SET09120 – Data Analytics

Coursework 1

Learning Outcomes Covered:	LO1, LO2, LO3 and LO5
Assessment Type:	Practical Assessment / Demonstration
Overall module assessment	40% coursework 1, 60% coursework 2
For this assessment:	40%
Assessment Limits:	1. Up to 6 pages report 2. Cleaned and formatted datasets
Submission Deadline:	Friday, 31 October 2025 at 17:00 BST
Submission Method:	Via Moodle
Turnitin on submissions:	Turnitin report not visible to students
Module Leader:	Dr Taoxin Peng
Tutor with direct responsibility:	Dr Taoxin Peng and Dr Kehinde O. Babaagba
Return of work and feedback:	Feedback on submissions will normally be provided within three working weeks from the submission date.
Notes:	• You are advised to keep a copy of your submitted assessment. • Please read and follow the 'Fit-to-Sit' guidance if you need to request an extension

Assessment regulations and academic integrity

The University rules on Academic Integrity apply to all submissions. The student academic integrity regulations contain a detailed definition of academic integrity breaches.

- You cannot knowingly permit another student to copy all or part of your work.
- You must not share your work with other students. This includes posting any of your work in any repository that is accessible to others (such as GitHub) and applies also after you have completed the course.
- Asking coursework-related questions in external online forums (such as Stackoverflow) is NOT permitted.

By submitting the report, you are confirming that:

- It is your own work except where explicit reference is made to the contribution of others.
- It has not been submitted for any module, programme or degree at Edinburgh Napier University or any other institution.
- If you have made use of generative Artificial Intelligence (AI) tools, you have done so only as allowed for this assessment, and have provided the relevant details in the coursework declaration.

<Assessment specification starts here>

(a) Academic skills support: In advance of submission, you can access the support of the academic skills team. They can help you with any aspect of the assessment that you might struggle with, that is not content related. For example, they can help with time-management, effective reading and note-making, and any aspect of academic writing that you might struggle with. This support is provided through workshops and individual appointments which are bookable online via MyNapier: Improve your Academic & Study Skills (napier.ac.uk). You can also directly email the Academic Skills Adviser, Hannah Awcock, h.awcock@napier.ac.uk for any specific academic skills support you require.

(b) **Use of generative AI:**

Please include the Assessment Declaration Cover Sheet in your submission. This is provided below.

Submissions must be accompanied by a declaration cover sheet to fulfil the requirements of the university Assessment Policy.

ASSESSMENT DECLARATION COVER SHEET

Please complete this cover sheet for each assessment submission. For group assessments, each group member must individually complete and submit a cover sheet.

STUDENT REGISTRATION NUMBER: Click or tap here to enter text.

MODULE TITLE: Click or tap here to enter text.

DATE OF SUBMISSION: Click or tap to enter a date.

I declare that, except where explicitly acknowledged*, this assessment is my own work and has not been submitted for any other module or degree programme at Edinburgh Napier University or any other institution. This declaration is made in compliance with Edinburgh Napier University's Academic Integrity Regulations.

***IMPORTANT**: Contributions from other sources include incorrectly cited quotations or content generated by Generative Artificial Intelligence (Gen AI) tools, such as ChatGPT. See Artificial Intelligence Tools and Your Learning for more information.

Note: It is also important to check the associated Assessment Brief to understand what is permissible. Unless stated in the Assessment Brief, use of Grammarly for checking spelling and grammar is allowable in this assessment, but it cannot be used to generate content.

Please declare here your use of such tools in this assessment submission:
(Select one of the two options below)

☐ Yes, I have used Gen AI tools for this submission, and I have described how below.

☐ No, I have not used Gen AI tools for this submission.

If you selected **Yes**, briefly describe (in less than 100 words) how you used these tools:

Use	Permitt-ed?	Advice	How to acknowledge use
As a search engine	Yes	Cross reference AI output for factual accuracy in authoritative texts e.g. text books, reading lists, peer-reviewed publications	Acknowledgment not required
As an ideas generator/conversational partner/debating partner	Yes	Cross reference for accuracy as above AND check for bias, irrelevant or too generalised ideas.	On cover sheet: "I used [tool name] on [date] with the question [insert question/prompt used] to give me ideas, of which I used/adapted into [idea name] in this submission"
To suggest a submission structure	With caution	Consult the assessment brief first to ensure your structure follows the recommendations and meets the learning outcomes.	On cover sheet: "I used [tool name] on [date] with the question [insert question/prompt used] to get a submission structure, which I used/adapted into [part name] in this submission"
To make suggestions to improve your communication of your ideas	With caution	Always start with your own writing first to develop your own thinking. Use the AI tool to get quick feedback and use your judgement whether its advice is appropriate for your submission. Work on one paragraph at a time.	Acknowledgment not required *Or* On cover sheet: "I used [tool name] on [date] with the question [insert question/prompt used] on [section name(s)/whole submission] to get feedback on my writing, which I then improved based on its advice on [spelling/grammar/ vocabulary/etc.]
To generate content (example 1: not allowed)	No	Never ask an AI tool to generate parts of your submission from scratch. Do not input assessment brief or rubric into AI tools and ask it to generate your submission.	
To generate content (example 2: allowed)	With caution	Always put copied-and-pasted AI content in quotations marks (in the case of text) or label other media appropriately.	On cover sheet: "AI generated content is indicated in this submission [within quotation marks/labelled] which gives the prompt used, the tool name and date used"

Data processing and transforming

For this assignment, you must use the dataset provided on Moodle. The use of other datasets will result in failing the module.

DATA DESCRIPTION

The file credits.xlsx contains historical observations on 10 variables (attributes) for 1,000 past credit applications. Each applicant was given a rate of "good" (700 cases) or "bad" (300 cases) credit. Based on the applicant's profile, a bank can make reasonable decisions about whether or not to award a loan. The table below shows the metadata about the dataset.

Table: Attributes and Values for the Credit Data

Attribute Name	Value	Code description
Case_no		Case number allocated to each applicant
	numerical	
checking_status		Status of existing current account
	< 0	Less than 0
	0<=X<200	Between 0 (inclusive) and 200
	>=200	Greater or equal to 200
	no checking	No current account in the bank
credit_history		Debt history
	no credits/ all paid	No debt taken or all debts paid back duly
	all paid	All debts at this bank paid back duly
	existing paid	Existing debts paid back duly till now
	delayed previously	Delay in paying off in the past
	critical/other existing credit	Critical account/other debts existing (not at this bank)
purpose		The purpose of a loan
	new car	New car
	used car	Used car
	furniture/equipment	furniture/equipment
	radio/tv	Radio/television
	domestic appliance	Domestic appliance
	repairs	repairs
	education	education
	vacation	holiday
	retraining	retraining

	business	business
	other	Other purposes
credit_amount		Debt amount
	numerical	
saving_status		Savings account/bonds
	<100	Less than 100
	100<=X<500	Between 100 (inclusive) and 500
	500<=X<1000	Between 500 (inclusive) and 1000
	>=1000	Greater or equal to 1000
	no known savings	unknown/no savings account
personal_status		Personal status and gender
	male div/sep	Male: divorced/separated
	female div/dep/mar	Female: divorced/separated/married
	male single	Male: single
	male mar/wid	Male: married/widowed
	female single	Female: single
age		Age in years
	numerical	
job		Job status
	unemp/unskilled non res	Unemployed/ unskilled – non-resident
	unskilled resident	Unskilled - resident
	skilled	Skilled employee / official
	high qualif/self emp/mgmt	Management/self-employee/officer
class		Decision – good or bad
	good	Safe to provide a loan
	bad	Not safe to provide a loan

Like much of the data that companies store in data warehouses, this is genuine historical data recorded by a German bank, and much of the interest lies in discovering patterns within it. For example, a bank loan customers need the monetary resources to meet their goals. In exchange for loans, the bank charges interest to customers. Repayment of the loan and interest is vital to the lending bank because the loaned money is the “raw materials” of their business, and the interest is the source of profit. How to increase profits is a big question for the bank. The bank managers have only a vague idea about their customers, who are good (safe, offer a loan) and who are not (risky, don’t offer any loan or offer a loan with caution, e.g., charge higher interest). Fortunately, the bank stores data about their customers, the status of existing accounts, savings status, credit history, job status, etc. Bank managers hope to improve their understanding of customers and seek specific actions to increase their profit. Analysing the data with a discovery tool will be convincing for managers.

YOUR TASK

You are asked to use OpenRefine and Weka to prepare the data for analysis and to produce a SHORT report describing the task.

Task and mark allocations are as follows (Total [40%]):

1. Understand and clean the data for analysis. At this stage, you are expected to undertake at least the following procedures: Carefully read through the metadata presented above and understand the meaning of each attribute and its values. You are expected to identify and correct all possible errors in the dataset. [12%]

2. Convert the data for analysis using Weka and Python. For this task, you are expected to convert the cleansed dataset (generated in task 1) from the XLSX format into the format that can be accepted by Weka (ARFF format) and Python. This might include transforming data from one type to another to use particular algorithms. you are expected to submit at least three datasets: one without any transformation, one with all nominal values, and one with all numeric values, all in ARFF format. [10%]

3. Create two visual representations illustrating the distribution of credit class (e.g. good and bad) for different ***personal status*** and ***saving status*** separately. To do this, start by restructuring the dataset into a new data frame for each condition (personal status and saving status) and include the total number of good and bad classes for each condition. [10%]

4. Please analyse the data using statistical techniques in Python to identify which variables (***personal status*** and saving status)) significantly influence the credit class and which do not. [8%]

You will submit:

1. An up to 6 pages **.pdf** document explaining your work:
    a. Use at least a 12-point font.
    b. should include a clear description of how you do the data cleaning and perform any data transformation.
    c. Errors identified and corrected accordingly should be included in a table format (suggested).
    d. Any necessary screenshots and tables can be put into an appendix not included in the 6-page limit.
2. Cleaned and formatted datasets
    a. All datasets that are ready for the analysis (ARFF versions) by Weka and Python should be zipped into a file called SET09120cw1_ and uploaded to Moodle by the submission

deadline, as per instructions. For example, if your matriculation number is 40123456, your zipped file should be SET09120cw1_40123456.

3. Python file **.ipynb**
    a. Please split each code step based on coursework tasks.
    b. Add a brief description to each code cell.
    c. Please provide readable comments on your code.

** Please submit two files: one PDF document and one ZIP file. The ZIP file should include the cleaned and formatted datasets, as well as the Python notebook (.ipynb).

Marking: 12% for Data Cleaning, 10% for Data Convention, 10% for Data Visualisation, and 8% for Statistical Analysis

Feedback: Apart from the marks, you will receive text feedback from Moodle 3 weeks after the submission deadline. The feedback will further explain what you have done well and what needs to improve, corresponding to your marks.

Deadline: 17:00, 31 October 2025

Late submission policy

- Coursework submitted after the agreed deadline will be marked at a maximum of 40%.
- Coursework submitted over five working days after the agreed deadline will be given 0%.

Extensions

Please contact the module leader before the deadline if you require an extension. Extensions are only provided for exceptional circumstances, and evidence may be required. See the Fit to Sit regulations for more details.

Plagiarism

Plagiarised work will be dealt with according to the university's guidelines (Please read - especially if this is the first time in a UK university): http://www2.napier.ac.uk/ed/plagiarism/

Report Template:

1. Introduction: Briefly introduce the aims of this coursework.
2. Data Preparation

2.1. Data Cleaning Describe the way OpenRefine cleans the data. Errors should be identified. Also, reasons for correcting the identified errors should be provided.
2.2. Data Transformation and Conversion Describe how the cleaned data is transformed/converted to data sets, which algorithms can analyse.
2.3. Data Framework and Visualisation plot a figure of the data requested above with a clear description/explanation.
2.4. Scientific Analysis, with code snippets, screenshots of the results, and thoughtful explanation.

3. Appendix (Optional)

Marking Scheme

Appendix A: Marking Scheme

	No Submissio n	Very poor	Inadequate	Adequate	Good	Very good
Task 1 Data Cleaning 12%	No work submitted	Cleaned Dataset with errors and Methods and Errors not well described	Cleaned Dataset with errors and Methods well described but Errors are not correctly identified	Cleaned Dataset with a minor error, but Methods well described and Errors are correctly identified	Cleaned Dataset without errors Errors are correctly identified, but Methods are not well described	Cleaned Dataset without errors and Methods well described and Errors are correctly identified
Task 2 Data Convention 10%	No work submitted	Missing converted datasets and methods/reasons for categorising not well- described	Either missing converted datasets or methods/reasons for categorising not well- described	All converted data is well presented, but methods/reasons for categorising are not well described	All converted data is well presented, and the method for categorising was described, but reasons were not justified	All converted data is well presented, the method for categorising is described, and the reasons justified
Task 3 Visualisation 10%	No work submitted	The solution was submitted but without any explanation	The solution was submitted with a description of the procedure but no explanation	The solution was submitted with a brief explanation	The solution has been clearly explained, but the code is not readable.	The solution has been clearly explained, and the code is clean and readable.

Task 4 Statistical Analysis **8%**	No work submitted	The solution was submitted but without any explanation	The solution was submitted with a description of the procedure but no explanation	The solution was submitted with a brief explanation	The solution has been clearly explained, but the code is not readable.	The solution has been clearly explained, and the code is clean and readable.