TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988

Student Performance Analysis Using Machine Learning

Sunil Bhutada, Professor, IT Department Sreenidhi Institute of Science and Technology Yamnampet, Hyderabad. (sunilb@sreenidhi.edu.in)
Sreelekha Komma, B.Tech IV year, IT Department Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabd (sreelekhakomma24@gmail.com)
Ruchika Bhutada, B.Tech III year, CSE Department Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabd (ruchikabhutada24@gmail.com)

Article Info
Volume 83
Page Number: 6982 - 6988
Publication Issue:
May-June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *18 May 2020*

Abstract:
For every student, academics play an important role in their life. So our project is mainly based on "Student performance analysis", which is a process of predicting the student performance based on the previous data about the students and their academics, so as to help them correct themselves. This lets both teachers and parents know about the student performance and by using certain indicators, we perform logistic regression using python which helps to easily predict the student performance and analyse the student data. Logistic regression model plays a major role in the prediction of correct data by using particular indicators or attributes from entire data to be trained to get the original data so that can be predicted easily. Later random forest and support vector machine also used to find accuracy and at last, we had compared all the algorithm to tell which performs well.

Keywords: *Prediction, Logistic Regression, Random Forest, Support Vector Machine, Python*

1. INTRODUCTION:

Students are an important asset in the educational process and based on their performance, their future will depend. As student performance analysis will analyse their performance based on their activities, we collect the data based on their daily participation and other works which helps us to analyse their performance and help make changes or suggest ways to improve themselves.

In this context, the information of each and every student should be collected and stored in any format as we can analyse easily and help them to improve easily. Logistic regression, random forest, support vector machine play an important role to analyse the data. As we collect a lot of data about the students that data will be unstructured, so a lot of pre-processing has to be done to make sure that the data can be worked with. With the structured data, we train different models and evaluate their performance using true classes and predicted classes.

This paper is composed of different sections. It starts with a literature survey as section 2 in the context of present work. Section 3 explains about problem statement. Section 4 contains the process flow. Section 5 about methodology. Section 6 about results that we had obtained by performing. Section 7 is about the conclusion and future scope.

2. Literature survey

The basic methodology proposed for prediction of the student performance analysis by data mining techniques, to analyse student performance [1] by the data mining techniques should use many algorithms like ms J48, NB tree, simple cart and many other tools are considered like WEKA [2]. Here are many significant factors that are used for

May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988

constructing the decision tree. Student performance analysis that shows a graph after performing the algorithm or data mining techniques [4].

Data management and data storage play an important role in analysing student performance. Data Storage can be started when the first student joins school or college by giving them an enrolment number based on that number we can start storing information related academics and other information about the student [2]. Even we can map based on the id of the student so we should consider it as the primary option in this type of cases and we can predict the overall data about the student as per class or semester wise [3]. As per the technology, the students started learning online and even they started writing exams online so it became easy to store student data before they start they should always enter their id so there marks or any other information will be automatically stored based on their id [4]. We can store the data in the many forms like excel sheets, CSV file, and many other formats.

We can also perform these predictions using the Latent semantic analysis it always shows the mathematical representation [5]. So the latent semantic is always based on the mathematical formulas whereas the data mining techniques help us to visualize the data by the help of graphs, histograms etc [1].

We can also perform using decision tree and Bayesian algorithm even it later compares both the accuracy to show which will perform better among them [6]. Whereas each method contains certain association rules to perform or to calculate accuracy and compare among each other [2]. The main observation is about which performances better among them so we can use for later as to get result good and accurate and can be used further.

3. Problem statement:

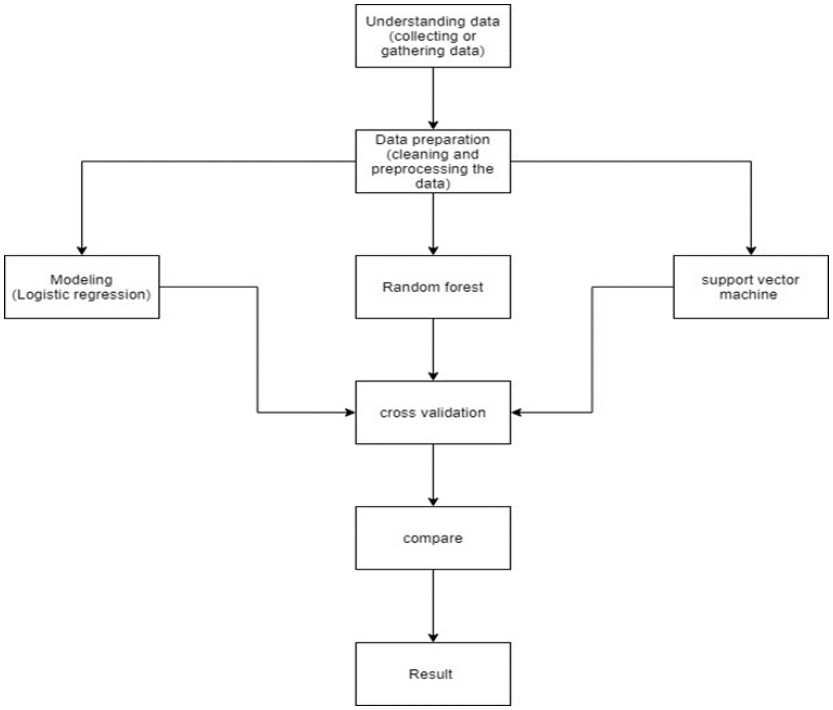
The general student performance analysis system considers only the student pass or fail. But focusing on just pass or fail does not represent the performance of the student precisely. We can also consider participating in other activities like course-based events or events in general. Considering other competitive exams, announcements etc. would help in determining the performance of the students precisely. Considering all the aspects and activities of a student will help in a more reliable and complete analysis of his/her performance.

It wouldn't be enough to judge the student based on the pass or fail, since predicting the student performance needs a lot of attributes to be considered. The main aim of the project is to predict student performance based on all activities belongs to their academics and other activities. Through this proposed method of ours, student performance can be easily measured and can also help teachers in taking steps to improve the performance of their students. Students can also get to understand themselves better.

4. Process Flow:

The process flow shows how the whole process of the project is done and the result is viewed. The following flow chart shows the process:

May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988



5. Methodology:

5.1 Understanding Data:

The data understanding is collecting or gathering of the data about the students. We took a dataset in the form of an excel sheet. The dataset contains information about the students like their id, gender, topics, announcements etc. So based on the student data we had we can predict the performance.

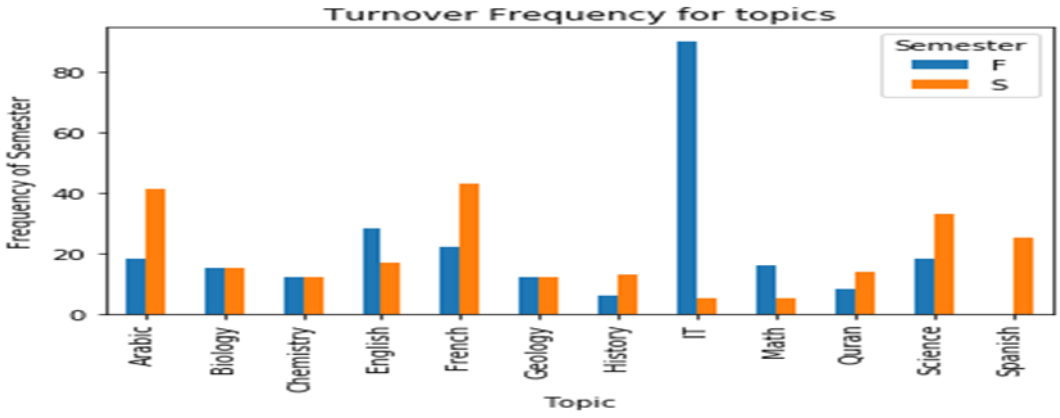
It causes to take place all activities related to building dataset. If there is no value for a particular attribute we should leave empty but we should not fill it with the wrong value.

5.2 Data Preparation:

Preprocessing plays an important role in the process. It deals with transforming data into an acceptable format. Cleaning data involves various steps including removing null values, combining similar values or attributes and using unique data. It is an important role and it is a very iterative and complicated stage.

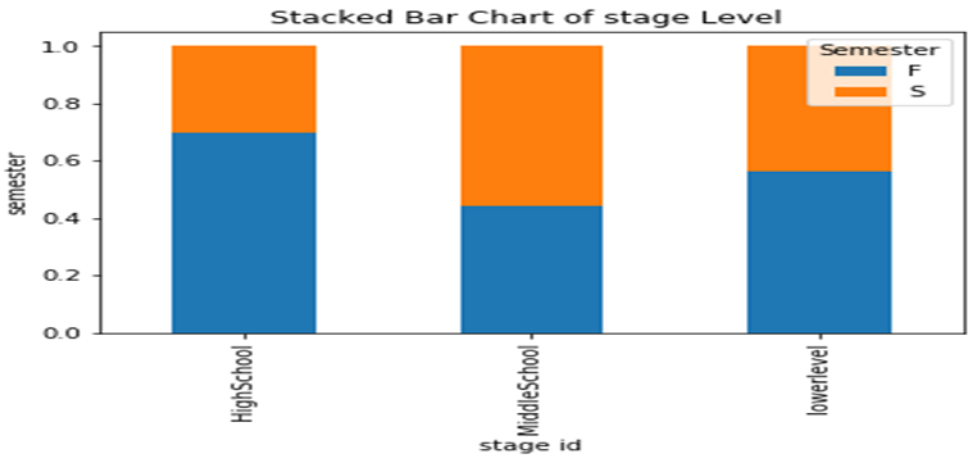
5.3 Data Visualization:

We visualize data to get a clear picture of the data as it shows the success rate and failure rate and failure of the topics in their semester.

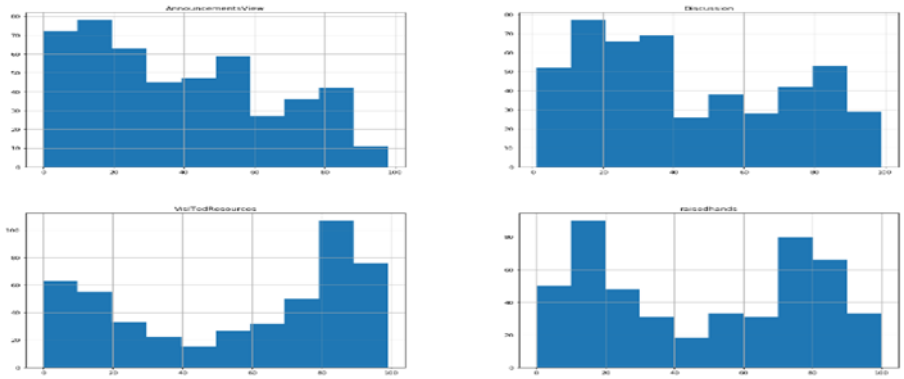


May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988

In each stage levels of data can be visualized based on their semester and their stage levels.



The other activity data like announcements, resources that are used, discussions.



The above chart shows for other resources how they performed.

5.4 Modeling:

5.4.1 Logistic Regression:

Here we used Logistic regression and convert categorical variables into dummy variables and it will perform well when data is linearly separable so it is easy to implement. We divided both testing and training data and used train dataset to build the model. The Recursive Feature Elimination(RFS) works by removing variables and building the model based on these variables. It uses model accuracy to identify which variable contributes to the most predicted target attribute. Here we will use Scikit-learn library which helps us to analyse easily

```
Logistic regression accuracy: 0.688
```

5.4.2 Random Forest:

Random forest is performed on the training dataset. Random forest runs efficiently on the larger datasets. It corrects the habit of overfitting the data as it is an ensemble classifier and easy to understand. It uses the accuracy that identifies the target attribute. It gives the appraisals which attribute is significant in

May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988

the classification. It always uses the labelled data to classify. It always produces a highly accurate classifier. It can handle lots of attributes without deletion.

Random Forest Accuracy: 0.611

5.4.3 Support Vector Machine(SVM):

Support Vector Machine is also performed on the training dataset. SVM is a supervised learning algorithm which contains labels. SVM works well when it has a clear margin of separation. It supports both regression and classification.it is most productive in greater dimensional space. In this case, the size of the dimensions is higher than the samples size. Here we calculated the accuracy for the target variable with less computational power. It is more robust than others. It is a method based learning model.

Support vector machine accuracy: 0.500

5.5 Cross-validation:

Cross-Validation is a statistical estimation for the machine learning algorithm. Here it is used to compare the models that we have performed by using k-fold cross-validation. It attempts to avoid the overfitting for the observations. Here we performed 10-folds cross-validation to find the average accuracy of models. In this case, cross-validation is used to find the average accuracy of all the three models that are logistic regression, random forest and support vector machine.

10-fold cross validation average accuracy: 0.604

6. Results:

The results show the precision and recall after the cross-validation as it calculates precision, recall, f1 score, support for fail rate, success rate and accuracy as shown below:

	precision	recall	f1-score	support
F	0.57	0.72	0.64	69
S	0.67	0.51	0.58	75
accuracy			0.61	144
macro avg	0.62	0.62	0.61	144
weighted avg	0.62	0.61	0.61	144

The precision and recall are defined in the terms of the confusion matrix the terms like true positive(TP), true negative(TN), false positive(FP) and false negative values are used.

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

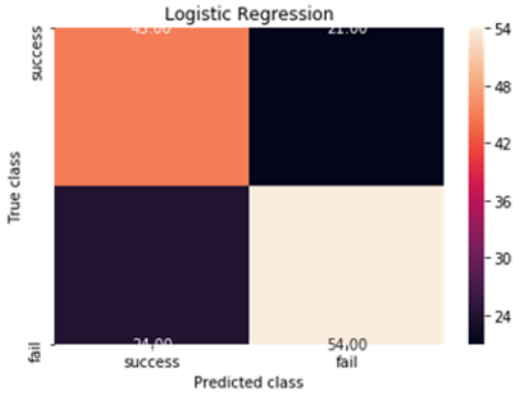
F1 score shows the equilibrium between precision and recall. Here we should calculate the f1 score for each and every model by their precision and recall values.

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

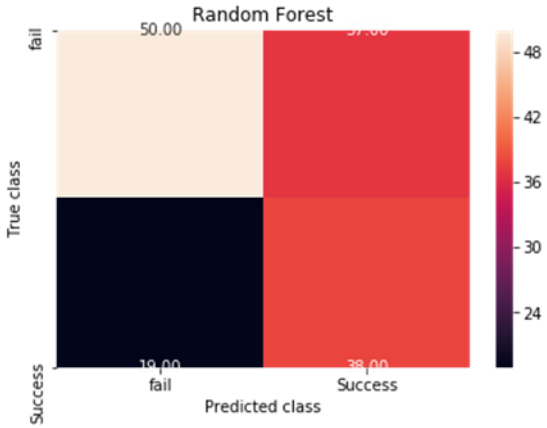
A confusion matrix is always used to trace the performance of the model for test data on predicted class and true class. It shows predicted result on a classification problem. Here the success and fail predictions summarized by count values broke down by each class.

Now we will build the confusion matrix for the logistic regression by the true class and actual class to show the success rate and failure rate

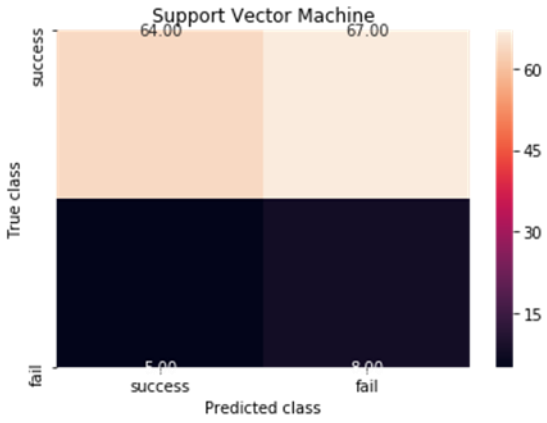
May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988



Then the confusion matrix for the random forest by the true class and actual class to show the success rate and failure rate of the student by their accuracy.



Later we build the confusion matrix for the support vector machine by the true class and actual class to show the success rate and failure rate of the student by their accuracy.



The above shows the result that is a comparison between all the models we got the good accuracy for the logistic regression so always consider the logistic regression to predict the performance the student as we get a good result.

7. Conclusion and Future Scope:

Predicting student performance is the most useful way to help both teachers and students improve their studies and learning process. It concentrates mainly on the development of student academics and students performance can be predicted based on the data and provide accurate results of their performance. The project is on the analysing and prediction of student performance analysis. A machine learning technique and a classification algorithm are applied to the project to make sure that the prediction shows the accurate value of the student performance will be calculated. By using this model we can consider many factors at a time.

REFERENCES

1. Mrinal Pandey, Vivek Kumar Sharma, “A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction,” International Journal of Computer Applications (0975 – 8887), Volume 61– No.13, January 2013
2. Abdelmajid Chaffai, Abdeljalil El Abdouli, Houda Anoun, Larbi Hassouni, Khalid Rifi, “Student Performance Analysis using Large-Scale Machine Learning” (IJCSIS) International Journal of Computer Science and Information Security, Vol. 13, No. 12, December 2015
3. Pauziah Mohd Arsad, Norlida Buniyami, Jamalul-Lail Ab Manan, Noraliza Hamzah, “ Proposed Academic Students’ Performance Prediction Model: A Malaysian Case Study” 2011 3rd International Congress on Engineering Education (ICEED), 7-8 December 2011, Malaysia
4. Zacharoula Papamitsiou, Anastasios A, “The effect of personality traits on students’ performance during Computer-Based Testing: a study of the Big Five Inventory with temporal learning analytics,” 14th IEEE International Conference on Advanced Learning Technologies (ICALT2014), Athens, Greece
5. Shaymaa E. Sorour, Tsunenori Mine, Kazumasa Goda, Sachio Hirokaw, “A Predictive Model to

May-June 2020
ISSN: 0193-4120 Page No. 6982 - 6988

Evaluate Student Performance," Journal of Information Processing Vol.23 No.2 192–201 (Mar. 2015)

6. Nguyen Thai Nghe, Paul Janecek and Peter Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance," 37th ASEE/IEEE Frontiers in Education Conference, October 10 – 13, 2007, Session T2G
7. https://www.kaggle.com/aljarah/xAPI-Edu-Data