ISE-291: Introduction to Data Science

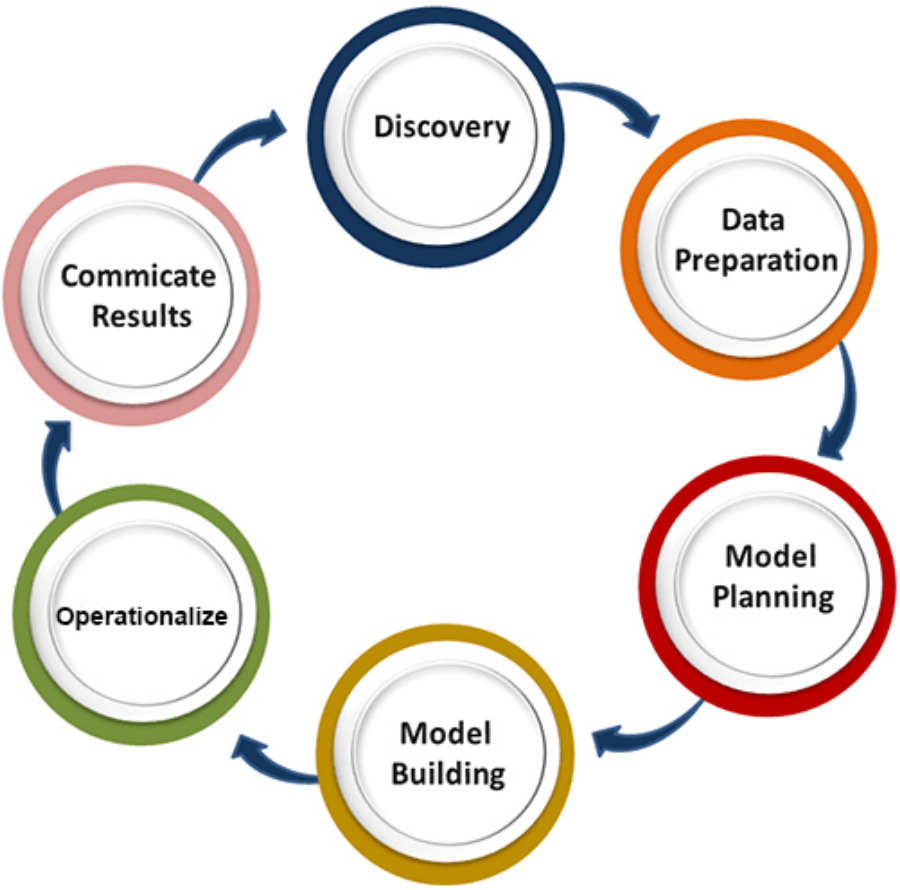
Term 231

Guidelines for the project 2023

Covers: Topics 2-9 Material **Deadline: 12th December 2023**

Problem description:

- You are a Data Scientist assigned to perform a comprehensive analysis to provide insights and predictions to the stakeholders for possible future policies/decisions.
- You are expected to establish a full design of the Data Analytics Lifecycle as follows:



Expected deliverables for the project

You have to run several analyses to fulfill the data analytics life cycle phases.

1. Discovery:

a. ***Select*** a field or topic of your interest. For example, an industrial engineering student is usually interested in supply chain applications. ***Find*** a relevant dataset that supports your topic of interest.,
b. ***State*** clearly the problem and ***set*** the objective of the analysis.

For your ease, few data websites are provided below:

Google Dataset: https://datasetsearch.research.google.com/
Kaggle: https://www.kaggle.com/datasets
Data.Gov: https://data.gov/
Datahub.io: https://datahub.io/collections
UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php
World Health Organization: https://apps.who.int/gho/data/node.home
KSA Open Data: https://data.gov.sa/en/home
Note: It is recommended to use the dataset, which may have **10 to 20 mixed variables** (i.e., numerical and categorical variables) having at **least 300 observations**.

2. Data Preparation:

a. ***Read*** the selected data, ***list*** the fields/variables, and ***identify*** their types.
b. ***List*** the inconsistencies, missing data, and outliers. ***Fix*** the inconsistencies, ***impute*** the missing data, and ***remove*** the outliers.
Note: If your data does not have inconsistencies, missing observations, or outliers, it is suggested to create them manually and apply methods to show your skills.

3. Model Planning:

a. ***Run*** the exploratory data analysis:
    i. ***Find*** the statistical summaries.
    ii. ***Make*** univariate graphs (i.e., graphs based on single variable).
    iii. ***Prepare*** bivariate plots (i.e., plots based on two variables).
    iv. ***Portray*** advanced graphs (i.e., graphs based on more than two variables).
    v. ***Assess*** the relationship between variables.
b. ***Summarise*** your findings.

In this course, we cover three model types: regression models, classification models, and clustering. ***Choose*** the suitable modeling technique for the successful achievement of your objectives and ***provide*** the justification.

4. Model Building:

a. ***Estimate*** the unknown model parameters (fitting) and ***evaluate*** the model (validation/cross-validation).

b. ***Compare*** different models based on the performance measures.
c. ***Interpret*** the findings and ***provide*** the details of the final selected model.

5. Communicate results:

a. ***Provide*** the summary and conclusion of your analysis.
b. ***Give*** possible future recommendations.

6. Operationalize:

a. ***Provide*** the general guidelines to adopt your methodology.
b. ***List*** the problems and issues in the implementation of the selected methodology.

General notes and guidelines:

i. You must prepare your report using a ***Jupyter Notebook*** and submit the report and your project's data (in CSV or XLS format).

ii. The report's structure should be outlined so that each section addresses one of the above-mentioned deliverables. Your work will be assessed on both the quality of the content as well as the presentation of your material.

It is crucial to support your analysis and choices by researching and citing your sources. State any assumptions you make during your analysis and justify them. The credibility of your analysis hinges on how thorough your research is.

Note: Plagiarism or copying of the project from any source is not allowed. If we find plagiarised and copied text, then it will be accounted as misconduct. Hence, the respective actions will be taken, and the students are solely responsible for any decision made by the authorities.

---------- **Best Wishes** ----------