Midterm Project 2

MTH 3270 Data Science

Due Mon., Apr. 27* (but will be accepated later – see below)

Rules: You must do your own work, and you're only allowed to speak about this project with the instructor (Grevstad) and your partner from the class (if you will be working with one).

If you will be working with a partner, you must notify the instructor (Grevstad) via email *prior* to beginning work together. The two of you will only submit one project (with both your names on it).

I'll be "flexible" with the due date. The projects are **due** in **Blackboard**, but I'll accept them any time **within one week** of the due date **Apr. 27, 2020** (i.e. any time up to **May 4** at **12:59 PM**).

Instructions: Check your **email** immediately for the **data sets** and also regularly during the week in case there are important announcements about this project.

This project will be a **secondary** analysis of the **education** data sets that were used in Project 1 (in your **MSU Denver email**):

API_4_DS2_en_csv_v2_820954.csv
Metadata_Country_API_4_DS2_en_csv_v2_820954.csv
Metadata_Indicator_API_4_DS2_en_csv_v2_820954.csv

The first one is the main data set, and contains data on education for each of 264 of the world's countries for the years 1960-2019. The second contains information about each country. The third contains information about the variables recorded for each country.

The data were obtained by going here:

`https://data.worldbank.org/topic/education`

and then clicking on "CSV" under "Download" on the right.

You'll need to either manually **delete** the **first four rows** of **API_4_DS2_en_csv_v2_820954.csv** *or* use the `skip = 4` argument in `read.csv()` before reading the data into R (and don't

forget `header = TRUE` and `stringsAsFactors = FALSE` in `read.csv()`). You may modify the code below.

```
# Read in data:
my.data <- read.csv("~/grevstad/mth3270/projects_mth3270/API_4_DS2_en_csv_v2_820954.csv",
                    skip = 4,
                    header = TRUE,
                    stringsAsFactors = FALSE)

# Read in country metadata:
my.countries <- read.csv("~/grevstad/mth3270/projects_mth3270/Metadata_Country_API_4_DS2_en
                         header = TRUE,
                         stringsAsFactors = FALSE)
```

You will use data from **only one of years** represented in the full data set (your choice). You'll need to do some **data wrangling** and **tidying** (which *may* involve combining data sets, selecting columns, adding new columns, filtering rows, grouping by a categorical variable, converting between wide and narrow formats, etc.). All **data wrangling** and **tidying** must be done in R (except by permission of the instructor). You may use the code below.

```
# For select(), left_join(), rename():
library(dplyr)

# For pivot_wider() (modern version of spread()):
library(tidyr)

# Rename the country code variable:
my.countries <- rename(.data = my.countries, Country.Code = ï..Country.Code)

# Left joint country codes to data:
countries <- left_join(my.countries, my.data, by = "Country.Code")

# Select one year.  I am using, as an example, 2000:
countries <- select(countries, Country.Code, Region, IncomeGroup,
                               Country.Name, Indicator.Code, X2000)

# Convert from narrow to wide form.  The pivot_wider() function is a more mod-
ern version of spread():
countries <- pivot_wider(data = countries,
                         names_from = Indicator.Code,
```

```
                            values_from = X2000,
                            id_cols = c(Country.Code, Region, IncomeGroup, Country.Name))
```

Your **tasks** are:

1. Carry out a **multiple regression analysis**. You may choose any response variable (Y) for your model, but it must be a numerical variable (not categorical). Likewise, you may use any explanatory (X) variables, but they too must be numerical (not categorical).

   Summarize your fitted model, and report at least one measure of **how well** the model **fits** the data.

2. Carry out a **logistic regression analyses** for predicting whether a country is in the **high income group** based on other explanatory variables from the data set.

   For the response (Y) variable, you'll need to create a *dichotomous* variable taking the value **1** if a country is **high income** and **0 otherwise** (e.g. using `mutate()` with `ifelse()`). You may use any explanatory (X) variables, but they must be numerical (not categorical).

   Summarize your fitted model.

3. At least one **machine learning** procedure (decision tree, random forest, k nearest neighbor, naive Bayes, or artificial neural network) for **predicting** the **categorical** variable **income group** (the original variable, not the one coded as **0** or **1**).

   Summarize the results of your procedure and provide an example of a **prediction** that it makes.

What to turn in: A **write-up** (perhaps 2-5 pages including graphs, if any) consisting of:

1. A **brief description** (e.g. 1-2 paragraphs) of any data wrangling and tidying you had to do in order to carry out tasks **1-3** above.

2. For task **1-3** above, a summary of the model you fitted or machine learning procedure you carried out (e.g. which variables were in the model, how well the model fitted the data or predicted the response variable, and which explanatory variables seemed important). of your comparisons/trend investigation (including the graphs).

3. Your **R code** with **comments** (use `#`) indicating **what** each chunk of code does and **why** it does it. The instructor (Grevstad) may request an electronic copy of your **R code** in a **.R file** (as produced by RStudio's editor), so please hold on to it.

Grading: Your grade will be based on:

1. Your attainment of **tasks 1-3** above.
2. Your **write-up**, including the graphs (as described above).
3. The inclusion of and correctness of your **R code**.