College Dublin
Computing • IT • Business

CCT College Dublin Continuous Assessment

Programme Title:	*MSc in Data Analytics*		
Cohort:	*MSc in Data Analytics SB+/FT (Feb 24 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_CA2*		
Lecturer(s):	*Marina Iantorno/Taufique Ahmed* *Sam Weiss* *Muhammad Iqbal* *David McQuaid*		
Issue Date:	18/04/2024		
Submission Deadline Date:	26/05/2024		
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** **Use the submission link on the Data Preparation & Visualisation Module page**		
Instructions for Submission:	*Please do not ZIP your files. ALL files must be uploaded individually (to a maximum of 20 files)* *Expected files : Written report (word document only, NO PDF's) ,Code files (Jupyter notebook (.ipynb) ONLY, NO PYTHON FILES), Data Files, Dashboard files. Note that the maximum number of Jupyter Notebooks is 4*		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	*After exam board June 2024*		

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).
3. Compare, contrast and select relevant libraries / techniques to process data from diverse sources. (Linked to PLO 5).

Statistics for Data Analytics

1. Formulate and test hypotheses using appropriate statistical techniques and evaluate and communicate the result effectively. (Linked to PLO 2, PLO 3, PLO6).
2. Utilise current software and language to produce the results of your analysis from existing data. (Linked to PLO 1, PLO 4).
3. Apply statistical analysis to appropriate datasets and critique the limitations of the model. (Linked to PLO 2, PLO4).

Machine Learning for Data Analysis

1. Modify and implement Machine Learning Algorithms to solve analytical problems. (Linked to PLO 1, PLO 2, PLO 5)
3. 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.
5. Formulate and evaluate a test and optimisation strategy for programmatic solutions. (Linked to PLO 5).

Data Preparation & Visualisation

1. Programmatically Implement graphical methods to identify issues within a data set (missing, out of range, dirty data)(linked to PLO 3, PLO 5)
2. Propose, design, develop, and implement an interactive data visualisation solution, for a given data set and potential audience, detailing the rationale for approach and visualisation choices made during development for a given use case, data characteristics and multiple transmission media (linked to PLO 2, PLO 5)
3. Collaboratively perform a critical analysis of a data set to optimise the data for a given problem space. Document the rationale behind the group's decisions to peers and stakeholders.(linked to PLO 5, PLO 6)

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.

Acceptable and Unacceptable Use of AI

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, papers, code fragments, functions, scripts 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.

Assessment Task

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

Scenario

"Today, big data is ubiquitous, machine learning applications are thriving, artificial intelligence appears in everyday conversations, and the internet of things is present even in household appliances. Businesses and organizations are increasingly managed through cloud computing and high-performance computing is progressively accessible as a service…More effective operations, reduced uncertainties, and real time decision-support could revolutionize agriculture to a great extent . Food could be produced more efficiently, of higher nutritional quality, in more stable supplies, with less environmental damage, and likely with additional economic, social, and ecological benefits."(Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022))

You have been tasked with analysing Ireland's Agricultural data and comparing the Irish Agri sector with other countries worldwide. This analysis should also include forecasting, sentiment analysis and evidence-based recommendations for the sector as well as a complete rationale of the entire process used to discover your findings. Your Research could include export, import, trade imbalance, arable production, animal stock, medicinal input, organic, gm products etc. (or any other relevant topic EXCEPT Climate change) with Ireland as your base line.

Note:

- **While topical, Agricultural impact on Climate Change SHOULD NOT be chosen as an area of research for this assessment.**
- **Members of the European Union implement the Common Agricultural Policy, and this should be researched as it has a significant statistical impact.**
- **The United Kingdom is NOT part of the European Union**

You must source appropriate data sets from any available repository to inform your research (all datasets MUST be referenced, and the relevant licence/permissions detailed).

Several Data Sets have been Supplied which you may use as you wish (You do not HAVE to use them)

LICENSES for supplied datasets

"FAO encourages you to use FAO databases for research, statistical, and scientific purposes. You may access, download, create copies and re-disseminate datasets subject to these Dataset terms.

Unless specifically stated otherwise, all datasets disseminated through the databases below are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) "

Criteria of Analysis

Discuss the choice of project management framework you have deemed suitable for this project.
It is Required that you use GitHub Classroom as your version control repository etc with regular commits of code and report versions. You may be called to a Viva to defend your work.

Please find the GitHub Classroom link below:
https://classroom.github.com/a/xAhQmKax

Programming for DA Tasks

YOU MUST ATTEMPT ALL 6 PARTS

1. **Programming**: 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. The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied. [30 marks]
2. **Data From Diverse Sources**: In a dedicated section of your report, compare, contrast, and select relevant libraries/techniques to process data from diverse sources [10 marks]
3. **Data manipulation**: In the same section as Part 2, critically appraise aggregation methods (eg combining data) to process and manipulate data from multiple data structures [0-20]
4. **Data structures**: You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format. [20 marks]
5. **Testing**: In a dedicated section of your report, you are required to document and evaluate a "testing" strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to. Note any trade-offs that you've made in these areas. [10 marks]
6. **Optimisation**: In a dedicated section of your report, you are required to document and evaluate an optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured that the code is making good use of your system's resources (eg CPU, RAM, time etc). Note any trade-offs that you've made in these areas. [10 marks]

Total Mark = 30+10+20+20+10+10=100:(100*0.5=50%)

Statistics for Data Analytics Tasks

- Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. **[0-20]**
- Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with international commerce, you could find a confidence interval for the population proportion of yearly dairy exports out of all agricultural exports). **[0-20]**
- Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests. **[0-40]**
- Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. **[0-20]**

Note: All your calculations and reasoning behind your models must be documented in the report and/or the appendix.

Total Mark = 20+20+40+20=100:(100*0.5=50%)

Machine Learning Tasks

Use of multiple models (at least two) to compare and contrast results and insights gained.

- Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.
  [0 - 30]
- Collect and develop a dataset based on the agriculture topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate agricultural topic (e.g., product price, feed quality etc…) for producers and consumers point of view in Ireland.
  [0 - 25]
- You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.
  [0 - 30]
- A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.
  [0 - 15]

  Total Mark = 30+25+30+15=100:(100*0.5=50%)

Data Preparation & Visualisation Tasks

- Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data (This will require research outside of class material). **[0-15]**
- Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.**[0-25]**
- Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.**[0-30**]
- Modern farming has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern farmers, using tufts principles, to showcase the information/evidence gathered following your Machine Learning

Analysis. Detail the rationale for approach and visualisation choices made during development making reference to Tufts Principles. **Note you may not use Powerbi, RapidMiner, tableau or other such tools to accomplish this (at this stage).[0-30]**

Total Mark = 15+25+30+30=100:(100*0.5=50%)

Additional notes :

All:

- Your documentation should present your approach to the project, including elements of project planning (timelines).
- Ensure that your documentation follows a logical sequence through the planning / research / justification / implementation phases of the project.
- Ensure that your final upload contains a **maximum of 1 Jupyter notebook per module.**
- Please ensure that additional resources are placed and linked to a logical file structure eg, Scripts, Images, Report, Data etc…
- Ensure that you include your raw and structured datasets in your submission
- 3000(+/- 10%) words in report (not including code, code comments, titles, references or citations)
- Your Word count MUST be included.

(it is expected that research be carried out beyond class material)

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:

- Jupyter Notebook, Word Document, Dashboard
- 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 *contacting Your 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.
- 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.

References

Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022): "Big data in agriculture: Between opportunity and solution" , https://doi.org/10.1016/j.agsy.2021.103298,(https://www.sciencedirect.com/science/article/pii/S0308521X21002511)