A Study of Academic Performance using Random Forest, Artificial Neural Network, Naïve Bayesian and Logistic Regression

Nurissaidah Ulinnuha1, Halimatus Sa'Dyah1, Muwardi Rahardjo1
1,2Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember
1nuris08@cs.its.ac.id

Abstract

Student admission still be an important problem of educational sector. There are several techniques that has been studied in order to solve admission problem. Hence, we choose 4 technique to be compared. In the present study, we compare random forest algorithm with artificial neural network, naïve Bayesian and logistic regression. Here we found that the composition of training dataset is important. We also found that random forest is often over fit and in accuracy is not better than other methods for dataset with small feature.

Keyword: Academic Performance, Random Forest Artificial Neural Network, naïve Bayesian, Logistic Regression.

1. Introduction

Random forest is a collection of decision trees built up with some element of random choice [1]. Literature points out the potential of random forest for classification, prediction and variable selection problem. Several interesting problem have been examined using random forest and it is evident that this technique has significant potential in providing useful classification model. Comprehensive review of applications of random forest have been provided by Galiano et al., Granitto et al., and by Genuer et al. [2] [3] [4]. Also, a number of researches have compared the performance of other data mining technique and random forest in different kinds of problems [5] [6] [7]. The present study aims to compare the performance of random forest and other technique to predict GPA score.

Over several past admission period, the number of applicants of graduate program always exceeds available seats. Therefore, it is important to have an adequate technique for measuring the potential of applicants especially their academic performance. Many researchers have studied the problem of prediction predicting academic performance in Educational institution.

Educational institution has many points of assessment when making admission decision. Some of points are overall undergraduate point average, graduate management, admissions test score, personal interview, work experience, age, sex, references, goals statement, and others [8]. Kabra and Bickhar [9] used decision tree to classify and predict students whether fail or success. They concluded that decision tree is successfully able to predict future student academic performance based on past one. Ngoc Hien and Haddawy [10] used bayesian network to predict student GPA and concluded that bayesian network has good performance for complete and balanced data. Zimmerman et al [11] used random forest decision tree to predict graduate program student academic performance based on its undergraduate academic performance. They found that 3rd year GPA on undergraduate program has high significant value in order to predict graduate GPA.

Several past study show that each technique has its own nonconclusive superiority. Due to importance of academic performance problem in the literature, this study focuses on comparison between random forest and other prediction techniques. In this study, we will analyze student performance using random forest decision tree, artificial neural network, Bayesian network and Logistic Regression thus compare the performance of these techniques.

In this paper, we deliver the result of our research in 5 sections. Section 1 introduces our research problem and objective. Section 2 describes several prediction techniques that are used in this study. Section 3 explains about the result of research. Section 4 explains the conclusion of research.

2. Review

This section briefly describes about prediction techniques used in this study. Those techniques are Artificial Neural Network (ANN), Bayesian Network, Logistic Regression and Decision Trees.

2.1 Artificial Neural Network

Artificial neural network is a computational model of that adopting from properties of neural networks. Neural network receives signals through synapses in the dendrites. When the received signal is strong...
enough (over threshold), then the neuron is activated and the signal on the axon is generated. This signal can be sent to other synapses and may activate other neurons.

ANN is generally composed from input(synapses) are multiplied by weight(the strength of each signal) is then calculated by a mathematical function that determines the activation of the neuron of the function will produce an output, which is described at figure 2.1:

\[
\text{output} = \text{sgn}(\sum_{i=1}^{n} \text{input}_i - \phi)
\]

If output = \text{sgn}(\sum_{i=1}^{n} \text{input}_i - \phi) < 0
then output = 0
if output = \text{sgn}(\sum_{i=1}^{n} \text{input}_i - \phi) >= 0
then output = 1

\[\phi = \text{threshold}\]

The resulting output is highly dependent on weight of each input. There are certain algorithms that can be used to obtain the weight value, which is where this process is called training or learning.

ANN has 3 different types namely Single-Layer Neural Network, Multi-Layer Perceptron and Recurrent Neural Network. In this study, we use Multi-Layer Perceptron.

Multi-Layer Perceptron is a kind of ANN that has high robustness in a complex problem such as linguistic problems, pattern recognition, and image processing. Multi-Layer Perceptron has 3 layers namely input layer, hidden layer and output layer. The architecture of multi-layer perceptron is shown in figure 2.2.

2.2 Logistic Regression

Logistic regression is one method used to determine the effect of one or more independent variables(\(x\)) to one dependent variable(\(y\)) that is a binary or continue with a single dependent variable is binary. Criteria variables are used:

- The dependent variable is a dummy variable that should have only two alternatives, such as yes or no, 1 or 0.
- The independent variables have interval or ratio scale data.

Logistic regression has standard equation shown on equation 2.1

\[Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\]  \hspace{1cm} (2.1)

In order know how independent variable contributes to determine the value of dependent variable, we can use log likelihood analysis and Wald analysis.

2.3 Naive Bayes

Naive Bayes is one of the inductive learning algorithms that effective and efficient for machine learning and data mining. These algorithms use the Bayesian probability of predicting the probability of the future based on past experience. Bayes theorem is a theorem in statistics used to calculate the probability of a hypothesis. While the optimal Bayes classifier calculates odds of a class from each group of attributes that exist and determine which one is most optimal class.

Based on the Bayes rule, the chances of \(E\) in the form of a set of attribute values of attributes(\(x_1, x_2, x_3, x_4 \ldots x_n\)) in a class can be formulated as follows:

\[P(C|E) = \frac{P(E|c)P(c)}{P(E)}\]  \hspace{1cm} (2.2)

To determine the choice of the class, use the maximum opportunities from all \(c\) in \(C\) with the function:

\[\arg \max_{c \in C} \frac{P(E|c)P(c)}{P(E)}\]  \hspace{1cm} (2.3)

Since \(P(E)\) is constant for all classes, then \(P(E)\) can be ignored so as to produce the function:

\[f_c(E) = \arg \max_{c \in C} P(E|c)P(c)\]  \hspace{1cm} (2.4)

Naive Bayes uses the concept of independence of attributes so that the need for the amount of training data needed to be much less.

\[f_c(E) = \arg \max_{c \in C} P(c) \Pi_{j=1}^{n} P(E_j | c)\]  \hspace{1cm} (2.5)
Naive bayes illustration:

![Naïve Bayes Illustration](image)

**Figure 2.3 : Naïve Bayes Illustration** Implementation of the naïve Bayesian filtering is the process of email, the multinomial logistic regression for the determination of a student majors, etc.

2.4 Random Forest

Random forest is a collection of decision trees built up with some element of random choice [1]. Literature points out the potential of random forest for classification, prediction and variable selection problem.

Random forest works by generating a number of trees to analyze the data then it combine all the output from tree and then through the process of vote (look for the classes who have the majority) to obtain the final result.

Random forest has high robustness for large data but it consumes much cost than other techniques.

3. Data Analysis

In present days, at educational system a student's performance is determined by GPA value. The main thing of its process is to predict GPA from students attributes such as their semester, gender, etc.

The dataset used in this study was obtained from academic institution of Institut Teknologi Sepuluh Nopember. Data collected from students from 2002 to 2009 who had graduated from the master. All the predictor and response variables are given in Table 1 for reference.

<table>
<thead>
<tr>
<th>No</th>
<th>Variable Name</th>
<th>Information</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Marital Status</td>
<td>Marital status when take magister college</td>
<td>0 = not married</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 = married</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>Gender of magister student</td>
<td>0 = woman</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 = man</td>
</tr>
<tr>
<td>3</td>
<td>Scholar University</td>
<td>Rating of university with scale 1-10 from survey of Webometrics</td>
<td>10 = 35 big first rank</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9 = 35 big second rank, etc</td>
</tr>
<tr>
<td>4</td>
<td>Period of Study</td>
<td>Time period for study</td>
<td>Nominal (2-4 years)</td>
</tr>
<tr>
<td>5</td>
<td>Work Status</td>
<td>Work status when take magister college</td>
<td>0 = not work</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 = work</td>
</tr>
<tr>
<td>6</td>
<td>Scholar GPA</td>
<td>GPA value at scholar</td>
<td>Nominal (0-4)</td>
</tr>
<tr>
<td>7</td>
<td>Age (new student)</td>
<td>Age when take magister college</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

Table 1. Information of dataset

Missing values of an attribute was cleaned by filling the value with average of fields on attribute. Meanwhile, the missing values that found at class involve that all field corresponding with that field of class was deleted.

To simplify prediction process, GPA as response variable was divided into two class. First class is for GPA less than 3.5 and the second one is for GPA equal or more than 3.5.

The software program chosen for the analysis is WEKA. WEKA provides state-of-the-art modeling methods including Random Forests, Logistic Regression, Multilayer Perceptron and Naïve Bayes.

3.1 Neural network analysis

This analysis considers a three layer feed forward. The input layer contains 18 nodes, 1 hidden layer with 9 nodes, that is calculated by number of attributes + number of classes / 2 and the output layer contains one node. It was decided to have 9 nodes in hidden layer because its performance was the best when comparing with the result of other number of nodes in hidden layer. The unipolar sigmoid activation is used at the hidden layer and at the output layer. The learning rate and momentum term are set to 0.3 and 0.2 respectively.

The result on Table 2 shows that, performance of Neural Network is quite good. Our experiment
dataset probably noisy and it influence NN performance [12].

3.2 Naïve Bayes analysis

A naïve Bayes classifier is based on the application of Bayes theorem that may predict class for classification [13]. Consider a set of training samples where each sample is made up from m discrete-valued attributes and a class from a finite set C. The naïve Bayes classifier may predict the class of an unknown sample using the available training set to calculate the most probable output.

The performance of the naïve Bayes method in GPA prediction depends on how the probability densities are estimated and how priors are distributed. Estimation of probability densities is important for the calculation of the likelihood and for estimation of the posterior distribution of GPA [14].

3.3 Logistic Regression analysis

Logistic regression considers Master GPA as independent variable and as marital status, gender, scholar university, period of study, work status, scholar GPA and age as dependent variable \( (x_1, x_2 \ldots x_7) \). The estimated model is:

\[
-7.8924 - 0.3235 x_1 + 0.2201 x_2 + 0.0716 x_3 + 1.1903 x_4 - 0.3563 x_5 + 0.3414 x_6 + 0.0836 x_7
\]  

(3.1)

These coefficients were estimated using maximum likelihood estimation with ridge parameter of 1.0E-8. The model found that period of study is the most influence variable to predict master GPA of student.

3.4 Random Forest analysis

In this study, we build 10 trees as element of random forest and we do 10 experiments for knowing its performance. The result of the experiment could be seen on table 2.

Here, the accuracy of random forest statistically low but in table 2, when the experiment use training set as input, the accuracy is 100%. It means that random forest is still over fit with data training.

In this study, the result of several runs is served in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Random Forest</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Use Training Set</td>
<td>Cross Validation</td>
<td>Percentage Split</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
<td>100%</td>
<td>68.27%</td>
<td>68.57%</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>0%</td>
<td>42.31%</td>
<td>31.43%</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>1</td>
<td>-0.0732</td>
<td>0.2159</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0933</td>
<td>0.3952</td>
<td>0.3686</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.1431</td>
<td>0.5162</td>
<td>0.5235</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>22.63%</td>
<td>95.82%</td>
<td>84.41%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Naïve Bayes</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Use Training Set</td>
<td>Cross Validation</td>
<td>Percentage Split</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
<td>76.92%</td>
<td>75.96%</td>
<td>68.57%</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>23.08%</td>
<td>24.04%</td>
<td>31.43%</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.2811</td>
<td>0.2415</td>
<td>0.2159</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.2585</td>
<td>0.2737</td>
<td>0.32</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.4504</td>
<td>0.4611</td>
<td>0.4952</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>62.73%</td>
<td>66.35%</td>
<td>73.28%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Use Training Set</td>
<td>Cross Validation</td>
<td>Percentage Split</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
<td>77.88%</td>
<td>75.96%</td>
<td>65.71%</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>22.12%</td>
<td>24.04%</td>
<td>34.29%</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.3197</td>
<td>0.2786</td>
<td>0.1286</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.3343</td>
<td>0.3658</td>
<td>0.3988</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.408</td>
<td>0.4447</td>
<td>0.452</td>
</tr>
</tbody>
</table>
4. Conclusion and Future Works

In this study, we have two point of conclusion. Hereby our conclusion:

1. Data training composition influence the performance of classifier technique.
2. Random forest analysis is over fit for some dataset.
3. Random Forest in accuracy is not better than other methods for dataset with small fitur.

Neural network still give better performance than others in several testing set. In addition, naïve bayes and logistic regression have same performance for this dataset.

In the future, we will try to improve the result of random forest by doing variable reduction using Principal Component Analysis to solve noisy dataset problem.

5. References

[12] Yu Zhou and Yali Wu. Analyses on Influence of Training Data Set to Neural Network Supervised