Backlog Prediction using Classification Techniques of Machine Learning

Aditi Parikh
Department of Computer Science & Engineering
Jaipur Engineering College and Research Center

Neelam Chaplot
Department of Computer Science & Engineering
Jaipur Engineering College and Research Center

Mukesh Agarwal
Department of Computer Science & Engineering
Jaipur Engineering College and Research Center

ABSTRACT
Every Educational organization's success rate depends highly on the success of the student. Many types of research are taking place in education field using machine learning techniques. Machine Learning has the ability to learn about the student and predict the performance of the student. In this research, analysis has been done to predict the possibility of the student getting backlog using various attributes related to the student and applying machine learning algorithms. Data of 648 students were collected containing 30 attributes. Preprocessing steps were adopted to convert the data in usable form. Once the data was ready then the Random Forest method was applied as learning algorithm. The accuracy of the random forest method was 94%. This type of analysis can help educational institutes take preventive measures for the improvement and monitoring of the student

General Terms
Machine learning, Classification Techniques, Data Mining, Artificial Intelligence

Keywords
Artificial Intelligence, Machine Learning, Supervised Learning, Classification, Exploratory Data Analysis (EDA), Prediction, Random Forest, Data Mining.

1. INTRODUCTION
Artificial Intelligence has become an important field for the researchers. A lot of research work is performed in this field. Industries and companies also use machine learning methods to trace its progress by`using all the different attributes available. Education institutes can also use machine learning algorithms to trace the progress of its students.

A student and the future will always be linked. No one can exist without the other. They are the building blocks of a nation. Education is the most powerful weapon of a student. Education system all over the world has changed rapidly since vast research in the field of Education Data Mining[EDM] and learning analytics. Educational Data Mining (EDM) is a field that makes full use of statistical, machine-learning, and data-mining (DM) algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational issues. EDM uses the statistical techniques to analyze the collected data from educational institutions to stumble on different patterns of student's behaviors and predict the performance of the students. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn. On one hand, the increase in both instrumental educational software as well as state databases of student's information has created large repositories of data reflecting how students learn. The main goal of this research field is to help the student improve their skills and to keep a check on things which hinders student from achieving success. Then one has to search for ways to improve it. This is done by using different DM techniques, machine learning algorithms, and statistical techniques. In this paper, Random Forest Technique has been used to identify that which student is likely to get backlogs in engineering degree so that preventive measures can be taken for the improvement of the performance of the student.

Predicting student's failure at educational institutes has become a difficult challenge due to the high number of factors that can affect the low performance of students and the imbalanced nature of these types of datasets. In this paper, a programming algorithm and different data mining approaches are proposed for solving these problems using real data of college students from Jaipur Engineering College and Research Centre, Jaipur(India). Firstly, the initial step was to select the best attributes in order to resolve the problem of high dimensionality.

The second step for analysis involves the imputation of the data collected using different methods, then the classification method which provides good result in different fields were tested to form a model which could judge which student is weak. Next step involved finding out the attributes that help in improving our model and gives good analysis using EDA(Exploratory Data Analysis) and it was updated when new attributes were introduced to the data.

2. LITERATURE REVIEW
Researchers these days are consistently attempting to analyze students' datasets using data mining and machine learning algorithms in order to understand how students learn and to ultimately increase the performance of students and the quality of learning. However, only a considerable amount of literature has been published on predicting the performance of the students based on different factors and attributes.

Cortez and Silva [1] in this paper carried out a study to predict the performance of school students based on school grades. They used Classification and Regression algorithms (Decision Trees, Random Forest, Neural Networks and Support Vector Machines). The result was that the past evaluation of the students was highly influenced by their performance. Preliminary work on analyzing the student datasets to predict their performance was undertaken by Aher and L.M.R.J. Han and Kamber [2] stated that data mining software that allows the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. Aher et.al [3] performed research in which they analyzed the examination performance of final year students using Weka mining tool. The algorithms

Association Rule, Classification (ZeroR), Prediction and Clustering (DBSCAN) were applied on student's examination data to check the growth of the students. Allan Tucker et. al,2018 [4] in this paper applied C4.5 and Naïve Bayes algorithms through which they revealed that Naïve Bayes performs better than C4.5 decision tree algorithm in predicting the students who have chances of failing the Module with an accuracy result of 88.48% for Naïve Bayes and 84.29% for C4.5 algorithm. Bharadwaj et. al, 2011 [5] conducted the study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application). They used Bayesian classification method on 17 attributes, it was discovered that the factors like students' grade in the senior secondary exam, living location, medium of teaching, mother's qualification, students other habits, family annual income, and student's family status were highly correlated with the student academic performance. In order to enhance the quality of education system Qasem A. Al Radaideh et. al,2006[1] used data mining function to evaluate student's academic data and stated that classification model can be used to enhance the courses. Al-Radaideh et. al. [6] in their paper stated that the analysis they performed can be used to give a deeper understanding of student's registering pattern in the course, and the faculty and managerial decision maker take all the necessary action to provide extra basic course skill classes and academic.

3. BASICS: MACHINE LEARNING

Machine Learning is the subset of Artificial Intelligence which uses different techniques to allow the computer to learn with the help of data, without being programmed. It has two parts that are Supervised machine learning algorithm and an unsupervised machine learning algorithm. The majority of machine learning uses supervised learning. The goal in this is to approximate the mapping function so well that when you have new input data that you can predict the output variables with maximum accuracy, which can be checked by making different matrices and check their value, for that data.

In Unsupervised learning, there is no supervisor. It is the training of the machine using information that is neither classified nor labeled and it allows the algorithm to act on that information without guidance. In this, the work of machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data. In this, no training will be given to the machine.

Unsupervised learning is grouped into two categories of algorithms:

Clustering: A clustering problem is where you want to discover the different groupings in the data, such as grouping customers by purchasing behavior.

Association: An association rule learning problem is where you want to discover rules that describe large portions of the data, such as people that buy A also tend to buy B.

Supervised machine learning algorithm as the name, indicates needs a supervisor who will tell or help to learn about what the conclusion will be. So supervised learning is a learning in which one has to teach or train the machine using data which is well labeled that means some data already have what one calls the correct solution. After this, the machine is given a new set of data so that supervised learning algorithm analyses the training data and produces a correct outcome from labeled data.

Supervised learning is then grouped into two categories of algorithms: **Regression** and **Classification** problems. Both problems have a goal to construct a clearly expressed model that can predict the value of the dependent attribute from other attributes. The difference between the two supervised learning problem is the fact that the dependent attribute is a numerical value for regression and categorical for classification.

Regression: A regression problem is a problem in which the predicted variable is a real value, such as "Price of the house (dollars)" or length of something (Km, m, cm)A regression problem is when the predicted variable is a real or continuous value, such as "salary" or "weight". Many different models can be used for regression, the simplest is the linear regression. It tries to fit data with the best hyperplane which goes through the points. For example when someone has to predict the price of a house so that the seller will have a maximum profit is an example of regression because the predicted values will be a real value. Regression model includes

- Linear Regression,
- Logistic Regression
- Polynomial Regression. etc.

Classification: A classification problem is a problem in which the output variable is a category, such as "Yes" or "No". A classification model draws some conclusion from observed values. If there is one or more inputs a classification model will try to predict the value of one or more outcomes. For example, predicting that a mail is a spam or not will be an example of classification because the value for the attribute will be in the form of a yes or a no. Thus Classification is used to create a model or create a categorical label based on the training set and the values of attributes available in the training set.

Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics to classify the given data. It is a supervised machine learning technique which works well with a large amount of data. This grabs all the information from previously available data and then apply the same to newly found observation. The information is grabbed from the training dataset and then it is tested on the test data to bring out the RMSE, MAE or the accuracy of the data. It is very difficult to judge that which classification technique should be used because the single method works well with a single dataset but not with others. All the different methods are tested and the one with more accuracy is considered the best to form the model. There are various algorithms that come under classification techniques. Some of them are :

- Logistic Regression
- Decision Tree
- Random Forest
- Naive Bayes
- k - Nearest Neighbours
- Stochastic Gradient Descent

In this paper, only the Random Forest algorithm has been used.

3.1 Random Forest

Random forests or random decision forests are the learning method for classification, regression, and other tasks, that

operate by constructing multiple numbers of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of each individual trees. The difference between the Random Forest algorithm and the decision tree algorithm is that in Random Forest, the process of finding the root node and splitting the feature nodes will run randomly. For applications in classification problems, the Random Forest algorithm will avoid the overfitting problem. For both regression and classification task, the same random forest algorithm can be used. The Random Forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering.

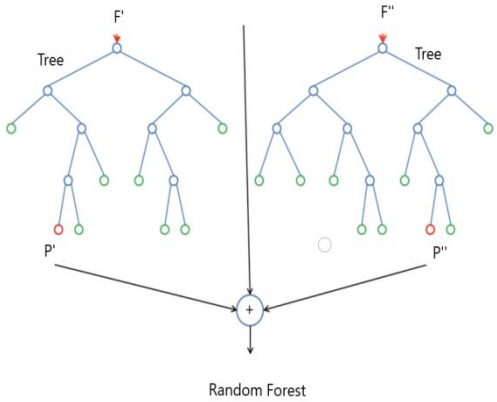

Fig 1: Random Forest adopted from Analytics Vidhya

4. EXPERIMENTAL SETUP

This research is performed to predict that a particular student will get backlogs in engineering or not. In this, a machine learning algorithm has been used to classify the weak students with the help of 26 attributes. The machine learning algorithm which has been used in this paper is the Random Forest method. The research started with the collection of the data. The data was collected from the students of Jaipur Engineering College and Research Centre (JECRC Foundation) from 2014, 2015, 2016, 2017 batch from Computer Science and Engineering branch. The data contained 648 rows and 30 attributes. The data collected initially contained many missing values. In the preprocessing phase of data, the missing values were filled using the median and mode method. The initial investigation has been performed by executing an algorithm by filling the missing values with mean and mode method and by filling the missing values with median and mode method and it was observed that the result generated after filling the missing values with median and mode were better. Example of the code for filling the value with the mode is as shown below.

Code for Mode
data.full[data.full$attribute1=='level2', "attribute1"] <- 'Level2'

Before filling the missing values with the median one needs to convert the character value into numeric value but if a person will convert character value directly into numeric then NAs will be introduced. So to solve this problem first the character value should be converted into factor and then factor into numeric so that NAs are not introduced.

Code for Conversion….**data.full$attribute2<- as.factor(data.full$attribute2)**
data.full$attribute2<-as.numeric(data.full$attribute2)

Code For Median
median<- median(data.full$attribute1, na.rm = TRUE)
data.full[is.na(data.full$attribute1),"attribute1"] <- median

Now the next step will be to choose attributes which will help in our prediction and will improve our accuracy. The Attribute selection was done through Exploratory data analysis(EDA) and the valid attributes were selected. Out of 30 attributes, 26 attributes were selected based on the number of levels, partial relation with backlog etc. Attributes like provisional degree(a student has taken his degree or not), TC(a person has taken a transfer certificate or not) were removed because if a person has taken his degree that means he has completed his engineering and has no backlog. The attributes which were used were Gender (the student is a male or a female), Rank (it is the rank which the student got in his competitive exam), Hostel (if the student lives in a hostel or not), Education (this states that the student has studied from which board) etc. Different attributes were plotted against the backlog to see the relation between the attribute and backlog. The Code used for creating the plot is given below.

Code For Graph
plot(train$attribute3, train$Backlog, main="Scatterplot", xlab = "A", ylab="Backlog ", pch=19)

There were two datasets: Training dataset and the test dataset The training data consisted of 70% of the total dataset and the test data consisted of 30% of the complete dataset. Firstly Machine Learning Algorithms were applied to training dataset for the learning of the system. The training was done to predict the value of backlog (ie. if a student got a backlog in any semester or not) and then the testing was done using the test dataset and the accuracy of the data is carried out.
Machine learning algorithm was applied to train the system to classify two classes ie. the students who are having the backlog or not having backlog using training data. Finally, test data was used to find the efficiency of the trained system. RMSE (Root Mean Square Error) is used to measure the standard deviation of the predicted value from the original value. There are different ways to measure the performance of the model. One way is through the Root Mean Square Error and the other one is by checking the error rate. Error Rate is the number of misclassification in the model. If the goal is to have the model with low Error Rate then using RMSE is not appropriate and vice versa.

5. RESULT AND CONCLUSION

Lastly, it was found that Random Forest has an accuracy of 94% which is shown in table 1. The out-of-bag(OOB) estimate of error rate which is equal to 5.39% and the confusion matrix is also shown in figure 2. Then the code for the graph of the model was written which is accurately shown in figure 3.

Table 1. The accuracy of the model

ALGORITHM	**ACCURACY**
Random Forest	94%

```
randomForest(formula = data.formula, data = data.train, ntry = 500
* nrow(data.test))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 5.39%
Confusion matrix:
    0  1 class.error
0 425 13  0.02968037
1  13 31  0.29545455
```

Fig 2: Results of Random Forest

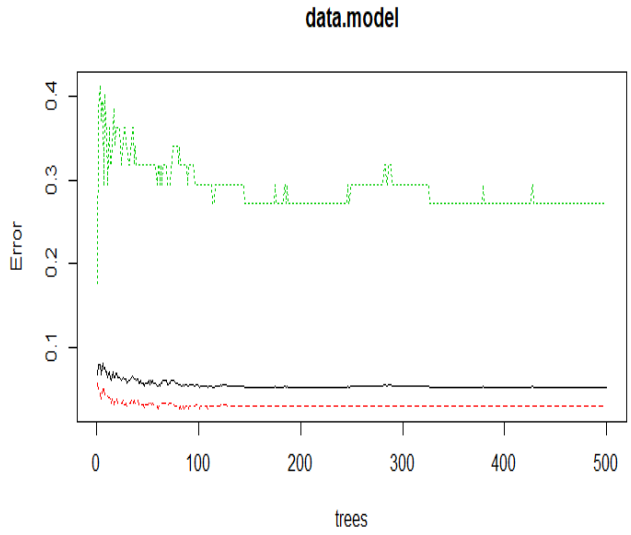

Fig 3: Graph describing error with respect to trees.

Then, at last, the ID from the test dataset was taken with the predicted values of the Backlog for the better view of the output.

```
## Code For Output Display and saving it
in csv file
```

Id<-test$attribute5
output.df<-as.data.frame(Id)
output.df$Backlog<-Backlog
write.csv(output.df , file = "Solution.csv", row.names = FALSE)

The results generated by the Random Forest Method are very promising. In this various other algorithms or hybrid algorithms can be performed to increase the accuracy of the prediction. Also, like the size of the data set is small it is required to check the accuracy of the algorithm with increasing and decreasing the size of the data set to see the effect. In this paper, only 26 attributes are considered. Other different attributes can also be considered. The data collected in this research was just for one college hence the data is collected from various colleges to see the change in accuracy. Further work in the direction provided above can result in a model which can be used as the standard model for prediction of backlog in the engineering students.

1.

6. ACKNOWLEDGMENT

Any research work requires the direct and indirect support of many people. We would like to thank them all. Also, heartiest thanks to the people who have provided us with details to carry forward with our research. Lastly special thanks to our friends and family for their moral support which helped us to carry forward with this research.

7. REFERENCES

[1] Cortez, P., Silva, A., (2008)" Using data mining to predict secondary school student performance." Presented at the 5th Annual Future Business Technology Conference, EUROSIS, pp. 5–12. J.

[2] Han and M. Kamber, "Data Mining: Concepts and Techniques,"

[3] Aher, S., L.M.R.J., L., (2011) "Data Mining in Educational System using WEKA." Presented at the International Conference on Emerging Technology Trends (ICETT), International Journal of Computer Applications® (IJCA), pp. 20–25.

[4] Mashael Al luhaybi, Allan Tucker and Leila Yousefi, 2018 "The Prediction of Student Failure Using Classification Methods: A Case Study"

[5] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 20116

[6] Al-Radaideh, Q., Al-Shawakfa, E., Al-Najjar, M., (2006) Mining Student Data Using Decision Trees

[7] C. Romero, S. Ventura, "Educational Data Mining: A Review of the State of the Art", IEEE Transactions on Systems Man and Cybernetics-Part c: Applications and Reviews, vol. 40, no. 6, pp. 601-618, 2010.]

[8] Sagardeep Roy, Anchal Garg, "Predicting academic performance of student using classification techniques"

[9] C.M. Vera, A. Cano, C. Romero, S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data."

[10] Neelam Chaplot, Praveen Dhyani, O. P. Rishi, "Astrological Prediction for profession doctor using classification techniques of artificial intelligence"