Aston University
BIRMINGHAM UK

College of Engineering & Physical Sciences
Assignment Brief

CS4850 Data Mining AM41UD Understanding Data	Final Coursework
Dr. Felipe Campelo f.campelo@aston.ac.uk	WASS Calendar: https://wass.aston.ac.uk/pages/viewcalendar.page.php?cal_id=2525
Mrs. Raakhi Rachel Jose r.racheljose@aston.ac.uk	

Assignment Brief/ Coursework Content:

This coursework consists of a full applied data mining activity, including exploratory data analysis, feature engineering, modelling, and derivation of conclusions and recommendations. Data from a real application is provided, together with a set of specific objectives.

Many of the coursework activities will feel familiar to students who have engaged with the lectures, but the activities will also require some independent reading, exploration, and critical thinking.

The coursework is aimed at assessing your theoretical and practical skills in applying a range of data mining techniques to solve a real-world complex problem, including exploratory data analysis, suitable pre-processing methods and modelling approaches in an original manner, and your ability to describe and evaluate your data mining application.

The central objective of this coursework is to give you experience with the full process of developing a solution for a complex applied data science problem. You are given a set of data for developing your solution, and your model will be evaluated based on its performance on a set of new data with undisclosed data labels, to simulate the real-world deployment of your solution.

Although the final coursework must be completed **individually**, you are encouraged to work together during the tutorials, exchanging ideas and insights on how to best approach the problem. Your solution and final report need to reflect your own thinking and understanding of the problem (i.e., they cannot be **written** as a group).

Please read these instructions carefully. There are relevant points throughout the sections, as well as some potentially interesting tips.

General Guidelines:

This coursework can be approached using either **Python** or **R** (or even **Julia**, if you're feeling adventurous!). A short series of **R** and **Python** tutorials are available in the Teaching and Learning tab of our Blackboard page, if you think you could benefit from it.

Your solution must be **easily reproducible:** the person assessing the report must be able to re-run your full analysis without difficulty, to reproduce and verify your results. To this end, your solution must be produced either as a **Jupyter Notebook** (if you're solving the task in Python or Julia) or an **R Markdown** notebook (if you're doing it in R). Besides the code blocks, your notebook must also contain some structure detailing the different parts of your analysis (e.g., Title, Introduction, Exploratory Data Analysis and Data pre-processing, etc.), not just a sequence of code blocks.

Some limited programming support can be provided in the tutorials, but you are expected to already have the required programming skills (or to be able to pick it up independently).

Carry out the data analysis process in a systematic way. Take good notes on what you have done and save your work regularly so that experiments can be repeated if necessary. You may find version control tools such as git/Github[1] useful. Also, please notice that for full reproducibility you may want to make a note of the seed used for pseudorandom number generators (and add this as part of your Notebook submission)

Exploratory data analysis and data pre-processing are two aspects of your solution that are as important as the models themselves. Be sure to spend some time exploring the training dataset (or, more specifically, the subset of this data that you will reserve for your modelling decisions) with adequate data visualisation and summarisation, and consider carefully how you will treat outliers, feature scaling, missing data, or any other specific aspect of the dataset. Document your thought process, as your decisions should be discussed in your report summary (see section **Submission Details**). It is always important to notice that distinct models may require different pre-processing steps. Take the time to explore your options before deciding which route to follow.

Task Description

Introduction to the Problem

Linear B-cell epitopes are short protein fragments that are recognized by certain components of the immune system, and their identification is often an important early step in the development of vaccines, diagnostic tests and therapeutic interventions against infectious diseases, allergies and even some cancers. However, experimental discovery of epitopes is often a laborious and resource-intensive process, and computational methods have been used for the past three decades to help prioritise candidates for characterisation in the lab, including recent developments [1-3] that make this process considerably more efficient.

The data

In this coursework you will explore a dataset that is directly related to this problem. We will be exploring epitope data and trying to predict epitopes for *Trypanosoma cruzi* (*T. cruzi*), a parasite that is transmitted through insect bites and which causes a disease known as Chagas' disease. There are currently millions of Chagas' cases in several countries, mainly in South America, and the disease can affect the heart, digestive system and nervous system, potentially leading to death.[2]

[1] Check https://www.theodinproject.com/lessons/foundations-git-basics for a quick guide to git.

[2] https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)

The dataset you'll be working with throughout this coursework was created by parsing and consolidating data retrieved from multiple online databases – mainly the Immune Epitope Database (IEDB) and NCBI Protein, using research tools developed by our team at Aston University [1]. Our goal in this coursework is to develop an efficient **data mining pipeline** to (potentially) predict new, previously unknown epitopes in the proteins of this virus.

The data set that is available for you to build your predictive pipeline is called **df.csv.** It has the following structure:

- Some ***Information*** columns: these have names starting with "*Info_*" and provide general information about the origin of the observations. These columns are not useful for prediction (**they should not be used for preprocessing/modelling**), but some of them contain information that is relevant for other parts of the data mining pipeline (more on that later).

- Over 1,200 ***feature*** columns: these have names starting with "*feat_*" and contain features that were calculated for each observation. The specific meaning of these features is not important for the development of this coursework – they were extracted using a state-of-the-art feature embedder for protein data, ESM-1b.[3]

- 1 ***Class*** column, containing the target class.

Besides the characteristics above, the data in this Coursework has another characteristic that makes it challenging: There are dependencies between rows which can break the assumptions of certain pre-processing and modelling approaches and needs to be kept in mind. More details are provided in the **Coursework Requirements** section.

Besides the data set available for your pipeline development, there is also one ***validation data set*** (file ***df_holdout.csv***) that is provided. This file has the same structure as the training data, except for the fact that ***the Class attribute is not provided in this set***. It is part of your task to develop a competent data mining pipeline capable of predicting it.

Coursework requirements

1. You are required to complete your analysis using the data provided for your pipeline development (file **df.csv**). Your analysis **must** include a well-reasoned application of the usual data mining steps:

    a. **Exploratory data analysis**, including investigating feature types, possible missing values or outliers, variable scales, class imbalance, and the use of visualisation to support your discussion and pre-processing decisions.
    TIP: you'll be working with a *high-dimensional* data set. Looking at features individually will not be a particularly effective strategy (there are over 1200 of them!). Look for smart ways to explore high-dimensional data.

    b. **Data pre-processing**, including variable scaling (e.g., normalisation), treatment of missing data and outliers, plus any other pre-processing choices informed by your EDA, including:
        i. **Feature reduction.** This could be achieved using a few different approaches. Some possibilities include: simple feature selection using correlations, mutual information or gain ratio; feature extraction + selection using, e.g., PCA; or more sophisticated selection methods such as BORUTA or MRMR.
        ii. **Class rebalancing**, using either simple under- or over-sampling approaches; synthetic data generation; or modelling approaches that explicitly take data imbalance into account (e.g., cost-sensitive learning).

[3] If you're curious, check https://www.pnas.org/doi/10.1073/pnas.2016239118

c. **Modelling:** you must fit **at least one** classification model for predicting the Class attribute based on the features. Trying multiple classification approaches (e.g., Random Forest, XGBoost, Logistic regression, etc.) before choosing one is strongly encouraged. The main performance metric that you should use to assess your model should be the *balanced accuracy* (the arithmetic mean of specificity and sensitivity). Other performance indices can also be used as you develop your solution.
TIP: when assessing the performance of your models, make sure that you're estimating the *generalisation* performance, not the training performance. The column **Info_cluster** in the training datasets should be used as the grouping variable when splitting the data – don't use random train/test splits or simple (ungrouped) cross-validation, as this can cause strong data leakage and result in artificially inflated performance metrics. Instead, use strategies such as GroupKFold if you're using the *Sklearn* (in Python) or *caret* (in R) packages, or simply split your training data into subsets for training and test based on the groups informed in variable **Info_cluster**.

d. **Model assessment:** you must report the estimated *balanced accuracy* for your model. This value must represent an estimate of the generalisation performance of the model, i.e., how well the model is expected to perform once it is presented with new data. See the tip provided in item (c).

2. You are required to **use your final selected data mining pipeline (pre-processing steps + model) to predict the classes for the holdout observations (from file *df_holdout.csv***). These predictions should be saved in a csv file and submitted together with your report. More details are provided in section **Submission Details**.

Your full analysis process should be documented in a *Jupyter Notebook* (or alternatively an *R Markdown Notebook*) in a way that is easily reproducible. Additionally, you will need to submit a 1,000-word *Analysis Summary* explaining your analysis choices and conclusions; and a CSV file containing your predictions for the entries in **df_holdout.csv**. More details are provided in Section "*Submission Details*".

A template for what a submission should look like (including a skeleton Notebook, Summary and predictions file) will be made available on Blackboard.

Assessment:
Your coursework will be assessed in three parts: Quality of analysis (60%), quality of summary report (25%) and final pipeline performance (15%). The details of each element of assessment are provided below.

The marks are attributed according to the following criteria.
(**important:** *marks will be awarded proportionally according to the quality of the noted aspects*):

Quality of analysis process (60 marks):
(based on your *Jupyter notebook* or *R Markdown* notebook)

Criterion	Weight
(EDA) Does the notebook present a reasonable exploratory data analysis (EDA) process? Does it make effective use of adequate graphical elements and statistical summaries of the data to generate insights for the subsequent analysis?	6
(EDA) Are the insights gained from the EDA adequately mentioned in the Notebook (in the form of a short paragraph or bullet list)?	3
(DPP) Does the notebook present reasonable data pre-processing (DPP) steps, based on the insights gained during EDA? This includes, e.g., data splitting strategies, variable	4

scaling, treatment of missing data and duplicated observations, feature engineering choices etc..	
(DPP) Does the notebook include a **feature reduction** step as part of the DPP stage?	4
(DPP) Does the notebook explicitly discuss and implement an approach to address the **class imbalance** problem, either as a pre-processing activity or by using a modelling approach that's appropriate for imbalanced data?	4
(DPP) Are all the DPP steps adequately mentioned in the notebook (in the form of a short paragraph or bullet list)?	3
(Mod) Does the notebook adequately describe tests and development of at least one classification model for this problem?	4
(Mod) Does the notebook include the development of at least one additional model, following appropriate pre-processing steps?	4
(Mod) Does the notebook provide a performance comparison of the model(s) against some type of baseline?	4
(Mod) Does the report describe some sort of parameter tuning for the models?	3
(Mod) Does the report clearly indicate and encapsulate the final pipeline developed (composed of the data pre-processing + modelling steps), and report the estimated Balanced Accuracy score for the selected model?	7
(Mod) Is there evidence that the calculated performance values refer to generalisation performance rather than training performance (e.g., was it calculated using an adequately defined split of data, distinct from the one used for model development)?	5
(Mod) Are all the modelling steps anc choices adequately mentioned in the notebook (in the form of a short paragraph or bullet list)?	3
(Extra) Is there evidence of innovative and original thought in the development of a solution? Are there interesting insights into the data suggesting additional research, exploration of distinct quality measures, or in-depth discussions of specific pre-processing or modelling aspects? This could include discussions on feature relevance, insightful ways to look at this data, deeper considerations on model adequacy, etc.	6

Quality of Summary Report (25 marks):
(based on your *PDF* summary report)

Criterion	Weight
Does the report follow a coherent structure (e.g., Executive Summary – Problem Definition – Description of solution - Results - Conclusions)?	5
Is the CW submission complete, i.e.: - Jupyter or R Markdown notebook (**IPYNB** or **RMD** file) - 1000-word Summary Report (**PDF** file) - Predictions file (a properly formatted **CSV** file)	8
Does the summary report use adequate figures, tables or other data-related elements to support the explanations?	3
Does the summary report describe the general background of the problem being explored and remains focussed on describing the student's solution to that problem?	4
Does the summary report clearly describe the chosen pre-processing and modelling approaches, providing a brief rationale for their choice?	5

Final pipeline performance:

NOTE: The final pipeline performance will be assessed on the predictions generated for a **holdout** data set. This set is intended to simulate the real-world scenario of having your model deployed to predict new data, for which the actual class is unknown.

You must use your final predictive pipeline to generate predictions for the observations provided in the file **df_holdout.csv**. These predictions must be submitted as a **CSV file** containing the following columns: *Info_PepID*, *Info_pos* (taken from **df_holdout.csv**) and *Prediction* (the Class value that your model predicted for each row). If your model outputs probabilities, these may be added (but this is not required) to the CSV as a column *Probabilities.*

The performance of your model will be calculated automatically based on your submission, so it is **very important** that this file is in the required format. **Only a single prediction file should be submitted**, containing the predictions provided by the final predictive pipeline that is presented in your report.

The final pipeline performance marks (15% of total CW marks) are attributed according to:

Criterion	Weight
Is the final *Balanced Accuracy* score achieved on the holdout set consistent with the estimated generalisation performance documented in your summary report or notebook? (this is a proxy for good process in preventing data leakage in your processes)	10
How did your final *Balanced Accuracy* on the holdout set compare to the rest of this year's cohort? (100% for best performance, 0 for worst performance or for non-submitted CSV / non-calculable from CSV; Linear scaling for intermediate values)	5

As part of your feedback report, you will receive the performance value of your model on the holdout set.

A final tip

This is a challenging data mining exercise that will require some creativity and innovative thinking to be successfully solved. Be extra careful with overfitting and data leakage when developing/evaluating your models. The **Tutorials** will be used to provide guidance on the development of your solution, so please make sure to engage with those!

Recommended reading/ online sources:

All course materials are available on Blackboard.

- Lecture notes and recordings are stored under *Learning resources.*
- Additional materials are available under **Talis Reading List** as well as under **Tutorials and Resources**

KDNuggets offers excellent tutorials and online resources for further learning:

https://www.kdnuggets.com/tutorials/index.html

Further references about the underlying problem:

[1] J. Ashford, J. Reis-Cunha, I. Lobo, F. P. Lobo, F. Campelo (2021): *Organism-specific training improves performance of linear B-cell epitope prediction*. https://doi.org/10.1093/bioinformatics/btab536

[2] J. Ashford, A. Ekart, F. Campelo (2022): *Estimating the limits of organism-specific training for epitope prediction*. https://doi.org/10.1101/2021.11.02.466801

[3] F. Campelo *et al.* "Phylogeny-aware linear B-cell epitope predictor detects candidate targets for specific immune responses to Monkeypox virus." https://doi.org/10.1101/2022.09.08.507179

Key Dates:

29 January 2024	Coursework set
22 April 2024 (17:00h UK time)	Submission date
22 May 2024	Expected feedback return date (individual coursework feedback summary + provisional marks, available via Blackboard)

Submission Details:

Coursework files must be submitted electronically on Blackboard, using the links available under the "**Assessment**" area. Late submissions will be treated under the standard rules of Aston University, namely: 10% penalty for each working day after the deadline, up to a maximum of 5 working days, after which submissions are no longer marked.

The following files must be submitted:

- A single **Jupyter / R Markdown Notebook**: please clearly indicate your name at the top of the notebook. The file will be submitted through a specific link which will be located under **Assessment Submission** area of the module's Blackboard page. The file name of the notebook should follow the naming convention ***Lastname_Firstname_ModuleCode_Notebook***. Unless an alternative submission format is previously agreed with the module leader, the notebook **must** be submitted as either an **ipynb** or an **Rmd** file. Solutions submitted as any document format other than these will not be marked. The notebook should be well-organised to describe your data analysis steps, with an appropriate mix of text/Markdown blocks and code blocks.
  (note: this file can be compressed as a ZIP file if needed for submission)
- A single **Summary Report**: Please clearly indicate your name at the top of the notebook. The file will be submitted through a specific link which will be located under **Assessment Submission** area of the module's Blackboard page. **The file name of the summary report should also follow the naming convention** ***Lastname_Firstname_ModuleCode_Report.pdf***. The summary report must be submitted as a **PDF file. Do not** submit it as a MS Word, ODS or other file formats. The total length of the summary report (excluding template section headers, title page, references, and tables) must not exceed 1,000 words. Submissions exceeding this limit by more than 10% will be subject to a flat penalty of 5 marks.
- A single **prediction file**: this is a file containing the predictions of your selected model on the **holdout** data. This must be a **CSV** file containing a data frame with the structure defined in section *Final Pipeline Performance*. The file will be submitted through a specific link which will be located under **Assessment Submission** area of the module's Blackboard page. The name of the file must follow the naming convention ***Lastname_Firstname_ ModuleCode_predictions.csv***.

General Marking Descriptors:
(refer to section *Assessments* for details)

Fail	
0 - 39	Little or no evidence of work. Report fails to address even basic aspects of preliminary data exploration, pre-processing, and modelling.
40 - 49	There is some evidence of work, but only preliminary. Incomplete data mining process, report is too brief or too poorly structured to cover all the aspects of data mining process.
Pass	
50	Data exploration finds a few key aspects of the data characteristics, pre-processing steps tested without justification, development and evaluation of some baseline models, missing important details in the report.
51 - 59	Report is generally well-organised; some preliminary pre-processing is performed. Models are deployed, but little or no evidence of further model exploration. Some mistakes present in the analysis.
Merit	
60 - 69	Data exploration finds several key aspects of the data characteristics. Some relevant pre-processing steps tested and rationale provided. Reasonable models developed.
Distinction	
70 - 79	Data exploration finds all key aspects of the data characteristics. Evaluation of data pre-processing steps in relation to the key aspects discovered is well documented. Systematic development and evaluation of some sensibly chosen models. All requirements are fulfilled.
80 - 89	Professionally presented data analysis. Evidence of critical evaluation of existing knowledge, some new ideas and insights into the problem.
90+	All aspects of the coursework are presented at a solid professional level, similar to what could be found in an academic publication. Evidence of insights that could suggest to new ways to explore this data.