CCT College Dublin Continuous Assessment

Programme Title:	*MSc in Data Analytics*		
Cohort:	*MSc in Data Analytics FT/SB+ (Sep23 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/ David Gonzalez* *Bharathi Chakravarthi/ Marina Iantorno* *Muhammad Iqbal* *David McQuaid*		
Issue Date:	03/10/2023		
Submission Deadline Date:	10/11/2023		
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:	***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 ONLY, NO PYTHON FILES), Data Files, GITHUB Link* *Note that the maximum number of Jupyter Notebooks is 4*		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	*3 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. Apply statistical analysis to appropriate datasets and critique the limitations of these models (PLO 2,4)
3. Utilise current software tools and languages to produce and document result sets from existing data (e.g., spreadsheets, R, Python). (PLO 1,4)

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: Population in Ireland

A large amount of data has been collected by The Central Statistics Office in Ireland in relation to the population of Ireland, This data is available at:

https://data.cso.ie/product/pme

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):

- Annual Population Change
- Immigration and Migration
- Population Forecasting
- 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 regard to the future. This analysis will rely on the available data from CSO 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 the CSO.

Criteria of Analysis

Statistics: (Graded out of 100)

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

1. Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs. [0-35]
3. Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]
4. Use Normal distribution to explain or identify some information about your dataset. [0-20]
5. Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

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. **[0-20]**
2. You must also rationalise justify and detail all the methods used to prepare the data for ML. **[0-30]**
3. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. **[0-20]**
4. All design and implementation of your visualizations must be justified and detailed in full. **[0-30]**

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

1. Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios. Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modeling. **[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

hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. **[0 - 30]**

3. Show the results of two or more ML modeling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**
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 of the relevance and effectiveness of your findings. **[0 - 20]**

Programming: : (Graded out of 100)

1. 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). **[0-50]**

   Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.

2. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

CA1 NOTE DO NOT ZIP YOUR SUBMISSION FILES, ALL FILES MUST BE SUBMITTED INDIVIDUALLY

Submissions that are suspected of plagiarism and/or inclusion of AI (CHATGBT, BARD etc…) Generated content will be referred to the college authorities.

Note ALL Students are required to use Git for any Assignments that they are working on.

This means that ALL changes must be committed to Git during your assignment. (Not just a single commit at the end!) This is to allow you to display your incremental progress throughout the assessments, give you practice for your capstone/thesis, allows you to create an online portfolio that can be used to showcase your work and to ensure that there are no problems with final uploads (as all your work will be available on GitHub). It is expected that there will be a minimum of 10 commits (with many of you making very many more). You may Only use your CCT email for your git account, private/work email-based accounts will not be accepted. You must also include ALL your lecturer's CCT emails as a collaborator on your account.

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:
  - 4000 (+/- 10%) words in report (not including code, code comments, titles, references or citations)

- Report submission MUST be a word document only (No PDF's!);
- Code in a Jupyter Notebook file only but may be referenced in the word document.
- GITHUB Link
- 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 the appropriate 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
- 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.