**Project Title**

|  |
| --- |
| *Data Science Methodology case study Emails* |

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

|  |
| --- |
| *I have selected “Emails” as my data science methodology topic. This will be the case study on the classification of emails, whether the email is spam or not.* |

Your answer needs to be a little bit longer. Write a few sentences to complete your assignment.

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?".

|  |
| --- |
| *We receive a lot of emails daily, but the main problem is that all the emails received are not from genuine sources, a lot of them are spam. So, we need to classify emails, whether it is spam or not. So, the question would be “can we automatically determine whether the email is spam or not?”* |

Your answer needs to be a little bit longer. Write a few sentences to complete your assignment.

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 Approach
2. Data Requirements
3. Data Collection
4. Data Understanding and Preparation
5. Modelling and Evaluation

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

|  |
| --- |
| 1. *Analytic Approach – Here we need to classify email into two parts: the email received is spam or not spam. So, it’ll be a yes/no type of answer. Therefore, we need to use the classification model here. If the result is yes then the received email is spam, and if the result is no, it is not spam.* 2. *Data Requirements – To create a classification model we need data that contains all types of emails, this data should contain all the emails that are spam and not spam as well.* 3. *Data Collection – We know what type of data we’ll require for our model. For collecting data we’ll collect emails from our mailboxes. For large quantities, we’ll also collect different email samples from the internet and other resources.* 4. *Data Understanding and Preparation – Now we have our data, but do we require all the collected data or some of the unnecessary data can be dropped. We will understand our data and the insights from the data. We will remove duplicate emails if any are present in the data. We will try to add emails with different types like emails containing discount offers, promotions, advertisements, etc. And finally, will prepare our data accordingly.* 5. *Modelling and Evaluation – Now we will try to develop our model using different algorithms available. We will create a classification model and lastly will evaluate the outcome and perform corresponding changes until we have a suitable model. We can compare the spam emails that the model predicts to the actual spam emails to evaluate the result.* |