Abstract

This report describes the use of classification algorithms in predicting the success of undergraduate engineering students of a private university of northern India. The success is measured in terms of academic performance and campus placements. The main objective of the study was to examine which classification algorithms gives best accuracy in present context and what the most sensitive predictor variables? The results indicate Neural Network gives best accuracy (77%) for predicting the academic performance and decision tree is suitable for predicting placement success or employability of students. Prediction of academic performance can be done using performance in high school and grades acquired during first semester. However, it was not possible to build a model for predicting placement success by using academic progress related attributes. Inclusion of marks awarded by industry experts lead to development of model with good accuracy of 75%.

Contents

1. Introduction and business context

The success of students plays a very crucial role in satisfaction of stakeholders of any University. This in turn is very useful in improving the quality of University as it will attract good students. This case study is of an Indian private university where the popular programmes are undergraduate programmes offered by school of engineering and technology. The success here is considered in two contexts. First is academic success, this indicates that a student is able to clear the courses/modules of the programmes with good grades. Second, a student is able to secure a job through campus placement. That is, for the University considered, this project aims to develop classification models to predict the success of undergraduate engineering student. The models developed will be of benefit to the University as these will help the strategic planning team in deciding admission policies, campus placement activities, teaching-learning processes, mentoring and grading policies.

The classification models will try to predict the academic success of the student based on attributes of previous education (high school performance). The study also aims to investigate if the model differs for gender or students belonging to different branches (like civil, computer science, electrical etc). For predicting the placement success, additional attributes from pre-placement interviews are also considered. Pre-placement interviews are structured processes followed by the department of computer science for past one year, where experts from Industry are called to assess students on a standard battery (parameters defined under two subheads). These experts rate the students in two domains: *Attitude & Skills* and *Technical Knowledge*.

The report is structured as follows: in section 2 data mining objectives are described. Section 3 contains project planning and execution methodology. In section 4 and 5 details of data exploration and data preparation are discussed. Finally, Modelling and evaluation of models are discussed in section 6.

2. Data mining objectives

Two main objectives are:

1. To predict the academic success of the students which means predicting the percentage marks with which a student is going to take the degree. Also to predict whether he/she will take the degree with {distinction, first-class, second-class}
2. To predict the placement success of the students which means predicting whether the student is going to get a job in first round of campus placements or not.

Two different datasets were constructed for these predictions. These are explained in section 4 and 5.

3. Project planning and execution

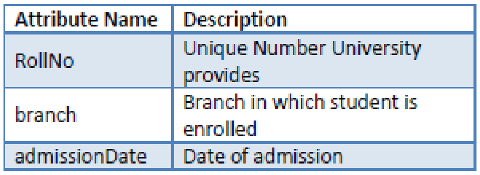
Table 1 Methodology used

Step	Phase	Work assigned
CRISP_DM 1	Business Understanding	Worked closely with two senior members of strategic planning team of the University for identifying few Key Performance Indicators of academic success and placement success of students
CRISP_DM 2	Data Understanding	The data was collected from four different functional units of the University, it was than integrated and explored
CRISP_DM 3	Data Preparation	After addressing the quality issues, initial classification algorithms were modelled with given KPIs, some of these were revised in the light of poor accuracy. Some more input attributes were added to improve the accuracy
CRISP_DM 4	Data Modelling	Decision Trees, K-NNs, Naïve-Bayes and Neural Networks were applied
CRISP_DM 5	Evaluation	The algorithms were evaluated using confusion matrix

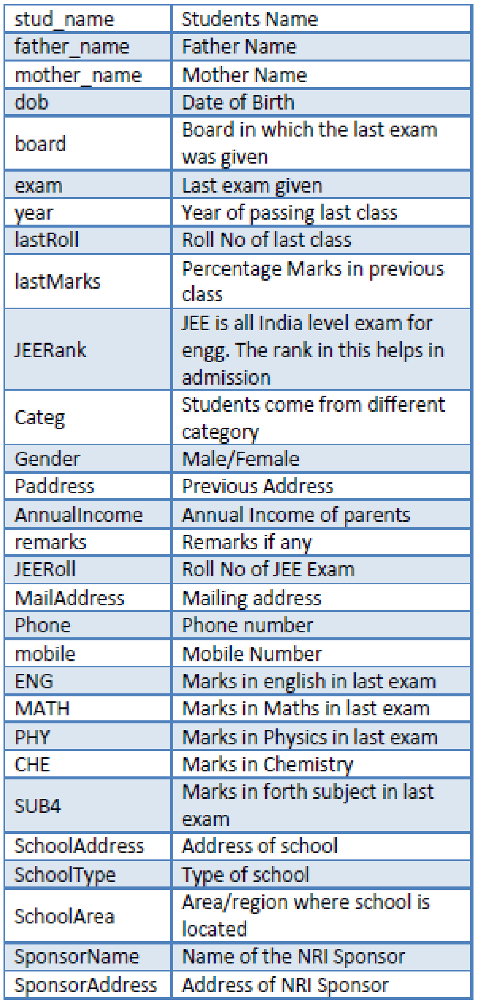
4. Data Understanding

3.1 Collect the data: The data was scattered in four different functional units of the university. The brief introduction to the units and data is as follows:

3.1.1 Controller of Records: This department collects and stores the data of students at the time of admission and is responsible for maintaining this information. The data for last one batch (students who took admission during last year) is depicted here. Some of the fields were different for previous batches (admitted during previous two years) as JEE marks(explained later) were not recorded for the earlier batches.



Attribute Name	Description
RollNo	Unique Number University provides
branch	Branch in which student is enrolled
admissionDate	Date of admission



stud_name	Students Name
father_name	Father Name
mother_name	Mother Name
dob	Date of Birth
board	Board in which the last exam was given
exam	Last exam given
year	Year of passing last class
lastRoll	Roll No of last class
lastMarks	Percentage Marks in previous class
JEERank	JEE is all India level exam for engg. The rank in this helps in admission
Categ	Students come from different category
Gender	Male/Female
Paddress	Previous Address
AnnualIncome	Annual Income of parents
remarks	Remarks if any
JEERoll	Roll No of JEE Exam
MailAddress	Mailing address
Phone	Phone number
mobile	Mobile Number
ENG	Marks in english in last exam
MATH	Marks in Maths in last exam
PHY	Marks in Physics in last exam
CHE	Marks in Chemistry
SUB4	Marks in forth subject in last exam
SchoolAddress	Address of school
SchoolType	Type of school
SchoolArea	Area/region where school is located
SponsorName	Name of the NRI Sponsor
SponsorAddress	Address of NRI Sponsor

4.1.2 *Dean Office:*

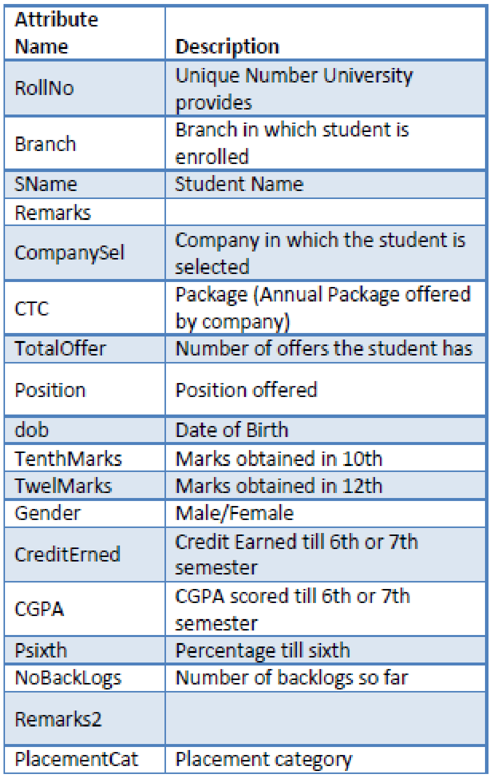
This office maintains the information about students' performance (grades) in various courses, SGPA and CGPA of the students. For each semester a separate table is maintained, and for a four year programme. 8 such tables are maintained for each batch. The number of columns in each table depends on the number of courses (modules) taken by the students. These are maintained as excel sheets.

4.1.3 Department Level:

The pre-placement interviews were conducted for one batch of 173 students and these are recorded as marks obtained in various parameters of Attitude & Skill and marks obtained in Knowledge domain. The aggregated sum of marks obtained in sub-parameters of attitude & skill and knowledge domain was considered for analysis.

4.1.4 School of professional attachment:

This school is responsible for helping students with placements and training in companies. The attributes recorded here are



Attribute Name	Description
RollNo	Unique Number University provides
Branch	Branch in which student is enrolled
SName	Student Name
Remarks	
CompanySel	Company in which the student is selected
CTC	Package (Annual Package offered by company)
TotalOffer	Number of offers the student has
Position	Position offered
dob	Date of Birth
TenthMarks	Marks obtained in 10th
TwelMarks	Marks obtained in 12th
Gender	Male/Female
CreditErned	Credit Earned till 6th or 7th semester
CGPA	CGPA scored till 6th or 7th semester
Psixth	Percentage till sixth
NoBackLogs	Number of backlogs so far
Remarks2	
PlacementCat	Placement category

For each batch a separate excel sheet is made. The next step was to integrate this data using common field RollNo. However, the rollno was in different format in some of these excel sheets (eg 10-csu-012 and 10csu012 or 10CSU012). This was corrected in the excel sheet first and then the Talend data integration tool was used to combine all the attributes together.

5. Data Exploration

To predict the academic success of a student, records of past three batches were considered, as the University was following similar grading pattern since then. Prior to that absolute grading was practiced and hence those records were not considered for analysis. A total of 1793 rows are recorded there in undergraduate programme of School of Engg and Tech. The details recorded at the time of admission were explored. The details are discussed as follows.

- **RollNo –** is a unique number given to each student and all 1793 students have *unique rollno*. This field is taken as *id*. Though this can be ignored but, it was considered for tracking the outliers (which can be future extension of the work done for this project)
- **Branch –** There are 9 branches. The datatype is polynomial. The distribution of students in these branches is shown in Figure1.

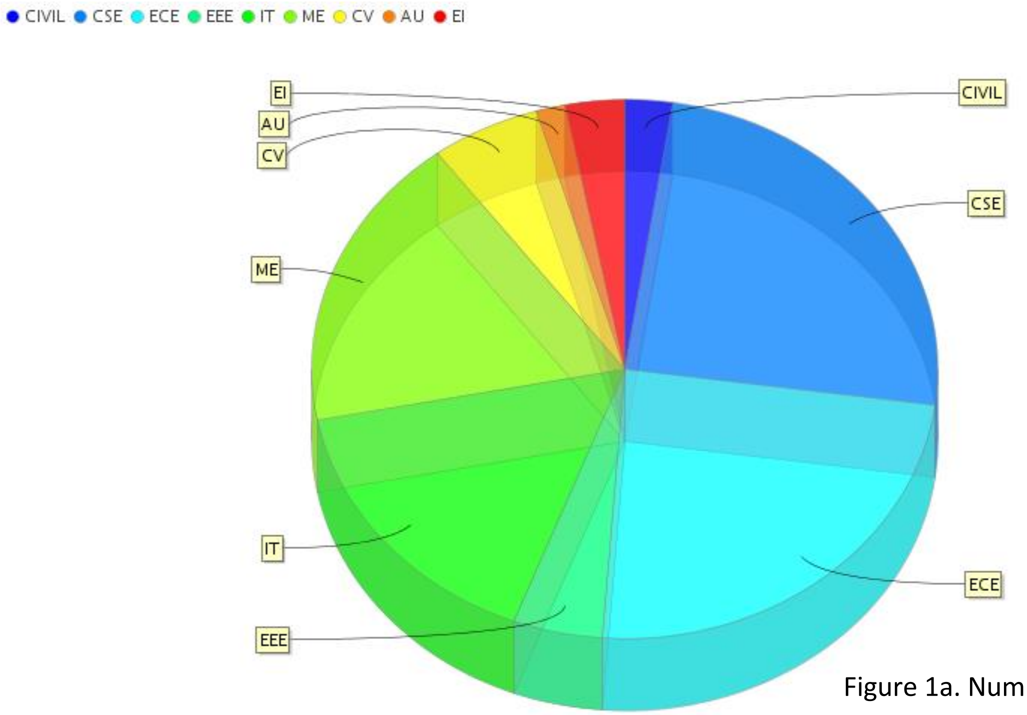


Figure 1a. Number of students in different branches

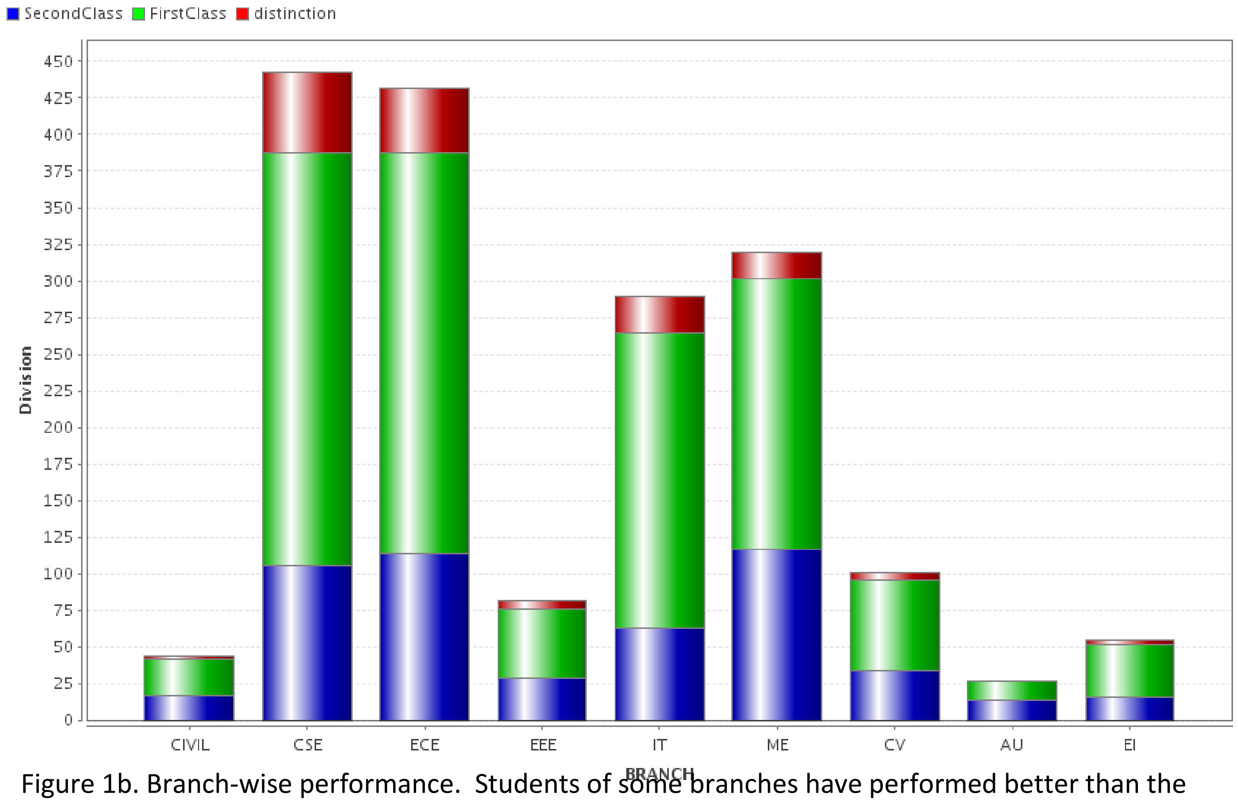


Figure 1b. Branch-wise performance. Students of some branches have performed better than the others. Thus attribute has prediction capacity

Intake in ECE, CSE and ME is higher as compared to other streams. Two streams EI and AU are discontinued for past two years, which is why number of students are less in these streams. Thus, during classification the branch wise model was not made for AU and EEE. However, the attribute was considered as it is interesting to examine whether the prediction model have different accuracy for different branches. This will helps stakeholders understand the variation in practices followed by different academic departments. Also, it is evident from figure1b that performances of students differ across branches, so it is important to include this attribute.

- **Admission Date:** This attribute is not included as there are three admission dates for past three batches. This data can be retrieved using rollno's as well as pattern indicated the admission year.
- **Student Name, Father Name and Mother Name:** will not add any value to the analysis and hence these attributes are not considered for analysis.
- **Date of Birthdate:** Due to eligibility criteria the range of ages are 18 to 19 years. The mean is 18.5 years. This will not be significant in predicting the success and hence it was not considered for dataset.
- **Board of last exam:** out of 1793 entries 1790 were CBSE and only 3 were from different board. Hence, variation being too low, the attribute will not contribute significantly in prediction. However, the good thing about this low variation is that the marks obtained in high school being from same board are comparable.
- **Last exam:** all the records have same value. Attribute ignored.
- **Last Rollno:** doesn't contribute in prediction. Attribute ignored.
- **Last Marks:** these indicate marks obtained in previous exam, which is high school, also called 12th in India. This attribute is taken. This is given in percentage. The mean value being 84.6% (+- 6), Range of the marks is [60.000 ; 95.200] and median is 79.2. The box plot and histogram are shown in figure2:

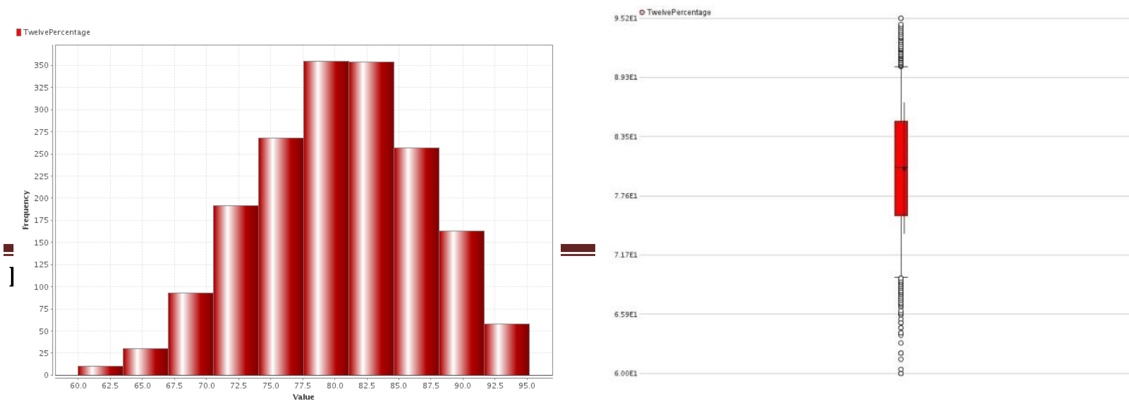


Figure2. Marks scored in 12th class (high school)

The histogram shows a normal distribution, which is also evident from skewness whose value is -0.2 (which is close to zero). The negative part shows it is slightly skewed towards right, this indicates that there are more values above the average. The wide range of the distribution makes it a good attribute for input to classifier.

- **JEE Rank-** this is rank in a National Level exam which is used for admission, though this input can be of help as the University considers this as one of the admission criteria, but the criteria used was different for all the three batches, hence it is not advisable to keep this in dataset used for prediction.
- **Category:** This indicates the reservation of seats in admission. There are only two categories: one being Haryana General and other is All India. The university no longer has reservations, as the policy is no longer functional. So including this attribute which not add any value in business context as it will not be helpful in designing new policies.
- **Gender:** This attribute is considered for inclusion in dataset. There are more number of males in undergraduate engineering programmes as compared to females. (as shown in figure3)

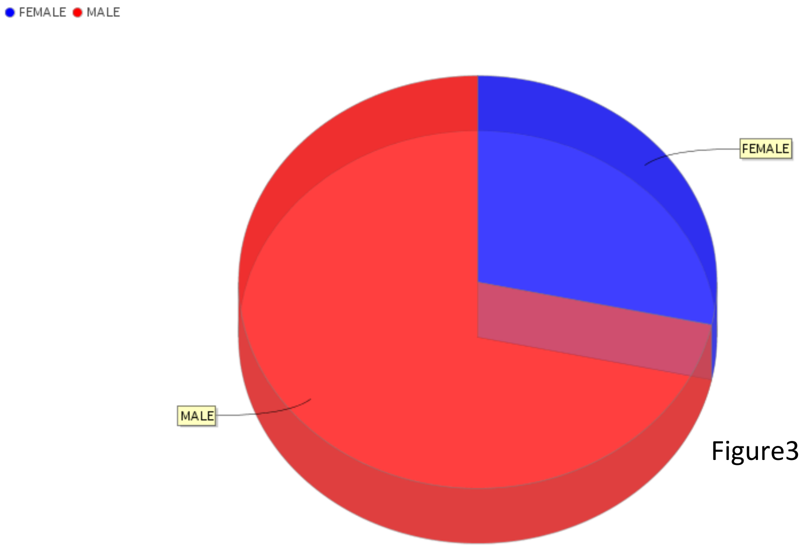


Figure3. Numbers of males and females

The possibility of this attribute in predicting success is good as average marks scored by girls are more in high school as well as engineering programme. This is visible from figure 4. The class imbalance can be addressed by duplicating some of the records of females.

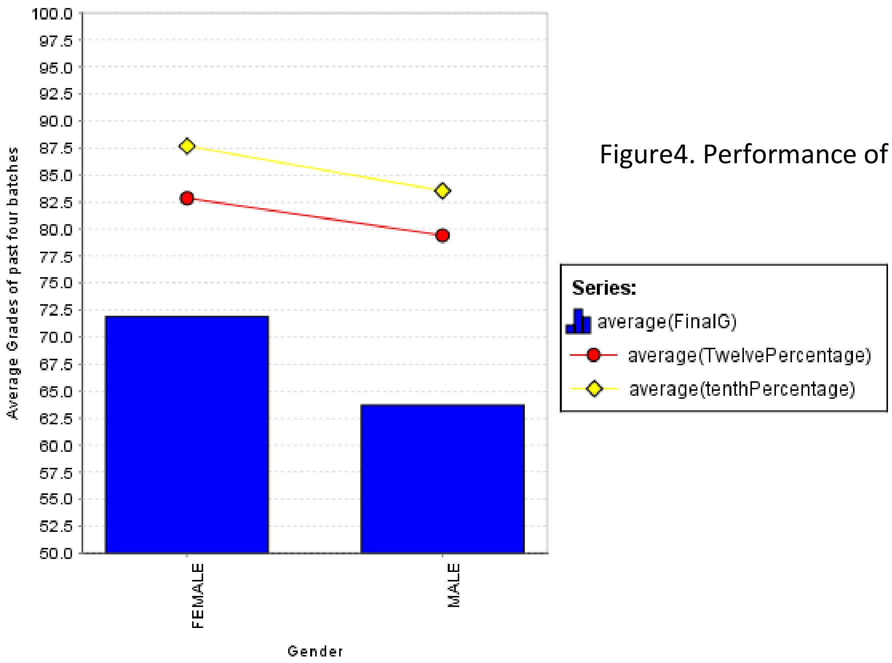


Figure4. Performance of males and females

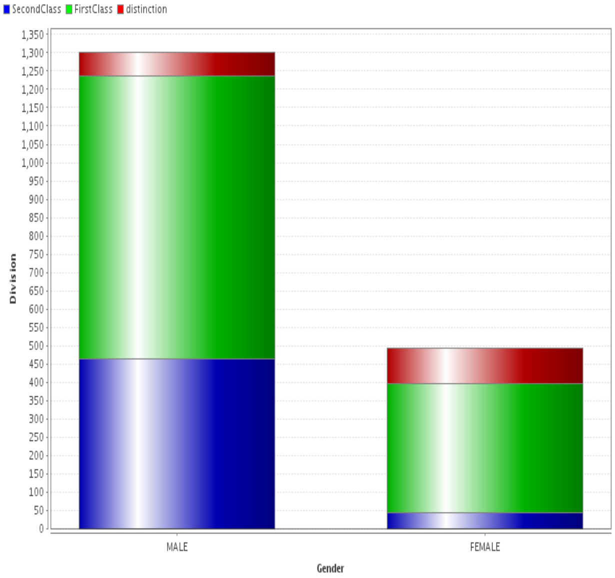


Figure5. Number of distinction, firstclass and secondclass in males and females

From figure5 it is evident that higher percentage of females achieved more number of distinctions and lesser percentage of these were having second-class. Thus, this attribute can play a significant role in prediction.

- **Permanent Address, Mail Address:** since the university is not residential, most of the students come from geographically closer location. And subgrouping of regions is not feasible; also it will not lead to meaningful business decision related to admissions or campus placements. Thus, this attribute is not selected for database construction.
- **Annual Income:** 42% of the data about annual income of the parents was missing the records hence this attribute was not selected.
- **JEE-rollno:** not meaningful and hence discarded.
- **Phone and mobile:** unique for each 1793 records, not selected for constructing dataset.
- **Marks in ENG, MATH, PHY, CHE and SUB4:** These attributes are marks scored in different subjects in high school. The aggregate marks are used for calculating the merit for admission. The only problem in taking this data is the university has started recording these details only since last year, prior to that only aggregate marks were recorded, so data is missing for previous batches. These attributes are thus not taken. However, in extension of the study next year, these can be considered.
- **School Type, address and name:** School type indicates whether the school is Govt. owned or private. All the students come from private school. 100% of the records contain same value and attribute will not contribute in prediction, thus it is not considered. School Name and address, only 12% schools are common, rest are all different. These attributes are discarded.
- **Sponsor Name and sponsor address:** Few seats are reserved for students who are NRI sponsored, the name and address are unique and these attributes are discarded.

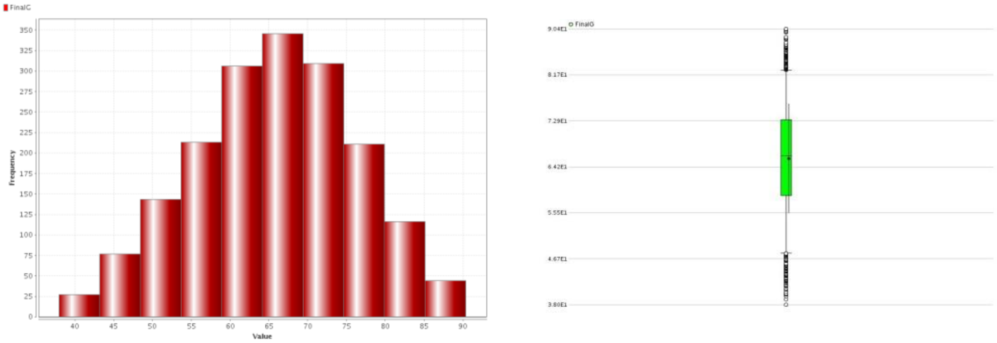
Dean Office related data

- **Grades in different modules and aggregate CGPA, SGPAs:** Aggregate SGPA (semester grade point average) and CGPA's (cumulative grade point average) in different semesters are kept. The final CGPA in degree is used a label. The final CGPA is converted into percentage using following formula:

(CGPA-0.75)*10

This was determined and a final-Percentage attribute was created.

Figure6. Distribution of final marks obtained in degree



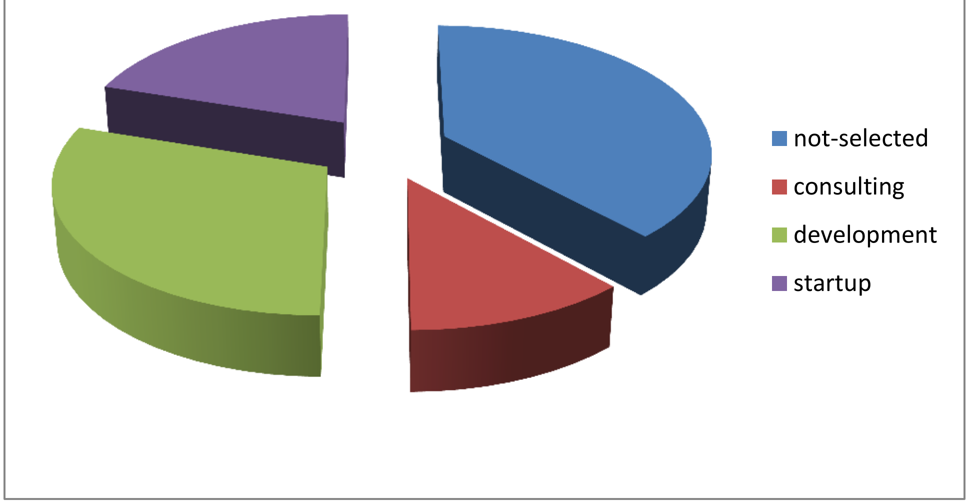
The distribution of final-percentage is almost normal with a skewness of -0.2. The mean value is 65.8+-10.4. The median is 58.6. A label attribute called division is created using this attribute which is explained in section 5.

School of Professional Attachment – dataset

In this office following details are available. (Only those attributes are mentioned which are not available in other offices)

- **CompanySel:-** this attribute contains the name of the company that visits the campus. There are 300 odd companies which have visited us in last three years. The name will not play significant role in prediction and it is discarded. However, label attribute is generated using this attribute. During the first iteration, a label called placement status was constructed where the companies were categorized into three types namely – development, consulting and start-up. The forth value of this multi-class label was not-selected. The distribution of values in this class label is shown in figure 7.

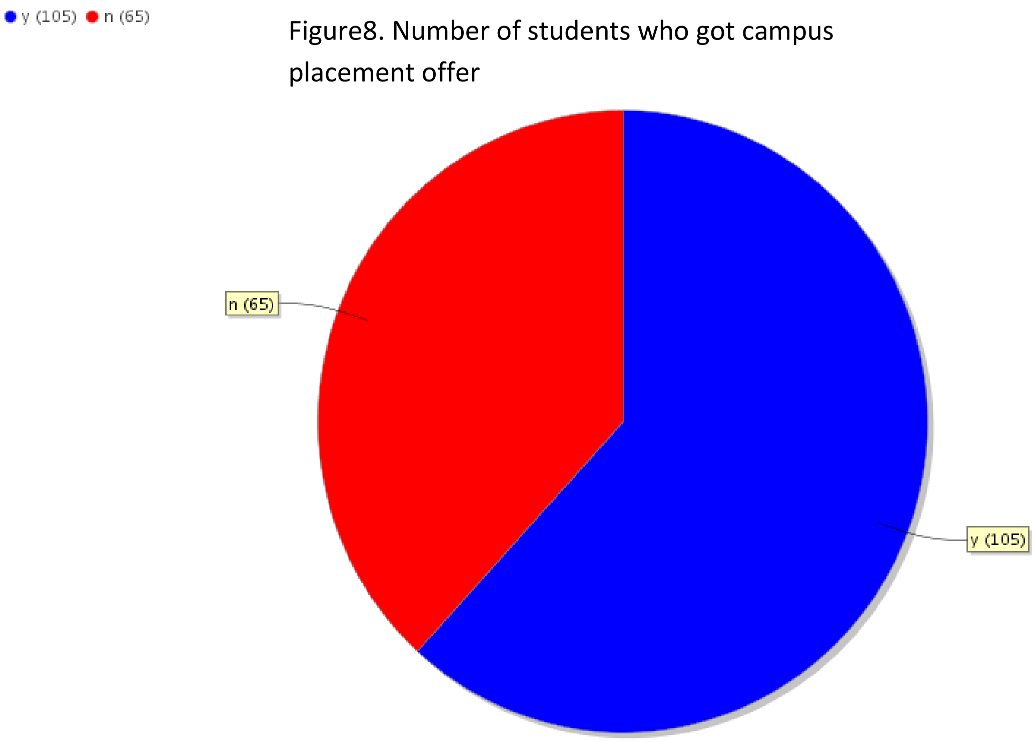
Figure7. Types of companies visited



However, none of the classification algorithm was able to exceed the accuracy of approximately 56% with this label. This means with given set of evaluations done at University it is difficult to predict whether a student will go to a development, or consulting or Startup Company. To simplify this, another label attribute was created which is explained as follows.

A binomial attribute called “Placed” is generated using CompanySel attribute. Missing values were replaced with “n” (indicates that student was not successful in securing a job) and rest all values are replaced with “y” (indicates that student was successful in securing a job). Row duplication will be tried to address the class imbalance.

Figure8. Number of students who got campus placement offer



- **CTC:** is the package offered by the company. Only two different companies offer high starting package rest all offer same starting salary. Thus, attribute has less variability and will not contribute in classification. This attribute was thus discarded.
- **Tenth Marks:** Most of the companies visiting the campus have a minimum eligibility criterion of having 60% marks in 10th, 12th and engineering. This is why 10th marks are also recorded in this office. Important point to note here is that this is not recorded at the time of admission but later during third year before the start of the campus season. The mean value of 10th marks is 84.6+-6.2. The histogram is shown as follows:

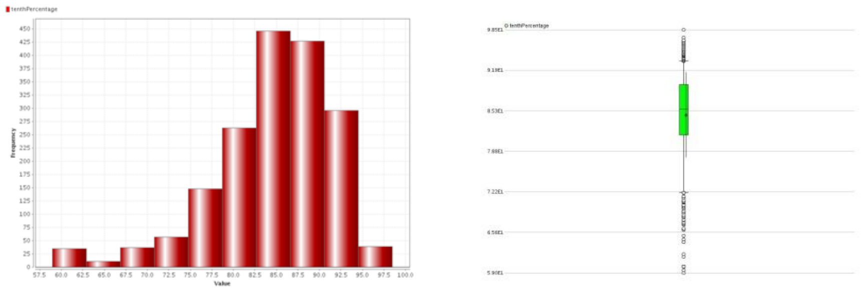


Figure9. Marks scored in 10th class (two years prior to 12th class)

This distribution is slightly skewed towards the right which is also evident from skewness value of -1.2. This indicated that greater number of students have got better than mean scores.

Department of Computer Science & Engineering

The department of computer science has started a new activity of conducting pre-placement interviews with the help of industry experts. A team of experts from industry is called to conduct the interviews prior to actual campus season and rate the student on two parameters "attitude & skills" and "technical-knowledge". This activity started one year back and results for one batch are available. These attributes are used for prediction of success in campus placement of computer science students.

- **Marks in Attitude & Skills through PPI:**

The mean is 50.6+-16.09, median=52.5, skewness=-0.06, kurtosis=-0.2. The histogram is

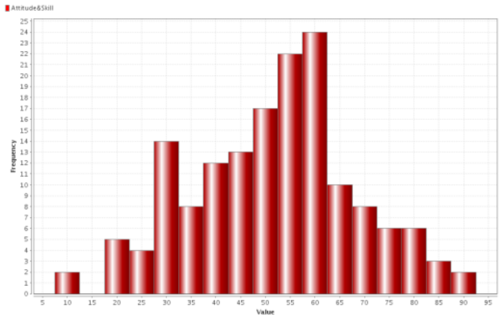


Figure10. Distribution of marks in attitude&skill parameter

- **Marks in technical knowledge through PPI:**

The mean is 41+-16.7, median is 40, skewness=0.55, kutosis=0.2. The histogram is

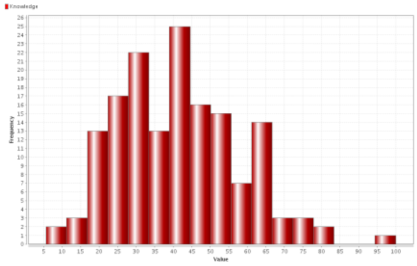


Figure11. Distribution of marks in technical knowledge parameter

The two histograms plotted together:

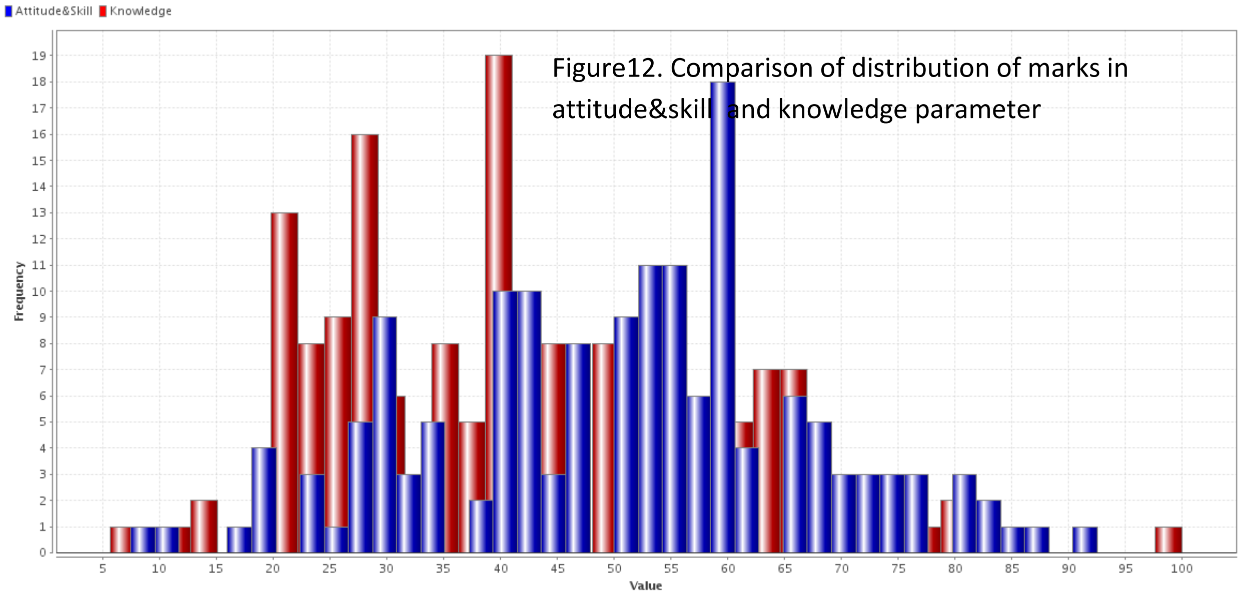


Figure12. Comparison of distribution of marks in attitude&skill and knowledge parameter

This clearly shows attitudes and skills are better than technical knowledge.

Averages of percentage-marks, attitude&skills and knowledge for placed and not placed students were compared. Substantial differences exist in the two categories. Results are shown in figure 13.

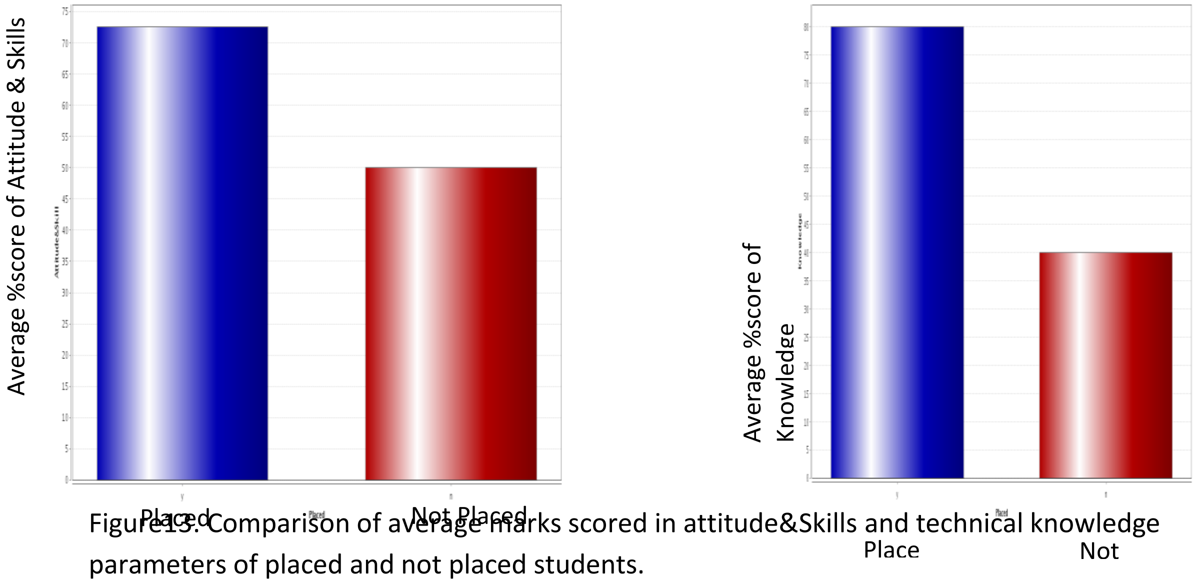


Figure13. Comparison of average marks scored in attitude&Skills and technical knowledge parameters of placed and not placed students.

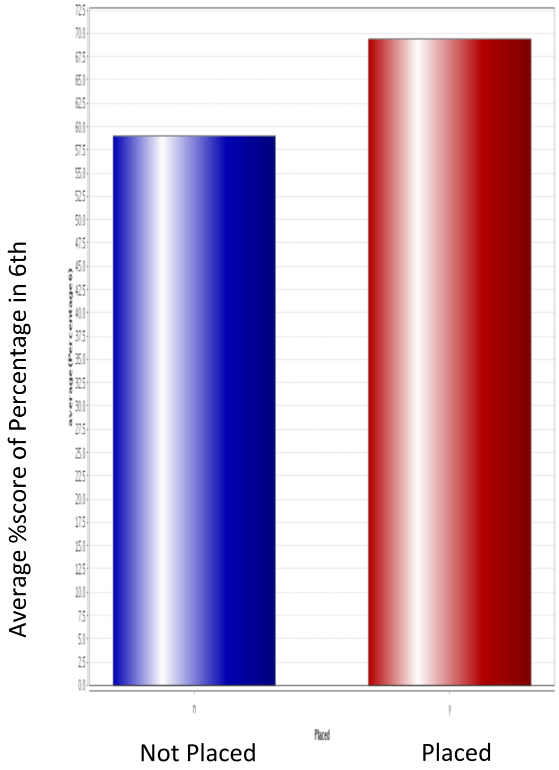


Figure14. Comparison of average final percentage of placed and not placed students.

The parallel plot also shows the discriminatory capability of these measures

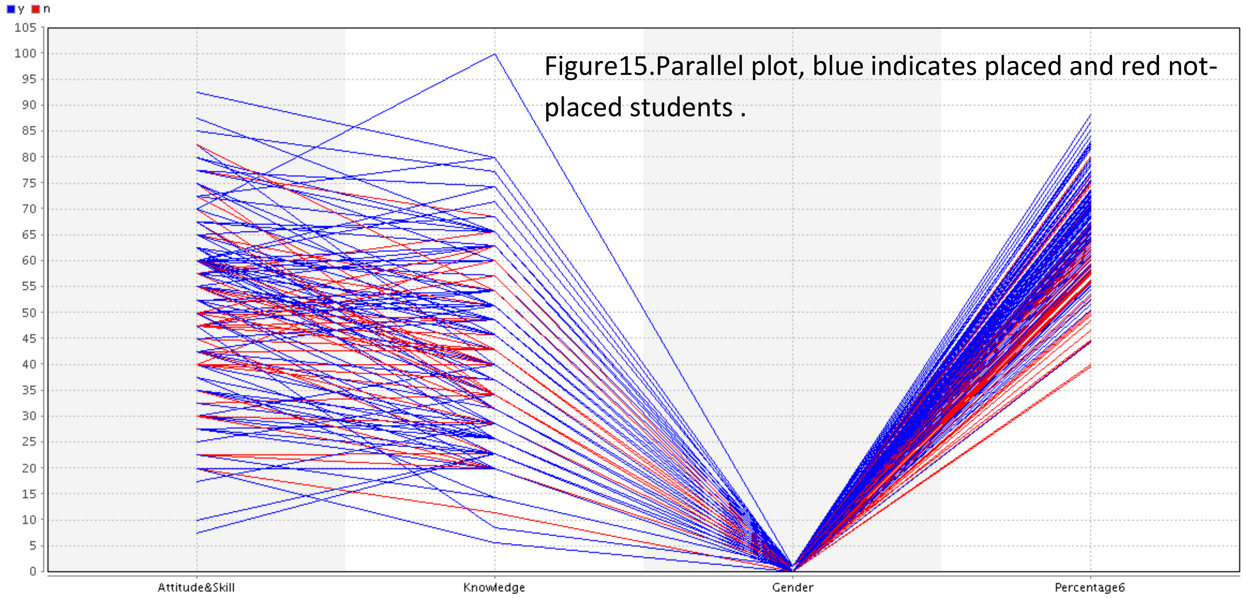


Figure15.Parallel plot, blue indicates placed and red not-placed students .

The main attribute for discrimination seems to be percentage marks obtained in 6th semester (before the start of the campus season).

To examine how campus placement success is different in males and females, Pareto plots were drawn. These are shown in figure 16.

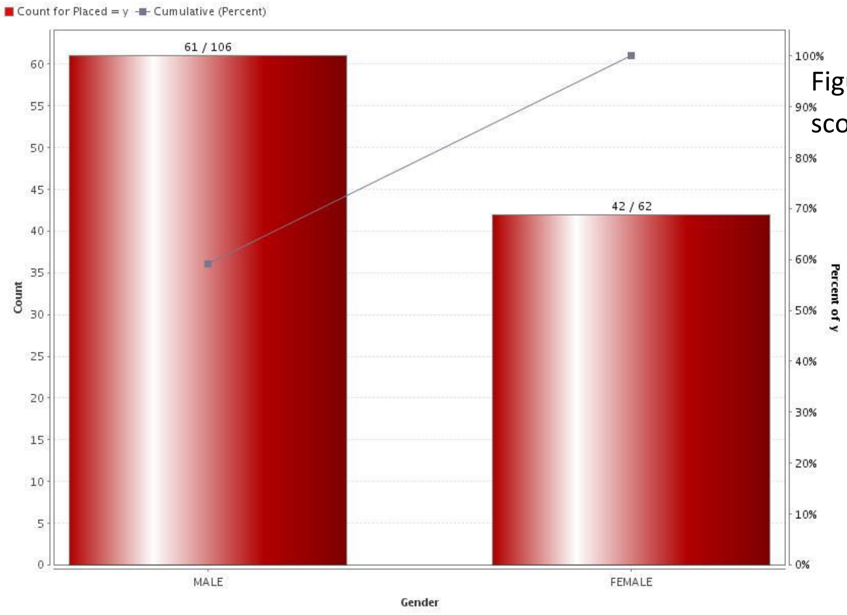


Figure16. Comparison of placements scored by males and females.

The graph here shows that 67% of the females are placed whereas only 58% males are placed. Thus, attribute "Gender" might be helpful in classification model which is trying to predict placement success of an undergraduate student.

Thus two separate datasets were created, one for predicting the academic success having 1793 records and other for predicting the placement success having 173 records. In next section, which is about data preparation, data quality issues as well as data transformation for generating new attributes are discussed.

6. Data Preparation:

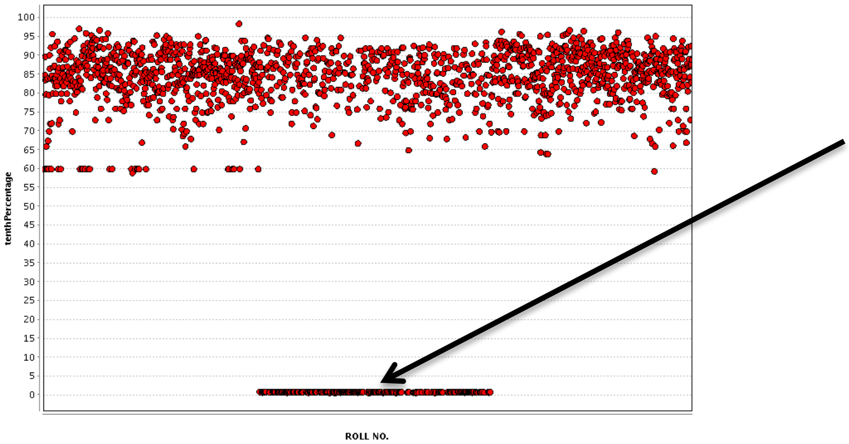
As discussed in previous section, for predicting the academic success, following attributes were considered for classification

Table 2 Data quality issues are visible in the table

Name	Type	Statistics	Range	Missing
Branch	polynomial	mode = CSE (442), least = AU (27)	CIVIL (44), CSE (442), ECE (431), EEE (82), IT (290), ME (320), CV (101), AU (27), EI (55)	0
tenthPercentage	numeric	avg = 73.267 +/- 29.325	[0.616 ; 98.500]	33
twethPercentage	numeric	avg = 68.959 +/- 28.403	[0.650 ; 95.200]	12
Gender	polynomial	mode = MALE (1280), least = MALE (1)	FEMALE (492), MALE (1280), MALE (3), MALE (1)	16
FinalPercentage	real	avg = 46.930 +/- 27.821	[5.000 ; 90.400]	0

From this table, following issues about data quality were observed

4.1 Tenth-Percentage (10^{th}) and Twelfth Percentage (12^{th}): the minimum values shown here were 0.616 and 0.650, this indicates that these values not in proper format. The excel sheet data format of these values were different for different set of cells/records. In some the records, this was percent format, in others it was number format. To examine this further, these were plotted.



These numbers were not in proper format and needs multiplication by 100.

Figure17. Outlier detection and correction in tenthPercentage.

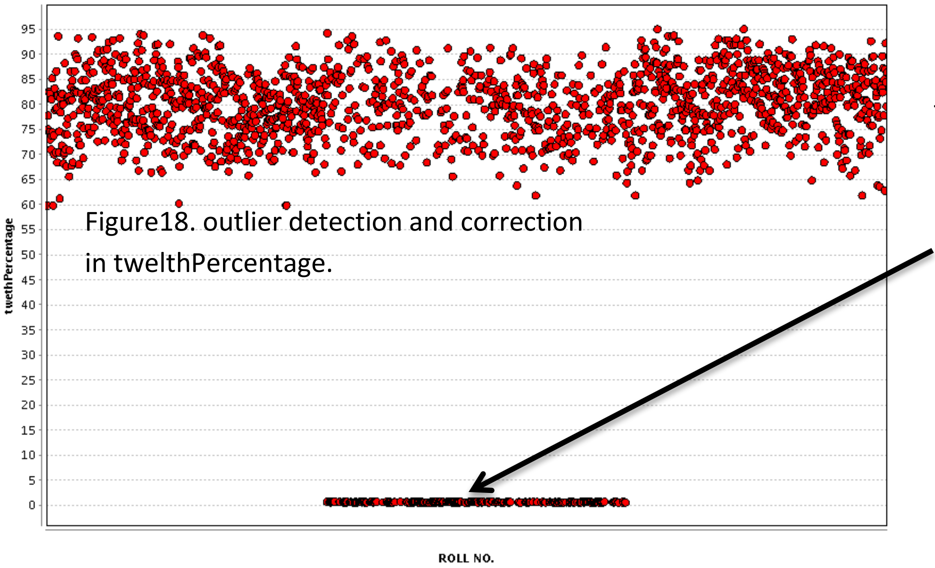
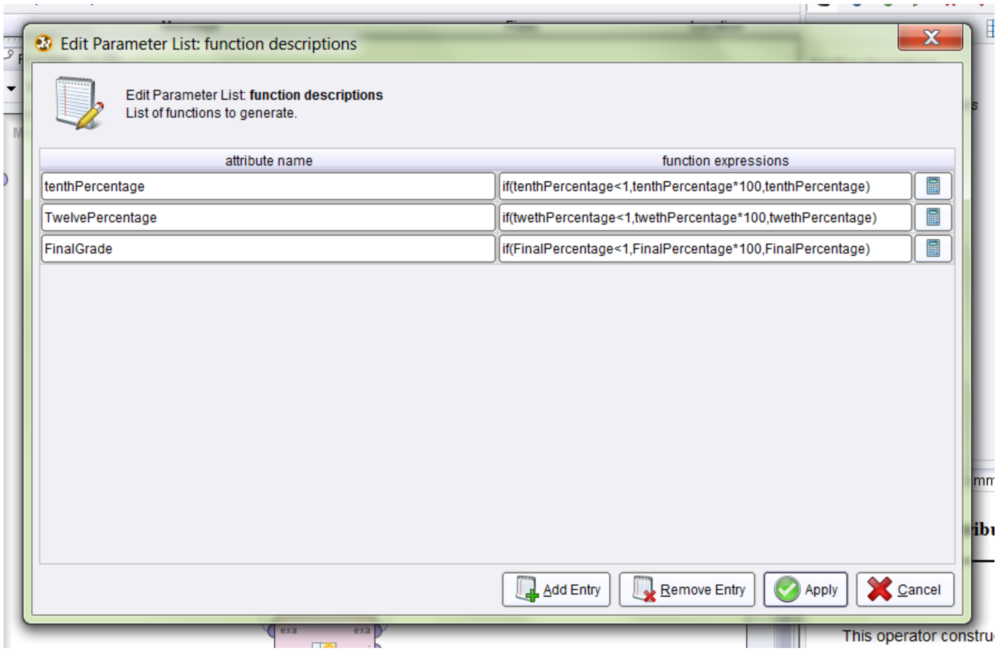


Figure18. outlier detection and correction in twelthPercentage.

These numbers were not in proper format and needs multiplication by 100.

This was done using generate attribute operator in RapidMiner.



Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
tenthPercentage	if(tenthPercentage<1,tenthPercentage*100,tenthPercentage)
TwelvePercentage	if(twethPercentage<1,twethPercentage*100,twethPercentage)
FinalGrade	if(FinalPercentage<1,FinalPercentage*100,FinalPercentage)

Figure19. condition used in generate attribute operator of RapidMiner

Similarly, the FinalPercentage attribute had some data quality issues. The University follows a 10-point grading system and percentage is calculated using following formula (CGPA-0.75)*10. In some of the excel sheets the formula was not copied and instead of percentage marks, the CGPA was provided. This is evident from figure 20.

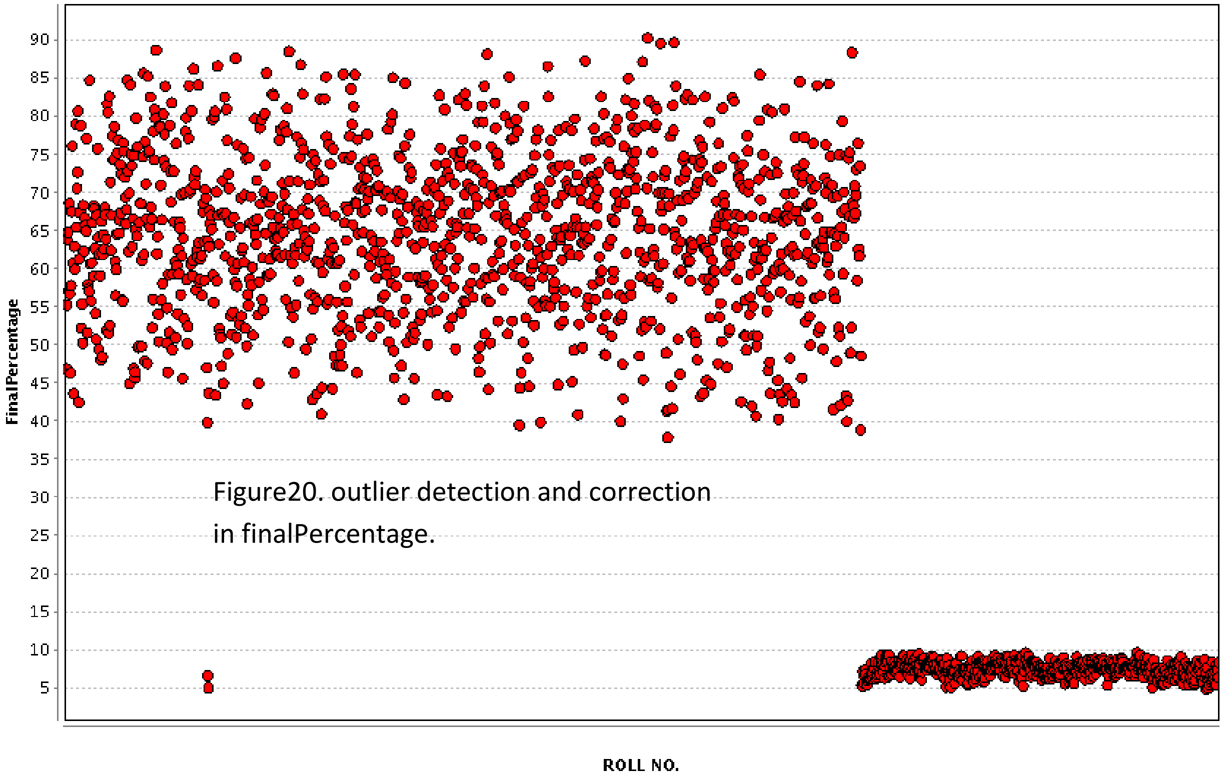


Figure20. outlier detection and correction in finalPercentage.

This was again done using generate attribute operator with function if(FinalGrade<10,(FinalGrade-0.75)*10,FinalGrade)

4.2 In gender attribute, some of the values MALE were with the spaces, and RapidMiner was taking it as different value. The spaces were removed.

4.3 Missing values: 33 and 12 values out of 1793 were missing in tenthPercentage and twethPercentage attributes. Since the mean is a good predictor of the population, the missing values are replaced with averages.

The three attributes after correction are:

Table 3 Numeric attributes after outlier detection and correction

	Attribute	Mean	Std-dev	Variance	Skewness	Kurtosis	Range	
	Tenth Percentage	84.6	6.9	48.2	-1.2	2	59%-98.5%	
	Twelfth Percentage	80.3	6.5	42.1	-0.2	-0.3	60%-95.2%	
	Final B.Tech Percentage	65.8	10.4	107.8	-0.2	-0.4	38%-90.4%	

After addressing these quality issues, the relationship between attributes was examined. The averages of marks obtained in 10th, 12th and finally in B.Tech were plotted branch-wise. These are as shown in figure 21.

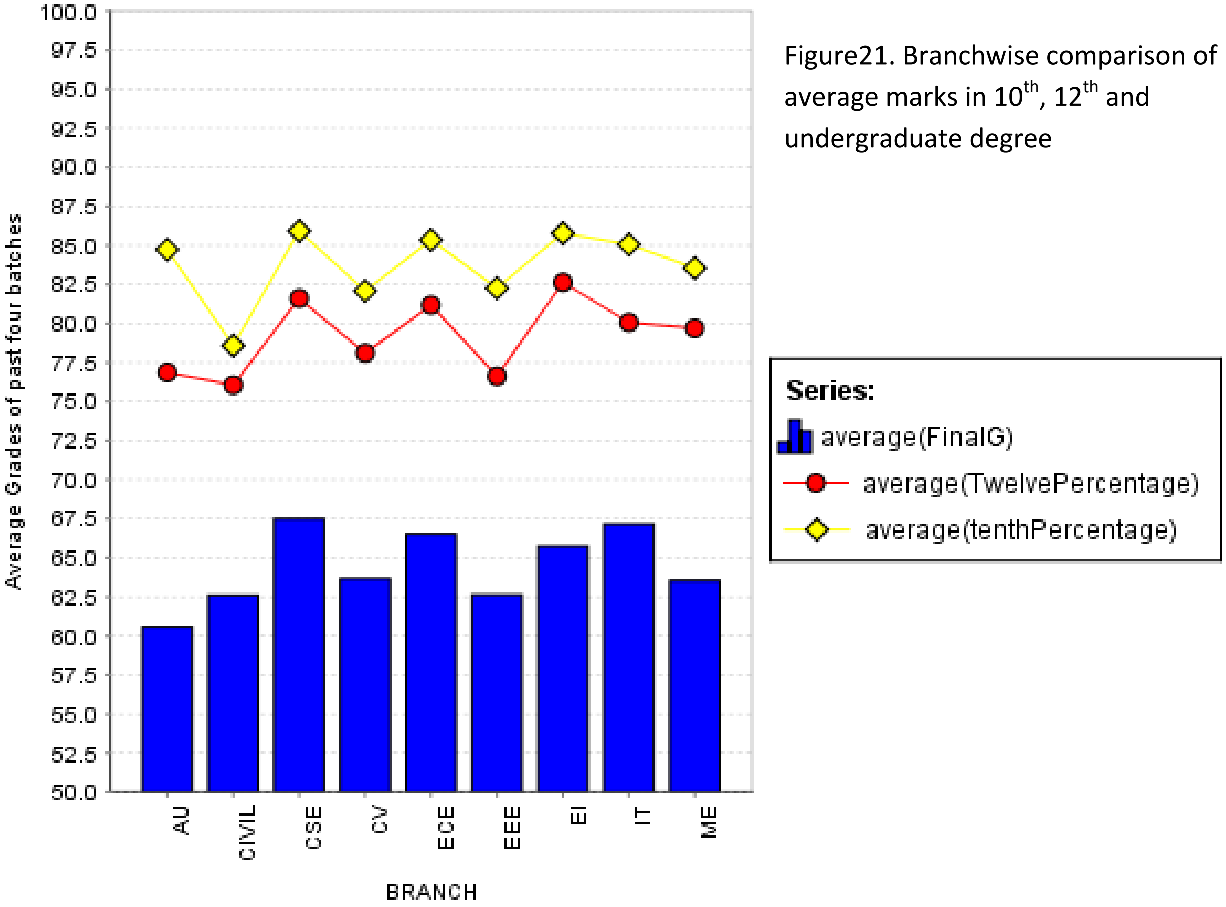


Figure21. Branchwise comparison of average marks in 10th, 12th and undergraduate degree

The figure 21 shows high correlation between performances of 10th, 12th and final-percentage and difference in performance of students across branches is also visible. To investigate the degree of dependence in performance in 10th, 12th and final-grades correlation coefficient was determined. The results are:

Correlation b/w 10th and Final-Percentage → 0.45

Correlation b/w 12th and Final-Percentage → 0.56

Slightly stronger correlation exists between grades of 12th and final-percentage than 10th marks and final-percentage.

Linear Regression was determined and results are as follows

Table 4 Linear regression model

Attribute	coefficient	Std-error	std-coeff	tolerance	t-stat	p-value
10th percentage	0.34433948	0.0323779	0.0279705	0.7613087	10.63503	0
12th percentage	0.72591583	0.0344073	0.0584439	0.7613087	21.09774	0

The high positive value of t-stat shows that these attributes are relevant for predicting the final-percentage (root-mean-squared-error is 0.95).

Equation of model for prediction is

0.344 *tenthPercentage + 0.726 *twethPercentage – 21.664=0 is the equation of the model

4.4 Data Transformation:

A new polynomial attribute called “Division” was constructed using generate attribute operator of RapidMiner.

if(FinalG>=80,"distinction",if(FinalG>=60,"FirstClass","SecondClass"))

The conversion formula is as per the norms of the Univeristy. Though this attribute is not available in electronic records but it is used while providing the degree to the student. The degree specifies whether the student has received the degree with Distinction or FirstClass or Secondclass. The distribution of values in division attribute is shown in figure 22.

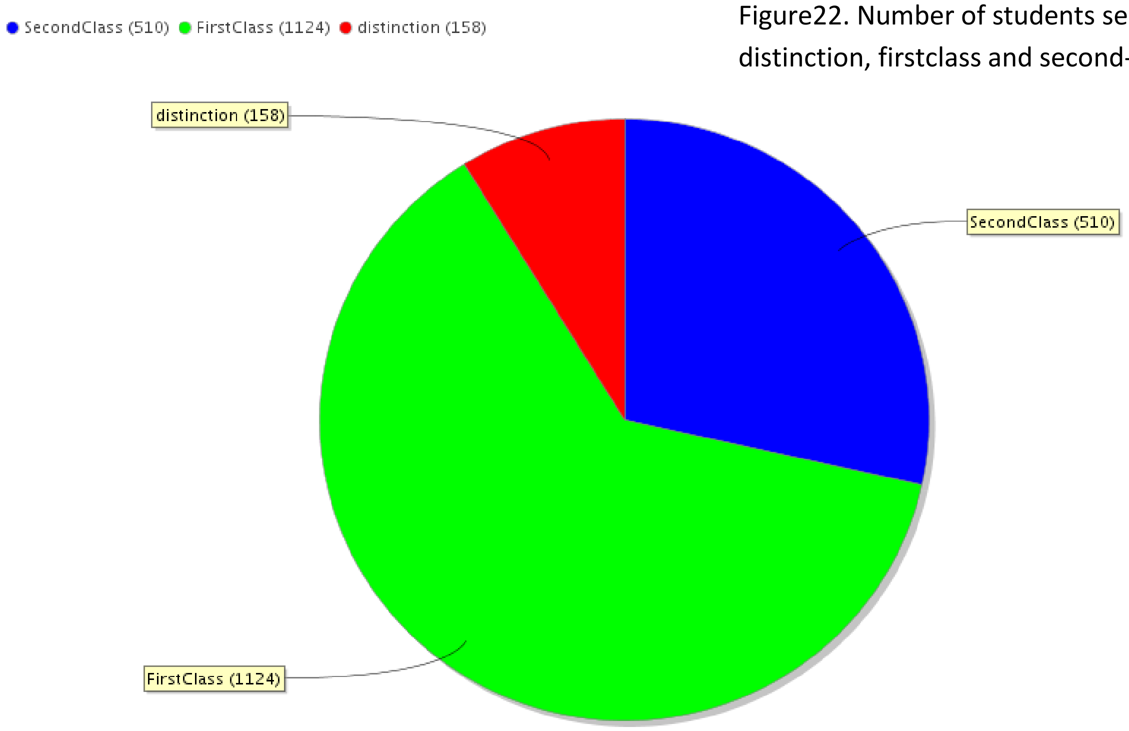


Figure22. Number of students securing distinction, firstclass and second-class

The final dataset used for predicting the academic success is as follows (1793 records):

Table 5 Dataset for predicting the academic performance

Type	AttributeName	Type2	Mode, Median	Ranges	Missing
Regular	BRANCH	polynomial	mode = CSE (442), least = AU (27)	CIVIL (44), CSE (442), ECE (431), EEE (82), IT (290), ME (320), CV (101), AU (27), EI (55)	0
Regular	Gender	polynomial	mode = MALE (1296), least = MALE (1)	FEMALE (492), MALE (1296), MALE (3), MALE (1)	0
Regular	tenthPercentage	real	avg = 84.648 +/- 6.876	[59.000 ; 98.500]	0
Regular	TwelvePercentage	real	avg = 80.316 +/- 6.466	[60.000 ; 95.200]	0
Regular	FinalG	real	avg = 65.786 +/-	[38.000 ; 90.400]	0

			10.384		
Regular	Division	nominal	mode = FirstClass (1124), least = distinction (158)	SecondClass (510), FirstClass (1124), distinction (158)	0

For predicting the placement success the dataset constructed for academic success was not appropriate as the accuracies of models was poor, it was not even touching 60%. This was discussed with the strategic planning team of the University which suggested doing this analysis using the performances of students in pre-placement interviews. However, these interviews were done for a small subset of students in department of computer science that too for only one batch. So a separate dataset was constructed for meeting this objective. Here the marks of 10th and 12th were not contributing significantly in predicting the model, hence these were not included for classification.

Following dataset was constructed: Here as mentioned in previous section, the attribute Placed is constructed from already existing attribute CompanySel. The missing values of company selected were replaced by “n” and rest all were replaced with “y” (173 records)

Table 6 Dataset for predicting the placement success

Name	Type of attribute	Data-type	Mean/Mode	Range	Missing
Placed	label	nominal	mode = y (105), least = n (65)	y (105), n (65)	0
Attitude&Skill	regular	numeric	avg = 50.593 +/- 16.690	[7.500 ; 92.500]	14
Knowledge	regular	numeric	avg = 41.062 +/- 16.799	[5.714 ; 100.000]	14
Gender	regular	binomial	mode = MALE (106), least = FEMALE (62)	MALE (106), FEMALE (62)	2

Percentage6	regular	numeric	avg = 65.390 +/- 10.174	[39.600 ; 88.300]	0

7. Modelling

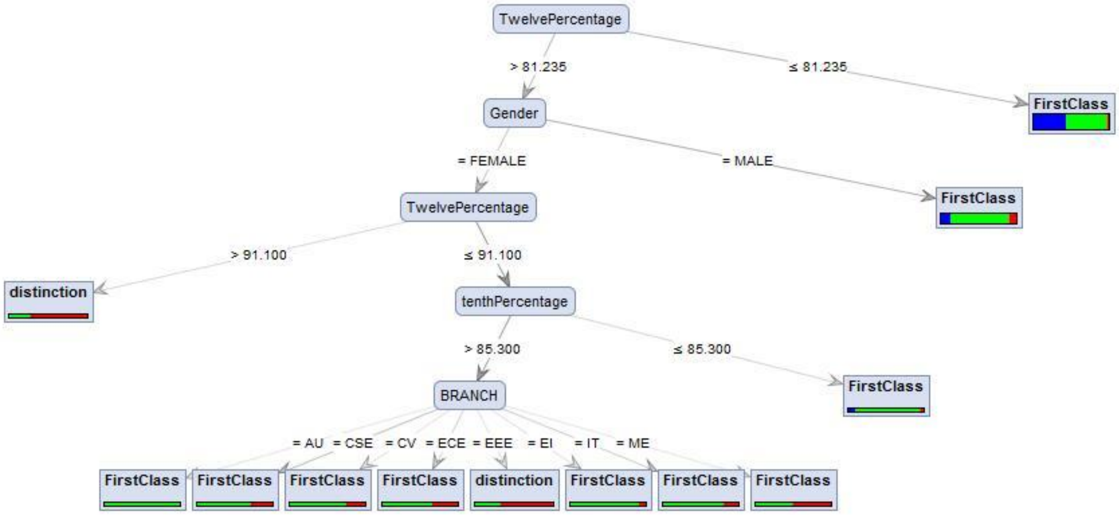
5.1 For predicting the academic success of the students

Initially all the five predictor variables (three numeric type variables, one binomial and one polynomial) were taken as input to three classification algorithms: decision tree, k-NN and Naïve-Bayes. The output variable was final performance that can have any of the three values – distinction, first-class and second-class.

After this, the polynomial and binomial predictor attributes were removed and Neural Network was developed with numeric predictor attributes.

5.1.1 Decision Tree

Figure23. Decision tree to predict academic success



Confusion Matrix and performance

Table 7 Confusion matrix for decision tree

Column1	true SecondClass	true FirstClass	true distinction	class precision
pred. SecondClass	1	1	0	50.00%
pred. FirstClass	509	1108	129	63.46%
pred. distinction	0	15	29	65.91%
class recall	0.20%	98.58%	18.35%	

Accuracy: 63.50% +/- 1.03% (mikro: 63.50%)

This is after optimizing all the parameters. It is evident from the confusion matrix that class recall of secondClass is very low, and distinction is also low. The accuracy is largely because of FirstClass. Since firstClass is more frequent value, the accuracy is biased. Hence, one can conclude the model is not a very good one. Even introducing the duplicate records did not help improve accuracy to great extent.

5.1.2 k-Nearest Neighborhood

With k=30, accuracy was highest accuracy: 67.64% +/- 3.47% (mikro: 67.63%)

Table 8 confusion matrix for k-NN

Column1	true SecondClass	true FirstClass	true distinction	class precision
pred. SecondClass	238	160	6	58.91%
pred. FirstClass	272	953	131	70.28%
pred. distinction	0	11	21	65.62%
class recall	46.67%	84.79%	13.29%	

The class recall of secondClass is improved and distinction and first class is reduced.

5.1.3 Naïve Bayes

Table 9 confusion matrix for Naive Bayes

	true SecondClass	true FirstClass	true distinction	class precision
pred. SecondClass	228	146	3	60.48%
pred. FirstClass	282	908	93	70.77%
pred. distinction	0	70	62	46.97%
class recall	44.71%	80.78%	39.24%	

accuracy: 66.85% +/- 3.43% (mikro: 66.85%)

5.1.4 Neural Network: the binomial values of branch and gender was removed and neural network was also tried.

Table 10 Confusion matrix for Neural Network

Column1	true SecondClass	true FirstClass	true distinction	class precision
pred. SecondClass	190	135	3	57.93%
pred. FirstClass	320	975	133	68.28%
pred. distinction	0	14	22	61.11%
class recall	37.25%	86.74%	13.92%	

Accuracy: 66.24% +/- 3.49% (mikro: 66.24%). Accuracy did not improve on adding the duplicate value for distinction. This clearly indicates that even without using the Gender and Branch, this model is able to match the accuracy of Decision Tree, k-NN and Naïve-Bayes.

These results tell us that based on the input parameters considered at the time of admission, the prediction about final grades can be made but with maximum 67% accuracy.

To extend this further and understand other factors during iterations of business understanding and modelling (as stated in CRISP-DM guidelines), it was decided to include SGPA of first semester as an input parameter in predicting the academic success of a student. This will help the University understand that how well a student has taken up any programme, might play a crucial role in predicting his/her success in the programme. Hence, attribute SGPA1 was added to the input of the dataset. The mean value of this attribute is 6.98+-0.96, Range [4.33 to 9.66] and none of the values were missing. Figure24 shows strong relationship between final-percentage and first semester SGPA.

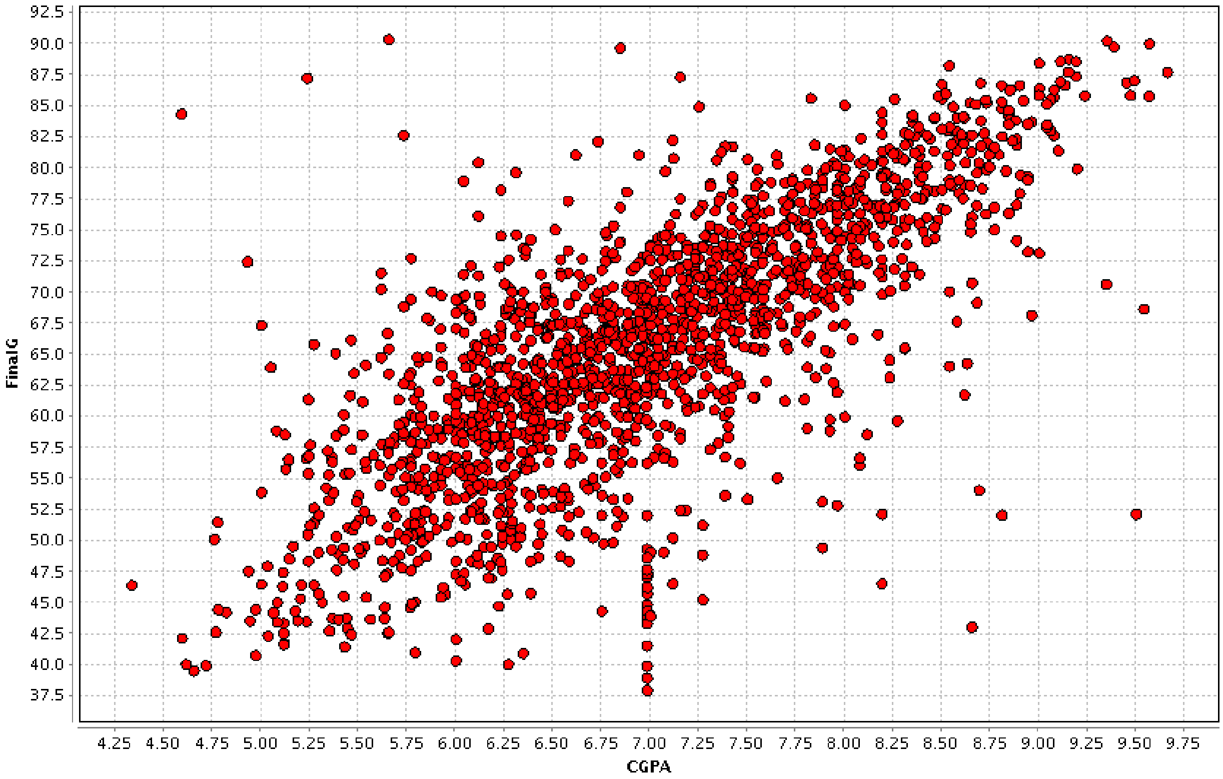
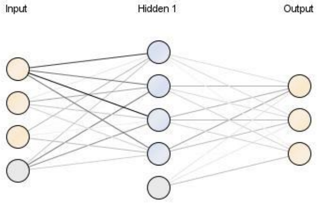


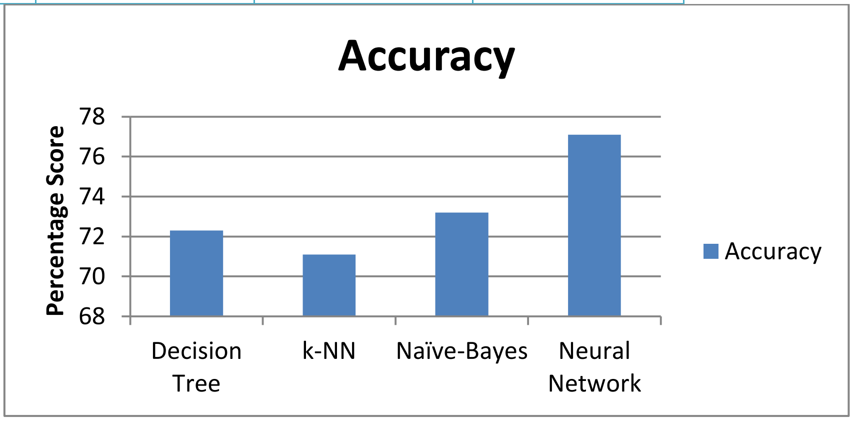
Figure24. Plot of CGPA or SGPA in first semester and final-percentage.

Amongst all the classification algorithms, highest accuracy of 77.3% was achieved by Neural Network (which is higher than all the three classification methods as shown in Figure 3). This change was mainly visible in improvement in prediction accuracy of distinction-class.

Table 11 Confusion matrix for Neural Net after including one more predictor variable

	true SecondClass	true FirstClass	true distinction	class precision
pred. SecondClass	316	105	4	74.35%
pred. FirstClass	193	976	63	79.22%
pred. distinction	1	43	91	67.41%
class recall	61.96%	86.83%	57.59%	





Thus, the academic success of students mainly depends on his grades in high school as well as on how well they are adjusted during the first semester which may include many factors like their interest in engineering, gaps between expectations and reality, huge pressure of evaluation in semester based system since the students come from annual system. Different departments are handling students in a different way which is also evident by looking at the branch-wise accuracy of the model (though it is also affected by the number of students in the branch). Branch-wise accuracies are as follows:

Branch CSE=77.37% +/- 3.09% (mikro: 77.38%)

CIVIL=71.50% +/- 18.58% (mikro: 70.45%)

ME=80.62% +/- 9.56% (mikro: 80.62%)

ECE=69.12% +/- 12.73% (mikro: 69.14%)

IT=82.41% +/- 6.44% (mikro: 82.41%)

EI=79.67% +/- 13.45% (mikro: 80.00%)

And accuracies of other models with inclusion of this attribute are as follows:

Accuracy with k-NN is 77.18% +/- 2.83% (mikro: 77.18%)

With Naïve-Bayes is 73.38% +/- 2.06% (mikro: 73.38%)

With decision tree is accuracy: 72.49% +/- 3.64% (mikro: 72.49%)

Business understanding and suggestions: University should introspect and come up with strategic plans like giving first year to enthusiastic faculty members who can create interest of students in engineering, or making the transition from school to college easy for students, or possibly counselling the students to set the expectations right etc. it is also advisable to monitor students engagement and have early detection mechanism to identify the students who are losing interest.

5.2 Predicting the placement success of a student:

It is important to state it again that dataset created for predicting academic success was not suitable for predicting the placement success as the models were giving poor accuracies. New attributes were added for this prediction

5.2.1 Decision Tree

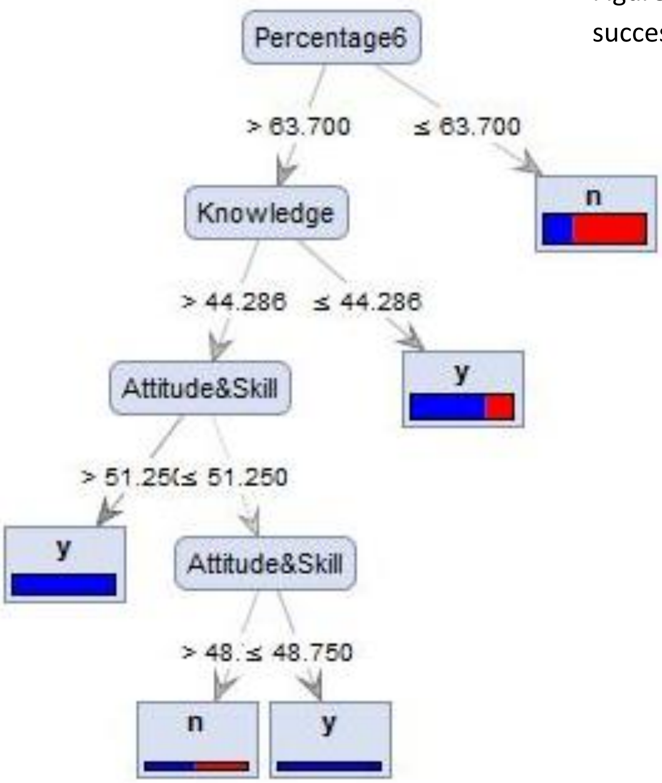


Figure26 Decision Tree for predicting the placement success of a student

Table 12 confusion matrix for decision tree to predict placement success

	true y	true n	class precision
pred. y	24	5	82.76%
pred. n	8	14	63.64%
class recall	75.00%	73.68%	

Accuracy: 74.51%

5.2.2 K-NN

With k=12,

Table 13 confusion matrix for k-NN for placement success

Column1	true y	true n	class precision
pred. y	25	6	80.65%

pred. n	7	13	65.00%
class recall	78.12%	68.42%	

Accuracy: 74.51%

The ROCs of three different models were compared as shown in figure 27

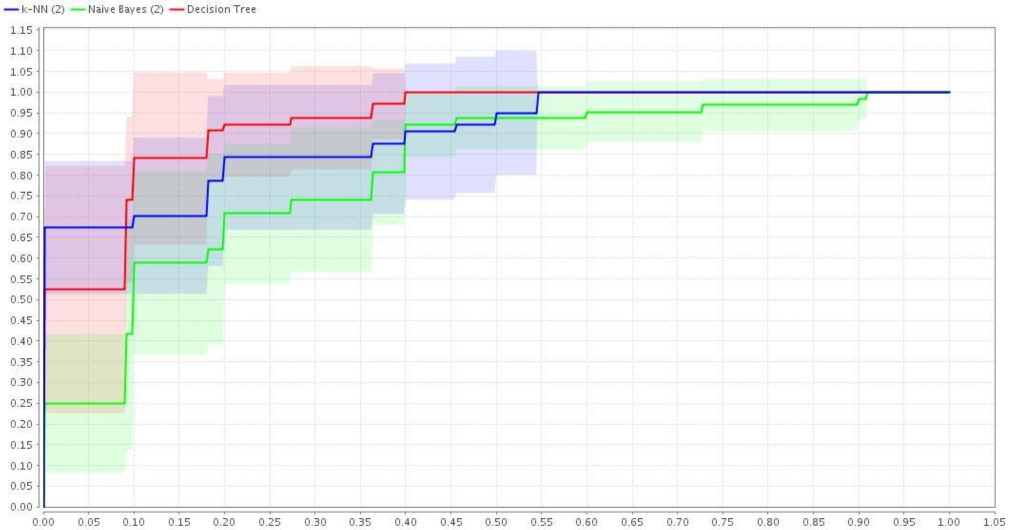


Figure27 ROC comparison of classification models for predicting placement success of students

AUC optimist – decision tree – 0.9
AUC – decision tree – 0.78
AUC optimist – k-NN – 0.9
AUC – k-NN – 0.53
AUC optimist – Naïve-Bayes – 0.74

Decision tree is a suitable model for this prediction.

Business Understanding: Engineering education mainly consists of three things – fundamental knowledge, skills and attitudes. Though strong focus of University is there in teaching fundamentals, but there seems to be gaps in judging the skills parts. It is evident that skills and attitudes are not measurable through existing evaluation methods used by University. This is evident from the fact that it is not possible to predict the placement success of a student through the grades he/she is securing in all types of modules covered throughout four years. University should train/expose the faculty members to new pedagogy and evaluation methods which can address these issues.

6 Conclusion

This main objective of the project was to develop classification models to predict a) academic performance b) placement success of the undergraduate engineering students of a University. Two main research questions which were addressed are: which model gives better accuracy in these predictions and which predictor variables contributed as input in developing the model. The latter part can help the University understand the important factors that lead to students' success.
It was found that neural networks gives best accuracy for prediction the class with a student clears his degree (with distinction or with first class or with second class). For estimating the marks in percentage the regression model can be used whose equation is provided. For predicting the placement success decision tree gave better accuracy which was visible from ROC curves and area under these curves. AUC of decision tree was 0.9 which means that classification algorithm is giving good performance.

In terms of predictor attributes, it was found that high school marks alone were not sufficient in making good prediction about academic progress. Students' performance during first year along with high school marks lead to a good prediction accuracy of 77%. This means it is very crucial for the University to provide good counselling services during first year to make transition easy for students. It is also evident from this analysis that to predict placement success of students of this University their academic performance alone was not sufficient. Scores of interviews conducted by industry experts along with academic performance helped in developing good model with 75% accuracy. This means that there are gap in the methods the University is evaluating the students. Proactive measures can be taken to fill these gaps as this will help in improving the placement success of the students.

List of figures