University of Suffolk
School of Engineering Arts, Science and
Technology

## Module: Data Mining and Statistics

| | |
|---|---|
| Module Tutor: | Dr Adnane Ez-zizi |
| Assessments: | Assessment: Technical Report<br>on or before 14/03/2024<br>Word count: 3000 |

## Module Rationale

Data science includes many techniques for classification, analysis and prediction. This module focuses on those techniques relating to data mining and statistically driven approaches. These techniques also have the advantage of being “explainable AI”, more so than deep learning approaches, and some are long established techniques of “business intelligence”.

## Aims

Through the completion of this module students will:

1. Develop a comprehensive understanding of data mining, business intelligence and statistical AI techniques, extending their research skills especially in terms of quantitative analysis
2. Learn how to discern which technique is most appropriate for a given scenario
3. Be able to apply and derive clear insights from relevant techniques, with the ability to articulate the limits, constraints and statistical soundness of the insights.

## Learning Outcomes

On successful completion of this module, a student will be able to:

1. Demonstrate substantial knowledge and systematic understanding of theory, techniques and applications of data mining, business intelligence, statistical AI and quantitative analysis
2. Assess which techniques are most appropriate for a particular scenario
3. Create an effective robust solution applying appropriate techniques to a particular scenario
4. Derive clear insights from the solution, articulating the limits and statistical soundness of the insights

## Assessment: Technical Report Overview

You are required to write a technical report in order to pass this module. The objective of the technical report is to explore a data set, define research question(s) based on research or business requirements and complete the analysis or model building to solve the problem you defined. The flow, look and feel of the technical report could be similar to a typical scientific paper or the template reports provided as part of this assessment brief (see https://brightspace.uos.ac.uk/d2l/le/lessons/76711/units/1186170).

Activities include:

- Defining the problem and motivating its importance (for this you will need to include an introduction and a literature review)
- Selecting and describing the data set
- Cleaning and preparing the data for modelling (e.g., dealing with missing values, transforming features, creating new features)
- Exploring the data using statistical summary measures (e.g., percentages, mean, median, standard deviation) and graphs (e.g., bar chart, line graph, histogram)
- Defining an experiment setup
- Implementing your proposed approach (typically you will compare multiple machine learning models)
- Evaluating and analysing your approach
- Showing statistical significance testing results if and as appropriate

## Data sets:

You may choose one of the following data sets:

- **IoT Intrusion Dataset:** The IoTID20 dataset is publicly available for applying intrusion detection techniques in IoT networks. The dataset is available at: https://sites.google.com/view/iot-network-intrusion-dataset/home

- **Student Dropout and Academic Success Dataset:** The dataset contains information related to students enrolled in different undergraduate degrees, such as agronomy, education, nursing and technologies. The data can be used to build machine learning models to predict students' dropout and academic success. The dataset is available at: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

If you wish to use a different data set, prior approval from the module leader is required. Your proposed dataset must be recent and not previously analysed in publicly available repositories.

## Structure of the report

Your report should include the following sections. As a minimum you need to have abstract, introduction, methods, results and discussion sections.

*Abstract (~250 words)*

Overall summary of your project including motivation, research problem, methods, main findings (e.g. model performance and evaluation) and any conclusions that could be drawn.

*Introduction and background (~500 words)*

Introducing the research/technical problem that you intend to solve and an overall summary of the issue with the set of research questions that you will be pursuing. You should also include a short critical literature review to motivate the research questions that you will be pursuing. The literature review should not be merely descriptive but should be critical in nature, with a view to highlighting the value and importance of your work.

*Methods and experimental setup (~1,000 words)*

Data description

Describe the data and any issues it might have (e.g., missing values, skewed data).

Data preprocessing

Describe any steps you undertook to prepare the data for statistical and machine learning analyses. This includes, for example, dealing with missing or incorrectly formatted values as well as feature engineering steps (e.g. encoding nominal variables, scaling continuous variables and crafting new features).

Statistical and machine learning methods

Describe the methods/tools you chose to address your research question(s). This could include:

- Stating null and alternative hypotheses.
- Describing how you will quantify reliability e.g. significance testing, confidence intervals. If appropriate, describe how you will measure effectiveness e.g. regression r-square, clustering evaluation, classification f1-score.
- Describe any machine learning or statistical techniques that you will be using, the selected features, model tuning and parameter selection approach.
- Explain how the analysis will answer your research question(s).

*Results (~750 words)*

- Exploratory data analyses
- Summarise results, reliability and comparison to any benchmark analyses or models.

- Critically analyse results e.g. limitations of data, setup or approach, characteristic errors, feature importance analysis, possible improvements
- Conclude with what you have learned from this study which would improve yourself as a data scientist. Would you recommend this as a solution to your problem? Provide reasons.

*Discussion and conclusion (~500 words)*

- Interesting discussion around the results
- Discussion around the limitations of the study
- Future directions
- Overall conclusions

You will also be marked based on the Python code that you use to run your analyses and generate your results as described below.

*Supporting Python code*

- The code should be submitted as a **SINGLE Jupyter Notebook**, and should replicate all the results presented in your report.
- Different parts of your Python script should include comments to make it easy to follow what you have done. Divide your notebook into separate sections, each referring to one part of the analysis (e.g. Data Preprocessing, Data Exploration, Data Modelling).
- All the outputs (e.g. tables, charts, results) should be generated within the Jupyter Notebook.

You will also be marked on the clarity and readability of your report. Therefore, use a formal academic writing style and make sure your final report is edited and does not contain grammatical or spelling issues. The report may include sub-level or other high-level headings as necessary.

## Submission Requirements

1. Ensure that your work is submitted as a SINGLE Word document or PDF with a filename matching the pattern sXXXXXX.doc or SXXXXXX.pdf where sXXXXXX is your UoS userid. Ensure that the document is marked with your UoS userid. Your name should not appear anywhere in the files.
2. Your report should use a line spacing of 1.5 and a body font size of 12pt.
3. Ensure that you submit your supporting Python code as a separate Jupyter Notebook (i.e., “ipynb” file) with all code outputs generated. In other words, submit your report and code on Brightspace as two separate files, so Turnitin can check your report for plagiarism.
4. If your submission exceeds the word count by up to 10% then there will be no penalty applied. Submissions that exceed the word count by more than 10% will be applied a fixed penalty of 5 percentage points (i.e., 5 marks). In all cases, the penalised mark will not be reduced below a pass level, assuming the work merits a pass. Tables, diagrams (including associated legends), appendices, reference lists, tables of contents, footnotes, and endnotes are excluded from the word count but should be used appropriately. It is for

the Module Leader to decide if there is excessive or inappropriate use of components excluded from the word count.

5. All bibliographies must be formatted according to Harvard style.
6. Your submitted report will be checked for plagiarism using Turnitin and any plagiarism detected will be dealt with seriously according to the University's policy for plagiarism, so paraphrase and cite your references carefully.
7. Remember that you can run a similarity report check once before submission by uploading your report document to the Similarity Report Check folder.
8. The use of AI to assist in completing the assignment is not permissible, except for the following tasks:
   i. Proofreading
   ii. Media creation for aesthetic purposes
   iii. Use as a learning tool, as you would use for example online tools like google search or Wikipedia
   iv. Code assistance using autocompletion function as with Github Copilot.
   v. Literature resource search (e.g. Elicit)

   If you have employed AI tools, you must include an acknowledgment section specifying how they were used (<u>the content of this section will not be included in the word count</u>). Select and adapt the appropriate statement below:

   1. "I attest that this project did not use AI at any stage in its development or in the creation of any of its components."
   2. "I attest that this project made use of AI in the following ways:". You must then use the following table to document your usage.

| **Usage** | **Tool Used (e.g., ChatGPT-4, Elicit)** | **How you edited the output, if at all** | **Conversation Link (If available)** |
|---|---|---|---|
| **Search of relevant literature** | | | |
| **Media Creation** | | | |
| **Proofreading** | | | |
| **Learning tool (e.g. research process, functioning of an AI algorithm)** | | | |
| **Code assistance** | | | |

## Assessment Rubric:

The following rubrics will be used in the marking of your assessment:

| Assessment Criterion | Max marks | > 75% | 50% – 74% | 25% – 49% | < 25% |
|---|---|---|---|---|---|
| **Abstract:**<br>The abstract succinctly summarises the research, including the purpose, methodology, and findings. It clearly outlines the scope and objectives of the study and provides a brief overview of the results and conclusions drawn. | 10 | Fully or mostly fulfilled (7.5 – 10pts) | Adequately fulfilled (5 – 7.49pts) | Partially fulfilled (2.5 – 4.99pts) | Minimally or none fulfilled (0 – 2.49pts) |
| **Introduction and background:**<br>The problem is well defined, and its significance is clearly and logically argued for. The topic is placed in context within its field, and the literature is well synthesised. | 10 | Fully or mostly fulfilled (7.5 – 10pts) | Adequately fulfilled (5 – 7.49pts) | Partially fulfilled (2.5 – 4.99pts) | Minimally or none fulfilled (0 – 2.49pts) |
| **Methods:**<br>The discussion of the methodology demonstrates an appropriate selection of research methods to answer the stated research questions, including the description of the data, data pre-processing and statistical and machine learning procedures | 30 | Fully or mostly fulfilled (22.5 – 30pts) | Adequately fulfilled (15 – 22.49pts) | Partially fulfilled (7.5 – 14.99pts) | Minimally or none fulfilled (0 – 7.49pts) |
| **Results:** | 30 | Fully or mostly fulfilled (22.5 – | Adequately fulfilled (15 – 22.49pts) | Partially fulfilled (7.5 – 14.99pts) | Minimally or none fulfilled (0 – 7.49pts) |

| | | | | | |
|---|---|---|---|---|---|
| Rigorous analysis of the data using tools and techniques appropriate to the field. Clear, concise and correct explanation of the results with the help of diagrams and tables. | | 30pts) | | | |
| **Discussion and Conclusion:**<br>The discussion and conclusion offer a concise yet thorough understanding and synthesis of the research outcomes, linking them effectively to the aims, objectives, and/or research questions. The discussion addresses the study’s limitations and future directions. | 10 | Fully or mostly fulfilled (11.25 – 15pts) | Adequately fulfilled (7.5 – 11.24pts) | Partially fulfilled (3.75 – 7.49pts) | Minimally or none fulfilled (0 – 3.74pts) |
| **Python Code:**<br>Code is submitted in the required format and is well documented. It can replicate all the analyses reported without issues and all the outputs have been successfully generated | 10 | Fully or mostly fulfilled (7.5 – 10pts) | Adequately fulfilled (5 – 7.49pts) | Partially fulfilled (2.5 – 4.99pts) | Minimally or none fulfilled (0 – 2.49pts) |
| Total mark (%) | | | | | |