**Data Science Program Final Project**

**Executive Summary**

At the end of the Data Science program, students are required to complete a final project of their choice. They are given six weeks to work on the project. Often times, they will be paired up with another fellow-student.

This document is dedicated to Michelle. It will explain the purpose and scope for the project.

**Business Objectives**

To showcase the skills that Michelle has acquired through the Data Science program. She will be using R, Python, Tableau, SQL, and other programs to wrangle, analyze, and visualize the “Classifying emails as Spam/No Spam” dataset SMSS Spam Collection assigned by Instructor Saleh. I will use NLP for preprocessing and Naïve Bayes for Modeling.

At the end of the project, Michelle should be able to explain their work in layman’s term, and present their findings to the students, faculty, staff, and potential employers, along with other interested parties via Zoom.

**Background**

As a way to activate and put practical use to what the students have learned, doing a final project is a good way to demonstrate that.

Michelle has been assigned the “SMSS Spam Collection” dataset by Instructor Saleh. (Spam or No Spam.)

**Scope**

Michelle will be using the software taught in the program to complete the project. I will be intentional on using tools of my interest or tools that may aid in finding a job. I may choose to use additional software/tools, but that is not required.

**Functional requirements**

Data Wrangling: The downloaded dataset should be successfully cleaned up for analyzing. Columns and unusable columns should be removed. As the dataset is fairly large, I should consider sub-setting the dataset in a proper manner, meaning the subset should be a random selection of the data. The datatypes for each column should also be converted to a usable format for the needed analysis.

Data Analysis: Michelle will familiarize herself with the dataset. I should have a good understanding of what each column means, and how the values are measured. I will then brainstorm on questions to ask, and what I might gather from the dataset. Then, I will identify the proper functions to create models, predictions, etc.

Data Visualization: Once I have a comprehensive understanding of and insight gathered from the dataset, I will work on visualizing the findings. I may decide to use Tableau or other graphing programs and compile the visuals and texts in a Power Point slideshow.

Presentation: Working with school leaders, Michelle will schedule a time to present their findings via Zoom. I should be able to communicate in a clear and easy-to-understand manner. The presentation should be kept around 20 minutes. I should be dressed professionally for this occasion.

**Personnel requirements**

I will touch base once a day on my work and determine next steps to ensure the I am progressing. I will continually work on my assigned project and transforms the data to properly calculate and manage the data. All aspects of the data will be reviewed to ensure that I properly process the data and to determine my findings at the end of the project. Once a week, I will plan out the next week and will complete the tasks and assignments that my instructor assigns. All assignments will be competed for the instructor to properly grade. I will report my progress to my instructor (Product Owner.)

Once a week, I will meet with my instructor. I should be prepared to ask questions and seek guidance for the next steps. All steps will be followed and completed in the expected time manner given.

I may also consult with my other mentors. I will not be afraid to ask questions and seek guidance to ensure that I complete this project.

**Delivery schedule**

Week 1: Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed. Educate myself on classifying emails. Set up Github.

Week 2: Study the dataset and ask questions. What are some possible correlations? Is the data normally distributed? What are some predictive models I can make from it? Visualize the data to see if there are any interesting findings.

Week 3: Modeling/Optimization (Combined Stepwise - Forward and Backward Selection) and Machine Learning (Random Forest.) Utilize NLP for preprocessing and Naïve Bayes for Modeling.

Week 4: Review and validate findings from the previous week and draw insights/conclusions.

Week 5: Compile findings into a Power Point slideshow. Go over it with my instructor and friend/family member to ensure that the presentation is clear and logical. Work on the style and layout of the presentation so it is delightful on the eyes.

Week 6: Make final touches to the Power Point presentation. Michelle should not attempt to come up with a brand-new analysis. There will not be enough time to verify the findings. I should practice presenting at least a couple times, and at least once with my instructor.

**Other requirements**

All programs used should be free of charge. Though I may decide to use a paid service, such as a more advanced version of Tableau, which is at my discretion.

**Assumptions**

The software programs and platforms Michelle uses should be available, up-to-date, and not broken. Michelle will utilize only those software programs and platforms that were taught during her Data Science Program. Michelle will utilize all available programs and software and ensure that she utilizes the programs specifically assigned by the instructor.

**Limitations**

If something should come up for Michelle during this six-week period, the project may be delayed. If the instructor or mentor has scheduled or unscheduled time-off, the project may be delayed as well. Michelle may experience a roadblock in her work, which may push back the completion date.

**Risks**

The risks that may arise are such like natural disasters, power outages, family emergencies or broken software/hardware. Michelle is eager to complete the program so there should be no motivation issues. The instructor and mentors are phenomenal so there is no concern of no help from her. The risk of this project being incomplete is minimal. I will be successful in completing this project!