Assignment 1 – Data Exploration and Feature Extraction

Due Date: 11PM Sun 09 Apr 2023
Word limit: ~1,500 words (excluding spaces, tables, and references)

Submission

The assignment should be completed individually and submitted through learnonline. Please make sure that you complete all the parts of the assignment.

Assignment Requirement

In this assignment, we will be focusing on the first two course objectives: i) apply standard processes to prepare large data sets for data exploration and ii) perform data exploration on large data sets using visualisation, statistical techniques, and data mining techniques to identify relationships and opportunities. This further means that in this assignment you will explore a dataset and prepare a set of features that can be used for a classification task and complete a report where you will discuss observed trends in your data. This dataset will be used in Assignments 2 and 3 to build and evaluate predictive models.

We will be working on a subset of a real-world dataset collected by 1) Hungarian Institute of Cardiology, Budapest, 2) University Hospital, Zurich, 3) University Hospital, Basel, and 4) V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The dataset you will be using in this assignment, has been used in several Kaggle challenges and numerous studies.

In this assignment, our goal is to **prepare a dataset that can be used to predict heart disease (i.e., `target`)**. This assignment will have three parts:

1. **Introduction** (5 marks). In this part you will briefly discuss the goal of the assignment and what is the role of data mining in addressing this and similar challenges (e.g., predicting disease spread). You should budget approximately **250-300 words** for this section.

2. **Related Work** (10 marks). This part is important as it will inform your feature extraction and help you to understand the problem better. As we discussed, predictive models depend on features you extract. To extract potentially relevant features, you need to know the domain. Here, you will **review at least 3 papers** that address a problem of predicting heart or other diseases. Please try to find another study that uses this same dataset so it would be easier to compare your results with previous work. This section should provide a brief review of the selected papers (approximately **500 words**) including:

    a. What kind of problem they were trying to solve (e.g., predicting heart disease or a tumor)?

    b. What kind of features they were using (e.g., a number of physical activities per week)?

    c. What are the predictive models used in those studies (e.g., SVM or neural networks)?

    d. How those studies informed **your research**? That is, you should finish this part with a short paragraph (or a sentence) that says something along those lines “Based on the previous research, we decided to focus on the following set of features…” (please refer to Part 3).

3. **Data exploration** (15 marks). Using R and RStudio for descriptive statistics and visualisations, in **approximately 700-800 words**, in this part you will:

    a. Explain the features you decided to extract or use from the obtained dataset. Please note that most (or all) of the variables in your dataset can be used directly as features. Some other variables can be used to derive additional features. Please read and think very carefully about the features and decide whether you use them directly, construct new features from them, or not to use them at all.

b. Once you have selected your features, provide descriptive statistics (using tables and figures) of your feature set. This **may include** value distributions, skewness, histograms, bar plots, box plots, and other details that you find relevant. In doing so, you are also expected to discuss possible impact of the observed properties on the classification problem.

c. Along with the univariate analysis (such as value distribution), you are also expected to investigate the correlation between variables that you find important (this does not have to include all the features). A useful tool here would be correlation plots or correlograms, bar plots of contingency tables or mosaic plots.

d. Prepare a CSV file with a following structure:

`id`, `feature_1`, `feature_2`, … `feature_n`, `target`

The CSV file **should be included in your submission** and is supposed to include only those features that you selected for classification. If you decide to use all the features from the original file, that is perfectly fine.

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.

Dataset description

Details for each of the columns included in the dataset are provided in Table 1.

Table 1. Structural data description

Variable	Description	Type
id	A unique ID that identifies a participant in the study	Numerical
age	Age in years	Numerical
sex	Male and Female were recorded	Categorical
cp	Chest Pain type: typical angina; atypical angina; non-anginal pain; and asymptomatic	Categorical
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Numerical
chol	Serum Cholestoral in mg/dl	Numerical
fbs	Fasting blood sugar > 120 mg/dl (True or False)	Boolean
restecg	Resting electrocardiographic results: normal; having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) or showing probable or definite left ventricular hypertrophy by Estes' criteria	Categorical
thalach	Maximum heart rate achieved	Numerical
exang	Exercise induced angina (True/False)	Boolean
oldpeak	ST depression induced by exercise relative to rest	Numerical
slope	The slope of the peak exercise ST segment: upsloping; flat; downsloping	Categorical
major_vessels	Number of major vessels (0-3) colored by flourosopy	Numerical
restwm	Rest wall motion abnormality: none; mild or moderate; moderate or severe; akinesis or dyskmem	Categorical
target	Heart disease diagnosed (disease/no disease)	Categorical

Marking Criteria

High Distinction

In meeting this level, you will address all three parts of the assignment, demonstrating clear understanding of the topics covered in the course. Data inspection will be supported with relevant, quality, figures and accompanied explanation of the observed trends. All the figures and tables will be labelled. Critical reflection on the existing work, including how this informed the extraction of feature you included in the analysis, would be provided. Clear explanations and summary statistics for each feature would be provided. 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. A CSV file with the final feature set will be submitted.

Distinction

In meeting this level, you will address all three parts of the assignment, demonstrating clear understanding of the topics covered in the course. Data inspection will be supported with relevant, quality, figures and accompanied explanation of the observed trends. All the figures and tables will be labelled. Summary of the previous work would be provided, including the conclusion how this previous work informed the features you extracted. Clear explanations and summary statistics for each feature would be provided. Overall, in meeting this level you will demonstrate a well-considered knowledge of the concepts through your descriptions and explanations. Comments from your previous report will be addressed. A CSV file with the final feature set will be submitted.

Credit

In meeting this level, you will address all three parts of the assignment. Data inspection includes basic plots (e.g., histograms) that are supported with relevant figures and accompanied explanation of observed trends. All the figures and tables will be labelled. Some literature has been referenced in Part 2. A list of features used in this work would be provided. Overall, in meeting this level, you will demonstrate a sound knowledge of concepts through your descriptions and explanations. A CSV file with the final feature set will be submitted.

Pass

In meeting this level, you will address at least two parts of the assignment. In doing so, you will demonstrate knowledge of the concepts through your descriptions and attempt to design meaningful set of features.

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.