A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

| **Programme Title:** | BSc in Computing in IT - Sept 2022 cohort [4th year] | | |
| --- | --- | --- | --- |
| **Cohort:** | FT | | |
| **Module Title(s)**: | Data Exploration and Preparation | | |
| **Assignment Type:** | Individual | **Weighting(s)**: | 40 |
| **Assignment Title:** | Exams – Alternative Assignment | | |
| **Lecturer(s)**: | Dr. Muhammad Iqbal | | |
| **Issue Date:** | 5th January 2023 (24 hours) | | |
| **Submission Deadline Date:** | 6th January 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** | | |
| **Instructions for Submission:** | Upload one zip file composed of pdf/ word file, Jupiter notebook, python files, dataset and any supporting information. | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | N/A | | |

**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:

* Choose and implement suitable data-encoding techniques for a variety of machine learning algorithms
* Engineer new features selection in data with the goal of improving the performance of machine learning models

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** | **Level 9 awards** |
| 90% + | Exceptional | Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this | 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 | 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 | Attains all the minimum intended programme learning outcomes |
| 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 | 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 | 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 experience of 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**

Attempt all questions and the marks for each question are mentioned separately.

**Question 1:**

Let us consider the following table as mentioned below

Table 2.1

Table

Description automatically generated

a) Refer to the income attribute of the five customers in Table 2.1, before pre-processing.

1. Find the mean income before pre-processing.
2. What does this number actually mean?
3. Now, calculate the mean income for the three values left after pre-processing. Does this value have a meaning?

b) Explain why zip codes should be considered text variables rather than numeric.

c) Identify which columns are suitable for the categorical datatypes in the Table 2.1 and provide the appropriate encoding scheme to replace the variables with numeric data.

d) Explain why a birthdate variable would be preferred to an age variable in a database based on the presence of data attribute in Table 2.1.

e) Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis. Briefly explain your answer.

(400 – 500 words, 5 \* 6 = 30 marks)

**Question 2:**

Forests are prime sources of the Earth's capacity to establish a balance in the climate, by the global impact of their photosynthesis. They provided a strong and common defence against climate change, taking out the greenhouse gas carbon dioxide and generating oxygen for the planet. This is useful in purifying the atmosphere and controlling jumping up temperatures. However, forest fires create a big challenge to climate change. A data set is provided on Moodle along with source link is mentioned below

* https://archive.ics.uci.edu/ml/datasets/Forest+Fires

Your responsibility is to explore this data by considering the following questions for Data exploration and preparation.

* Discuss and describe the purpose and need for data exploration and preparation within data analytics projects. Use illustrations to justify your answer.

(10 marks)

* Choose a relevant encoding technique to transform the categorical data to numeric format.

(10 marks)

* Formulate the questions for EDA (Exploratory Data Analysis) and perform EDA based on the data provided and address the issues to understand the impact of fire and the resulting burned area.

(20 marks)

* Analyse the benefits and drawbacks for feature selection and extraction techniques. Which features are important from the provided dataset? (Feature Selection Method)

(20 marks)

* Prepare the data for the machine learning model after encoding and feature selection, predict the burned area of forest fires using the machine learning model.

(10 marks)

* You are free to perform any process to enhance an understanding of data preparation and exploration of the Forest data set.

(1000 - 1200 word, 70 marks)

**Submission Requirements**

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

* The code and datasets should be provided and uploaded in zip format on Moodle.
* Must be clearly specified the number of words used in the report.
* Number of Words for the report (Min: 2000 words +- 5%) excluding diagrams and code.
* 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.
* Be submitted by the deadline date specified or be subject to late submission penalties
* Must be clearly specified the number of words used after each section in the report.
* Include the CCT assessment cover page.
* We can check the authenticity of practical work on your Virtual machine (Do not destroy your VM at least 8 weeks after the submission of this CA).