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

Reducing the outpatient department (OPD) duration at hospitals with data science by better management of staff time for more serious patients by initial illness diagnosis at home, in developing countries.

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?".

Government Hospitals in developing countries are often overcrowded with patients and are understaffed. Often the reason is people cannot afford the care in private hospitals (provided for free by the government and insurance is still far from reach of common people) and the government cannot afford to hire permanent staff at such hospitals. Heres an example of a government hospital in Nepal facing same trouble:   
<https://kathmandupost.com/health/2019/08/24/how-nepal-s-oldest-hospital-and-the-government-that-runs-it-continue-to-fail-the-country-s-poor>  
  
What can we do to reduce the number of patients at such hospitals or use the time of available staff more effectively?  
If we can reduce the number of patients reaching the hospital for OPD (Outpatient Department) Doctors will be able to provide more time to patients in need of intimate consultation.  
  
But when people use search engines like google to perform diagnosis, it is often wrong. Some of the major reasons are patient being unaware of the right questions to ask, neglecting treatment history, information from a person lacking credibility or insufficient information which may mislead the patient.  
  
Can initial diagnosis be performed at home?  
Yes, the answer is symptoms but evaluated in a controlled environment keeping external factors like environment(temperature, mosquitoes,etc.) in mind.  
  
By performing diagnoses at home or at health posts we can decrease the number of patients reaching government hospitals for OPD services.

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. Modeling and Evaluation

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

1. Analytical approach Symptoms are very much like Ingredients in a recipe but instead help to diagnose an Illness instead. So, the machine learning algorithm, decision trees may be used. Features like fever, loose stool, etc. can be used to split the date and gain more information. Moreover, the tree can also be used to suggest certain laboratory tests, instead of pointing towards a disease when certain symptoms are evident. 2. Data Requirements In the Business Understanding and Analytical Approach stage, we determined that automating the process of initial diagnosis or laboratory test suggestion is potentially possible using the symptoms the patient is showing. To build a model, we need extensive data on different diseases and their symptoms including lab tests, patient's features(age, height, etc) that are to be performed for confirmation. Data about steps followed in diagnosis can also be helpful. 3. Data Collection Hospitals have always recorded their patient's(age, height, weight,etc.) information electronically or on paper along with a series of steps in the diagnosis, medications provided and lab tests suggested. These pieces of information can be helpful for diagnosis and suggesting medicines and laboratory tests. These pieces of information can be obtained from hospitals (private or public, nationwide or from all over the world) 4. Data Understanding and Preparation Data Understanding: The data can be stored in tabular form where each row represents illness and for each illness, the corresponding presence of symptoms, age groups, weight groups and need for laboratory correlations are documented. We know that symptoms of diabetes include: 1. Increased thirst 2. Frequent urination 3. Extreme hunger 4. Unexplained weight loss 5. Presence of ketones in the urine (ketones are a byproduct of the breakdown of muscle and fat that happens when there's not enough available insulin) 6. Fatigue 7. irritability 8. Blurred vision 9. Slow-healing sores source: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444 So maybe if a patient contains all these symptoms, then we can say that the illness is diabetes. Data Preparation: In this stage, data is prepared for the next stage in the data science methodology, which is modeling. This stage involves exploring the data further and making sure that it is in the right format for the machine learning algorithm that we selected in the analytic approach stage, which is decision trees. i. First, look at the data to see if it needs cleaning For example, the same illness may be known by different names in different languages spoken in different regions of the country and we would have to change all those names to one more recognized name. ii. Convert all Yes's to 1's and the No's to 0's 5. Modeling and Evaluation Modeling: The decision tree can be used to form rules to model the data that can be used to predict an illness : For instance: i. If symptoms contain fever and wound and no other problem, then it is most likely an infection of the wound. ii. If symptoms contain fever but no wound and no other problem, then it is most likely Tiredness. We can analyze all the branches of the tree to come up with similar rules for determining the disease for different diseases. Evaluation: To evaluate our model of Illness, we will split our dataset into a training set and a test set. We will build the decision tree using the training set. Then, we will test the model on the test set and compare the illness that the model predicts to the actual illness. To quantify how well the decision tree is able to determine the illness for different symptoms correctly, we will create a confusion matrix that presents a nice summary of how many illnesses from symptoms were correctly classified. It also sheds some light on what illnesses are being confused with what other illnesses.