Assignment 3 – Classification

Due Date: 11PM Jun 30, 2023
Word limit: ~2,000 words (excluding spaces, tables, and references)

Submission

The assignment should be completed **INDIVIDUALLY**. In this assignment you are supposed to build on your previous work that you did in groups. In so doing, you are welcome to reuse everything you developed previously. However, you are also free to change whatever you feel can be improved. For example, you can use different set of features for your models.

Assignment Requirement

In this assignment, our goal is to i) reflect on the process we applied in previous assignments, ii) develop a series of classification models, and iii) compare results obtained from the selected models. This assignment will have the following parts:

1. **Introduction and Recap** (3 marks). In this part you are expected to briefly summarize (not more than 3 paragraphs) what has been done in previous assignments.
2. **Data exploration and Feature Selection** (5 marks). Explain data exploration and feature selection processes that you performed in previous two assignments.
    a. Choose ONE plot you presented in previous submissions and explain what this plot represents. What does it tell you about the variable(s) presented? Did you consider any other way to represent this variable(s)?
    b. Explain the feature selection process. How did you decide which features to include in your final dataset? Have you performed outlier detection (in case there was a need for that)? Did you scale your variables and why? Have you used any algorithm for feature selection (such as those we covered in practical work)?

    This section should be structured around the content you provided in Assignments 1 and 2. However, this section is not necessarily about copying text you previously had. It is more about your reflection on the process and presenting the understanding of the tasks you performed in previous two assignments. It is perfectly fine if you decide to use different set of features from what you agreed with your team in Assignment 2.
3. **Building Classification Models** using R (15 marks). Apply the classification algorithms to build at least four classifiers to analyze the heart disease data, using the algorithms we introduced in the course - decision tree, neural networks, SVM, random forests, Naïve Bayes, K-Nearest Neighbors or Ensemble methods. (10 marks)

    For each method, briefly report settings you used to run the algorithm (that is, parameter optimization) and the performance measures that you select (e.g., accuracy, misclassification error rate, precision, recall). As you choose your optimal model for each method, you might end up having several options for a given method, such as the case in Assignment 2 where we compared several decision trees models. Here, you should report only one model per given method. That is, if you decide to use SVM, decision tree, KNN and neural network, you should report four models – one per each method.

4. **Model Comparison and Conclusion** (7 marks). In this part, the focus should be on describing performance indicators and selecting the best performing model. For the model that yields the best performance, you should provide feature importance analysis and discuss the variables that are most predictive of the heart decease.

While there is a word limit assigned to this assignment, your submission may be longer (within reason).

The report should include a cover and table of content.

Marking Criteria

High Distinction

In meeting this level, you will address all the parts of the assignment, demonstrating clear understanding of the topics covered in the course, also taking into the account feedback received on the previous submission. Data inspection will be supported with relevant, quality, figures and accompanied explanation of observed trends. All the figures and tables will be labelled. Clear description of the selected classification methods would be provided, including parameter optimization. Finally, you will identify relevant metrics for model comparison and provide the analysis of feature importance for the best performing model. This level requires clear and coherent writing, with the concise narrative. Overall, in meeting this level, you will demonstrate a comprehensive knowledge of the concepts through your descriptions, explanations, and discussions of the content.

Distinction

In meeting this level, you will address all the parts of the assignment, including the feedback you received on the previous assignment. Data inspection includes basic plots (e.g., histograms, correlation matrix) that are supported with relevant, quality, figures and accompanied explanation of observed trends. All the figures and tables will be labelled. Clear description of the selected classification methods would be provided, including parameter optimization. Basic metrics (e.g., accuracy or recall) for model comparison are being used for model comparison and basic analysis of feature importance for the best performing model is provided. Writing is clear and the narrative is coherent across the report. Overall, in meeting this level you will demonstrate a well-considered knowledge of the concepts through your descriptions and explanations.

Credit

In meeting this level, you will address at least two parts of the assignment, where initial inspection would be a mandatory part, addressing parts of the feedback you received on your previous assessment. Data inspection includes basic plots (e.g., histograms) that are supported with relevant figures and accompanied explanation of observed trends. Clear description of the selected classification methods would be provided, however most of the models are applied with default configuration. Overall, in meeting this level, you will demonstrate a sound knowledge of concepts through your descriptions and explanations.

Pass

In meeting this level, you will address at least two parts of the assignment, where building various models is mandatory. In doing so, you will demonstrate knowledge of the concepts through your descriptions. Clear description of the selected classification methods would be provided, however all or most of the models are applied using default configuration.

Academic integrity

You are expected to reference and cite all resources mentioned using a selected referencing convention (e.g., UniSA Harvard, or APA).

Extensions

Extensions for assignments are available under the following conditions

- permanent or temporary disability, or
- compassionate grounds

In all cases, documentary evidence (e.g. medical certificate, road accident report, obituary) must be presented to the Course Coordinator. A medical certificate produced on or after the due date will not be accepted unless you are hospitalized.

If you apply for extension within 24 hours before the deadline, you must see the course coordinator in person unless you are in an emergency like being admitted in a hospital.

Late Penalties

Unless you have an extension, late submission will incur a penalty of 30% deduction per day (or part of it) of lateness.