# Project 1

Machine Learning and Data Mining [02450]

**Mia Due Paarup - s213304**

**Joaquim Siqueira Lucena - s232535**

**Gokul Desu - s242580**

| Section | Mia | Joaquim | Gokul |
|---|---|---|---|
| Data Description | 30% | 30% | 40% |
| Data Attributes | 40% | 30% | 30% |
| Data Visualization | 30% | 40% | 30% |
| Discussion | 40% | 30% | 30% |
| Exam problems | Equally | Equally | Equally |

Table 1: Contribution (%) to each section of the report.

**DTU** Technical University of Denmark

# 1 Data Description

Our dataset contains 299 instances of patients who had heart failure. The data was collected during their follow-up period, where each patient has 13 clinical features(12 features and 1 target classification). The overall problem of interest is if we can predict the survival or not of a patient, based on their clinical data. The dataset was obtained on UC Irvine's Machine Learning Repository, together with one introductory paper[1].

The research at [1] explains that the survival of patients can be predicted by using ejection value and serum creatinine alone. Moreover, using only these two features can lead to more accurate predictions than using the whole dataset. They employed ten different Machine Learning methods, including linear regression, neural networks, and a Naive Bayes classifier. For the whole dataset, the best approach was using random forests, achieving an accuracy of 0.73. However, when analyzing only serum creatinine, ejection factor, and follow-up time, they achieved 0.83 using logistic regression.

Our goal for this project is to hopefully find other useful correlations that could be used to predict patients' survival(or death). Finally, we hope that we can use this data to create our own model with a decent accuracy

# 2 Data attributes

The dataset contains 13 clinical attributes which are the following:

**Age** represents the patients age in Years. It is discrete because it is recorded as integers in this dataset. It is also ratio since it has a true zero point, and differences between values are meaningful. **Anaemia** signals if the patient has anemia or not and is discrete and nominal, as it falls into two categories: either yes or no.

**Creatinine Phosphokinase (CPK)** represents the level of the CPK enzyme in the blood, measured in mcg/L. It is discrete and ratio because it is measured in integers and has a true zero point. **Diabetes** show if a patient has diabetes or not. It is a discrete and nominal attribute, as it is binary. **Ejection Fraction** represents the percentage of the blood leaving the heart at each contraction, measured in percentages. It is discrete and ratio, as it is measured in integers and has a true zero point.

**High Blood Pressure** signals if the patient has high blood pressure or not. It is discrete and nominal (the patient either has high blood pressure or not). **Platelets** show the amount of platelets in the blood, measured in kiloplatelets/mL. It is continuous and ratio. **Serum Creatinine** measures the mg/dL of serum creatinine in the blood. This attribute is continuous and ratio, as it can be measured in fractions and has a true zero point. Similarly, **Serum Sodium** measures the quantity of serum sodium in the blood, using mEq/L. However it is discrete and ratio because it is measured in integers and has a true zero point.

The **Sex** attribute is discrete and nominal, as patients are categorized as either male or female. **Smoking** is also discrete and nominal, as patients are either smoking or not. **Time** is the follow-up period and is measured in whole days, so it is discrete and ratio. The **Death Event** is discrete and nominal, as it is binary (the patient either died or survived during the follow-up period).

There are no missing values or corrupted data in this dataset. See table 2 for basic summary statistics of the attributes.

| Attribute | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|
| Age | 60.83 | 11.89 | 40 | 95 |
| Anaemia | 0.43 | 0.49 | 0 | 1 |
| Creatine Phosphokinase | 581.84 | 970.30 | 23 | 7861 |
| Diabetes | 0.42 | 0.49 | 0 | 1 |
| Ejection fraction | 38.08 | 11.83 | 14 | 80 |
| High blood pressure | 0.35 | 0.48 | 0 | 1 |
| Platelets | 263358.03 | 97804.24 | 25100 | 850000 |
| Serum creatinine | 1.39 | 1.03 | 0.5 | 9.4 |
| Serum Sodium | 136.63 | 4.41 | 113 | 148 |
| Sex | 0.65 | 0.48 | 0 | 1 |
| Smoking | 0.32 | 0.47 | 0 | 1 |
| Time | 130.26 | 77.61 | 4 | 285 |
| Death event | 0.32 | 0.47 | 0 | 1 |

Table 2: Basic summary statistics of the 13 attributes including mean, standard deviation and the range.

# 3 PCA and Data Visualization

## 3.1 Data Visualization

Before performing the PCA, we prepared some plots to understand and visualize our data better. After standardizing it, we can say that many features of our data are very different from a normal distribution, except platelets in the blood and serum sodium, which closely resemble a normal distribution. Of course, several of our attributes will not be normally distributed since they are reported as binary. This data is displayed in Fig. 1.

A boxplot can be seen in Fig. 2, and by analyzing it, our data seems to have some outliers. However, because of our lack of knowledge of the medical field and the relatively small amount of observations, we cannot conclude whether they are true outliers.

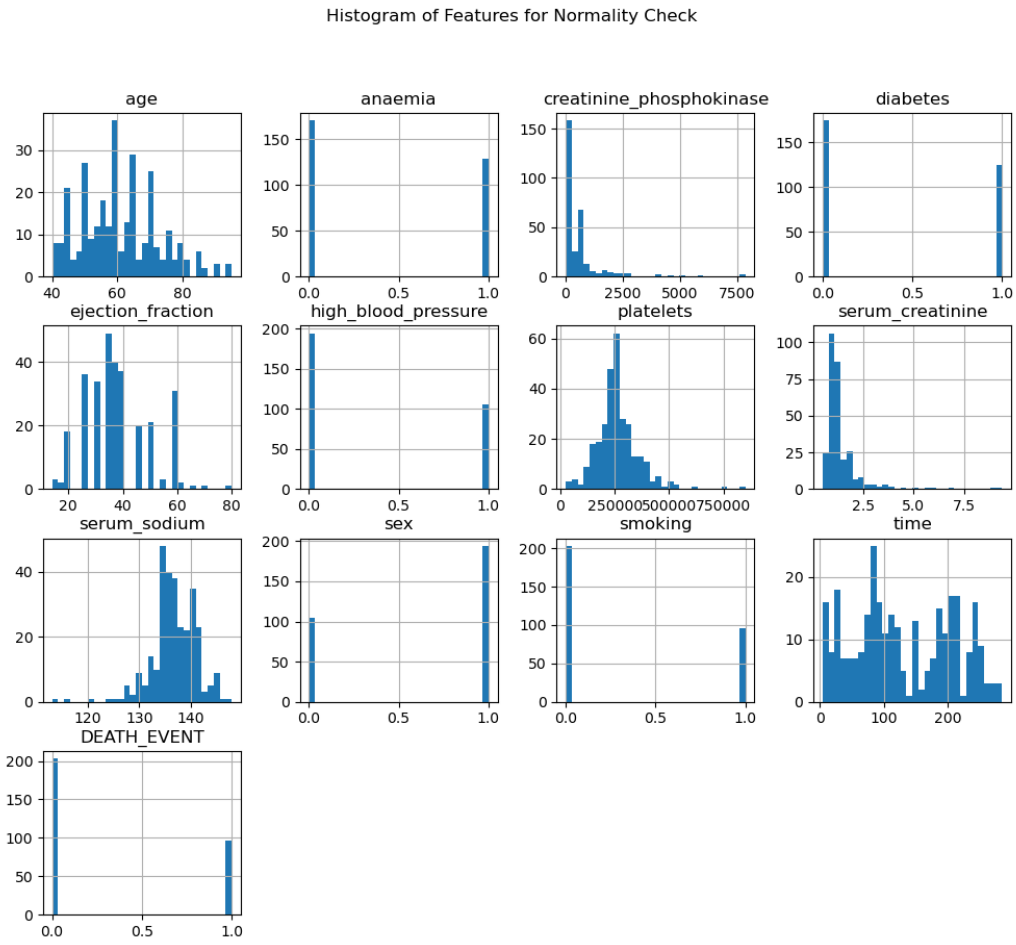


Figure 1: Histograms of each feature in our dataset, including target.

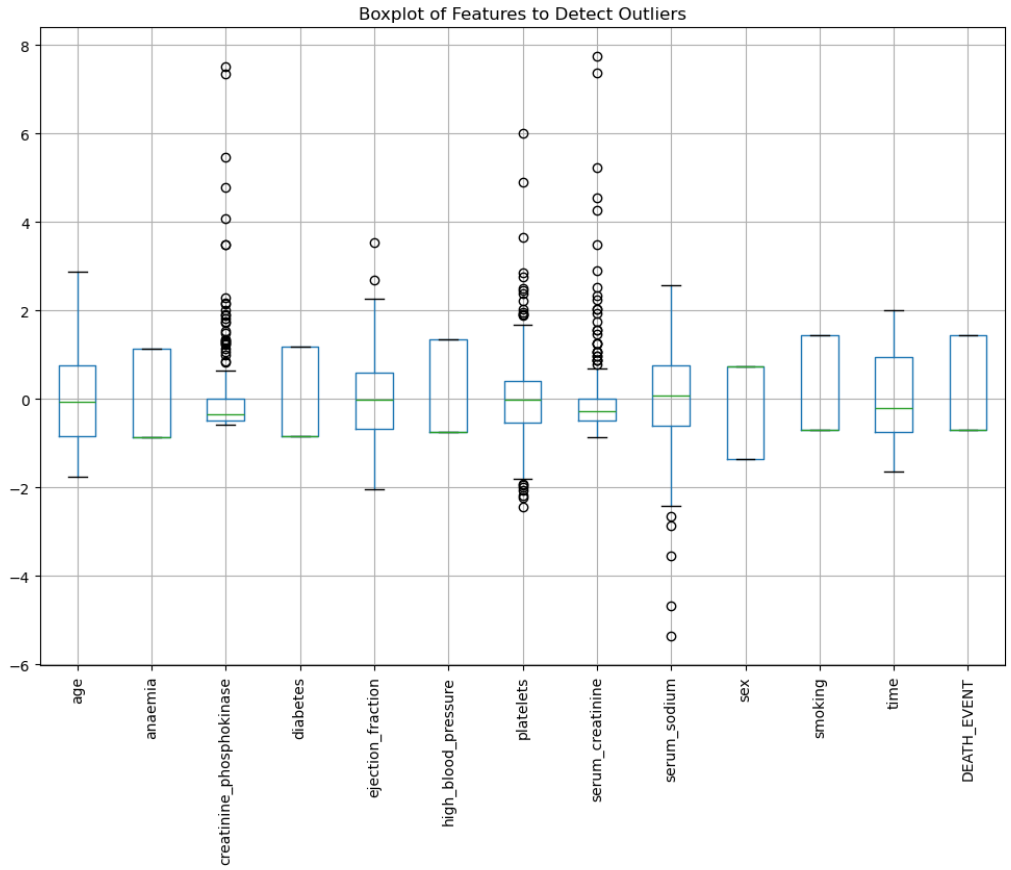


Figure 2: Boxplots of each feature in our dataset, including target.

Lastly, before conducting the PCA, we checked the correlation matrix. As we can see in Fig. 3, the biggest correlation is between features sex and smoking, suggesting that the majority of the patients that smoke are males.

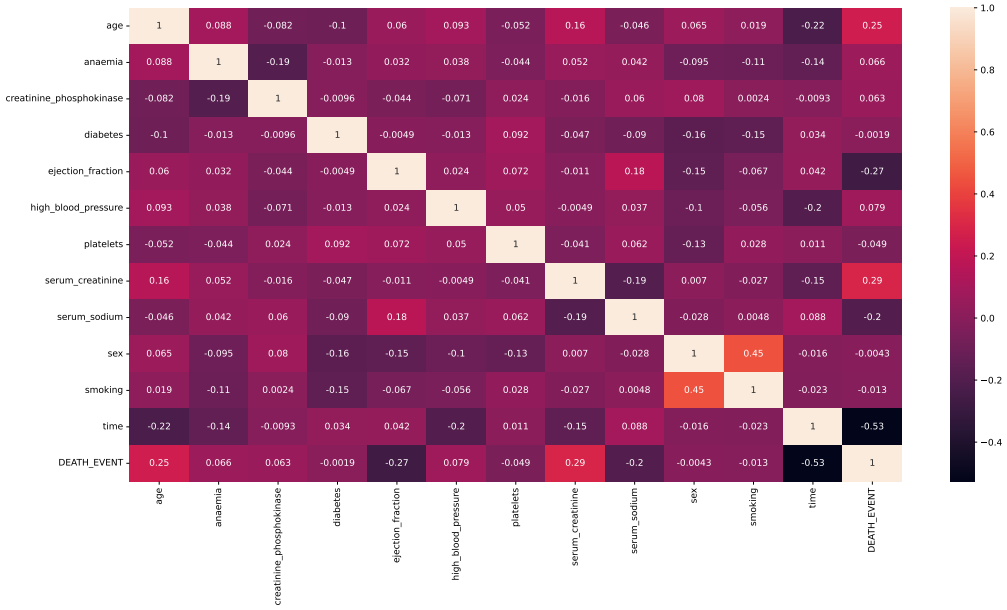


Figure 3: Correlation matrix

From the correlation matrix figure 3 we are able to observe a moderate positive correlation between age and death event (0.25) suggesting that older patients may have a slightly higher chance of experiencing a death event. There is a negative correlation between ejection fraction - a key measure of heart performance - and death event (-0.27), indicating that lower ejection fractions are associated with a higher likelihood of a death event. There is a strong negative correlation (-0.53) between time and death event suggesting that patients who have a shorter follow-up period are more likely to experience death event.

Notably, both ejection fraction and serum creatinine show strong correlations with death events, which aligns with findings from the previous study [1] that highlight the significance of these factors in predicting patient survival.

Most other attributes, like creatinine phosphokinase, diabetes, high blood pressure, and platelets, show weak correlations with death event, meaning they have little to no linear relationship in this dataset.

## 3.2 Principal Component Analysis

After standardizing and centering our data, we conducted the PCA. We experimented using 12 PCA components, as our dataset has 12 features. The explained variance represents the proportion of the dataset's variability captured by each component. As shown in Fig. 5, the first two are the most significant components, with 13.8% and 13.1% of the total variance. The blue line indicates the cumulative sum of the components variances. To achieve more than 90% explained variance, we need to use ten or more principal components, which suggests that our data is structured around many dimensions and that many features could contribute to the variability. Additionally, we have plotted the feature contribution of each feature to the first three principal components in Fig 4.

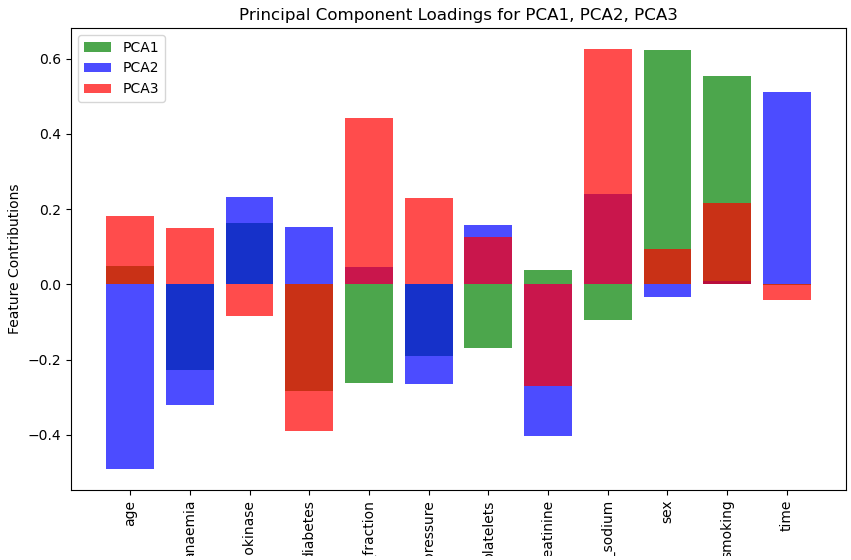


Figure 4: Feature contribution of the first 3 Principal Components

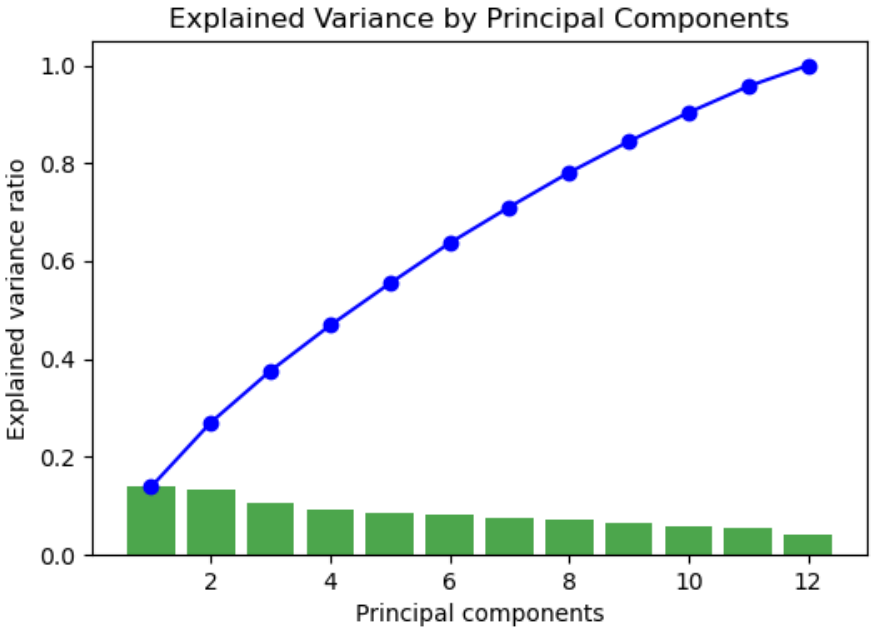


Figure 5: Correlation matrix

Lastly, we have also plotted our target data(DEATH_EVENT), projected onto the first two components to illustrate the effects of the PCA. Fig. 6 illustrates it. As we can observe, using only PCA1 and PCA2, we are not able to separate our data entirely; therefore, again, it suggests that we require additional components to capture underlying patterns and facilitate our future classification fully.

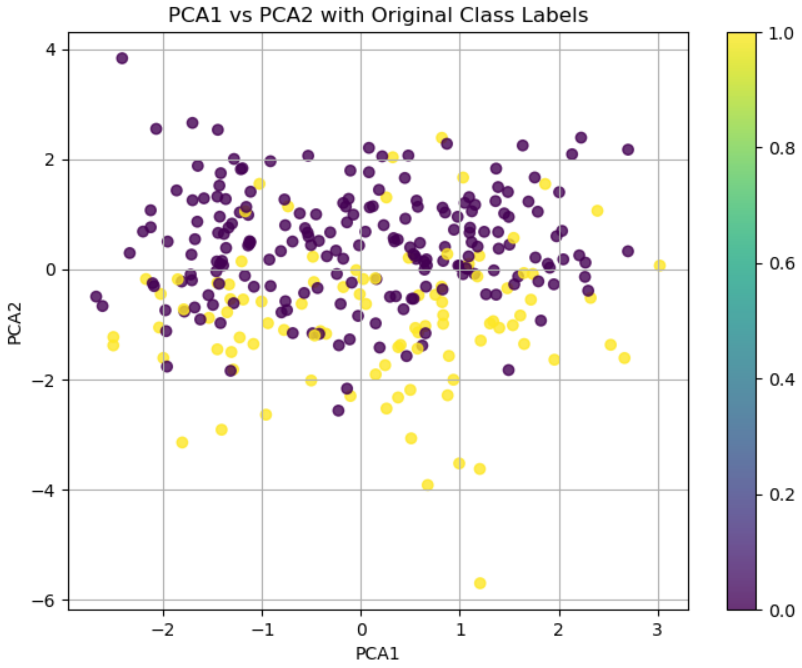


Figure 6: Projection of our target data onto PCA1 and PCA2

# 4 Discussion

Our correlation analysis identified notable relationships between age and death event, ejection fraction and death event, and time and death event. By performing Principal Component Analysis (PCA) and plotting the explained variance, we found that the first two components account for approximately 26.9% of the total variance, while around ten principal components are needed to explain more than 90% of the total variance. This result indicates that the dataset is highly multidimensional, with many features contributing to its overall variability.

The feature contribution plot highlights the most influential features of the first three principal components. For PCA1, the most significant positive contributions come from sex and serum sodium, while ejection fraction shows the most significant negative contribution. In PCA2, age and anemia have the largest negative contributions, while time makes the most significant positive contribution. These results suggest these features are crucial in shaping the dataset's structure within the first two principal components.

Lastly, when we project the target variable (death event) onto the first two principal components, we observe that using only PCA1 and PCA2 is insufficient to separate the classes entirely. This implies that additional principal components may be required to capture the underlying patterns necessary for effective classification.

# 5 Exam problems

## **Question 1** - Spring 2019 question 1

Reviewing the key attributes we see that $x_1$ is coded an an integer to represent different 30-minute intervals. While the intervals themselves are measured on an interval scale, this suggest that the attribute may be considered interval. Attributes $x_2$, $x_4$, $x_6$ and $x_7$ are all counts of occurrences, which are typically measured on a ratio scale because they have a true zero point. The congestion level $y$ are taking discrete ordered values suggesting this attribute is ordinal. In conclusion, the correct answer must be **D**.

## **Question 2** - Spring 2019 question 2

Using definition 4.17 and 4.18 from Chapter 4, we are able to evaluate the $p$-norm distances:

$$d_{p=\infty}(\mathbf{x_{14}},\mathbf{x_{18}}) = \max\{|26-19|, |2-0|\} = 7, \tag{1}$$

Making option **A** the correct answer.

## **Question 3** - Spring 2019 question 3

Using definition 3.18 from Chapter 3, we are able to calculate the explained variance. Here we are calculating the explained variance for the first four principal components:

$$\text{Total variance} = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 = 670.4 \tag{2}$$

$$\text{Explained variance} = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{670.4} = 0.87 \tag{3}$$

The variance explained by the first four principal components is greater than 0.8, thus making option **A** the correct answer

## **Question 4** - Spring 2019 question 4

Concerning option A, high values of **Accident victims** and **Immobilized bus** means large negative values, therefore a negative projection onto principal component number 5, meaning A is not the correct answer.

Having a low value of **Accident victim** and a high value of **Immobilized bus** which has a large negative coefficient would typically result in a negative projection onto principal component number 3. Option B is therefore not the correct answer

Having low values of **Time of Day**, **Accident victim** and **Defects** which all have large negative coefficients, while also having a high value of **Broken truck** which has a large positive coefficient, would typically results in a positive projection onto principal component number 4 instead of a negative value of the projection. C is not the correct answer

An observation with a low value of **Time of day** which has a large negative coefficient, while also having high values for both **Broken Truck**, **Accident victim** and **Defects** which all have large positive coefficients. This means that this observation would typically have a positive value of the projection onto principal component number 2, making option **D** the correct answer

## **Question 5** - Spring 2019 question 14

Using definition 4.23 from Chapter 4 we are able to calculate the Jaccard Similarity. Firstly, we count the entries which is similar for $s_1$ and $s_2$ which is two words: "the" and "words". Secondly, the entries for which $s_1$ is dissimilar to $s_2$ being the following entries: "bag", "of", "representation", "becomes", "less", "parsimoneous" being a total of 6 entries. Now counting the words for which $s_2$ is dissimilar to $s_1$ which are the entries: "if", "we", "do", "not", "stem", being a total of 5 entries. Finally we are able to calculate the Jaccard similarity:

$$\frac{2}{2+6+5} = 0.153846, \tag{4}$$

We are able to conclude that option **A** is the correct answer

# References

[1] Davide Chicco and 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*, 20:1–16, 2020.