Computer Based Assessment

Heart Disease Diagnosis with Analytics

Introduction

A hospital would like to assess the usefulness of Analytics models in predicting heart disease and provided a set of data as recorded in AHD.csv.

A data dictionary is provided in Appendix A. AHD is the outcome variable to be predicted. A recent research paper on this dataset is included as a PDF file.

Part A: Data Exploration and Preparation (30%)

1. Conduct data exploration. Show and explain interesting findings. What are the data quality issues (if any)?
2. Create a summary table that shows important information. Explain the findings.

Part B: Analytics and Model Based Insight (40%)

3. Execute CART and another model of your choice taught in this module. Which model perform better? Explain.
4. Which of the two models would you recommend the Hospital to use? Explain.
5. Based on your chosen model above, explain the key findings to Hospital management.

Part C: Advanced Concepts (30%)

6. The cross validation error in CART is reported in the rpart package cp table. **If the outcome variable is continuous, this is fine. But if the outcome variable is categorical, an important information is missing.** What is the missing important information? Propose a way to obtain this information.
7. Read the research paper S.Chellammal, R. Sharmilax (2019). Comment on their approach. *Note: It is not required to know the techniques (MLP, SMO and NB) used in the research paper. Comment on their approach based on techniques and concepts learnt in any first course in Statistics module (e.g. AB1202 or equivalent) and this module.*

Age:	Age of person in years.
Sex:	Gender (1: male; 0: female).
ChestPain:	Chest Pain Type (1: typical angina; 2: atypical angina; 3: non-anginal pain; 4: asymptomatic).
RestBP:	Resting Blood Pressure.
Chol:	Serum Cholesterol in mg/dl.
Fbs:	Fasting Blood Sugar > 120 mg/dl. (1: true; 0: false).
RestECG:	Resting ElectroCardioGraphic Results (0: normal; 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2: showing probable or definite left ventricular hypertrophy by Estes' criteria).
MaxHR:	Maximum Heart Rate achieved.
ExAng:	Exercise Induced Angina (1: yes; 0: no).
Oldpeak:	ST Depression Induced by Exercise Relative to Rest.
Slope:	Slope of the Peak Exercise ST Segment (1: upsloping; 2: flat; 3: downsloping).
Ca:	Number of Major Vessels (0 to 3) colored by Flourosopy.
Thal:	Thalassemia (normal; fixed defect; reversable defect).
AHD:	Angiographic Heart Disease status (No: < 50% diameter narrowing; Yes: > 50% diameter narrowing in any major vessel).