Midterm Project 3

Racial and Ethnic Representativeness Data Sets
MTH 3270 Data Science
Due Fri., May 13

Rules

You may work alone or with a partner from the class. You're only allowed to communicate about this project with the instructor (Grevstad) or your partner if you are working with one. If you work with a partner, the two of you will submit the same project and receive the same score.

All analyses (data wrangling, visualizations, statistical summaries, etc.) must be done using **R** (except by permission of the instructor).

The projects are **due** in **Canvas** as a **pdf** file no later than **Friday, May 13, 2022** at **11:59 PM**.

Instructions

The project will use the following data sets:

1. The **Racial and Ethnic Representativeness of US Postsecondary Education Institutions** data sets from the annual Data Challenge Expo contest sponsored by the American Statistical Association.
2. Another **Supplemental 2017** data set, posted on the **MTH 3270** course page in **Canvas**.

The **Racial and Ethnic Representativeness of US Postsecondary Education Institutions** data sets are:

> **HEsegDataviz_CollegeData_4-year_v5.csv** This dataset combines public data from the Integrated Postsecondary Education Data System and the US Census Bureau's American Community Service in an index of racial and ethnic representativeness of US postsecondary education **four-year** institutions. The data link college racial composition to the racial composition of an institution's "market," defined geographically according to institutions' level, degree of selectivity, and urbanicity.
>
> **HEsegDataviz_CollegeData_2-year_v5.csv** The same as HEsegDataviz_CollegeData_4-year_v5.csv, but for **two-year** institutions.

The **data sets** and a **data dictionary** (**HEsegDataviz_Dictionary.xlsx**) containing **descriptions** of the **variables** in the data sets are obtained via the link below. Save one or the other of the **csv** files containing the data and read it into R using `read.csv()` (and don't forget `header = TRUE` and `stringsAsFactors = FALSE`).

`community.amstat.org/dataexpo/home`

The **Supplemental 2017** data set is:

> **Supp2017Data.csv** This dataset contains **more** public data from the Integrated Postsecondary Education Data System on US postsecondary education **four-year *and* two-year** institutions (`https://nces.ed.gov/ipeds/`). The data contain, for each institution, the institution ID number, institution name, year (all 2017) undergraduate enrollment, student to faculty ratio, 12-month undergraduate headcount, graduation rate.

The **data set** and a **data dictionary** (**Supp2017Data_Dictionary.xlsx**) containing **descriptions** of the **variables** in the data set are obtained via the **course site** in **Canvas**. Save the **csv** file containing the data and read it into R using `read.csv()` (and don't forget `header = TRUE` and `stringsAsFactors = FALSE`).

Note: Each college appears in multiple rows of **Racial and Ethnic Representativeness of US Postsecondary Education Institutions** data sets (once for each of the years 2009-2017). You **must** filter out just the rows corresponding to the **year 2017** (using `filter()`), and do the entire project using data for just that one year (which is the year corresponding to the **Supplemental 2017** data set).

You *might* need to do some further data wrangling and tidying (which *might* involve selecting columns, adding new columns, filtering rows, grouping by a categorical variable, recoding, etc.).

Check **Canvas Announcements** and/or your **email** regularly in case there are important announcements about this project.

Tasks

Your **two tasks** are:

T1 Every machine learning procedure has at least one **tuning parameter** (or **complexity parameter**), whose value you choose, that controls the **model complexity**, that is, how closely the fitted model is able to conform to the data:

- → ***Decision tree***. The tuning parameters are: 1) The **minimum size of a node** in order for a split to be attempted; 2) The **complexity parameter**, **cp**, for which a split is only performed if it decreases the misclassification rate by this percent or more.
- → ***Random forest***: The tuning parameter is the **number of variables** to use in each tree.
- → ***K nearest neighbor***: The tuning parameter is the **number of neighbors**, k.
- → ***Artificial neural network***: The tuning parameter is the **number of hidden units**, k.

A *poorly* chosen **tuning parameter** value leads to **overfitting** or to **underfitting**. A *good* tuning parameter value does neither. In other words, a *good* tuning parameter value produces a fitted model that classifies or predicts **out-of-sample** observations well.

The **tuning parameter** value can be selected via splitting the data into **training** and **testing** sets. The **test** set serves as the **out-of-sample** observations.

Your **first task** consists of the following **three steps**:

(a) First **join**, ***by institution***, either the ***four-year* 2017** or the ***two-year* 2017 Racial and Ethnic Representativeness of US Postsecondary Education Institutions** data set to the **Supplemental 2017** data set (use one of the `*_join()` functions from `"dplyr"`).

(b) Next, separate the resulting data set ***randomly*** into **75% training** and **25% testing** sets.

(c) Then choose **two** of the above ***machine learning* classification** procedures (either **decision tree**, **random forest**, ***k* nearest neighbor**, or **artificial neural network** – your choices) for **predicting** the **Four-year Institution Category** (`fourcat`).

For **each** of the **two** procedures you chose to use:

i. Carry out the procedure on the ***training set***, using a **minimum** of **three explanatory (*X*) variables** in the model, at least ***two of which*** must come from the **Supplemental 2017** data set. The choice of *which* explanatory (X) variables to use is yours to make, but they must be numerical (*not* categorical).

ii. Fit the model to the (*training set*) data using **at least three** different values of the **tuning parameter**.

iii. For each of the **three** different **tuning parameter** values, apply the ftted model (based on the *training* set) to the ***test set*** to **classify** institutions (in the *test set*) into **four-year institution categories**, and compute the **accuracy** (e.g **correct classification rate**) of the procedure.

Then, for **each** of the **two** procedures:

- **Summarize** your procedure: Indicate *which* **classification procedure** you used and *which* **explanatory variables** you used.

- **Report** the **values** of the **tuning parameter** you evaluated, and indicate **which** of these values resulted in the **best *out-of-sample* classifier** of institutions, e.g. which one had the highest **correct classification rate** for the ***testing set***.

T2 Your **second task** is to carry out a ***cluster analysis*** (**hierarchical** or ***k* means**, your choice) to group the institutions into ***k*** clusters, where ***k*** is in the range **2-5** (your choice). You must use a **minimum** of **four explanatory (*X*) variables** in the procedure, at least ***two of which*** must come from the **Supplemental 2017** data set. The choice of *which* explanatory (X) variables to use is yours to make, but they must be numerical (*not* categorical).

You're *strongly encouraged* to **standardize** (center and rescale) the explanatory (X) variables prior to carrying out the cluster analysis so that the results are more meaningful (use `rescale()`), but it's *not* a requirement.

Next, inspect whether the clusters seem to correspond to the **Four-year Institution Categories** (`fourcat`). To decide, look for whether institutions within clusters are largely in one **Four-year Institution Category** or another (use the variable `fourcat`). This can be an informal inspection or something more formal (e.g. computing a measure of "purity" for each cluster) – your choice.

(It's okay if the institutions *don't* cluster according to Four-year Institution Categories.)

Then

- **Summarize** your procedure: Indicate *which* **cluster analysis procedure** you used, *how many* **groups** k you used, *which* **explanatory variables** you used, and *how many* **institutions** ended up being in each of the k clusters (groups).

- **Report** the results of your assessment of whether the clusters seem to correspond to **Four-year Institution Categories**.

What to Turn In

1. A well-organized **write-up** as a **pdf** file (perhaps 3-7 pages) containing:

   (a) A **brief description** (e.g. 1-2 paragraphs) of any data **wrangling** and **tidying** you had to do in order to carry out tasks **T1** and **T2**.

   (b) Your **responses** addressing the **bullet items** under tasks **T1** and **T2** above (*four* bullet items total).

2. Your **R code** with **comments** (use `#`) indicating **what** each chunk of code does and **why** it does it, either as an **appendix** in your **write-up pdf** or as a separate **.R file** (as produced by RStudio's script editor).

Grading

Your **grade** will be based on:

1. Your level of attainment of **tasks T1** and **T2**.

2. Your **write-up**, and in particular, the inclusion and depth of your **responses** addressing the four **bullet items** (as described above).

3. The inclusion of and correctness of your **commented R code**.