NOTTINGHAM
TRENT UNIVERSITY

School of Science and Technology

COURSEWORK ASSESSMENT ELEMENT

Module Code	COMP40721
Module Title	Foundation of Artificial Intelligence
Module Leader	Dr Salisu Yahaya
Module Team	Dr Salisu Yahaya
Coursework Title	Implementation and Written Report
Module Learning outcomes assessed.	[ML02 - ML06]
Contribution to element	70%
Date set	9 September 2024
Deadline for submissions	**Monday 13 January 2025 2:30pm**
Method of Submission	NOW Dropbox
Deadline for Feedback	3 February 2025
Method of Feedback	Via NOW Dropbox

Work handed in up to five working days late will be given a maximum Grade of Low Third whilst work that arrives more than five working days will be given a mark of zero.
Work will only be accepted beyond the five working day deadline if satisfactory evidence, for example, an NEC is provided. https://www.ntu.ac.uk/studenthub/my-course/student-handbook/submit-a-notification-of-extenuating-circumstances

The University views **plagiarism and collusion** as serious academic irregularities and there are a number of different penalties which may be applied to such offences. The **Student Handbook** has a section on Academic Irregularities, which outlines the penalties and states that **plagiarism** includes:

'The incorporation of material (**including text, graph, diagrams, videos etc.**) derived from the work (published or unpublished) of another, by unacknowledged quotation, paraphrased imitation or other device in any work submitted for progression towards or for the completion of an award, which in any way suggests that it is the student's own original work. Such work may include printed material in textbooks, journals and material accessible electronically for example from web pages.'

Whereas **collusion** includes:
"Unauthorised and unacknowledged copying or use of material prepared by another person for use in submitted work. This may be with or without their consent or agreement to the copying or use of their work."

If copied with the agreement of the other candidate both parties are considered guilty of Academic Irregularity.
Please remember submitting portions of work already assessed is **Self-Plagiarism** and is also a serious academic irregularity.

Penalties for Academic irregularities range from capped or zero grades for elements of modules, to dismissal from the course and termination of studies.
To ensure that you are not accused of plagiarism, look at the NOW page Plagiarism and Academic Integrity at NTU. for guidance.

To help you avoid plagiarism and collusion, you are permitted to submit your work **once** to a separate drop box entitled "Draft report" to view both the matching score and look at what areas are affected. It is then down to you to make any changes needed.

Turnitin cannot say if something has been plagiarised or not. Instead it highlights matches between your text and other Turnitin content. There is no Good or Bad score, it depends on the piece of work.

If you find your text matching there may be a problem, see the examples below.

1) The reference section is highlighted. This may mean you have referenced correctly, and this has been matched with other well referenced documents online.

2) A table containing class data is highlighted. This is acceptable as long as any text accompanying the table is not similar picked up as identical

3) Paragraphs of text in the introduction or conclusion sections are highlighted. This may mean they have been copied exactly from another source. Even if this source is referenced this is bad practice, see advice below.

4) A sentence, or part of a sentence is highlighted. Sometimes there are few ways to write a sentence, especially straightforward ones. As long as this does not occur throughout a paragraph this may be acceptable. There will be occasions where a few words within a sentence produce a match. This is acceptable but ensure that this not a common occurrence or a patchwork of copied statements from different sources.

Chat GPT and other AI-powered language models
It is important to note when using any AI platform that they generate the most common responses to questions, not necessarily the correct ones. They also fabricate evidence. The material they produce is not your own words. Assessments require you answer questions giving your own view and in your own words. The outputs from Chat GPT do not provide that.

By presenting such material as your own words you are violating Academic Integrity policy, a matter that NTU takes very seriously.

The skills you develop during your time with us allow you to interrogate material and evaluate it, important skills in all careers. Chat GPT does not allow you to develop these.

I. Assessment Requirements

This coursework is an individual assignment. You must **implement it independently**.

This coursework is a mini project which applies machine leaning techniques to analyse a real-world data set. You are asked to implement several machine learning tasks using the data from "Stack Overflow 2024 Annual Developer Survey". This survey will help you understand the state of software developers, you can read more about the survey here: Stack Overflow Developer Survey 2024. The dataset can be downloaded from the stack overflow CDN and imported into Kaggle:

- https://cdn.sanity.io/files/jo7n4k8s/production/262f04c41d99fea692e0125c342e446782233fe4.zip/stack-overflow-developer-survey-2024.zip

The following tasks are required in the coursework.

(1) Implement exploratory data analysis and gain an understanding of the data set and its features. These concepts will be covered as part of the module content.

(2) Implement cluster analysis and understand the characteristics of developers in each cluster.

(3) Implement classification and build machine learning models for predicting whether a developer is in high income (compensation) based on developers' information. Note: for developers in this data set, low compensation is defined as annual compensation less than $55,000, otherwise, as high compensation. The compensation is provided in different denominations based on the developer's location. However, the converted value in USD is in column **DI**: ConvertedCompYearly

(4) Implement a regression model that predicts the salary of a person given some

attributes. Compare your result with the actual salary and ensure that the prediction error is as minimal as possible.

All tasks must be implemented using Python programming.

You will present the coursework in the form of a technical report containing the six sections listed in Table 1. A report template is provided at the end of this specification. Please follow the template to write your report.

Table 1 Structure of report, weights, and recommended pages

Section	Weighting	Recommended Pages per Section
1. Introduction	0.1	1-2
2. Data Understanding and Exploratory Data Analysis	0.2	2-4
3. Cluster Analysis	0.1	2-3
4. Machine Learning Methods and their Implementation	0.4	5-6
5. Evaluation Machine Learning Models	0.1	2-3
6. Discussions and conclusions	0.1	1
Appendix	0	No limit

A report with 15 pages is recommended. The report in total, however, must not exceed 20 pages (excluding title page, contents page, references, and appendices) with the font Arial and size 10 in the main text. A penalty of a single grade will be incurred if you exceed the 20-page limit. You may put extra information in appendices which is not counted in the 20-page limit.

You are asked to write the report with the provided report template at the end of the template. It is recommended to cite and list referees using Harvard Referencing style (see https://www.ntu.ac.uk/m/library/referencing-made-easy). However, other (author, year) styles like APA are also accepted.

By the submission deadline, you are expected to submit both your report (in MS Word or PDF format) and your Python source code (in *.ipynb or *.py format) to NOW Dropbox.

Your work will be assessed according to the assessment criteria provided in Section II.

The remainder of this specification provides you with detailed requirements for each area of content – you should read it very carefully.

1. Introduction

- State the coursework tasks and state the insight you intend to gain in the coursework.

- Introduce the CRoss Industry Standard Process for Data Mining (CRISP-DM) methodology. Explain its application and importance with appropriate reference to the literature.
- Discuss how you are applying CRISP-DM to the project in your coursework.

2. Data Understanding, Data Preprocessing, Exploratory Data Analysis

- Describe the background information of the data, such as how the data are collected and what is the purpose of the data.
- There are many columns in the data set. Select appropriate features from the columns for your analysis with justification. Obviously, some features have a bigger impact on the compensation than others. You will decide by yourself which features should be adopted in your project.
- Describe the selected features such as (though not limited to) their name, description, and data type. For numeric attributes, provide descriptive statistics. It is sufficient to describe only those attributes used in your analysis. Select at least 3 or more features and plus the target variable Compensation. For a Distinction coursework, it is expected that at least seven predictors will be selected, including both numeric and categorical features.
- Describe the quality of the data set, such as (though not limited to) determination of the number of all flawed instances, which include duplicate or conflicting instances as well as instances with missing values, erroneous values, or outliers.
- If any duplicate or conflicting instances, missing values, outliers/erroneous values, outliers exist, demonstrate how you clean these values.
- Conduct the exploratory data analysis for understanding the data, such as (though not limited to), identify outliers using a histogram or box plot; visualise the distribution of one categorical attribute using a pie plot or bar plot; explore the relationship between two features using a scatter plot; explore the relationship among three features by a scatter plot.

3. Cluster Analysis

- Describe the process of data transformation and normalization used in cluster analysis.
- Perform cluster analysis of the data set using some clustering methods (such as k-Means and hierarchical clustering). Implement cluster analysis using Python language. Describe parameter setting, initialisation, stopping criterion and discuss how you choose the optimal number of clusters.
- Describe the characteristics of each cluster that are generated in cluster analysis.

4. Machine Learning for Classification and their Implementation

- Describe the workflow of machine learning for classification with a flow-chart.
- State and describe classification methods that are used in your coursework. The

methods may be chosen from those taught in this module, such as k-Nearest Neighbour, Decision Trees, Logistic Regression. It is also allowed to choose methods that are not taught in this module. For a Distinction coursework, at least 3 classifiers should be chosen for classification.

- State and describe the regression models that are used for the salary estimation. You may choose from the methods covered in the module as well as those not covered in this module.
- Describe parameter setting in your classification and regression method(s).
- Describe the process of data transformation and normalization for the tasks.
- Build and implement machine learning models and tune hyper-parameters in these models for good performance. You may implement these models using Scikit-Learn modules or other Python libraries that are not taught in this module.
- Implement ensemble learning for classification. Describe the ensemble method(s) that you are using.

5. Evaluation Machine Learning Models

- Evaluate and compare the performance of different machine learning models. You should at least use one or more of the performance metrics (as appropriate), such as accuracy, confusion matrix, recall and precision, or Receiver Operating Characteristic Curve (ROC curve), Error Rate etc.
- Explain results using appropriate tables or figures.
- Critically review which model performed best and how hyper-parameter tuning change the performance of the models.

6. Discussions and Conclusions

- Summarise your work and findings in this mini project, such as how the selected features influence the developers' compensation.
- Describe what kind of insight that you have gained from this module.
- Explain whether and how well has the module developed your understanding of AI and Machine Learning.

Finally, it must be pointed out that there exist some online Jupyter notebooks on this data set. It is allowed for you to study these notebooks, but you must implement your own code in your coursework and cite these notebooks in your bibliography (if you used any). While you can use ChatGPT or other Large-Language Models (LLMs) to better understand the module and assessment, you should be careful not to copy the content as this will be flagged as generated by ChatGPT and could lead to academic irregularities. Therefore, you should ensure that the report and implementation are your own work.

You have one chance to check the similarity between your work by submitting your report and source code to Draft folder on NOW Dropbox. Turnitin similarity score should be somewhere around 30% for the report and around 60% for the code.

II. Assessment Criteria

Class/ Grade/ Assessment Criteria	Distinction Low 13 | Mid 14| High 15 *Exceptional Distinction 16	Commendation Low 10 | Mid 11| High 12	Pass Low 7 | Mid 8| High 9	Fail Mid 4|Marginal 6	Fail Low 2 *Zero 0	Comments	Grade 0
	* Excellent use of sources that evidence independent study and, in some cases, content that is not taught. Writing and content are nearly perfect				* A section is missing		
Section 1. Introduction Weighting is 0.1	Excellent description of the machine learning tasks in the coursework. The insight intended to gain in the coursework is excellent. Deep understanding of the CRISP-DM methodology. Excellent discussion of applying CRISP-DM to the project in the coursework.	Good description of the machine learning tasks in the coursework. The insight intended to gain in the coursework is good. Good understanding of the CRISP-DM methodology and its applications but may not in depth. With appropriate reference to the literature. Good discussion of applying CRISP-DM to the project in the coursework but may not in depth.	Reasonable description of the machine learning tasks in the coursework but without details. The insight intended to gain in the coursework is reasonable. Reasonable description of the CRISP-DM methodology and its applications but may miss some details. Lacks appropriate reference to the literature. Reasonable discussion of applying CRISP-DM to the project in the coursework but may be brief.	Insufficient description of the machine learning tasks in the coursework. The insight intended to gain in the coursework is insufficient. Insufficient description of the CRISP-DM methodology and its applications. Insufficient discussion of applying CRISP-DM to the project in the coursework.	No meaningful information about the machine learning tasks in the coursework and the insight intended to gain in the coursework. No meaningful description of the CRISP-DM methodology and its applications. No meaningful discussion of applying CRISP-DM to the project in the coursework.		0

Section 2. Data Understanding, Data Pre-processing, Exploratory Data Analysis Weighting is 0.2	Excellent selection of features with justification in depth. Excellent description of the data and features with details. Excellent statistical summary. Excellent understanding of data quality in depth. Excellent description of the relevant data pre-processing with excellent critical consideration and justification. Excellent exploratory data analysis using different methods with detailed explanations.	Good selection of features with good justification. Excellent description of the data and features but may be not in depth. Good statistical summary. Good understanding of data quality. Good description of the relevant data pre-processing with some critical consideration and justification. Good exploratory data analysis with good use of graphics but may miss some detailed explanation.	Selection of reasonable features but with some justification. Reasonable description of the data and features but may be not in details. Lack statistical summary. Reasonable understanding of data quality but may miss some details. Reasonable description of the relevant data pre-processing but with insufficient critical consideration and justification. Reasonable exploratory data analysis but with insufficient use of graphics and miss some details.	Selection of simple features with little justification. Some description of the data and features. No statistical summary. Some understanding of data quality but insufficient. Some description of the relevant data pre-processing but without critical consideration and justification Some exploratory data analysis but without use of graphics and miss some details.	No meaningful selection of features. No meaningful explanation of the data and features. No meaningful description of the relevant data pre-processing and no critical consideration and justification. No meaningful exploratory data analysis.		0
Section 3. Cluster Analysis Weighting is 0.1	Excellent cluster analysis with detailed explanation. Excellent description and justification of parameter setting and stopping criterion in depth. Excellent explanation of clustering results in	Good cluster analysis but may missing some details. Good description and justification of parameter setting and stopping criterion but may not in depth. Good explanation of clustering results but	Reasonable cluster analysis with some omissions. Reasonable description and justification of parameter setting and stopping criterion but with many omissions. Reasonable explanation of clustering results but	Some cluster analyses but not insufficient. No description and justification of parameter setting and stopping criterion. Lack explanations of results.	No meaningful cluster analysis.		0

	depth.	not in depth.	with many omissions.					
Section 4. Machine Learning for Classification and their Implementation Weighting is 0.4	Excellent description of machine learning workflow with excellent flow chat. Excellent explanation and justification of the selected machine learning methods and if applicable, with excellent use of mathematical formulas or diagrams. Excellent description of data transformation and normalization with details. Excellent implementation with excellent hyper-parameter tuning and with detailed explanation. Excellent use of ensemble learning with detailed explanation.	Good description of machine learning workflow with flow chat but may miss some details. Good explanation and justification of the selected machine learning methods but miss some details, and if applicable, may without use of mathematical formulas or diagrams. Good description of data transformation and normalization but may miss some detail. Good implementation with good hyper-parameter tuning methods but not with detailed explanation. Good use of ensemble learning but not with detailed explanation.	Reasonable description of machine learning workflow but may without flow chat and miss many details. Reasonable explanation and justification of the selected machine learning methods but may miss many details and if applicable, without use of mathematical formulas or diagrams. Reasonable description of data transformation and normalization but may miss many details. Reasonable implementation with little use of hyper-parameter tuning methods and without little explanation. Reasonable use of ensemble learning but not with detailed explanation.	Some description of machine learning workflow but may be incomplete. Insufficient explanation and justification of the selected machine learning methods. Insufficient description of data transformation and normalization. Little implementation without hyper-parameter tuning methods and without explanation. No use of ensemble learning.	No meaningful description of machine learning workflow. No meaningful explanation and justification of the selected machine learning methods. No meaningful description of data transformation and normalization. No meaningful implementation.			0
Section 5. Evaluation Machine Learning Models	Excellent and critical evaluation and comparison of the performance of different machine	Good evaluation and comparison of the performance of different machine learning models through	Reasonable evaluation and comparison of the performance of machine learning models but use only use	Some evaluation and comparison of the performance of machine learning models but use only	No meaningful evaluation and comparison of the performance of machine learning			0

Weighting is 0.1	learning models through use of different performance metrics. Excellent explanation of the performance metrics. Excellent use of excellent tables or data visualisation.	use of different performance metrics but may miss some critical evaluation. Good explanation of the performance metrics but may miss some detail. Good use of tables or visualisation.	one or two performance metrics and may miss some critical evaluation. Reasonable explanation of the performance metrics but without details. Insufficient use of tables or data visualisation.	one performance metrics and may miss critical evaluation. Some explanation of the performance metrics. No use of tables or data visualisation.	models. No meaningful explanation of the performance metrics.		
Section 6. Discussions and Conclusions Weighting is 0.1	An excellent, clear and concise summary of the work and findings is given. An erudite account of the insight gained is provided. Clear indications as to how to advance the work further.	A good comprehensive summary and concise discussion is provided that clearly articulates the purpose and findings of the analytics project. Good insight into how their data analytic skills have been developed. Good indications as to how to advance the work further.	A reasonable summary and discussion is provided that clearly articulates the purpose and findings of the analytics project. Reasonable insight into how their data analytic skills have been developed. Reasonable indications as to how to advance the work further.	Little summary of the work and findings is given. Some insight gained is provided. Insufficient indications as to how to advance the work further.	No summary of the work and findings is given. No insight gained is provided. No indications as to how to advance the work further.		0

III. Feedback Opportunities

Formative (Whilst you're working on the coursework)

You will frequently be given informal verbal or written feedback regarding your performance on tasks relating to the coursework assessment during the lectures, surgeries, and/or laboratory sessions. Attendance is therefore important for your development and thus coursework success. In addition, your Tutor may provide you with additional interim formative check points depending on the delivery pattern of the course.

Summative (After you've submitted the coursework)

You will receive specific feedback regarding your coursework submission together with your awarded grade when it is returned to you. Your assessor will provide you with the following as a minimum:

- Your grade;
- A feedback comment (a statement regarding the quality of your work)
- A feed forward comment (a statement regarding how you could improve your data analytic knowledge and skills for the future)

IV. Resources that may be useful

Referencing styles please use Harvard as detailed here
Guide to planning your time here and an automated planner here
Further guidance on avoiding cheating is here

Remember to use Outlook or physical calendars to block out time between lectures and labs to work on this coursework.

V. Moderation

All assessments are subject to a two-stage moderation process. Firstly, any details related to the assessment (e.g., clarity of information and the assessment criteria) are considered by an independent person (usually a member of the module team). Secondly, the grades awarded are considered by the module team to check for consistency and fairness across the cohort for the piece of work submitted.

VI. Aspects for Professional Development

ALL aspects of this report will provide meaningful evidence of a wide range of academic, technical and 'soft skills' such as Python programming, organisation and planning, analytical reasoning, reflection and effective report writing. It also provides potential employers with clear evidence of your ability to manipulate and intelligently analyse large data sets to identify business value. The report itself covers examples of writing a scientific-style report, researching existing literature, programming projects, referencing appropriately, construction and proper labelling of figures

Many of these are useful transferable skills for employment applications or your Skills Portfolio. Similarly, the practical class protocols provide several examples appropriate for use in the Skills Portfolio as Python programming and data analysis skills.