Research Project
CSF415 Data Mining
Student Enrollment in college degrees

Team number - 7

Name	**ID number**
Ravi Bharadwaj C	2016AAPS0244H
Shekhar Somani	2016B3A70347H
Maneesh Sistla	2017A7PS0238H

Problem definition

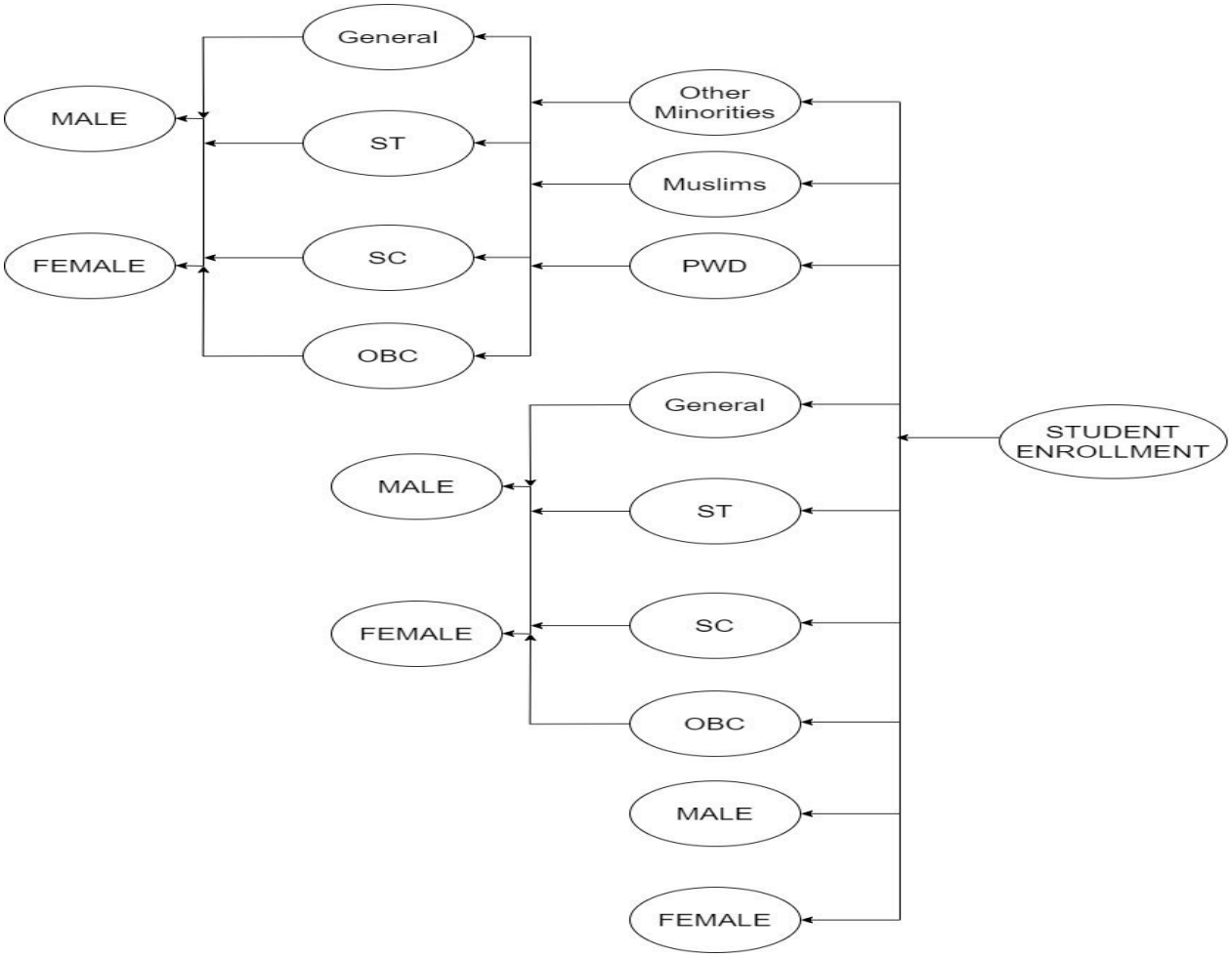
The aim of this project is to find relationships and draw different insights on the kind of higher education preferred by various sectors of our society. We wish to explore the following insights in detail:

1. What colleges take what percentage of each category like:
    a. Muslims
    b. Women
    c. SC/ST/OBC
    d. PWD, etc
2. If various sectors and minorities are split up or located in nearby places
3. How many years of higher education do each of the sectors pursue
4. What kind of sector prefers what kind of degrees
5. What kind of sectors look for self financing colleges and what kind of sectors don't
6. What kind of people go into IT(CS degree)
7. What kind of people go into medicine, etc.
8. Train a classifier to predict whether or not a student will get into a certain college based on his ethnicity and gender.

Different techniques like classification and clustering will be used to achieve satisfactory results for each of the insights.

Data description

Link to the dataset - [Student Enrollment for college degrees 2015-16]
The dataset contains data of 400,000+ higher education programmes offered in various colleges in our country.
The entire dataset is classified into minorities (PWD, Muslims and other minorities), categorized based on caste - (General, SC, ST and OBC) and further based on gender.
The dataset gives numerical info about the following divisions of society:



The dataset also provides detailed information about the course in a particular college, which contains features like:

- Discipline & discipline group
- Level of degree
- Number of years studied per degree
- Self financed vs general college

Development tools to be used

The following list details the use of various development tools which will be used throughout the project. This is list is not exhaustive and we may use other relevant tools as we come across them.

- Python3
    - Python 3.6 will be used throughout the project
- Jupyter
    - Used as a smooth interface for rapid prototyping and POC generation
- Pandas
    - A python library that will be used for data handling and organization
- SciKit-learn
    - A python library that will be used for verification and implementation of simple classification models
- Matplotlib
    - A python library to generate plots for data analysis
- Seaborn
    - Another python library used for generation of plots for data analysis