TITLE : **STUDENTS PERFORMANCE IN EXAMS**

DONE BY : **NEDHUNURI ROHITH REDDY**

Abstract

The ultimate goal of any educational institution is offering the best educational experience and knowledge to the students. Identifying the students who need extra support and taking the appropriate actions to enhance their
performance plays an important role in achieving that goal. In this research, four machine learning techniques have been used to build a classifier that can predict the performance of the students in a computer science subject that is offered by Al-Muthanna University (MU), College Of Humanities. The machine learning techniques include Artificial
Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression. This research pays extra attention to the effect
of using the internet as a learning resource and the effect of the time spent by students on social networks on the students' performance. These effects introduced by using features that measure whether the student uses the internet
for learning and the time spent on the social networks by the students. The models have been compared using the ROC
index performance measure and the classification accuracy. In addition, different measures have been computed such
as the classification error, precision, recall, and the F measure. The dataset used to build the models is collected based
on a survey given to the students and the students' grade book. The ANN (fully connected feed forward multilayer ANN) model achieved the best performance that is equal to 0.807 and achieved the best classification accuracy that is equal to 77.04%. In addition, the decision tree model identified five factors as important factors which influence the
performance of the students.

Introduction

The economic success of any country highly depends on making higher education moreaffordable and that considers one of the main concerns for any government. One of the factors thatcontributes to the educational expenses is the studying time spent by students in order to graduate.For example, the loan debt of the American students has been increased due to the failure of manystudents in getting graduated on time [1]. Higher education is provided for free to the students inIraq by the government. Yet, failing of graduating on time costs the government extra expenses.

To avoid these expenses, the government has to ensure that the student graduate on time. Machinelearning techniques can be used to forecast the performance of the students and identifying the atrisk students as early as possible so appropriate actions can be taken to enhance their performance.One of the most important steps when using these techniques is choosing the attributes or thedescriptive features which used as input to the machine learning algorithm. The attributes can becategorized into GPA and grades, demographics, psychological profile, cultural, academic progress, and educational background [2]. This research introduces two new attributes that focus on to the effect of using the internet as a learning resource and the effect of the time spent by students on social networks on the students’ performance. Four machine learning techniques, fully connected feed forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been used to build the machine learning model. ROC index has been used to compare the accuracy of the four models. The dataset used to build the models is collected from the students at the College Of Humanities during 2015 and 2016 academic years using a survey and the student’s grade book. The dataset has the information of 161 students. The activities of this research include feature engineering to create the students dataset, data collecting, data preprocessing, creating and evaluating four machine learning models, and finding the best model and analyzing the results.

Literature Review

Much research has been done in the area of educational data mining where a predictive model is built to forecast the performance of students to identify the at risk students. This problem can be considered a hard problem because the performance depends on many characteristics related to the students. These characteristics can be categorized into student's GPA and grades, demographics, psychological profile, culture, academic progress, and educational background [2]. The student's GPA is the most important attribute used to predict the performance. The GPA can represent the real value for the future educational and career possibilities and progression. In addition, the academic potentials can be evaluated by the student GPA. The demographics information that consists of the family background, the gender, disability, and age is also considered an important attribute [3]. This research introduces two new attributes that focus on using descriptive features related to the internet and social network usage and their effect on the performance. On the other hand, many machine learning and data mining techniques have been used to predict the students' performance such as: Artificial Neural Network (ANN); K-Nearest Neighbor (KNN); Support Vector Machine (SVM); Linear Regression; Logistic Regression; Decision Tree (DT); Random Forest (RF); Principal Component Analysis (PCA); Naïve Bayes (NB); Neuro-Fuzzy classification (NF); Decision List (DL); Bayesian Network (BN); and Discriminant Analysis (DA).

Artificial Neural Networks

Artificial Neural Network represents a set of input unites and output unites that are connected to each other by weighted connections. The ANN learns by changing the weights of the connections in a way so it is able to predict the right target label for some input data instances. One of the famous learning algorithms used to train the ANN is Backpropagation Algorithm. ANN has many advantages such as its high resistance to noisy datasets and its well performance on classifying patterns that has not been trained on so it’s used in situations when there is a little knowledge of the relation between the class label and the features in the dataset. There are many real world applications of the ANNs such as image and handwritten recognition, speech recognition, laboratory medicine and pathology. There are many types of the ANNs which can be classified based on their architecture and design. One type is a fully connected multilayer feed forward ANN in which the network has an input layer, one or more hidden layers, and the output layer. In addition, its connections never cycle back to an input unit or to an output unite located in the previous layer. Also, each unit in a layer L provides input to each unit in the layer L+1.   A three layer fully connected feed forward ANN has been used in this research. The network consists of an input layer, two hidden layers, and the output layer. The input layer has twenty input unites, neurons, while the first hidden layer has six hidden unites. The second hidden layer has three hidden unites. The fourth layer is the output layer which has only one output unite. The Rectifier Linear Unit has been used as the hidden unites' activation function .

The Experiment

Dataset and Data Sources

The dataset used in this research is collected from the Archeology department and the Sociology department of the college of Humanities at Al-Muthanna University during the 2015 and 2016 academic years. Two data sources have been used, survey collected from the students and the students' grades data records. The dataset contains 161 student records, 76 male and 85 female. The dataset contains twenty attributes. The attributes can be divided into five categories which are personal and life style, studying style, family related, educational environment satisfaction, and student's grades. Table2 shows the attributes used in order to construct the dataset. Each student has been labeled as Weak or Good based on his/her final grade in the computer science subject. The weak student is the student who has a final grade less than sixty out of 100. On the other hand, the Good student is the student who has a final grade equal or greater than sixty.

Validation Method and Accuracy and Performance Measures

In this research, three folds cross validation method has been used. In this method, the dataset is divided into three equal size sets. The learning and testing are executed three times. At each fold or execution, the machine learning algorithm selects one set to be the test set and the remaining two sets as the training sets. The accuracy and the performance measures is aggregated over all the folds in order to calculate the final performance and the final accuracy of the model. The ROC index, the area under the curve, performance measure has been used to evaluate the performance of the classification models. This measure is a well-known measure that is relying on the ROC curve and it is calculated by using the prediction scores. Equation 3 is used to calculate the ROC index [21]. In addition to the ROC index, many important measures have been used such as the accuracy, the classification error, and the F Measure. Equation 4 is used to calculate the F Measure. The F Measure is a useful alternative to the misclassification rate measure.

= ∑ (([ ])− ([−1]))×(([ ])+ ([−1])) 2 ⁄| | =2
(3)

Where | | represents the number of thresholds that are used, ([ ]) represents the false positive rate at the threshold i, and ([ ]) represents the true positive rate at the threshold i. A larger ROC index indicates a better classification model. A model with ROC index above 0.7 considered a strong model while a model with ROC index below 0.6 considered a weak model. [21].

= 2∗ ( )by(+)
= ( )by(+)
= ( )by(+)

TP, True Positives, is the number of data rows in the test set which had a positive target and that were predicted to have a positive target. TN, True Negatives, is the number of data rows in the test set that had a negative target and that were predicted to have a negative target. FP, False Positives, is the number of data rows in the test set which had a negative target but that were predicted to have a positive target. FN, False Negative, is the number of data rows in the test set that had a positive target but that were predicted to have a negative target .

Models Implementation

All the models have been implemented by the RapidMiner Studio software. A Cross Validation operator has been used in order to execute the three folds validation operations during the training and the testing phases. The operator sampling property set to linear sampling. In order to find the best set of the models' parameters, Optimize Parameters (Grid) operator has been used. The ANN operator has been configured to use the Rectifier activation function and the number of hidden layers sizes set to be 6 and 3 consecutively. The ANN model used 100 epochs in the training phase. All the other parameters has been set to the default values. The Optimize Parameters operator has been set to find the best value of the learning rate and the L2 regularization. For the learning rate and the L2 regularization, the configuration set to use 100 steps on a linear scale from 0 to 1. For building the DT model, the Optimize Parameters operator has been set to find the best value of the splitting criterion, and the minimal size for split properties. Also, apply pruning property has be set by the optimization operator. All the other parameters has been set to the default values. The Logistic Regression operator has been set to use regularization and the optimization operator set to find the best value for the solver method and the lambda. The lambda search property set to use 60 steps on a linear scale starts from 0 to 1.797. All the other parameters has been set to the default values. For building the Naïve Bayes model, the optimization operator has been set to find the best values for the Laplace correction, the estimation mode, using the application grid, the bandwidth selection, the number of kernels, and the size of application grid. The number of kernels search property set to use 10 steps on a linear scale starts from 1 to 20. The application grid size search property set to use 10 steps on a linear scale starts from 1 to 40.

THE RESULTS

Based on the given inputs like gender,race,parental level of education,lunch,math score,reading score,writing score,math passstatus,reading passstatus,writing passstatus,overall status,total score,percentage the output is grade.

The Conclusion

To solve the problem of identifying the students who have a poor academic performance in the computer science subject offered by Al-Muthanna University, College Of Humanities, four classification models have been built to predict the performance of the students. Four machine learning techniques, fully connected feed forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been used. The models have been compared to one another using the ROC index performance measure and the classification accuracy. ANN model has the highest ROC index that equals to 0.807 and accuracy of 77.04. In addition, the decision tree model showed that not all the attributes involve in the classification process. Computer Grades-Course1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency are the attribute used by the decision tree model.