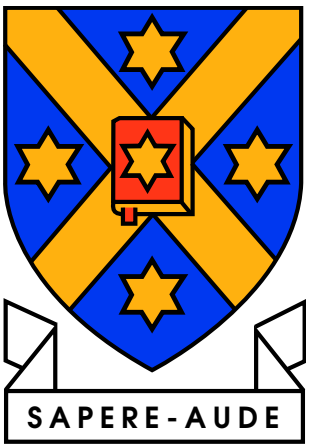
INFO411 ASSIGNMENT 2 REPORT



Josh Signal

A report submitted in partial fulfilment of the paper INFO411 as part of the requirements of the degree of Master of Applied Science in Artificial Intelligence

October, 2023

EDA and Attribute Selection

First I explored the clean DS1 dataset (See Figure 1, 2 and 3). I then used this to compare with statistical properties of the 4 location datasets in DS2 and made inferences to make decisions pertaining to how reasonable it is to use or cut parts ot the messier datasets. A detailed version of analysis can be found in the "JoshSignal-EDA.jl" pluto notebook.

14 Attribute meanings (according to the UCI Machine Learning Repository):

- sex (1 = male, 0 = female)
- cp = chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
- trestbps = resting blood pressure (in mm Hg on admission to the hospital)
- chol = chol: serum cholestoral in mg/dl
- fbs = (fasting blood sugar ¿ 120 mg/dl) (1 = true; 0 = false)
- restecg = resting electrocardiographic results
- thalach = maximum heart rate achieved
- exang = 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 (0-3) colored by flourosopy
- thal = categories of complication from mid to severe (3 = normal; 6 = fixed defect; 7 = reversable defect)
- num = diagnosis of heart disease (for DS1 0 = absence, 1 = presence. For DS2 0 = absence, 4 = severe)

Key findings:

- ca and thal have high correlation with the target (See Figure 2).
- chol has a small correlation with target.
- Switzerland and Va have over 25% of chol entries equal to zero (missing entries).
- ca and thal are largely absent from Switzerland, Va and Hungary. If we want to use DS2, we almost necessarily must not use ca and thal. This greatly reduces model performance later.

- 50% of Switzerland rows are missing fbs. Since Switzerland is small, we can impute this.
- About a 1/3rd of DS2 is missing the slope attribute. Slope is moderately correlated with the target. Instead of dropping it and further reducing performance, we try to impute it.

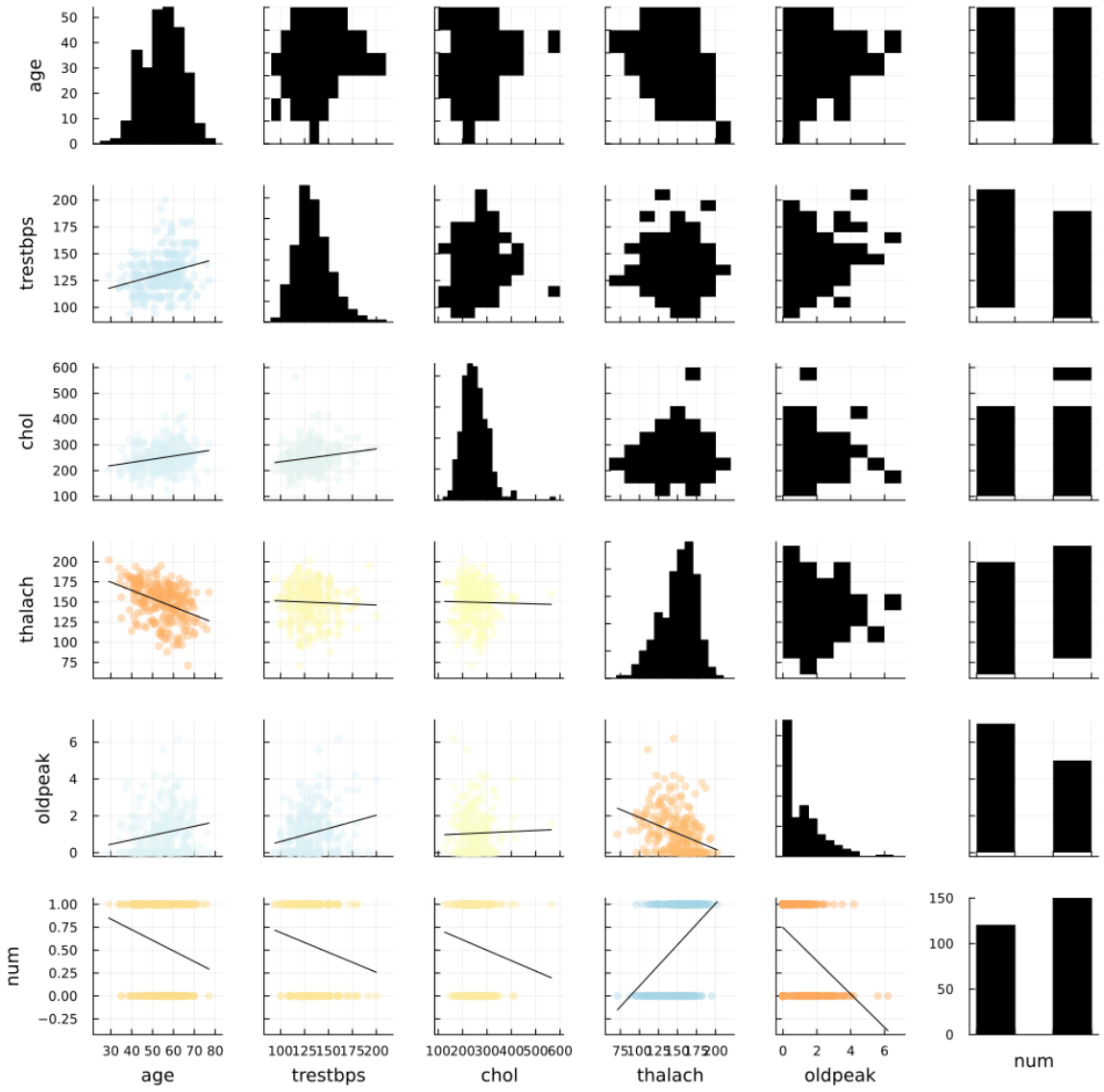


Figure 1: Corrplot of DS1.

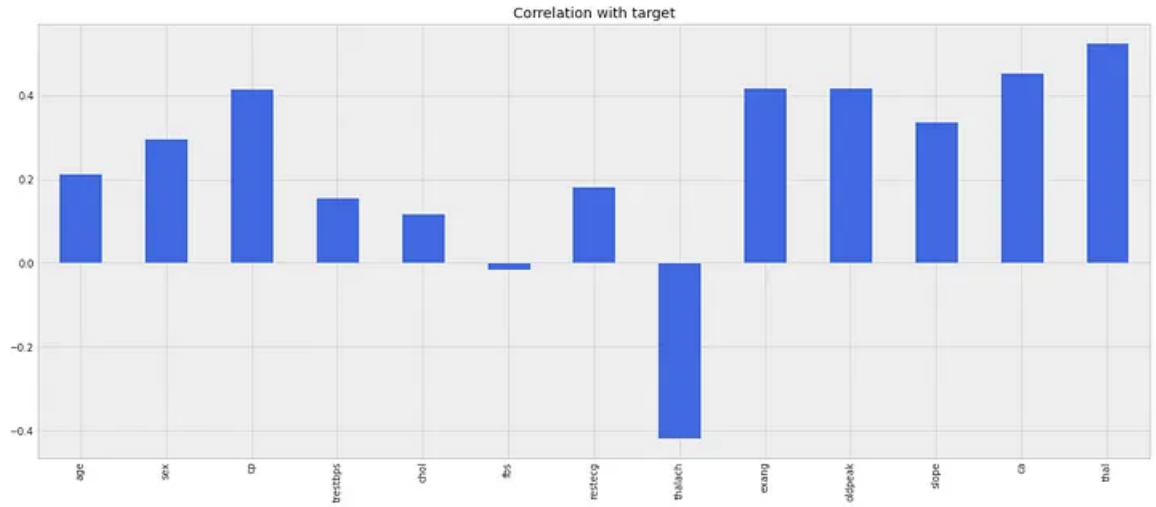


Figure 2: Correlations of DS1 attributes with target.

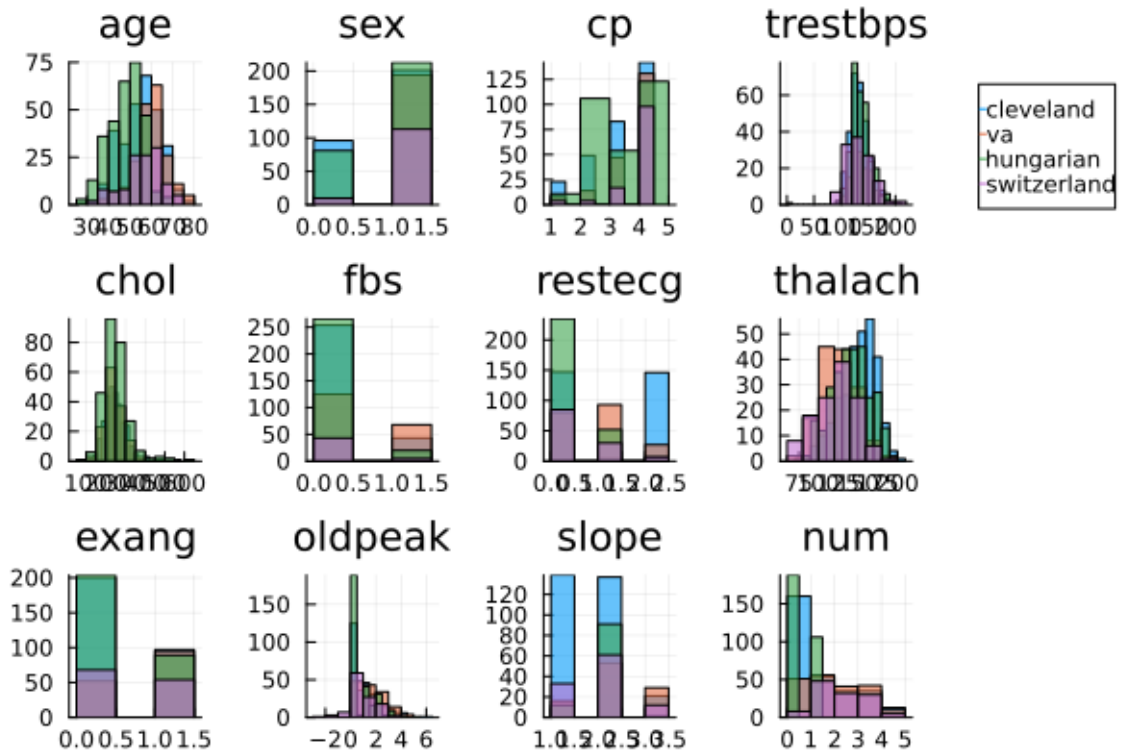


Figure 3: DS2 location distribution comparisons.

Imputation

To impute the dataset, I used a multiple imputation method. I implemented a Random Forest imputator (using the Impute.jl package) on the missing values 10 times and took the mean.

Modelling

I conducted 3 main experiments (including some tuning):

- fit decision tree and ridge regressors on DS2
- Compared decision tree regression classifiers trained on DS1 and DS2 (binary target)
- Fit decision tree regressjon classifier on DS1 and tested generalization on DS2

Main findings:

- fit decision tree and ridge regressors on DS2: RMS 0.932
- Compared decision tree classifiers trained on DS1 and DS2 (binary target): DS1 validation RMS = 0.395, DS2 validation RMS = 0.380239 (slight improvement on DS1)
- Fit decision tree classifier on DS1 and tested generalization on DS2: DS2 test RMS = 0.7

DS2 performs slightly better than DS1 alone (not statistically significant but intuitive as DS2 3 times the data points) but DS1 does not generalize to DS2.