PROJECT RESPONSES

1. Import your datasets (one for fitting regression models and one for fitting classification models) into R and perform the necessary clean-up operations. Provide a short summary of what clean-up operations were needed (e.g., changing the type of one or more variables, computing new variables, removing observations with missing values, etc.).

⇨ The two data sets are imported and cleaned up. For the classification analysis, the outcome variable in the dataset fetal_health and another variable histogram_tendency is converted into factors, and the variables with non-zero variance variables are removed. For the regression analysis, the variables country and status are converted into factors, NAs are removed and the variables with non-zero variance variables are removed.

2. Provide names, descriptions for each variable in the datasets (you can use to the output from summary() here).

⇨ Summary of the variables in both the datasets are produced using summary().

3. Provide details of the context in which the datasets were collected.

⇨ Fetal Health data set:

a. Reduction of child mortality is reflected in several of the United Nations' Sustainable Development Goals and is a key indicator of human progress.
The UN expects that by 2030, countries end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce under-5 mortality to at least as low as 25 per 1,000 live births.

b. Parallel to notion of child mortality is of course maternal mortality, which accounts for $295\,000$ deaths during and following pregnancy and childbirth (as of 2017). The vast majority of these deaths (94%) occurred in low-resource settings, and most could have been prevented.

c. In light of what was mentioned above, Cardiotocograms (CTGs) are a simple and cost accessible option to assess fetal health, allowing healthcare professionals to take action in order to prevent child and maternal mortality. The equipment itself works by sending ultrasound pulses and reading its response, thus shedding light on fetal heart rate (FHR), fetal movements, uterine contractions and more.

Life expectancy data:

a. Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries.

b. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well.

c. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

4. Explain whether performing dimension-reduction is appropriate in the case of each of your datasets.

⇨ As the datasets contain less dimensions than observations and there thousands of data points dimensionality reduction is appropriate.

5. Explain whether using cluster analysis and using the cluster identifier column as a predictor is appropriate in the case of each of the datasets.

⇨ The cluster analysis is not appropriate in case of datasets as there are multiple values for outcome variable

6. Conduct the specific set of analyses that were described in this week's lecture video on a subset of each dataset, to ensure that the modeling process could be completed after solving any data-related and/or model-fitting-related issues.

⇨ Some of the analysis are conducted prior to fitting the models

7. As part of your modeling process, ensure that you are able to compute variable importance for each of the modeling approaches (refer to the lecture video for details).

⇨ The variable importance is performed

8. As part of your modeling process, ensure that you have identified appropriate measures of performance, and are able to compute variations in model performance using the approach described in section 5 of the caret package's documentation site. Specifically, in regression models, begin working on the code needed for comparing MLR, Lasso, GAM, Random Forest, Boosted Tree, and SVM models via numerical and graphical summaries. Analogously, on the classification side, begin working on code for comparing results from Logistic regression, LDA or QDA (pick one that is more appropriate), KNN, Random Forest, Boosted Tree, and SVM approaches. You don't have to have this code ready yet, but getting started now will be useful. Note: once we cover neural networks, you will consider their applicability in building predictive models on your datasets, too. That is something you would do during week 13.

⇨ All the models are fitted, but there are some problems in accuracy computation and certain necessary steps and there is a necessity to check those.