A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Title:** | HDIP Data Analytics | | |
| **Delivery Mode:** | FT | | |
| **Cohort Details:** | *HDipData\_Sept23\_FTStage 1 Sem1* | | |
| **Module Title(s)**: | *Data Preparation*  *Programme schedules are all published on the* [*CCT IQR Provider Profile*](https://irq.ie/providers/cct-college-dublin?id=fec9ea7a-ace4-42c7-9fd5-7fccb6f0a53a&ref=%257B%2522search%2522:%2522cct%2522%257D) | | |
| **Assignment Type:** | *Individual* | **Weighting(s):** | *50%* |
| **Assignment Title:** | CA1\_DataPrep\_HDip | | |
| **Lecturer(s)**: | David McQuaid | | |
| **Issue Date:** | *10/03/2025* | | |
| **Submission Deadline Date:** | *06/04/2025 at 11:59pm* | | |
| **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:** | **This assignment is submitted via Moodle.** | | |
| **Instructions for Submission:** | The Jupyter Notebook File Must be saved as “YourName\_DPrepHDip\_CA1.ipynb” | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | Approx. 3 weeks after FINAL submission (inc PMC cases) | | |

# Assessment Outline

## Description of Assessment Task

“Spam” emails are a major issue in everyday life and spam filters have now become a necessity in all aspects of life. The data set attached is real life example of emails that have been sent to a company and have been labelled as True (is spam) and False (not spam). This is a high dimensional dataset that needs to be prepared for classification, in order to create a ML mode for spam filtering.

Minimum Requirements

You are required to use the dataset contained within the file “spambase\_v5.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. and what these Characterisation mean in the context of this data.
* 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 in the context of this data.
* 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 in the context of this data.
* Explain **in your own words** what the “Curse of Dimensionality ” is and how it may affect your analysis going forward in the context of this problem.
* Testing your dimensionally reduced dataset practically and discuss your results in the context of this data.

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

**Data description is available in the spambase\_v5\_documentation.txt file**

**Note ALL Students are required to use Git Classroom for any Assignments that they are working on. This assignments Git Classroom link is** [**https://classroom.github.com/a/Z0ErlJzG**](https://classroom.github.com/a/Z0ErlJzG)

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 (No Python or Excell Files).

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

## Assessment Requirements

All assessment submissions must meet the following minimum requirements:

* Be submitted in the format outlined in the assignment summary table.
* Be submitted by the deadline date specified or be subject to late submission penalties.
* Be submitted via Moodle upload
* 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.
* Include the CCT assessment cover page.

## Learning Outcomes:

This assessment addresses the following module learning outcomes for this module:

* 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)
* 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)

**Statement of Acceptable Use of Artificial Intelligence**

|  |
| --- |
| **Acceptable and Unacceptable Use of AI**  *This statement is useful when you are allowing the use of AI tools for certain purposes, but not for others.* |
| * 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   + Generating code for your assignment   + Writing a draft of a writing assignment   + Writing entire sentences, paragraphs or papers 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. |

## Grading Criteria

This grading rubric sets out the marking criteria for your assignment.

*[insert grading rubric- This set of marking criteria should be developed into a detailed rubric for each module- support resources are available]*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criteria** | ***Characterization of the Data Set*** | ***Application of Data Preparation and EDA*** | ***PCA Implementation*** | ***Explanation of the “Curse of Dimensionality”*** | ***Testing and Discussion of Results*** |
| **Weighting per criteria** | **10 marks** | **30 marks** | **30 marks** | **10 marks** | **20 marks** |
| *Excellent (+70%)* | Comprehensive and insightful analysis; thoroughly contextualized characterization with significant implications. | Excellent preparation with multiple insightful visualizations; rationale is thorough and contextualized effectively. | Comprehensive PCA analysis; thorough rationale for variance retention; insightful application of dimensionality reduction. | Excellent, insightful explanation; thoroughly connects the concept to the data analysis context and implications. | Comprehensive testing with thorough discussion; insightful analysis of results in the context of the data. |
| *Very Good (60 - 69%)* | Thorough characterization with good context; clear explanation of implications. | Good preparation and a variety of clear visualizations; rationale is mostly well explained. | PCA performed effectively; clear explanation of feature selection and rationale provided. | Good explanation with clear relevance to the analysis; shows understanding of implications | Good testing with relevant discussion; insights drawn from results. |
| *Good (50 - 59%)* | Adequately describes size, attributes, missing values, and observations, but lacks contextual analysis. | Basic preparation and some EDA visualizations; rationale is present but lacks depth. | PCA performed; explanation of variance retention is present but lacks detail in rationale. | Satisfactory explanation of the concept; some relevance to the analysis context. | Adequate testing with some discussion of results; relevance is somewhat clear. |
| *Acceptable (40 - 49%)* | Some relevant details provided but lacks clarity or depth. | Some data preparation done, but visualizations are unclear or poorly explained. | PCA attempted but lacks clarity in rationale or explanation. | Basic understanding, but explanation is unclear or incomplete. | Some testing done, but results discussion is weak or unclear. |
| *Fail (> 39%)* | Incomplete characterization; lacks key details. | Minimal data preparation; no visualizations or rationale. | No PCA conducted or poorly explained; no rationale. | Incorrect or vague explanation; lacks understanding of the concept. | Minimal testing; discussion lacks clarity or relevance. |

**The Irish Grading System**

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.

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, and review the marking criteria outlined in this assignment brief for a breakdown of the marking criteria for this specific assignment.

**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 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](https://moodle.cct.ie/course/view.php?id=827)
* For additional support with IT subject content, students are advised to access the [CCT Support Hub](https://moodle.cct.ie/course/view.php?id=1861).