Scoil na Faisnéisíochta & na nEalaíon Cruthaitheach
School of Informatics & Creative Arts

Department of Visual and Human Centred Computing

Assignment Brief

Name of module	Data Science		
Module code	COMP I8026		
Lecturer(s)	Kevin McDaid		
No. of credits for module	10		
Breakdown of module	**CA - 100%**	**Exam**	**Total 100%**
Percentage of overall module CA for this assignment	35%		
Name of assignment	Data Science Project		
Indicate if the assignment is recoverable or non-recoverable	Recoverable		
Due date	*23rd November 2025*		
Assignment description	See description below.		
Extent of AI use narrative[1] **(This will indicate if AI use is unrestricted, restricted, prohibited by the lecturer(s))**	**AI use is only allowed to be used for the Data Mining stage of the assessment. If it is used at this stage then this should be explained in a markdown block at the start of the Data Mining Jupyter Notebook.** **AI cannot be used for any other elements of the assessment.**		
Format of submission (if relevant) **Deliverables** **(What must be included in submission)**	See description below		
Mode of submission	Moodle submission		

[1] Consult the Generative Artificial Intelligence Guidance for Staff for narrative to be included here.

(If submission is in hardcopy a student must also provide evidence via Moodle (photo/video) of the submission by the due date for submission)	
Special requirements as per ACS (if relevant). **Indicate if reattendance of the module is required if the overall module is failed.**	
Stages of delivery of proposals / draft versions / subcomponents with staggered delivery (where applicable) **Percentages of marks for items delivered in this manner (where applicable)**	See description below
Marking criteria (details on rubric / criteria)	See description below
Method of feedback to be provided to students	Marks for each section through Moodle.

Notes to Students

This information on this assignment brief may represent only one component of the overall module. Please ensure that you contact the other lecturers teaching on this module for the information on assessments relating to the other components of the module. It is your responsibility to ensure that you complete all assessments for every module.

- Please note that all work submitted as partial fulfilment of this module must adhere to the Institute Academic Integrity Policy. The Generative Artificial Intelligence (AI) and Your Assessments A Guide for Students provides useful support information for students.

- If you have been permitted to use AI in the preparation of your assignment you should include a Declaration of Use acknowledging that you have used generative artificial intelligence in the creation of material for the assessment. Material that has not been adapted/modified should be referenced using existing reference styles. The Declaration of Use[2] should be included in an appendix in the assessment submitted by the learner and should:
    - Provide a written acknowledgment of the use of generative artificial intelligence.
    - Specify which technology/technologies were used.
    - Identify the prompts used.
    - Provide the resulting outputs.
    - Explain how the output was used in the submitted work (used directly or modified).
- If generative artificial intelligence (AI) is permitted in the assessment and the learner has chosen not to use it, the following disclosure in the learner submission is recommended: "*No content generated by AI technologies has been used in this assessment*".
- All work submitted must be accompanied by a Continuous Assessment Cover Sheet.
- We recommend all students consult the DkIT Guide to Harvard Referencing which provides useful information on referencing and how to avoid plagiarism.
- All work must be submitted by the due date indicated by your lecturer(s). Any work submitted after the due date is subject to the penalties outlined in the Institute Continuous Assessment Procedures
- The terms recoverable and non-recoverable are detailed in the Institute Continuous Assessment Procedures.

[2] Consult The Generative Artificial Intelligence (AI) and Your Assessments A Guide for Students for example of wording to be used for Declaration of Use statement.

Data Science – Project

CA2 (35%) – DATA SCIENCE PROJECT 2025– VERSION 2

Summary

For this assessment you are required to source a real data set clearly related to an industry problem. You should then solve this problem by working through each of the steps of the Data Analytics Lifecycle we have studied in class using Python to analyse the data.

This assignment should be completed in pairs and details of the pairings should be communicated to the lecturer on or before October 23rd. You may be allowed to complete the project on your own - please talk to the lecturer asap if this is something you wish to do.

The problem you should attempt to solve should be a classification or a regression problem.

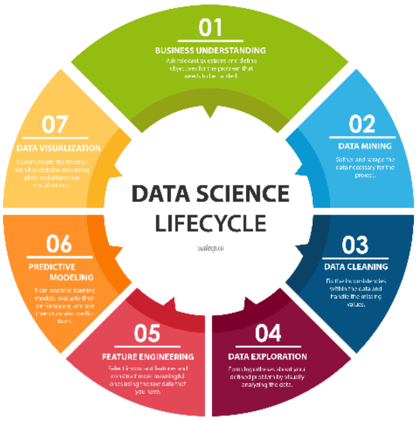
At the end of the work you should develop a jupyter notebook detailing your work and findings.

This continuous assessment will contribute **35%** towards the final mark for this module. The project is due by **November 23rd 2025**. You will be required to demonstrate progress towards completion of the project on an ongoing basis. This element will be worth 5%. You will be informed in advance when and how this progress will be assessed.

There may be an interview on your work in the weeks following submission.

Description

In this project you are required to source a real data set clearly related to a specific problem or problems and to apply each of the steps in the Data Analytics Lifecycle we have studied to solve the problem(s). In solving the problem(s) you should make use of machine learning regression or classification methods. The data set would ideally include at least 6 columns which can be used for prediction with a mixture of numerical and categorical.


DATA SCIENCE LIFECYCLE
sudeep.co
01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Based on the steps in the Data Analytics Lifecycle you should complete each of the following:

Step 1 - Business Understanding: Clearly identify and specify what problem(s) the data can be used to solve and explain why this might be important to a particular company and to a broader industry.

In specifying the problem, you should have a single response variable that you are trying to predict and this should represent the final and key problem you intend to solve. It is important that you specify the type of response variable as this will dictate the type of analysis you can perform. In addition, you will also have associated problems you will need to solve such as the profile of the predictor variables and the relationship between the predictor variables and between the predictor variables and the response variable.

Note that the problem set may be expressed as a set of questions or as a set of objectives.

Step 2 - Data Mining: Using a suitably titled jupyter notebook collect the primary data for the project and locate and collect any additional data that might help you solve the problem. Export and store this data as a csv file (see details below). Details of the steps and difficulties associated with this stage of the analysis should be included as comments in the analysis notebook you submit.

In sourcing the data, you are encouraged to use the web parsing and API methods discussed in class. You might also consider constructing your own data set to solve a novel problem which could have future business potential! Note that if you simply choose a data set from Kaggle which does not require any mining element then you will not receive any marks for this element.

Step 3 - Data Cleaning: Importantly, you must characterise all the variables within your data set as type[Numerical/Categorical] and purpose[response/explanatory/not relevant]. Following this you should clean the data using the methods we explored through the course. This activity will naturally require you to deal with missing values and outliers. You can also at this stage group variables together if it is clear that you need to do so to apply the predictive models later in the section.

Again, all activity, issues and decisions should be clearly documented in your code using markdown blocks.

Step 4 - Data Exploration: You are required to perform Exploratory Data Analysis following the methods we have developed through the course.

Your analysis workbook should contain sections for Univariate, Bivariate and Multivariate Analysis.

Through this activity you may well answer some of the questions driving this project including questions relating to profiling the data and understanding the relationship between variables. You should include detailed comments in your notebook on your interpretation at each stage of this analysis.

Step 5 - Feature Engineering: You are required to specify the variable you are wishing to predict and to identify the important features and the key variables for completion of further predictive analysis.

Here, you are also required to add additional variables if required and/or to construct more meaningful ones. Importantly, this is where you might construct numerical versions of categorical variables and also may identify variables which are multi-colinear and may therefore be ignored.

It is very important that you produce a correlation matrix with the variables that you intend to use in the modelling phase.

Note that in this step you may well end up restating some of the findings from the exploration step.

Step 6 – Predictive Modelling: You must use appropriate regression or classification methods to deliver a prediction model for the response variable you are seeking to model. You may use a number of models and explain clearly the structure and output of each model. It may be helpful at this point for each member of the pair to apply a different set of models to the problem.

Most importantly, you should examine the quality of the predictions and make conclusions as to the best model for practical use and its limitations.

Step 7 – Findings: Throughout the notebook you should use appropriate visualisations. This final section should summarise the work including the key findings and the key visualisations. In this section you should make conclusions as to the best model for practical use and its limitations. You visualisations and conclusions should relate very clearly to the problems you outlined in the Business Understanding step of this project.

Deliverables

You are required to submit the following **four** files only through Moodle:

1. <u>Data Mining Jupyter Notebook</u>

A single notebook named ***DataMining.ipynb*** that implements any web scraping or other code used to source the data for this project. This code should output the data in the form of a csv file named *dataProject.csv* when the code is run. The file should include appropriate comments to explain the workings. If there are problems with the code then you should include a final section which explains the issues. Note that this file should not exceed 10MB in size and you can remove any data that is not of value.

2. Data File

A single csv data file named ***dataProject.csv*** which contains the data used for analysis in the project. This will be the same as the outputDataMining.csv file is your data came from a mining activity. There is a limit of 10Mb on the size of this file.

3. Analysis Jupyter Notebook

A single Jupyter notebook named ***DataAnalysis.ipynb*** which should include only the seven numbered sections as set out in the steps detailed earlier in the this document as well as a final section which outlines the work completed by each member of the team. You can add subsections. Note the following

- Each of the steps should be clearly labelled in a markdown block using # formatting with subsections labelled using ### formatting.
- Any descriptions or interpretations of output comments should be contained in markdown blocks, not as comments in the code. The only comments in the code should relate to the code.
- All graphs should be included as outputs from the code sections.

For Step 2:Data Mining, the section in the analysis file will contain a short summary of the source of the data and should refer back to the Data Mining notebook if your data was mined.

4. Html Output Document

A single html file generated from the jupyter notebook which shows the output from the analysis file. The file should include all text and graphics relevant to your answer including clear interpretations and findings.

Indicative Marking Scheme (Out of 100%)

Note that if there is a clear difference in the output of each member of the team then marks will awarded on the basis of the work completed by each member.

1. Business Understanding: Identification and explanation of problem and related companies and industry: 5%
2. Data Mining: 10%
   Note that if you simply choose a data set from Kaggle which does not require any mining element then you will receive 0% for this section.
3. Data Cleaning: 7.5%
4. Exploration of data: 20%
5. Feature Engineering: 5%
6. Predictive Modelling and Assessment of Model Quality: 35%
7. Conclusions and findings: 7.5%
8. Overall document quality: 5%

9. Ongoing Progress: 5%

Interview

Note that you must attend for interview if requested to do so. Failure to attend for interview will result in a grade of 0.

Rubric

Grade (%)	Description
80+	Project demonstrates mastery of subject matter with novel/original work applied to a complex problem. Data is sourced correctly and has been appropriately structured and analysed. Analysis of data and model development is accurate, complete, thoroughly explained, and appropriate with clear link to aims. Analysis shows original thinking and implementation beyond what was learned in course. All documentation is well-structured and very well-written. Student has presented with full knowledge of both the problem and the solution including appropriate critical analysis.
70-79	Project demonstrates thorough understanding of subject matter. Data is sourced correctly and has been appropriately structured. Analysis of data and model development is accurate, complete, thoroughly explained, and appropriate with clear link to aims. Functionality is well developed with only minor issues. Document is well-structured, with complete version history. Student has presented with excellent knowledge of both the problem and the solution including appropriate critical analysis.
60-69	Project demonstrates good understanding of subject matter. Data is sourced correctly and has been appropriately structured and analysed. Analysis of data and model development is accurate, substantially complete, well explained, and appropriate with good link to aims. May be missing some appropriate analysis. Document is reasonably structured, substantially complete and well delivered. Student has presented with reasonably good knowledge of problem and solution but lacks some critical analysis and in depth understanding.
50-59	Project demonstrates reasonable understanding of subject matter.

	Data analysis and model development is accurate, partially complete, reasonably well explained, and appropriate with some link to aims. May be missing some appropriate analysis and links to aims and objectives may not be complete. Document is adequately structured, mostly complete and reasonably well delivered. Student has presented problem and solution satisfactorily, but with some issues relating to problem and solution. Lacks critical understanding.
40-49	Project demonstrates partial understanding of subject matter. Analysis is relatively simple allowing very basic exploration and model development. Basic but incomplete analysis based entirely on functionality provided in class with significant omissions. Poor link to aims and objectives and weak understanding of logic and findings of the model. Document missing some key elements. Student has presented poorly. While presentation has some relevant structure, it lacks completeness and coherence.
30-39	Project demonstrates little understanding of subject matter. Little evidence of exploration of problem and model development. Incomplete analysis based entirely on functionality provided in class with weak links to aims. Documentation poor and missing significant elements. Presentation is very confusing and lacks clarity and focus. No critical analysis.
0-29	Project demonstrates almost no understanding of subject matter. No serious attempt to address the problem.

Additional Notes

You are encouraged to explore applications and data sets which may be significantly different to the ones we investigated in the course. In this case it may be necessary to adjust the marking scheme to reflect the project application chosen.

Plagiarism

PLEASE PAY SPECIAL ATTENTION TO THE ISSUE OF PLAGIARISM. The DkIT policies are available at https://www.dkit.ie/system/files/academic_integrity_policy_and_procedures.pdf

In summary, all work submitted by learners for assessment purposes, or for written or oral publication, must be their own work. Where this is informed by the work of others, the source must be properly referenced using the accepted norms and formats of the appropriate academic discipline.

Late Submission

The policy for late submission will apply and is available at the link below. Any legitimate late submission must be accompanied by explanation and supporting documentation as per the policy.

https://www.dkit.ie/system/files/continuous_assessment_procedures_document_v4.pdf