AY22 Sem2 BC2407 Computer Based Assessment Question Paper

Breast Cancer Diagnosis with Predictive Analytics

Introduction

A recent research paper by Manas and Veeky (2023)[1] claimed that Random Forest is the superior method for breast cancer diagnosis after comparing 5 Machine Learning models.

In this assignment, you will read this paper, conduct analysis on the same dataset, propose and answer additional research questions. How much do you agree with Manas and Veeky?

Part A: Review the Research Paper (30%)

1. Read the paper and state concisely the key contributions of this paper in bullet point format.
2. Explain where you will/might conduct this research work[2] differently from Manas and Veeky (2023). Conclude with a list of research questions in bullet point format that can be answered from the dataset.

Part B: Analytics and Insight (40%)

3. For each of your research question in (2), excluding those covered in (5) below, write down a data analytics plan (a sequence of **key steps** in bullet point format) that will produce results and a conclusion.
4. Execute each analytics plan in (3) using a software of your choice, present the key software outputs and your conclusions[3].
5. Do a 70-30 train-test split and produce a table, comparing the testset performance of Logistic Regression, Random Forest, and one other suitable technique learnt in BC2406 or BC2407. The table must include overall accuracy, false positive rate, false negative rate, precision and recall metrics. Describe the dataset used for train-test (especially if it is not the original dataset provided but modified in some way). State your key findings.

Part C: Conclusions (30%)

6. How much do you agree with Manas and Veeky (2023)[2]? Explain.
7. Suggest ways to improve/extend the research/analysis. Explain.

[1] Manas and Veeky (2023). Diagnosis of Breast Cancer Using Random Forests. International Conference on Machine Learning and Data Engineering, Procedia Computer Science 218 (2023) 429–437. [Paper and dataset are provided as part of this assignment].

[2] You are not required to consider models outside BC2406 and BC2407. Other models are out-of-scope for this assignment.

[3] Your code (e.g., Rscript, Python script) must be submitted as separate files but all answers should be complete in the CBA Submission word document file without having to read/execute any other files.

Appendix A

Sample Rcode for Sampling the Majority to Create Balanced Trainset (Down Sample the Majority)

Sources:

- Chew C.H. (2021) AI, Analytics and Data Science, Vol. 1, Chap 8 (CART), Cengage.
- RScript default CART.R from BC2406 Analytics I CART topic.

```r
# Random sample from majority class Default = No and combine with Default = Yes to form new trainset -----
majority <- trainset[Default == "No"]

minority <- trainset[Default == "Yes"]

# Randomly sample the row numbers to be in trainset. Same sample size as minority cases.
chosen <- sample(seq(1:nrow(majority)), size = nrow(minority))

# Subset the original trainset based on randomly chosen row numbers.
majority.chosen <- majority[chosen]

# Combine two data tables by appending the rows.
trainset.bal <- rbind(majority.chosen, minority)
summary(trainset.bal)
## Check trainset is balanced.
```