**DETERMINATION OF STUDENT RETENTION RATES AND FUTURE CAREERS**

By piquing into the text data for students on social media channels and communication networks, it has been established that students in Asian countries face issues with applying learned concepts and retaining them in situations where required.

The project has previously produced keyword results for the same and has even looked for a correlation between student performances in the classroom and their correlation with text data in the form of essays and reviews left by teachers.

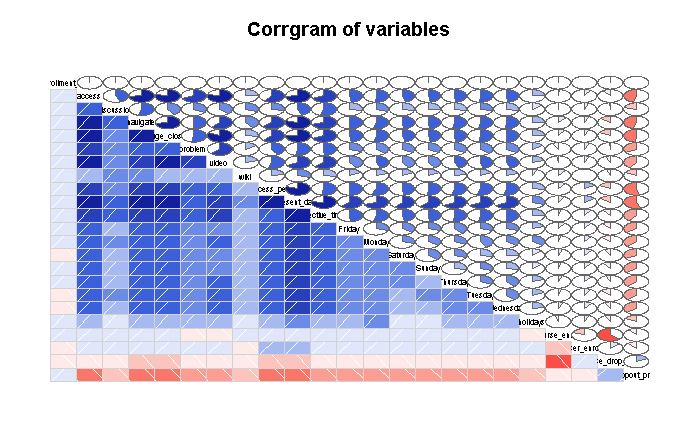
This document concerns with the results of techniques applied in understanding another dynamic problem of student learning systems( particularly MOOCs) in India to see and determine factors that affect student graduation in key components as well as find red flags for students who may be at risk.

To compare these results, student performances from schools in the west have been benchmarked and find causes for retention in them as well.

**Chapter 1-Studying At Risk Students**

Based on the first data set used, it was surmised that certain aspects of learning such as absent rates, proximity of classrooms and quarterly quiz and test marks will determine the possibility of a student dropping out. The dataset shows the required data for over a thousand Indian students. The results of the study are available in the **simulated\_data.csv** dataset.

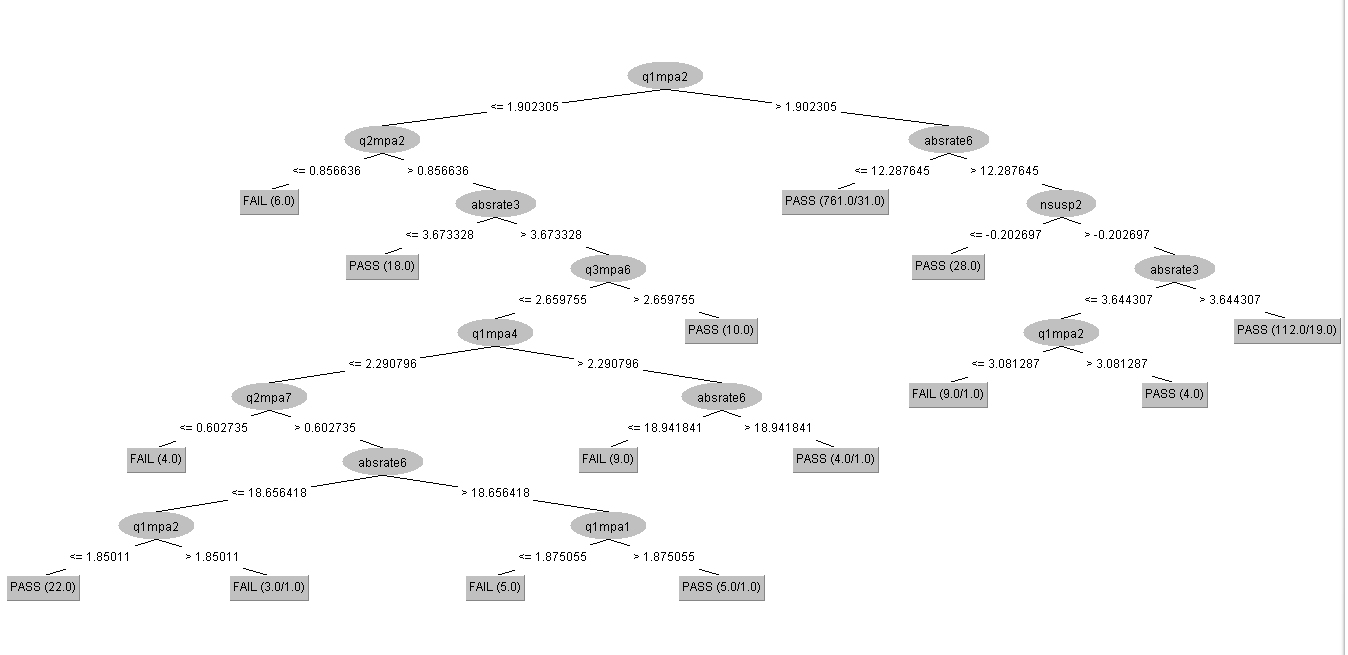
The student performances was measured at regular intervals along with factors such as time present, number of suspensions and test performances along with the dropout possibility( 1 representing that the student has dropped out and 0 representing that the student has not dropped out and continued with the course).



For illustrative purposes, a correlation database was created that showed the relation between the parameters and the dropout statistics which showed that absent rates and suspensions did have a negative effect on students dropping out but not as much compared to mobility.

The correlation parameters can be viewed in the **Correlation of student variables.xlsx** table.

The project then moved on to creating a decision or treeing algorithm that showed the best accuracies for identifying the instances correctly. The methods of SMO and Logistic regression the best accuracies of 85 percent. The figure below shows a representative result of one of the models using a J48 algorithm. All cases were tested for various cross validations and training to testing ratios. A 10 fold cross validation proved to produce the best accuracies for a majority of the cases. The results for the models can be viewed in the folder **Chapter 1-Results.**



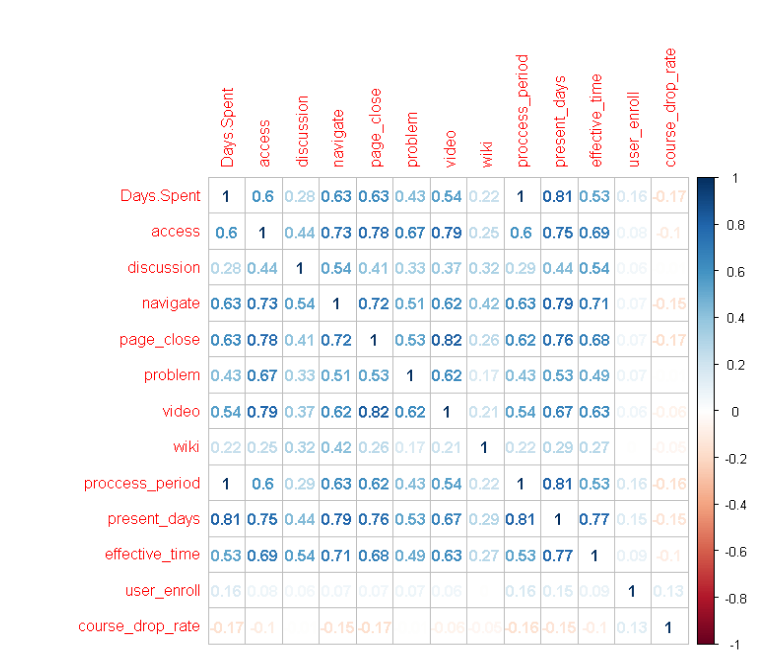
**Chapter 2-Tracking MOOCs**

To further track the effects of the learning environment on students, a dataset containing information for students in an online learning platform was used titled **MOOC\_Visual.csv.** The dataset shows parameters such as enrollment periods, time spent, resources used on the online modules and the chances of the students dropping out of the course. The objective here was to find a way to see any warning signs for students in other modules which can allow course creators to help mitigate such students and prevent such chances of total course cessation and retention.

In this case, course dropout rates were measured from 0 to 1 with 1 representing a complete possibility of a dropout and numbers below it and 0- at risk students.

Correlation studies as represented from the table below shows that the days spent by the student, time taken for modules(process period) and number of page closes, had the best correlation against avoiding chances of drop outs. Time spent on the discussion forums, asking problems and effective time(time spent between study videos) had the least negative correlation.

Prediction models were created for the same by using the Weka and RStudio platforms to produce trees with massive leaves and branches owing to the large number of instances. A sample treeing classification model can be seen in the **TREE(86% ACCURACY).txt** file. Additional models were created, among which the xgboost and Hoeffding Tree models produced the best accuracies. The models can be seen in the **Chapter 2-Results** folder.



**Chapter 3- American Student Retention Rates**

To compare the effects of certain pedagogical elements in the retention of students, the students of American Institutes were studied as well as the relation to scores in tests such as the SAT, ACT, GRE etc. The results and source code were compiled in Python.

The results for this component can be viewed in the folder **Chapter 3-Results.**

**Chapter 4- Student Performance- Career Predictor**

To further allocate a proper means of understanding where students can be best placed based on their contextual information, some analysis was performed on the data for college students and their personal details.

This was done with the intention of pressing issues faced by universities and companies in understanding where interns and new employees should be placed depending on features beyond what’s seen on the CV or resume.

The results for the same can be seen on the folder **Chapter 4-Results**.