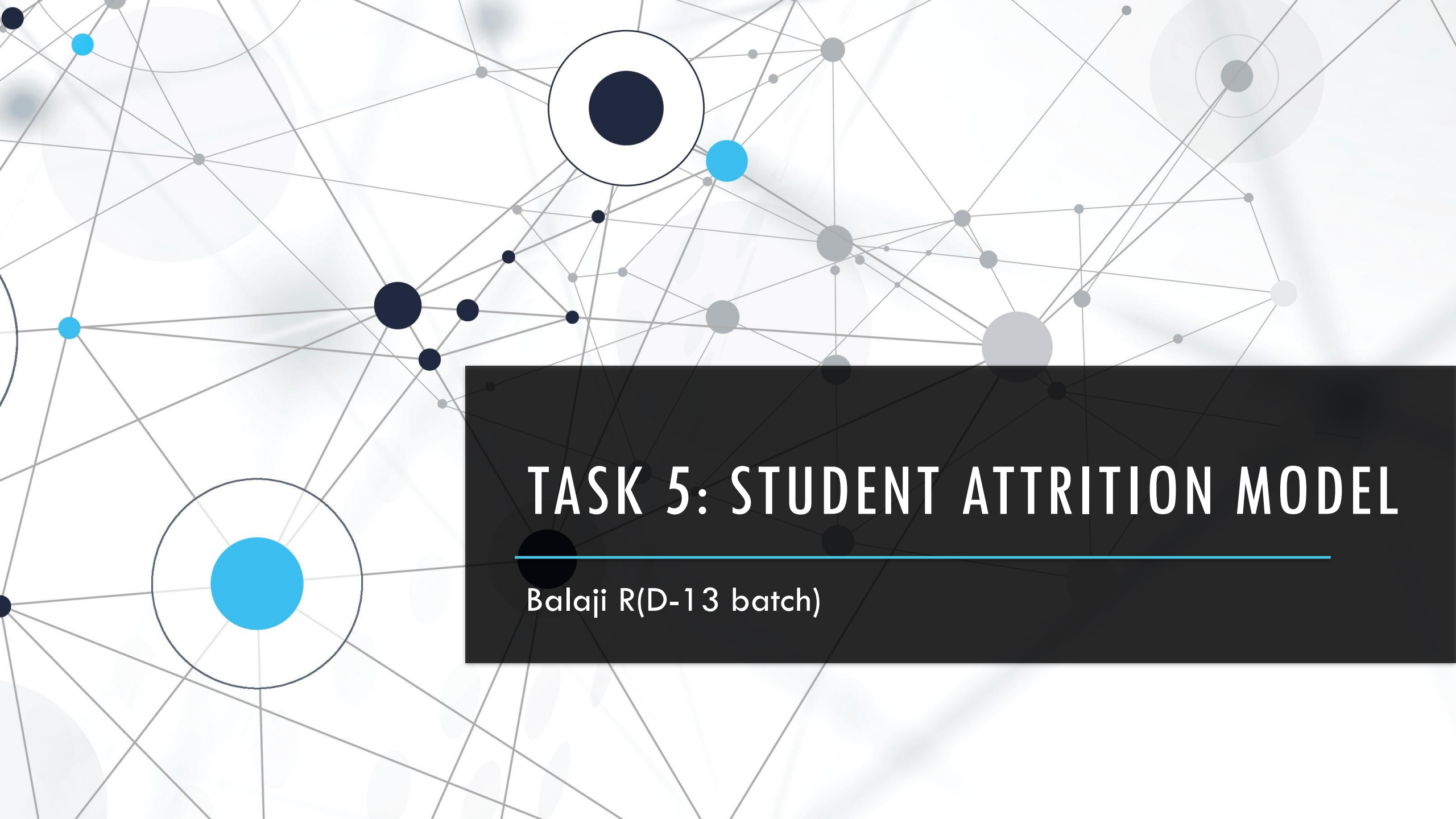
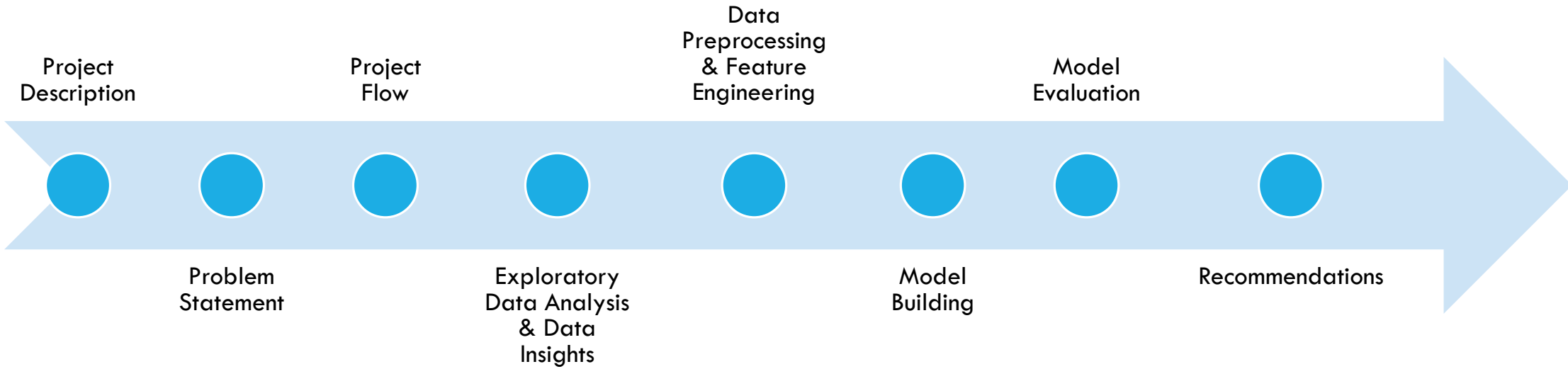


TASK 5: STUDENT ATTRITION MODEL

Balaji R(D-13 batch)

TABLE OF CONTENTS



PROJECT DESCRIPTION

Given the dataset of Clearwater State University's student information, a column contains data about student's early attrition is taken as target and an end-to-end machine learning model is built using 'Random Forest Classification' algorithm which has accuracy of 84%.

PROBLEM STATEMENT

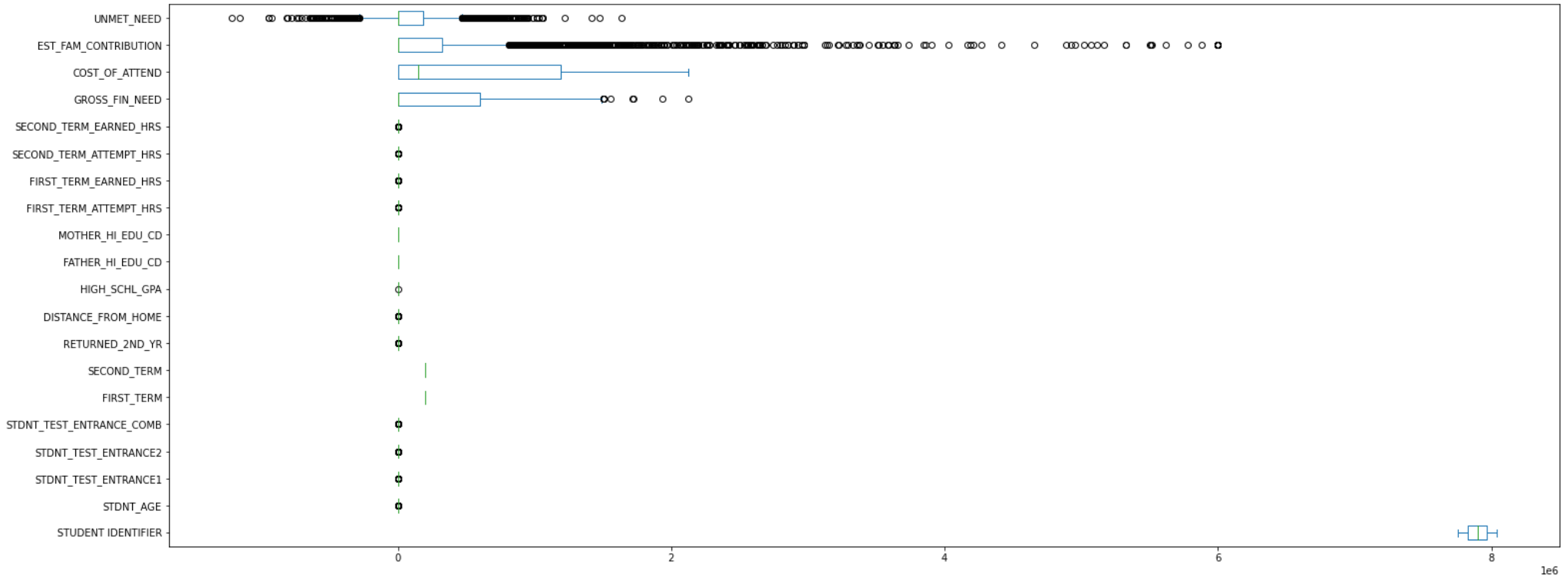
Clearwater State University has collected student data to check the enrolment to the curriculum programs and the percentage of students continuing to the next year. It was observed that many students are opting out from continuing the next year. The rate of attrition was very high.

Now, the goal is to,

- Identify key drivers of early student attrition,
- Build a predictive model to identify students with higher attrition risk,
- Recommend appropriate interventions based on analysis.

PROJECT FLOW

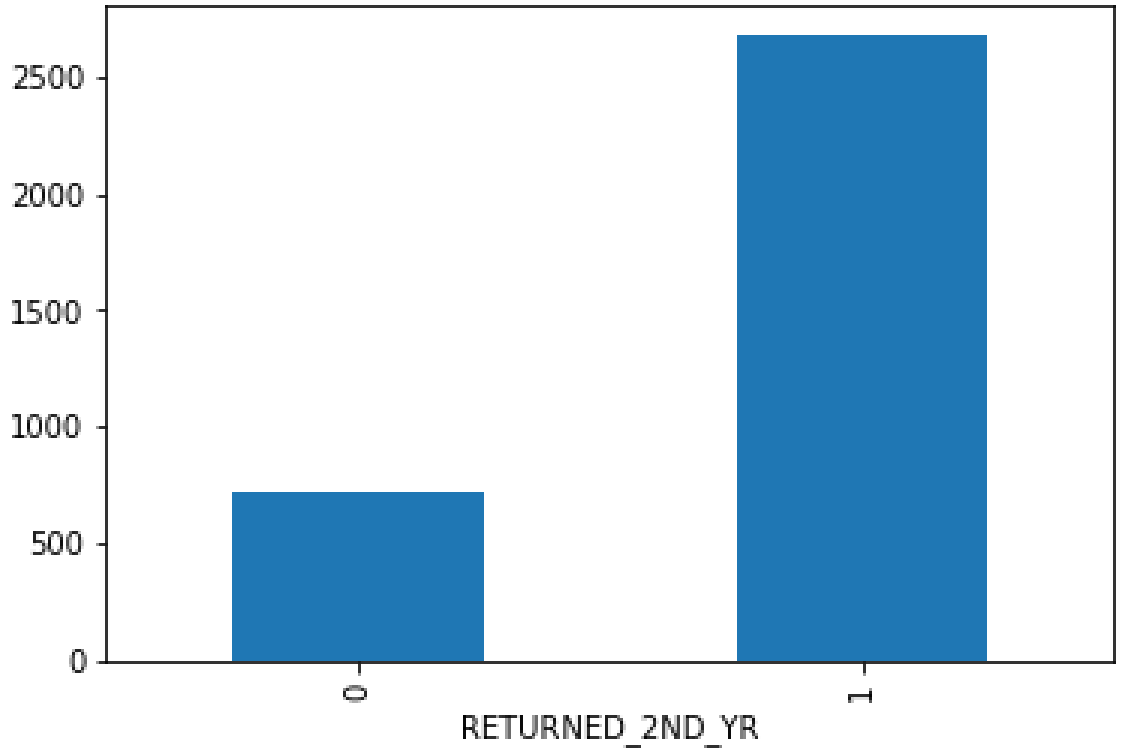
EXPLORATORY DATA ANALYSIS & DATA INSIGHTS



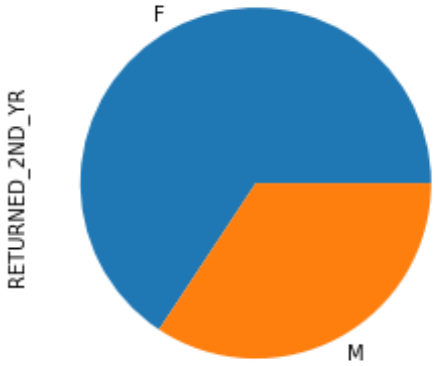
Box plot displaying the outliers in numerical columns.

EXPLORATORY DATA ANALYSIS & DATA INSIGHTS

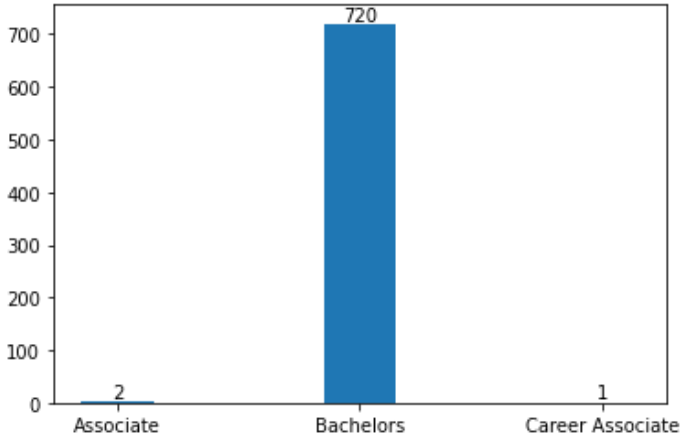
- 78.74% of students returned back to second year.
- 21.26% of students attrited from the university after first year.



EXPLORATORY DATA ANALYSIS & DATA INSIGHTS



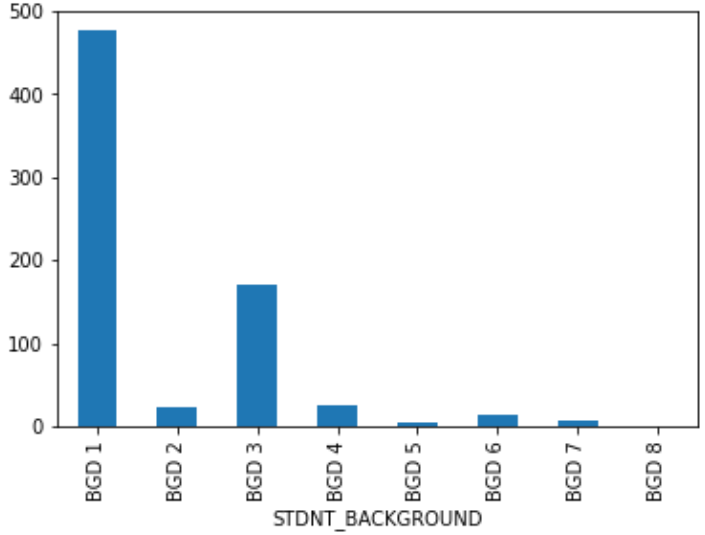
- Female students have more attrition rate than male students.



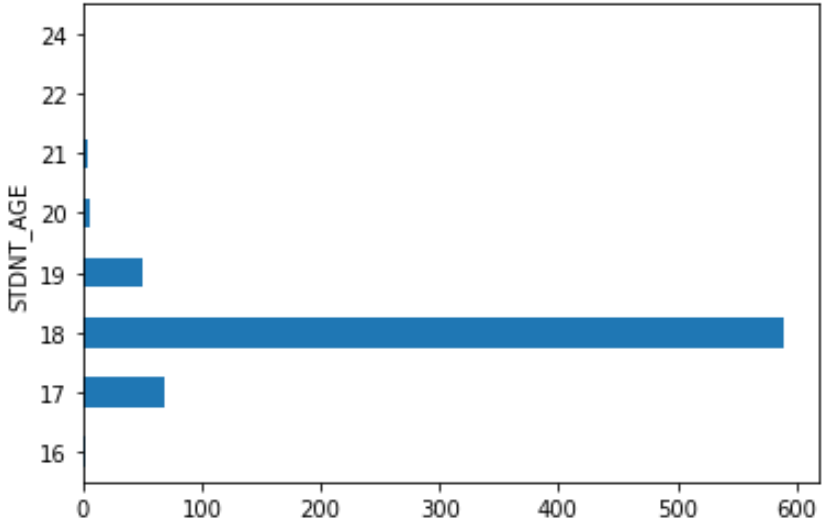
- The students who enrolled for bachelors degree drops out more than others.

EXPLORATORY DATA ANALYSIS & DATA INSIGHTS

- Students from first and third background have comparatively more attrition rate than the students from other backgrounds.

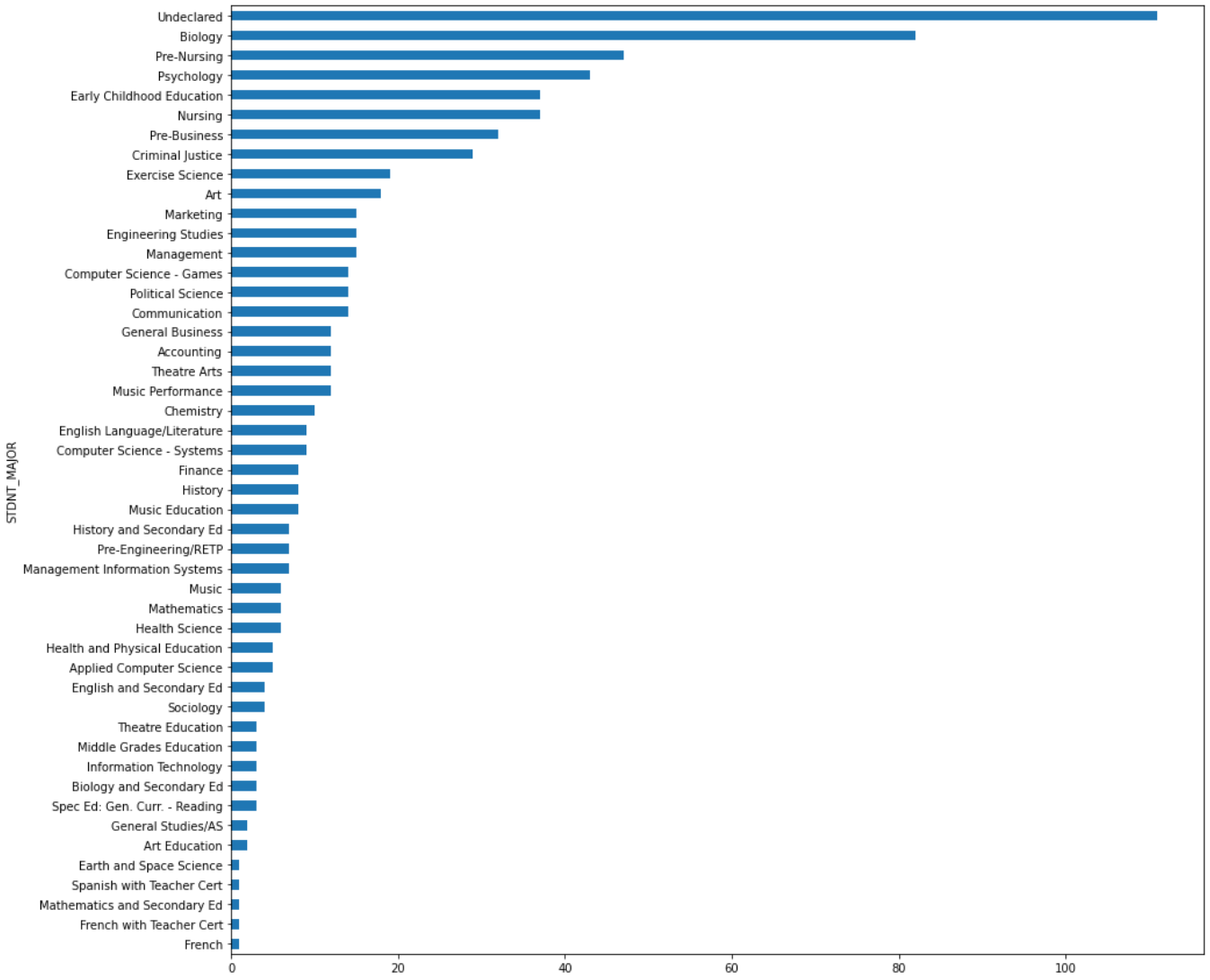


- Students whose age is lesser than or equal to 18 has more attrition.

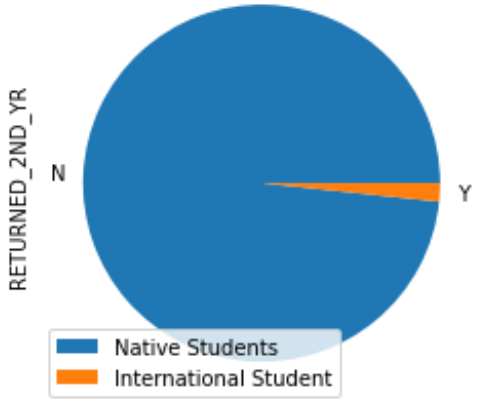


EXPLORATORY DATA ANALYSIS & DATA INSIGHTS

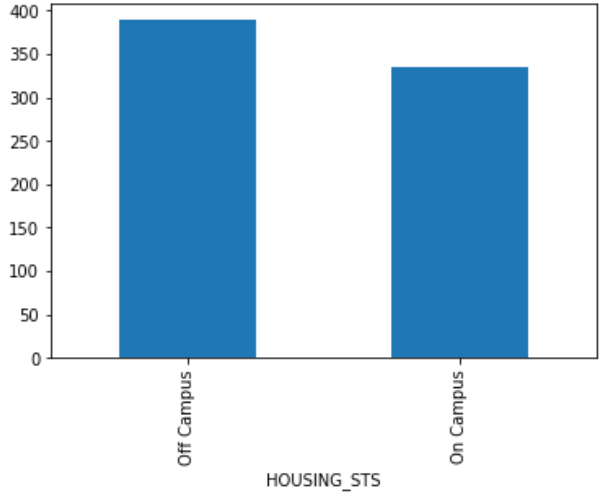
- Students from the departments namely biology, pre-nursing, psychology, early childhood education, nursing have more attrition.



EXPLORATORY DATA ANALYSIS & DATA INSIGHTS



- International students attrits lesser than the native students

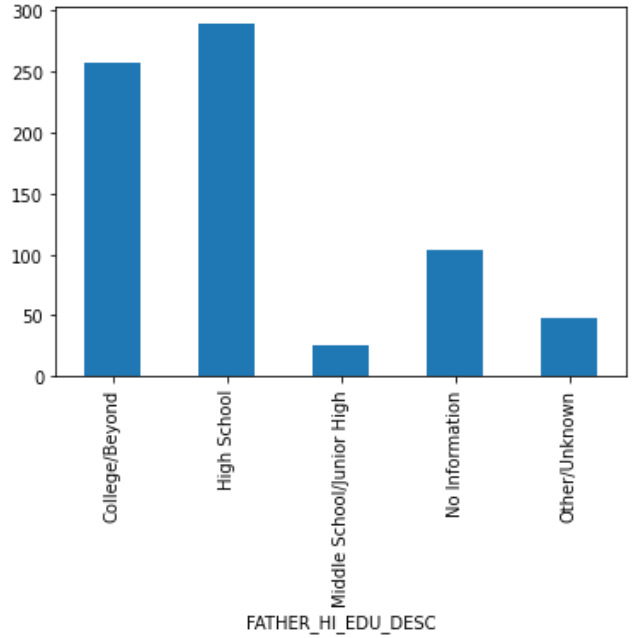


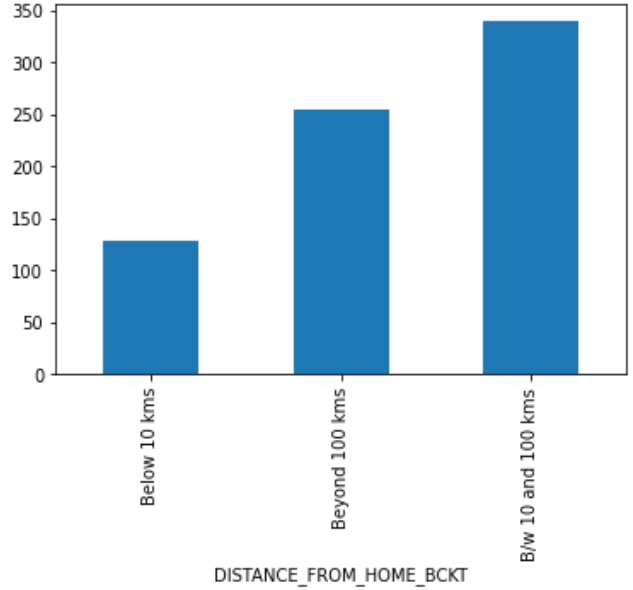
- The students who stays off campus drops out more than the ones who lives on campus.

EXPLORATORY DATA ANALYSIS & DATA INSIGHTS

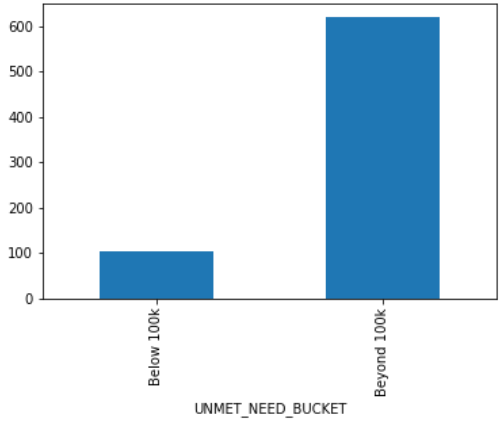
- Students whose father went for high school and college has more attrition than others.

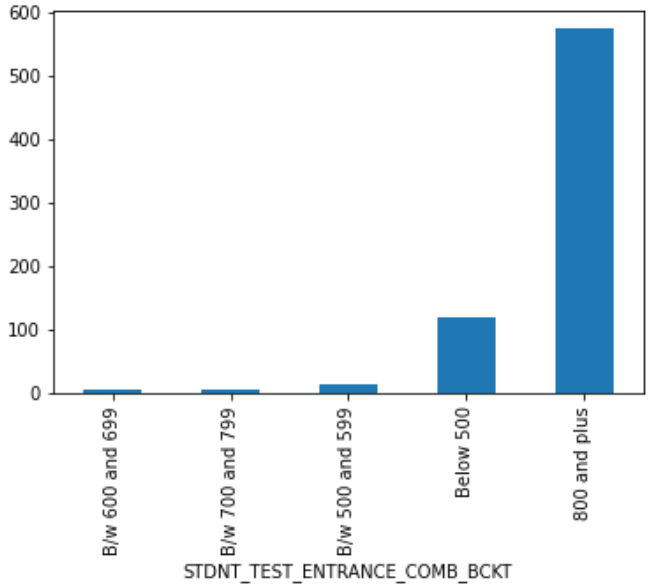
- Students who staying between 10 and 100 kilometers from the university is dropping out more than others.





EXPLORATORY DATA ANALYSIS & DATA INSIGHTS

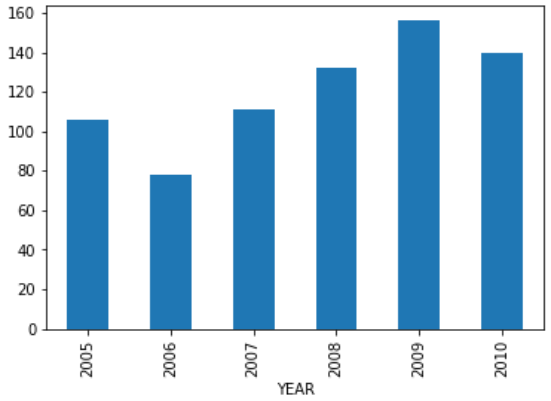


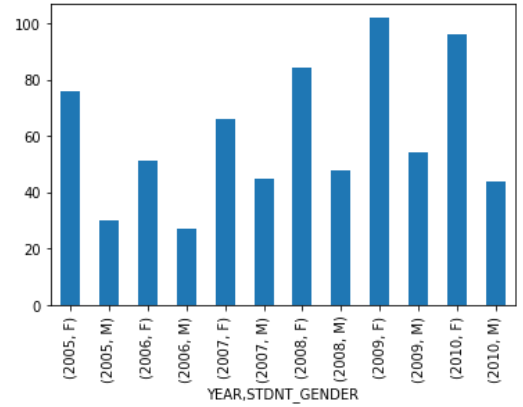


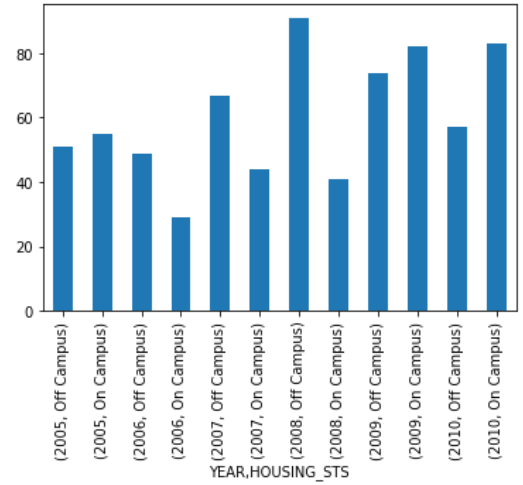
- Students who scored 800 and above in entrance test have more attrition than others.
- Students whose unmet financial need is beyond 100k drops out more.

EXPLORATORY DATA ANALYSIS & DATA INSIGHTS

- The attrition rate is getting increased in recent years.
- On campus student attrits more in past two years.
- Female students attrition count is increasing significantly every year.







DATA PRE-PROCESSING & FEATURE ENGINEERING

The given dataset contains,

- 3400 entries of data;
- 56 columns of which has,
  - 10 are int datatype,
  - 10 are float datatype,
  - 36 are object datatype
- Out of which, 32 columns have null values.

DATA PRE-PROCESSING & FEATURE ENGINEERING

Dropping the columns which has null value count more than 25%, viz.,

- CORE_COURSE_NAME_6_S
- CORE_COURSE_GRADE_6_S
- CORE_COURSE_NAME_6_F
- CORE_COURSE_GRADE_6_F
- CORE_COURSE_NAME_5_S
- CORE_COURSE_GRADE_5_S
- CORE_COURSE_NAME_5_F
- CORE_COURSE_GRADE_5_F
- STDNT_TEST_ENTRANCE1
- CORE_COURSE_NAME_4_S
- CORE_COURSE_GRADE_4_S
- CORE_COURSE_NAME_4_F
- CORE_COURSE_GRADE_4_F
- CORE_COURSE_NAME_3_S
- CORE_COURSE_GRADE_3_S
- STDNT_TEST_ENTRANCE2

DATA PRE-PROCESSING & FEATURE ENGINEERING

Dropping the ineffectual columns namely,

- STUDENT IDENTIFIER
- FATHER_HI_EDU_DESC
- MOTHER_HI_EDU_DESC
- DEGREE_GROUP_DESC
- FIRST_TERM
- SECOND_TERM

DATA PRE-PROCESSING & FEATURE ENGINEERING

Null Value Treatment:

Filling the null values with most common values in the following columns,

- FATHER_HI_EDU_CD
- MOTHER_HI_EDU_CD
- HIGH_SCHL_NAME

DATA PRE-PROCESSING & FEATURE ENGINEERING

Null Value Treatment:

Filling the null values with their mean value in the following columns,

- SECOND_TERM_ATTEMPT_HRS
- SECOND_TERM_EARNED_HRS
- STDNT_TEST_ENTRANCE_COMB
- DISTANCE_FROM_HOME
- HIGH_SCHL_GPA

DATA PRE-PROCESSING & FEATURE ENGINEERING

Null Value Treatment:

Filling null values in course name as 'ENG 1101 & ENG 1102' and grade as 'B & C‘ in the following columns,

- CORE_COURSE_NAME_2_F
- CORE_COURSE_GRADE_2_F
- CORE_COURSE_NAME_3_F
- CORE_COURSE_GRADE_3_F
- CORE_COURSE_NAME_1_S
- CORE_COURSE_GRADE_1_S
- CORE_COURSE_NAME_2_S
- CORE_COURSE_GRADE_2_S

DATA PRE-PROCESSING & FEATURE ENGINEERING

- Replacing 'INCOMPL' as 'NOT REP' in below course grade columns,
  - CORE_COURSE_GRADE_1_F
  - CORE_COURSE_GRADE_2_F
  - CORE_COURSE_GRADE_1_S
- Converting the following categorical columns into numerical columns,
  - HIGH_SCHL_NAME
  - STDNT_BACKGROUND

DATA PRE-PROCESSING & FEATURE ENGINEERING

- Labelling the 'UNMET_NEED' column as,
  - 0 if UNMET_NEED = 0
  - 1 if UNMET_NEED > 0
  - 2 if UNMET_NEED < 0
- Replacing 1 as attriting and 0 as not attriting in 'RETURNED_2ND_YR' column.

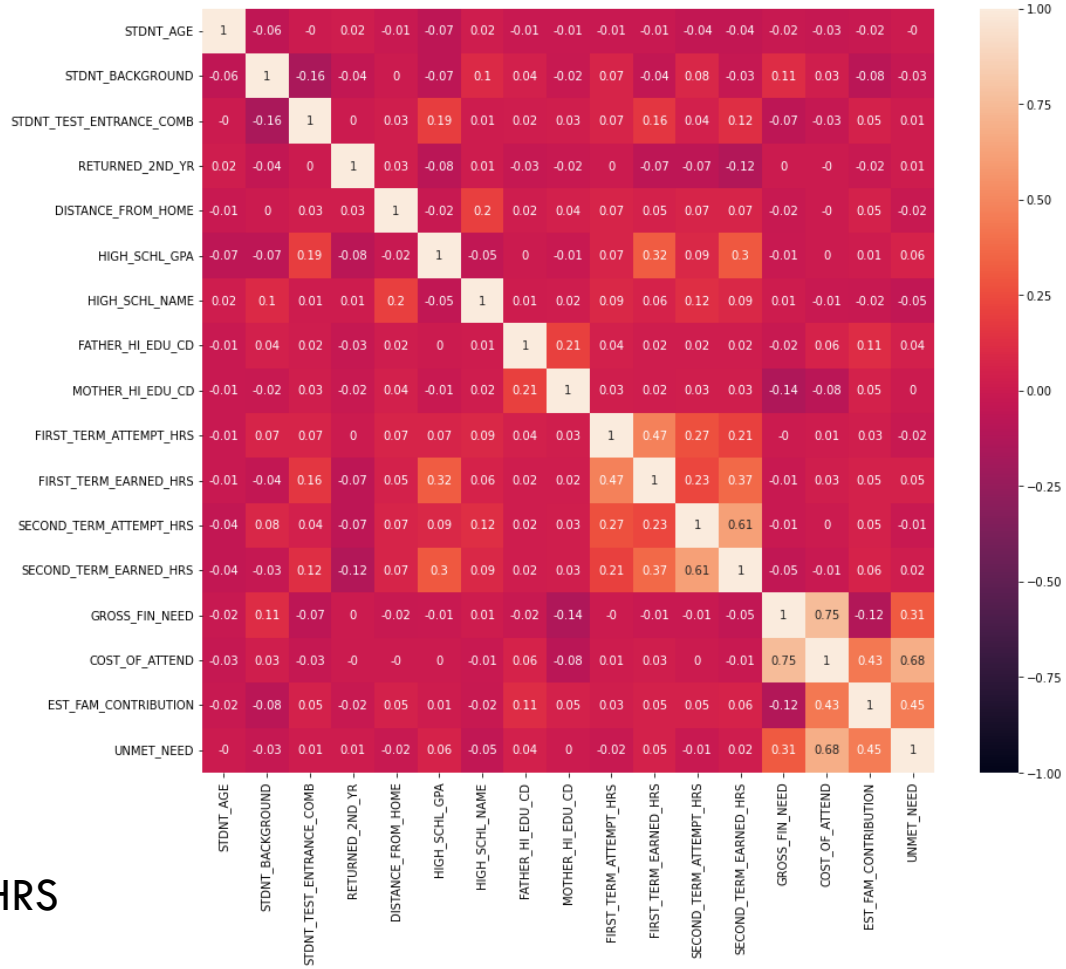
DATA PRE-PROCESSING & FEATURE ENGINEERING

From the above map,

- COST_OF_ATTEND is highly correlated with GROSS_FIN_NEED & UNMET_NEED
- FIRST_TERM_ATTEMPT_HRS is highly correlated with FIRST_TERM_ATTEMPT_HRS
- SECOND_TERM_ATTEMPT_HRS is highly correlated with SECOND_TERM_ATTEMPT_HRS

So, the below columns are dropped.

- UNMET_NEED
- GROSS_FIN _NEED
- FIRST_TERM_ATTEMPT_HRS
- SECOND_TERM_ATTEMPT_HRS



MODEL BUILDING

Classification algorithms such as Logistic Regression, Random Forest Classifier, XGB Classifier and Decision Tree Classifier are considered and their F1 score is as follows,

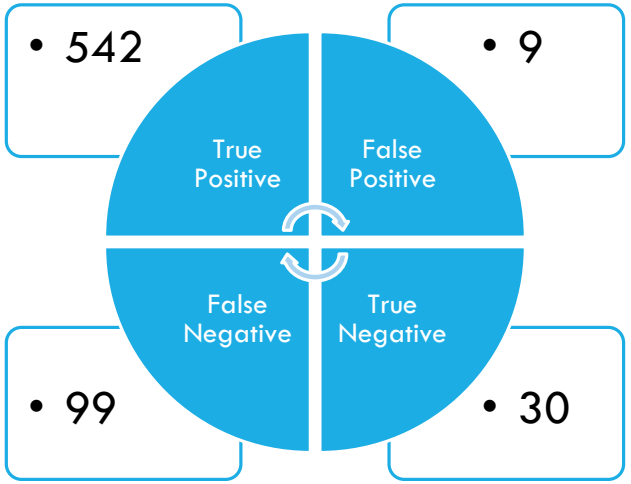
Algorithm	F1 Score
Logistic Regression	0.00275
Random Forest Classifier	1.00
XGB Classifier	0.94538
Decision Tree Classifier	1.00

Out of this four algorithms, Random Forest Classifier is chosen and model is built.

MODEL EVALUATION

- The given data is split into training and testing data with test size of 20% and the results were predicted.

  - Confusion Matrix:



  - Classification Report:

	precision	recall	f1-score	support
0	0.85	0.98	0.91	551
1	0.77	0.23	0.36	129
accuracy			0.84	680
macro avg	0.81	0.61	0.63	680
weighted avg	0.83	0.84	0.80	680

CONCLUSION – RECOMMENDATION - 1

Gender

- Consistenly female students attrition increase every year.
- 66% female students left the college

Recommendation:

Special Care on Female Students:

1. Provide Safe environment, feel good factor to attend college
2. Empower women by boosting their moral
3. Special Scholarship for female students
4. Job opportunities
5. Implementation of POSH(Prevention of Sexual Harassment Act)
6. Identify female representative and have regular meeting to understand the real issue.

CONCLUSION – RECOMMENDATION - 2

Student Major Course & Entrance Score

- More than 33% attrition contributed by Biology. Pre-Nursing, Psychology, Early Child Education & Nursing
- 'Undeclared' Stream accounts for – 15% of admission rate. Better stream declaration at the time of enrolment will help to improve
- High Attrition from students those who scored more than 900 marks in their entrance exam

Recommendation

Course Consulting facility & Monitoring Systems:

1. Arrange Course Consulting before the admission to help the student to select the right course
2. Special focus on Biology, Pre-Nursing, Psychology, Early Child Education & Nursing
3. Assess student performance more frequently by interactive sessions and conducting tests
4. Arrange Student consulting/Mentorship (one to one discussion) throughout the course to provide the guidance to the students whenever required.

CONCLUSION – RECOMMENDATION - 3

Family contribution & Unmet financial need of the student

- 456 students left due to one of the reason of family contribution.
- More than 63% students who does not get any support from family or society for fees
- High No. of Students left due to unmet of financial needs.

Recommendations:

Provide Scholarship & Loan facility:

1. identify potential students who are good in studies and offer the scholarship
2. Provide Education loan facility
3. Earn and learn scheme so that they can also support to their families.
4. Enable Part-time jobs options.