A REPORT

ON

Data Insights on K-12 Education System

BY

Name	**ID Number**
Amartya Ayushi	2019B2A41467H
B Shreyas Bhat	2019B1A80969G
Vaibhav Rana	2019A1PS1113G
Jathin Narayan	2019A7PS1001G
Himanshu Jain	2019A3PS0423G

Project Report

BITS-PS-1-004

AT

LearningMate, Mumbai

A Practice School–I Station of

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

June 2021

ACKNOWLEDGMENT

We would like to thank BITS Practice School Division for this excellent opportunity to pursue our Practice School-I program at LearningMate, Mumbai.

We would like to express my heartfelt gratitude towards Meenakshi Nagdev, Manager - Talent Acquisition,LearningMate, for giving us an opportunity to be part of an elite organization and contribute to the development of Frost, Data Insights on K-12 Education.

We would also like to thank Roshini Tiwari, Talent Acquisition Lead, for facilitating the process and for assigning the project based on my interest. We would also extend our gratitude to our project guide, Vinay Agrawal, Associate Solution Architect, who is not only overseeing the project personally but also guiding us whenever we are in need of any help.

We are also grateful to our PS Faculty, Ashutosh Bhatia Sir, for his dedication towards students currently in the PS and for his quick response to all our queries.

Furthermore, we would like to acknowledge with much appreciation the crucial role of Ajay Kumar, Usha Rani, Siddhartha, Prashanth, and Anand who have helped us immensely during the course of the project. Finally, we also thank all the staff from LearningMate, who have indirectly contributed to the success of this project.

Without the help and contribution of the above-mentioned people, this project would not have been a success.

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI
Practice School Division

Station: LearningMate
Centre: Mumbai
Duration: 53 days
Date of Start: 31st May 2021 **Date of Submission**: 22nd July 2021

Title of the Project: Data Insights on K-12 Education

Name and Designation of the Expert: Vinay Agrawal, Associate Solution Architect, LearningMate, Mumbai
Name of the PS Faculty: Ashutosh Bhatia

Key Words: Engagement, At-risk, Supervised learning, Data Analysis, Dropout rates

Project Areas: Measuring Student Engagement and Predicting At-risk Students

.

Signature of Faculty Date:

The K-12 education system in the United States

In the United States, education is compulsory for students aged from 5-8 to 16-18 years by state law.

Students can go to Public schools where they do not have to pay any tuition fees as these schools are funded by the federal government and sometimes even the state government. The state entirely decides the curriculum, and thus it varies from state to state.

There are also private schools, but these schools can be expensive, costing up to 12 thousand dollars every year; these expenses grow exponentially every year as the child climbs the education ladder.

This is the reason why in 2012, 87% of students attended public schools, and only the remaining 13% distributed between private schools and homeschooled.

To support this extensive schooling system, the federal government allocates 6-7% of the US GDP to the education sector. In 2016-17, the budget allocated to the education sector was around 1.3 trillion dollars, approximately 7.2% of the GDP.

In our project, the critical component of the K-12 system that we are focusing on is the standardized test conducted annually by the respective State governments. These exams are compulsory for students undertaking the public school system, and this test makes sure that the student is getting the desired minimum level of education.

There's been a bit of inflation in the education system as students' average grades are rising, but the average performance in SAT exams is dropping every year. This led to significant concerns among schools providing K-12 education, resulting in an increased number of dropout students over the years.

The dropping out and this dropping performance are directly related to student's engagement in the classroom, especially now that the classes are being conducted online. Our project is about predicting these students who are not properly engaging in the school and are potentially at risk of dropping out to make special arrangements for the students in need and help them get through this.

The frost insights in k-12 education are developed by the Learning Mate, serving as a one-stop tool for these schools.

Student disengagement

With the exponential evolution of science and technology, educational tools have made dramatic changes in recent decades. Virtual Learning Environments (VLEs) like Massive Open Online Courses **(MOOCs**), which provide lecture videos, online assessments, discussion forums, and even live video discussions via the Internet, have become commonplace especially in the period of the COVID-19 outbreak. Two of the benefits it brings account for the increasing adoption of online learning. Firstly, **VLEs** provide convenience for participants to enroll in courses by breaking time and distance limitations. Moreover, online learning platforms based on the Internet are able to record a type of data, including data from a user's VLEs and other learning systems, which is called **trace data**

Hence, identifying final students' outcomes in a timely manner ensures no delay in helping online platforms to make an instant intervention, which can assist underachievers to improve their performance.

Data Mining (DM) has emerged and gained an important role in data analysis, which is properly defined as the extraction of data from a dataset, and discovering useful information from it is not always straightforward or simple to promise absolute privacy, confidentiality, and anonymity while using open VLE. Hence, the platforms are usually reluctant to publish their data. Actually, the datasets with **anonymous processing** and high privacy level are likely to be adopted for studies.

EDM can also be used to identify aspects related to participants' dropout intention and classify students who tend to drop out based on their historic data Many studies have applied **machine-learning approaches and deep-learning approaches** for predicting the students' performance [29,30] and optimizing the learning settings when more students' behavior data are available due to the development of technology-enhanced learning environments like MOOCs and

Learning Management Systems (LMSs).

Predicting students' outcomes during the course is crucial in MOOCs and LMSs because it can help teachers recognize at-risk students and assist them in passing the course. Several works have been done using the difficulty of the question, how long the student took to respond, and whether the response was correct as input to identify at-risk students.

IEP stands for **Individualized Education Program**. An **IEP** lays out the special education instruction, supports, and services a student needs to thrive in school

A variety of previous research on predicting students' performance used traditional **machine learning approaches** to fit demographic information, interaction logs, or both.

Logistic Regression (LR) was typically employed in models predicting students at risk of failure and showed promising predictive results. **the interaction of students with the VLE** and the interaction was logged in the **number of clicks daily** for each course.

The type of interaction was categorized into 20 classes, meaning different click actions, such as visiting the recommended URL and resource, *completing quizzes, and filling in questionnaires.*

It is clear that students *who get higher scores in assessments and interact with the learning environment frequently* are more expected to pass the course, while those who fail the course tend to have lower scores in assessments and fewer clicks to the VLE. As a result, assessment performance history and mutual history can be used to predict whether a **student is at risk** during any online courses.

Meaningful data that can be extracted from LMS

Completion Rates

Whether online learners are actually completing the eLearning course?

How long does it take them to finish up each task or module?

Online Learner Performance and Progress

insight into learning behaviors, experience, and proficiency

eLearning Assessment Scores

Provides measurable data

A high percentage of passing grades indicates that your eLearning course is on-target. The opposite is also true. eLearning assessment data also gives you the power to identify online learners' strengths and weaknesses, which you can then use to create personal learning paths. For example, invite your online learners to take a **pre-assessment** before the eLearning course. Use your findings to choose eLearning activities and modules that suit their needs.

Online Learner Surveys

Online learners share their honest opinions and recommendations. You use this data to create a plan of action and modify your eLearning design. Surveys also help you determine whether your eLearning course is relevant and relatable for your audience.

Every piece of Big Data that you collect helps you personalize your eLearning approach. Online learner satisfaction ratings, completion times, and assessment scores reveal strengths and areas for improvement. As such, you have the opportunity to create eLearning content that offers the most value. For example, Big Data may reveal that online learners lack the necessary motivation. Therefore, you can include **gamification** elements to reward them for their efforts and motivate them to succeed

Overall course completion rates and online learner performance provide you with a complete picture, but they aren't necessarily useful for individual evaluation. For that, you'd need to look at user progress reports to identify knowledge and **performance gaps**.

Existing tools and platforms doing EDM

Caliper Analytics enables institutions to collect learning data from digital resources to better understand and visualize learning activity and product usage data, and present this information to students, instructors, and advisors in meaningful ways to help inform:

· Student recruitment and retention plans

· Program, curriculum, and course design

· Student intervention measure

Metric Profiles provide a common language for describing student activity across multiple learning environments. By establishing a set of common labels for learning activity data, the metric profiles greatly simplify the exchange of this data across multiple platforms. Many different products can be created using the same labels established by the standard. The IMS learning Sensor API™ is designed to define basic learning events and to standardize and simplify the gathering of learning metrics.

Measuring Student Engagement

We define **student engagement** as **a meta-construct that includes behavioural, emotional, and cognitive engagement**

Behavioural engagement draws on the idea of participation and includes involvement in academic, social, or extracurricular activities and is considered crucial for achieving positive academic outcomes and preventing dropping out. Following the rules, adhering to classroom norms, and the absence of disruptive behaviour such as skipping school or getting into trouble

Emotional engagement focuses on the extent of positive (and negative) reactions to teachers, classmates, academics, or school.

identification with the school, which includes belonging, or a feeling of being important to the school, and valuing, or an appreciation of success in school-related outcomes'

Cognitive engagement is defined as a student's level of investment in learning. It includes being thoughtful, strategic, and willing to exert the necessary effort for comprehension of complex ideas or mastery of difficult skills

Methods for Studying Engagement

Student Self-report

Students are provided items reflecting various aspects of engagement and select the response that best describes them.

It is critical to collect data on students' subjective perceptions, as opposed to just collecting objective data on behavioural indicators such as attendance or homework completion rates, which are already commonly collected by schools.

Useful for assessing emotional and cognitive engagement which are not directly observable and need to be inferred from behaviours.

Experience Sampling

In this methodology, individuals carry electronic pagers or alarm watches for a set time period. In response to ESM signals, students fill out a self-report questionnaire with a series of questions about their location, activities, and cognitive and affective responses allow researchers to collect detailed data on engagement at the moment rather than retrospectively (as with student self-report), which reduces problems with recall failure and the desire to answer in socially desirable ways

Teacher Ratings of Students

offer an alternative perspective on student engagement from that reported by the students themselves. Some teacher rating scales include items assessing both behavioural and emotional engagement.

It is useful for studies with younger children who have more difficulty completing self-report instruments due to the reading demands and limited literacy skills. Some studies have included both teacher ratings and students' self-reports of engagement in order to examine the correspondence between the two measurement techniques

Interviews

Participants are asked to tell their stories in more open-ended and unstructured ways

They can provide insight into the reasons for variability in levels of engagement to help understand why some students do engage while others begin to withdraw from school. Interviews can provide a detailed descriptive account of how students construct meaning about their school experiences, which contextual factors are most salient, and how these experiences relate to engagement

Observations

To assess individual students’ on and off task behaviour as an indicator of academic engagement

Academic engagement refers to a composite of academic behaviours such as reading aloud, writing, answering questions, participating in classroom tasks, and talking about academics

Used by school psychologists to screen individual children in both typical and special needs populations, especially those at risk for disengagement and academic failure

Narrative and descriptive techniques

The quality of instructional discourse in the classroom is an indicator of substantive engagement.

The frequency of high-level evaluation questions, authentic questions, and uptake (i.e., evidence that teachers incorporate students’ answers into subsequent questions) was observed as indicative of substantive engagement.

Noted behaviours such as relating the task to prior knowledge, requesting clarification, and using analogies as measures of cognitive engagement

Engagement Algorithm

After carefully inspecting the available dataset, we decided to finalize on some parameters to be considered to measure an engagement index. These parameters are measurable and meaningful metrics that we can extract from the dataset, either directly or through some manipulation.

Parameters decided:

- **Frequency_interaction** : Number of clicks on the Virtual learning environment.
- **Performance_index** : Index providing performance of students.
- **studentAssessment['score']** : Scores on final assessment.
- **Student_info['attendance']** : Attendance.
- **Registration Dates** : Early registration/Late registration

An algorithm has been proposed to measure student engagement by assigning point values or weightings to various assessment measures. It is desired to determine whether or not this algorithm is an effective measure of engagement. The algorithm combines faculty perception, student participation, and student self perception. The formula for calculating the engagement index is as follows:

Engagement Index

(K1*frequency_interaction) + (K2*_Performance_Index_) + (K3*student_assessment) + (K4*(Date_registration - Date_unregistration)) + (K5*attendance)

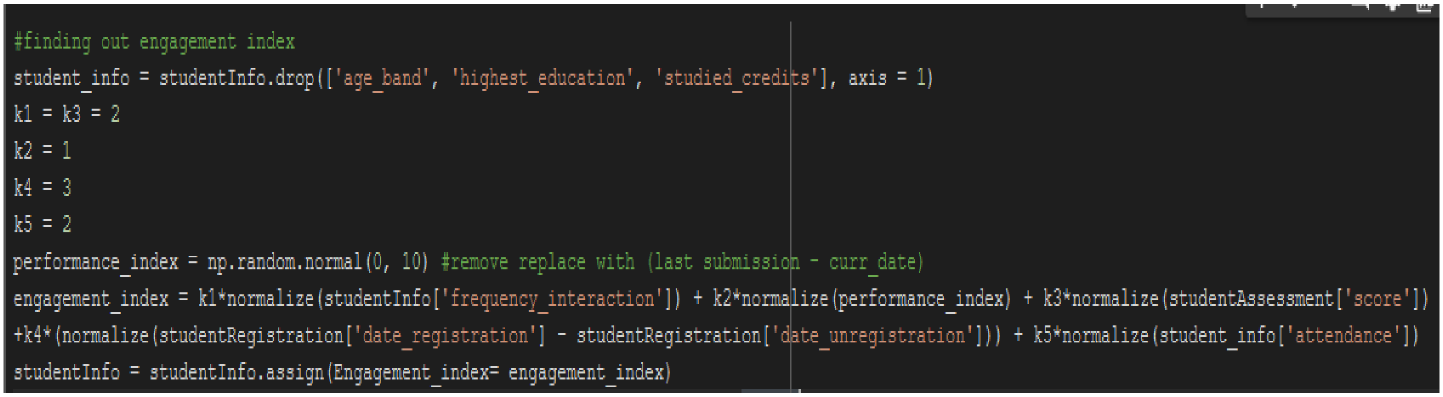
Where K1 , K2 , K3, K4 and K5 are weighting constants used to give different weights to individual parameters depending upon their importance .

Data Normalization

Normalization refers to rescaling real valued numeric attributes into the range 0 and 1.

It is useful to scale the input attributes for a model that relies on the magnitude of values, such as distance measures used in k-nearest neighbors and in the preparation of coefficients in regression.

We normalized our parameters to simplify and make our index more understandable.



```
#finding out engagement index
student_info = studentInfo.drop(['age_band', 'highest_education', 'studied_credits'], axis = 1)
k1 = k3 = 2
k2 = 1
k4 = 3
k5 = 2
performance_index = np.random.normal(0, 10) #remove replace with (last submission - curr_date)
engagement_index = k1*normalize(studentInfo['frequency_interaction']) + k2*normalize(performance_index) + k3*normalize(studentAssessment['score'])
+k4*(normalize(studentRegistration['date_registration'] - studentRegistration['date_unregistration'])) + k5*normalize(student_info['attendance'])
studentInfo = studentInfo.assign(Engagement_index= engagement_index)
```

Missing Data

In real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction.

Deleting Column not needed for analysis

```
studentVle.drop(['date'], axis =1)
```

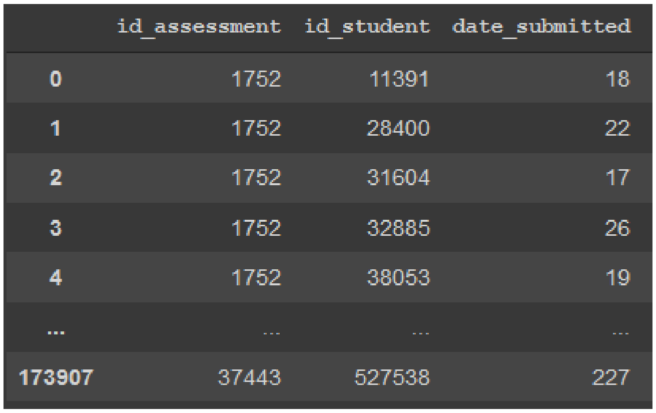
Replacing missing data

```
studentRegistration['date_unregistration'].replace('NaN', '0',
inplace=True)
```

Generating Parameters by Data Manipulation

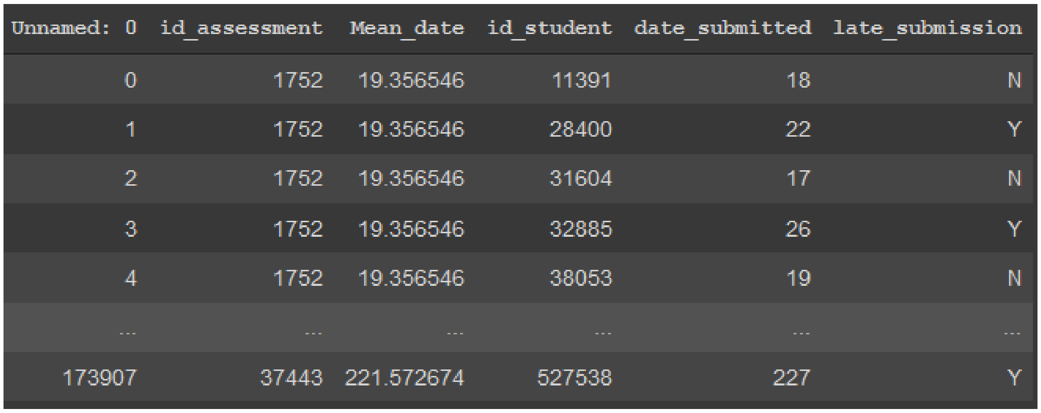
In a real world dataset, more often than not, we find that we can’t directly obtain all the necessary information and parameters.

By using some already existing functions in the pandas library of python we were able to perform some data manipulation and extract relevant parameters.



	id_assessment	id_student	date_submitted
0	1752	11391	18
1	1752	28400	22
2	1752	31604	17
3	1752	32885	26
4	1752	38053	19
...	...	...	...
173907	37443	527538	227

Elaborating one example of manipulation. Here we had the data of date of submission of an assignment and its assessment id but we had to judge whether the submission was made late or not so we can group the assignments by their assignment id and apply the .mean() function to them, to find average time taken to submit an assignment and then compare individual submission dates in a new column.



Unnamed: 0	id_assessment	Mean_date	id_student	date_submitted	late_submission
0	1752	19.356546	11391	18	N
1	1752	19.356546	28400	22	Y
2	1752	19.356546	31604	17	N
3	1752	19.356546	32885	26	Y
4	1752	19.356546	38053	19	N
...	...	...	...	...	...
173907	37443	221.572674	527538	227	Y

Predicting risk of students dropping out

Students in this category are determined to be at risk of dropping out of their courses. It is crucial to have in place a dropout warning framework which would preemptively identify K-12 students who fall into this category. In spite of the advantages of this new learning opportunity, a large group of online K-12 students fail to finish course programs with little supervision either from their parents or teachers. Students drop out of the class may be due to many reasons such as lack of interests or confidence, mismatches between course contents and students' learning paths or even no immediate grade improvements from their parents' perspectives. Therefore, it is crucial to build an early dropout warning system to identify such at-risk online K-12 students and provide timely interventions.

Parameters

After comparing the existing factors on a correlation heatmap, the following parameters were deemed to be the most efficient and linearly independent.

- **Date** - date of registration
- **Sum_click** - number of clicks on the VLE
- **Engagement_index** - index calculated from the previous section
- **Attendance -** number of days attended
- **Frequency_interaction** - number of times students interact with VLE

Challenges

It is important to study approaches to identify at-risk K-12 online students and build an effective yet practical warning system. However, this task is rather challenging due to the following characteristics:

- Multiple modalities - Difficult to determine uniform metrics due to subjective factors like quality of teachers, emotional attachment, etc.
- Length variability - Should be able to differentiate students on length of learning history and whether they are newly enrolled
- Data imbalance - K-12 online interactive courses have relatively low dropout rates in comparison to other forms of learning.

Calculation of risk algorithm

To predict whether a student is at risk of dropping out of school, we can represent this as a multi-class classification problem, the students can be segmented into three classes of ‘High Risk’, ‘No Risk’, and ‘Low Risk’. For this, we use various supervised and unsupervised learning algorithms such as ‘K-Means Clustering’, ‘Decision Tree’, ‘RandomForest’, ‘Naive Bayes Classifier’ based on metrics such as performance, attendance, and engagement index. We can also define thresholds in the engagement index to classify students in the three classes.

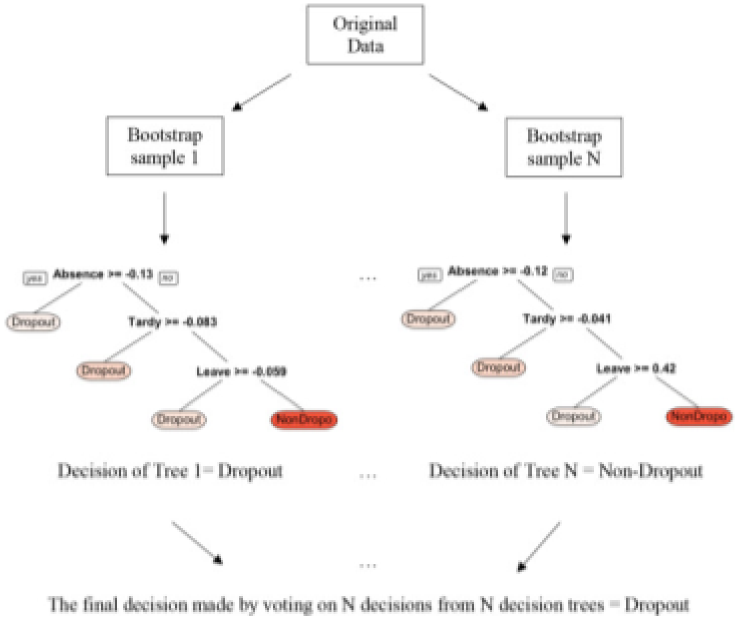


Fig 1 Example of a Decision Tree for students dropping out. Source 2

Algorithms

Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible features, including events like attendance, faculty_Score, and performance It is one way to display an algorithm that only contains conditional control statements.

Random Forest

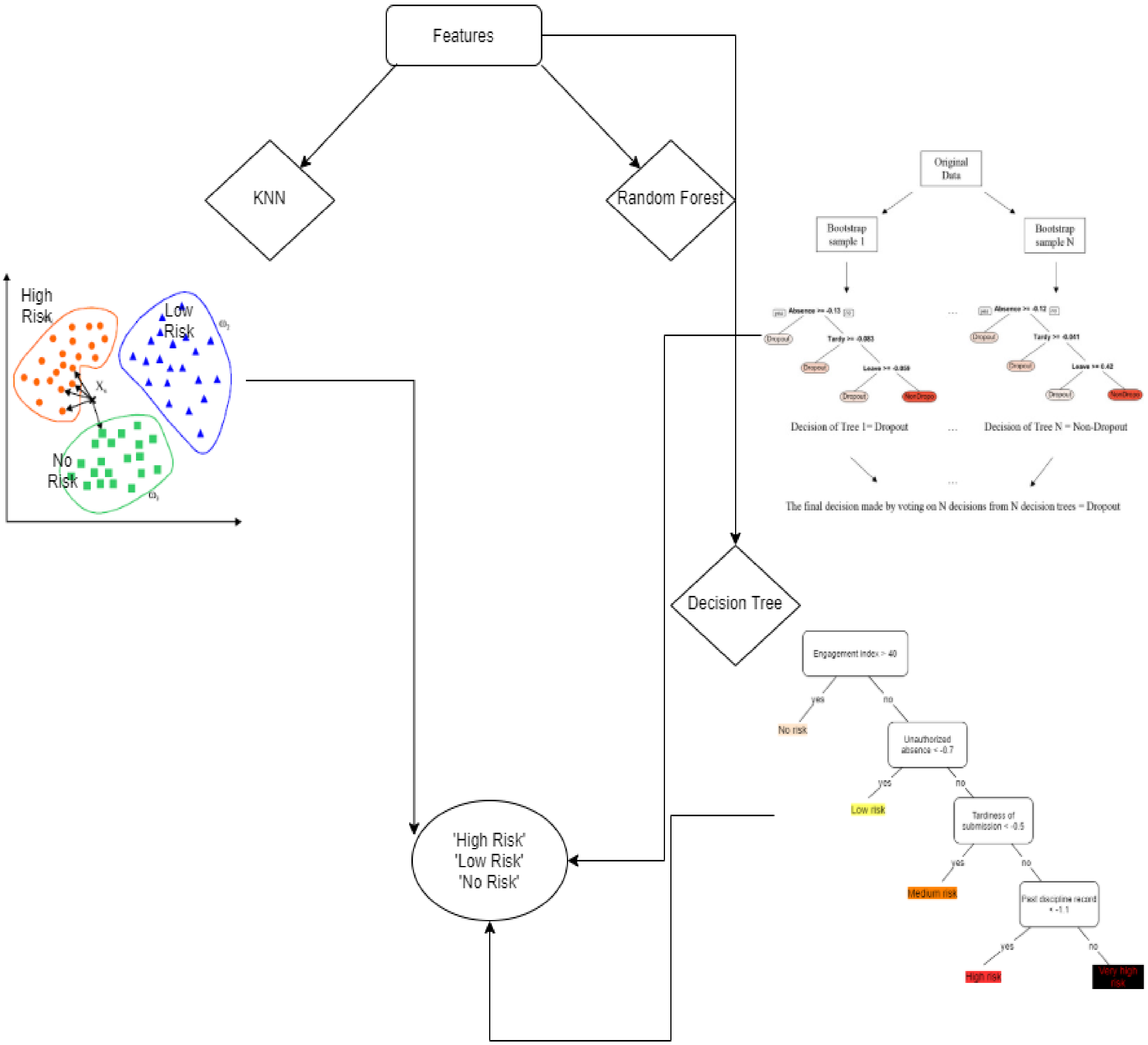
It's an ensemble of Decision Trees. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Naive Bayes Classification

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
p(student=dropout/features) = p(attendace/student drops out) x p(high score/student drops out) x p(late submission / student drops out) ….

K means Clustering

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. Hence, all points are classified as 'high risk', 'low risk', 'no risk'


Features
KNN
Random Forest
Decision Tree
High Risk
Low Risk
No Risk
X_u
ω_1
ω_2
ω_3
Original Data
Bootstrap sample 1
Bootstrap sample N
yes
Absence >= -0.13
no
Dropout
Tardy >= -0.083
Dropout
Leave >= -0.059
Dropout
NonDropo
yes
Absence >= -0.12
no
Dropout
Tardy >= -0.041
Dropout
Leave >= 0.42
Dropout
NonDropo
...
Decision of Tree 1= Dropout
Decision of Tree N = Non-Dropout
The final decision made by voting on N decisions from N decision trees = Dropout
Engagement index > 40
yes
no
No risk
Unauthorized absence < -0.7
yes
no
Low risk
Tardiness of submission < -0.5
yes
no
Medium risk
Past discipline record < -1.1
yes
no
High risk
Very high risk
'High Risk'
'Low Risk'
'No Risk'

Results

- Got 99.81% test accuracy using KNN Classfier
- Got 99.1 % test accuracy using Random Forst Classifier

Dataset

OULAD: Open University Learning Analytics Dataset

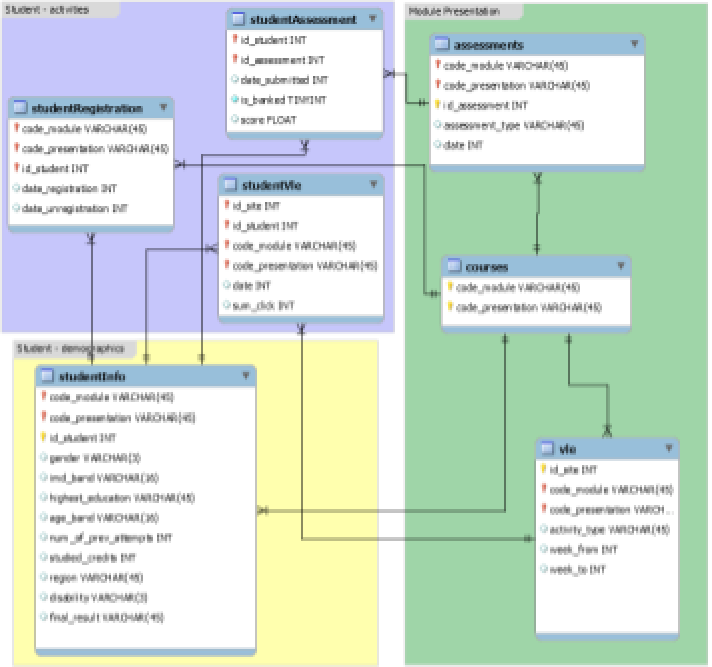


Fig 2 Database Schema of OULAD
Source 1

This is an anonymised dataset that has a data frame with 32593 rows and 12 variables like the final result of a student in the courses taken, the region in which the student lives, gender, VLE engagement, disability of students, code of module for which the student has registered, and so on. The dataset has an equal representation of Male and Female students and students from different regions of the UK.

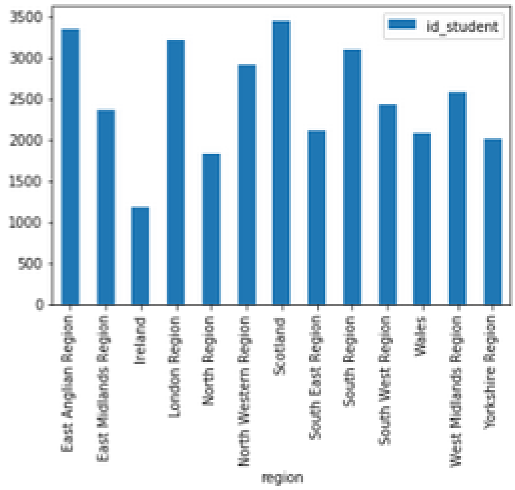


Fig 3 Distribution of regions

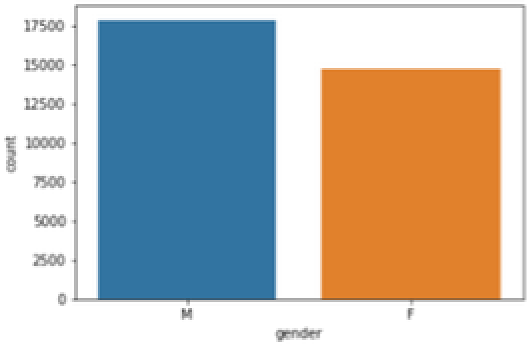


Fig 4 Distribution of Male and Female.

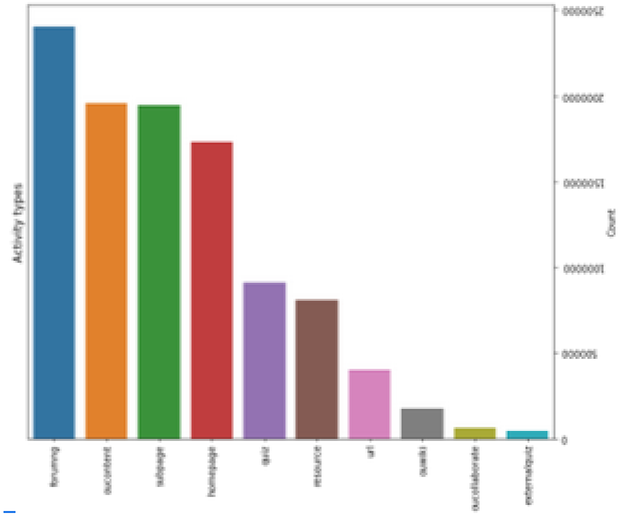


Fig 5 Distribution of activity type in data.

References

1. https://analyse.kmi.open.ac.uk/open_dataset
2. Dropout early warning system for high school students using machine learning, Chung and Lee. Children and Youth services review, 2018
3. https://arxiv.org/abs/2003.09670
4. https://elearningindustry.com/types-big-data-extract-lms-how-use
5. http://www.imsglobal.org/caliper-analytics-v11-introduction
6. Beck, Joseph. (2005). Engagement tracing: using response times to model student disengagement.. 88-95.
7. Fredricks, Jennifer & Mccolskey, Wendy. (2012). The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments. 10.1007/978-1-4614-2018-7_37.
8. Wilson, S., George, D., & Cambron, M. (2008, June). Algorithm for Defining Student Engagement. In *2008 Annual Conference & Exposition* (pp. 13-165).
9. https://machinelearningmastery.com/rescaling-data-for-machine-learning-in-python-with-scikit-learn/