cct | College Dublin
Computing • IT • Business

CCT College Dublin Continuous Assessment

Programme Title:	*MSc in Data Analytics*		
Cohort:	*MSc in Data Analytics FT/SB+ (Sep22 start)*		
Module Title(s):	*Programming for DA* *Statistics for Data Analytics* *Machine Learning for Data Analysis* *Data Preparation & Visualisation*		
Assignment Type:	*Individual*	**Weighting(s):**	*Programming for DA **50%*** *Stats for Data Analytics **50%*** *ML for Data Analysis **50%*** *Data Prep & Vis **50%***
Assignment Title:	*MSC_DA_CA1*		
Lecturer(s):	*Sam Weiss* *John O'Sullivan/Marina Iantorno* *Muhammad Iqbal* *David McQuaid*		
Issue Date:	*04/10/2022*		
Submission Deadline Date:	*11/11/22*		
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** **Use the submission link on the** **Data Visualisation and Preparation Module page**		
Instructions for Submission:	*Place all files into a **standard .zip file** and upload.* *Expected files : Written report (pdf / word) ,Code files (jupyter notebook), Data Files*		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	*2 weeks after the last submission including PMC's*		

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:

Programming for DA

1. Debate the selection of programming concepts in the design of programmatic solutions, in terms of paradigm and language selection. (Linked to PLO 1).
2. Design and implement algorithms for use within the context of data analytics. (Linked to PLO 2).

Statistics for Data Analytics

1. Explore and evaluate datasets using descriptive statistical analyses. (PLO 1)
2. Formulate and test hypotheses using appropriate inferential statistical techniques and evaluate and communicate the results effectively to peers and team members. (PLO 2,3,6)

3. Apply statistical analysis to appropriate datasets and critique the limitations of these models (PLO 2,4)

Machine Learning for Data Analysis

2. Develop a machine learning strategy for a given domain and communicate effectively to team members, peers and project stakeholders the insight to be gained from the interpreted results. (Linked to PLO 1, PLO 4, PLO 6)

3. Implement a range of classification and regression techniques and detail / document their suitability for a variety of problem domains. (Linked to PLO 5)

4. Critically evaluate the performance of Machine Learning models, propose strategies to optimise performance. (Linked to PLO 3)

Data Preparation & Visualisation

1. Discuss the concepts, techniques and processes underlying data visualisation to critically evaluate visualisation approaches with respect to their suitability for different problem areas. (linked to PLO 1)
2. Programmatically Implement graphical methods to identify issues within a data set (missing, out of range, dirty data) (linked to PLO 3, PLO 5)
3. Engineer new features selection in data with the goal of improving the performance of machine learning models. (linked to PLO 2, PLO 4)

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 9 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	Attains all the minimum intended programme learning outcomes
40 – 49%	Acceptable	
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.

Scenario: Transport and Infrastructure

A large amount of data has been collected by Dublin City Council (DCC) regarding Transport and Infrastructure in the Greater Dublin Area, This data is available at:
https://data.gov.ie/organization/dublin-city-council?tags=Transport+and+Infrastructure

You are required to choose a particular area of interest and formulate the appropriate questions for modelling and analysis. For Example (but not limited to):

- Clamping Appeals
- Multistorey Car Parking Space Availability
- Telecoms Underground Infrastructure DCC
- etc…

You are required to collect, process, analyse and interpret the data in order to identify possible issues/ problems at present and make predictions/ classifications in regards to the future. This analysis will rely on the available data from DCC and any additional data you deem necessary (with supporting evidence) to support your hypothesis for this scenario.
This will require you to employ critical analysis of not only the domain of choice but also for the regression and or classification that you undertake.

Note: This is an academic exercise and not a hypothetical report to DCC

Criteria of Analysis

Statistics: (Graded out of 100)

You need to analyse the data using statistical logic and statistical techniques. Note: ALL Statistical work MUST be carried out using Python.

You are required to:

1. Summarise your data using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You are required to plot at least two graphs. **[0-50]**
2. Use at least one discrete distribution (Binomial/Poisson) to explain/identify some information about your data. Explain any decisions carefully. **[0-25]**
3. Use at least one Normal Distribution to explain/identify some information about your data. You must justify the use of the measures you calculated and the techniques you used. You must work with Python, but your mathematical reasoning must be documented in your report.**[0-25]**

Data preparation and Visualization : (Graded out of 100)

1. You must perform appropriate EDA on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained. **[0-30]**
2. You must also rationalise and detail all the methods used to prepare the data for ML. **[0-20]**
3. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. **[0-30]**
4. All design and implementation of your visualizations must be justified and detailed in full. **[0-20]**

Machine learning for Data Analytics:(Graded out of 100)

1. Explain the reasoning for selecting one of the following machine learning approaches for the chosen dataset (supervised/ unsupervised/ semi-supervised). Discuss and explain the rationale for choosing the appropriate project management framework/ activities (CRISP-DM, KDD or SEMMA). **[0 - 20]**
2. Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate parameters based on hyperparameters, and then analyze the chosen approaches. **[0 - 30]**
3. Perform the training and testing of the machine learning models, with cross validation/ GridsearchCV, to demonstrate the authenticity of the modelling outcomes. Display a comparison of the results of two or more ML modeling using a table or graph representation. Examine the performance of the machine learning models based of the chosen metric for supervised/ unsupervised/ semi-supervised approaches and analyze it critically. **[0 - 20]**
4. Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations. Provide a report along with an explanation and interpretation to convince DCC of the relevance and effectiveness of your findings. **[0 - 30]**

Programming: : (Graded out of 100)

The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices. (code quality standards should also be applied) **[0-100]**

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:
  - 5000 (+/- 10%) words in report (not including code, code comments, titles, references or citations)
  - Report submission MUST be a word document; Code in a Jupyter Notebook file only but may be referenced in the word document.
  - Be submitted by the deadline date specified or be subject to late submission penalties
  - Be submitted via Moodle upload
  - Use Harvard Referencing 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. 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 requested by *attending the next class,* 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.
- 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.