Final Project: Guide Lines

MATH 5530/4530-Statistical Computing

Spring, 2024

1 Overview

- In this assignment, you'll analyze the sani dataset provided on Canvas. Your task involves applying classification techniques covered in our course, like logistic regression, KNN, Naive Bayes, Support Vector Machine, and bagging or random forest, to determine if a person has CAD or not.
- You have the option to work either alone or in a team of up to two people for this final project. If you wish to form a group of three, you'll need prior approval.
- Every member of the team will submit an identical report, and each member will be awarded the same score.
- You need to write name of each team member in the submitted report.

2 Deliverables

- You are required to turn in two files: a written report in PDF format and an HTML document created using R Markdown.

3 Report guidelines

Please refer to the provided research paper on Canvas for the formatting guidelines. Additionally, you should explore other papers related to the sani dataset on Google Scholar by searching for terms like "sani data machine learning," etc.

1. **Title**: Give your report an informative title.
2. **Abstract**: Explain briefly what you want to do in this project. Write a paragraph.
3. **Introduction**: Read some papers and write one or two paragraph about your work relating with others.
4. **Data Description and Pre-processing**: Utilize descriptive statistics and visualizations such as plots and tables to analyze the dataset. Not all variables in the dataset may be relevant, so you'll select a maximum of 20 variables. Describe the method you used to choose these variables.

5. **Machine Learning Algorithms**: Describe the different models , methods you are using in this project.
6. **Model Comparison**: Explain about the methods about your model comparison. How are you compare your models.
7. **Results**: Write about your findings by tables and curves (ROC-curve).
8. **Conclusion**: Provide a summary of your findings by addressing the question: which technique or model demonstrates the most effectiveness for this dataset and what factors contribute to its success?