Students Performance Prediction in Exams

Abstract

From primary education to university education organizations scope of evaluating student's data has increased for the past three decades across the globe. Multiple techniques have evolved to uncover patterns from data collected over the students Many data mining techniques are proposed to extract the hidden knowledge from educational data, so the observations discovered during such analysis would be used to enhance student performance, retention and intern change the way institutions approach learning based on the outcome of the analysis. Student predictive model built in this project is to classify if the student is classified in to Low, Middle and Hight level performers in exams based on the set of attributes gathered prior to exams. This is evaluated by classifiers using SVM, Neural Network MLPClassifier, LGBMRegressor, Bayesian and Decision tree, that produced an accuracy rate of up to 85% for students performing at the low,72% for Hight performers with 69% at the middle performers, these are based on the f1-score that gives harmonic mean of precision and recall. Used a fictious dataset and a secondary data set sourced from The University of Jordan student system.

Background

Sixty percent of college students attend more than one institution over the course of their academic career (Adelman, 2006; Peter & Forrest Cataldi, 2005). This study describes processes involved in analyzing and predicting student performance rate for educational institutions and see fit and be scalable to programs offered in both private and government sectors of higher education, including community colleges. For each student, the institution admitted are increasing creating a profile of students' academic caliber and concentrate on students that would fall behind on academic progress & goal of the purpose of joining and pursuing the courses. Orgs in collaboration with institutional research plans to create profile of each student to understand if student would complete the program on time, and if they would drop off program midterms/year. And or students looking for other opportunities in a different institution by transferring. Creating such profile ahead of time using predictive analytics by Educational Orgs help both students in custom approach in resolving issues & difficulties in completing the program and intervein if such students would be counseled for a better outcome of their educational goals. At the same time help their own institutions on loosing monitory value in foregone revenue otherwise would

have gained if student does not drop off midway or planning to transfer to other institutions, because such Students did not learn to fit into the society of the institution removed themselves from the institution or many other factors that could be mitigates if intervened early.

Business Understanding: Defining the Problem

Prior such studies have focuses mainly on quantitative, focusing primarily on sociodemographic variables and not custom approach at student individual level and factors particular to the program's students enrolled to identify the issues. Students who are having difficulties completing the programs show signs by many factors, including and not limited, by not attend the classes on a regular basis, falling behind on schoolwork evident from grades on a regular basis and or having difficulty paying tuition fees for subsequent terms, interaction effects such as on-campus versus off campus residence or institution type. Many other factors effect academic preparation, student disposition, the student peer environment, individual student experiences, organizational factors, and external pressures. Student counselors would have to interfere into such students only after the fact they started falling behind and it may be too late to bring back such students on track and counselors may not have all reasons causing the issues in the first place. Or on the other hand, there are students who are looking for more challenging programs and are planning or already requested to transfer their credits to other programs outside the existing institutions. Having such student information along with factors causing the behavior to student counselors would be beneficial to intervene before irreversible damage has already been done.

Defining the target variable:

Colleges and Universities have been collecting large amounts of data for "conducting business" in student information systems. Those includes xAPI-Edu-Data, which contains more Gender, Nationality, Educational Stages, Grade Levels, course topic, school year semester, Parent responsible for student, raised hand, visited resources, viewing announcements, Discussion groups Parent Answering Survey, Parent School Satisfaction, Student Absence Days etc. And this practice has been in place for many decades for statutory reporting requirements. Also, organizations like the National Student Clearinghouse have enabled colleges to track students who attended their institution, left for another institution, and graduated and other meta data is available. Also, other sources of data are annual student surveys sent to students. And the target variable is the grade a student would get which are classified in to Low-Level: interval includes values from 0 to 69. Middle-Level: interval includes values from 70 to 89. High-Level: interval includes values from 90-100.

Data Understanding:

Data columns are.

```
data.columns
```

```
Index(['gender', 'NationalITy', 'PlaceofBirth', 'StageID', 'GradeID',
       'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands',
       'VisITedResources', 'AnnouncementsView', 'Discussion',
       'ParentAnsweringSurvey', 'ParentschoolSatisfaction',
       'StudentAbsenceDays', 'Class'],
      dtype='object')
```

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','Middle-School','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', 'Bi-ology', '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(nu-meric: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')

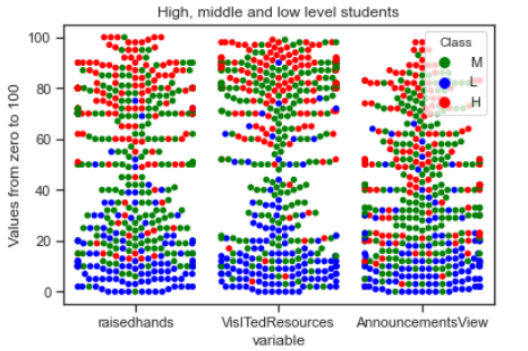
15 Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:'Yes','No')

16 Student Absence Days-the number of absence days for each student (nominal: above-7, un-der-7)

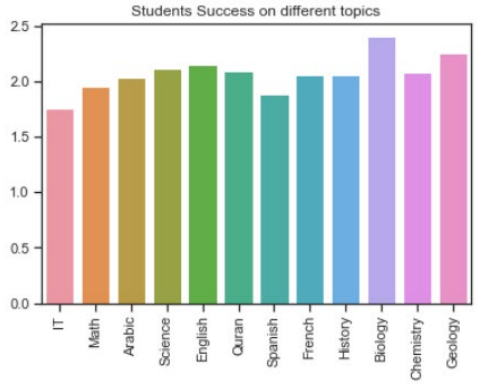
Data Exploration:

First step upon importing the dataset was to convert parameters/features into a more useful format and add some columns that may be useful for visualization, modelling and analysis. Based on the information in dataset, a correlation study was conducted to see how strongly, or weakly variables were related and the magnitude of change in one variable with respect to target variable was calculated. We plotted bar charts to understand the mean value of variables. In the process changed the DataFrame format from wide to long and a identifiers for 'class' with 'raisedhands', 'VisITedResources' & 'AnnouncementsView' to make swarmplots.

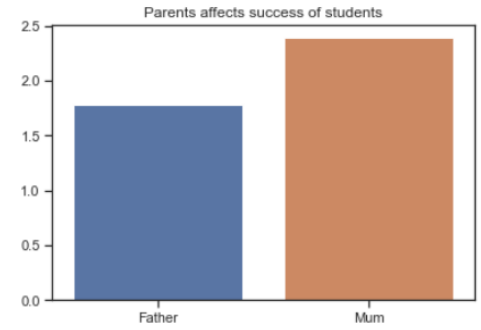
```
Text(0.5, 1.0, 'High, middle and low level students')
```

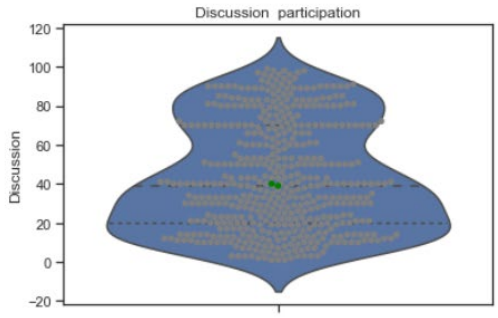


This swarm plot shows how students have higher values of raising hands, visiting resources, and viewing announcements take high level values. At the same time, we can see students who take l ow level although they have higher values of raising hands, visiting resources, and viewing annou ncements. This leads us to explore why these students have low levels despite having higher valu es of raising hands. To do this we have to give class numeric values and then check if there is pari ty for boys and girls, which shows 1.88 for boys while gals have 2.29 average. But it does not tell us much about the parity between the raised hands that are higher for low levels. To find other c orrelation factors plotted the students by the nationality that show how it fares along with topics of study.

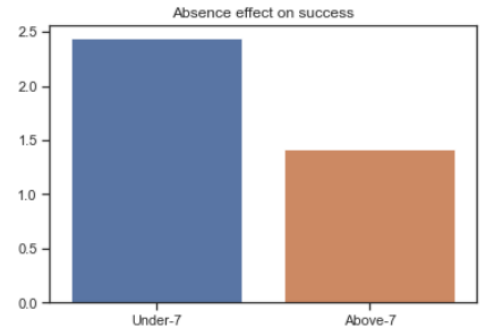


There is no clear-cut evidence even after these EDA graphs, so digging further into family members on their effect, that show some sign of mother influence on success rates while discussion rates of participation have effect on students.





And on absence effect on success rates is illustrated by the graph, that show if absence is higher for having lower levels.

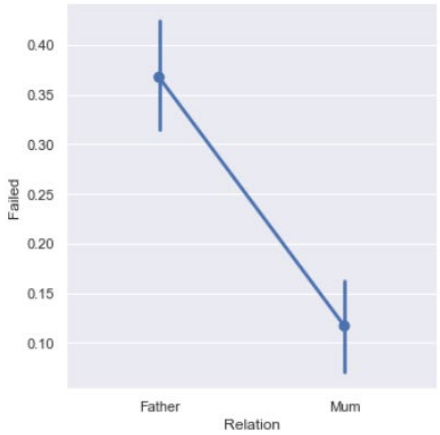


Up on looking up further on failed classes it is evident that Geology has no failed students, while IT, French, Arabic and Science has high failed rates.

t[31]:

Topic Class	Arabic	Biology	Chemistry	English	French	Geology	History	IT	Math	Quran	Science	Spanish
H	19	16	10	17	20	6	4	15	6	8	16	5
L	17	4	8	10	16	0	3	38	7	6	10	8
M	23	10	6	18	29	18	12	42	8	8	25	12

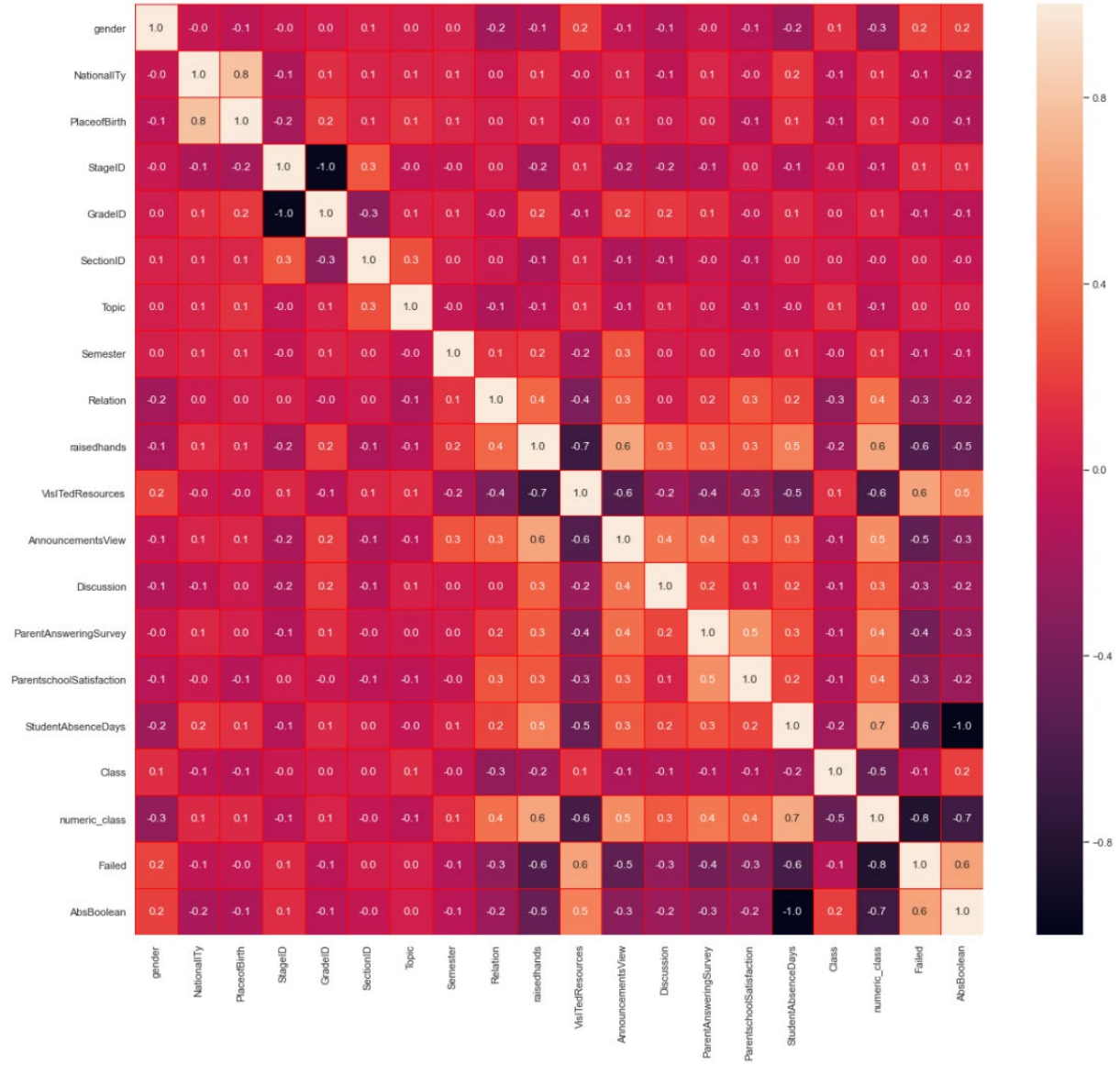
And when studied the rational of students whose parents are generally satisfied with the education they received, while parents were least satisfied with the school performed much in lower levels.



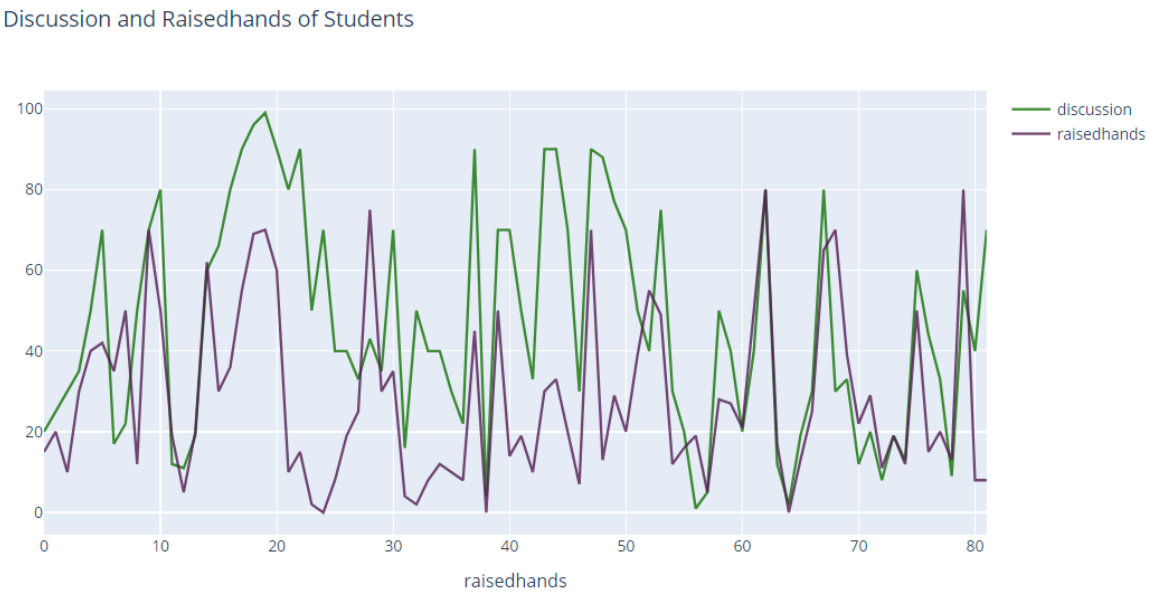
This graph also says the influence of mothers on student's failure rates, that is mothers influence has lower failure rates. Also, lowest performers did not visit the course resources as much as high performers.
Looking deeper in to StudentAbsenceDays with respect to courses, we can see geology students seemed to participate more frequently compared to other subjects in attending the class more than those in any other subject, which could explain why none of the Geology students failed.

```
Out[35]: Topic
         Arabic       0.389831
         Biology      0.266667
         Chemistry    0.500000
         English      0.422222
         French       0.323077
         Geology      0.250000
         History      0.473684
         IT           0.473684
         Math         0.476190
         Quran        0.318182
         Science      0.450980
         Spanish      0.320000
         Name: AbsBoolean, dtype: float64
```

Below diagram shows the correlation plot of different input which indicate strong correlation between place of birth, raised hand and announcements checking, absence days to parent's school satisfaction.



We can also see Discussion and Raisedhands of students goes hand in hand.



Data Preparation, transformations & Imputation:

Our data did not have any missing values. Input dataset already had numerical values, so less effort was done to prepare the Data. Some of the variables like 'TotalQ' derived from input dataset variables class with low, mid, and high level and used melt to widen the data frame.

Since the data is from many different sources of student information systems, surveys and other sources, the data needs to be cleaned and imputations like Converting all features in all the data extracts to strings.

Post string conversion make numeric and nominal variables for character response and numeric responses.

Grade level standardization for each course

Removing rows with empty target variable

Creating different binning response variables in both train and test data sets.

Create median, IQR features, for resources visited.

Replacing features with features with wider granularity to reduce missing values.

Imputing missing values in most of the features (most common)

One Hot Encoding for students' absences for each course.

Creating response variables in both train and test without outliers as per IQR range, creating test and train data sets removing outliers

Winsorizing response variables in both train and test data sets

Modeling:

To train the algorithm, we implement a split train/test methodology to prevent bias in the learning. The dataset is split into a randomly sampled pool of datapoints. 70% of those points are used for training, the remaining 30% is used for validation of the training data. So, the test data is not necessarily continuous time, but rather a random selection of points from the set. Then Fit data, predict data, and find accuracy of the models. Tried several models and compared their performance to evaluate the best algorithm to predict student performance. We trained data to fit on different models like StandardScaler, SVM & MLPClassifier for example.

Results:

SVM classifier results using sklearn library has produced an accuracy of 67% with 'rbf' kernel, while 'linear' kernel produced an accuracy of 64%, with highest precision of 71 for Low performers, middle performers at 65 and high performers at 54

```
print(classification_report(y_test, y_pred))
```

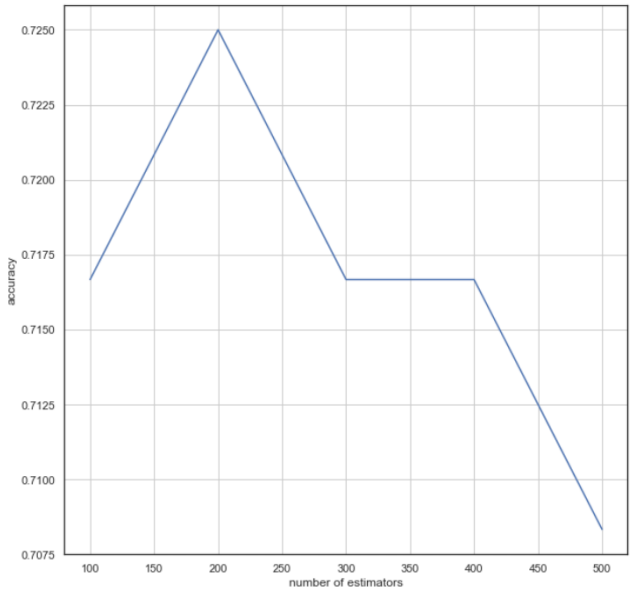
```
              precision    recall  f1-score   support

           H       0.54      0.49      0.51        39
           L       0.71      0.85      0.77        34
           M       0.65      0.62      0.63        71

    accuracy                           0.64       144
   macro avg       0.63      0.65      0.64       144
weighted avg       0.63      0.64      0.63       144
```

By replacing y values with derived 'TotalQ' values of Low, Mid and High values, the accuracy went up to 74%. While each value of H, L and M went up to 0.72,0.85 and 0.69 respectively.

```
print(classification_report(y_test, y_pred))
```

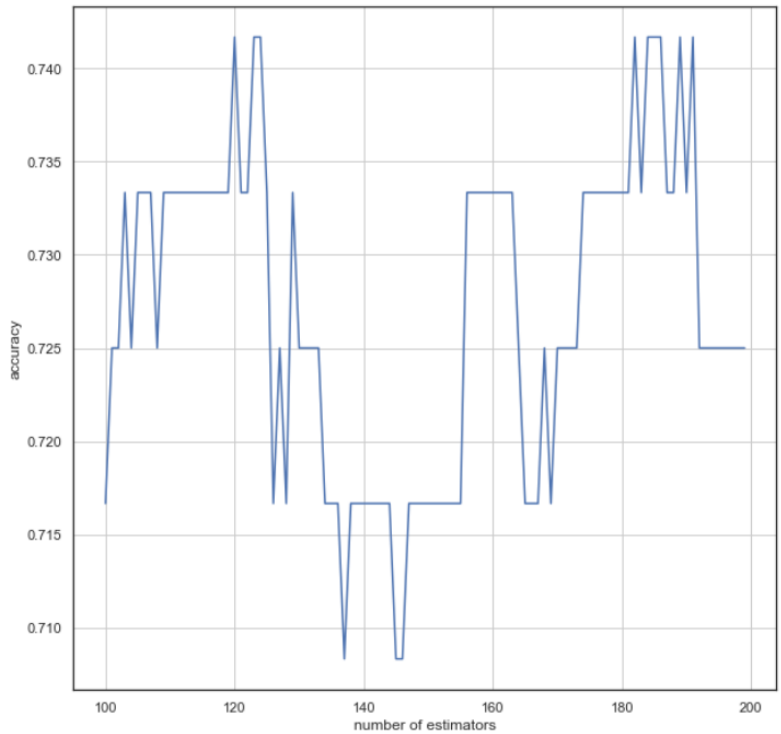
```
              precision    recall  f1-score   support

           H       0.72      0.46      0.56        39
           L       0.85      0.85      0.85        34
           M       0.69      0.83      0.76        71

    accuracy                           0.74       144
   macro avg       0.76      0.72      0.72       144
weighted avg       0.74      0.74      0.73       144
```

Random Forest Classifier




```
Maximum value of accuracy is 0.725
when n_estimators= 200.
```

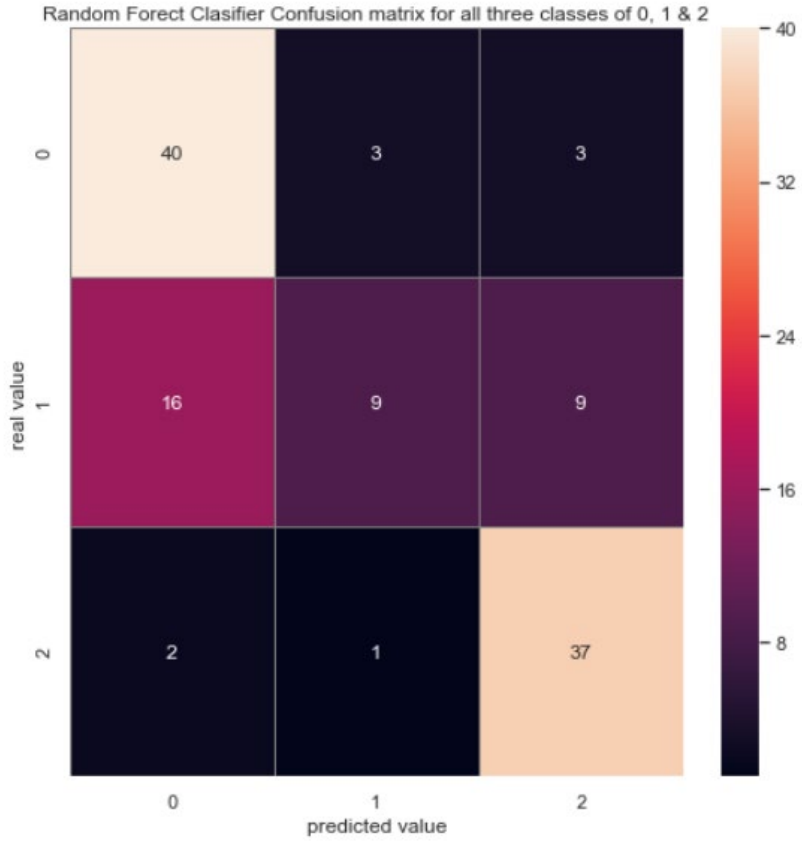
From the graph we can see n_estimators=200 for the best accuracy result and so redo the estimator 100 and 200, which went by accuracy to 0.74




```
Maximum value of accuracy is 0.7416666666666667 when number of estimators between 100 and 200
```

Based on the model derived below table tabulates each algorithm results in accuracy

Algorithm Name	**Accuracy**
SVM:	67%
MLPClassifier	67%
Liner Model	62%
GridSearchCV	70%
RandomForestClassifier	74%
GaussianNB	65%
DecisionTreeClassifier	65%



	precision	recall	f1-score	support
0	0.69	0.87	0.77	46
1	0.69	0.26	0.38	34
2	0.76	0.93	0.83	40
accuracy			0.72	120
macro avg	0.71	0.69	0.66	120
weighted avg	0.71	0.72	0.68	120

It is evident Random Forest Classifier is the best performing classifier at 74% and would be a best fit for the given data and based on the overall performance of the classifier will be determined by average Precision and Average Recall.

Conclusion/Discussion:

Factors effecting student performance from EDA is evident with nationality (Jordan nationals) performed better, working on Chemistry, and having relation with mother has positive effect on the levels. While being male, participate(less) in discussion and absence more than 7days has negative effect on the scores.

Based on above accuracy metrics Random Forest Classification at 74% is more suited for the data set in predicting students' performance.

Acknowledgements & References:

1) Student Performance Data Set

https://archive.ics.uci.edu/ml/datasets/student+performance

2) Students' Academic Performance Dataset xAPI-Educational Mining Dataset

https://www.kaggle.com/aljarah/xAPI-Edu-Data

3) Why is Educational Data Mining important in the research?

https://towardsdatascience.com/why-is-educational-data-mining-important-in-the-research-e78ed1a17908

4) Educational Data Mining (EDM)

https://www.cmu.edu/datalab/getting-started/what-is-edm.html

4) International Journal of Database Theory and Application. Mining education data.

http://article.nadiapub.com/IJDTA/vol9_no8/13.pdf

5) Educational Data Mining & Students' Performance Prediction

https://pdfs.semanticscholar.org/b280/216a1d63015afc6a3d1aac9595aeb2b7dd5a.pdf

6) Educational Data Mining: Student Performance Prediction in Academic

https://www.researchgate.net/profile/Mukesh-Kumar-234/publication/332369964_Educational_Data_Mining_Student_Performance_Prediction_in_Academic/links/5cb03346299bf120975f8dc1/Educational-Data-Mining-Student-Performance-Prediction-in-Academic.pdf

7) Educational Data mining for Prediction of Student Performance Using Clustering Algorithms

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.567.8824&rep=rep1&type=pdf

Q & A

1. What is xAPI used for?

   This is an API that can get student related data sets that allows learning content experiences a students have either by online or offline learning activities.

2. Are there datasets to get class test scores, sports activity, prior education pattern, activity level in class, social behavior, sports activity, parent education for the prediction model to be comprehensive & get better accuracy?

   Not a lot of such data is publicly available, since student data contains lot of sensitive information and is protected by law both federal, state and other local laws.

3. Do student repeat for each course or only one student one row in the data set?

   One student per course

4. Why use individual encoding vs LableEncoder?

   I want to know how each feature encoding works by each model, LableEncoder would equaly work for categorical features encoding.

5. How did you segregate the interval starting 0-69 into low level and so on?

   No specific rule outlined, but this is a general rule to segregate low, mid and high. It can be adjusted as required.

6. What is the purpose of swarm plot?

swarm plots are to avoid obscuring points by calculating non-overlapping positions instead of adding random jitter, very much like strip plots. Here in my analysis this shows discussion participation in relation to others.

7. What effect does the models have effect on the outliers?

   Models like RandomeForest, SVC & Grid search model performed the best on the dataset. One possible contributor to this could be that no outliers were removed.

8. What model evaluation techniques used for all models?

   Confusion matrix evaluation is done for all models. The avg of overall performance of the model was determined by average Precision and Average Recall values for each model.

9. How well will it scale to more general data and or fictional data?

   It is a challenge to overlay this as a generalize model, because of many factors influence student performance and we must factor in each student case (data) to formulate conclusions.

10. What are next steps?

    Science finding real life data is a challenge, I'm trying to approach institutional research of multiple public universities to see if they can mask sensitive data and provide near real time data for further analysis and fine tuning the model.