**ML Project**

*Dataset and analysis*

* Choose a health-related dataset among the ones listed below
* Think of a question you are interested in. This is entirely up to you. The aim should be prediction (not inference).
* Run at least two supervised and one unsupervised algorithm. The unsupervised algorithm can be used to give a descriptive overview of your data. The supervised algorithms should be at least two to compare results.

*Project presentation*

* The project presentation is due on 02/05/2022. Please keep in mind that it takes time to run things on the HPC, so aim to finish your project a couple of weeks in advance to give yourself time to prepare the presentation. You don’t need to write a report, just prepare the presentation.
* We might optionally ask to review your code, so please keep it tidy.
* In your oral presentation please stick to the following rough structure: 1-3 slides of background, 1-5 slides of methods including data description, pre-processing, model tuning, up to 10 slides of results, 1 slide for conclusions/future work (no discussion or public recommendations are needed). Try to keep the total around 15 slides.

*Various*

* It is ok to apply new methods to an existing paper/study in order to improve it

*What are we evaluating?*

* Clear and concise presentation of the rationale and question you chose
* Justification of methods chosen and detailed knowledge of both theory and best practices for each method (e.g. data processing required, parameter tuning etc.)
* Clear presentation of the results and appropriate comparison metrics
* Appropriate choice of visuals i.e. plots, tables, diagrams etc.
* Confident conclusions

***Datasets***

*Tabular*

* UK Biobank: access was given to all of you during TDS. The same scientific questions as in TDS are not permitted of course.
* MIMIC-III, please see below (text section)

*Imaging*

* Breast cancer imaging <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=64685580>
* Colorectal Cancer histology https://zenodo.org/record/53169#.Yi8qRBDP1MA

*Text*

* MIMIC-III data from intensive care patients including electronic health records and also tabular data. PLEASE NOTE: a request for access is needed, so it you choose this dataset request the data immediately to get it in time. You can list me (Dragana) as your reference. It is a very large data collection, so it will be helpful to limit your scope and it will require some confidence with large data handling.
* 200k Pubmed abstract, with every sentence labelled in one of 8 categories https://github.com/Franck-Dernoncourt/pubmed-rct

*Audio*

* cough recordings from COVID patients and controls https://zenodo.org/record/4498364#.Yi8m2RDP1MD

**Do imaging data for colorectal cancer?**

Analysis Plan

*Which machine learning method is most accurate in the prediction of colorectal tumours among different classes of colorectal histological images ?*

*Which machine learning method is most accurate in the prediction of different classes of colorectal histological images such that different types of tumours can be identified from images?*

*Are we able to differentiate different types of colorectal histological images to identify tumours and predict them in unseen data?*

Colorectal cancer data

* Two supervised and one unsupervised algorithm
  + Unsupervised as descriptive overview of data
  + Supervised to then compare between the two models

**Methods to use**

Supervised

* Random forest
* CNN

Unsupervised

* K-Means Clustering

**Images available**

* **Images\_tiles\_5000**
  + Contains 5000 histological images
  + 150px \* 150px
  + Each image belongs to one of 8 tissue categories
    - Adipose, Debris, Lymphoma, mucosa, complex, stroma, tumour and empty
* Larger\_images\_10
  + 10 larger histological images of 5000 x 5000 px each
  + Images contain more than one tissue type
* All images are RGB, 0.495 micro metre per pixel

*What are we evaluating?*

* Clear and concise presentation of the rationale and question you chose
* Justification of methods chosen and detailed knowledge of both theory and best practices for each method (e.g. data processing required, parameter tuning etc.)
* Clear presentation of the results and appropriate comparison metrics
* Appropriate choice of visuals i.e. plots, tables, diagrams etc.
* Confident conclusions

Clustering

* Have to first convert 4D arrays to 2D to allow clustering to take place
* This is due to having images converting to arrays containing 4 dimensions
  + No. of data, pixel length, pixel width, colour (RGB = 3)
* Then have to standardise the data

Presentation

**Description of data**

* Show example images of each of the categories

**Data Preprocessing**

* Convert 4D arrays to 2D for clustering and random forest
* Also standardised the data for clustering, random forest and SVM
* Extracted labels from the folder names

**K-means Clustering**

* Have to identify the number of clusters that are optimal
  + So looked at elbow and silhouette score
* Used PCA before clustering to reduce dimensionality

**Random Forest**

* While we lose interpretation through random forest, our focus is on prediction, to see if we are able to identify the right classification for each image
* Used random forest compared to gradient boosting as it is more computationally efficient and still accurate
* Used 5-fold cross validation using GridSearchCV

Note: For calculating F1 score, precision and recall, consider using average = “macro”.

**SVM**

* Conducted SVM with and without PCA prior to svm training
* Also normalised the data
* 5-fold cross validation using GridSearchCV

**Deep Learning (CNN)**

* Have to convert data to floats
* Normalise data by dividing by 255