# Project_report

Baldeep Dhada,Han Chen,Somya Nagar

2024-03-21

## Introduction and Motivation

---

## Data description

1. Raw data

This dataset includes the medical histories of 299 patients diagnosed with heart failure, with the target variable being ‘death_event’. As the *table 1* shows each patient’s profile consists of 13 clinical attributes, 5 categorical variables and 7 numerical variables.

| variables | Type | description | values |
|---|---|---|---|
| age | Integer | age of the patient (years) | [40,95] |
| anaemia | Binary | decrease of red blood cells or hemoglobin | 0,1 |
| creatinine_phosphokinase | Integer | level of the CPK enzyme in the blood(mcg/L) | [23,7861] |
| diabetes | Binary | if the patient has diabetes | 0,1 |
| ejection_fraction | Integer | percentage of blood leaving the heart at each contraction (%) | [14,80] |
| high_blood_pressure | Binary | if the patient has hypertension | 0,1 |
| platelets | Continuous | platelets in the blood(kiloplatelets/mL) | [25100,850000] |
| serum_creatinine | Continuous | level of serum creatinine in the blood (mg/dL) | [0.5,9.4] |
| serum_sodium | Integer | level of serum sodium in the blood (mEq/L) | [113,148] |
| sex | Binary | woman or man | 0,1 |
| smoking | Binary | if the patient smokes or not | 0,1 |
| time | Integer | follow-up period (days) | [4,285] |
| death_event | Binary | if the patient died during the follow-up period | 0,1 |

Table 1. A table provides a description of the variables included in the dataset. The variables are categorized by type, including Integer, Binary, and Continuous, along with a brief description of each variable’s meaning. The ‘values’ column indicates the range or possible values for each variable.

1.1 Target variable

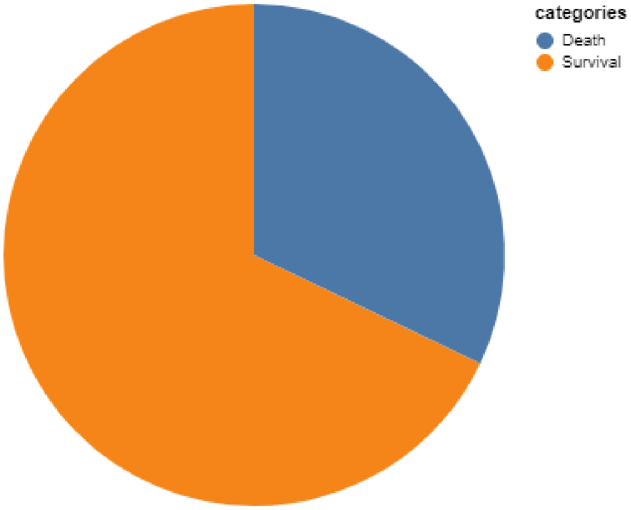
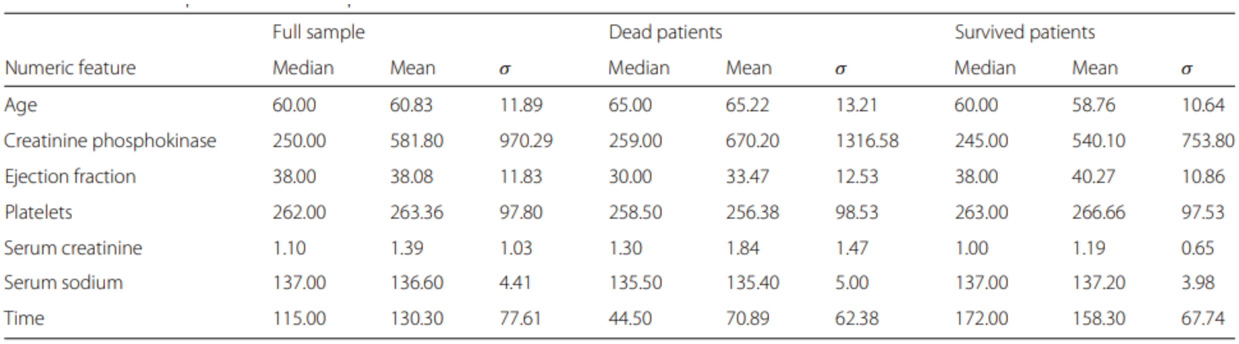


Figure 1: A pie chart depicting the distribution of survival and mortality rates, crucial target variables in our analysis. There are 203 survival observations and only 96 dead observations signifying imbalance in the data.

As Figure 1 shows below, our targeted variables `death_event` is unbalanced with 203 survival observations vs 96 dead observations.

1.2 Numerical variables

As the Figure 2 shows, these variables exhibit different scales, indicating the need for standardization. Based on descriptive statistics, the variables do not appear to have significant differences between the deceased and surviving groups, except for `time`.



| | Full sample | | | Dead patients | | | Survived patients | | |
|---|---|---|---|---|---|---|---|---|---|
| Numeric feature | Median | Mean | $\sigma$ | Median | Mean | $\sigma$ | Median | Mean | $\sigma$ |
| Age | 60.00 | 60.83 | 11.89 | 65.00 | 65.22 | 13.21 | 60.00 | 58.76 | 10.64 |
| Creatinine phosphokinase | 250.00 | 581.80 | 970.29 | 259.00 | 670.20 | 1316.58 | 245.00 | 540.10 | 753.80 |
| Ejection fraction | 38.00 | 38.08 | 11.83 | 30.00 | 33.47 | 12.53 | 38.00 | 40.27 | 10.86 |
| Platelets | 262.00 | 263.36 | 97.80 | 258.50 | 256.38 | 98.53 | 263.00 | 266.66 | 97.53 |
| Serum creatinine | 1.10 | 1.39 | 1.03 | 1.30 | 1.84 | 1.47 | 1.00 | 1.19 | 0.65 |
| Serum sodium | 137.00 | 136.60 | 4.41 | 135.50 | 135.40 | 5.00 | 137.00 | 137.20 | 3.98 |
| Time | 115.00 | 130.30 | 77.61 | 44.50 | 70.89 | 62.38 | 172.00 | 158.30 | 67.74 |

Figure 2: Statistical quantitative description of the numeric features that shows the mean, median and standard deviation for each of the quantitative variables. The mean, median and standard deviation values are further grouped into ‘Full sample’, ‘Dead patients’ and ‘Survived Patients’ to give a description based on the target variable.

1.3 Categorical variables

We visualize 5 categorical variables against the target variable “death_event.” As Figure 3 shows below, the number of males is larger than females, and individuals with a history of ‘anaemia,’ ‘high blood pressure,’ ‘diabetes,’ and ‘smoking’ are more prevalent than those without such histories. These variables demonstrate bias. Therefore, we need to perform oversampling to preprocess our data.

2. Data Preprocessing

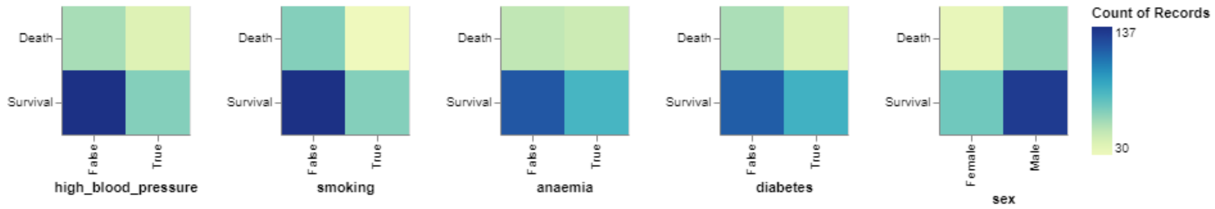


Figure 3: Heatmap of Categorical Variables (high_blood_pressure, smoking, anemia, diabetes, sex) that shows the trend of counts of records with the target variable death_event. Count of records can be seen for 'Death' and 'Survival' for each of the variables.

### 2.1 Standardization

Based on Table 1 and Figure 2, the numerical variables have vastly different scales, so we need to scale our data to prevent some features with larger scales from dominating the learning process. Also, we can observe the presence of outliers in the data. Therefore, choosing Z-score Normalization would be a suitable method for standardization. Additionally, from the distribution of continuous variables, it is evident that the distribution of data varies with different values of the target variable for some continuous variables. Z-score Normalization can effectively preserve the information about the original variable distributions.

Mathematically, for a feature $X$, the Z-score normalization formula is:

$$Z = \frac{(X - \mu)}{\sigma}$$

Where:

- $Z$ is the standardized value (Z-score).
- $X$ is the original data point.
- $\mu$ is the mean of the feature.
- $\sigma$ is the standard deviation of the feature.

The resulting Z-scores have a mean of 0 and a standard deviation of 1, making it easier to compare and interpret different features with varying scales.

### 2.2 Oversampling

As Figure 1 shows, there is a huge imbalance in the response variable `death_event` with 203 survival observations vs 96 dead observations. Dealing with imbalanced datasets poses a challenge because the majority of machine learning methods tend to overlook the minority class, resulting in poor prediction. Hence, we will attempt to balance using the SMOTE (Synthetic Minority Oversampling Technique) algorithm.

SMOTE works by synthesizing new samples of the minority class to balance the disparate sample sizes between different classes. The core idea of the SMOTE algorithm is to generate synthetic samples by interpolating between existing minority class samples, thus increasing the quantity of minority class samples and creating a more balanced distribution between the minority and majority classes.

As shown in Figure 4, even after oversampling, each category of response remains equal (203 occurrences each), yet the distribution of categorical variables remains uneven, except for the variable 'anaemia'

## Hypothesis

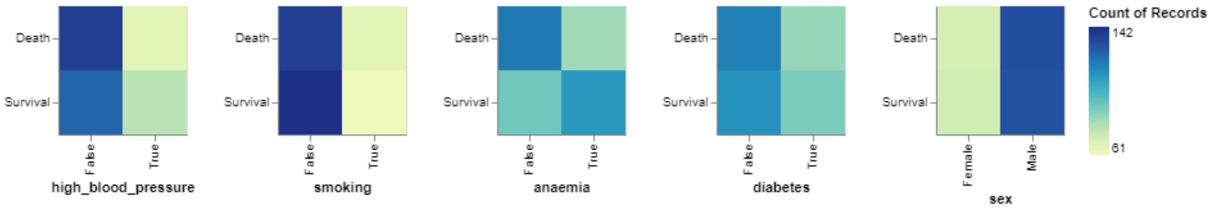


Figure 4: Oversampling Heatmap of Categorical Variables (high_blood_pressure, smoking, anemia, diabetes, sex) depicting the counts of records with the target variable death_event. Counts of records can be observed for 'Death' and 'Survival' for each of the variables.

## Feature selection

As shown in Figure 5, there exists some correlation between variables. We will conduct feature selection to identify the most important variables for model prediction.

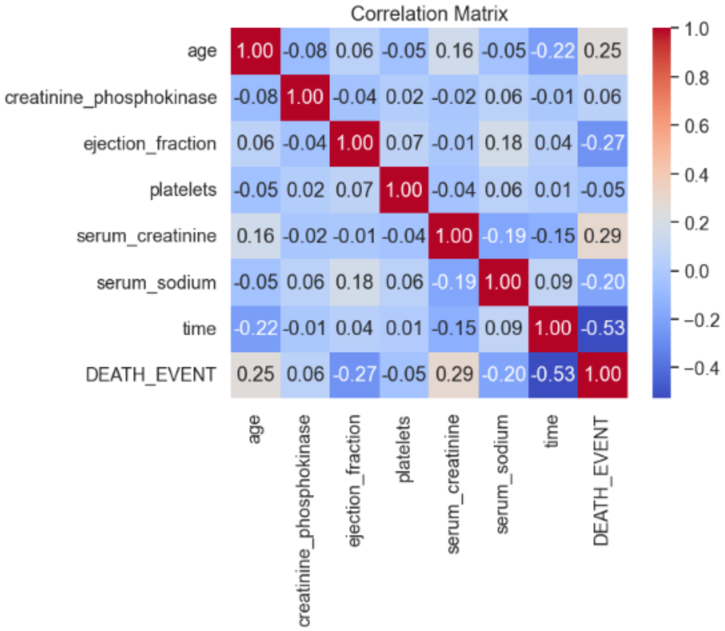


Figure 5: A correlation Matrix that shows the collinearity between the predictors and the reponse variable. A value close to -1 or 1 is indicative of high collinearity between predictors. A value close to 0 is indicative of low collinearity between predictors. The more saturated a box is relative to its color the higher the collinearity. The less saturated a box is relative to its color the lower the collinearity.

1. Z-test

A Z-test is a statistical method used to assess whether there's a significant difference between the means of the populations under consideration. By formulating the null hypothesis (H0) suggests that the population means are equal. Next, the test statistic (Z) is calculated using the formula:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here, $\bar{X}_1$ and $\bar{X}_2$ represent the sample means of the two populations, while $\sigma_1$ and $\sigma_2$ denote the known population standard deviations. $n_1$ and $n_2$ are the sample sizes for the respective populations.

As we can see on the table 3, variable `time` has significant high z-score and `ejection_fraction`, `age` and `serum_creatinine` are the following important variables

# Model

Extreme Gradient Boosting (XGBoost) is an ensemble learning method that employs gradient boosting to train decision trees sequentially. The optimal parameters obtained through grid search are ('colsample_bytree': 0.6, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 150, 'subsample': 0.8)

Generalized Additive Models (GAMs) are a class of generalized linear models that allow for flexible modeling of non-linear relationships. The model utilized B-spline basis functions with the number of degrees of freedom (df) and degrees of the spline set to [3, 3, 4, 4, 4, 3, 4] and [2, 2, 3, 3, 3, 2, 3], respectively, allowing for the capture of non-linear relationships between the predictors ['age', 'creatinine_phosphokinase', 'ejection_fraction', 'platelets', 'serum_creatinine', 'serum_sodium', 'time'] and the response variable.

# Result