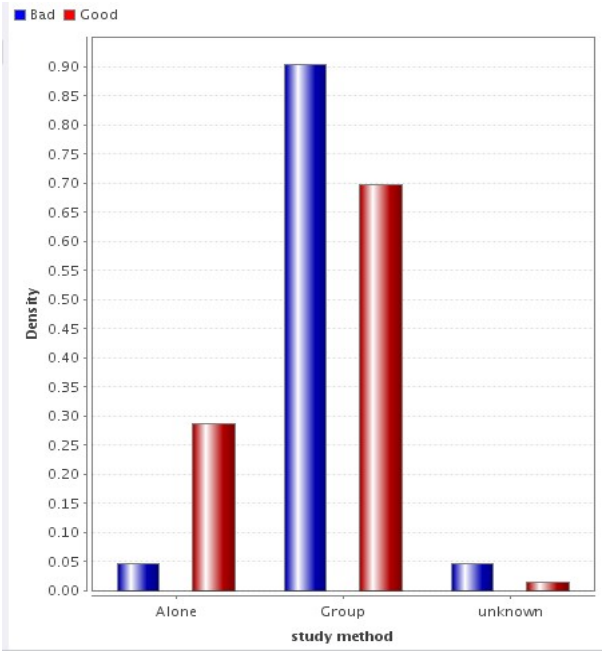

Fig.6 Study method analysis

Based on monthly living expenses, students who have a monthly fee of more than Rp. 1.000.000 more have no achievements.

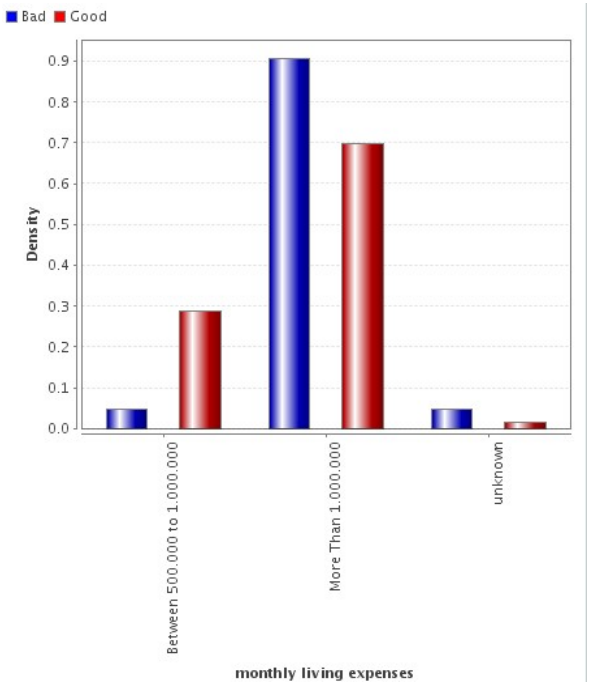

Fig.7 Monthly living expenses analysis

In the analysis of the relationship variables with friends, students are more accomplished and have good grades if they have good relations with their friends.

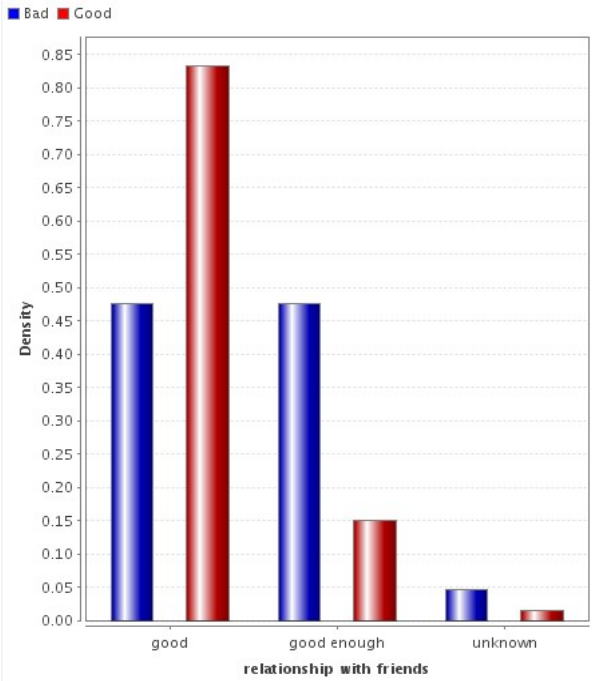

Fig.8 Relationship with friends

Based on the relationship variable with family too, students have good achievements if they have good realtionships with their families.

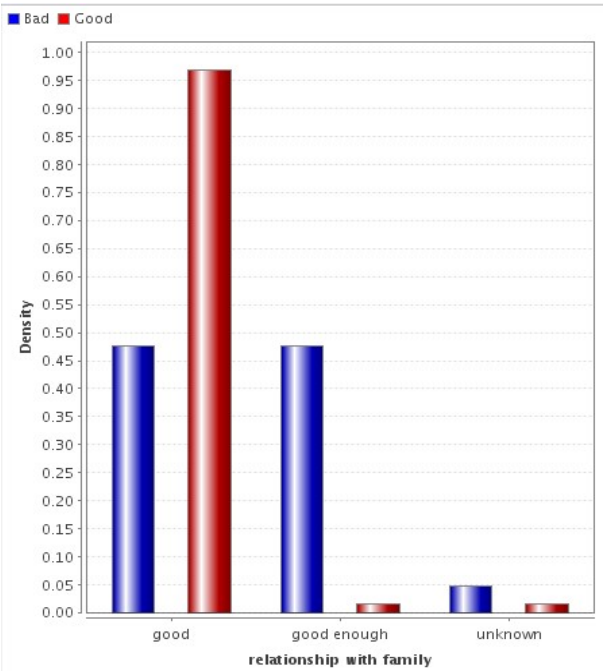

Fig. 9 Relationship with family analysis

Based on the motivation to learn variable, students who have moderate learning motivation are also able to get good achievements and IPK scores.

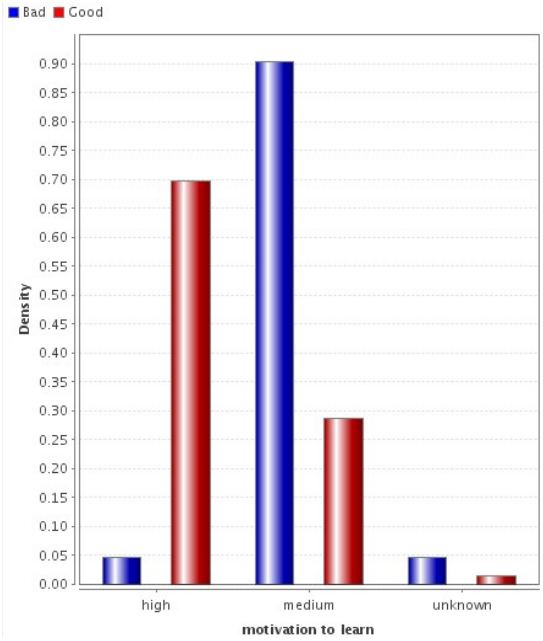

Fig.10 Motivation to learn analysis

Based on the variable studying while working, it can be seen that students who study not while working have higher grades than students who study while working.

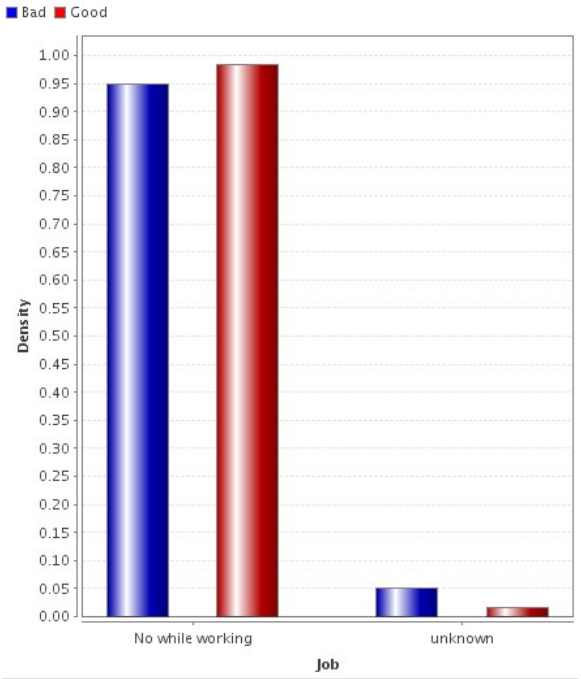

Fig.11 Job analysis

Based on figure 12, many students who get scholarships are more accomplished than students who don't get scholarships:

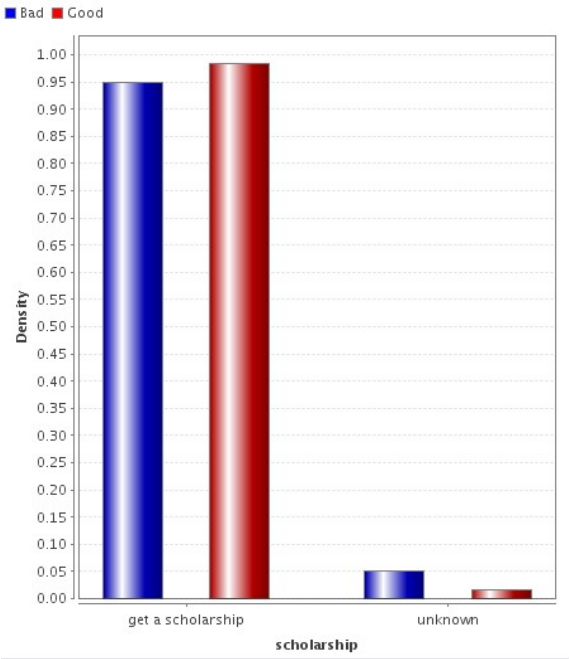

Fig.12 scholarship analysis

For the transportation variable, students who have good achievements and grades are supported by good transportation are private motorbikes or taxi.

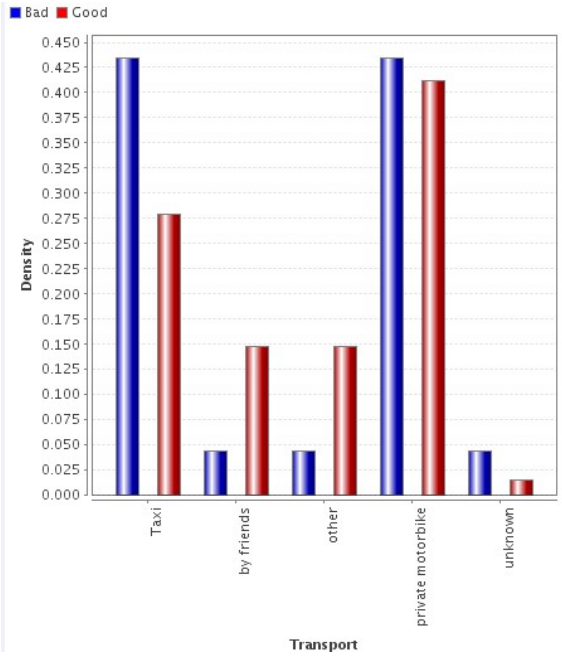

Fig.13 Transport analysis

In the analysis of the last variable is internet service, students must be supported by good internet services and at least have a personal quota to get achievements.

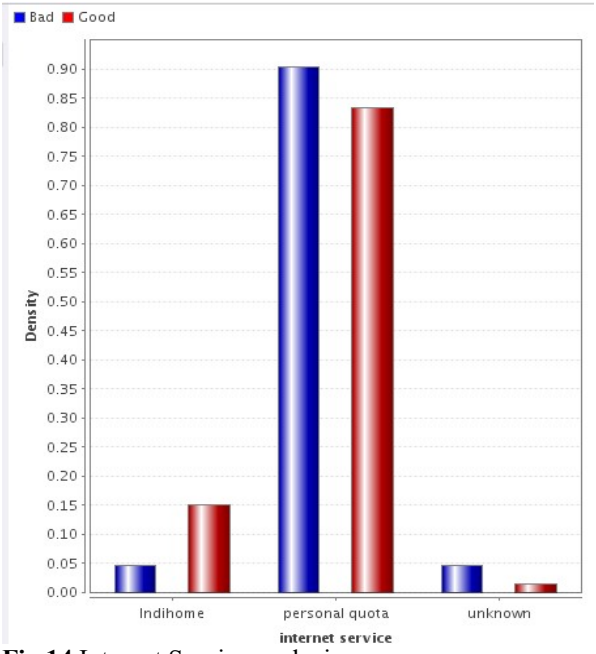

Fig.14 Internet Service analysis

The following is the result of calculating the accuracy of the Naïve Bayes Algorithm on the student achievement dataset model.

Table 3. Performance Vector for the model

Accuracy = 60.00%	True Good	True Bad
Pred. Good	3	1
Pred. Bad	1	0
Precision = 25.00% (positive class: Good)	True Good	True Bad
Pred. Good	1	0
Pred. Bad	3	1
Recall = 100.00% (positive class: Good)	True Good	True Bad
Pred. Good	1	0
Pred. Bad	3	1

4 Conclusion

The result shows the naïve bayes algorithm can be used to predict student achievement based on learning the student achievement dataset. Analysis of variables that affect student achievement has not been fully obtained because the student dataset is small and requires more analysis related to the variables that really affect student achievement according to needs, location of case studies and more in-depth analysis of each student. Based on the research, it was found the variables that affect student achievement include the choice of majors, residence, relationships with friends, relationship with family, job, and scholarships. The accuracy of the naïve bayes algorithm for the case study model of this student achievement reached 60%, precision 25%, and recall 100%.

Acknowledgments

Thanks to The Chancellor of the Musamus University for facilitating and supporting this publication as well as head of the musamus university computer education study program who has given permission to carry out this research.

References

[1] H. Y. Safitri Daruyani, Yuciana Wilandari, "Faktor-faktor yang mempengaruhi indeks prestasi mahasiswa fsm universitas diponegoro semaster pertama dengan motode regresi logistik biner," *Pros. Semin. Nas. Stat.*, pp. 185–193, (2013).

[2] K. Daely, U. Sinulingga, and A. Manurung, "Analisis Statistik Faktor-Faktor," *Saintia Mat.*, vol. **1**, no. 5, pp. 483–494, (2013).

[3] Y. N. Febianti and M. Joharudin, "Faktor-Faktor Ekstern Yang Mempengaruhi Prestasi Belajar Mahasiswa," *Edunomic J. Pendidik. Ekon.*, vol. **5**, no. 2, p. 76, (2018), doi: 10.33603/ejpe.v5i2.246.

[4] A. Pujadi, "Faktor-Faktor Yang Mempengaruhi Motivasi Belajar Mahasiswa: Studi Kasus Pada Fakultas Ekonomi Universitas Bunda Mulia, " vol. **3**, no. 2, pp. 40-51,(2007)

[5] D. Firdaus, "Penggunaan Data mining dalam kegiatan pembelajaran," vol. **6**, no. 2, pp. 91–97, (2017).

[6] M. L. C. Buono, N. Pandiangan, and S. H. D. Loppies, "The Implementation of An Expert System in Diagnosing Skin Diseases Using the Dempster-Shafer Method," *J. Phys. Conf. Ser.*, vol. **1569**, no. 2, (2020), doi: 10.1088/1742-6596/1569/2/022028.

[7] S. H. Dolfi Loppies, F. Xaverius, and N. Pandiangan, "Andorid-based Diet Guide for Diabetes Melitus, Heart, Maag, Kidney, and Impaired Liver Function Disease," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. **1125**, no. 1, p. 012032, (2021), doi: 10.1088/1757-899x/1125/1/012032.

[8] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. **2**, no. 2, pp. 213–219, (2017), doi: 10.22202/ei.2016.v2i2.1465.

[9] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. **10**, no. 2, pp. 421–432, (2019).

[10] T. Imandasari, E. Irawan, A. P. Windarto, and A. Wanto, "Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. **1**, no. September, p. 750, (2019), doi: 10.30645/senaris.v1i0.81.

[11] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," vol. **2**, no. 3, pp. 207–217, (2015).

[12] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. **10**, no. 2, pp. 160–165, (2018), doi: 10.33096/ilkom.v10i2.303.160-165.