# Assignment 2

**Biomedical Data Science (MATH11174), 22/23, Semester 2**

April 6, 2023

## Due on Thursday, 6th of April 2023, 5:00pm

! Pay Attention

The assignment is marked out of 100 points, and will contribute to ***30%*** of your final mark. The aim of this assignment is to produce a precise report in biomedical studies with the help of statistical and machine learning. Please complete this assignment using **Quarto/Rmarkdown file and render/knit this document only in PDF format** (rendering while solving the questions will prevent sudden panic before submission!). Submit using the **gradescope link on Learn** and ensure that **all questions are tagged accordingly**. You can simply click render on the top left of Rstudio (`Ctrl+Shift+K`). If you cannot render/knit to PDF directly, open **Terminal** in your RStudio (`Alt+Shift+R`) and type `quarto tools install tinytex`, otherwise please follow this link. If you have any code that does not run you will not be able to render nor knit the document so comment it as you might still get some grades for partial code.

Codes that are **clear and reusable will be rewarded**. Codes without proper indentation, choice of variable identifiers, **comments**, efficient code, etc will be penalised. An initial code chunk is provided after each subquestion but **create as many chunks as you feel is necessary** to make a clear report. Add plain text explanations in between the chunks when required to make it easier to follow your code and reasoning. Ensure that all answers containing multiple values should be presented and formatted only with `kable()` and `kable_styling()` otherwise penalised (**no use of `print()` or `cat()`**). All plots must be displayed with clear title, label and legend otherwise penalised.

This is an **individual assignment**, and **no public discussions** will be allowed. If you have any question, please ask on Piazza by specifying your `Post to` option to `instructors`. To join Piazza, please follow this link.

# Problem 1 (27 points)

File `wdbc2.csv` (available from the accompanying zip folder on Learn) refers to a study of breast cancer where the outcome of interest is the type of the tumour (benign or malignant, recorded in column `diagnosis`). The study collected 30 imaging biomarkers on 569 patients.

## Problem 1.a (7 points)

- Using package `caret`, create a data partition so that the training set contains 70% of the observations (set the random seed to 984065 beforehand).
- Fit both a ridge and Lasso regression model which use cross validation on the training set to diagnose the type of tumour from the 30 biomarkers.
- Then use a plot to help identify the penalty parameter $\lambda$ that maximises the AUC and report the $\lambda$ for both ridge and Lasso regression using `kable()`.
- ***Note : there is no need to use the `prepare.glmnet()` function from lab 4, using `as.matrix()` with the required columns is sufficient.***

```
## Answer in this chunk
```

## Problem 1.b (2 points)

- Create a data table that for each value of `lambda.min` and `lambda.1se` for each model fitted in **problem 1.a** that contains the corresponding $\lambda$, AUC and model size.
- Use 3 significant figures for floating point values and comment on these results.
- ***Note : The AUC values are stored in the field called cvm.***

```
## Answer in this chunk
```

## Problem 1.c (7 points)

- Perform both backward (we denote this as **model B**) and forward (**model S**) stepwise selection on the same training set derived in **problem 1.a**. Mute all the trace by setting `trace = FALSE`.
- Report the variables selected and their standardised regression coefficients in increasing order of the absolute value of their standardised regression coefficient.
- Discuss the results and how the different variables entering or leaving the model influenced the final result.
- ***Note : You can mute the warning by assigning `{r warning = FALSE}` for the chunk title***

```
## Answer in this chunk
```

###Problem 1.d (3 points)

- Compare the goodness of fit of **model B** and **model S**
- Interpret and explain the results you obtained.
- Report the values using `kable()`.

```
## Answer in this chunk
```

### Problem 1.e (2 points)

- Plot the ROC curve of the trained model for both **model B** and **model S**. Display with clear title, label and legend.
- Report AUC values in 3 significant figures for both **model B** and **model S** using `kable()`.
- Discuss which model has a better performance.

```
## Answer in this chunk
```

### Problem 1.f (6 points)

- Use the four models to predict the outcome for the observations in the test set (use the $\lambda$ at 1 standard error for the penalised models).
- Plot the ROC curves of these models (on the sameplot, using different colours) and report their test AUCs.
- Display with clear title, label and legend.
- Compare the training AUCs obtained in **problems 1.b and 1.e** with the test AUCs and discuss the fit of the different models.

```
## Answer in this chunk
```

# Problem 2 (40 points)

File `GDM.raw.txt` (available from the accompanying zip folder on Learn) contains 176 `SNP`s to be studied for association with incidence of gestational diabetes (A form of diabetes that is specific to pregnant women). SNP names are given in the form `rs1234_X` where `rs1234` is the official identifier (rsID), and `X` (one of A, C, G, T) is the reference `allele`.

## Problem 2.a (3 points)

- Read in file `GDM.raw.txt` into a data table named `gdm.dt`.
- Impute missing values in `gdm.dt` according to `SNP`-wise median `allele` count.
- Display first 10 rows and first 7 columns using `kable()`.

```
## Answer in this chunk
```

## Problem 2.b (8 points)

- Write function `univ.glm.test()` where it takes 3 arguements, `x`, `y` and `order`.
- `x` is a data table of `SNP`s, `y` is a binary outcome vector, and `order` is a boolean which takes `false` as a default value.
- The function should fit a logistic regression model for each `SNP` in `x`, and return a data table containing `SNP` names, regression coefficients, odds ratios, standard errors and p-values.
- If order is set to `TRUE`, the output data table should be ordered by increasing p-value.

```
## Answer in this chunk
```

## Problem 2.c (5 points)

- Using function `univ.glm.test()`, run an association study for all the `SNP`s in `gdm.dt` against having gestational diabetes (column `pheno`) and name the output data table as `gdm.as.dt`.
- Print the first 10 values of the output from `univ.glm.test()` using `kable()`.
- For the `SNP` that is most strongly associated to increased risk of gestational diabetes and the one with most significant protective effect, report the summary statistics using `kable()` from the GWAS.
- Report the 95% and 99% confidence intervals on the odds ratio using `kable()`.

```
## Answer in this chunk
```

## Problem 2.d (4 points)

- Merge your GWAS results with the table of gene names provided in file `GDM.annot.txt` (available from the accompanying zip folder on Learn).
- For `SNP`s that have p-value $< 10^{-4}$ (`hit SNPs`) report `SNP` name, effect `allele`, chromosome number, corresponding `gene` name and `pos`.
- Using `kable()`, report for each `snp.hit` the names of the genes that are within a 1Mb window from the `SNP` position on the chromosome.
- ***Note: That are genes that fall within +/- 1,000,000 positions using the pos column in the dataset.***

```
## Answer in this chunk
```

## Problem 2.e (8 points)

- Build a weighted genetic risk score that includes all `SNP`s with p-value $< 10^{-4}$, a score with all `SNP`s with p-value $< 10^{-3}$, and a score that only includes `SNP`s on the FTO gene
- ***Hint: ensure that the ordering of `SNP`s is respected.***
- Add the three scores as columns to the `gdm.dt` data table.
- Fit the three scores in separate logistic regression models to test their association with gestational diabetes.
- Report odds ratio, 95% confidence interval and p-value using `kable()` for each score.

```
## Answer in this chunk
```

## Problem 2.f (4 points)

- File `GDM.test.txt` (available from the accompanying zip folder on Learn) contains genotypes of another 40 pregnant women with and without gestational diabetes (assume that the reference allele is the same one that was specified in file `GDM.raw.txt`).
- Read the file into variable `gdm.test`.
- For the set of patients in `gdm.test`, compute the three genetic risk scores as defined in **problem 2.e** using the same set of `SNP`s and corresponding weights.
- Add the three scores as columns to `gdm.test` ***(hint: use the same columnnames as before).***

```
## Answer in this chunk
```

## Problem 2.g (4 points)

- Use the logistic regression models fitted in **problem 2.e** to predict the outcome of patients in `gdm.test`.
- Compute the test log-likelihood for the predicted probabilities from the three genetic risk score models and present them using `kable()`

```
#Answer in this chunk
```

## Problem 2.h (4points)

- File `GDM.study2.txt` (available from the accompanying zip folder on Learn) contains the summary statistics from a different study on the same set of `SNP`s.
- Perform a meta-analysis with the results obtained in **problem 2.c** (***hint : remember that the effect*** **`alleles`** ***should correspond***)
- Produce a summary of the meta-analysis results for the set of `SNP`s with meta-analysis p-value $< 10^{-4}$ sorted by increasing p-value using `kable()`.

```
#Answer in this chunk
```

## Problem 3 (33 points)

File `nki.csv` (available from the accompanying zip folder on Learn) contains data for 144 breast cancer patients. The dataset contains a binary outcome variable (`Event`, indicating the insurgence of further complications after operation), covariates describing the tumour and the age of the patient, and gene expressions for 70 genes found to be prognostic of survival.

### Problem 3.a (6 points)

- Compute the correlation matrix between the gene expression variables, and display it so that a block structure is highlighted using the `corrplot` package.
- Discuss what you observe.
- Identify the unique pairs of (distinct) variables that have correlation coefficient greater than 0.80 in absolute value and report their correlation coefficients.

```
## Answer in this chunk
```

### Problem 3.b (8 points)

- Perform PCA analysis (only over the columns containing gene expressions) in order to derive a patient-wise summary of all gene expressions (dimensionality reduction).
- Decide which components to keep and justify your decision.
- Test if those principal components are associated with the outcome in unadjusted logistic regression models and in models adjusted for `age`, `estrogen receptor` and `grade`.
- Justify the difference in results between unadjusted and adjusted models.

```
## Answer in this chunk
```

### Problem 3.c (8 points)

- Use PCA plots to compare the main driverswith the correlation structure observed in **problem 3.a**.
- Examine how well the dataset may explain your outcome.
- Discuss your findings in full details and suggest any further steps if needed.

```
## Answer in this chunk
```

### Problem 3.d (11 points)

- Based on the models we examined in the labs, fit an appropriate model with the aim to provide the most accurate prognosis you can for patients.
- Discuss and justify your decisions with several experiments and evidences.

```
## Answer in this chunk
```