***Which topic did you choose to apply the data science methodology to? (2 marks)***

Topic: Emails

***Next, you will play the role of the client and the data scientist.***

***Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. (3 marks)***

***You are required to:***

1. ***Describe the problem, related to the topic you selected.***
2. ***Phrase the problem as a question to be answered using data.***

***For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".***

1-) Problem:

The spam emails are the most problematic things. These emails spammed the inbox and then respective person find difficulty in order to sort out the important emails from unnecessary.

2-) Question:

 Can we figure out whether the email is spam or not?

Involvement of Email Servers Administrators:

* Set overall direction
* Remain engaged and provide guidance
* Ensure necessary support where needed

Identify business requirements

* Logistic Regression
* Yes or No for each email
* What type and specific words will be used to mark the email spam or not

***Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. (5 marks):***

1. ***Analytic Appsaroach***
2. ***Data Requirements***
3. ***Data Collection***
4. ***Data Understanding and Preparation***
5. ***Modeling and Evaluation***

1-) Analytic Approach:

• Machine Learning approach will be used to address the problem

• Classification model is used for spam or unspam emails. Predicted the likelihood of problem and based at the dominant income.

2-) Data Requirements:

• Verified the header of emails.

• Properly format the emails data as emailed-by, signed-by, encrypted-by.

• Check the subject, and spelling mistakes in email as well as repetition of same words.

• Make dictionary of some words, if we see those special words more than some specific number, we classify the email as spam.

• One record for one email.

• One column covering all aspects of email.

3-) Data Collection:

• Collect the email data from different email-servers and use them to train your machine.

4-) Data Understanding and Preparation:

• Make the models of data i.e. Classification Models.

• Refined the solution of question.

• Filling the missing values in a data.

• Take decision about unavailable data as whether to put some value or put just NULL.

• Automating some of the data collection and preparation processes in the database.

• Choose Table Format to structure the data.

• Remove some unnecessary features from data

5-) Modeling and Evaluation:

• Predictive Modeling to classify the emails.

• Constant refinement, adjustments and tweaking are necessary within each step to ensure the right outcome.

• Improve the model by giving more spammed email to train the machine more.

• Do statistical significance testing.

• Find the optimum values of the parameter

• Determine the optimal model

• Find the optimal model by ROC curve by check the separation of lines.

• Ensure the data is properly handled and cleaned.