A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Title:** | *HDIP Data Analytics* | | |
| **Cohort:** | *HDip Data Analytics - HCI/FT - Feb22 cohort* | | |
| **Module Title(s)**: | *Data Preparation*  *Machine Learning* | | |
| **Assignment Type:** | *Individual* | **Weighting(s)**: | *50% (DP)*  *50% (ML)* |
| **Assignment Title:** | *CA2\_DP\_ML\_HDip\_Lvl8* | | |
| **Lecturer(s)**: | *David McQuaid, Marina Soledad, Muhammad Iqbal* | | |
| **Issue Date:** | *15th April 2022* | | |
| **Submission Deadline Date:** | *15th May 2022* | | |
| **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:** | *Assessment must be submitted before 15th May 2022 as a Jupyter Notebook file and a PDF.*  *The Jupyter Notebook File Must be saved as “YourName\_DP\_ML\_HDip\_CA2.ipynb”* | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | *After the Exam Boards Early July 22* | | |

**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:

**Data Preparation**

**MLO 4** - Select and perform appropriate feature selection and/or dimensionality reduction techniques on a variety of wide datasets.

(Linked to PLO 3 (Stage 4 SLO 3))

**MLO 5** - Develop strategies for identifying and handling missing and out-of-range data, as well as feature engineering as part of the preparation phase of data analysis.

(Linked to PLO 4 (Stage 4 SLO 4))

**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)

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**

Students are advised to review and adhere to the submission requirements documented after the assessment task.

**Note: All data wrangling, analysis, implementation of Machine Learning and visualizations must be generated using python**

Online retailer, eBay is providing an option of bidding to their customers globally. Bidding is employed to find the real price of items in the market based on the demand. The price offered by anyone participating in this process is termed as a 'bid'. A dataset is available at the following link and on Moodle for the bidding of customers

https://archive.ics.uci.edu/ml/datasets/Shill+Bidding+Dataset

Normal bids are classified as ‘0’ bids in the data set and anomalous bids as ‘1’. Your goal is to use classification or clustering algorithms to predict the bids in the future. You would need to clean and prepare the dataset for the machine learning modelling under the following guidelines for Data Preparation and Machine Learning modules.

**Data Preparation**

* Characterisation of the data set: size; number of attributes; has/does not have missing values, number of observations etc.**[0-10]**
* 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]**
* 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]**
* 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 or clustering the normal as compared to anomalous biddings.**[0-40]**

**Machine Learning**

* Provide a logical justification based on the reasoning for the specific choice of machine learning approaches (supervised/ Unsupervised) for the provided Shill bidding data set. **[0-20]**
* Machine Learning models can be used for Prediction, Classification, Clustering and Time series analysis. You should plan on trying multiple models (at least two) with proper parameter-selection using hyperparameters and show a comparison between the chosen modelling approaches. **[0-30]**
* You should train and test the Machine learning models in the case of supervised learning and use other metrics for unsupervised learning if appropriate. Use cross validation to provide authenticity of the modelling outcomes. Perform a comparison of ML modelling outcomes using a Table or graph visualization.**[0-30]**
* Describe the rationale for the selection of models and justify the choice of hyperparameters.**[0-10]**
* A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes**.[0-10]**

**Report**

A report is required to provide the details of work performed in all tasks. The report should be based on Introduction, rationale of data preparation techniques, machines learning models and conclusions. Illustrations should be used to highlight the details of any section.

This should also include evidence to support your data wrangling and analysis through the use of references and citations.

Conclusions, Findings of data set and references/citations in (HARVARD style).

This should be completed in the Jupyter Notebook Markdown.

**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:

* Be submitted before *15th May 2022* as a Jupyter Notebook file.
* 2000(+/- 10%) words in report (not including code, code comments, titles, references or citations)
* The Jupyter Notebook File Must be saved as “YourName\_DP\_ML\_HDip\_CA2.ipynb”
* Be submitted by the deadline date specified or be subject to late submission penalties
* Be submitted via Moodle upload
* Use [Harvard Referencing](http://40.115.124.2/sp/subjects/guide.php?subject=harvardref) when citing third party material
* Be the student’s own work.
* Include the CCT assessment cover page.

**Additional Information**

* Lecturers are not required to review draft assessment submissions.
* 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 requested by contacting the Lecturer,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 or access the [CCT Learning Space](http://learningspace.cct.ie/subjects/index.php).
* For additional support with subject matter content students are advised to contact the [CCT Student Mentoring Academy](https://moodle.cct.ie/mod/forum/view.php?id=55148)
* For additional support with IT subject content, students are advised to access the [CCT Support Hub](https://moodle.cct.ie/course/view.php?id=1861).