MA335 Final Project

Due in: 12pm Friday 9th of July 2021

June 21, 2021

The project involves implementation of clustering and classification techniques on the data set which includes a selection of the World Development Indicators (WDI), derived from a primary World Bank database, and the information about the casualties of the COVID-19 pandemic. The *project_data.csv* file stores the data set used for this assignment and *project_data_info.csv* provides the definitions of the variables.

Task: Investigate the relationship between WDI indicators and the severity of the COVID-19 pandemic. The task is designed to test student's ability to apply clustering and classification algorithms and the ability to evaluate and interpret their results. Since we want to perform classification the response variable (*covid_deaths*) will have to be converted to a categorical one. Clustering should be implemented to understand how WDI indicators could be used to group countries, and later considered when discussing the results of classification models. Use **R** or **Python** in order to conduct your statistical analysis. **The implemented code should be included as part of an Appendix of your report and should run without errors**.

For your analysis of the relationship between WDI indicators and COVID-19 causalities complete the following tasks:

1. Analyse using descriptive statistics (both graphical and numerical representations) project_data.csv dataset. Generate an appropriate table as summary and appropriate graphs - for example boxplots, histograms or scatterplots. [15 marks]

2. Implement Clustering Algorithms with the first 10 variables of the datasets (leaving out *continent* and *covid_deaths* variables) using K-means and Agglomerative Clustering algorithms. Comment on the results of clustering in relation to the *continent* variable. [20 marks]

3. Transform *covid_deaths* into a binary variable. Fit logistic regression model using the remaining 11 variables to predict high COVID causalities. Describe the produced model and comment on what it demonstrates. [15 marks]

4. Transform *covid_deaths* into a categorical variable with 4 possible labels. Consider whether some manipulation of the data-set should be implemented before applying your learning algorithms. Implement QDA, LDA and logistic regression for this multiclass classification problem. Compare the results using appropriate validation techniques and performance metrics. [30 marks]

5. Discuss to what extent WDI data can be used to predict the causalities of similar pandemics. Is it realistic that the same relationships between the predictor and response variables should

be expected in the future? Summarise what can be learnt from this data about the response to COVID-19 pandemic of countries with similar economic profiles. [10 marks]

The remaining 10 marks will be allocated based on the following report guidelines (see *Writing Reports: a brief guide* document on moodle).

General rules and hints:

- Plan and structure your work. Structure your report, for example, Page 1: cover page (title, your name, date, and so on). Page 2: abstract, contents and word count. Pages 3-7: introduction; preliminary analysis; analysis; discussion; conclusion; references. Page 8- 10: appendix: R-code with explanations, etc.
- Put all the code that was necessary for your report into an appendix, explaining what you are doing (add comments within the R or Python script). Do not include code of an analysis which is not used for your report. Make sure, that YOU wrote the code (the use of someone else's work, without citing the source, can be viewed as plagiarism).
- Use an appropriate word processor (MS Word or Open office) or type setter (Lyx or Latex).
- Use point size 12, Times New Roman; line spacing 1.5.
- The report should contain between 1600 and 2400 words (without the cover page and appendix). It should not be more than 8 pages long without counting the cover page and the appendix. More than 2400 words, or more than 8 pages, might result in a reduction of marks.
- Do not use more than 5 figures and 4 tables within the main text. You may include further figures and tables in the appendix, if necessary.