STA4026S 2024 – Statistics Honours Analytics
Assignment 1 – Applied Supervised Learning

Due Date: Monday 11 March 2024, 12:00 (noon)

Please read the following instructions carefully:

- You may complete this assignment either in pairs or individually.
- Answer all questions using R. You may use any packages and functions you like.
- You must submit a .pdf report to **Assignment 1 – Write-up** in Vula. Use the following naming convention:
  `STDNUM001_Analytics_A1.pdf` for individuals
  `STDNUM001_STDNUM002_Analytics_A1.pdf` for pairs.
- Upload all relevant R files (.R, .Rmd, .RData) to **Assignment 1 – Code & Predictions** in Vula. Use similar naming convention to the report. Be sure to use comments in your code! Also upload the .csv file with the predictions for Question 3 (see that question's instructions).
- Your report may be compiled using any software you like, although you are encouraged to use RMarkdown or LaTeX. Some marks will be awarded for communication and presentation, so make sure everything in your report is legible and neat.
- You are encouraged to include figures and tables, but only do so if they are relevant to your discussion. Also be sure to interpret them sufficiently; marks will be deducted for any output placed in the report without discussion.
- Do NOT paste raw R output verbatim, this will be penalised. If you want to include R output, typeset it properly or present it in a table.
- To help the reader easily assimilate the information, round values to a small number of decimal places (unless there is a good reason for expressing a more exact value).
- The report should NOT contain any code – use `echo = FALSE` for .Rmd files and do NOT attach any code as an appendix.
- Set seeds appropriately to ensure that all results are exactly reproducible.
- **The page limit is 15 pages**. Anything beyond the 15$^{\text{th}}$ page will be ignored.

Question 1 [35 marks]

Both the question setting and data are completely fictitious

Frikkie the Free State farmer acquired a new plot of land on which he planted some mealies at the start of the season. The plot had previously been divided into regions and crop rotation[1] was applied for decades, yielding vastly different soil quality in the different regions.

Having heard of the usefulness of UAV's (drones) in farming, Frikkie unwisely decided to plant the seeds by dropping them from a high altitude and letting the wind distribute them. He did the same when applying pesticides, causing some plants to have pests whilst others do not.

Before harvesting, each plant is inspected and the following characteristics measured:

- The exact longitude position of the plant, given in decimal degrees.
- The exact latitude position of the plant, given in decimal degrees.
- The number of mealies the plant yielded.
- Whether or not the plant had pests.
- The plant's height, in metres.
- The mealie quality – various measurements are combined into a single index with levels A – D. A represents the highest quality, D the lowest.

These data have been collated in the file `Q1dat.csv`. You are now tasked with modelling the **mealie quality**.

a) Divide the data into training and testing sets using an 80/20 split. Fit an overfitted classification tree on all features such that all the terminal nodes are homogenous. Briefly describe this tree, explain how you achieved leaf homogeneity, and report the testing misclassification rate. (5)

b) Use cross-validation (CV) to decide on a tree size (motivate your decision), and report this tree's test misclassification rate. Compare this with the training/CV error rates, as well as with the results from a). Interpret any possible discrepancies in context of the problem. (10)

c) Next, plot the quality against the location data (latitude and longitude). We see that the different regions yielded different quality mealies, noting that these regions are split by perpendicular lines. The task is to rotate the location features sub-space such that the decision boundaries are orthogonal to the axes. Apply 10-fold CV to determine the ideal amount of rotation, θ, searching over `thetas <- seq(0, pi/2, by = 0.01*pi)`.

[1] Crop rotation refers to the agricultural practice of alternating the use of pieces of arable land across seasons such that the soil can replenish nutrients and to avoid pest and weed immunity developing.

For each transformed feature set (rotated location + all other features), fit a classification tree pruned to 4 terminal nodes, then plot the CV misclassification rate as a function of θ. Report the the best rotation, fit a tree to this transformed feature space (with brief interpretation), and report its testing error with comparison to previous results. (20)

Question 2 [20 marks]

The Indian Premier League (IPL) is an annual T20 cricket tournament that generates the third-highest revenue per match of all sports tournaments (as of February 2024). The file `Q2dat.csv` contains data from 743 completed matches over several iterations of the IPL. In each match, there is a defending team (that bats in the first innings) and a chasing team (bats second). We are interested in predicting after the first innings whether the defending team will win the match, given by "Defending Result" (1 = win, 0 = loss).

The features include the names of the two teams playing; whether the defending team is playing home, away, or at a neutral venue; whether the match is played during the day or day and night; whether the defending team won or lost the toss (determining who gets to decide which team bats first); batting and bowling strength measurements for the chasing and defending teams respectively; the score being defended; the venue's median first and second innings scores; and the venue's altitude and dimensions.

a) Use logistic regression to model the target variable and apply elastic-net regularisation to this model, motivating for the choice of α and λ. Interpret the coefficients of the variable(s) left in the model and report the testing accuracy and F_1-score. (10)

b) Plot the selected model's ROC curve and give the area under the curve. Using out-of-sample data, determine the minimum decision rule threshold, τ, that will result in a recall value of at least 0.75. Indicate this point on the ROC curve. (10)

Question 3 [45 marks]

In this question we will examine data first presented and analysed by Tsanas et al. (2009), provided in `Q3dat.csv`. The goal of the exercise is to predict the value of the Unified Parkinson's Disease Rating Scale (UPDRS) measurement for patients with early-stage Parkinson's disease. This metric is often used to track the disease's progression.

The data were recorded via a telemonitoring device for remote symptom progression monitoring, allowing medical staff to obtain speech recordings captured in the patients' homes, thereby avoiding time-consuming physical examinations.

The following features were extracted from the original dataset:

Variable	**Description**
age	Subject age in years
sex	Biological sex – 0 = male; 1 = female
total_UPDRS	Clinician's total UPDRS score, linearly interpolated
Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP	Several measures of variation in fundamental frequency
Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA	Several measures of variation in amplitude
NHR, HNR	Two measures of ratio of noise to tonal components in the voice
RPDE	A nonlinear dynamical complexity measure
DFA	Signal fractal scaling exponent
PPE	A nonlinear measure of fundamental frequency variation

Note that step-by-step instructions are not provided; you are expected to follow appropriate procedures. You are welcome to include an exploratory analysis, although this should be brief. Make sure that you summarise the approach taken for each model and provide and discuss the relevant results.

a) Use the following models to predict the total UPDRS score: (30)

- A linear model with variable selection/regularisation applied.
- K-Nearest Neighbours (KNN), motivating for the choice of k.
- A random forest. Briefly explain the hyperparameter selection process and report variable importance.
- Any boosted tree model, again detailing your hyperparameter selection. Report variable importance and provide insights into the relationship between the most influential features and the predicted total UPDRS score.

b) Compare the models and select (with motivation) the one you would use on unseen data. Evaluate the chosen model's out-of-sample Root Mean Square Error (RMSE) relative to the observed distribution of total UPDRS scores. Also provide a visual illustration of the out-of-sample errors and comment on this model's weaknesses (if any). (10)

c) Finally, use the selected model to predict the total UPDRS scores for the observations contained in the file `Q3testing.csv`. Write your predictions to a .csv file as follows, to be uploaded to Vula:

- The file must only contain one column, namely the total UPDRS predictions.
- Do **NOT** include a column header.
- No blank cells to the left or above this column.
- The order of the predictions must correspond to the order of the observations in `Q3testing.csv`.
- The file name must be your student ID(s), all uppercase. E.g. STDNUM001.csv or STDNUM001_STDNUM002.csv.

You will only lose marks if you fail to adhere to these instructions, or if your predictions are severely wrong/nonsensical. The top 10 RMSE values calculated from this will be posted, although note that this is purely for fun and will have no bearing on your assessment marks. (5)

$\mathcal{END}$

TOTAL MARKS = 100

References

Tsanas, A., Little, M., McSharry, P. & Ramig, L. (2009), 'Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests', *Nature Precedings* pp. 1–10.