Assignment 21: Comparison of Feature Space for Heart Disease Prediction

35 Points scaled to 20 Points

Introduction

In this assignment you will compare different predictor variable combinations for predicting the occurrence of hear disease. The data for this exercise ("heart.csv") were obtained from Kaggle (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction). Please visit the page to learn more about the data and the provided variables.

Objectives

- *Prepare data as input to machine learning algorithms*
- *Visualize relationships between variables*
- *Train and assess models using different feature spaces*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and output embedded. Files can be rendered to HTML webpages if your instructor requires this. Questions should be stated and answered within Markdown cells.*

Questions

This assignment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following task and answer the associated questions.

Task 1. The "HeartDisease" column is the dependent variable. It is currently coded such that 0 = No Heart Disease and 1 = Heart Disease. Make sure this variable is treated as a nominal variable in the model. The following predictor variables should be treated as continuous: "Age", "RestingBP", "Cholesterol", "MaxHR", and "Oldpeak". The following predictor variable should be treated as nominal data: "Sex", "ChestPainType", "RestingECG", "FastingBS", "ExerciseAngina", and "ST_Slope". Make sure all variables are treated appropriately in the model. You do not need to convert the nominal predictor variables to dummy variable since this is not required to use Random Forest. (5 Points)

Task 2. Subset the variables to create models using (1) all the available predictor variables, (2) just “Cholesterol” and “MaxHR”, and (3) “Age”, “RestingBP”, and “MaxHR”. (5 Points)

Task 3. Split the data into separate and non-overlapping training and testing sets using the “HeartDisease” variable for stratification. (5 Points)

Task 4. Train three Random Forest models using the three different feature spaces. You can use the default hyperparameter settings. (10 Points)

Task 5. Assess the models using the withheld testing data. Calculate Confusion Matrices and Overall Accuracy metrics. Using the precited probabilities, calculate the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC ROC). Use the results to discuss and compare the different feature spaces. (10 Points)