Overview

You will be provided a dataset that contains answers from a survey given to 2 Computer Science Classes, one from year 2017 and the other from 2018. Given this dataset, you will be executing the following:

- Exploratory Data Analysis and Feature Engineering
- Unsupervised Learning
- Supervised Learning

Assignment

You may use any programming language and API you prefer to complete this assignment (R, Python, Scala, etc).

Exploratory Data Analysis and Feature Engineering

Perform Exploratory Data Analysis and Feature Engineering. Provide a zip file containing:

- Brief PDF report highlighting observations or patterns in the data that you found. Include charts/plots illustrating your observations/patterns (minimum of 1 chart)
- Source code files

Unsupervised Learning

Identify subgroups within the data and visualize results; perform any method of unsupervised learning you wish. Provide a zip file containing:

- Brief PDF report explaining your approach that include charts/plots illustrating your results
- Source code files

Supervised Learning

Construct a model that will predict what year an individual student is from. Perform any method of supervised learning you wish. Use 3-fold cross validation in conjunction with Area under ROC Curve as a performance metric for your model. Include a confusion matrix of your best run.

Provide a zip file containing:

- Brief PDF report explaining your approach; include Area under ROC Curve Value and Confusion Matrix
- Source code files