Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera
Marlon Ferrari

Table of contents

1. Introduction

A description of the problem and a discussion of the background

The Internet revolution brought more than social medias and faster information exchanges. It brought also a generation of people who studies through the digital environments. Under this context, the online education evolved quickly and the transformation of the societies really started. Nowadays, people in distant places, poor countries can benefit from technology to achieve information and in this case, the Massive Open Online Courses, MOOCs had a major role. MOOCs can join people all around the world to achieve understand in a wide range of areas, delivering science and culture.

It is known, also, that online learning suffers massive unenrollment. The logical border and the lack of motivation can make the students leave. Under this context, what are the related features which causes it? How understand the student scenario and predict his churn or low grades? I think that is a relevant point. If MOOCs platforms achieve student understanding and predicting, I think it's possible to menage the student's churn and find a way to give them the needed motivation.

With this set in mind, I started a search for MOOCs generated Students Data to investigate and prepare some conclusions about the theme.

2. Data

A description of the data and how it will be used to solve the problem
To guide my investigation, I was looking for a Set to help to understand the student's behavior, motivation and correlated characteristics in order to better understand why or how is the result of an enrollment. So, it is important to find a dataset with some key features like grade, gender, enrollment levels, and so on. Location data is also important to understand cultural marks, which will be explored by locations APIs. Guided by the analysis exploration, I'll be able to build a model to predict student's behavior or results. After querying correlated datasets in order to find those with better columns, I found a nice DataSet from Kaggle called "Students' Academic Performance Dataset". You can check it here https://www.kaggle.com/aljarah/xAPI-Edu-Data.

The data compounds 16 columns with aggregated informations about over 480 students of a Learning Platform called Kalboard360. The datails will be shown next section.

2.1 Data Structure

As previously mentioned, this dataset includes 16 columns:

1. Gender - student's gender (nominal: 'Male' or 'Female')
2. Nationality- student's nationality (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')
3. Place of birth- student's Place of birth (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')
4. Educational Stages- educational level student belongs (nominal: 'lowerlevel','MiddleSchool','HighSchool')
5. Grade Levels- grade student belongs (nominal: 'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12 ')
6. Section ID- classroom student belongs (nominal:'A','B','C')
7. Topic- course topic (nominal:' English',' Spanish', 'French',' Arabic',' IT',' Math',' Chemistry', 'Biology', 'Science',' History',' Quran',' Geology')
8. Semester- school year semester (nominal:' First',' Second')
9. Parent responsible for student (nominal:'mom','father')
10. Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)
11. Visited resources- how many times the student visits a course content(numeric:0-100)
12. Viewing announcements-how many times the student checks the new announcements(numeric:0-100)
13. Discussion groups- how many times the student participate on discussion groups (numeric:0-100)
14. Parent Answering Survey- parent answered the surveys which are provided from school or not (nominal:'Yes','No')
15. Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:'Yes','No')
16. Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

The most important characteristic of this dataset is that it has included the parent's data, which is a nice approach to understand the student.

3. Methodology

The first steps are the data exploration and insight-taking approach. This will understand the data and the columns. The purpose of this exploratory analysis is to identify hidden features and understand the relations between the features. Next, I will do a descriptive analysis by building a dataset for a clustering algorithm. This way, the data understanding will become a more powerful decision-making, focused on student's behaviors. Finally, I will create a predictive analysis by building a dataset with the best features for a supervised learning algorithm to predict the student's behavior under certain conditions, which will achieve my final objective.

4. Analysis

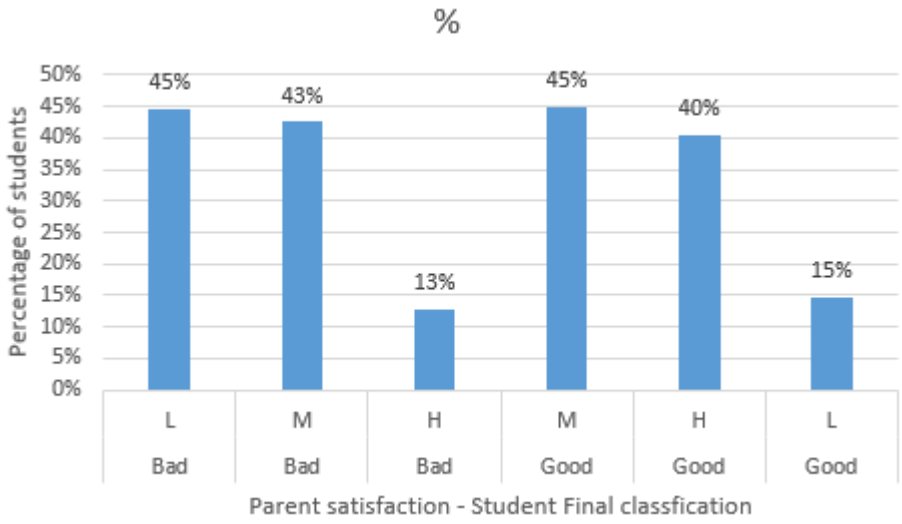
As mentioned, this section will understand the data in order to compose the clustering dataset.

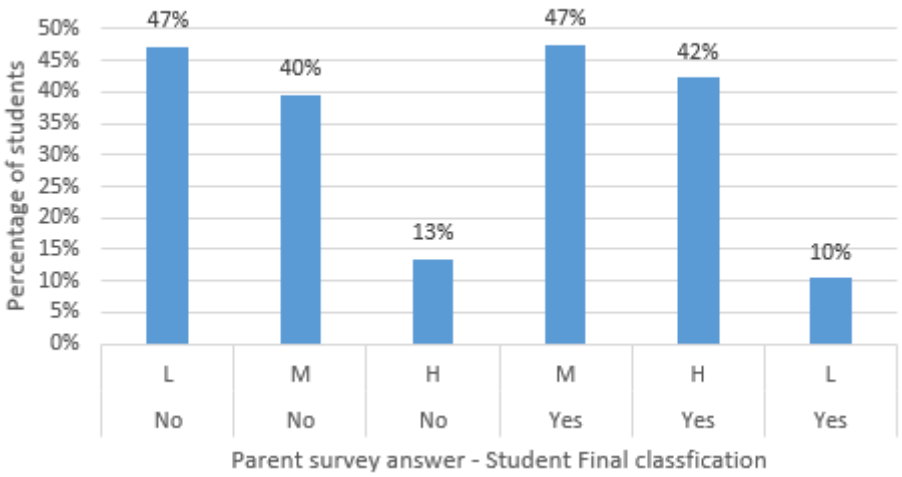
4.1 Exploratory Analysis

In the context to understand the student and his results, setting up a dataframe with the following columns:

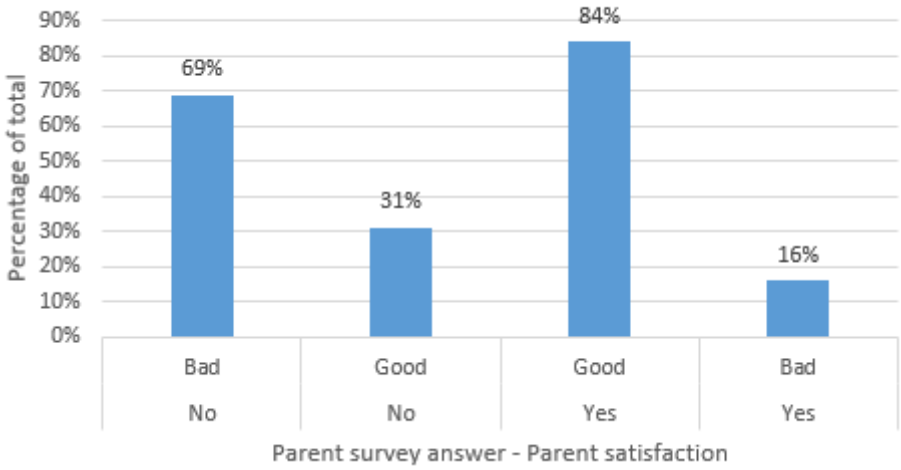
'gender','PlaceofBirth','StageID','Topic','raisedhands','VisITedResources','AnnouncementsView','Discussion', 'ParentAnsweringSurvey','ParentschoolSatisfaction','StudentAbsenceDays', 'Class'

Then, we are guiding some analysis based on attributes and student final score level. The next two results shows the percentage of student classification and parent behavior (satisfaction and answering surveys):





The next analysis shows that parents not involved in answering the scholar's surveys, are likely to become unsatisfied with the School. This can mean that well-informed parents can better understand the student's enrollment and reality and are better satisfied.



4.1.1 Understanding student's behavior

The features related to student behavior are:

1- Raised hands
2- View course resources
3- View course announcements
4- Join discussion groups

We are going to classify the data into 3 bins in order to normalize our analysis. Those bin are based of frequency of an action:
0 – Low frequency
1 – Medium frequency
2 – High frequency
Related to raised hands attribute, the results shows that students with high raised hands actio ns are most like to have mid-high classifications. On the other hand, low raised hands frequenc y tends to a mid-lower classification.

raisedhands	Class	
0	L	0.534314
	M	0.392157
	H	0.073529
1	M	0.577778
	H	0.288889
	L	0.133333
2	H	0.543011
	M	0.424731
	L	0.032258

About the frequency of course resources visit, there is a positive link between high frequency visits and high classification levels. The opposite is also true (lower frequency linked to low classification).

VisITedResources	Class	
0	L	0.656250
	M	0.293750
	H	0.050000
1	M	0.560976
	H	0.231707
	L	0.207317
2	M	0.495798
	H	0.483193
	L	0.021008

In addition, the announcements visualization is also related to the same as shown.

AnnouncementsView	Class	
0	L	0.468354
	M	0.388186
	H	0.143460
1	M	0.506667
	H	0.393333
	L	0.100000
2	H	0.526882
	M	0.462366
	L	0.010753

The next attribute, the discussion frequency shows that there is a link to the classification, but not so strong linked related to previous attributes. This can indicate that discussion, despite relevant, is not directly or, in other words, mainly linked to the student classification and we can conclude that discussions are important to the secondary learning process (reinforcement of content knowledge).

Discussion	Class	
0	M	0.416290
	L	0.371041
	H	0.212670
1	M	0.538462
	H	0.253846
	L	0.207692
2	H	0.480620
	M	0.379845
	L	0.139535

As shown in the next results, the lower the absence of the student, the higher tends to become their classification.

StudentAbsenceDays	Class	
Above-7	L	0.607330
	M	0.371728
	H	0.020942
Under-7	M	0.484429
	H	0.477509
	L	0.038062

4.1.2 Clustering DataSet

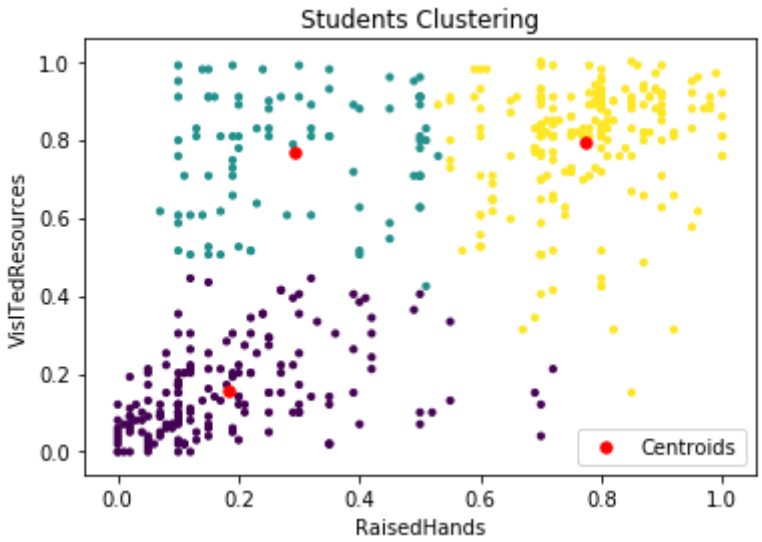
The following analysis identifies what are the correlation between the dataset features, which will guide the build of the clustering model.

	raisedhands	VisITedResources	AnnouncementsView	Discussion
raisedhands	1.000000	0.691572	0.643918	0.339386
VisITedResources	0.691572	1.000000	0.594500	0.243292
AnnouncementsView	0.643918	0.594500	1.000000	0.417290
Discussion	0.339386	0.243292	0.417290	1.000000

As shown, the best correlation is between raised hands and visited resources. So, we are going to build the K-Means clustering with those attributes as features.

The best K factor found for data is 3:

- High applied Students
- Mid Applied Students
- Low Applied Students



4.2 Building a supervised algorithm

As mentioned before, one of the purposes of this research is build a model to predict student classification in order to help teachers, parents and students improve their actions. So, this section will discuss the build and performance of the following algorithms:

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

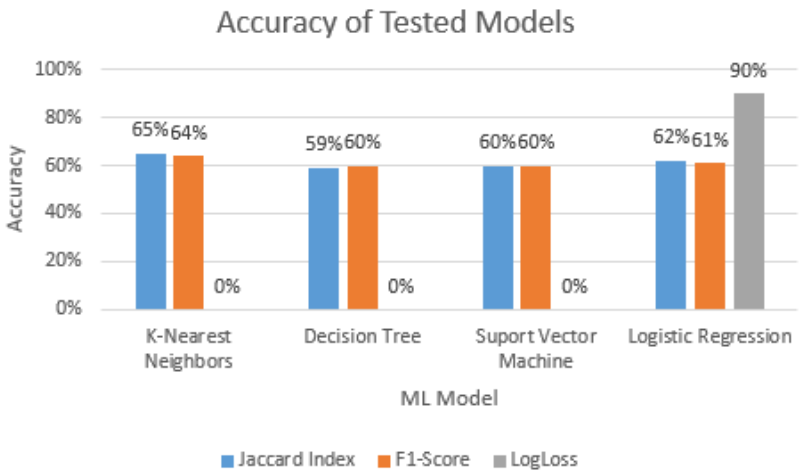
The train and test sets configuration is 80 (train)-20 (test):

Train set: (384, 4) (384,)
Test set: (96, 4) (96,)

5. Results and Discussion

The following table shows the performance of the supervised algorithms. The result analysis involved three score systems:

1- Jaccard Index
2- F1-Score
3- LogLoss (For logistic regression)



The score index shows that the best was K-nearest neighbors. This model can now help to predict students classification and help them to achieve better results. Also, it is needed model tuning in order to get better prediction accuracy.

6. Conclusion

Data analytics and Data Science are vital fields for improving Online Courses experience. Set the right content for the right student is a complex but essential task to keep the students enrolled, motivated and getting high classification. This will, with no doubt, improve education levels of their countries and help to improve their economy. Despite that, online learning has a potential value for increasing society levels.

The presented research focused in data analytics and building a machine learning model to understand the student's behavior and classification under online learning courses. It concludes, at its full scope, some points:

- Parent's active participation and tracking are important. Absent parents are linked to absent students and more unsatisfied.
- Students who read announcements and visit the course resources are most likely to have higher classification.
- Actions related to discussions are less likely to improve student's classification.

As a result, a predictive model can help the online platforms to understand the student's acts and take decisions. The best model was the K-Nearest neighbors with k=4 and accuracy of 0.65 Jaccard Index. Finally, it is important to mention that location data was not possible to be used. That is because it refers to born location of the student and this is not a important feature. Therefore, it could be a more important data the place where the student was connected, because the high absence levels could be related to poor Internet connection areas, like conflicted-areas and under development countries. Therefore, this research is a starting point for further works and model adaptation.