STUDY OF PATTERNS IN HIGHER EDUCATION ENROLLMENT

Bharadwaj. R
Student
BITS Pilani, Hyderabad Campus

Somani. S
Student
BITS Pilani, Hyderabad Campus

Sistla. M
Student
BITS Pilani, Hyderabad Campus

***Abstract*—The aim of this project is to find patterns, relationships, and draw insights from the types of higher education preferred by different sections and castes of society. This paper details the different methods used in the cleaning, preprocessing and visualzation of the dataset.**

***Keywords*—Data mining, higher education, preprocessing, visualization**

I. Introduction

This research project aims to analyze the student enrollment patterns in higher education programmes across India in the year of 2015. The dataset was obtained from Indian government data webite – https://data.gov.in. This dataset contains 400,000+ entries of programmes which show the types of students enrolled in differnet higher education programmes.

Pre-processing techniques like data cleaning, data reduction and data transformation were applied on the data.

Various visualizations were also plotted to give us a better understanding of the dataset.

This project was done in python3 using jupyter notebooks.

II. Problem Definition

We aim to find patterns and draw insights from the student enrollment in higher education programmes for the year of 2015 in India. We aim to use this dataset to draw the following insights:

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

III. Data Description

The dataset contains data of 400,000+ higher education programmes offered in various colleges in our country in the year of 2015. The dataset initially (before pre-processing) contained 58 features. The dataset is divided into three levels - 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 values on the enrollment statistics per minority, caste, gender, and a combination of the above three.

The dataset is very sparse and contains a lot of redundant features which were removed during data pre-processing.

IV. Data Pre-processing

Data pre-processing on the dataset involved data cleaning, data reduction and data transformation to prepare the dataset for analysis tasks further on.

A. *Data Cleaning*

a) An essential part of data cleaning included the processing of empty data. All blank values were replaced with zeros as zeros represent no enrolled students of the particular minorty, caste or gender.

b) *As another part of data cleaning, some features were converted to the same case to avoid duplicates of different case present*

B. *Data Reduction*

As a part of data reduction, we removed unecessary columns and columns which were blank. We further removed columns which contained ids and remarks. These columns would not affect the data analysis in any way and hence were removed.

C. *Data Transformation*

a) As an essential part of data transformation, we used min-max normalization. This was used to map the number of students of each minority, caste or religion to 0 to 1. The min value was 0 and max was the max number of students enrolled in the programme

b) As another important part of data transformation, we used binarization to convert textual data into vectors to help with classification tasks. We used the one-hot encoding method to convert textual categorical data into binary vectors

V. Data Visualization

The visualization techniques were used to draw insights about the magnitude of population that studied to various higher education levels, the distribution of castes in student enrollment(Fig. 2), and the distribution of minorities as well. A few graphs were plotted to indicate the number of students studying in various fields like Computer Science(Fig. 1), Medicine and Arts, with additional information about how many of the aforementioned students studied at various education levels. Seaborn library was used to plot the barplots.

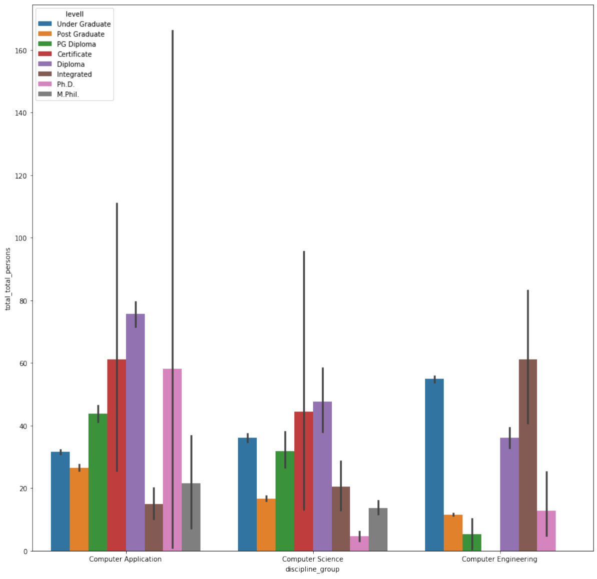

Fig. 1

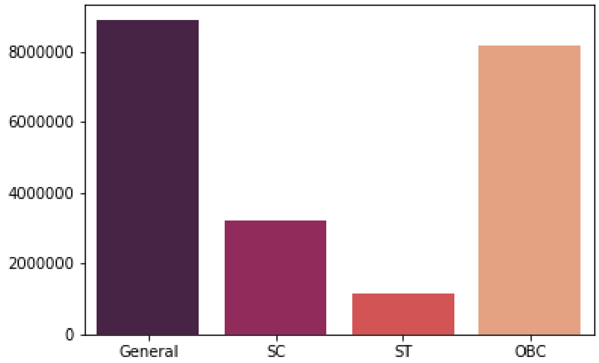

Fig. 2