Learning Outcomes to be assessed

1.	Analyze a dataset from a problem domain in depth, and select appropriate statistical models, tools, and techniques to derive insights regarding the dataset and domain.
2.	Effectively extract, transform, interrogate, and analyze large datasets.
3.	Construct, refine, interpret, and critically evaluate predictive analytical and machine learning models.
4.	Critically evaluate and utilize hyperparameter search strategies for optimizing machine learning models.

Supervised Machine Learning – Classification (100 Marks)

Dataset

Each row in ***QualityPrediction.csv*** corresponds to a drinking water sample taken from a different urban location. Independent variables correspond to content of different elements present in water. Target variable 'is_safe' tells us whether the drinking water is safe for drinking or not.

Task

Using the above dataset, Municipal Corporation in a particular city wants to construct a classification model that can label any new water sample as either safe or unsafe for drinking. Construct such a model in Python by trying two classification algorithms - random forest and logistic regression.

In addition to providing the python code, you are required to provide *critical analysis* of your approach and results in a pdf report. [*Important* – Critical analysis does not mean merely describing things. It means discussing the why behind doing things]

Your code and analysis should cover the following points:

1. Data Preparation (What steps would you take to prepare your data and why?) [20]

2. Model Hyperparameter Tuning (Which hyperparameters would you tune and why? How would you tune them?) [20]

3. Choice of Evaluation Metric (Which metric would be suitable for model evaluation and why?) [20]

4. Overfitting avoidance mechanism (Which mechanism (feature Selection/ regularization) would you use and why?) [20]

5. Results analysis
   a). Which of the two models (random forest or logistic regression) would you recommend for deployment in the real-world?
   b). Is any model underfitting? If yes, what could be the possible reasons?
   [20]

You must submit the following in a zipped folder:

1. Critical Analysis Report (.pdf)

2. Python Code (.py)

Naming convention:

Report should be named as –

Report_Firstname_Surname.pdf

Code should be named as –

Code_Firstname_Surname.py

Zipped folder should be named as –

Firstname_Surname.zip

There is no prescribed word-count for the report. It will be assessed on quality, and not quantity of content.

Assessment Criteria

Each part will be graded according to the following criteria:

1. Quality of code (correctness and completeness) [Weightage – 80%]

2. Quality of analysis in report (critical analysis of approach, presentation and interpretation of results, conclusion) [Weightage - 20%]