Data Science Methodology (Problem Solving approach)

The aspect of data science is an area which seek to ensure that problems in organizations are solved using the modern machine learning skills and expertise. The new trend in data analysis is the use of machine learning to analyze data. However, we want to examine the data science methodology prescribed by **John B. Rollins**, a Data Scientist for IBM analytics. These questions will guide us through this topic and will help us develop a problem solving analytical approach in the business environment. The highlighted headings will guide our answers to the questions outlined below.

According to Rollins, the Data science Methodology aims to answer 10 questions in the following order:

1. What is the problem you are trying to solve? That is, ***problem approach***.

Here, the problem at hand is specified in a language that is understandable to every stakeholder interested in the project. The problem is explained and the data scientist must get to the root of the problem to be solved. Moreover, the problem must be asked in form of a question so as to make it easier for people to understand. We need a title for the problem, please don't forget.

Let us look at this case study:

Title: **Effect of overcrowding and congestion on hospital service delivery in Bowen University Hospital**.

However, we must give a brief explanation of the problem.

The problem that this study seek to solve bothers on the increased number of patients who have to wait for a long time before they see a medical personnel especially doctors. There have been complaints from people concerning the hospital space, number of medical personnel, long queues, and increased waiting time. All these can be summarized into two words which are overcrowding and congestion.

The next thing is to have a question that summarizes the problem, that is, the problem must be formulated into a question. The question in this regard will be:

How can overcrowding and congestion be reduced in hospitals to ensure efficient service delivery?

The next step is to select an ***analytical approach*** based on the question formulated.

There are different analytical approaches to be used, nevertheless, they must be in line with the question asked. They include:

Descriptive (deals with the current situation) approach: use descriptive model if the question is to show relationships.

Diagnostics (statistical analysis approach): this approach as questions about what happened and why the situation is happening.

Predictive (deals with forecasting): Questions that are asked here include: What if these trends continues? What will happen next? Use predictive model if the question is to determine the probability of an action.

Prescriptive: this approach deals with how to solve a problem.

Classification approach deals with situation where the question requires a yes/no answer to predicting situation.

*****The correct approach to use depends on the business requirements for the model**.

2. How can you use data to answer the question?
3. What data do you need to answer the question?
4. Where is the data coming from (identify all sources) and how will you get it?

Here, we must specify the **data requirements** for the problem we are investigating.

Data requirements involves asking questions about specific data needed for the study, sourcing for the data, how to understand or work with them and how to prepare the data to meet the desired outcome?

The data needed for our case study will include the following:

a. The number of medical personnel (doctors, nurses, medical lab scientist, pharmacists and so on)
b. The number of bed spaces available
c. The record of patients which will include their names, arrival time, waiting time, time spent for consultation and departure time.
d. The space available for patient to stay before they are being attended to.

These are the specific area where data is needed, the next step is to determine how to source for the data (data gathering). As regards our case study, the data will be sourced

from the hospital in-patient record system at the general outpatient department (GOPD) record section and the human resource unit (where data on medical personnel will be collected).

5. Is the data that you have collected representative of the problem to be solved?

To answer question 5, you must **understand the data** you've collected. This can be done through:

Descriptive Statistics: univariate statistics (mean, median, minimum, maximum, SD), pairwise correlations (close correlation and look for highly correlated variables), histogram (shows how values or variables are distributed). This will help you examine the data and know whether you have representative data or need to expand your scope.

Data quality must also be assessed: univariate and histograms are used to assess the quality of a data set. Can also be used to identify missing values, invalid or misleading values.

Data Preparation is the next phase which involves: data cleansing and transformation.

Ways by which data are prepared include: removing invalid values, missing data, remove duplicates, and formatting. Moreover, feature engineering (this is critical when Machine Learning is used to analyze data) and text analysis can also be carried out.

Additional literature review of important factors can be carried out in cases where additional information about the problem is needed.

6. What additional work is required to manipulate and work with the data?

This question can be answered by completing **the data set** which can be carried out by merging all data into one table so as to have a comprehensive data set for analysis.

7. In what way can the data be visualized to get to the answer that is required?

Modelling seek to answer the question: what way can the data be visualized to get to the answer that is required?

Two questions are vital here: (1) what is the purpose of data modelling? (2) What are some characteristics of this process?

Data modelling focuses on developing models that are either **predictive** (yes or no or stop/go outcomes) or **descriptive (**. These models are based on the analytical approach selected either statistically driven or machine learning driven.

Data modelling focuses on developing models that are either **predictive model** (yes or no or stop goal outcomes) or **descriptive model**; these models are based on the analytical approach selected either statistically driven or machine learning driven.

Training set is a set of historical data in which the outcome is already known and are used for a predictive model. It determines whether the model need to be calibrated or not. A test set can be used for a descriptive model.

Successive data compilation, preparation and modelling depends on the understanding the question at hand, select an analytical approach or method to solve the problem and obtain, understand, prepare and model the data according to John Rollins descriptive data science methodology framework. It is just to build a model and answer the question.

8. Does the model used really answer the initial question or does it need to be adjusted?

Evaluation seek to answer the question: Does the model used really answer the initial question or does it need to be adjusted?

Evaluation! A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are done iteratively. Model evaluation **is performed during model development and before the model is deployed**. Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.

Model evaluation can have two main phases:

The **first** is the **diagnostic measures phase**, which is used to ensure the model is working as intended. **If the model is a predictive model, a decision tree** can be used to evaluate if the answer the model can output, is aligned to the initial design. It can be used to see where there are areas that require adjustments. Please note that decision tree is not the only tool that can be used here. **If the model is a descriptive model, one in which relationships are being assessed, then a testing set** with known outcomes can be applied, and the model can be refined as needed.

The **second** phase of evaluation that may be used is **statistical significance testing**. This type of evaluation can be applied to the model to ensure that the data is being

properly handled and interpreted within the model. This is designed to avoid unnecessary second guessing when the answer is revealed.

9. Can you put the model into practice?

Deployment is the ninth stage of this methodology. A data science model will provide an answer, however, the usefulness of the model in giving answers to the initial question is to involve the stakeholder and ensure they are familiar with the different tools. The next thing is to put the model into a test by deploying it so far the data scientist is sure it will work. However, we must not rush things, we can decide to deploy it to a small group of users in a test environment and gradually build the trust of stakeholders all around.

10. Can you get constructive feedback into answering the question?

Obviously the last stage is the feedback point. Getting feedback as regards the model from users will help in adjusting and refining it. Furthermore, the performance of the model can be assessed to determine its impact in the organization. The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required. Throughout the Data Science Methodology, each step sets the stage for the next. Making the methodology cyclical, ensures refinement at each stage in the game. The feedback process is rooted in the notion that, the more you know, the more that you'll want to know. Once the model is evaluated and the data scientist is confident it'll work, it is deployed and put to the ultimate test: actual, real-time use in the field.

Project/assignment: Using the knowledge from what we have discussed, choose one of the following topic areas and explain how you will solve a problem.

1. Hospitals
2. Banks
3. Communication
4. Courier service

Please note that a topic cannot be chosen for more than one group. This means that a topic per group.

The topic chosen must be carried out using the following heading:

a. Title of the work
b. Explain the problem by giving brief details that will ensure a better understanding of the problem by stakeholders.
c. Formulate the problem into a single question that will guide your work

d. Select an appropriate analytical approach for the problem.
e. Data requirements (type, source and where to get it)
f. Develop a model based on the two types specified
g. Select an evaluation method for your model.