COMP1804	**Applied Machine Learning**	**Faculty Header ID:**	**Contribution 100% of course**
Course Leader **Dr. Stef Garasto**	**COMP1804 Coursework**		**Deadline Date:** **26/04/2022 (23:30 UK time)**
This coursework should take an average student who is up-to-date with the lectures and the labs approximately 50 hours Feedback and grades are normally made available within 21 days of thecoursework deadline			
Learning Outcomes: 1. Rationalise appropriate scenarios for Machine Learning applications and evaluate the choice of machine learning methods for given application requirements. 2. Demonstrate competency in using appropriate libraries/toolkits to solve given real-world Machine Learning problems and develop and evaluate suitable application. 3. Understand and apply the relevant input data preparation and processing required for the Machine Learning models used, and quantitatively evaluate and qualitatively interpret the learning outcome. 4. Recognise and critically address the ethical, legal, social and professional issues that can arise when applying Machine Learning technologies.			

Plagiarism is presenting somebody else's work as your own. It includes: copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as your own work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by the University. Please see your student handbook for further details of what is / isn't plagiarism.

All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.

Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.

Coursework Submission Requirements

An electronic copy of your work for this coursework must be fully uploaded on the Deadline Date using the link on the coursework Moodle page for COMP1804. For this coursework you must submit 4 separate files:

- A single pdf file named 'report.pdf' which will be the written report; the written report must have a maximum limit of 3500 words **including** references. It is also recommended for the report to have at least 2000 words.
- A single csv file named 'exclusions_dataset_taskX.csv' (X is the number of the task you chose). The csv file will include **all** data from your dataset with an annotation as to whether each data point has been excluded from further analysis and why. Regarding the format:
    - the first line should be:
        - *data_id,excluded,reason_for_exclusion*
    - next lines should list the ID for the data point & respective completed fields.
        - The data_id should be the same as shown in the dataset or should be the image filename (whichever applies).
        - Use 1/0 to indicate whether a specific review was excluded (1) or not (0).
        - The field "reason_for_exclusion" can have one label from the following: ["duplicate", "missing_value_or_label", "invalid", "other", "N/A"]. "N/A" indicates the non-excluded rows.
        - For example:
            - *doc_2017_001,1,duplicate*
              doc_2018_051,0,N/A
- A single .zip file containing an ipython notebook file and a pdf (or html) file, showing the Machine Learning implementation (2 methods) on your dataset. The files should be named 'code_ML.ipynb' and 'code_ML.pdf'.
    - The pdf file should be generated from the python notebook via "File/Download As/PDF" after executing **all** the cells. It should show exploratory data analysis, data cleaning, data pre-processing, model training with curves, systematic experimentation and evaluation of the test set.

Some general notes:

- Any text in the document must not be an image (i.e. must not be scanned) and would normally be generated from other documents (from the latex template in this case).
- There are limits on the file size (see the relevant course Moodle page).
- Make sure that any files you upload are virus-free and not protected by password or corrupted otherwise they will be treated as null submissions.
- Your work will not be printed in colour. Please ensure that any pages with colour are acceptable when printed in Black and White.
- You must NOT submit a paper copy of this coursework.
- All courseworks must be submitted as above. Under no circumstances can they be

accepted by academic staff.

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances, and penalties for Assessment Offences. See http://www2.gre.ac.uk/current-students/regs

Detailed Coursework Specification

Designing a machine learning solution requires considering several aspects of the problem, the availability and cleanliness of data and corresponding annotations, nature of the problem addressed, methodology choice, evaluation among others. It is important to be up to date with current practices and Machine Learning (ML) techniques used in the modern software that drives many computers and devices today and be familiar with their strengths and limitations. It is of equal importance to familiarize with the whole data processing and evaluation pipeline enabling successful implementation of Machine Learning techniques. Adding these skills to your portfolio will increase your employability and will help you to aim for higher paying jobs in industry, as well as academia.

The task is to implement ML solutions for your chosen problem using the dataset assigned to you. You should choose one of the sub-tasks below. Note that for each sub-task a related labelled datasets will be distributed. Which dataset you should use depends on the first letter of your surname. You should implement the whole procedure for designing a ML approach for solving the problem and produce a written report individually.

A list of references for each dataset made available is provided at the end. While most of the datasets listed are available for download in their original form, please do *not* use data downloaded from anywhere else. This is because these datasets have been slightly modified to be better tailored to this specific module.

Available sub-tasks

Sub-task 1: Text Classification/regression – peer reviews.

This task is to implement a ML solution for text classification/regression (long texts). It uses a dataset of ML paper peer reviews from the International Conference of Learning Representation (in the years between 2017 and 2020) [1,2].
Specifically, you will use as input a text document concatenating: the title of the paper, the abstract of the paper, the review comments, the final acceptance/rejection comment. Such input should be used to predict the following attributes:

- Acceptance status (‘Accept’ or ‘Reject’)
- Review score (Integer number between 1 and 10).

Note that for the latter attribute you can choose whether to use multiclass classification or regression. You can choose whether to predict both features simultaneously or separately.

Additionally, the dataset is provided with a further attribute, the reviewer confidence score (an integer number between 1 and 5), which is optional to use. If you want to explore the

data further, a separate dataset with the text field split into the original fields “review comments”, “paper title”, “paper abstract” and “final acceptance/rejection comment” can be provided upon request.

Sub-task 2: Image classification – skin lesions.

This task is to implement a ML solution for a classification problem from images. Specifically, you are provided with images of skin lesions [3] and your task is to correctly predict the following attributes:

- Whether a skin lesion is benign or malign (1 for ‘is_benign’, 0 for ‘is_malign’)
- The fine-grained diagnosis for the skin lesion (7 possible categories).

You can choose whether to predict both features simultaneously or separately. Additionally, the dataset is provided with a further attribute, the location of the skin lesion (for example, “scalp”), which is optional to use. If you want to explore the data further, a separate dataset with more attributes can be provided upon requests. The dataset has been adapted to the requirements of this module; the original dataset was released under the terms of the CC BY-NC 4.0 licence by Tschandl et al. [3].

Sub-task 3: Image classification - advertisements.

This task is to implement a ML solution for a classification problem from images. Specifically, you are provided with images of advertisements [4] and your task is to correctly predict the topic of each advertisement.

- Images are of different sizes and there are 39 possible topic categories.
- You may choose to group together some of the categories (keeping no less than 12 categories). You should thoroughly discuss (and will be evaluated on) the reasons behind and the implications of grouping together different categories.

Sub-task 4: Text classification – amazon reviews.

This task is to implement a ML solution for a multi-task classification problem from text data (mostly short texts). Specifically, you are provided with Amazon reviews [5] (the text is the review title and the review main body joined together) and your task is to predict the following attributes:

- The number of stars associated with the review (on a scale of 1 to 5).
- Whether a product is from the category “Video Games” (“video_games”) or “Musical Instrument” (“musical_instrument”).

Note that for the first attribute you can choose whether to use multiclass classification or regression. You can choose whether to predict both features simultaneously or separately.

Additionally, the dataset is provided with a further attribute: whether the review is verified or not (either True or False), which is optional to use. If you want to explore the data further, a separate dataset with the text field split into the original fields “review title”, “review main body” can be provided upon request.

Tasks:

1. **Practical Assignment** (complete code that is executed without errors). The source code must be **well documented and error free** (i.e. no debugging necessary to run). For each dataset, the assignment includes:
   - Exploratory Data Analysis (e.g. label distributions per attribute and per set).
   - Data cleaning.
   - Data Splitting (in training and test sets, but see below) and Data Pre-processing (where appropriate: normalization/standardization, data augmentation, over/under-sampling, text processing).
   - 2 ML Methodologies (a basic one & an additional one): appropriate ML methods should be used that have coherent implementations and sound pipelines, without any errors; (if the basic ML method is a Neural Network, the additional one can be another Neural Network).
   - Systematic experimentation: you should choose one parameter/attribute to change for each ML methodology (the attribute/parameter can be the same or different across the two methodologies) and show how it affects the results using clear and well formatted figures and tables. Bonus points are given for experimenting using a validation dataset.
   - Evaluation of the 2 methods using at least 2 metrics and showing 3-10 examples from the test dataset.

2. **Written Report:**
   - Document in IEEE conference format. Use template available on Moodle or on Overleaf (make a copy of the Overleaf template).
   - Should include references (citing other work) where appropriate (when images, data, code, or any other resources have been used from other sources)
   - Document structure:
     - **Abstract:** Briefly summarise what the report contains. That is: the task you are solving and why it is important; the outline of the ML methods you implemented and the systematic experimentation performed; the summary of your results and your conclusions. The abstract should be between 100 and 200 words.
     - **Introduction and related work**: This section should talk about the following:
       - The problem to be solved, why and to whom it matters, why it is challenging.
       - Existing work related to your chosen task (it can be about the exact same task or a similar one).
       - A brief overview of the dataset and the data pre-processing steps implemented.
       - Your chosen ML implementations and a brief overview about why they are appropriate.
       - What your systematic experiment is.
     - **Ethical discussion**: Identify and discuss some of the social, ethical and legal implications of your chosen task, from data collection and processing to the ML prediction. The discussion should take into account communities and people that may be affected by the ML system.
     - **Dataset preparation**: Describe exploratory data analysis, data cleaning, splitting and pre-processing and the reasons behind your design choices.

- **ML methods**: Describe and explain the 2 methods used and the reasons behind your design choices.
- **Experiments and evaluation.** Describe the systematic experimentation implemented for each ML method. Based on the experiments, evaluate, present, analyse and explain method performance and metrics used (why are the metrics appropriate?).
- **Discussion and future work**: Reflections on a) what worked well and what worked less well; b) reasons behind the performance obtained; c) how your work could be extended in the future and what addition can be made to it.
- **Conclusions**: A brief summary of the work done and what the main highlights were.
- **References**: All existing works and resources (code/images/etc) you used or talked about in your report must be cited properly.

Deliverables:

An admissible coursework submission needs to include:

- All coursework submission requirements as specified in the Coursework Submission Requirements section above should be uploaded by the Deadline Date using the link on the coursework Moodle page for COMP1804.

References

[1] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (PeerRead): Collection, insights and nlp applications," 2018. [Online]. Available: https://arxiv.org/abs/1804.09635.

[2] *Openreview*, github. Accessed: 11/01/2022. [Online]. Available: https://github.com/Seafoodair/Openreview.

[3] P. Tschandl, C. Rosendahl, and H. Kittler, H, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". *Scientific data*, vol. 5, no. 1, pp.1-9., 2018.

[4] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka, "Automatic understanding of image and video advertisements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 1705-1715.

[5] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* November 2019, pp. 188-197.