CCT College Dublin Continuous Assessment

Programme Title:	HDip in Science in Data Analytics for Business		
Cohort:	FT/ PT		
Module Title(s):	Machine Learning (10 ETCS)		
Assignment Type:	Individual	**Weighting(s):**	50%
Assignment Title:	CA1 Project		
Lecturer(s):	Dr. Muhammad Iqbal		
Issue Date:	6th March 2024		
Submission Deadline Date:	21st April 2024		
Late Submission Penalty:	Late submissions will be accepted up to **5** calendar days after the deadline. All late submissions are subject to a penalty of **10%** of the mark awarded. Submissions received more than 5 calendar days after the deadline above **will not** be accepted and a mark of 0% will be awarded.		
Method of Submission:	**Moodle**		
Instructions for Submission:	Upload separate files including MS word file, jupyter notebook, dataset and any supporting information on Moodle.		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	Three weeks after submission		

Learning Outcomes:
Please note this is not the assessment task. The task to be completed is detailed on the next page.
This CA will assess student attainment of the following minimum intended learning outcomes:

1. Develop a machine learning strategy for a given domain, communicate this strategy effectively to team members, peers and project stakeholders (CRISP-DM)
(Linked to PLO 1, PLO 4, PLO 6)
2. Implement a range of classification and regression techniques and detail /document their suitability for a variety of problem domains.
(Linked to PLO 5)
3. Critically evaluate and optimise the performance of Machine Learning models.
(Linked to PLO 3)

Attainment of the learning outcomes is the minimum requirement to achieve a Pass mark (40%). Higher marks are awarded where there is evidence of achievement beyond this, in accordance with QQI *Assessment and Standards, Revised 2013*, and summarised in the following table:

Percentage Range	**CCT Performance Description**	**QQI Description of Attainment**
		Level 6, 7 & 8 awards
90% +	Exceptional	Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this
80 – 89%	Outstanding	
70 – 79%	Excellent	
60 – 69%	Very Good	Achievement includes that required for a Pass and in **many** respects is significantly beyond this
50 – 59%	Good	Achievement includes that required for a Pass and in **some** respects is significantly beyond this
40 – 49%	Acceptable	Attains all the minimum intended programme learning outcomes
35 – 39%	Fail	Nearly (but not quite) attains the relevant minimum intended learning outcomes
0 – 34%	Fail	Does not attain some or all of the minimum intended learning outcomes

Please review the CCT Grade Descriptor available on the module Moodle page for a detailed description of the standard of work required for each grade band.

The grading system in CCT is the QQI percentage grading system and is in common use in higher education institutions in Ireland. The pass mark and thresholds for different grade bands may be different from what you have experienced in the higher education system in other countries. CCT grades must be considered in the context of the grading system in Irish higher education and not assumed to represent the same standard the percentage grade reflects when awarded in an international context.

Assessment Task

This is a project for machine learning using the PYTHON programming language. Develop and deploy machine learning models in any one of the following areas only, analyse and subsequently interpret the results.

- Education
- Justice, Legal system, and Public Safety
- Housing and Zoning

You can find any public dataset from an authentic resource repository and the dataset should have at least 5 columns after cleaning and more than 300 rows.

The type of question(s) that you should formulate for the project will depend on the chosen area of the dataset that you are considering for the machine learning project.

Suggested possible analysis / project questions are mentioned below (this is a small, suggested, sample of questions, other questions may be more appropriate to your project)

- What are the most important features for predicting X as a target variable?
- Which classification approach do you prefer for the prediction of X as a target variable, and why?
- How to classify the loyal and churn customers using Support Vector Machines?
- Why is dimensionality reduction important in machine learning?

The student would need to consider the following instructions (a - d) during the development of this project.

a) Logical justification based on the reasoning for the specific choice of machine learning approaches.

b) Multiple machine learning models (at least two) using hyperparameters and a comparison between the chosen modelling approaches.

c) Visualise your comparison of ML modelling outcomes. You may use a statistical approach to argue that one feature is more important than other features.

d) Cross-validation methods should be used to justify the authenticity of your ML results.

You will present the findings and defend the results in the report (MS Doc) by highlighting your work. Your report should capture the following aspects that are relevant to your project investigations.

1. A clear introduction, motivation, a description of the problem domain, and an explanation of how the project's goals are justified using Prediction / Classification algorithms.

   (20 marks)

2. Characterization of data, pre-processing, explanation and description of techniques used for the variation in the accuracy across three training splits (20%, 25% and 30%) using cross validation techniques.

   (30 marks)

3. What is the primary purpose of hyperparameter tuning in machine learning? Could you elaborate on specific hyperparameter tuning techniques (e.g., GridSearchCV) applied to machine learning models to find optimal parameters?

   (25 marks)

4. Interpret and explain the results obtained, discuss overfitting / underfitting / generalisation, provide a rationale for the chosen models and use visualisations to support your findings. Comments in Python code, conclusions of the project should be specified at the end of the report. Harvard Style must be used for citations and references.

   (25 marks)

Submission Requirements

All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the marks awarded.

- All files should be uploaded separately on Moodle.
- Clearly detail the number of words used in the report.
- Number of Words in the report (1250 words +/-10%) excluding diagrams, code, references, citations and titles.
- Use version control like Github or any other tool to show the progress in CA1. You should have at least 5 commits on Github before submission.
- The rubric is provided for the detailed breakdown of marks at the end of this CA.
- Use Harvard Referencing when citing third party material.
- Be the student’s own work.
- Include the CCT assessment cover page.
- Be submitted by the deadline date specified or be subject to late submission penalties.
- Must be clearly specified the number of words used after each section in the report.

Acceptable Use of AI for Assignment at CCT

Acceptable and Unacceptable Use of AI	● The use of generative AI tools (e.g. ChatGPT, Dall-e, etc.) is permitted in this assignment for the following activities: ○ Brainstorming and refining your ideas; ○ Fine tuning your research questions; ○ Finding information on your topic; ○ Drafting an outline to organise your thoughts; and ○ Checking grammar and style. ● The use of generative AI tools is not permitted in this course for the following activities: ○ Impersonating you in classroom context ○ Completing group work that your group has assigned to you ○ Writing a draft of a writing assignment ○ Writing entire sentences, paragraphs or papers to complete class assignments. ● You are responsible for the information you submit based on an AI query. Your use of AI tools must be properly documented and cited. ● Any assignment that is found to have used generative AI tools in an unauthorised way will be subject to college disciplinary procedures as outlined in the QA Manual. ● When in doubt about permitted usage, please ask for clarification.	This statement is useful when you are allowing the use of AI tools for certain purposes, but not for others. Adjust this statement to reflect your particular parameters of acceptable use, and your discipline context.

GRADING RUBRIC – Machine Learning - 2022/2023

GRADE	90-100%	80-90%	70-79%	60-69%	50-59%	40-49%	35-39%	<35%
Performance	**Exceptional**	**Outstanding**	**Excellent**	**Very Good**	**Good**	**Acceptable**	**Fail**	**Fail**
Introduction to problem Description, Motivation and Objectives (20%)	An exceptional introduction to problem description and motivation that provide a concise and clear case for the proposed Machine Learning project. An exceptional specification of objectives concisely.	An outstanding introduction to problem description and motivation that provide a compact and clear case for the proposed Machine Learning project. An outstanding specification of objectives precisely.	An excellent introduction to problem description and motivation that provide a precise and clear case for the proposed Machine Learning project. An excellent specification of objectives succinctly.	A very good introduction to problem description and motivation that provides a very convincing case for the proposed Machine Learning project. A very good specification of objectives.	A good introduction to problem description and motivation that furnishes a largely convincing case for the proposed Machine Learning Project. A good specification of objectives.	An acceptable introduction to problem description and motivation that offers a somewhat weak case for the proposed Machine Learning Project. An adequate specification of objectives.	A poor introduction to problem description and motivation that fails to motivate the problem or provide a case for the proposed Machine Learning Project. A poor specification of objectives.	An impecunious introduction to problem description that fails entirely to motivate the problem. An impecunious specification of objectives.
Characterization and cleaning of Dataset, Training and Testing of Models (30%)	An exceptional characterization and cleaning of a dataset that abstracts all details from source to fields. An exceptional 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 exceptional way.	An outstanding characterization and cleaning of dataset that highlights all details from source to fields. An outstanding 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 outstanding way.	An excellent characterization and cleaning of the dataset that summarizes all details from source to fields. 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 characterization and cleaning of the dataset that summarizes all details from source to fields. 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 characterization and cleaning of the dataset that summarizes all details from source to fields. 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 acceptable characterization and cleaning of the dataset that summarizes all details from source to fields. An adequate accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is used.	A poor characterization and cleaning of the dataset that summarizes all details from source to fields. A poor accuracy obtained based on the training and testing of ML models using three logical splits. Cross-validation is not used.	An impecunious characterization and cleaning of the dataset. An impecunious obtained based on the training and testing of ML models using three logical splits. Cross-validation is not used.
Purpose of hyperparameter tuning and application of hyperparameter tuning technique (25%)	Clearly articulates the primary purpose of hyperparameter tuning and explains specific hyperparameter tuning techniques applied to machine learning models, demonstrating advanced knowledge and understanding.	Provides a comprehensive explanation of the purpose of hyperparameter tuning and provides a detailed and accurate description of hyperparameter tuning techniques, showcasing a strong understanding of the topic.	Describes the purpose of hyperparameter tuning with clarity and describes hyperparameter tuning techniques with clarity, though may have some minor gaps or less detailed explanation.	Conveys a basic understanding of the purpose of hyperparameter tuning but lacks depth or may contain inaccuracies. Conveys a basic understanding of hyperparameter tuning techniques but lacks depth or may contain inaccuracies.	Mentions the purpose of hyperparameter tuning but with significant gaps or inaccuracies, demonstrating a limited understanding. Mentions hyperparameter tuning techniques but with significant gaps or inaccuracies, demonstrating a limited understanding.	Provides a vague or incomplete explanation of the purpose of hyperparameter tuning, indicating a need for improvement. Provides a vague or incomplete explanation of hyperparameter tuning techniques, indicating a need for improvement.	Contains major inaccuracies or misunderstandings about the purpose of hyperparameter tuning. Contains major inaccuracies or misunderstandings about hyperparameter tuning techniques.	No meaningful response or completely incorrect information. No meaningful response or completely incorrect information.

Interpretation of results, Code description and comments, Conclusions, citations, and references (25%)	An exceptional interpretation and explanation of the results, code description, comments, conclusions, citations, and references based on problem specification and objectives. The results clearly state that the models are neither overfitted nor underfitted. An exceptional justification is provided.	An outstanding interpretation and explanation of the results, code description, comments, conclusions, citations, and references based on problem specification and objectives. The results clearly state that the models are neither overfitted nor underfitted. An outstanding advocacy is provided.	An excellent interpretation and explanation of the results, code description, comments, conclusions, citations, and references based on problem specification and objectives. The results clearly state that the models are neither overfitted nor underfitted. An excellent defence is provided.	A very good interpretation and explanation of the results, code description, comments, conclusions, citations, and references 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, code description, comments, conclusions, citations, and references based on problem specification and objectives. The results state that the models are overfitted but not under fitted. A good justification is provided.	An acceptable interpretation and explanation of the results, code description, comments, conclusions, citations, and references 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, code description, comments, conclusions, citations, and references based on problem specification and objectives. No clear results obtained.	An impecunious interpretation of the results. No clear results obtained.