

Student Profile Evaluator

Submitted in partial fulfillment of the requirements
of the degree of

Bachelor of Engineering

by

Krupa Shah (SAP ID:60001170030)

Lalita Takle (SAP ID:60001170032)

Project Guide:

Prof. **Mayur Parulekar**



Electronics Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai University

2020-2021

Certificate

This is to certify that the project entitled “**Student Profile Evaluator**” is a bonafide work of “**Krupa Shah and Lalita Takle**” (SAP ID.**60001170030, 60001170032**) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of “**Bachelor Of Engineering**” in Electronics Engineering.

Internal Guide

Prof. **Mayur Parulekar**

Internal Examiner

Head of Department

Dr. Prasad S. Joshi

External Examiner

Principal

Dr. Hari Vasudevan

Project Report Approval for B. E.

This project report entitled (*Student Profile Evaluator*) by (**Krupa Shah and Lalita Takle**) is approved for the degree of **Electronics Engineering**.

Examiners

1.-----

2.-----

Date:

Place: Vile Parle(west), Mumbai.

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

Krupa Shah- 60001170030

Lalita Takle-60001170032

(Name of student and SAP ID)

Date:

Abstract

Every year thousands of engineering students do not get placed or are also not able to apply for masters due to low GPAs. This can be avoided by providing special attention and assistance to such weak students well in advance, so that they do not suffer in future. The partial aim of the project thus enables us to predict the final Cumulative Grade Point Average (CGPA), taking into consideration the grades for first 2 to 3 years of study.

This project also aims in proving a full profile evaluation of a student and thus giving out suggestions, in order to improve his/ her profile. This can be useful for everyone, including those who want to pursue higher studies and also those who wish to take up jobs. The evaluator will take various inputs from the user such as GPA, internship experience, courses completed, skills, interests etc. Taking into consideration all these factors, it will give out specific programs and courses which the user might be interested in. On the other hand, it will also have a feature through which one can enter the targeted program and the tool will provide suggestions and various tips in order to be eligible for that particular program.

Acknowledgments

The making of project involves teamwork, vision, patience and perseverance on the part of group members. But a team can achieve a greater height only by proper guidance from faculty members and friends. Hence, we would like to express our gratitude to all those who in instrumental during the course of developing of the project.

First, we would like to thank our internal guide **Prof. Mayur Parulekar** for his valuable inputs and timely guidance. She always had very practical suggestions and sound knowledge.

We would like to thank our **Laboratory-Assistants** and **Laboratory-Attendants** for the help and co-operation offered by them along every stage of the project.

Lastly, we would like to thank our head of department, **Dr. Prasad S. Joshi** for extending his support towards the completion of this project and for the timely encouragement.

(Signature)

Contents			
CHAPTER NO.	SECTION NO.	TOPIC	PAGE NO.
		Certificate	ii
		Project Report Approval for B. E.	iii
		Declaration	iv
		Abstract	v
		Acknowledgments	vi
1.	1.1 1.2	Introduction Preface Plan of Action	1 2 2
2.	2.1 2.2	Review of Literature Review for Predictive Analysis Review for Profile Evaluator	4 5 13
3.	3.1 3.2	System Model/Architecture Components required Data and Information required	15 16 17
4.	4.1 4.2 4.3	System Implementation Data Pre-processing Training of Model Analysis of Results	20 21 22 34
5		Conclusions & Scope	41
6		REFERENCES & BIBLIOGRAPGHY	43

Chapter 1

Introduction

1.1. Preface

The performance of students during their studies is observed in many ways. Thus, both teachers and students have major problems in monitoring educational results and taking the necessary steps to avoid academic failures. Early predictions of the final CGPA can therefore resolve the issue and help teachers identify at risk students and also give students a strong belief in their studies.

Choosing the right university, or course plays huge importance in shaping the professional career of a student. Such decisions are to be made at an early age. Student Profile Evaluator is a tool, which will aid the students in taking crucial career decisions and making difficult choices. The tool will take various factors of the students' profile as input and based on the training data analysis; it will give out the chances of that student getting admit in the particular university. It will also have an option of working towards your dream university. When a student enters his/ her dream university and program, the tool will provide suggestions to improve his/her chances of getting admitted in that particular program. Considering all the profile details, it will have a wide variety of editable templates which can be used by the students for making their personalized SOP and Resumes. This can be very useful for those who dream of a particular university since childhood. They can take necessary efforts towards the fulfilment of their dreams well beforehand. May it be the CGPA requirement, or competitive exams scores such as GRE, TOEFL, IELTS. They will have their goal set, and will work rigorously towards it.

1.2. Plan of Action

We plan on using Microsoft Azure for Predictions and Tableau for all the main Data Analysis. Various Classification and Regression Models will be used to Train the data in Microsoft Azure ML Studio. We are currently using the data of students from Dwarkadas J. Sanghvi College of Engineering for the training of the model. For the prediction of CGPA, we

have collected the Results data of students from the Electronics department, since the year 2014. Out of these, all batches till 2016 are graduated and hence can be used to test the model. We will consider their GPA for the first two years and then will predict their final CGPA. And the accuracy of the model can be found out by comparing the predicted results with the actual values.

For the Profile Evaluator, we are collecting data from various sources such as Coursera. The courses which the student took on the online platform Coursera will help in categorizing the students on the basis of their interests and thus recommending programs considering their areas of interest. GRE/ GMAT and TOEFL/IELTS scores of the alumni are also being collected for this purpose. Various other details like the No. of Internships the students have completed, TOEFL/IELTS writing section marks, Final CGPA are being considered during the Analysis. The data missing for some alumni will be collected by personally interacting with the students.

Chapter 2

Review of Literature

2.1. Review for Predictive Analysis

[1] stated that the predictive prediction accuracy used by classification methods grouped in algorithms for predicting the performance of students for 2002-2015 is shown in meta-analysis based on highly precision prediction methods and essential factors that can impact students' academic achievements. It was quoted that "Neural Network has the highest prediction accuracy by (98%) followed by Decision tree, Support vector machine and K- nearest neighbour having same accuracy, Naive Bayes". The attributes are hybridization of two features, which are internal and external assessments, external being of highest accuracy which is the marks obtained in final examination, plays an important role in predicting students' performance. The maximum error of prediction is less than (10%) in neural networks along with the ability to capture nonlinear relationships easily. Lastly, the method that has lowest prediction accuracy is Naive Bayes by (76%).

The variables used are CGPA, student demographic, high school background, scholarship, social network interaction. Students who were admitted would accept this as a binary problem, metrics like precision, recall, F-measure and area under the receiver operator curve were used to evaluate the performance of these algorithms. To prepare them for entry into machine learning algorithms, categorical variables required special treatment. In addition to the accuracy of precision, recalling, F-measure and AUC or ROC score, a cross-validated accuracy score as well as a cross-validated AUC score, the problem was resolved by analysing the performance of our machine learning algorithms. Inferring how close its training and testing results are is the logistics regression which is the best algorithm for our problem. A measure of whether two categorical values for a group of individuals classified according to both variables are independent investigates the dispersion of numbers. For this test the null hypothesis assumes that the two categorical variables are separate. The alternative hypothesis is that the functionality is correlated with the "Great Commit Indicator."

[2] used a feature selection method to minimize the number of variables in each model. The Naïve Bayes Classifier along with an Ensemble model with a sequence of models such as SVM, K-nearest neighbours and Naïve Bayes classifier were used for best accuracy. Six distinct

predictive modelling techniques using only academic factors such as grades or scores presented in this paper are:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Multi-Layer Perceptron (MLP)
- Naïve Bayes Classifier
- K-nearest Neighbour

The data was collected from the Spring 2013 and Spring 2014 semesters of first-year engineering (FYE) students from the large Midwestern U.S University. Around 3000 students enrolled during these years for the FYE hence this generated a large dataset for the research. The data from Spring 2013 was distributed randomly into three datasets: 50% for training, 25% for comparing different techniques and tuning, 25% for secondary testing. And finally, data from Spring 2014 was used for testing. The same dataset was used for all the six classifiers and the individual accuracy, F-1 score was calculated for each classifier technique. The table below shows the results and accuracy of all the classifiers:

Method	Log reg	KNN	MLP	DT	SVM	NBC
$F_{1,S}$	0.56	0.43	0.50	0.46	0.53	0.59
Accuracy	0.926	0.949	0.931	0.923	0.872	0.869
Accuracy-Pass	0.953	0.997	0.967	0.961	0.884	0.870
Accuracy-Fail	0.586	0.345	0.483	0.448	0.724	0.862
True Negative	344	360	349	347	319	314
False Positive	17	1	12	14	42	47
False Negative	12	19	15	16	8	4
True Positive	17	10	14	13	21	25

Fig 2.1. Metrics for models [2]

Then an Ensemble model was created. Towards the end, it was concluded that Naïve Bayes and the ensemble model with three models (NBC, Support Vector Machine, and K-Nearest Neighbour gave the most accurate results.

[3] shows the techniques of using the databases to predict and classify students on the basis of academic performance measured using Cumulative Grade point average grades

(CGPA). The prediction aids the university to step in and help the low performing students well beforehand. Most commonly used classifiers along with Neuro-Fuzzy classification are used here. The dataset contains electrical engineering students studying at a Malaysian public university for intakes 2005, 2006, and 2007. Complete numeric data set of 391 students collected from the university database system was used. This paper proposes a Neuro-Fuzzy classification method to predict and classify the electrical engineering students' future academic performance based on their past academic performance, which is then categorized or graded into several bands ranging from poor performing students to outstanding students. This prediction can result in helping the low performing students with some extra tutorials before the future examinations and thus enabling them to score good in the coming exams. Adaptive Network based Fuzzy Inference System (ANFIS) was used as a Data Mining tool. The conclusion stated that the classifiers are needed before selecting a system to decide on various information. The tool was developed in order to help the academically lacking students and thus motivate students to achieve good grades in the final examination.

[4] studied student performance approaches using machine learning and obtained useful results. The main objective of the prediction system was to predict the grade of the student taking various factors as input. Many of the commonly found prediction systems the major importance is given to the machine learning algorithms used, but the data on which these algorithms run is of equal importance. So, having data with all the required features that affect students' performance is crucial. Various factors other than the grades, which includes demographic factors, socio-economic factors affect the academic performance of a student.

School reports and questionnaires compiled by [5] were used for the data for two different subjects namely Mathematics and Portuguese language . The dataset consisted of 33 attributes with 678 instances and was taken from the UCI repository. The 33 attributes had a significant direct or indirect impact on the performance and the features were classified using Support Vector Machine (SVM), Decision Tree and Naïve Bayes. Accuracy was measured using the following metrics:

- True Positive rate
- False Positive Rate

- Precision
- Recall
- F Measure

True Positives (TP) refer to the data instances that were properly labelled. False Positives (FP) are incorrectly labelled as positive. Given below are the formulae for precision, recall and F-score (F measure):

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{P}$$

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

85% of the data is properly classified by the SVM. The Decision Tree classes approximately 83 percent with accuracy. Naïve Bayes obtained accuracy of 81.5%. Classification results have been effectively strengthened by [4] Data Pre-Processing (DPP) steps such as discretizing, class balancing and deletion of invalid entries. For class balance, an algorithm for Synthetic Minority Oversampling (SMOTE) was used. The [4] system specifically aimed at making precise predictions for weaker students in order to be cared for well ahead of time. The [4] framework particularly intended to make precise predictions for weaker students in order to be cared for well ahead of time. [4] DPP techniques have been implemented for student data, usually highly unbalanced by over sampling techniques as well as by sampling techniques. The [4] over-sampling improved the classifier's accuracy compared to the sample. The DPP methods have therefore improved the efficiency and also reduced error of all classifiers.

[6] analysed the undergraduate students' performance using EDM. Mainly, two aspects were studied i.e. predicting the final grades at the end of a four-year study and studying progressions and combining them with prediction results. These results aid in warning low performing students in advance. Several Classifiers were built to predict the performance using the admission marks from high school certificates and final marks of first and second-year courses at university. This led to derivation of courses using these classifiers which can serve as powerful measures of student success in a degree programme. Thus, helping at-risk students

or to further encourage the promising students. Students are divided into groups using Clustering techniques. The study[6] answers the following three questions: Question 1: “Can we predict students’ performance with a reasonable accuracy at an early stage of the degree programme using marks only?” Question 2: “Can we identify courses that can serve as indicators of a good or low performance at the end of the degree?” Question 3: “Can we identify typical progressions of students’ performance during their studies and relate them with the indicator courses?”

This was the workflow of the research mentioned by [6]:

- (1) Using classifiers for predicting graduation performance
- (2) Extracting implementable performance predictor for graduate students
- (3) Scrutinising the progress of students' academic performance over time
- (4) Linking the results of prediction and progression.

The following conclusions were drawn out from this research which answer the questions mentioned to a certain extent in the paper [6]:

- The findings show that the graduation performance in a four-year university program can only be estimated with a fair precision using pre-university marks and marks of first- and second-year courses.
- The second question is about the illustration of courses which can be used as indicators of good or poor graduate performance. Four courses that can serve as such indicators were put to the test by decision trees
- The third question is how the academic achievement of students develops during the four-year degree. Astonishingly, students tend to have the same marks every year: low, medium or high marks in all classes. Thus, this paper uses the Educational Data Mining techniques to predict the students’ performance.

[7] presented a new approach to predict the students' academic performance using Regression. The main aim was to develop a set of multivariate linear regression models to predict the academic performance of Engineering students based on the category of CGPA. The final outcome will be the students end semester exam grades using the inputs provided. The following variables were considered for the regression:

- Usage of Internet for the academic/education purpose
- Usage of Internet for the Entertainment purpose
- Usage of Internet for the Communication purpose
- Active duration in social media networks
- Usage of Internet before the end semester exam

Cumulative GPA where the variable from to represents the student's behaviour on the usage of the internet for various activities and y represents the outcome of the model. These factors were used to observe patterns using the internet used by various undergraduate engineering students to predict their final CGPA. The dataset contained academic performance of 150 students from undergraduate engineering disciplines collected through questionnaires.

The workflow of the research was as follows:

- Collection and understanding of data of engineering students' seven semesters named as S1, S2, S3, S4, S5, S6, and S7 and then analysing it.
- Splitting the data into train and test data.
- Applying multivariate regression on the training dataset and then running it to predict on the test dataset.
- The training dataset had 3 categories of students based on the performance

Thus, a different approach of using regression for the prediction was found here.

[8] helps in understanding various classification techniques of each of the 5 data mining models (k-nearest neighbour, neural networks, decision trees, support vector machines, naive bayes) used efficiently on the basis of types of variable needed and shows us extent of proficiency of these techniques along with this it also helps us with some introduction of few open-source software such as WEKA, rapid miner, KNIME and SSDT. Here [8] took into consideration the educational model for an Engineering college where they are using these technique to better understanding the estimated Marks for final year students with classifying them into two distinctive parts which are inter related with 1 similar character and it uses decision making algorithm which gives them an accuracy of 60% and with this they make a proper use of artificial neural network to build their classifier model. They use two phases of classification procedure such as Development of a model for training and evaluating the model using testing data through which they get their highest accuracy with Multiplayer Perception Model and they also depict various models working which are used in educational data mining with the help of a graph depicting their accuracy with the given variables. [9] investigated the dataset that contains the GPA data for the first three academic years and the final CGPA of 1,841 students from 2002 to 2014 across various engineering streams from the Covenant University in Nigeria.

Six different data mining algorithms namely:

- Random Forest
- Tree Ensemble
- Logistic Regression
- Decision Tree
- PNN
- Naive Bayes

“89.15% were the highest accuracy in the logistic regression forecast, and 87.88% in the Tree Ensemble. The 3rd best accuracy was 87.85% for the Decision Tree predictor, and the Random Forest predictor had the 4th highest precision with 87.70% accuracy. The Naive Bayes predictor was 86.438% accurate, and the PNN predictor was 85.89% less accurate”. Applications like MATLAB and KNIME (Konstanz Information Miner) were used. It was concluded that the grades in the First three years of the engineering course in Nigeria played a huge role in the final grades. Thus, [9] concluded that a maximum accuracy of 89.15% was achieved, and using regression models for performance validation, R² values of 0.955 and 0.957 were achieved using both linear and pure-quadratic based regression models.[10] has presented a novel hybrid algorithm, HLVQ to predict student academic performance and employability chances.

2.2.Review for Profile Evaluator

[11] developed a Student Placement Analyser using Machine Learning Algorithms and predicted the chances of a student getting placed. The prediction was made in five placement statuses namely: Dream Company, Core Company, Mass Recruiters, Not Eligible and Not Interested in Placements. Specifically, Decision Tree Algorithm was used for the prediction. The algorithm was implemented using the Scikit-Learn library in Python. To find the relationship among the variables, Logistic regression was used. Following variables were considered for the prediction process.

Variables	Description	Possible Values
Dept	Department	{AE, CHEM, CIVIL, CSE, ECE, EEE, EIE, MECHANICAL}
Gender	Gender of Student	{M,F}
Board	12 th Board of studies	{SB-HP, SB-GJ, SB-MH, ST-AP, ST-TN, SB-BI, ST-KARNATAKA, CBSE, ICSE, SB-RJ, ST-KL}
Location	Location where 12 th is completed	Pin code
12 th	Marks percentage in 12 th	Marks in %
NA	Number of	Integer

Variables	Description	Possible Values
	Standing Arrears	
AH	Arrear History	Integer
CGPA	Cumulative Grade Point Average	Float value out of 10
COMP	Placement Status	{Dream Company, Core Company, Mass Recruiter/Common Company, Not Eligible, Not Interested, Not Placed}

Fig 2.2. Variables for prediction

And using Scikit-Learn on Python the following predictions were made.

	Mass Recruiter/ Common Company	Core Company	Dream Company	Not Eligible	Not Interested	Not Placed	Total
Girls	35	30	24	6	5	14	114
Boys	63	26	27	23	9	27	175
Total	98	56	51	29	14	41	289

Fig 2.3. Prediction

Data mining was used to provide recommendations for various university courses in [12]. The recommendation system architecture is stated below:

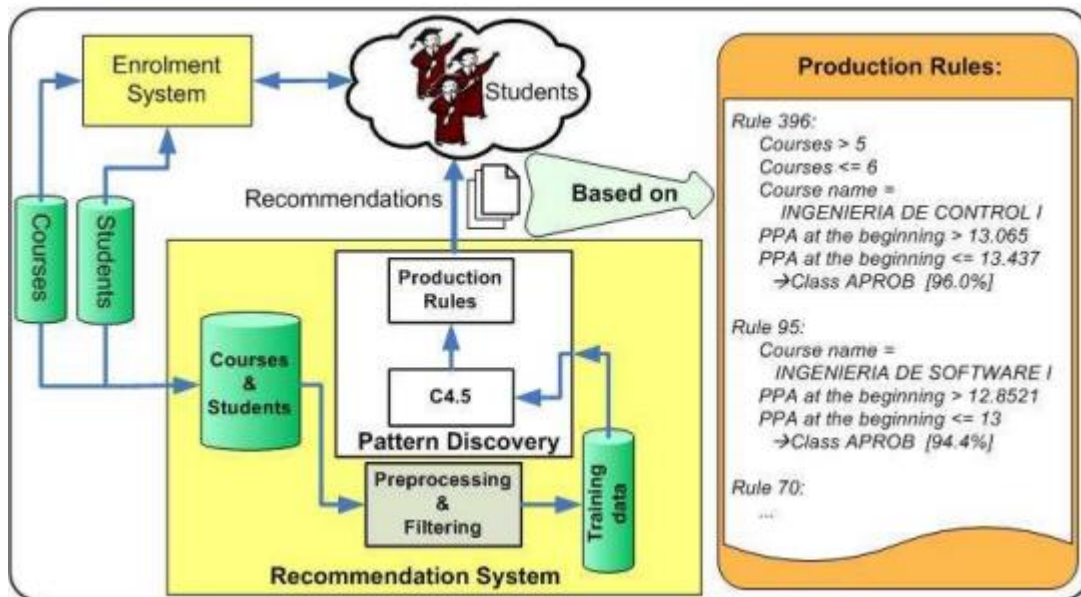


Fig 2.4 Recommendation system Architecture

The data used was of the students from the School of Systems Engineering, enrolled through the years 2002 to 2008. The steps in the process were the pre-processing of data, Pattern extraction and evaluation and finally the analysis of the results.

Chapter 3

System Model /Architecture

3.1. Components required

- a. **Microsoft Excel:** Microsoft Excel has the fundamental characteristics of every table using a cell grid in numerical letter-named columns and rows to manipulate data such as arithmetic operations. It can also present data as line graphs, histograms and maps with a restricted 3D graphics display. It enables data to be segregated to display its reliance on different variables (by using the pivot tables and scenario manager) from different perspectives. The programming element Visual Basic for Applications enables users, for example for the resolution of differential equations of mathematical physics using a wide range of computational methods and then report the results back to the spreadsheet. It also has a number of interactive features that allow user interfaces to conceal the table from the user, so that the table presents itself as a so-called app, through a custom user-designed interface, e.g. a stock analyser, or in general as a design tool which asks the user questions and offers responses and reports.

- b. **Tableau Desktop and Tableau Prep Builder:** In the business intelligence sector, Tableau is the most powerful and fastest growing data visualisation tool. It helps to simplify raw data into a format that is simple to understand. The data analysis with the tableau is very simple and the visualisations are provided as dashboards and worksheets. The data generated with the Tableau can be understood at any level of an organisation by professionals. It can also build a personalised dashboard from a non-technical user. The best characteristics are:
 - Data Blending
 - Real time analysis
 - Collaboration of data

Tableau is outstanding because it does not require any technical or programming abilities to function. The tool has attracted interest from all sectors, such as businesses, investigators, various industries and so on.

Tableau Prep Builder is one of the best tools available for Data Cleaning and Manipulation. It provides a complete picture of the data.

c. **Microsoft Azure ML Studio (Classic):** ML Studio (classic) has a Drag and drop interface and has Scalable (10-GB training data limit) and offers CPU support only. Machine Learning Studio (classic) provides an interactive, visual workspace to easily build, test, and iterate on a predictive analysis model. Datasets and analysis modules are to be dragged-and-dropped onto an interactive canvas, connecting them together to form an experiment, which is run in Machine Learning Studio (classic). To iterate on model design, edit the experiment, save a copy if desired, and run it again. It can also convert a training experiment to a predictive experiment, and then publish it as a web service so that the model can be accessed by others. There is no programming required, visually connecting datasets and modules to construct your predictive analysis model.

3.2. Datasets and Information required:

I. CGPA Dataset: Features of Dataset

- Results of Electronics department students from batch 2014 to 2019
- Admission details of 2014 to 2019
- Coursera Enrollments and completion data.
- AMCAT scores
- Health
- Travelling Time
- Absences
- Higher Education plans
- Interest in Electronics
- No. of KT

- HSC and CET marks

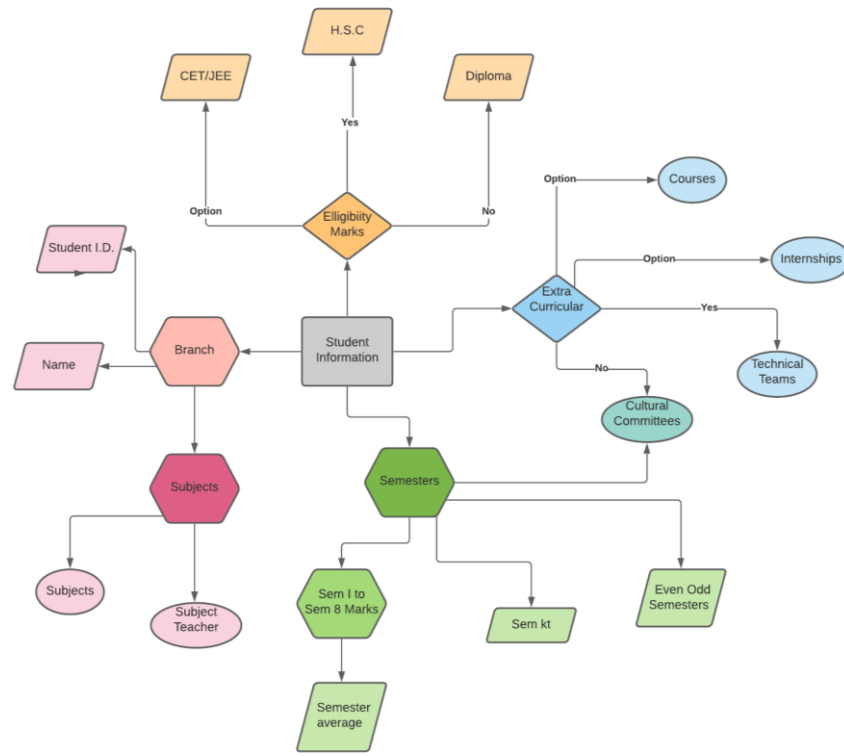


Fig 3.1. CGPA prediction Flowchart

In this model we are taking into consideration the change in the trends of students' marks starting from Semester 1 to Semester 8 of previous year students and building a Predictive model for the same. We are taking into consideration firstly the students previous scores from their H.S.C/ CET/JEE and Diploma which would help us understand the pre degree status of a student. After which we classify them in Branches and evaluate them based on their subjects, subject teacher and their topic of knowledge. We also take in the Semester marks from each and every semester along with their backlogs. We do not take into consideration the students who are Dropouts. We also see the changes happening due to their Extracurricular activities and the number of Semesters that affect the student performance.

II. Masters Dataset:

Features of the Dataset:

- University Studying at
- GRE and TOEFL/ IELTS scores
- SOP and LOR rating
- Early applicant?
- No. of Internships
- No. of Courses

In the ER Diagram below, as we can see our focal point is Student Permit wherein, we first take into consideration their GRE/GMAT/IELTS/TOEFL scores and the particular University that the student selects for his higher studies along with the Universities that he applied and the Permits and Rejects. This can also be correlated with the student's Extra-curricular activities based on their internships, Courses in the field along with the Technical and Non-Technical activities that the student took up in their college. We also take into consideration their semester average with branch, subjects in correlation with the Subjects that they take up in their Masters Programme. As we also know growing issues with Visas of different countries and after COVID effects. And we are strictly taking into consideration the students who have no GAP between their undergraduate education and Masters/MBA.

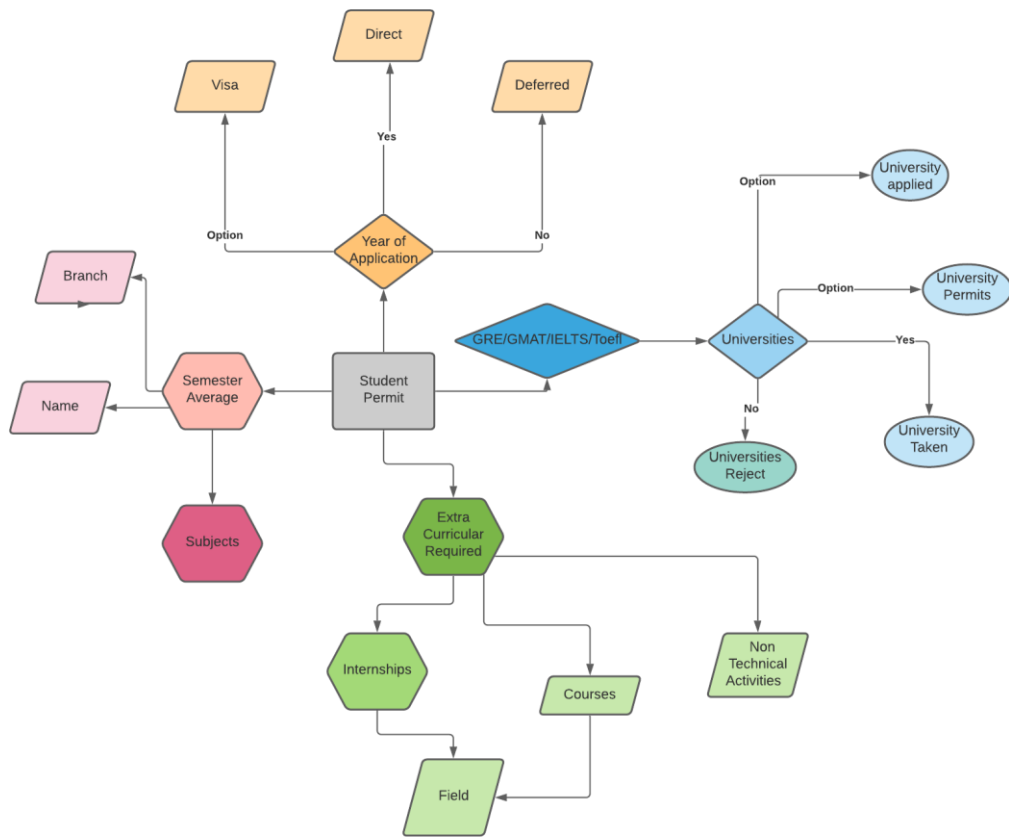


Fig 3.2 Flowchart for Admission Prediction

Chapter 4

System Implementation

4.1. Data Cleaning

- For the features with no information collected, we tried to assume certain values using some correlation with the existing available data.
- Each student either appeared for TOEFL or IELTS exam and hence we scaled TOEFL marks to IELTS band to avoid confusion. This scaling was done by referring to ETS official website. The table below represents it.

TOEFL Score	IELTS Band
0-31	0-4
32-34	4.5
35-45	5
46-59	5.5
60-78	6
79-93	6.5
94-101	7
102-109	7.5
110-114	8
115-117	8.5
118-120	9

Fig 4.1. Scaling TOEFL and IELTS

- Tableau Prep Builder was used to perform Joins and Unions on the dataset. The Entry (Admissions Data), Results Data and the Higher Studies Data was joined and its schematic is shown below.



Fig 4.2. Tableau Prep Builder Flowchart

4.2. Data Pre-processing

For pre-processing of data, we started to visualize the data on Tableau, to find insights of the data.

4.2.1. Admissions data for the 2017-2021 batch.

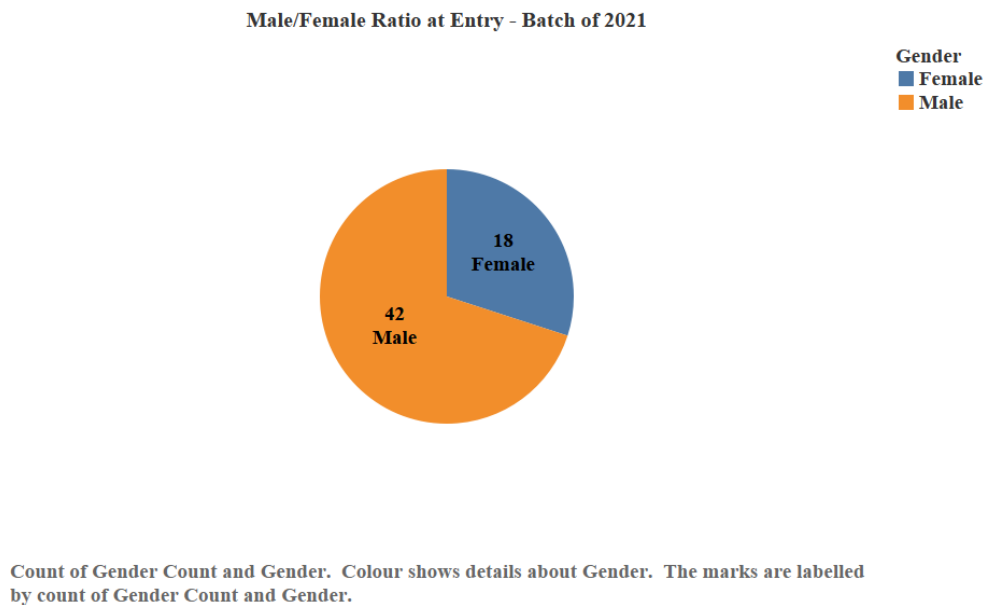
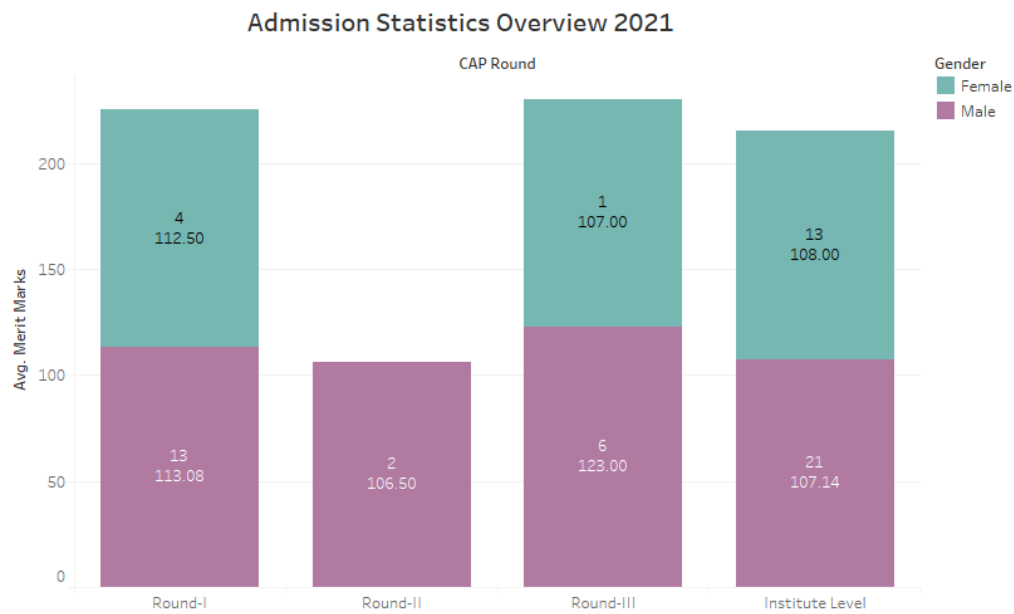


Fig 4.3. Male to Female Ratio 2021

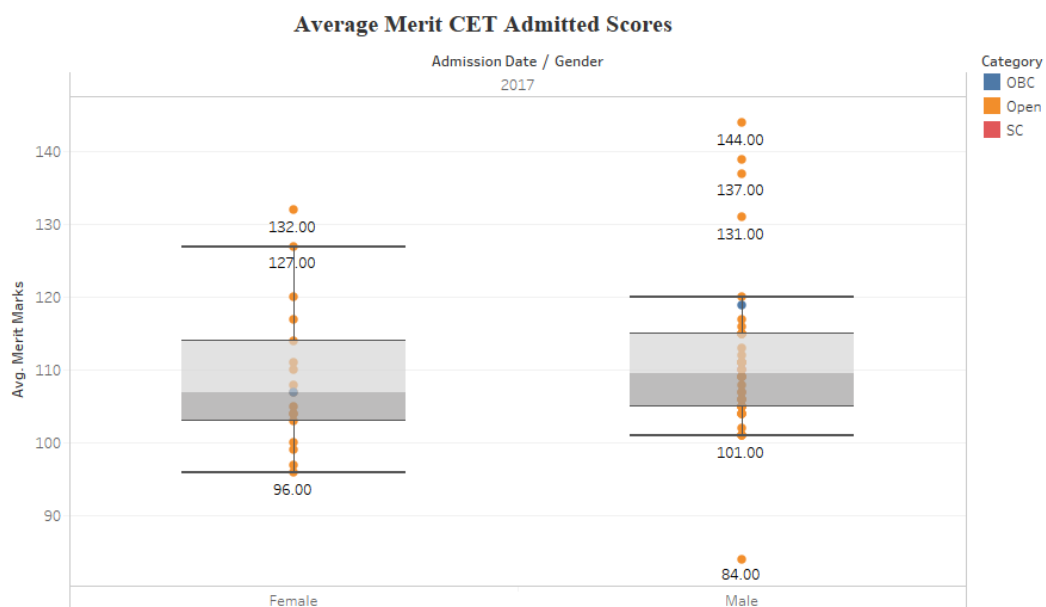
Male to Female ratio could be calculated based on this. This showed that in the 2017 admissions for the Electronics Department, more males were admitted in comparison to the number of females.



Average of Merit Marks for each CAP Round. Colour shows details about Gender. The marks are labelled by distinct count of Application ID and average of Merit Marks.

Fig 4.4. Admission Stats

The above graph shows that, maximum students were admitted through the Institutional Rounds. 17, 2, and 7 students were admitted in the Cap Rounds I, II and III respectively whereas 34 students were admitted through the Institutional Quota. It is observed that the males admitted through Cap rounds had a greater mean of CET merit marks while in the case of Institutional Quota, Females had a greater average.



Average of Merit Marks for each Gender broken down by Admission Date Year. Colour shows details about Category. Details are shown for Candidate Name.

Fig 4.5. Average CET scores

Top 3 highest marks in CET were observed in the Male category. The female marks ranged from 96 to 132 while those of males had a wider range from 84 to 144. The color shows different categories like OBC, Open and SC.

4.2.2. Results Analysis for 2012-2013 batch.

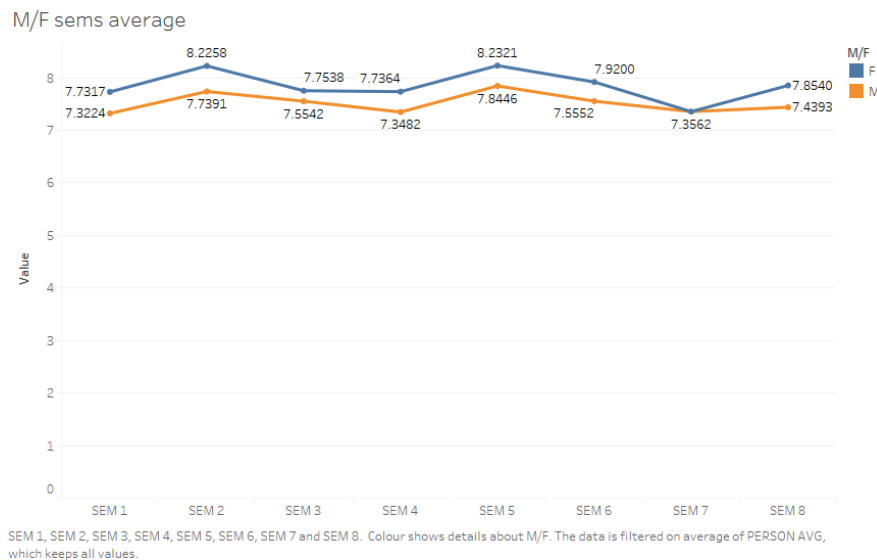


Fig 4.6. Males/ Female CGPA Average

From the above chart, it can be seen that on an average female perform than males in each semester, except for the 7th semester, where both are seen to perform almost same.

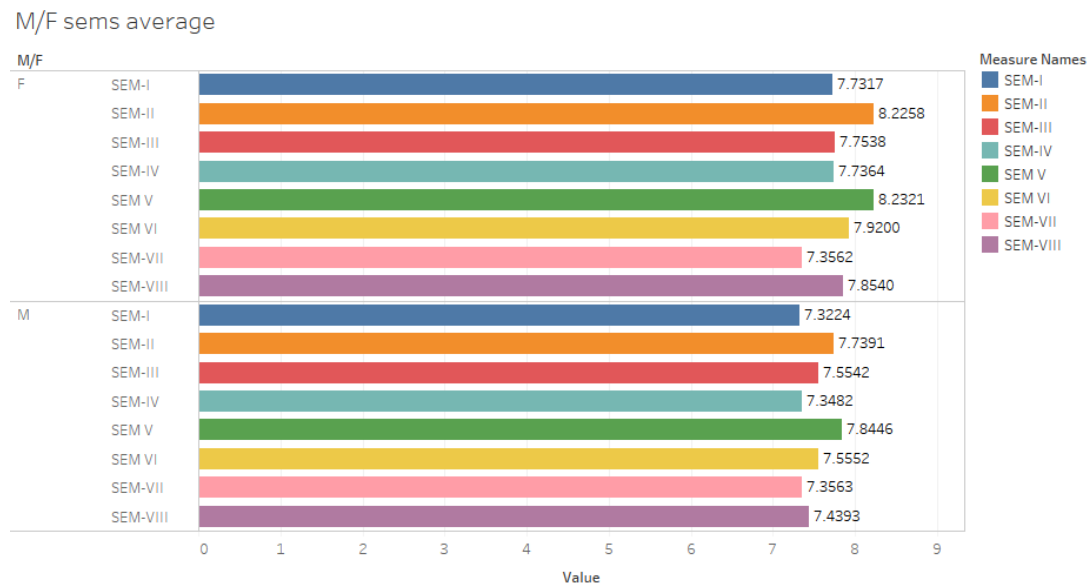


Fig 4.7. CGPA average

It is observed that Males and Females of the 2012 batch, performed the best in their 5th Semester with an average GPA of 8.23 for girls and 7.84 for boys. We will gain further insight about why students performed well in this particular semester when we look into the courses and the faculty for every semester.

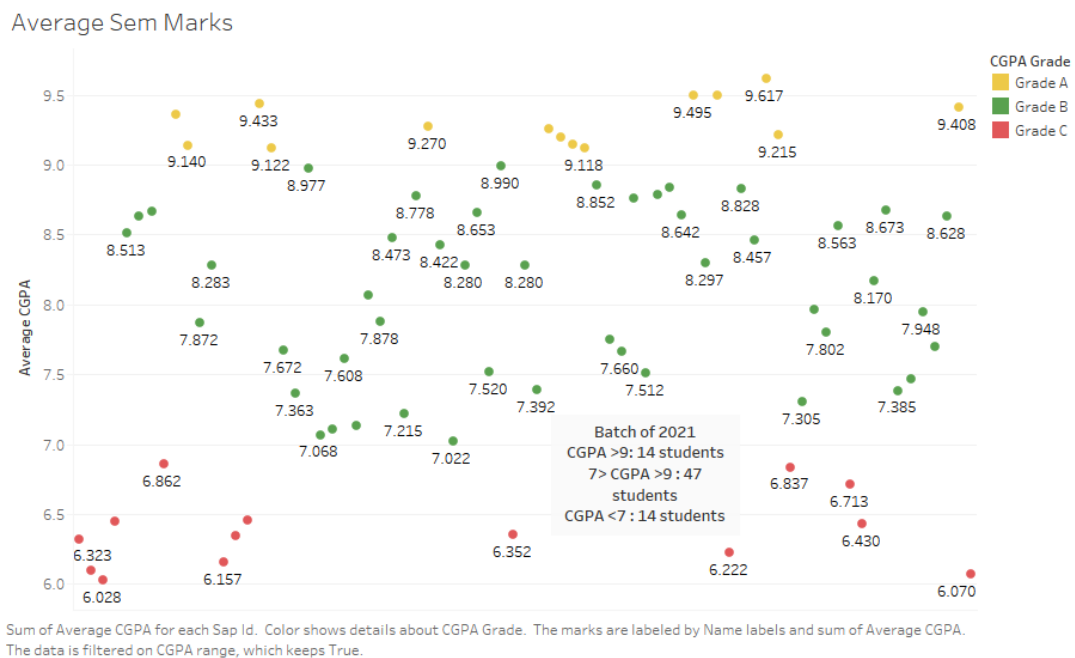


Fig 4.8. CGPA scatterplot

The CGPA for 2021 Batch lies between 6.07 to 9.617 in which no. of students with distinction (Grade A) i.e. CGPA greater than 9 are 14, with Grade B i.e. CGPA between 7 and 9 are 47 students, and lastly with Grade C i.e. CGPA below 7 are 14. Parameters were used for this to view the students between a certain range of CGPA.

CGPA Box and whisker

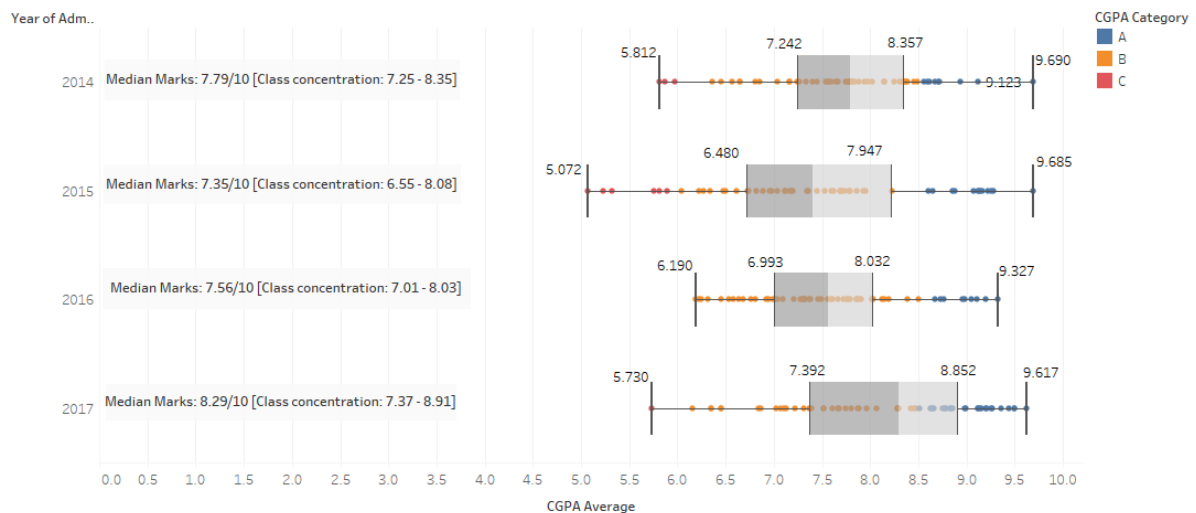


Fig 4.9. CGPA of 4 Batches

CGPA was plotted for 4 batches using box and whisker and the median and concentration of each batch was specified.

CGPA vs. HSC vs. AMCAT vs. CET

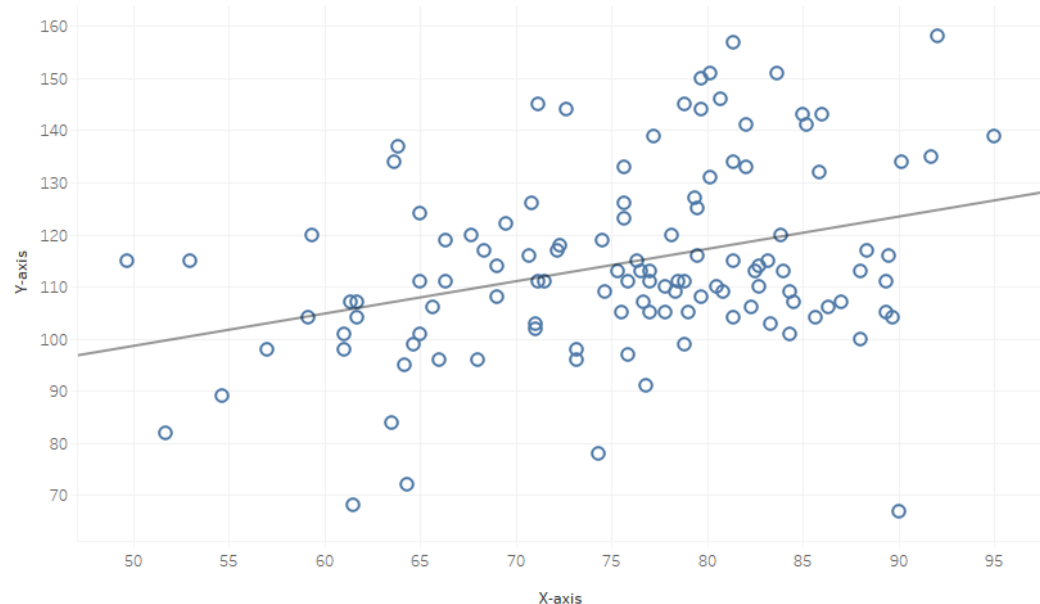


Fig 4.10. CGPA vs. HSC% vs. CGPA vs. AMCAT

The above worksheet contains parameter to choose X and Y axis parameter. The viewer gets four options of HSC %, CET scores, CGPA and AMCAT scores to view on X and Y axis. This makes it easy to find correlation between any of these two features.

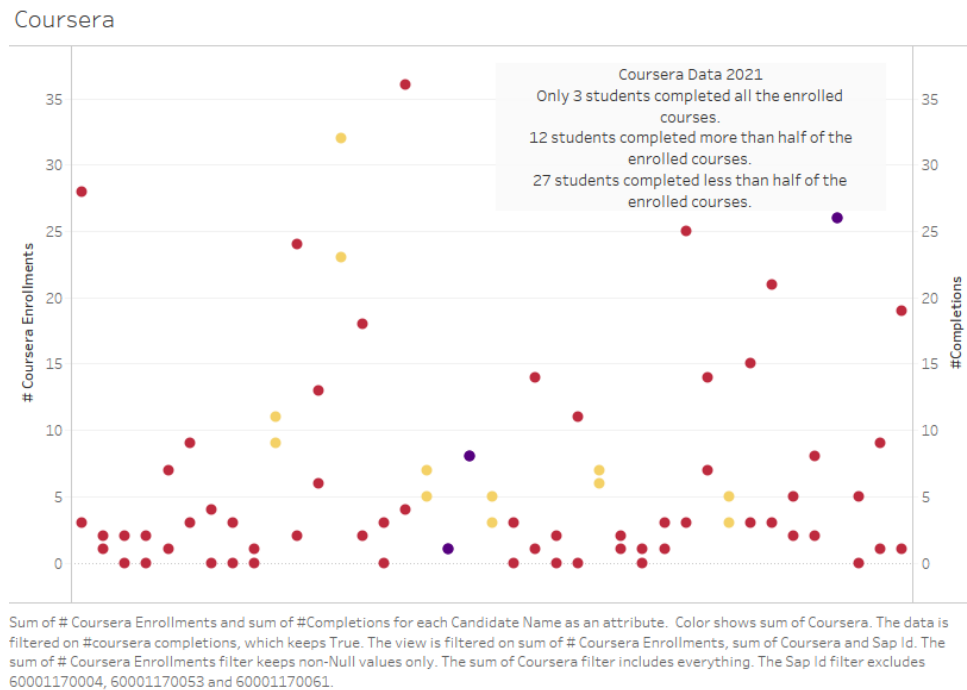
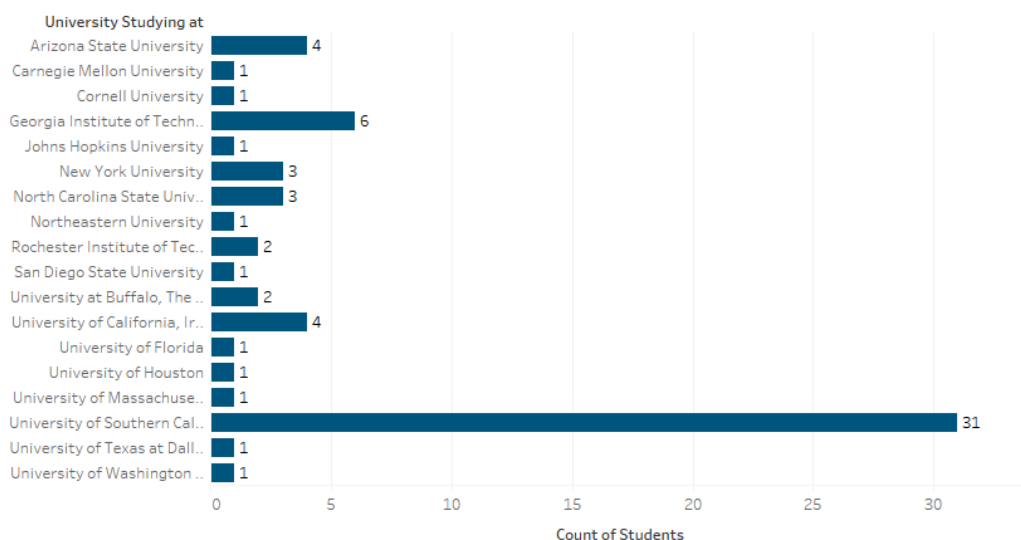


Fig. 4.11. Coursera Enrollments and Completions

The dual-axis chart for plots No. of Coursera Enrollments and No. of Coursera Completions and thus shows that 3 students completed all courses they enrolled for, 12 students completed more than half the courses they enrolled for and 27 students showed no credibility by not even completing half of the courses they enrolled for.

Students studying in USA from 2019 Batch studying Computer Science



Distinct count of Candidate Name for each University Studying at. The marks are labeled by distinct count of Candidate Name. The data is filtered on Batch to display, Country to display and Dept. to display. The Batch to display filter keeps True. The Country to display filter keeps True. The Dept. to display filter keeps True.

Fig 4.12. Department and Country wise student distribution

Parameters let the viewer choose the Batch, Country and Department and then shows the students from that Batch, studying in the selected Country and Department and their distribution over universities. Students from 2019 batch studying CS in USA chose University of Southern California on a huge scale.

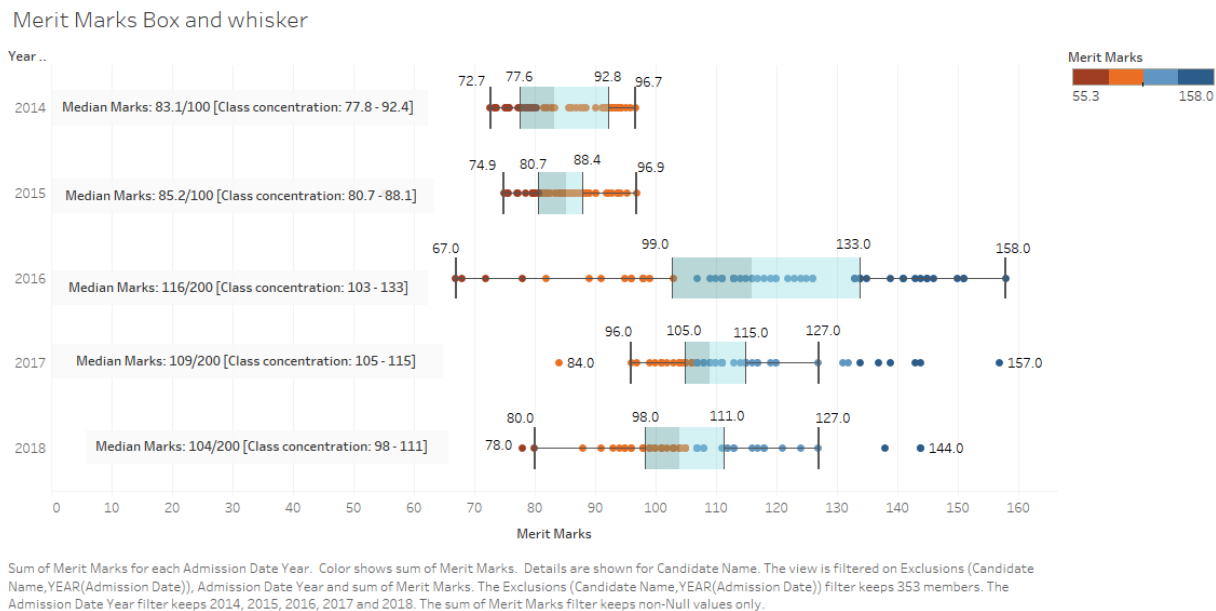


Fig 4.13. CET box and whisker

CET scores was plotted for 5 batches using box and whisker and the median and concentration of each batch was specified. It is observed that the performance of students in CET degraded in each batch.

State-wise distribution on Map

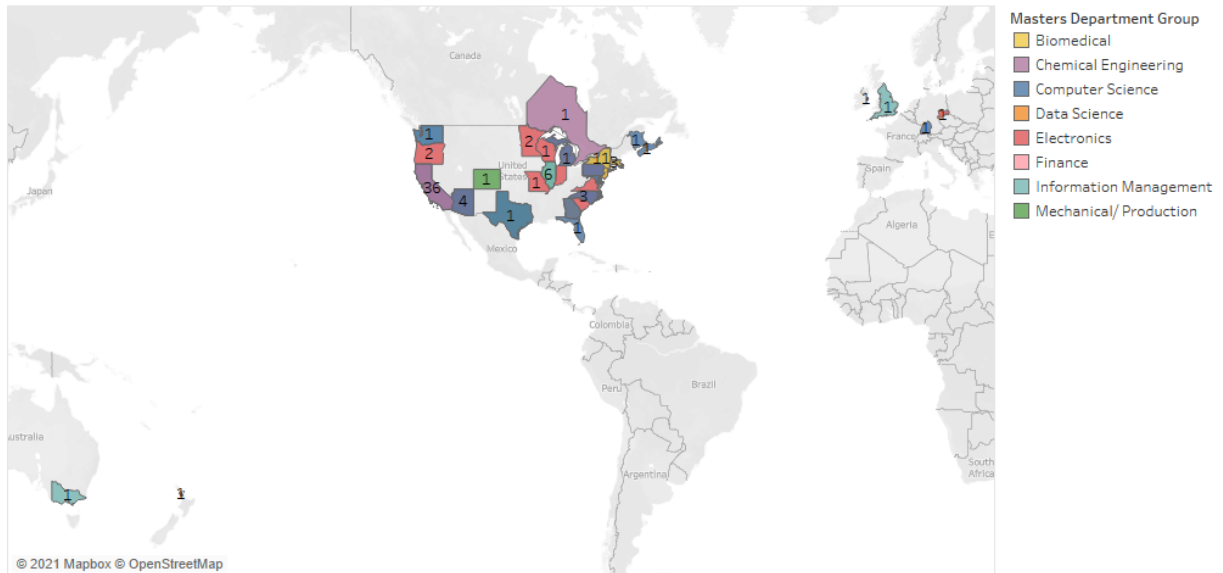


Fig 4.14. Student distribution on Map

Distribution and count of students over various states and their department can be deduced from this chart. It is observed that maximum student Concentration was observed in California.

Stream changed?

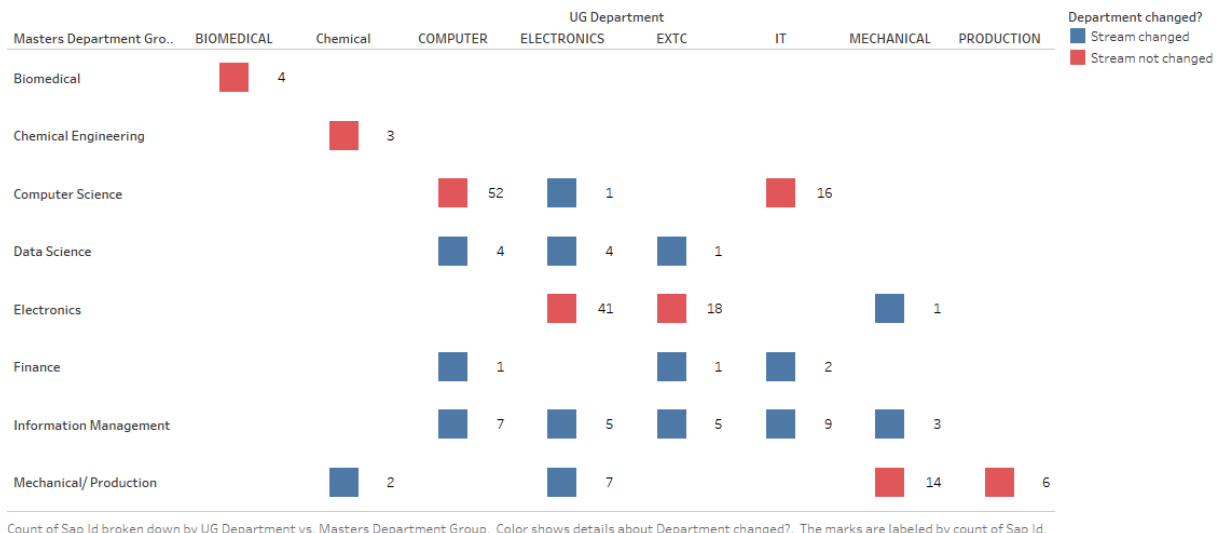
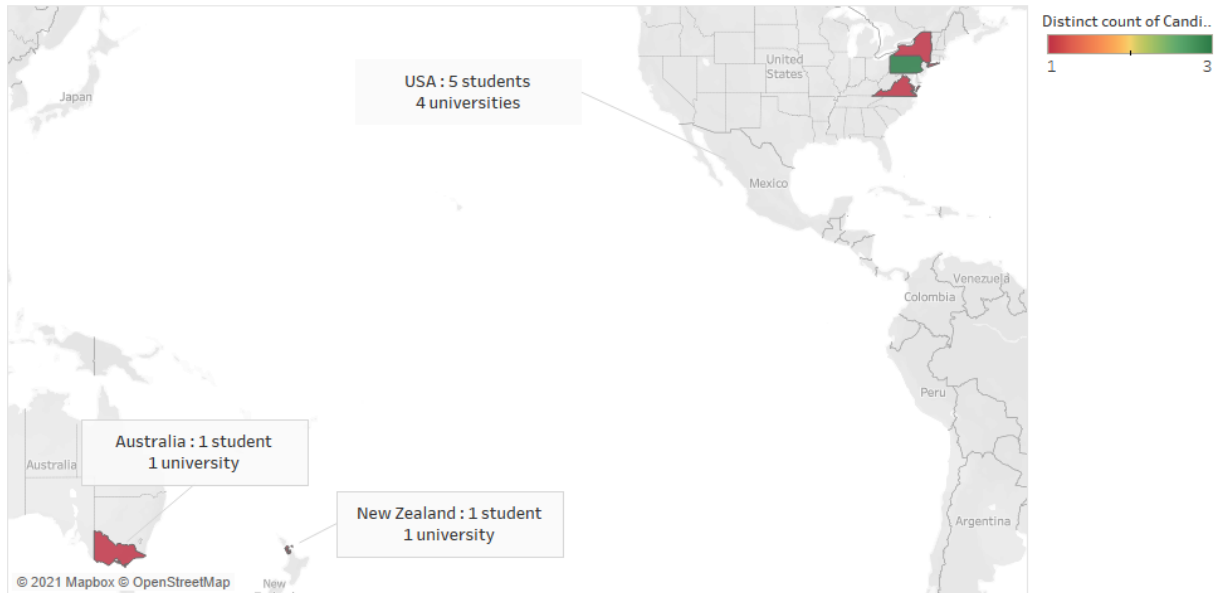


Fig 4.15. Stream changes

The above chart specifies the number of students from each department who changed their stream while going for Masters. All students from Biomedical and Production Department pursued the same stream for their Masters.

2020 Batch



Map based on Longitude (generated) and Latitude (generated). Color shows distinct count of Candidate Name. Details are shown for Country and State. The data is filtered on Batch to display, which keeps True. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Fig 4.16. Country-wise Distribution

The parameter lets you choose the batch and shows the distribution of students over Country and shows color legend on the basis of count.

4.3. Training the Model

The model was trained and tested on Microsoft Azure ML Studio (Classic).

Steps:

1. Uploading the dataset
2. Select Columns from the dataset which are required for the analysis.
3. Edit Metadata for converting the Numerical columns to Categorical Data.
4. Clean Missing Data for cleaning the Null values.
5. Split Data for splitting the data into Training and Testing set. Here we set the split percentage to 0.7 i.e., 70% Training set and 30% Testing set.
6. Training the data consists of 3 steps:
 - i. Choose a model, Classification or Regression model. Here we are using the following models for Classification:
 - Two-class Logistic Classification
 - Two-class Boosted Decision Tree
 - Two-class Boosted Decision Forest
 - Two-class Neural Network
 - Two-class Support Vector Machine
 - And for the regression we are using following Models:
 - Linear Regression
 - Boosted Decision Tree Regression
 - Neural Network Regression
 - Decision Forest Regression
 - ii. Connect Training output of Split data and the selected model to the Train Model.
 - iii. Use a Score Model to use it for predictions.
7. Score Model is used for scoring predictions for a trained classification or regression model. Connect output of the Train Model and the Test output of Split data to the Score Model.
 - i. For classification models, Score Model outputs the probability of the predicted value as well as the predicted value for the class.

- ii. For regression models, the Score Model gives just the predicted numeric value.
8. Evaluate Model is used to calculate a set of metrics used for evaluating the model's accuracy (performance). Right-click the module and select Visualize to see a sample of the results.
9. Add Rows is used to concatenate two datasets, here it concatenates output and thus provides us a comparison for all the metrics.
10. Permutation Feature Importance computes the permutation feature importance scores of various feature variables when given a trained model and a test dataset.
11. Finally, Convert to CSV is connected to the Score Model of the Best Model to obtain a CSV file of the predictions.

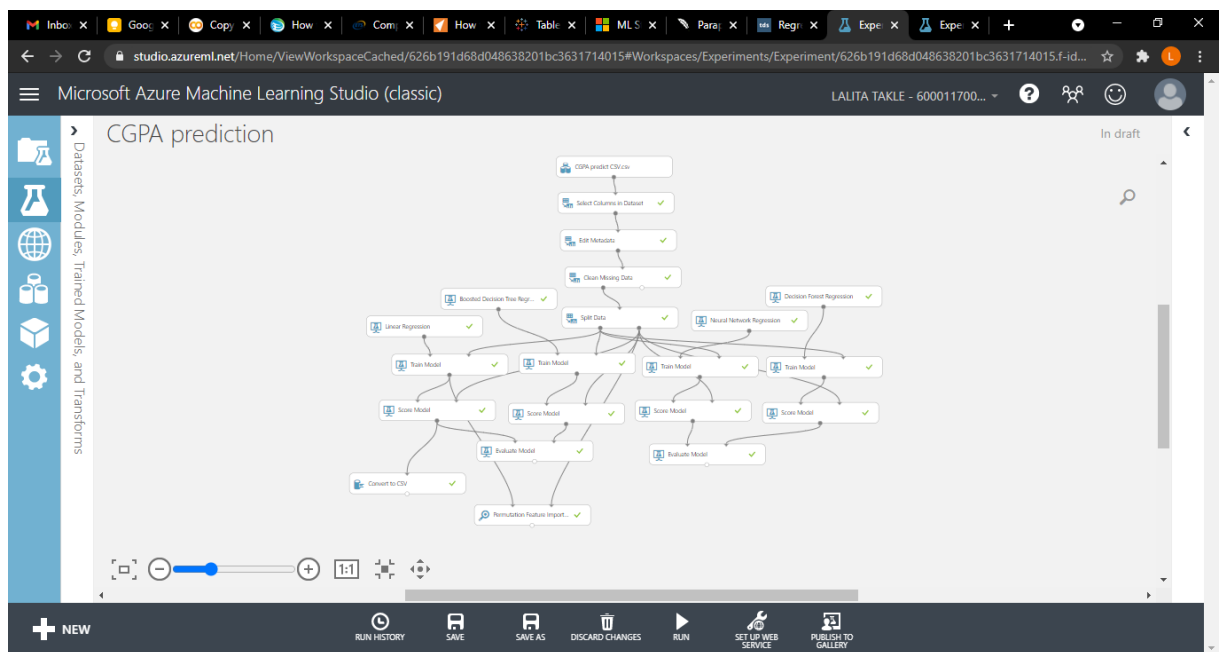


Fig 4.17. CGPA prediction

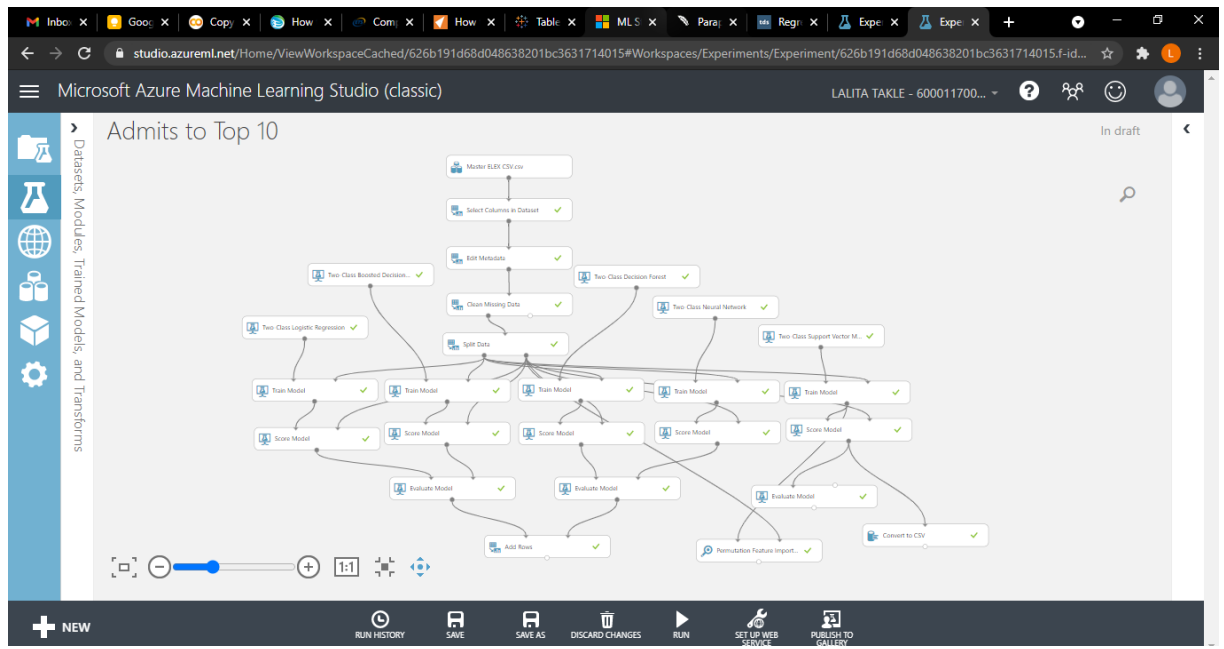


Fig 4.18. Admission prediction

4.4. Analysis of the Results

The results from all the models for both the predictions are compared on the basis of their metrics. The following metrics are used to compare models.

i. Regression Models – CGPA Prediction

- **Mean Absolute Error (MAE):** This metric determines how similar the forecasts are to the real outcomes; a lower score indicates better accuracy.
- **Root Mean Squared Error (RMSE):** It produces a single value that sums up the model's error. The metric ignores the difference between over-prediction and under-prediction by squaring the difference.
- **Relative Absolute Error (RAE):** RAE between predicted and observed values is defined as the mean difference divided by the arithmetic mean.

- Relative squared error (RSE): Similarly, the total squared error in the predicted values is normalized by a relative squared error (RSE) by dividing the actual values into a complete squared error.
- Coefficient of determination (R^2): It denotes the model's prediction performance as a value between 0 and 1. A value of 0 indicates that the model is random (and thus explains nothing); a value of 1 indicates that there is a perfect fit. Nevertheless, when evaluating R^2 values, keep in mind that low values can be completely normal, while high values can be suspicious.

Table 4.1 CGPA prediction Metrics

Metrics	Linear Regression	Boosted Decision Tree Regression	Neural Network Regression	Decision Forest Regression
MAE	0.6421	0.7185	0.7161	0.6687
RMSE	0.7656	0.9163	0.8648	0.8517
RAE	0.8771	0.9816	0.9782	0.9134
RSE	0.7102	1.1017	0.9062	0.8788
R^2	0.2897	-0.0173	0.0937	0.1211

RMSE is the most preferred metric because the errors are squared before averaging which means that large errors pay a high penalty. This means RMSE is useful when major mistakes are unwanted. Thus, considering RMSE, it is observed that Boosted Decision Tree Regression is the best Model for CGPA prediction.



Feature	Score
	
HSC Eligibility Percentage	0.269672
Interest in Electronics	0.052828
Gender	0.044887
CET Marks	0.018805
No. of KT	0.002212
Health	0.001821
Extracurriculars	0.001343

Fig 4.19. CGPA Features

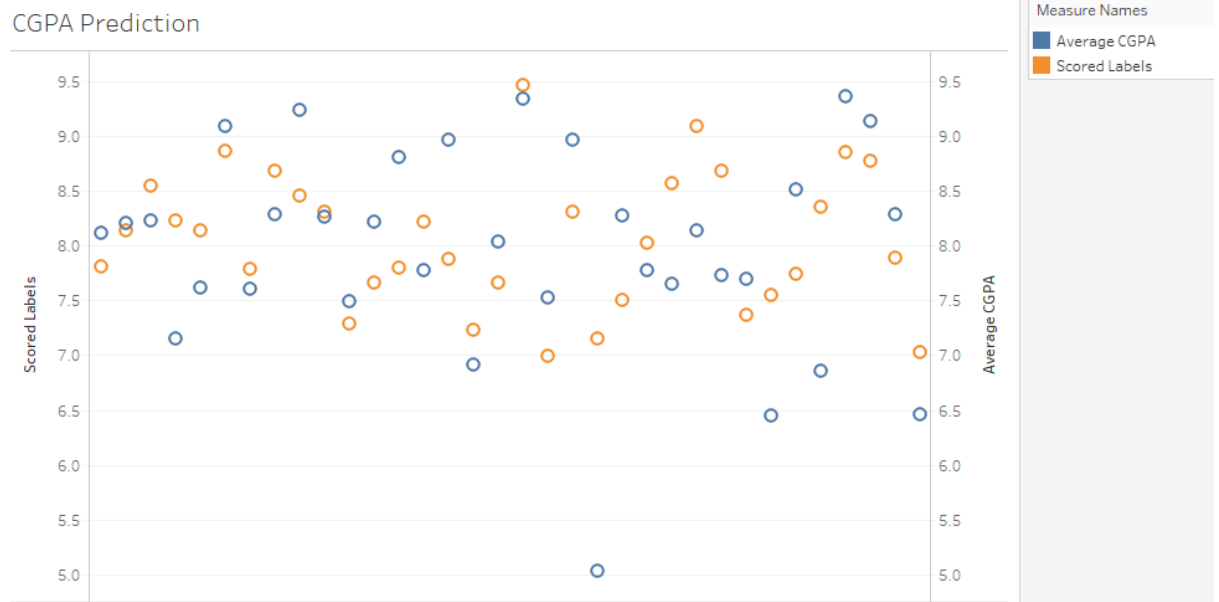


Fig 4.20. CGPA prediction

ii. Classification Models – Admit chances prediction

- Accuracy: A classification model's accuracy is measured as the proportion of true results to total cases.
- Precision: It is the ratio of actual results to all positive outcomes.
- Recall: This is the proportion of valid results restored by the model.
- F-score: It is calculated as the weighted average of precision and recall between 0 and 1, with 1 being the ideal F-score.
- AUC: The Area Under the Curve (AUC) is a metric of a classifier's ability to differentiate between classes and is used to summarize the ROC curve.

For Classification Models, after Accuracy AUC is considered to be the best metric to decide the best model because it is scale-invariant and measures how well predictions are ranked and not their absolute values. It is classification-threshold-invariant, which means that it measures the quality of predictions done by the model irrespective of the chosen threshold, unlike the rest of the metrics which depend on the value of the threshold. Although Accuracy is threshold-variant, but it is also considered along with AUC.

A. Prediction of Admit Chances into Top 10 Universities

Table 4.2. Admits in Top 10 Universities prediction

Metrics	Two-Class Logistic Regression	Two-Class Boosted Decision Tree	Two-Class Decision Forest	Two-Class Neural Network	Two-Class Support Vector Machine
Accuracy	0.9193	0.9516	0.9193	0.9516	0.968
Precision	0	0.5	0	0.5	0.6
Recall	0	0.6667	0	0.6667	1.00
F-score	0	0.5714	0	0.5714	0.75
AUC	0.9661	0.9774	0.9661	0.9661	0.972

The above table shows that AUC and Accuracy for Two-Class Support Vector Machine is the highest and proves that for predicting the chances of getting an admit in the Top 10 university, Two-Class Support Vector Machine is the Best Model.

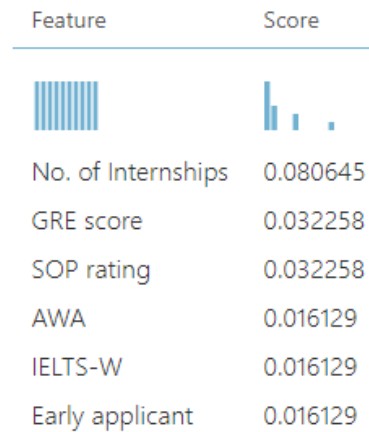


Fig 4.21. Admits Top 10 Features

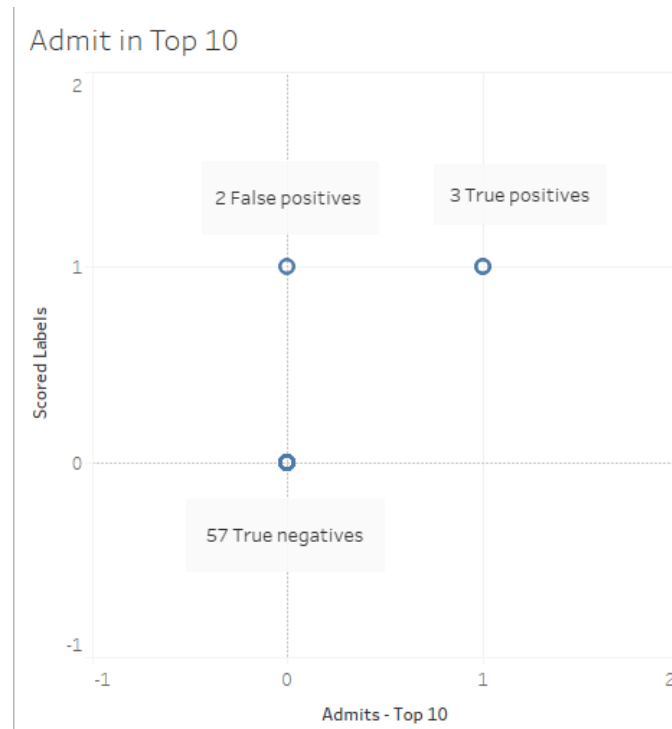


Fig 4.22. Admits Top 10 prediction

B. Prediction of Admit Chances into Top 20 Universities

Table 4.3. Admits in Top 20 Universities prediction

Metrics	Two-Class Logistic Regression	Two-Class Boosted Decision Tree	Two-Class Decision Forest	Two-Class Neural Network	Two-Class Support Vector Machine
Accuracy	0.9354	0.9032	0.9516	0.9677	0.935
Precision	1	0.5714	1	1	0.714
Recall	0.4285	0.5714	0.5714	0.7142	0.714
F-score	0.6	0.5714	0.7272	0.8334	0.714
AUC	0.9350	0.8961	0.9740	0.9298	0.960

The above table shows that AUC for Two-Class Decision Forest is the highest and proves that for predicting the chances of getting an admit in the Top 20 university, Two-Class Decision Forest is the Best Model.

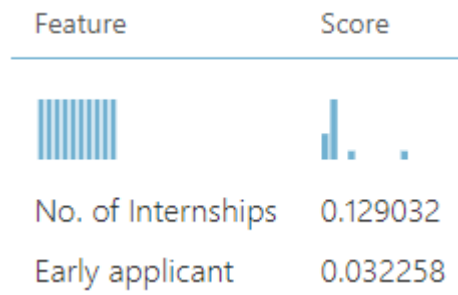


Fig 4.23. Admits Top 20 Features

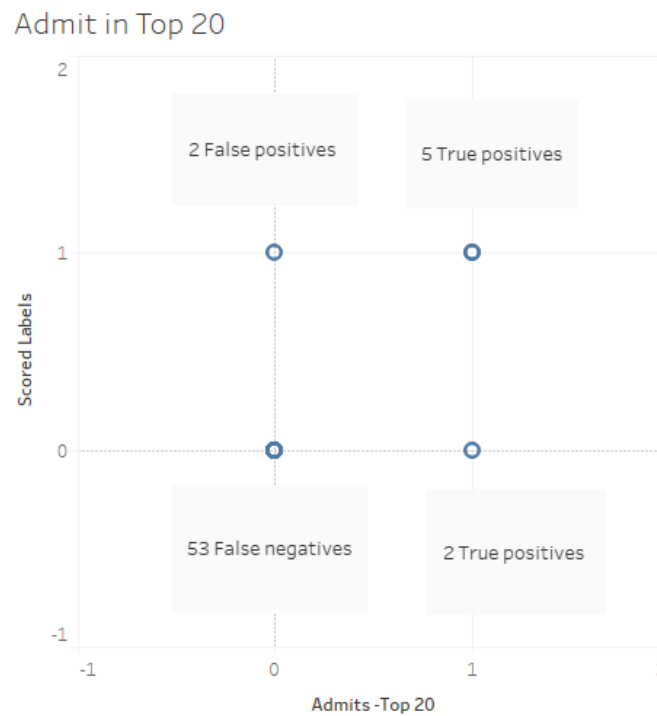


Fig 4.24. Admits Top 20 prediction

C. Prediction of Admit Chances into Top 50 Universities

Table 4.4. Admits in Top 50 Universities prediction

Metrics	Two-Class Logistic Regression	Two-Class Boosted Decision Tree	Two-Class Decision Forest	Two-Class Neural Network	Two-Class Support Vector Machine
Accuracy	0.9032	0.8870	0.9193	0.9193	0.855
Precision	0.8181	0.7142	1	0.7857	0.667
Recall	0.6923	0.7692	0.6153	0.8461	0.615
F-score	0.75	0.7407	0.7619	0.8148	0.640
AUC	0.9372	0.9403	0.9120	0.9387	0.934

The above table shows that AUC for Two-Class Boosted Decision Tree is the highest and proves that for predicting the chances of getting an admit in the Top 10 university, Two-Class Boosted Decision Tree is the Best Model.



Feature	Score
	
IELTS	0.112903
Early applicant	0.096774
GRE score	0.048387
No. of Internships	0.048387
No. of Courses done	0.048387
No. of KT's	0.032258

Fig 4.25. Admits Top 50 Features

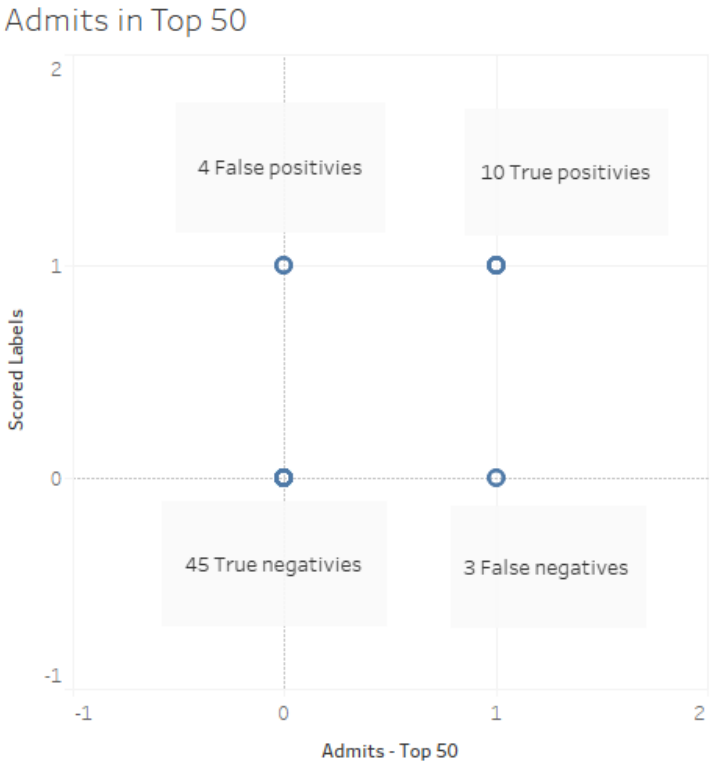


Fig 4.26. Admits Top 50 prediction

D. Prediction of Admit Chances into Top 100 Universities

Table 4.5. Admits in Top 100 Universities prediction

Metrics	Two-Class Logistic Regression	Two-Class Boosted Decision Tree	Two-Class Decision Forest	Two-Class Neural Network	Two-Class Support Vector Machine
Accuracy	0.7903	0.7741	0.7903	0.8064	0.839
Precision	0.7906	0.7608	0.8048	0.8095	0.833
Recall	0.8947	0.9210	0.8684	0.8947	0.921
F-score	0.8395	0.8333	0.8354	0.85	0.875
AUC	0.8859	0.7730	0.8514	0.8925	0.895

The above table shows that AUC for Two-Class Support Vector Machine is the highest and proves that for predicting the chances of getting an admit in the Top 10 university, Two-Class Support Vector Machine is the Best Model.

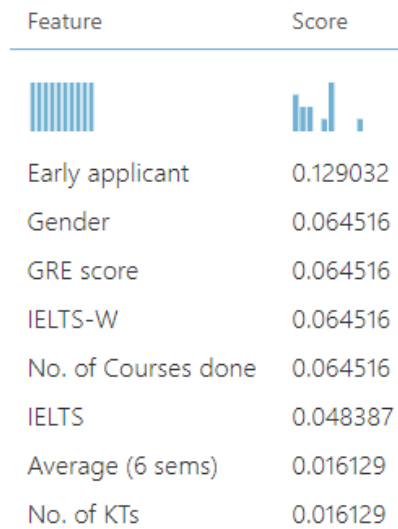


Fig 4.27. Admits Top 100 Features

Admits in Top 100

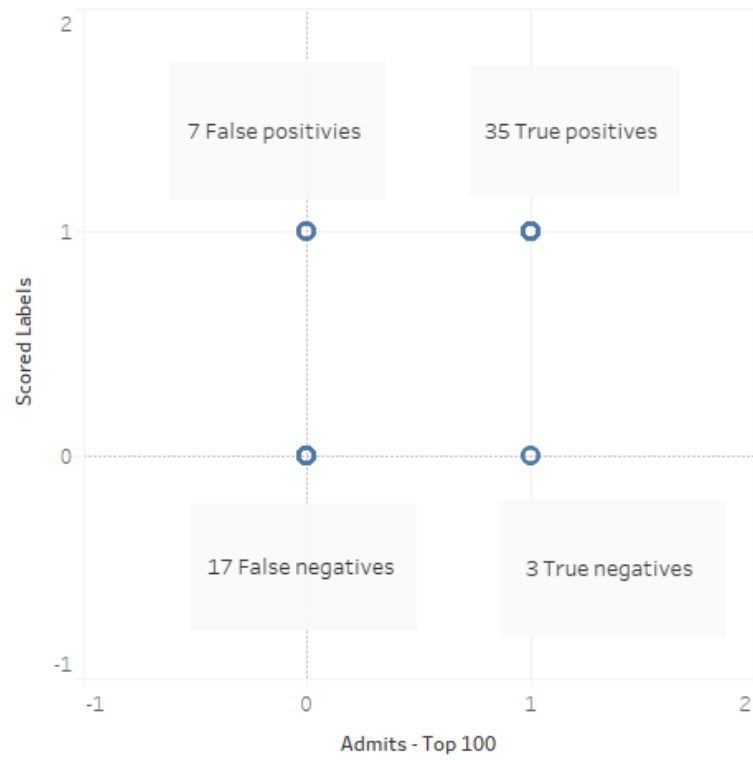


Fig 4.28. Admits Top 100 prediction

Chapter 5

Conclusion & Scope

From over years, the organisation of higher education systems has evolved from reactionary to assertive in decision-making and system performance analysis. The reliability of the education system and the overall value of knowledge is essential for student performance. There are many advantages for early diagnosis of student problems and learning problems, as it offers a unique chance of tackling causative factors on time to prevent student failure and to drop out trends. In determining the final CGPA and grade course, the performance of engineering students within the first two years of study is often said to be the main factor due to the difficulties of increasing grades significantly at higher levels with robustness and intensity. The whole model can assist many students plan and work towards their career paths. The early prediction of CGPA not only leads to making additional effort in the course of their engineering work, but also to maintain an average CGPA overall. The profile evaluator will help students understand the position of their dream university competition and thus encourage them to go an extra mile.

This study has a number of constraints. The datasets are from one engineering university and only students from the Department of Electronics, other departments and universities can be included in further research. Furthermore, educational research shows that certain factors, such as learning, self-efficacy, motivation and interest, as well as the environment of teaching and learning play a role in the learning process and thus affect student achievement. Future studies should therefore include these variables in the models so that prediction accuracy is improved.

Chapter 6

References

- [1] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in *Procedia Computer Science*, 2015, doi: 10.1016/j.procs.2015.12.157.
- [2] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Comput. Educ.*, 2016, doi: 10.1016/j.compedu.2016.09.005.
- [3] N. Buniyamin, U. Bin Mat, and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," in *2015 IEEE 7th International Conference on Engineering Education, ICEED 2015*, 2016, doi: 10.1109/ICEED.2015.7451491.
- [4] P. Chaudhury, S. Mishra, H. K. Tripathy, and B. Kishore, "Enhancing the capabilities of student result prediction system," in *ACM International Conference Proceeding Series*, 2016, doi: 10.1145/2905055.2905150.
- [5] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008*, 2008.
- [6] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, 2017, doi: 10.1016/j.compedu.2017.05.007.
- [7] R. R. Rajalaxmi, P. Natesan, N. Krishnamoorthy, and S. Ponni, "Regression model for predicting engineering students academic performance," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 71–75, 2019.
- [8] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," in *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects, COMITCon 2019*, 2019, doi: 10.1109/COMITCon.2019.8862214.
- [9] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, 2019, doi: 10.1016/j.heliyon.2019.e01250.
- [10] K. S. Bhagavan, J. Thangakumar, and D. V. Subramanian, "Predictive analysis of student academic performance and employability chances using HLVQ algorithm," *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-019-01674-8.
- [11] S. K. Thangavel, P. D. Bkaratki, and A. Sankar, "Student placement analyzer: A

- recommendation system using machine learning,” in *2017 4th International Conference on Advanced Computing and Communication Systems, ICACCS 2017*, 2017, doi: 10.1109/ICACCS.2017.8014632.
- [12] C. Vialardi, J. Braver, L. Shafr, and Á. Ortiaosa, “Recommendation in higher education using data mining techniques,” in *EDM’09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*, 2009.