CCT College Dublin Continuous Assessment

Programme Title:	Higher Diploma in Data Analytics for Business		
Cohort:	PT		
Module Title(s):	Data Preparation & Visualisation Statistical Techniques for Data Analytics Machine Learning		
Assignment Type:	Individual	**Weighting(s)**:	50%
Assignment Title:	Continuous Assessment		
Lecturer(s):	Muhammad Iqbal David McQuaid Marina Soledad Iantorno		
Issue Date:	22nd Nov 2023		
Submission Deadline Date:	3rd Jan 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** ● **Report in Word Format** ● **Full code base for each module** ● **Upload your code on Github** ● **Provided one zip folder of all documents, images and code files**		
Instructions for Submission:	Upload all files MS word, jupyter notebook, dataset and any supporting information separately. No zip files are allowed.		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	Three weeks after submission date		

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

- 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)

Statistical Techniques for Data Analytics

- Explore and evaluate datasets using descriptive statistical analyses. (Linked to PLO 1)
- Formulate and test hypotheses within a business context using appropriate statistical techniques and both evaluate and communicate the results effectively to peers and team members. (Linked to PLO 3, PLO 6)
- Use and understand current software tools and languages to produce result sets from existing data(e.g. Excel, R, Python). (Linked to PLO 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:

<table>
<tr><th rowspan="2">Percentage Range</th><th rowspan="2">CCT Performance Description</th><th colspan="2">QQI Description of Attainment</th></tr>
<tr><th>Level 6, 7 & 8 awards</th><th>Level 9 awards</th></tr>
<tr><td>90% +</td><td>Exceptional</td><td rowspan="3">Achievement includes that required for a Pass and in most respects is significantly and consistently beyond this</td><td rowspan="3">Achievement includes that required for a Pass and in most respects is significantly and consistently beyond this</td></tr>
<tr><td>80 – 89%</td><td>Outstanding</td></tr>
<tr><td>70 – 79%</td><td>Excellent</td></tr>
<tr><td>60 – 69%</td><td>Very Good</td><td>Achievement includes that required for a Pass and in many respects is significantly beyond this</td><td>Achievement includes that required for a Pass and in many respects is significantly beyond this</td></tr>
<tr><td>50 – 59%</td><td>Good</td><td>Achievement includes that required for a Pass and in some respects is significantly beyond this</td><td rowspan="2">Attains all the minimum intended programme learning outcomes</td></tr>
<tr><td>40 – 49%</td><td>Acceptable</td><td>Attains all the minimum intended programme learning outcomes</td></tr>
<tr><td>35 – 39%</td><td>Fail</td><td>Nearly (but not quite) attains the relevant minimum intended learning outcomes</td><td>Nearly (but not quite) attains the relevant minimum intended learning outcomes</td></tr>
<tr><td>0 – 34%</td><td>Fail</td><td>Does not attain some or all of the minimum intended learning outcomes</td><td>Does not attain some or all of the minimum intended learning outcomes</td></tr>
</table>

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.

Background:

A company has collected data for their employees and is planning to use this data to identify patterns and trends that can help to improve employees satisfaction and productivity. The dataset contains information about their employees, including their age, gender, education level, job role, hourly pay rate, work experience, job satisfaction, and more. The purpose is to use your knowledge in decision making for the optimisation of company operations and welfare of the employees.

You are responsible to analyse the dataset provided, and to identify the role of statistical techniques, need for the data preparation and exploration and machine learning models for prediction/ classification/ clusterings. You will present the influence of data preparation, statistical techniques and machine learning to communicate your findings effectively to the company stakeholders. Also, you will need to use critical thinking skills and problem-solving abilities to identify relevant patterns and trends in the data that can help the company to improve employee satisfaction and productivity.

The dataset is provided on Moodle along with this integrated CA2.

Task:

As a data analyst, your task is to prepare and analyse the data set using appropriate data preparation, statistical techniques and ML models. Your analysis should aim to identify any relationships or trends in the data that can be used to improve employee satisfaction and productivity. Evaluate the influence of data preparation for the use of Statistical Techniques and ML modelling outcomes. Communicate the results in a form of a report in the Jupyter notebook markdown of the analysis to stakeholders using clear and concise explanations, visualisations, and appropriate statistical terminology. All work should be stored on github and perform commits.

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]**

Statistical Techniques:

- Use descriptive statistical analyses to explore and evaluate the data set, including measures of central tendency and dispersion and frequency distributions. Correlation matrices are also accepted. Provide a summary of your findings. (0-30 marks)
- Formulate and test hypotheses within a business context using appropriate statistical techniques like t-tests or ANOVA to identify significant relationships between variables. Provide a summary of your findings. Use at least two statistical tests. (0-40 marks)
- Use a Jupyter notebook to produce result sets from the provided dataset, such as scatter plots or regression models. Provide a summary of your findings. (0-10 marks)
- Write the results of the analysis of your findings to stakeholders using clear and concise explanations, visualisations, and appropriate statistical terminology. (0-20 marks)

Machine Learning:

- 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. (0-20 marks)
- 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)
- 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)
- Exhibit a comparison of ML modelling outcomes using a Table or graph visualisation. Identify the possible similarities and contrast of the Machine Learning modelling outcomes based on chosen metric and discuss their statistical understanding. (0-20 marks)

Submission Requirements

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

- Upload all parts of CA (MS Word/Jupyter Notebook) as separate files on Moodle.
- Must be clearly specified the number of words used in the report.
- Number of Words in the report (Min 3000 words / Max 4000 words) excluding diagrams and references.
- The marking scheme is provided based on the breakdown of marks for each section in this CA.
- Use Harvard Referencing when citing third party material.
- Be the student's own work.
- Include the CCT assessment cover page.
- Be submitted by the deadline date specified or be subject to late submission penalties.
- The Word Document must be saved as "**YourName_StudentID_Lvl8_CA2_Integrated.zip**"

- **Use any version control system (for example, Github)** to show the commits of your progress for integrated CA2. You should have at least 5 commits after the release date of your assignment.
- **Note:** Make sure you will upload correct material on Moodle before the specified deadline.

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.