

---

<b>Module Title:</b>	Machine Learning I
<b>Assessment Type:</b>	Group practical and report
<b>Assessment Title:</b>	CA2
<b>Release Date:</b>	7 <sup>th</sup> May 2021
<b>Submission date:</b>	13 <sup>th</sup> June 2021
<b>Assignment Compiler:</b>	Dr. Muhammad Iqbal
<b>Method of Submission:</b>	Upload one zip file (Report, code and datasets) submitted to Moodle
<b>Group/ Individual:</b>	Group (One person can upload on Moodle)

---

**Learning Outcomes Assessed:** Machine Learning I

**List the module learning outcomes to be assessed (delete as necessary)**

**MLO 1** - Employ data mining frameworks to the solving of analytical problems

(Linked to PLO2)

**MLO 4** - Explore a range of classification and regression techniques and ascertain their suitability for a variety of problem domains.

(Linked to PLO 5)

**MLO 5** - Evaluate and optimize the performance of classification and regression models

(Linked to PLO 3)

This is a group-based project (2 – 4 students) using PYTHON programming language to analyse a specific problem in the following areas, such as Pharmacy, Library, Holiday Booking System, Medical Practice, Concert hall, Motor mechanic, Sales, Customers behaviour, Primary School, Role-playing game and manufacturing etc. Your group may choose data of any other category based on your interest from Kaggle ([www.kaggle.com](http://www.kaggle.com)) or UCI (<https://archive.ics.uci.edu/ml/index.php>) or any other repository. The dataset should have at least 4000 rows and 10 columns (for example, type of variables may be categorical, continuous and discrete) after cleaning and there is not any maximum limit. Your group would need to formulate a set of objectives in the domain of chosen dataset and the ML project should address the achievement of these objectives. Fundamentally, the objectives should provide a clear outline about your project. For example, which features are the most important for predicting target variable (X)? You can start with a simple approach so you can achieve something quickly and then progress to more complicated approaches during this group project.

The group should consider the following guidelines during the development of Machine Learning (ML) project.

1. Justification for the selection of machine learning approaches for the chosen problem using any data mining framework, such as CRISP-DM, KDD or SEMMA for the implementation.

2. For ML techniques (Classification, Regression and Clustering), you should plan on trying multiple approaches (at least two), with proper parameter-selection techniques and a comparison between the chosen modelling approaches.
3. You should train ML modelling techniques and subsequently, test the models. Perform a comparison of two or more ML modelling techniques. You may use a statistical approach to argue that one feature is more important than some other feature.
4. Depending on the complexity of the problem, you should use cross-validation approach to justify the authenticity of your ML modelling results.

Your group will present the findings and defend the results in the report (MS Doc/ pdf or any other readable format). Your report should capture the following aspects that are relevant to your approach.

- i. Brief description and motivation of the problem for Machine Learning.  
(250 words, 10 marks)
- ii. What is/are the objectives of the problem(s) that are addressed in your project (Classification/ Regression/ Clustering Rules/ Information extraction etc..)  
(100 words, 10 marks)
- iii. Characterization of the data set: source URLs; size; number of attributes; has/ does not have missing values; number of examples etc. Clean and remove the missing values from the dataset. Provide a clear strategy.  
(100 words, 10 marks)
- iv. Train the ML models based on three different splits and discuss the variation in accuracy/ score obtained from the models in the training as well as testing.  
(400 words, 30 marks)
- v. Interpret the results based on problem specification and objectives. The ML modelling results should neither overfitted nor underfitted. Justify with arguments.  
(500 – 750 words, 20 marks)
- vi. Provide the explanation of code that will be used to solve the problem. Comments must be provided along with code.  
(200 words, 10 marks)
- viii. Conclusions based on the predictions and classification. Harvard style citations and References must be provided in the report.  
(200 words, 10 marks)

#### **Assessment Details**

**Provide a full summary of the assessment**

- The code and datasets should be provided in zip format.
- Maximum Number of Words for the report (2000 words excluding diagrams, code and HARVARD style References).
- Must be clearly specified the number of words used in the report.
- Describe the contribution of each team member in the project clearly and use a bar chart to represent the effort and time spent during this project.
- The rubric is provided for the detailed breakdown of marks.

**Note:** The names of group members must be uploaded on the link provided on Moodle until 11<sup>th</sup> May 2021 (23:59).

GRADING RUBRIC – Machine Learning I - 2021					
CRITERION	H1 ( $\geq 70\%$ )	H2.1 ( $\geq 60$ and $< 70$ )	H2.2 ( $\geq 50$ and $< 60$ )	PASS ( $\geq 40$ and $< 50$ )	FAIL ( $< 40$ )
Introduction to problem Description and Motivation (10%)	An excellent introduction to problem description and motivation that provide a precise and clear case for the proposed Machine Learning project.	A very good introduction to problem description and motivation that provide offers a very convincing case for the proposed Machine Learning project.	A good introduction to problem description and motivation that furnishes a largely convincing case for the proposed Machine Learning Project.	An adequate introduction to problem description and motivation that offers a somewhat weak case for the proposed Machine Learning Project.	A poor introduction to problem description and motivation that fails to motivate the problem or provide a case for the proposed Machine Learning Project.
Project Objectives (10%)	An excellent specification of objectives succinctly.	A very good specification of objectives.	A good specification of objectives.	An adequate specification of objectives.	A poor specification of objectives.
Characterization and cleaning of Dataset (10%)	An excellent characterization and cleaning of dataset that summarizes all details from source to fields.	A very good characterization and cleaning of dataset that summarizes all details from source to fields.	A good characterization and cleaning of dataset that summarizes all details from source to fields.	An adequate characterization and cleaning of dataset that summarizes all details from source to fields.	A poor characterization and cleaning of dataset that summarizes all details from source to fields.
Training and Testing of Models (30%)	An excellent accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is used to test the generalizability of the model and It should justify the results in an excellent way.	A very good accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is used to test the partial generalizability of the model and It should justify the results.	A good accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is used to test the partial generalizability of the model.	An adequate accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is used.	A poor accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is not used.
Interpretation of results (20%)	An excellent interpretation and explanation of the results based on problem specification and objectives. The results clearly state that the models are neither overfitted nor underfitted. An excellent justification is provided.	A very good interpretation and explanation of the results based on problem specification and objectives. The results state that the models are neither overfitted nor underfitted. A very good justification is provided.	A good interpretation and explanation of the results based on problem specification and objectives. The results state that the models are overfitted but not underfitted. A good justification is provided.	An adequate interpretation and explanation of the results based on problem specification and objectives. The results state that the models are adequate. An adequate justification is provided.	A poor interpretation and explanation of the results based on problem specification and objectives. No clear results obtained.
Code description and comments (10%)	An excellent description of code using comments. The comments are detailed and provide an explicit understanding of the functionality of the code.	A very good description of code using comments. The comments are brief and provide a clear understanding of the functionality of the code.	A good description of code using comments. The comments are very brief and provide an understanding of the functionality of the code.	An adequate description of code using comments. The comments are not satisfactory and provide a partial understanding of the functionality of the code.	A poor description of code using comments. The comments are not satisfactory.
Conclusions, citations, and references (10%)	An excellent demonstration of conclusions. An excellent report along with proper citations and references in all sections.	A very good demonstration of conclusions. A very good report along with proper citations and references in all sections.	A good demonstration of conclusions. A good report along with citations and references in some sections.	An adequate demonstration of conclusions. An adequate report along with incomplete citations and references.	A poor demonstration of conclusions or no conclusions. A report along with errors.