Final Project Memo

Mason Ma

10/2/2022

Heart Failure Prediction

An overview of your dataset

1. About this Dataset

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Number of observations: 299

Number of predictors: 13

We will be working with numeric, boolean predictors (anaemia, diabetes, high_blood_pressure, sex, DEATH_EVENT), and binary predictors (sex). The outcome will be a boolean variable where 1 denotes death and 0 otherwise.

The data set is listed as follow: (first 6 rows)

```
original_data<-read.csv("/Users/mason/Desktop/heart_failure_clinical_records_dataset.csv")
head(original_data)
```

```
##   age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1  75       0                     582        0                20
## 2  55       0                    7861        0                38
## 3  65       0                     146        0                20
## 4  50       1                     111        0                20
## 5  65       1                     160        1                20
## 6  90       1                      47        0                40
##   high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1                   1    265000             1.9          130   1       0    4
## 2                   0    263358             1.1          136   1       0    6
## 3                   0    162000             1.3          129   1       1    7
## 4                   0    210000             1.9          137   1       0    7
## 5                   0    327000             2.7          116   0       0    8
## 6                   1    204000             2.1          132   1       1    8
##   DEATH_EVENT
## 1           1
```

```
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

2. Link and Source:

https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

3. Explanations of predictors

age : age of each person

anaemia: decrease of red blood cells or hemoglobin (boolean)

creatinine_phosphokinase: Level of the CPK enzyme in the blood (mcg/L)

diabetes: If the patient has diabetes (boolean)

ejection_fraction: Percentage of blood leaving the heart at each contraction (percentage)

high_blood_pressure: If the patient has hypertension (boolean)

platelets: Platelets in the blood (kiloplatelets/mL)

serum_creatinine: Level of serum creatinine in the blood (mg/dL)

serum_sodium: Level of serum sodium in the blood (mEq/L)

sex: Woman or man (binary)

3. Citation

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

An overview of your research question(s)

——What variable(s) are you interested in predicting? What question(s) are you interested in answering?

We are interested in the outcome variable–DEATH_EVENT. We aim to predict how much these predictors contribute to the much likely a person will be dead due to the values of predictors.

——Name your response/outcome variable(s) and briefly describe it/them.

DEATH_EVENT: a boolean variable where 1 denotes death, 0 otherwise.

——Will these questions be best answered with a classification or regression approach?

Classification, since the task is to predict a discrete class label.

——Which predictors do you think will be especially useful?

high_blood_pressure and diabetes.

——Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

It will be a combination of inferential and descriptive model. We aim to find the model to best visually emphasize a trend in data, discover what features are significant, and state relationship between outcome & predictor(s).

Your proposed project timeline

—-When do you plan on having your data set loaded, beginning your exploratory data analysis, etc?

The data is obtained from Kaggle. I will do the data cleaning and modeling (Logistic Regression) next week.