**BLM19432E Introduction to Data Science Final Project**

Please submit all your project files (data + jupyter code file + project document file) in a single archive file like: **name\_surname.zip**

|  |  |
| --- | --- |
| **Name Surname:** |  |
| **Dataset definition:** |  |
| **Dataset source**  **(web address):** |  |
| **Aim of the project:** |  |

**Step0: Project Proposal – 10p**

1. Find your project data on [Kaggle](https://www.kaggle.com/datasets), [UCI](https://archive.ics.uci.edu/ml/index.php) etc.
2. Explain why you have chosen to work on this data set?
3. Do basic exploratory data analysis (EDA) to understand your data.
4. Briefly explain your aim. Which machine learning model do you to decide and why?

**Step1: Exploratory Data Analysis – 10p**

1. Explain the shape of the dataset (restriction of at least 10 columns, 1000 rows)
2. Explain the column types in the dataset
3. Explain the distribution of only 2 features in your dataset (one numeric, one categorical) using visualizations

**Step2: Preprocessing – 10p**

1. How many columns include missing values
2. Explain your method to handle each of those missing values
3. Explain if you needed to apply any kind of transformations.

**INTERIM REPORT SUBMISSION – April 26**

**Step3: Feature Engineering – 10p**

1. Determine which features are the most valuable and whether you need to create a new feature?
2. Select (i.e., filter) or create features that make machine learning algorithms work
3. Explain your aim in detail.

**Step4: Machine Learning Problem – 10p**

1. Select the machine learning problem (Regression, Classification or Clustering) that is appropriate for your data set.
2. Select 2 evaluation methods that you will use in your project, and explain them in detail by giving proper formulation.
3. Show example visualizations associated with your selected problem.

**Step5: Algorithms, Implementation and Performance Comparison – 35p**

1. Select 2 machine learning algorithm. For example: if you choose a classification problem, you can use Logistic Regression and Naïve Bayes Classifier.
2. Give brief definition of each of two algorithms along with the parameters need to be tuned.
3. Select your base evaluation technique. Explain why did you choose this method?
4. Using your base evaluation technique as the quality criterion, try to find the best parameters for each of your clustering method.
5. Compare the two algorithms based on the selected evaluation technique, numerically and also visually.
6. Which algorithm performs the best? Can you explain why?

**Step6: Further Performance Improvement (Your best algorithm) – 15p**

1. Remove outliers (if there are any) in your data. Check if this improved your algorithm’s performance. Show before and after performance of your algorithm (both visually and numerically)
2. Apply a feature selection method. Check if this improved your algorithm’s performance. Show before and after performance of your algorithm (both visually and numerically)
3. Can you propose any other improvement techniques?

**Step7: Inference – 10p**

Give comments on your findings. Did you achieve your goal in the project?