CSCI316
Big Data Mining Techniques and Implementation

Project Specification

You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) description of your approach and how the methodology was implemented; (4) the strengths and weaknesses of the approach or implementation; (5) your results and an analysis of the results; (6) a brief summary and a conclusion. The summary should state new and interesting things that you learned and discovered while working on this project. The conclusion should summarize your main findings and statements about possible future work (e.g., how you plan to improve your models and approach in future).

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (approach, and discussion of strengths and weaknesses)
- Implementation (methods, key-issues, how these were addressed and sample codes)
- Results (include illustrative Figures and Tables and explanations)
- Discussion and Conclusions

The Task

Definition of the task:
You are to implement an end-to-end data mining project to analyse the provided dataset. The objective is to implement a workflow to predict the targe variable of the data (i.e., classification or regression). This workflow must include two stages, as illustrated in Figure 1.

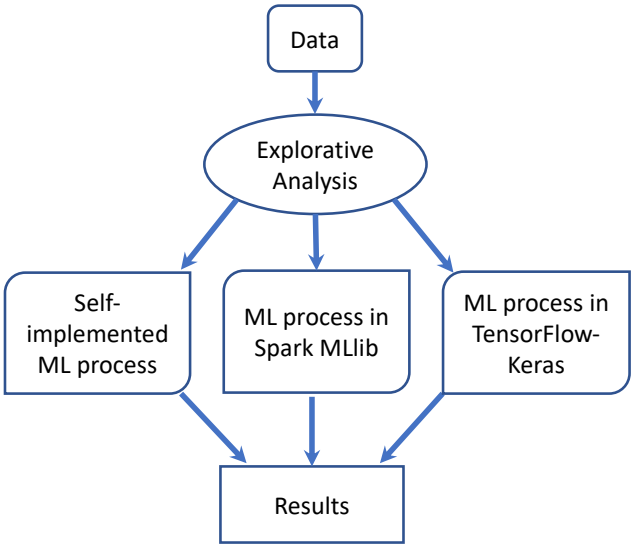

Figure 1. The workflow stages

<u>Stage1:</u> Data exploration
This stage includes the use of Spark's DataFrame and RDD APIs in Python to explore the data. Understand the dataset by querying a few important statistic measures of the data. Visualise the data and explain your findings. It is important that you demonstrate an in-depth understanding on the data that you are analysing. (Note. You *cannot* use Pandas and Scikit-Learn in this stage.)

Stage2: Predictive analysis

This stage includes three machine learning (ML) processes. Each process must include ***at least three*** kinds of ML models (such as decision tree, random forest, naïve Bayesian, feedforward network, etc.). The models must be evaluated with common metrics (such as accuracy, precision, recall and ROC).

Specific requirements of Process One:

- All ML models and evaluation methods in this process must be implemented *from scratch* in Python. You can use any Python modules except the ML libraries.

Specific requirements of Process Two:

- This process is built with the ML library of Spark (i.e., pyspark.ml and pyspark.mllib)

Specific requirements of Process Three:

- This process is built with TensorFlow and Keras.

Deliverables

- Slides
- Presentation
- Python Source Codes

In your slides, you must explain the detailed pipeline design and evaluation outcomes, as well as any other interesting findings or lessons learned. Any claim that you make in the slides must be supported by the implementation in your submitted Python source codes.