Final Project Proposal AMAT 502

Krishna Chaitanya Yejju, Guru Datta Abhishek Yejju, Tejas Patel, Mahesh Utlapalli

November 20, 2021

1 Introduction

The dataset that we want to investigate in our final project is Bachelors Degree Majors. This dataset consists of 612 rows and 9 columns with almost all of the values being unique. This dataset majorly deals with the data of the bachelors degree holders from different states of the United States of America by the different subcategories which are: State, Sex, Age group, and different fields of students with their major degree. There are 51 unique values in the state column, 4 unique values in the age group, and 3 unique values in the sex column.

This dataset is much interesting based on the above-mentioned features. And there are even more features that make this dataset to be unique from other datasets those are the fields of major degree that students are obtaining from the particular state in their field of interest. For example: In this dataset, there are five different types of fields were wherein any of the fields that student can get his/her major degree based upon their interest.

2 Methods

2.1 Classification

Classification uses the input training data, here we use decision tree algorithm where the data is continuously split according to a certain parameter to predict the likelihood that subsequent data will fall into one of the bachelor's degrees by taking unique features like age group, sex, and state.

2.2 Clustering

Clustering is the way of grouping the data points into different clusters, consisting of similar data points or behavior.

2.2.1 Method-1: K-Means

By K-Means algorithm we have so many advantages like less time complexity and more approachable, we will recompute the group center points using the mean value with the number of groups represented by the variable K. Finally data points will be clustered based on feature selection.

2.2.2 Method-2: Gaussian Mixture Models(GMM)

Gaussian mixture model involves the mixture (i.e., superposition) of multiple distributions. Here rather than identifying clusters by ”nearest” centroids, we fit a set of K Gaussian's to the data. And we estimate Gaussian distribution parameters such as mean and variance for each cluster and weight of a cluster. After learning the parameters for each data point we can calculate the probabilities of it belonging to each of the clusters.

2.3 Regression

To predict value based on inputs or independent variables we use regression where one or more than one attribute(s) can be used to predict using binary output.

3 References

[1] https://www.kaggle.com/tjkyner/bachelor-degree-majors-by-age-sex-and-state