CCT College Dublin Continuous Assessment

Programme Title:	*Higher Diploma in Data Analytics for Business*		
Cohort:	*FT Sept 2024*		
Module Title(s):	*Data Preparation* *Machine Learning*		
Assignment Type:	*Individual*	**Weighting(s)**:	*50%*
Assignment Title:	*Integrated CA2*		
Lecturer(s):	*David McQuaid* *Muhammad Iqbal*		
Issue Date:	*11th November 2024*		
Submission Deadline Date:	*22nd December 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:	*You must submit a zip file containing a Word with the report and a Jupyter Notebook file with the code and the outcome*		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	*After the approval from the exams board.*		

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	**Level 9 awards**
90% +	Exceptional	Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this	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	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	Attains all the minimum intended programme learning outcomes
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	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	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 experience of 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.

Learning Outcomes:

Data Preparation

- Develop strategies, incorporating basic programming skills (input / output and basic data structures) for identifying and handling missing and out-of-range data. (linked to PLO 4)
- Programmatically implement graphical methods to identify issues within a data set (missing, out of range, dirty data). (linked to PLO 2, PLO 3)
- Engineer new features selection in data with the goal of improving the performance of machine learning models. (linked to PLO 2, PLO 4)
- Critically evaluate and implement suitable data-encoding techniques for a variety of machine learning algorithms. (linked to PLO 1, PLO 5)

Machine Learning

- MLO 1 - Implement Machine Learning Algorithms to solve analytical problems. (Linked to PLO 1, PLO 2, PLO 5)
- MLO2 - Determine whether a given data analysis problem requires the use of supervised, semi-supervised or unsupervised learning methods. Develop and implement the chosen learning method. (Linked to PLO 2, PLO 4, PLO 5)
- MLO4 - Implement a range of classification and regression techniques and detail / document their suitability for a variety of problem domains. (Linked to PLO 5)

Background:

Assume you are given the opportunity to work as a Data Scientist for a global bike rental firm that has operations in New York, London, Berlin, Chicago, San Francisco, Dublin, and other locations. The company offers three different bike kinds for rent: Standard, Mountain, and Electric. With 900 rows and a variety of features/ attributes/ columns, including City, Weather, Rider Age, Bike Model, Ride Length, Distance Covered, Rider Satisfaction, and Bike Rental Cost, the dataset is available on Moodle.

The company is interested to leverage the provided dataset to improve operational efficiency, customer satisfaction, and pricing strategies. Your task is to understand and analyze the patterns in the customer behavior and bike rental usage across different cities, weather conditions, bike models, and age groups. Your analysis will guide the company to plan for the future and make data-driven decisions on bike fleet management, pricing adjustments, and customer satisfaction improvements.

Assessment Details:

Your responsibility is to analyze the provided dataset to address key business based objectives. The analysis will be splitted into two parts: Data Preparation and Machine Learning Modeling. These analyses will help the company understand customer preferences and enhance their services.

Data Preparation

1. Characterisation of the data set: size; number of attributes; has/does not have missing values, number of observations etc.[0-10]
2. Application of Data preparation/evaluation methods (Cleaning, renaming, etc) and EDA (Exploratory Data Analysis) visualizations (plural), including a clear and concise explanation of your rationale for what you are doing with the data and why you are doing it.[0-20]
3. Apply encoding, scaling and feature engineering as and if required, detailing how and why you used these techniques and the rationale for your decisions.[0-30]
4. Explore the possibility of using dimensional reduction on the dataset. Employ both LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis) and compare the separation of classes through visualization. Explain the difference between both techniques in your own words and discuss in detail how your results may affect your analysis of classifying, clustering or prediction.[0-40]

Machine Learning

1. Provide a conceptual understanding and logical justification based on the reasoning for the specific choice of machine learning approach (supervised/ Unsupervised) for the provided data set. You can discuss the pros and cons of both approaches based on your understanding. Make a set of independent and dependent variables based on your objectives for ML tasks. (0-20 marks)

2. Machine Learning models can be used for Prediction, Classification, and Clustering. You can choose suitable features for the machine learning models based on feature selection methods, such as random forest or any other method. The selection of hyperparameters for the ML models should be performed by using hyperparameter tuning, such as GridSearchCV. Obtain the best accuracy using optimal values of the hyperparameters. (0-30 marks)
3. You should train and test the Machine learning models in the case of supervised learning for different splits (at least 2 splits) and use appropriate metrics for unsupervised learning. Use k-fold (10 or 20 or 30) cross-validation to provide authenticity of the modelling outcomes. (0-30 marks)
1. Exhibit a comparison of ML modelling outcomes using a Table or graphical visualisation. Identify the possible similarities and contrasts in the Machine Learning modelling outcomes based on chosen performance metric and discuss their statistical implications. (0-20 marks)

Submission Requirements

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

All assessment submissions must:

- Maximum number of words should be in the range of 2000±5%.
- Add required word count if applicable.
- Submission the report in Word format and. IPYNB file containing the code. The name of the file should be Hdip_YOUR_NAME_Year.
- Be submitted by the deadline date specified or be subject to late submission penalties.
- Use Harvard Referencing when citing third party material.
- Be the student’s own work.
- Include the CCT assessment cover page.
- Students must use the classroom GitHub link (https://classroom.github.com/a/ywEzhmog) and you should have more than 7 commits on GitHub on different dates and times.
- Use of Gen AI (chatGPT or Copilot or others) must be clearly stated for any part of your assignment.

Statement of Acceptable Use of Artificial Intelligence

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
  - Generating code for your assignment
  - 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.

Data Preparation

Marking Criteria	**Weighting**	**Excellent (+70%)**	**Very Good (60-69%)**	**Good (50-59%)**	**Acceptable (40-49%)**	**Fail (<39%)**
Data Preparation	**0-10**	Thoroughly describes the dataset's characteristics (size, attributes, observations, and missing values) with high precision.	Accurately describes the dataset's characteristics, with minor omissions.	Provides a good overview of dataset characteristics with some relevant details.	Basic dataset description provided, lacking key details in one or more areas.	Insufficient or incorrect description of the dataset characteristics.
Data Preparation/Evaluation Methods & EDA	**0-20**	Clear, comprehensive application of data preparation techniques and EDA visualizations, with a well-justified rationale for each method used.	Strong use of data preparation and EDA with mostly clear rationale for the chosen methods.	Uses good data preparation and EDA methods; rationale provided, though limited in depth.	Basic application of data preparation methods with minimal or unclear rationale.	Limited or no use of data preparation or EDA methods; lacks clear rationale.
Encoding, Scaling & Feature Engineering	**0-30**	Effective and well-explained application of encoding, scaling, and feature engineering, with detailed rationale for each technique chosen.	Correct application of encoding, scaling, and feature engineering, with mostly clear rationale for each technique.	Good application of encoding, scaling, and feature engineering with some rationale for chosen techniques.	Basic use of encoding, scaling, or feature engineering; limited explanation or rationale.	Little or no evidence of encoding, scaling, or feature engineering; lacks rationale.
Dimensional Reduction (LDA & PCA) & Visualization	**0-40**	Comprehensive comparison of LDA and PCA, with well-illustrated visualizations and clear, detailed explanation of how each technique affects data analysis.	Strong comparison of LDA and PCA with well-presented visualizations and a good explanation of their effects on data analysis.	Good comparison of LDA and PCA with visualizations; provides a basic explanation of each technique's effect on data analysis.	Limited comparison of LDA and PCA with basic or unclear visualizations; explanation lacks detail.	Minimal or incorrect comparison of LDA and PCA; lacks meaningful visualizations or explanation.

Machine Learning

Marking Criteria	**Weighting**	**Excellent (+70%)**	**Very Good (60-69%)**	**Good (50-59%)**	**Acceptable (40-49%)**	**Fail (<39%)**
Conceptual Understanding of ML Approaches	20 Marks	Comprehensive conceptual understanding of supervised and unsupervised learning, well-reasoned justification for chosen approach, insightful discussion of pros and cons, clear identification of independent and dependent variables relevant to ML objectives.	Very good understanding of ML approaches with logical justification for choice, very good discussion of pros and cons, independent and dependent variables identified appropriately.	Good understanding of ML approaches, some justification for choice may lack depth, pros and cons mentioned but insufficient detail, variables identified but may not align well with objectives.	Basic understanding of ML approaches, weak or unclear justification for choice, minimal discussion of pros and cons, independent and dependent variables may be poorly defined.	Poor or missing understanding of ML approaches, no justification for choice, little to no discussion of pros and cons; independent and dependent variables not identified.
Feature Selection and Hyperparameter Tuning	30 Marks	Thorough selection of relevant features using appropriate methods, excellent execution of hyperparameter tuning (e.g., GridSearchCV), optimal hyperparameter values identified and leading to the best accuracy with clear justification.	Very good feature selection with relevant methods, effective hyperparameter tuning performed, and optimal values identified with some justification for accuracy achieved.	Good feature selection with limited methods, hyperparameter tuning attempted but may lack thoroughness, and accuracy results present but without clear justification.	Basic feature selection evident, hyperparameter tuning poorly executed or missing, and accuracy results not clearly linked to the chosen parameters.	Poor or missing feature selection, no evidence of hyperparameter tuning, and accuracy results not provided or lack relevance to the task.
Training and Testing ML Models	30 Marks	Comprehensive training and testing of ML models with multiple splits, appropriate metrics used for unsupervised learning, robust k-fold cross-validation applied (10, 20, or 30), clear documentation of outcomes and methods used.	Very good training and testing with at least two splits, relevant metrics used, k-fold cross-validation applied,and more details required in documentation of methods and outcomes.	Good training and testing of ML models, at least two splits attempted, metrics used but may be appropriate to some extent, and k-fold cross-validation mentioned but required clarity.	Basic training and testing, incomplete splits, metrics may be poorly defined or not appropriate, and minimal reference to k-fold cross-validation.	Poor or missing training/ testing of ML models, no splits, metrics not provided, and no proper evidence of k-fold cross-validation.
Comparison and Discussion of ML Outcomes	20 Marks	Detailed and insightful comparison of ML modeling outcomes using tables/graphs, thorough identification of similarities and contrasts, excellent discussion of statistical understanding and implications of results.	Very good comparison of outcomes with appropriate visualizations, some identification of similarities and contrasts, and reasonable discussion of statistical understanding and implications of results.	Good comparison present, but required more depth, visualizations may be more clear, limited discussion of similarities/contrasts, and minimal statistical understanding demonstrated.	Minimal comparison with vague or unclear visualizations, incomplete discussion of similarities/contrasts, and little to no statistical insight provided.	Poor or missing comparison, no visualizations provided, no discussion of similarities or contrasts, and no statistical understanding evident.

The Irish Grading System

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.

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, and review the marking criteria outlined in this assignment brief for a breakdown of the marking criteria for this specific assignment.

Additional Information

- Lecturers are not required to review draft assessment submissions. This may be offered at the lecturer's discretion.
- In accordance with CCT policy, feedback to learners may be provided in written, audio or video format and can be provided as individual learner feedback, small group feedback or whole class feedback.
- Results and feedback will only be issued when assessments have been marked and moderated / reviewed by a second examiner.
- Additional feedback may be provided as individual, small group or whole class feedback. Lecturers are not obliged to respond to email requests for additional feedback where this is not the specified process or to respond to further requests for feedback following the additional feedback.
- Following receipt of feedback, where a student believes there has been an error in the marks or feedback received, they should avail of the recheck and review process and should not attempt to get a revised mark / feedback by directly approaching the lecturer. Lecturers are not authorised to amend published marks outside of the recheck and review process or the Board of Examiners process.
- Students are advised that disagreement with an academic judgement is not grounds for review.
- For additional support with academic writing and referencing students are advised to contact the CCT Library Service.
- For additional support with subject matter content students are advised to contact the CCT Student Mentoring Academy
- For additional support with IT subject content, students are advised to access the CCT Support Hub.