CCT College Dublin Continuous Assessment

Programme Title:	HDIP Data Analytics		
Cohort:	HDipData_Sept22_FT		
Module Title(s):	Data Preparation		
Assignment Type:	Individual	Weighting(s):	50%
Assignment Title:	CA1_DataPrep_HDip		
Lecturer(s):	David McQuaid		
Issue Date:	21/03/2023		
Submission Deadline Date:	21/04/2023 11:55pm		
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:	Assessment must be submitted before 11.55pm 21/04/2023 as a Jupyter Notebook file The Jupyter Notebook File Must be saved as "YourName_DPrepHDip_CA1.ipynb"		
Feedback Method:	Results posted in Moodle gradebook		
Feedback Date:	Approx. 2 weeks after FINAL submission (inc PMC cases)		

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:

1. Programmatically implement graphical methods to identify issues within a data set (missing, out of range, dirty data) (linked to PLO 2, PLO 3)
2. Engineer new features selection in data with the goal of improving the performance of machine learning models.
   (linked to PLO 2, PLO 4)
3. Perform a critical analysis of a data set to optimise the data for a given problem space. Document the rationale behind the decisions to peers and stakeholders. (linked to PLO 1,PLO 3, 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:

<table>
<tr><th rowspan="2">Percentage Range</th><th rowspan="2">CCT Performance Description</th><th>QQI Description of Attainment</th></tr>
<tr><td>Level 6, 7 & 8 awards</td></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></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></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></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></tr>
<tr><td>0 – 34%</td><td>Fail</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 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
Students are advised to review and adhere to the submission requirements documented after the assessment task.

Scenario:

You have been retained by a haulage company to analyse a dataset based on data collected from heavy Scania trucks in everyday usage. The system in focus is the Air Pressure system (APS) which generates pressurised air that are utilized in various functions in a truck, such as braking and gear changes. The dataset's positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS. The data consists of a

subset of all available data, selected by experts. This analysis will help determine the investment strategy for the company in the upcoming year.

All data wrangling, analysis, and visualizations must be generated using python.

The companies CTO also requires you to rationalize all the decisions that you have made in your report.

Requirements

You are required to use the dataset contained within the file "aps_failure_set.csv", conduct the following analysis and report on your findings:

Characterisation of the data set: size; number of attributes; has/does not have missing values, number of observations etc.

Application of Data preparation/evaluation methods (Cleaning, renaming, etc) and EDA 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.

Use PCA to establish the minimum number of features needed for retaining 99.5% variance in the data and then implement PCA to dimensionally reduce the data to the number of features that you have discovered. Include a clear and concise explanation of your rationale for what you are doing with the data and why you are doing it.

Explain **in your own words** what the "Curse of Dimensionality " is.

Conclusions, Findings of data set and references (HARVARD style).

Note that all written work MUST be completed in Jupyter Notebook Markdown (please review "Jupyter Notebook Tutorial" Notes in Moodle if you are unsure of this).

All Code must be included in code blocks (As normal). No other upload will be accepted.

All written work MUST be detailed in your Jupyter Markdown (NOT in code comments).

Data Dictionary

Columns	Value	D type	Description
Class	neg/pos	String/Object	The dataset's positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS
All Other Columns	0 to 8.584298e+09	float	Component Sensor result

Note that 0 values are perfectly valid sensor returns and what the sensors are measuring is unimportant.

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:

- Be submitted before 11.55pm 21/04/2023 as a Jupyter Notebook file.
- The Jupyter Notebook File Must be saved as "YourName_DPrepHDip_CA1.ipynb"
- 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.