Final Project Proposal

Note: Only one person for each team needs to submit the proposal on Gradescope.
See the submission guidelines for more details.

1) Team members

name	**BU email**
Janaki Riji Nair	janaki@bu.edu
Hongyi Yu	yuhy@bu.edu
Corey Evans	ceboston@bu.edu

2) Description of dataset

Our project aims to use various characteristics of universities in order to predict the range of tuition charged by universities in an academic year.

Dataset:

One dataset that will be used for the project is the National Universities Rankings which gives information on universities in the United States provided by the U.S. News & World Report in 2017. This dataset available as a .csv file includes some of the variables and can be accessed from data.world using the following link:

https://datasetsearch.research.google.com/search?query=Best%20National%20University%20Rankings&docid=L2cvMTFqb pnbHg1dg%3D%3D

Tables will also be created for some variables using data provided by the U.S. News & World Report in the following link:
https://www.usnews.com/best-colleges/rankings/national-universities

Input Attributes:

- Ranking: This is a numeric attribute that refers to the national ranking of universities in the United States.
- Setting: This is a nominal attribute that gives information on whether the university is situated in an urban or suburban or rural setting.
- Location: This is a nominal attribute that refers to the state in which a university is located.
- Undergraduate enrollment: This is a numeric attribute that refers to the total number of students enrolled in undergraduate programs in a university.

- School type: This is a nominal attribute that describes whether the university is a public or private university.

Output Attribute:

- Tuition range: This is a nominal attribute that would show the range of tuition charged by a university over the course of an academic year.

3) Type(s) of data mining (Note: at least one of the types must be classification learning or numeric estimation)

The type of data mining we intend to use for this project is Classification learning. This approach is chosen because there is a single output attribute which is the range of tuition and thus a nominal attribute.

4) Goal of the resulting model

Our data model would predict a university's tuition range based on various input attributes pertaining to the university. The input attributes would include ranking, location, setting, school type, and undergraduate enrollment of the university.

5) Description of transformations to your data

- The input attribute of university's ranking is provided as a numeric attribute in the dataset we obtained. Python would be used to discretize the numeric attribute of ranking into a nominal attribute of ranking ranges.
- The dataset currently gives the US states in which a university is located under the location attribute. We would use python to categorize the data on this input attribute into four distinct location types: West, East, Central and South.
- Undergraduate enrollment is given as a numeric attribute in our dataset. This input attribute would be transformed to a nominal attribute by changing it into ranges using Python.
- Relational tables will also be created on input attributes like setting and school type.
- The output attribute of tuition charged is presently a numeric attribute. This would be discretized using Python into a nominal attribute that provides tuition in ranges in order to use Classification learning in our project.

6) Plan for other required components

- We will use Python to discretize numeric variables into nominal variables as explicated above.

- We will utilize SQL queries to perform computations like averages, summation etc on the various variables identified in the project
- Discovering patterns in the data will help us determine how the attributes predict tuition range of universities in the United States. We would specifically learn decision tree models using the training data to predict tuition ranges of universities.
- We will be creating confusion matrices to assess the model's overall accuracy and compute goodness scores.
- We will use WEKA for the visualization of our data models.

7) Division of work among team members

All team members would collaborate for all steps in the project. However, each team member would be taking the lead for the different steps as mentioned below:

- Janaki: Data Transformation using Python
- Hongyi: Finding Dataset, Data Mining
- Corey: Data Visualization using WEKA