BIOL63162 Scientific Programming, Computational Tools and Machine Learning: Statistical Learning Assessment

For this assessment you will be trying to build the best predictive model you can for two data sets, consisting of a *regression* and *classification* problem. You will need to write your analysis up as an Rmarkdown document that details *how* you built your model, giving all stages of data visualisation, exploration, model building and model testing. Ultimately, you need to make clear the logic of how you arrived at your final models and need some final indication of the model performance.

Regression problem (50%)

The regression problem is based on the `Boston` data set from the `ISLR2` package. To create it, use the following code:

```
RegressionData <- Boston[,c(5,7,8,12,13)]
```

This should create a new data set containing the following variables:

1. `nox` - nitrogen oxides concentration (parts per 10 million)
2. `age` - proportion of owner-occupied units built prior to 1940
3. `dis` - weighted mean of distances to five Boston employment centres
4. `lstat` - lower status (lower education or manual labour jobs) of the population (percent)
5. `medv` - median value of owner-occupied homes in $1000s

Your job is to try and predict the concentration of nitrogen oxides (`nox`) using `age`, `dis`, `lstat` and `medv` as input variables. You do not need to use all the variables, but your choice should be based on model comparisons and the aim to create the most accurate *and* parsimonious model you can. The context for this is consideration of safe levels of `nox` in the air in different areas of Boston. In particular, trying to determine whether levels of `nox` can be predicted given the demographic information from different areas. If it can then this has important implications for development of new housing and businesses in those areas.

Classification problem (50%)

The classification problem is based on the `Carseats` data set from the `ISLR2` package. To create it, use the following code:

```
ClassificationData <- Carseats[Carseats$ShelveLoc=="Good"|Carseats$ShelveLoc=="Bad",]

ClassificationData <- ClassificationData[,c(1,2,6,7,8)]
```

This should create a new data set containing the following variables:

1. `Sales` - unit sales (in thousands) at each location
2. `CompPrice` - price charged by competitor at each location
3. `Price` - price company charges for car seats at each site
4. `ShelveLoc` - a factor with levels *Bad* and *Good*, indicating the quality of the shelving location for the car seats at each site
    a. You may wish to re-coding this so that 0 = “bad” and 1 = “good”
5. `Age` - average age of the local population

Your job is to try and predict whether stores will place the car seats in a *good* or *bad* shelf location using what is known about current sales, competition, and local demographics. As with the regression problem, you do not need to use all the variables, but your choice should be based on model comparisons and the aim to create the most accurate *and* parsimonious model you can. The context for this is the car seat company wants to expand into new areas and wants to predict whether a new shop will put the car seats in a good or bad shelf location, given the different variables. They also want to know how the probability of putting the seat in a "good" location will change with changing sales, pricing and competitor pricing.

Formatting

You should submit a **single PDF document** that contains both the regression and classification analyses. The document should read like a report that you could hand to another data analyst for their assessment of your work. As such, it must be more than just the R output and should contain sentences and paragraphs explaining *what* you are doing, *why* you are doing it and drawing attention to the *important elements* of the R output. You should also make it clear when decisions are made based on your own interpretation or reading of the data/figures, compared to when you are using conventions to make decisions (e.g. what VIF value counts as high). Ideally, these conventions would also be referenced. There is **no word or page limit**, however, excessive detail or overly long outputs will be penalised so make sure you only include what is necessary and do not have redundancies or unnecessary details included. Do not include your name in the submission or submission title, only your student number, as these will be marked anonymously.

Deadline: **29th April 2022** by **16:00**