Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes

A Saifudin[1*], Ekawati[1], Yulianti[1], T Desyani[1]
[1]Informatics Engineering, Universitas Pamulang, Jalan Raya Puspitek 46, Tangerang, Banten 15310, Indonesia

[*]aries.saifudin@unpam.ac.id

Abstract. Supervision of academic performance is very important to ensure that students can complete their education on time. There have been many proposed applications of machine learning algorithms to predict students' academic performance. Prediction is done by analyzing a dataset of historical academic of the student's grade. The dataset which analyzed has many variables (features), this can increase complexity and decrease model performance because maybe not all features are relevant. We propose to implement the forward selection algorithm to select features that can improve model performance. The result shows that the performance of predictive models of students academic scores can improve with the application of feature selection.

1. Introduction

In the world of Education predicting academic performance is the most important step to find out the quality of students. Institutions can improve academic quality and optimize available resources to help students complete their studies[1]. Steps to help students improve academic performance include, data variables, identification of features or factors that influence learning performance, models in prediction using classification techniques based on easily identified variables, model validation was developed for Universities with achievements[2].

Naïve Bayesian algorithm is a classification of algorithms that have proven its simplicity and efficiency [3]. Naïve Bayes is one of the simplest probabilistic classifiers and often performs very well in many real-world applications, although there is a strong assumption that all features depend on class conditions. Naïve Bayes has better result than KNN (k-Nearest Neighbors) in student performance prediction[4]. The results obtained to determine that the proposed model can significantly improve the appearance of the Naïve Bayes Classifier. In practice, the assumption of attribute independence in Naïve Bayes is often violated on high-dimensional data, so the results are often less than optimal [5].

Most of the datasets contain the number of attributes because the accuracy results may not be much better, so to make the selection of the best result attributes is very important. Forward Selection is one way to determine the most influential attributes in a dataset by negotiating the attributes one by one until the relevant attributes are obtained, because not all attributes are relevant to the problem. Forward Selection is used in the data pre-processing step to select the appropriate features for building models in

data mining[6]. But the Forward Selection algorithm is used to improve prediction accuracy and reduce computational complexity.

The Forward Selection method will be applied to the prediction of student academic achievement using the Naïve Bayes algorithm. Forward selection is used to select attributes that meet the criteria, so only the selected attribute will enter the classification process. It is expected that the selection of attributes using Forward Selection can overcome the problem of class imbalance and improve prediction accuracy.

2. Methodology

This study, the data used are secondary data. General data obtained from the UCI dataset (University of California) makes it easy to compare again with other studies.

The data used are academic and personal data of secondary school students about the value of mathematics in Portugal. The dataset consists of 395 sample data with 33 variables and 1 label to evaluate students' academic performance in mathematics. Detailed information on the attributes of the dataset used is shown in table 1.

Table 1. Attribute information

No	Attribute Name	Description
1	School	school of student containing a GP for Gabriel Pareira, and MS for Mousinho da Silveira
2	Sex	gender of student containing M for Men, and F for Women
3	Age	student ages range from 15 to 22 years
4	Address	the addres of student U for Urban and R for Rural
5	Famsize	student family sizes consist of LE3 for less than or equal to 3 and GT3 for greater than 3
6	Pstatus	the status of living with the parents of students consists of T to live together and A to separate
7	Medu	mother's education of students consists of 0 for none, 1 for basic education (grades 4), 2 for grades 5-9, 3 for secondary education, and 4 for higher education
8	Fedu	father ‘s education of students consists of 0 for none, 1 for basic education (grades 4), 2 for grades 5-9, 3 for secondary education, and 4 for higher education
9	Mjob	the work of mothers of students consists of teachers, related to health care, civil services (administration or police), at home, and others
10	Fjob	the father's work consists of teachers, related to health care, civil service (administration or police), at home, and others
11	Reason	student's reason choose this school because it is close to home, the school's reputation, course choices, and other
12	Guardian	student guardians consisting of mothers, fathers, and others
13	Traveltime	the time to go home to school students consist of 1 for less than 15 min, 2 for 15-30 min, 3 for 30 min-1 hour, or 4 for 1 hour
14	Studytime	weekly student study time consists of 1 for less than 15 minutes, 2 for 15-30 minutes, 3 for 30 minutes-1 hour, or 4 for 1 hour
15	Failures	the number of students who have failed in their class
16	Schoolsup	additional support for education to students
17	Famsup	support for education to students from family
18	Paid	additional paid classes in student subjects
19	Activities	additional school activities
20	Nursery	kindergarten that students have attended
21	Higher	students want to pursue higher education
22	Internet	internet facilities at home
23	Romantic	having romantic relationship

No	Attribute Name	Description
24	Famrel	quality of student family relationships
25	Freetime	free time students get after school
26	Gout	playing with friends
27	Dalc	alcohol consumption on weekdays
28	Walc	alcohol consumption on weekends
29	Health	current health status of students
30	Absences	the number of school absences obtained by students
31	G1	first class period for students
32	G2	second class period of students
33	G3	the final grade obtained by students
34	classification	prediction results obtained

The purpose of this study is to identify relevant factors using the Forward Selection technique and applied to the Naïve Bayes algorithm so that performance comparisons from the classification results can be made before and after feature selection is performed on students academic data. Following is the proposed framework of the model for this prediction shown in f1.

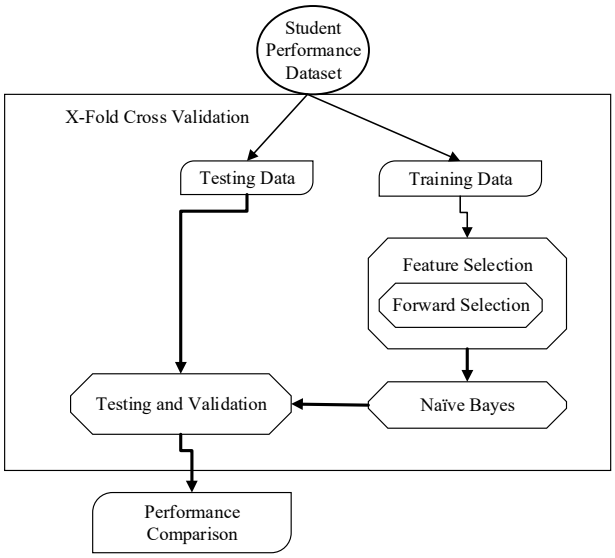

Figure 1. Student performance prediction model

Naïve Bayes is a classification algorithm used to predict future opportunities based on past experience using probability and statistical methods so that they are known as Bayes theorem [5]. Basically Naïve Bayes uses the Bayes theorem with the following general formula:

$$\boldsymbol{P(C|x)} = \frac{\boldsymbol{P(C)P(x|C)}}{\boldsymbol{P(x)}} \quad (1)$$

Where *C* is a class and *x* is a feature value. For features with continuous values it has a Gaussian distribution with mean (μ) and standard deviation (σ)[7]. So, the equation is as follows:

$$\boldsymbol{P(x_i|C_j)} = \frac{\boldsymbol{1}}{\sqrt{\boldsymbol{2\pi}\cdot\boldsymbol{\sigma_j}}}\boldsymbol{e}^{-\frac{(x_i-\mu)}{2(\sigma)^2}} \quad (2)$$

The purpose of this model is to predict student academic performance. Data sets are selected alternately as test data and the others as training data until all datasets have tested data. The distribution of datasets as training data and test data is shown in Figure 2.

Figure 2. Dataset cross validation

In the first validation, the first dataset is used as test data, while the second to fifth dataset is training data. In the second validation, the second dataset is used as test data and the other as trained data. Validation is repeated until all datasets are used as test data.

The validation results are used to measure the performance of the model. Confusion Matrix is a useful tool for analyzing how well the classifier can recognize tuples / features from different classes. Confusion Matrix is a 2 dimensional matrix shown in Table 2 [8].

Table 2. Confusion matrix

Class		Actual	
		True	False
Prediction	True	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

The performance of the model can be seen from the value of Accuracy or AUC. To calculate the performance of a model, the following equation can be used:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \tag{3}$$

$$\boldsymbol{TP_{rate}} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{4}$$

$$\boldsymbol{FP_{rate}} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \tag{5}$$

AUC can be calculated based on the estimated average trapezoidal plane for curves made by TPrate and FPrate [9]. AUC is calculated as a measure of the area of the ROC (Receiver Operating Characteristic) curve using equation [10].

$$\boldsymbol{AUC} = \frac{\boldsymbol{1+TP_{rate}-FP_{rate}}}{\boldsymbol{2}} \tag{6}$$

3. Results and Discussion

To find out the performance of the basic model that applies the Naïve Bayes algorithm as a classification without being optimized, the dataset is applied alternately as test data and training data. The second model is to apply naïve bayes with forward selection.

Model performance is calculated based on the results of validation and the number of attributes using the forward selection. The computation results of the model performance are compiled in Table 3. For visualization, a comparison of model performance is presented using the graph in Figure 3.

Table 3. Model performance

Number of Feature	Accuracy		AUC	
	NB	SFS+NB	NB	SFS+NB
1	85.56%	90.12%	0.857	0.861
2	85.56%	94.43%	0.857	0.919
3	85.56%	94.43%	0.857	0.921
4	85.56%	93.92%	0.857	0.913
5	85.56%	87.84%	0.857	0.870
6	85.56%	87.59%	0.857	0.870
7	85.56%	88.35%	0.857	0.879
8	85.56%	87.84%	0.857	0.874
9	85.56%	87.84%	0.857	0.874
10	85.56%	87.84%	0.857	0.874
11	85.56%	88.10%	0.857	0.876
12	85.56%	88.35%	0.857	0.879
13	85.56%	87.84%	0.857	0.876
14	85.56%	87.34%	0.857	0.870
15	85.56%	88.10%	0.857	0.878
16	85.56%	88.35%	0.857	0.881
17	85.56%	87.08%	0.857	0.864
18	85.56%	87.08%	0.857	0.866
19	85.56%	88.35%	0.857	0.883
20	85.56%	88.35%	0.857	0.885
21	85.56%	87.84%	0.857	0.880
22	85.56%	87.59%	0.857	0.880
23	85.56%	87.59%	0.857	0.876
24	85.56%	86.32%	0.857	0.860
25	85.56%	88.35%	0.857	0.877
26	85.56%	88.10%	0.857	0.881
27	85.56%	87.84%	0.857	0.880
28	85.56%	86.83%	0.857	0.876
29	85.56%	86.58%	0.857	0.874
30	85.56%	86.32%	0.857	0.870
31	85.56%	86.07%	0.857	0.870
32	85.56%	85.82%	0.857	0.861
33	85.56%	85.56%	0.857	0.857

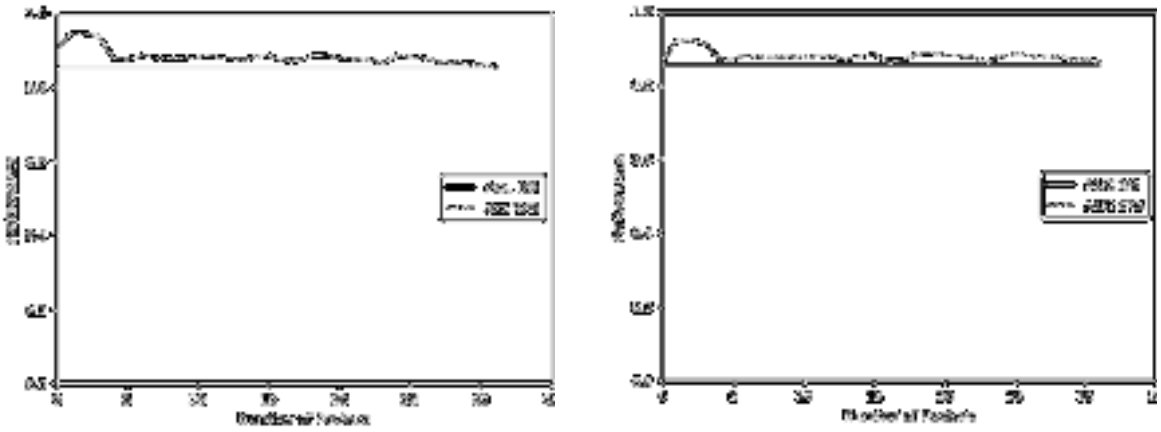

Figure 3. Model accuracy and AUC

The chart in figure 3 shows the performance of the model which implement Naive Bayes classification and the model which integrated Naive Bayes classification and Forward Selection. The chart shows an improvement in performance when the number of features is small. Performance becomes decreased by an increasing number of features.

In table 3, Forward Selection can choose the best feature and improve classifier performance. The highest performance can achieve until 8.86% accuracy and AUC 0.064 when choosing 3 features.

4. Conclusion

When FS is used to select the 3 best features and applied to a prediction model, it can provide better performance. As the number of features selected increases, the performance of the prediction model decreases. It's mean that many features are irrelevant or cause bias. Forward selection can use to choose the best feature and improve the performance of Naive Bayes classifier on students' academic performance prediction. The best performance can reach when Feature Selection chooses 3 features. For further research, it is recommended to use ensemble techniques to reduce misclassification.

References

[1] Lei C and Li K F 2015 Academic Performance Predictors *Proc. - IEEE 29th Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2015* 577–81

[2] Devasia T, Vinushree T P and Hegde V 2016 Prediction of students performance using Educational Data Mining *Proc. 2016 Int. Conf. Data Min. Adv. Comput. SAPIENCE 2016* 91–5

[3] Karthika S and Sairam N 2015 A Naïve Bayesian classifier for educational qualification *Indian J. Sci. Technol.* **8**

[4] Amra I A A and Maghari A Y A 2017 Students performance prediction using KNN and Naïve Bayesian *ICIT 2017 - 8th Int. Conf. Inf. Technol. Proc.* 909–13

[5] Kaviani P and Dhotre S 2017 International Journal of Advance Engineering and Research Short Survey on Naive Bayes Algorithm 607–11

[6] Zaffar M, Hashmani M A, Savita K S and Rizvi S S H 2018 A study of feature selection algorithms for predicting students academic performance *Int. J. Adv. Comput. Sci. Appl.* **9** 541–9

[7] Jain M M and Richariya P V 2012 An Improved Techniques Based on Naive Bayesian for Attack Detection *Int. J. Emerg. Technol. Adv. Eng.* **2** 324–31

[8] Lin L C, Yeh Y C and Chu T Y 2014 Feature selection algorithm for ECG signals and its application on heartbeat case determining *Int. J. Fuzzy Syst.* **16** 483–96

[9] Dubey R, Zhou J, Wang Y, Thompson P M and Ye J 2014 NeuroImage Analysis of sampling

techniques for imbalanced data : An n = 648 ADNI study *Neuroimage* **87** 220–41
[10] López V, Fernández A and Herrera F 2014 On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed *Inf. Sci. (Ny).* **257** 1–13