National College of Ireland

Masters in Artificial Intelligence (MSCAIB)

Programming for Artificial Intelligence (H9PAI)

Project (70%)

Release Date: *1st November 2024*

Submission Date: *18th December 2024 @ 23.55 hrs*

1. Introduction

This project is designed to evaluate the learning objectives of the Analytics Programming and Data Visualisation module as outlined below:

LO1 Analyse, compare, contrast and critically evaluate the characteristics of programming languages and environments commonly utilised for AI solutions implementation.

LO2 Critically assess the challenges associated with implementing AI solutions for various problems.

LO3 Critically assess methods and practices for software development in order to design and implement AI solutions requirements.

LO4 Evaluate, design and implement AI solutions by using key algorithms, data structures, and relevant programming languages.

2. Project Requirements

The objective of this project is to analyze and visualize data using programming techniques applicable to AI solutions. Each team will work with complex and semi-structured data, applying AI-focused preprocessing and visualization to support model evaluation and interpretation.

This project is a **group assessment**, with each team comprising **2-3 members**.

Your project must incorporate the following elements:

1. Each team member must select one structured or semi-structured data set (e.g., XML, JSON, web scraping) for analysis. Data should be stored programmatically in appropriate databases, such as MongoDB for semi-

structured data or PostgreSQL for structured data. Data sets must be programmatically stored in appropriate database(s) prior to processing.

2. Programmatically pre-process, clean, and transform each dataset as needed to prepare for analysis and visualization. Store processed data in suitable databases for analysis.
3. Conduct analysis to uncover patterns or insights within the data, and present these findings through visualizations. The following are some visualizations that you can employ
   a. histograms or box plots to display the distribution of key features, noting any trends, outliers, or central values (e.g., median, quartiles).
   b. scatter plots with trend lines to examine relationships between numerical features, helping to identify correlations or trends.
   c. for categorical data, create bar charts to compare values across different categories, providing a clear visual comparison
   d. a correlation heatmap to highlight relationship between numerical features.

(Optional) AI Model Application:

- If relevant to the data, apply a pre-trained AI model to analyze or classify data. For example:
  - Use a pre-trained text model like BERT for sentiment analysis if the data includes text.
  - Apply an image classification model like ResNet or VGG16 if the data includes images.

Note: This component is optional and does not require training models from scratch- only applying pre-trained models to the data.

Each data set should contain at least 1,000 records. Some appropriate data sets may be found at:

- https://catalog.data.gov/dataset?res_format=XML
- http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/
- https://data.gov.ie/dataset?res_format=JSON
- https://catalog.data.gov/dataset?res_format=JSON
- https://data.worldbank.org/

A list of other potential sources will be posted on Moodle.

3. Deliverables

3A. Project Report

The project report should clearly present your objectives, methods, results, and interpretations, emphasizing data storage, visualization, and, if applicable, the AI model application. The report should be around 3,500 words (excluding references) and follow the IEEE format. The report should contain the following sections:

I. Abstract

Summarize the project's objectives, methods, and key findings in a concise paragraph. Briefly outline any insights gained from data analysis, pre-processing, and visualization. If the optional AI model application was used, include a brief mention of the model applied and its outcomes.

II. Introduction

Provide background context for the project and explain the motivation for selecting the data and analysis approach. Clearly state the objectives of the project, including the primary goal of uncovering patterns or insights within complex and semi-structured data.

Present any research questions guiding the analysis, especially if the optional AI model application was used (e.g., "Can sentiment analysis provide insights into customer feedback?").

III. Related Work

Discuss relevant academic or industry work that guided the project, particularly regarding data storage, pre-processing, visualization techniques, and optional AI model applications. It should be more than a mere summary of the works and should discuss their limitations and implications.

IV. Methodology

1. **Data Selection and Storage:** Provide details on each selected dataset, including the source and rationale for selection. Describe how each dataset was programmatically stored in an appropriate database (e.g., MongoDB for semi-structured data like JSON/XML or PostgreSQL for structured data). Also provide a brief justification for the chosen database, libraries, etc.

2. **Pre-processing and Transformation:** Describe the pre-processing steps, such as cleaning, formatting, or restructuring data, conducted to prepare it for analysis. Include any transformations applied (e.g., removing duplicates, handling missing values) and a brief explanation of why these steps were necessary.

3. **Data Analysis and Visualization:** Describe each visualization created, for each visualization type, explain the purpose it serves and why it was chosen for this data.

(Optional) AI Model Application: If a pre-trained AI model was used, describe the model, its purpose, and how it was applied to the dataset. Justify the model's relevance to the project and discuss any tools or libraries used (e.g tensorflow, pytorch, etc.). Outline the steps for applying the model, such as loading the pre-trained model, formatting the data for input, and interpreting the output.

V. Results and Evaluation

1. **Visualization Results:** Present the main findings from each visualization, including patterns or insights covered. For example, discuss notable trends, correlations or distributions observed in the data.

2. **AI Model Results (if applicable):** Summarize the findings from the optional AI model application, such as sentiment trends in text data or classification accuracy in image data. Include visualizations (if relevant) of model outputs, such as bar charts summarizing sentiment scores or accuracy metrics.

3. **Evaluation:** Evaluate how well the project objectives were met, including any limitations in data quality, pre-processing, or model application that may have influenced results.

VI. **Conclusions and Future Work**

In this section you should summarize the main findings and reflect on their implications within the context of the research question or project objectives. Discuss any challenges encountered, such as data limitations or model performance issues, and how they might be addressed in future work. Suggest possible extensions or improvements, such as applying additional models, testing alternative datasets, etc.

VII. **Bibliography**

Here you should provide a **complete list** of the academic works cited in the report, including online datasets, or technical documentation for libraries and tools used in the project. Use **IEEE in-text citation and bibliography.**

3B. Project Presentation

You should create a video presentation (maximum 10 minutes long) that will act as a discussion point for your work. It should be used to provide a discussion on what you did, how you did it, why you did it and what you discovered. Note that although individual members will be presenting different parts of the video, each member of the team is expected to be able to present all aspects of the work individually and without assistance from other group members, if required.

3C. Code Artefact

You should create a *zip* or *gz* archive all assets such as program code, data and system configuration details.

3D. Project Journal

Each member of each project team should maintain their own journal over the course of the project. The journal should provide a brief description of each task carried out by the team member, the time spent carrying out the task, and a description of any challenges or difficulties encountered and how they were addressed.

Note: The submission of the journal is mandatory. Zero marks will be awarded to any team member for the entire project if a journal is not submitted. The individual project journal will be used to weight the marks awarded to each member of the team, as follows:

WEIGHT	CRITERIA
60%	A poor journal that fails to provide sufficient information on the team member’s contribution to the project, the time spent working on the project or the challenges encountered.
70%	An adequate journal that provides rudimentary information on the team member’s contribution to the project, the time spent working on the project or the challenges encountered.
80%	A good journal that provides reasonably in-depth information on the team member’s contribution to the project, the time spent working on the project or the challenges encountered

90 – 100%	An excellent journal that provides comprehensive information on the team member’s contribution to the project, the time spent working on the project or the challenges encountered.

4 Submission

The project carries 70% of the total marks for the module.

The submission should consist of:

- A **project report** that must include the names and student numbers of all team member (as per NCI official documents) These must be clearly visible on the front page of the report. The report should be named *TeamX.pdf* replacing *X* with your team number, and should be uploaded as a PDF document to the **Project Report** Turnitin link on Moodle.
- The **project cover sheet** as available here on NCI library website declaring any usage of AI tools in the assessment.
- A **code artefact** that should be uploaded as a *zip* or *gz* archive to the **Code Artefact** link on Moodle. This should be named *TeamX.zip* or *TeamX.gz*, replacing *X* with your team number.
- A **video presentation** that must include the names and student numbers of all team member (as per NCI official documents). These must be clearly visible at the start of the video. This should be uploaded as a **mp4** video named *TeamX.mp4* to the **Project Presentation** link on Moodle, where X is the number allocated to your team.
- A **project journal** (in a pdf format) providing an insight into the contribution each individual team member made to the project. One such report should be written and submitted separately by each team member. This report should be uploaded as **x12345678_teamX.pdf**, replacing *x12345678* with your student number and X with your team no.

Please following the file naming conventions given above.

Late submissions will not be accepted unless an extension has been requested through NCI360 and officially approved.

5 Marking

The project will be marked according to the grading rubric provided in the last two pages of this document.

6 Academic Integrity

Any written work created by others must be properly cited and should be paraphrased or summarised where possible, otherwise it should be included in quotes. Figures not created by you should include an acknowledgment detailing the name(s) of the creator(s). Code found on the internet should not be claimed as your own, but instead a comment should be included in the source code indicating where you obtained it.

Students are strongly advised to familiarise themselves with the Guide to Academic Integrity produced by the NCI Library[1].

Note: All submissions will be electronically screened for evidence of academic misconduct, e.g. plagiarism, collusion and misrepresentation. Any submission showing evidence of such misconduct will be referred to the college's academic misconduct committee for disciplinary action.

[1] https://libguides.ncirl.ie/academicintegrity

Grading Rubric – Programming for AI - Project

Semester 1 - 2024/25

Criterion	**Solid H1** ≥ 80%	**H1** ≥ 70% < 80%	**H2.1** ≥ 60% < 70%	**H2.2** ≥ 50% < 60%	**Pass** ≥ 40% < 50%	**Fail** < 40%
Project Objectives **(10%)**	Very challenging project objectives are exceptionally well presented, fully met and thoroughly discussed	Challenging project objectives are well presented, are fully met and thoroughly discussed.	Reasonable project objectives are well presented, fully met and adequately discussed.	Reasonable project objectives are clear, are mostly met and adequately discussed.	The objectives are clear, if unambitious and are at least partially met and briefly discussed.	The objectives of the project are unclear, have not been discussed. It is not possible to discern if the objectives have been met.
Literature Review **(10%)**	An excellent critical analysis of substantive and highly relevant literature.	A very good critical analysis of substantive and relevant literature.	A good analysis of relevant literature. The critical analysis aspect could be somewhat stronger.	An adequate analysis of mostly relevant literature. The critical analysis aspect could be significantly stronger.	A limited analysis of some relevant literature but it lacks evidence of understanding.	Little or no relevant literature reviewed. Very limited evidence of understanding.
Data Complexity and Handling **(15%)**	The data sets have been well prepared and meaningfully explored. All data sets were stored in appropriate databases before and after processing. At least two data sets have a high degree of complexity. At least one data set was programmatically retrieved - through an API or by web scraping.	The data sets have been well prepared and meaningfully explored. All data sets were stored in appropriate databases before and after processing. At least two data sets have a high degree of complexity.	The data sets have been well prepared and explored. At least one data set was stored in an appropriate database. At least one data set has a high degree of complexity.	The data sets have been appropriately prepared for analysis. At least one data set was stored in an appropriate database. At least one of the data sets is non-trivial.	The data sets were appropriately handled given the objectives. The use of databases is very basic and some inappropriate choices may be evident. The data sets are somewhat trivial.	Only one somewhat trivial data set was used. No database was used to store the data sets. No obvious development was carried out.
Data Processing **(20%)**	The data processing algorithms used play a well conceived and essential role in meeting the project objectives. The implementation significantly exceeds the stated minimum requirements.	The data processing algorithms used play a well conceived and essential role in meeting the project objectives. Multiple data processing techniques / languages were employed.	The use of data processing algorithms is well-thought and appropriate for the project objectives. Comprehensive use of at least one data programming language and multiple techniques.	The use of data processing algorithms is meaningful and appropriate for the project objectives. There is evidence of appropriate use of at least one data programming language and a small number of appropriate techniques.	Appropriate but basic use of data processing algorithms. Basic use of data programming languages and a limited number of techniques.	Poor or no implementation. If an implementation is provided, it demonstrates inappropriate use of data processing algorithms.

Grading Rubric – Programming for AI- Project

Semester 1 - 2024/25

Criterion	**Solid H1** ≥ 80%	**H1** ≥ 70% < 80%	**H2.1** ≥ 60% < 70%	**H2.2** ≥ 50% < 60%	**Pass** ≥ 40% < 50%	**Fail** < 40%
Data Visualisation **(15%)**	Visualisation choices are highly appropriate, exceptionally well-presented and are fully justified using relevant theory.	Visualisation choices are very appropriate, well-presented and are well justified using relevant theory.	Visualisation choices are appropriate, adequately presented and are accompanied by a basic justification that draws on mostly relevant theory.	Visualisation choices are somewhat appropriate, adequately-presented but lack a solid justification using relevant theory.	Visualisation choices appear to be random and are not justified using appropriate theory.	Visualisations (if any) are very poorly conceived, are illegible and are not discussed in the context of appropriate theory.
Results and Conclusions **(20%)**	Three or more insightful findings are excellently presented and thoroughly discussed in the context of the domain using appropriate references to prior work.	Three or more interesting and non-arbitrary findings are presented and thoroughly discussed the context of the domain using appropriate references to prior work.	Three or more interesting non-arbitrary findings are presented and thoroughly discussed.	Two or more interesting nonarbitrary findings are presented and appropriately discussed.	Two or more interesting nonarbitrary findings are presented but are poorly discussed.	Little to no non-arbitrary results and/or findings are presented.
Quality of Writing **(10%)**	Exceptionally well written, with no language errors. All figures are well conceived, readable and correctly captioned. The IEEE template is strictly adhered to. The report does not exceed the length limits. All references are appropriately and correctly used.	Well written, with no significant language errors. All figures are well conceived, readable and appropriately captioned. The IEEE template is adhered to. The report does not exceed the length limits. References are appropriately and correctly used.	Well written, but has a few significant language or style errors. Figures are well presented. The IEEE template and length limit are adhered to. References are complete and correctly used.	Adequately written. but as a few significant language and/or style errors. Some figures are may be hard to read. The IEEE template and length limit are mostly adhered to. References are complete, and correctly used.	Adequately written, with some significant language and/or style errors. Figures may be hard to read or presented in a sub-optimal manner. The IEEE template may not have been followed. References are mostly complete and correctly used.	Poorly written and littered with typographical errors and/or poor use of English. The IEEE template was not used. Figures may be hard to read. References (if any) are largely incomplete.