STATEMENT OF THE PROBLEM

The purpose of this project will be to discover the relationship between high school curriculum and student behavior with student performance in universities using educational data mining techniques to help the education community make better informed decisions to improve the chances of success of students at universities.

1.1. Data Collection

For data collection, approval was sought from the Institutional Research Ethics Board (IREB). The academic data collection approval from IREB is given in Appendix 10.1 and the approval form for non-cognitive data collection is given in Appendix 10.2.
The approval was given for collection of anonymous data of students.
For non-cognitive skill data was collected using a self-assessed survey among the undergraduate students.

The research population consists of students from RIT Dubai who started their undergraduate studies at RIT Dubai in 2014, 2015, 2016 and 2017.

The research population size used was different for academic data and for non-cognitive data.

Academic data

The academic data contained 213 rows, which represent 213 undergraduate student records.
The 31columns represent 31 attributes of each record. The attributes are a mix of numeric and character datatypes.
Academic data collected from RIT had high school scores, placement exam scores and undergraduate scores.

Non-Cognitive data

The non-cognitive data was collected using a survey which was created using a combination of 2 survey instruments:
the Grit Scale (Duckworth et al., 2007) and Self-control survey (Tangney et al., 2004).
The survey developed and the informed consent documents are given in Appendix 10.3.

Some of the other instruments that were referenced and considered were Cattel's 16 Personality Factor Test (Cattell et al,1972), Big Five Inventory (John and Srivastava, 1999), surveys from Search INSTITUTE and SuccessNavigator (Rikoon et al., 2015).

Non- Cognitive skills was measured in terms of 2 non-cognitive skills, Grit and Self Control.
The research population for the non-cognitive data were the respondents to the survey whose academic records were part of the academic data records used in the project.
The non-cognitive data set contained 67 records representing 67 such students.

1.2. Data Preparation

Academic data

The academic data obtained from RIT needed cleaning, standardization and feature engineering before any analysis could be done.

Only relevant features were selected for analysis from the 31 features. Hence irrelevant features such as exam dates, High School names etc. were removed.

Data completeness: Missing values of numerical input variables were filled with the mean of each field in order to ensure data completeness.

The features used in analysis can be summarized as

Curriculum	High school curriculum
MathGrade	High school Mathematics scores
EnglishGrade	High school English scores
PhysicsGrade	High school Physics scores
MPE	Mathematics placement scores
PPE	Physics placement scores
EPE	English language Proficiency scores
GPA	Cumulative GPA score from RIT
GritScale	Grit calculated from survey results
SelfControlScale	Self-Control calculated from survey results

The feature engineering and preprocessing done for each variable is given below:

1. Grade Point Average (GPA)

Grade Point Average is the dependent variable. It represents the 4-point Cumulative grade point average (GPA) of each student. This is the latest GPA of the student, i.e., for a student from 2017 application term, it will be the 1st Year GPA score and for a student from 2014 application form, it will be the 4th Year GPA score. GPA is thus a calculated field.

Note: An undergraduate student must have a minimum of 2.0 GPA to graduate

2. Curriculum

Data had to be cleaned for conformity and consistency. 7 records in the dataset had more than 1 curricula in the Curriculum field. A single instance of Curriculum was found for each of these records using values from other fields from the data file such as High School Name and Description/Qualification.

Records with wrong and multiple spellings for the curriculum were corrected and standardized.

In order to have a single instance for the Indian curriculum, all entries with CBSE, Indian, ICSE were standardized by assigning a single name “Indian”.

There were 1 or 2 records from different countries in the African continent which were grouped together and renamed “African”. The records included Angolan, Eritrea, Egypt, DR Congo, South Africa, Ethiopia, Nigerian and Kenyan curricula.

As shown in Figure 2 British, American, Indian, MOE, International Baccalaureate, African and SABIS have more than 10 records.

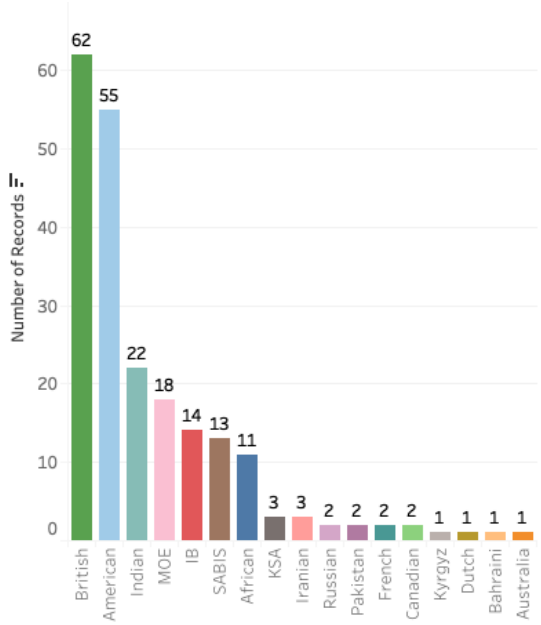


Figure 2. Distribution of academic records

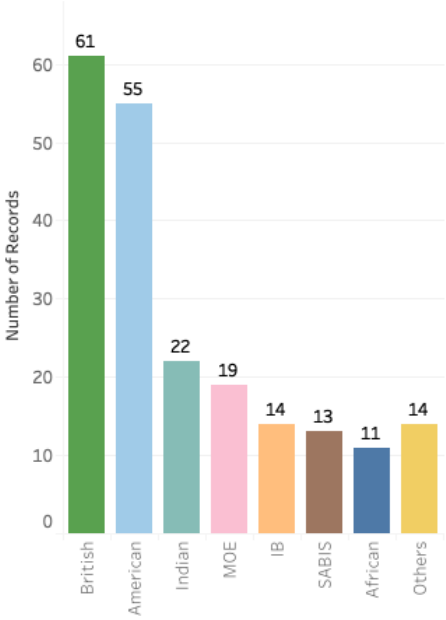


Figure 3. Distribution with Others

There are many curricula with 2 or 3 records including national curriculum of Saudi Arabia, Iranian, Pakistan, Russian, French and Canadian. These were combined and named "Others". The total number of records is 14.

Curricula with 1 student record were removed from the analysis as a single student's academic record will not be representative of the trends of students of that curricula. These include Kyrgyzstan, Dutch, Bahraini and Australia.
Figure 3 shows the new distribution of records after removing curriculum with fewer than 2 records and grouping curricula with 2 or 3 records as "Others".

3. High School scores: MathGrade, EnglishGrade, PhysicsGrade

Different curriculum scores students based on different grading and scoring systems, for example, letter grades for British curriculum, percentage for Indian, 8-point grading in International Baccalaureate etc.

In order to make comparisons, it was essential to normalize or standardize these scores.
All numerical values, including the high school scores, were set to a standardized scale, i.e., to 4.0 scale, so each field varied between 0 and 4.0. This will ensure that no variable gets a higher or lower weight in the regression coefficient due to its larger or smaller scale of measurement.

The standardization was done using R programming based on the conversion charts found on the cited sources. These charts are given in Table 2 for each of the different curricula.

Table 2: High Score conversion charts

LETTER GRADE	GRADE POINTS	NUMERICAL GRADE
A+	4.0	97-100
A	4.0	94-96
A-	3.7	90-93
B+	3.3	87-89
B	3.0	84-86
B-	2.7	80-83
C+	2.3	77-79
C	2.0	74-76
C-	1.7	70-73
D+	1.3	67-69
D	1.0	64-66
D-	0.7	60-63
F	0.0	0-59

American conversion chart
Source: *Princeton review*

GRADE	US GRADE
A*	A
A	A
B	B
C	B
D	C
E	C
F	D

British conversion chart
Source: *scholaro.com*

IB SCORE	GRADE
7-8	A
5-6	B
3-4	C
2	D
0-1	F

IB conversion chart
Source: *ibo.org*

GRADE	SCALE	US GRADE
A	90-100	A
B	80-89	B
C	70-79	C
D	60-69	D
F	0-59	F

MOE conversion chart
Source: *scholaro.com*

PERCENTAGE	GRADE	US GRADE
60-100	A	4
55-59	B+	3.5
50-54	B	3
43-49	C+	2.5
35-42	C	2
0-34	F	0

Indian conversion chart
Source: *scholaro.com*

4. Mathematics Placement exam score (MPE) and Physics placement score (PPE)

MPE and PPE are scores obtained by students at placement exams conducted by RIT Dubai. Each of these exams have a maximum score of 30. The placement exams have equal number of easy, medium and difficult questions. The format followed is objective and multiple choice.

These scores were standardized to 4.0 scale as the other numerical variables.

The conversion formula used was: (Placement Score/30) *4

5. English Placement Score (EPE)

All EPE scores were converted to IELTS scores to get equalized scores and then standardized to 4.0 scale. All EmSat scores were given in IELTS standardized scores in the collected data.

The different TOEFL exam scores were converted to IELTS using Table 3.

The conversion formula used to standardize the scores to 4.0 was - (Placement Score/9) *4

Table 3: English Placement score conversion

TOEFL Paper	TOEFL CBT	TOEFL IBT	IELTS
0 - 310	0 - 30	0 - 8	0 - 1.0
310 - 343	33- 60	9 - 18	1.0 -1.5
347 - 393	63 - 90	19 - 29	2.0-2.5
397 - 433	93 - 120	30 - 40	3.0 – 3.5
437 - 473	123 - 150	41 – 52	4.0
477 - 510	153 – 180	53 – 64	4.5 – 5.0
513 - 547	183 – 210	65 – 78	5.5 – 6.0
550-587	213 – 240	79 – 95	6.5 – 7.0
590-677	243 – 300	96 – 120	7.5 – 9.0

Source: *englishcollege.pl*

Non-Cognitive data

The results from the survey are used to calculate 2 predictor variables: GritScale and SelfControlScale

6. GritScale

Grit is the tendency to sustain interest in and effort toward very long-term goals (Duckworth et al., 2007).
GritScale is a numerical variable that is calculated according to the method specified by Duckworth et al (2007).
The survey instrument is a 12-point self-report questionnaire which was distributed to all undergraduate students at RIT Dubai.

Responses to the questions are scored according to the scoring proposed by Duckworth et al (2007). All scores are added up and divided by 12. This score is then normalized or standardized to 4.0 scale.
A maximum score of 4 indicates "extreme grit", and the lowest score indicate "no grit".

7. SelfControlScale

Self-control is the voluntary regulation of impulses in the presence of momentarily gratifying temptations (Duckworth and Steinberg, 2015).

Like GritScale, SelfControlScale is also a numerical variable calculated according to the methodology specified by Tangney et al (2004).
The survey instrument is a 10-point self-report questionnaire which was dispersed to undergraduate students along with the GritScale questionnaire.

Responses to the questions are scored according to the scoring proposed by Tangney et al (2007). All scores are added up and divided by 10. This score is then normalized or standardized to 4.0 scale.
A maximum score of 4 indicates "extreme self-control", and the lowest score indicates "no self-control at all".

2. ANALYSIS AND RESULTS

2.1. GPA distribution analysis

The Statistical summary of the dependent variable, GPA shows that average GPA of the obtained records is 2.759.

In order to study further the different statistical metrics of the records, boxplots of the records by curriculum is used as shown in Figure 4.

The distribution of records of each curriculum is represented by the length of the boxes and the whiskers. The Inter Quartile Range (IQR) shows that GPA range between the 25th percentile and 75th percentile.

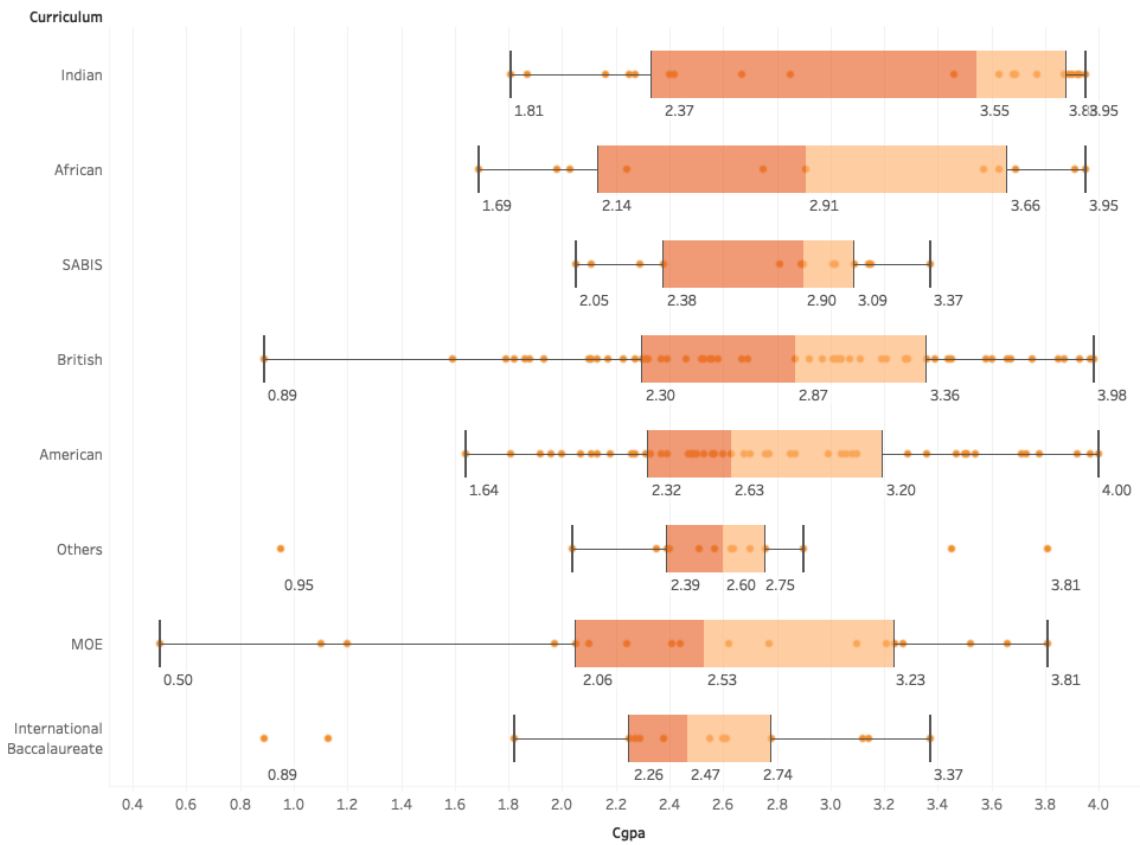


Figure 4. Box plot distribution of records

Indian curriculum has the highest median GPA and has a left skewed distribution. This curriculum has the largest percentage of students with GPA over 3.5 (50% students scoring between 3.55 and 3.95).

African Curriculum which is a conglomeration of curricula from the African continent has the second highest median GPA of 2.91. The distribution of records is close to that of Indian curriculum with 50% students scoring 2.91 and 3.95.

SABIS curriculum has a lower variance and fewer outliers => higher consistency in GPAs with 50% students have a GPA between 2.9 and 3.37.

2.2. Correlation analysis

The relationship between different features or variables were done to identify correlations between the dependent variable and independent variables and also between the different independent variables.
Figure 5 shows the correlation matrix (left) and the Scatterplot matrix(right) of the independent variable, GPA, and all the predictors with the exception of Curriculum.

As Curriculum is a categorical variable, it isn't used in this method of correlation analysis.

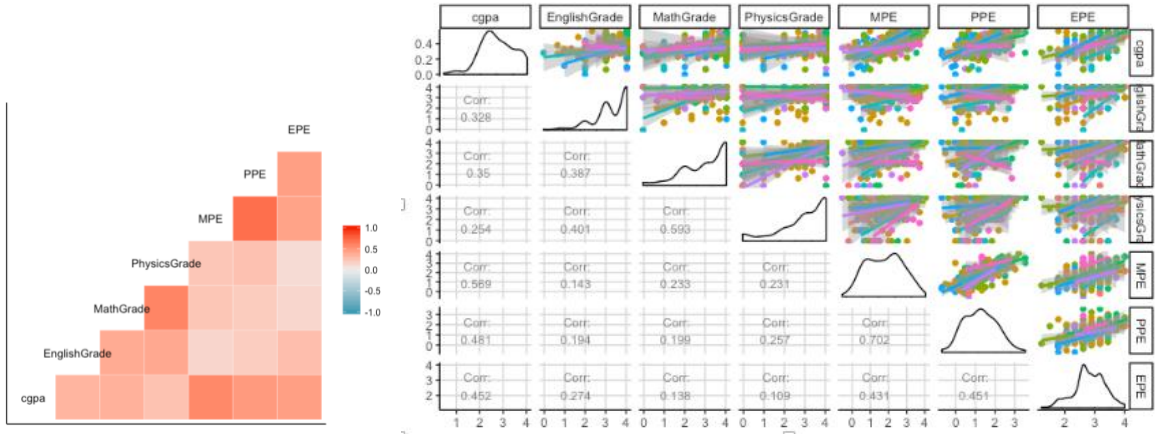


Figure 5. Correlation matrix and Scatterplot matrix

From the correlation matrix, moderate predictors for GPA seem to be MPE, followed by weaker correlations with other the placement exams, PPE and EPE; followed by even weaker correlations with EnglishGrade and MathGrade.
PhysicsGrade seem to have very low correlation with GPA.

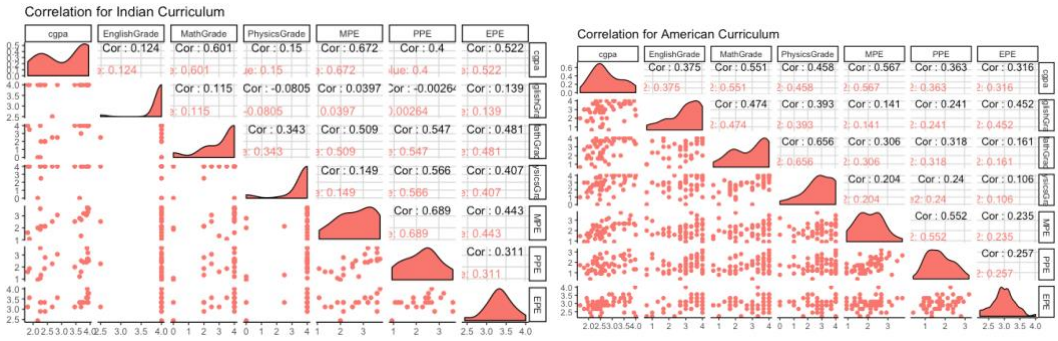
To confirm this, the scatterplot and the correlation coefficients can be used. All the predictors have a positive linear uphill relationship with each other and with the dependent variable, GPA.

GPA is moderately correlated with MPE (0.57), PPE (0.48) and EPE (0.45).

Correlation between Mathematics and Physics: Mathematics and Physics scores have stronger correlation indicating that a student's performance in Mathematics in High school is highly correlated (~0.6) to the student's performance Physics in high school. The same goes for Mathematics and Physics placement exams (0.7).

Correlation analysis by curriculum

For advanced analysis, correlation of the variables for each curriculum will be used as the diagnostic.
The correlation coefficients were compared for each curriculum and is given in Figure 6. Table 1 gives the curriculum-wise predictors with positive correlation to GPA.



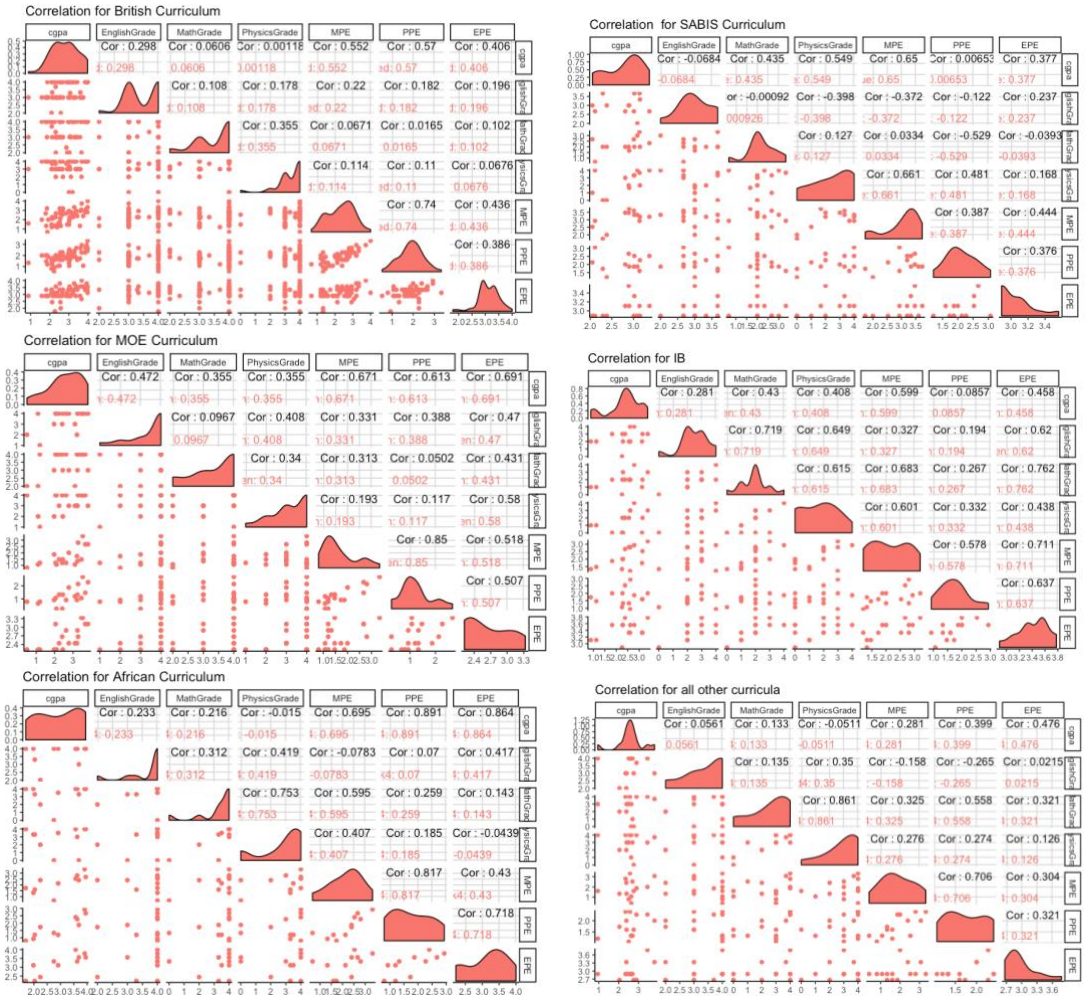


Figure 6. Correlation matrix by high school curriculum

All predictors with a coefficient above 0.6 can be considered as having a strong correlation to GPA and have been highlighted.

Table 4. Curriculum-wise predictors with strong correlation to GPA

INDIAN		**MOE**		**AFRICAN**		**BRITISH**		**AMERICAN**		**SABIS**		**IB**		**OTHERS**	
MPE	0.67	EPE	0.7	PPE	0.89	PPE	0.57	MPE	0.57	MPE	0.65	MPE	0.6	EPE	0.48
Math Grade	0.6	MPE	0.67	EPE	0.86	MPE	0.55	Math Grade	0.55	Physics Grade	0.55				
EPE	0.52	PPE	0.6	MPE	0.7										

Except for “Others”, all other curricula have MPE as a predictor with positive correlation to GPA. This is followed by EPE.
MathGrade and PPE have moderate correlation with GPA except for African curriculum where PPE is strongly correlated.
Curriculum with predictors with strongest correlation to GPA is African curriculum.
EnglishGrade shows no correlation with GPA except for MOE where there is a weak correlation of 0.47.

The correlation analysis of the whole data set and subsets by curriculum show that placement exam scores are a stronger indicator for undergraduate performance than scores obtained in high school. One reason could be that examinations conducted by different curricula are scored differently and hence cannot fully be standardized for comparison.

2.3. Subject proficiency analysis

The student performance in the 3 core subjects, i.e., English, Physics and Mathematics are the main criteria based on which undergraduate admissions take place.

As seen in the previous section, placement scores are better indicators of performance at university. The placement exams and proficiency test instruments are the same for all students irrespective of curriculum, hence, the results of these exams will be considered more comparative between students than the high school exam scores.

To study the relative difference and identify any subjects a curriculum is weak in, the statistical comparison is done using box plots.

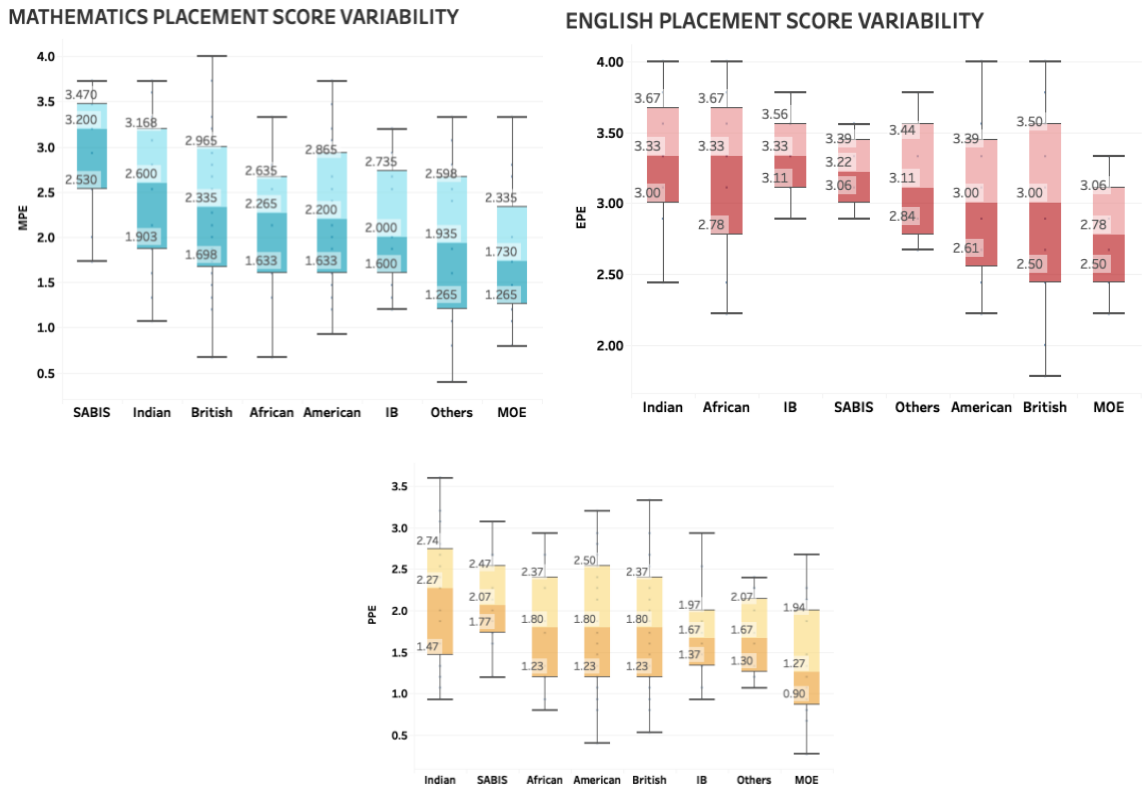


Figure 7. High School score distribution by curriculum

As seen in the box plots in Figure 7 of the 2 high school scores, Indian curriculum students have the highest scores in English and Physics and SABIS curriculum students have the highest scores in Mathematics.

The average scores from English proficiency exam, Mathematics placement exam and Physics placement exam of different curricula are given in Table 7. The variation of average scores is higher in Mathematics and Physics than in English.
In the case of Physics, the maximum average score is 2.26, much lower than that of English and Mathematics.

In order to assess a need for support in a certain subject, the percentage of students scoring lower than the average score for each exam were calculated as shown in Table 8.

A curriculum will be considered weak in a subject if more than 50% of their students score less than the average score in that subject.

Table 7. Average scores by Curricula

Curriculum	English proficiency	Mathematics placement	Physics placement
American	2.97	2.08	1.72
Indian	3.23	2.67	2.26
MOE	2.64	1.66	1.2
British	3.11	2.33	1.9
African	3.15	2.14	1.75
SABIS	3.04	3.03	2.1
IB	3.41	2.13	1.74
Others	2.96	1.86	1.75

Table 8. Percentage of students below the average score

Curriculum	English proficiency	Mathematics placement	Physics placement
American	52%	56.3%	69%
Indian	13.6%	36%	27%
MOE	77.7%	83.3%	83.3%
British	41.9%	40.3%	40.3%
African	27.2%	45.5%	54.5%
SABIS	53.8%	15.3%	23.0%
IB	7.1%	57.1%	57.1%
Others	78.5%	64.2%	57.1%

More than 50% American, MOE and Others curriculum students perform lower than average in English proficiency and placement exams.
Majority of the MOE curriculum students and more than half of the American curriculum students' scores below average in English proficiency and placement exams. This is also the case with "Others".

More than half of SABIS curriculum students score lower than average in English Proficiency exam, although majority of the SABIS students (over 75%) score over average in Mathematics and Physics placement exams.

On the other hand, over 90% IB students score above average in English proficiency tests but more than half of the students score lower than average in Mathematics and physics placement exams.

College readiness in terms of academic proficiency is lower for MOE curriculum and American curriculum who need to prepare better in all the 3 subjects Mathematics, Physics and English.

SABIS curriculum prepare students well in Mathematics and Physics but not well in English.

The opposite is applicable to IB, where students will need support in Mathematics and Physics.

2.4. GPA evolution with years of study

71 student records with GPA data for first 3 years of study were analyzed for performance evolution trends. Diagnostic was done using the average or mean GPA for each curriculum for the first, second and third years of study.
A visual representation is given in Figure 8 and Table 8 gives the percentage change associated with it. The Inter Quantile Range was also used to study the evolution of variability between the first 3 years of study.

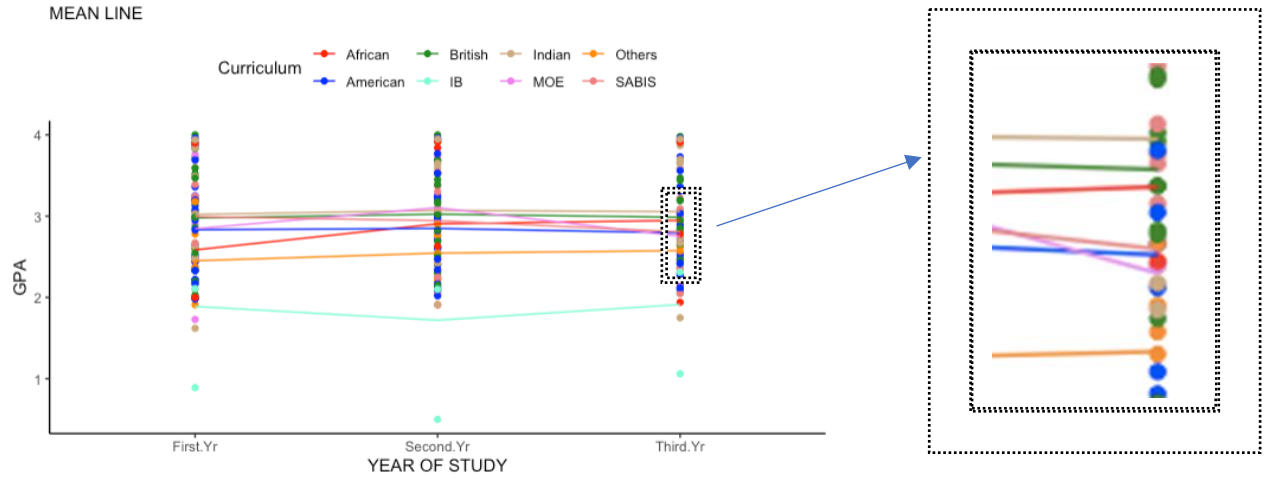


Figure 8. Performance evolution of mean scores between first and third year of study

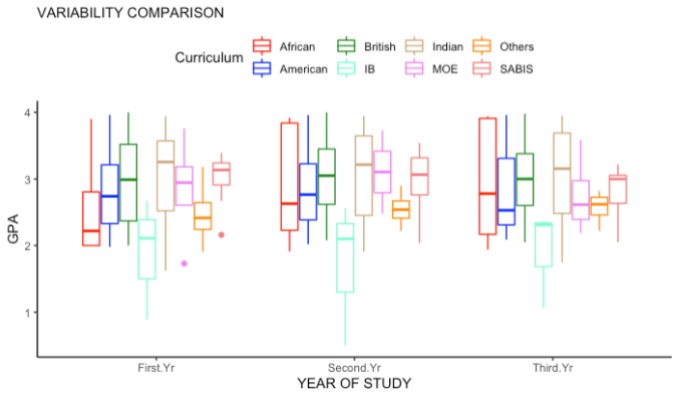


Figure 9. Evolution of IQR of GPA between first 3 years of study

Table 5. % change in performance

Curriculum	Change from 1st & 2nd year	Change from 2nd & 3rd year	Change from 1st and 3rd year
British	1.5%	-1.2%	0.2%
American	2.4%	-8.4%	-10.7%
Indian	1.6%	-0.4%	1.1%
MOE	9.1%	-11%	-3.2%
IB	8.9%	-2%	1%
SABIS	-1.7%	-4.7%	-6.3%
African	12.4%	1.4%	14%
Others	3.9%	1.1%	5%

Figure 8 shows African curriculum having the steepest mean line between the 1st and 3rd year of study. Figure 9 shows variability in GPA widening in the 2nd and 3rd year which means that high variance in GPAs among students. Hence, the steep increase could be attributed to large increase in GPAs of few students.
This could be an indication that students from African Curriculum are able to improve their performance especially in the second year, where the mean GPA sees an increase by 12.4%.

This is also evident in British and Indian curricula show consistency in GPA scores and variability in all the 3 years of study. This is also the case for “Others”.

In the 3rd year of study, Indian, British and African curriculum have mean GPAs very close to each other. The variability of GPAs though is much better for Indian and British curriculum whereas it is much larger for African curriculum.

Drop in performance is observed for American, SABIS and MOE curricula. The drop is most apparent from the 2nd to 3rdyear. It is interesting to note that the mean GPA of these 3 curricula are also close to each other by the third year.

It can also be noted in Figure 8 that the mean GPAs show an increase from the 1st to the 2nd year all curricula, with the exception of SABIS. Reasons for this could be higher motivation to

do well in University when joining or academic objectives in the 1st year are less challenging than those of the subsequent years of study.

On the other hand, the mean GPA drops from the 2nd to the 3rd year for all curricula, except for African and Others. The reason for this drop needs to be further studied with student interaction and/or surveys. As students start co-op in the 3rd year, any impact of the same on GPAs due to stress or time management issues should be studied. This is a very useful insight for Academic Student support departments and faculty.

This insight should be used to help students in the 2nd year of study with counselling and trainings such as time management training etc. to help them sustain their academic performance.
Students who attend preparatory or refreshment classes in the 1st year should also be given this support as there is an overlap between these 2 use cases with American and MOE curriculum students appearing in both.

2.5. Non-Cognitive data analysis

The survey for collecting non-cognitive skills in terms or Grit Scale and Self Control evoked a response from 67 students whose academic records were used for this project. These records were analyzed for their impact and correlations. The analysis and conclusions drawn will be treated as a pilot study as the sample population is small.

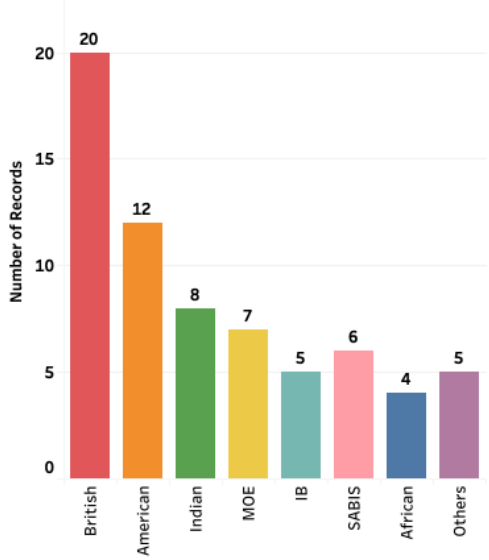


Figure 10. Distribution of records

Table 6. Non cognitive skills scales by curriculum

Curriculum	Avg. Grit Scale	Avg. Self Control Scale
American	2.4458	2.8600
IB	2.3980	2.5600
British	2.2920	2.6080
Indian	2.2688	2.5600
Others	2.2660	2.7840
MOE	2.2300	2.5829
African	2.1350	2.1800
SABIS	1.8900	2.4800

The distribution of these 67 records by the different curricula are shown in Figure 10. It can be noted that British curriculum student records are the largest in this set with 20 records. Due to low number of records, the results will not be conclusive for IB, African and Others.

The average Grit Scale and Self Control Scale for each curriculum is given in Table 6.

American curriculum students have the highest values for both these metrics of non-cognitive skills. This could be explained with the fact that irrespective of accreditation to NEASC, the American curriculum schools adopt the American educational culture of promoting self-esteem and self-expression among their students. As shown in Figure 11, this does not

translate to higher GPA. This gap between self-perception and academic performance among American curriculum students has been seen in US students in the results of Trends in International Mathematics and Science Study (TIMSS) and published by Sanandaji (2013).

The average Self Control Scale is the same for Indian and IB curricula.
Indian and British Curricula have GritScale and Self-Control Scale very close to each other, reinforcing their similarity.

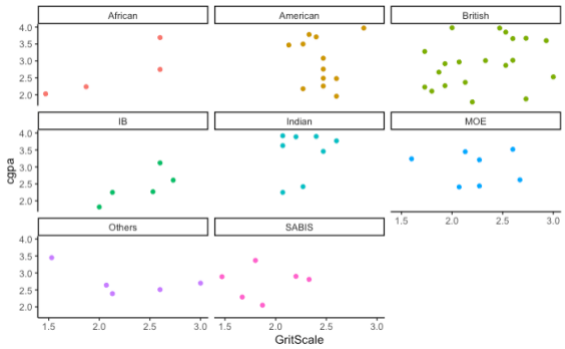


Figure 11. Grit Scale distribution by GPA

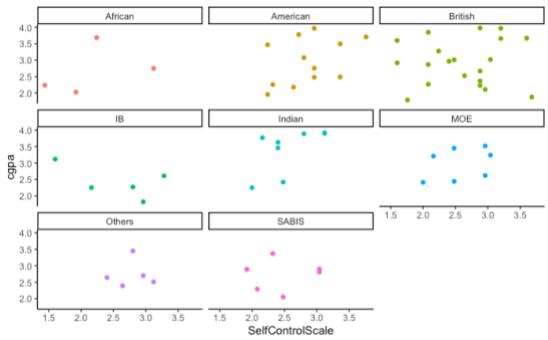


Figure 12. Self-Control Scale distribution by GPA

From Figure 11, the impact of grit scale on GPA is positive and visible for African, Indian, IB and SABIS curriculum where a higher Grit Scale indicates a higher GPA. The relationship could be termed close to linear for African and IB curricula as records with lower Grit Scale has lower GPA. From Figure 12, Indian curriculum student records show a positive relationship between Self Control Scale and GPA.

The mean or average of Grit Scale and Self Control Scale will be used to further analyze any relationship and is given in Figure 13.

If the average values for all the 67 student records are analyzed, it can be seen that the average self-control is higher than the average Grit scale irrespective of the academic performance of the students. This could mean that students feel they have better ability to control inhibitions and stay focused than the ability to persevere for a long-term goal. This belief could be due to self-discrepancy fueled by peer pressure.

The average Grit Scale is higher by 0.3 and Self Control Scale by 0.2 for students with a GPA of around 2.5 compared to those with a GPA of around 2.
Similarly, the average Grit Scale and Self Control Scale are higher by 0.15 for students with GPA of around 3.5 when compared to those with a GPA of around 2.

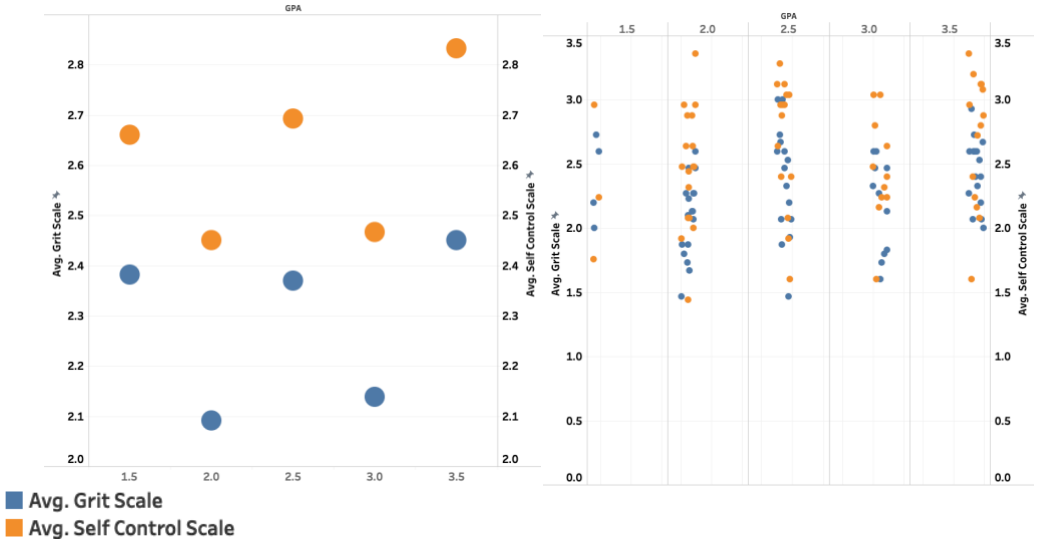


Figure 13: Average Grit Scale and Self Control Scale over GPA

The 2 metrics are exceptionally low for students with GPA around 3.0 as can be visualized from the plot on the right of Figure 12. This could mean that these students have lower perseverance compared to those scoring higher and lower than them. This could be due to the counterfactual thinking of students as proposed by Aronson et al (2007). In counterfactual thinking, students who are in 3.0 tend to compare themselves to the 3.5+ GPA students making them less confident about their own grit and self-control. It could also indicate disinterest in the topic of study for some students who are doing the course due to social pressures. Interviewing these students and counselling these students could improve their grit and self-control which could yield higher scores and better success in life for these students.

Grit Scale and Self-control are exceptionally high for student records with GPA around 1.5 which could be attributed to disinterest in the course of study.

Higher Grit and Self Control Scales indicate a better performance in college as seen for mean values. This correlation is also visible for the Indian curriculum students. The impact of Grit Scale on GPA is also visible for African, IB and SABIS curricula although as the sample population for these curricula are small, further confirmation using a larger population would be needed before using the metric in GPA prediction models.

American Curriculum students have the highest average Grit Scale and Self Control Scale which demonstrates that irrespective of accreditation to NEASC and following the US benchmark in academic, these students have been imbibed with the aptitude to persevere for long term goals and staying focused.

2.6. GPA Prediction Modeling

The objective of creating the machine learning model is to develop a model which can predict the future university GPA of a high school student from his high school curriculum scores and placement exam scores.

As the GPA is a continuous variable, prediction of GPA will require regression modeling.

Regression diagnostics

As one of the objectives of this project is to prediction GPA, the linearity of the relationship between the 7 predicator variables and outcome variable, GPA need to be studied.

The diagnostic plots given in Figure 14 give the following information:

- Variables have a Linear relationship as indicated by the almost horizontal line in Residual vs Fitted plot
- The residuals are normally distributed as can be seen from the QQ plot
- Residual spread is homogeneous in the Scale location plot indicating homoscedasticity.
- There aren't any influential datapoints that will influence the regression results as seen in the Residuals vs Leverage plot.

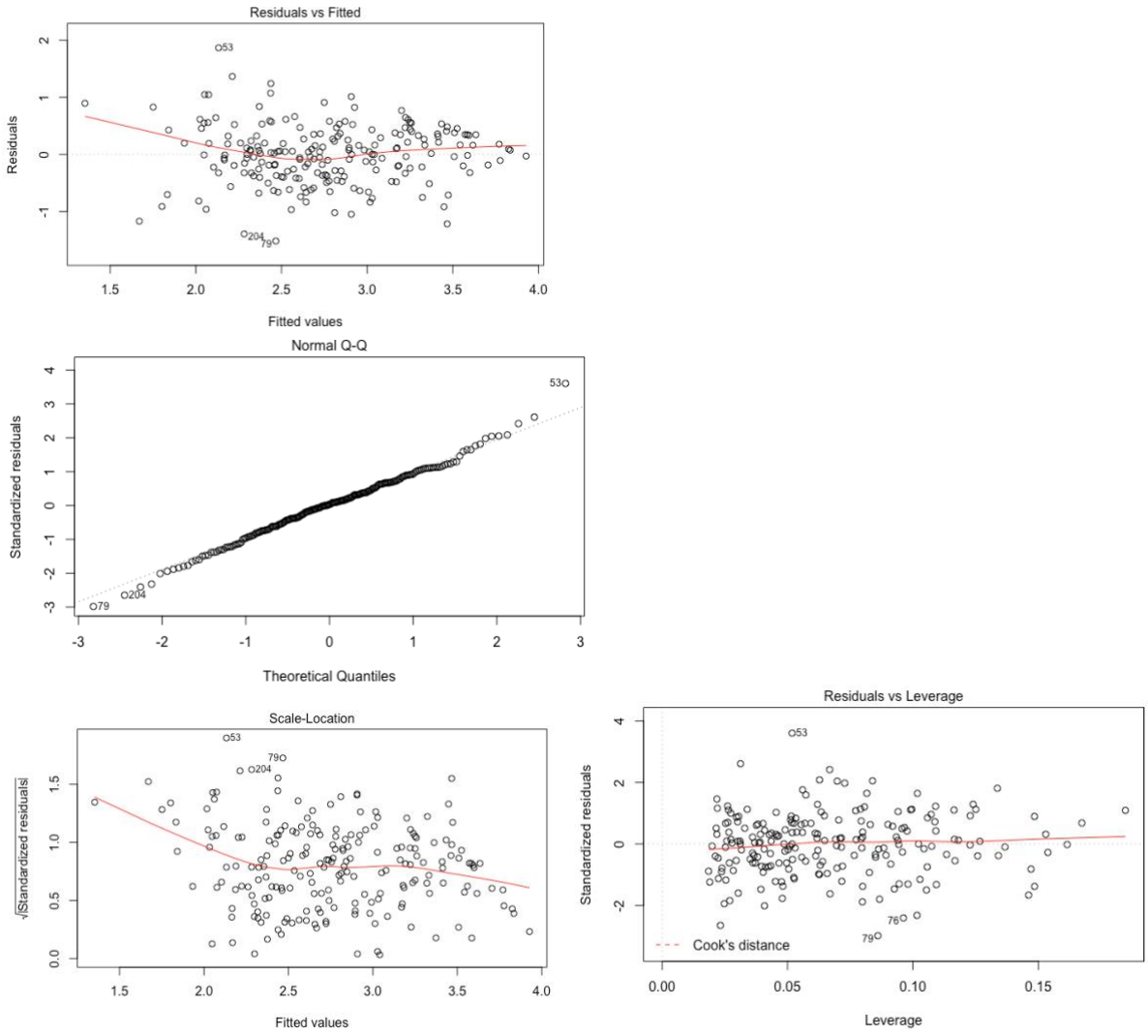


Figure 14. Regression diagnostic plots

Regression, hence, is applicable to this dataset for predictions and all records can be included due to the absence of influential ones.

Regression is used for prediction of GPA as GPA is a continuous numeric variable and not a discrete categorical variable.

Each predictive model was evaluated using the following four criteria that involved the use of either training or testing datasets:

R squared or Coefficient of determination: This metric indicates the percentage that a model can explain its output based on a training dataset.

The mathematical formula for r-squared is 1- $\frac{\sum(Actual\ values - Predicted\ values)^2}{\sum(Actual\ values - Mean\ values)^2}$
The higher the R-squared value, the better the model.

Root Mean Squared Error (RMSE): This metric indicates the absolute fit of the model and how close the actual data points are to the model's predicted values.
The mathematical formula for RMSE is $\sqrt{\frac{\sum(Actual\ values - Predicted\ values)^2}{Number\ of\ values}}$
The lower the RMSE, the better the model.

Data was split into training and test set in 80:20 proportion. Each of the regression models used were trained using training dataset consisting of 167 student records and tested using the test dataset consisting of 42 student records.

The training dataset and the testing dataset used for all regression models were kept the same.

Cross validation with 10 folds were used for training the models. To ensure that the samples remain the same for all models, the value of the random seed was kept fixed at 1 in the R program.

3 regression models were trained for testing prediction.

- Multiple linear regression model
- Decision Tree model using Anova method.
- Random Forest

For the decision tree, the complexity parameter was adjusted to prune the tree to avoid overfitting.

All the regression models were run on the training set using different set of predictors.
The predictor selection was done using the results of the correlation analysis described in section 7.2. Stepwise selection method was followed for all the 3 Prediction model types, starting with MPE. The different combinations of predictor or independent variables were used for training the models and r squared and RMSE compared.

Table 9 gives the best 3 r-squared value and RMSE values obtained for different set of predictors for each model.

Table 9. Regression models and metrics

Model#	Predictors	Multiple Linear Regression Model		Decision tree		Random Forest	
		R-squared	RMSE	r-squared	RMSE	r-squared	RMSE
#1	MPE, PPE, EPE	0.38	0.51	0.27	0.57	0.38	0.52
#2	MPE, PPE, EPE, MathGrade	0.40	0.50	0.34	0.54	0.46	0.48
#3	All predictors	0.43	0.49	0.33	0.54	0.44	0.49

As seen in Table 9, decision trees do not perform as well as Random Forest and Multiple linear regression models. Random Forest gives the largest r-squared value of 0.46 and lowest RMSE of 0.48 when the predictors used are all placements scores and High School Mathematics score.

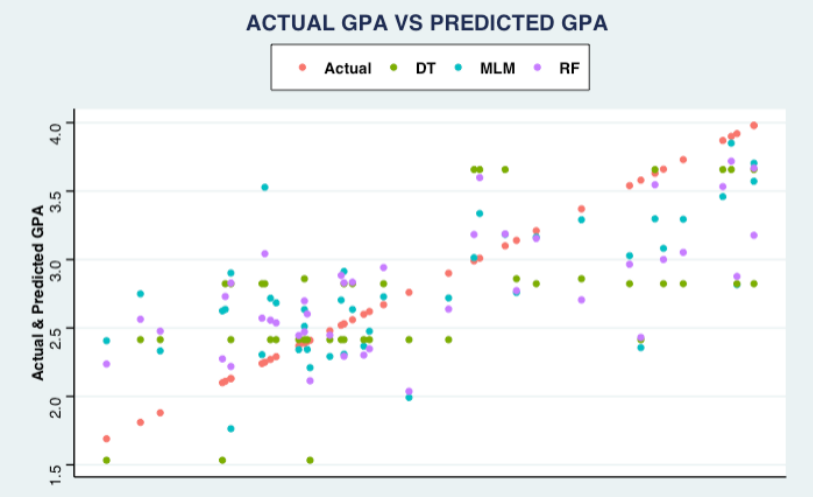


Figure 15. Actual Vs Predicted GPA plot

The Actual Vs Predicted values are shown in Figure 15. The actual GPAs are given in red and have been sorted in the ascending order. The datapoints for the different algorithms in Figure 15 correspond to Model#3 for Multiple Linear Regression, Model#2 for Decision Tree and Model#2 for Random Forest.

Random Forests have performed the best among all the different models trained for the academic data. The stepwise feature selection helped in identifying the best performance of the model with the 3 placement scores, EPE, MPE and PPE along with High School Mathematics score as the best combination for predicting GPA.

It needs to be noted that the predictive models developed were based on the data collected from RIT Dubai. More importantly, only the academic data was used in the model predictions due to the limited sample of the non-cognitive data available at the time of this project.

As pointed out in various studies in the past (Noftle et al, 2007, Horn et al, 2001 and West et al, 2016), non-cognitive skills influence academic performance. Hence one can expect better the prediction and higher r squared if GritScale and SelfControlScale can be added to the prediction models.

3. CONCLUSIONS

Different aspects of the relationship between High School curriculum and undergraduate performance were studied. Similarities and differences between curricula were also studied.

The study found that students from different curricula perform differently at RIT Dubai. This was validated with differences in their GPA variabilities. According to the population studied, Indian curriculum students, followed by African curriculum students, perform better than other curricula as the largest proportion of students scoring very high GPAs, students.
Indian, African and British curriculum students demonstrate a consistent and stable performance over the first 3 years of study. Their placement scores also demonstrate better college readiness than other curricula as they have fewer students scoring lower than the average scores. The non-cognitive scales are also comparable although African curriculum students scored lower than Indian and British curriculum students.

American, SABIS and MOE curricula show drop in performance especially in the 3rd year of study. The reason for this needs to be studied with more records and student interviews, but one explanation could be the start of co-op causing time management issues. MOE and American curriculum students also are less college ready than others as found in the placement score analysis. This finding could be used by Academic student support at RIT and High Schools to support the students with Time management trainings and additional supported learning in Mathematics and Physics. There is a gap between the self-perception and reality in performance in American curriculum students as seen from the non-cognitive skill analysis. This needs to be taken into account while training these students.

However, it needs to be noted that the study was conducted on a population of 213 students for academic performance, 67 students for non-cognitive skill analysis and 71 students for performance evolution analysis. Hence the results could be biased according to the performance of the population used for the study. The findings cannot be conclusive especially for the curricula with fewer records like the curricula joined together as “Others” and “African”.

This project hence should be considered as a pilot study done on the topic. Adding more data, in terms of student records at RIT as well as student records from other universities will aid in making insights which are applicable to a wider population.