Technical report

Heart attack risk assessment system

Amine Miadi, Lucía Rangel Pereyra, Mariia Shpir

Our main motivation to do this project is to save precious time for people at high risk of suffering a heart attack, the goal of our project is to prevent fatal cases and to notify patients of the seriousness of his situation based on medical indications.

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups. Heart disease causes 1 out of every 4 deaths. By 5 years after a heart attack, high percent of people die, develop heart failure, or have a stroke, according to the American Heart Association.

The main idea of the project is to build a risk assessment system for the patients so they don't have to wait or “loose” time while they get the testing results and go again to the doctor so they can check it.

Data Collection

The data we used for this mini project was obtained on the Machine Learning Repository from UCI, **“Echocardiogram Data Set”**.

The data set contains multivariate variables, three different types of attributes: categorical, integer and real, the number of attributes is **13** and it has **132 entries**. The main description of the data set points out that it is about patients who suffered heart attacks at some point in the past.

survival - The number of months patient survived (has survived, if patient is still alive).

still-alive - 0 = dead at end of survival period, 1 = still alive.

age-at-heart-attack - Age in years when heart attack occurred.

pericardial-effusion - Pericardial effusion is fluid around the heart. 0 = no fluid, 1 = fluid.

fractional-shortening - A measure of contractility around the heart, lower numbers are increasingly abnormal.

epss - E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.

lvdd - Left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.

wall-motion-score - Measure of how the segments of the left ventricle are moving.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 131 entries, 0 to 130
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   survival               130 non-null    object
 1   still-alive            131 non-null    int64
 2   age-at-heart-attack    126 non-null    object
 3   pericardial-effusion   131 non-null    int64
 4   fractional-shortening  124 non-null    object
 5   epss                   117 non-null    object
 6   lvdd                   121 non-null    object
 7   wall-motion-score      128 non-null    object
 8   wall-motion-index      130 non-null    object
 9   mult                   128 non-null    object
 10  name                   131 non-null    object
 11  group                  109 non-null    object
 12  alive-at-1             74 non-null     object
dtypes: int64(2), object(11)
memory usage: 13.4+ KB
```

Imported dataset

wall-motion-index - = wall-motion-score divided by the number of segments seen. Usually 12-13 segments are seen in an echocardiogram.

mult - A derivate var which can be ignored.

name - The name of the patient (anonymized).

group - The group of the patient.

Unfortunately, the dataset is quite small, so our team faced many difficulties in the development of models. In the future, in order to launch such a product, the dataset needs to be improved by the specialists for more accurate results.

Data Preprocessing

Dropping following columns:

`alive-at-1, mult, name, group` = don't have useful information for the project goals.

`wall-motion-score` = can be replaced with more detailed 'wall-motion-index' attribute.

Changing the data type.

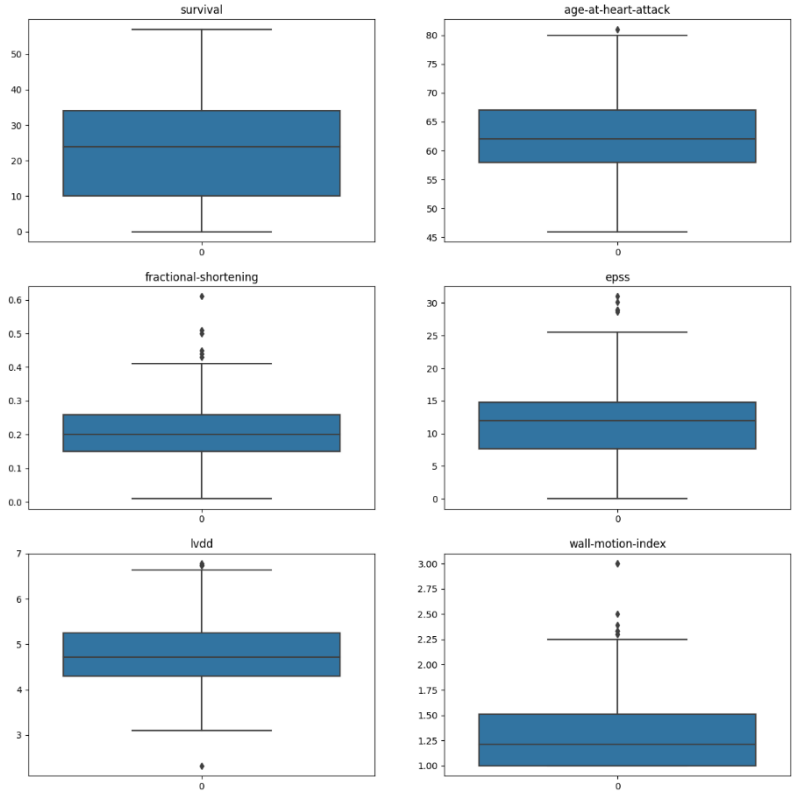
Removing all lines where `'survival'` is not specified (only 1).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 130 entries, 0 to 130
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   survival               130 non-null    float64
 1   alive                  130 non-null    int64
 2   age-at-heart-attack    130 non-null    int32
 3   pericardial-effusion   130 non-null    int64
 4   fractional-shortening  130 non-null    float64
 5   epss                   130 non-null    float64
 6   lvdd                   130 non-null    float64
 7   wall-motion-index      130 non-null    float64
dtypes: float64(5), int32(1), int64(2)
memory usage: 8.6 KB
```

The result after preprocessing

Filling all None values in the following columns using mean function: `age-at-heart-attack, fractional-shortening, epss, lvdd, wall-motion-index.`

EDA & Visualizations



Attributes analysis using Boxplot

First, we analyze the data using the Boxplot to determine the median and outliners (without categorical values).

Unfortunately, dataset contains many outliners that will further degrade the learning process of the model.

Our team decided to denoise the data using special algorithms provided in the Scikit-Learn library.

It significantly improves the accuracy of the models and does not greatly reduce the already small data.

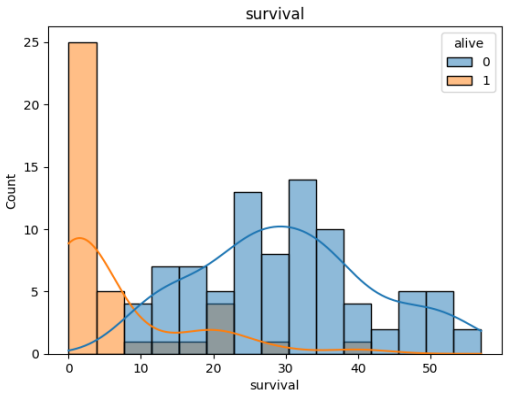
Comparison of distribution of the data based on the target variable

When analyzing the data, we noticed a strange discrepancy between the normal range of the values and the survival rate of patients.

`age-at-heart-attack:` For example, based on the ratio, we can conclude that patients aged 45 to 65 were more likely to die. But older patients (65 and older) were more likely to survive.

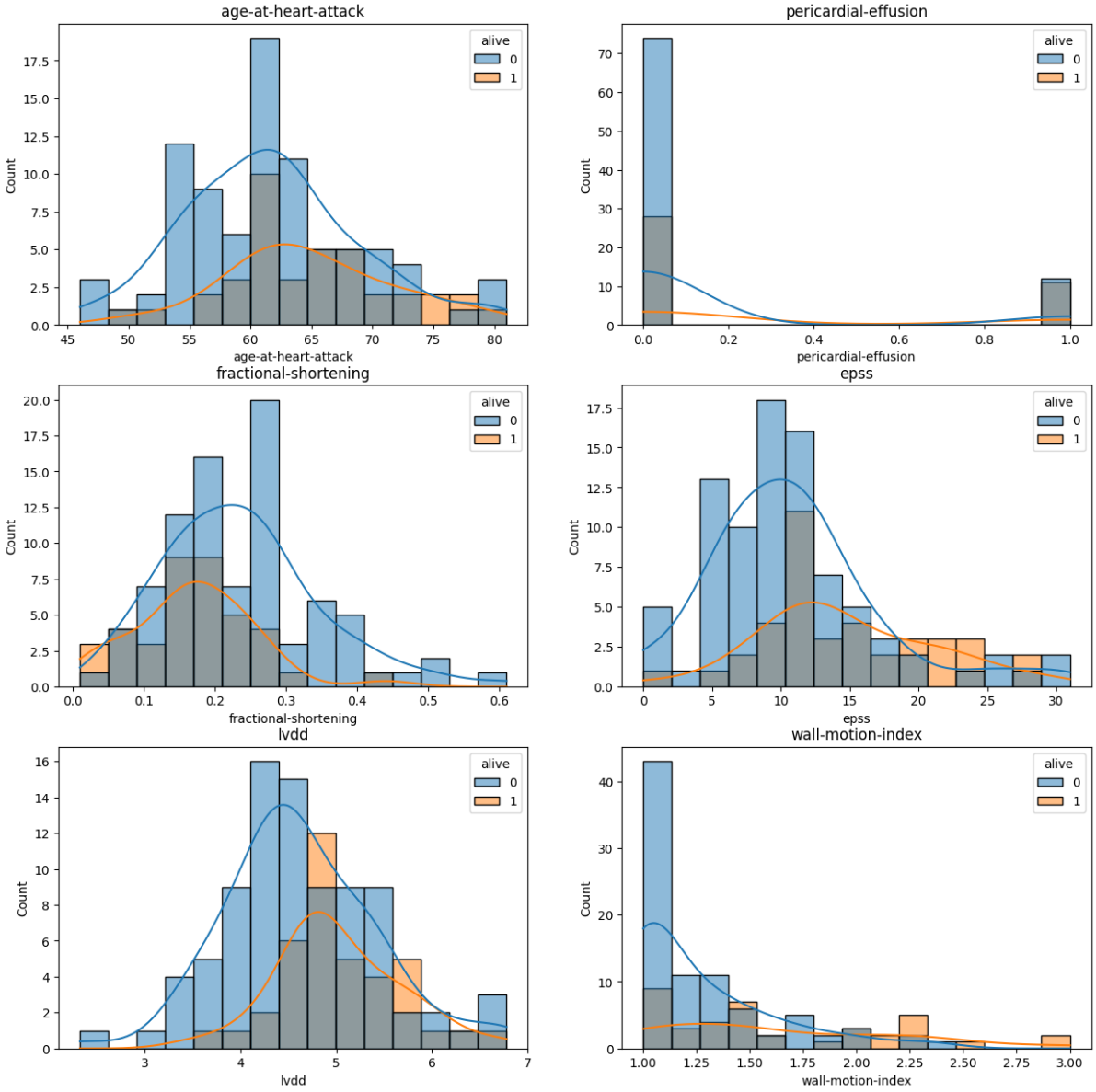
`pericardial-effusion:` Even though normally people should have this index at 0, in such case in our sample patients were more likely to die.

`fractional-shortening:` Normal range is from 0.2 to 4, and the lower the value, the more serious the situation. Unfortunately, there is no such patterns in the data (in our data with lower attribute values, patients were more likely to survive).



Similar conclusions can be made for other attributes besides **`lvdd`**, where indeed with the value in the normal range the person had a better chance of survival.

`survival:` If the patient has had a heart attack, there is a good probability that everything will be fine in the next couple of months.



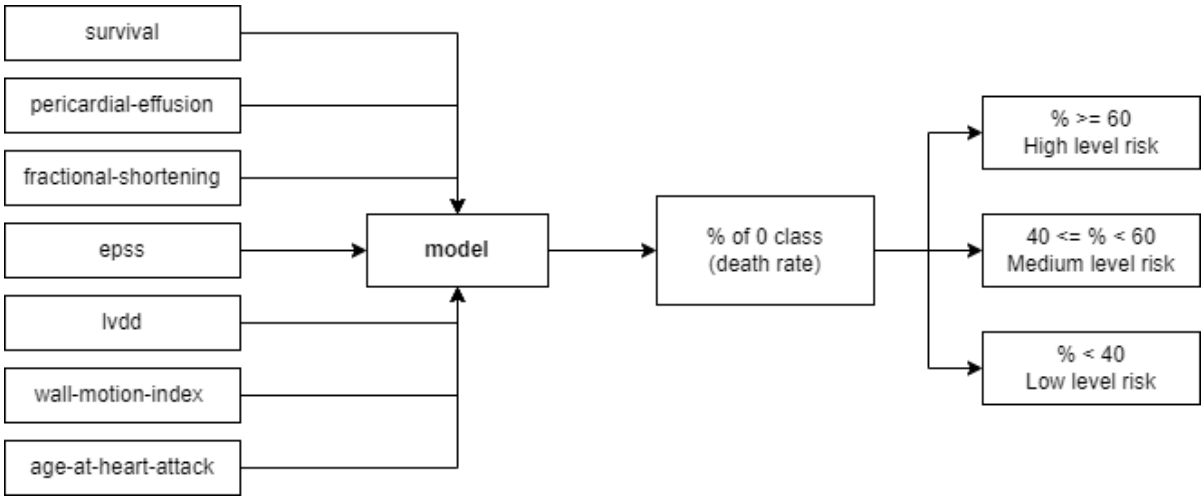
Comparison of distribution of the data based on the target variable

To summarize, we can conclude that the output values can be not 100% accurate, due to insufficient data size. Also, such a size makes it difficult to find a suitable model for the project. Dataset requires refinement by specialized medical staff along with data processing developers.

Learning Task

Our main task is to find out in how dangerous the patient's medical results are (in other words, the danger of his body condition). To do this, we have to predict how high the probability of death is. So, our final goal is a **supervised** (because the model is trained on the proposed dataset) **classification** (because we must "classify" the patient's condition).

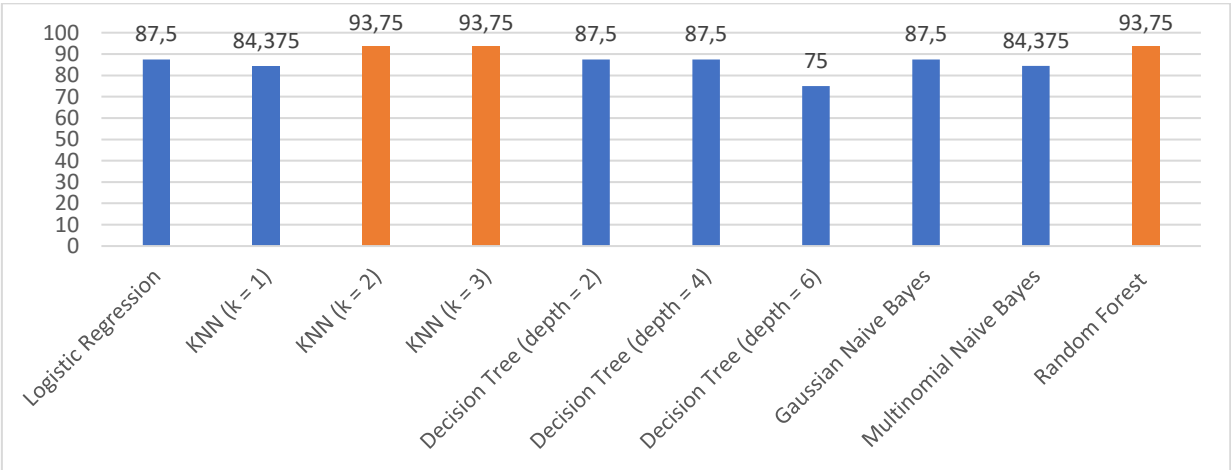
More closely, the learning task of our project is presented in schema below:



Scheme of data input, processing and output

Learning Approach

The data for the training were divided as follows: **75% training set, 25% test set**. As a result, training set`s shape is (93, 7) and test set's shape is (32, 7). The following algorithms will be used for classification (from Scikit-Learn library): K-Nearest Neighbors, Decision Tree Classifiers/Random Forests, Naive Bayes, Logistic Regression.



Comparison of the accuracy of different models

For analyzing Random Forest models and selecting hyperparameters, the dataset will be divided as follows: **60% training set, 25% validation set, 15% test set**. As a result, corresponding shapes are (75, 7), (32, 7), (18, 7).

Thus, in the max_depth range from 1 to 10 and in the n_estimators from 1 to 16 the best values of hyperparameters are **1** and **11**, model accuracy will be 94.4% (with 60/15) and **93.75%** (with 75/25).

Unfortunately, again the problem was the small size of the dataset. This made it difficult to choose the right hyperparameters and to train the model. In the long term, with the sufficient data, the **Random Forest** model will show better results than other models.

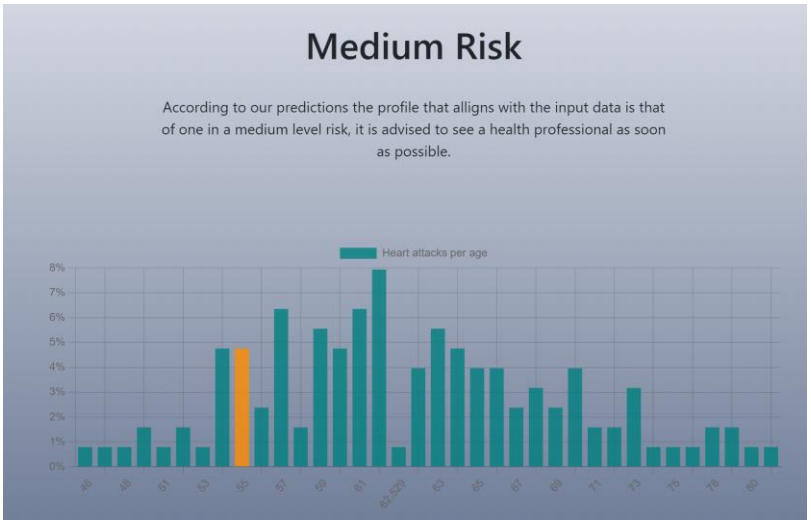
Data Privacy

As a socially responsible company we base our privacy of data use in according with GDPR laws that look to protect the privacy of the users. Data will be handled anonymously and all data collected will not store sensitive user information, we share the complete consent information with the users so they can read it before using the app and as Art 12 GDPR says, users should have completely access to their information and how it is managed so we added an email where they can contact the company to get their information. Compliance with the law allows us to have a reliable website where users can protect themselves according to the law and also as a company to protect ourselves from any mishaps.

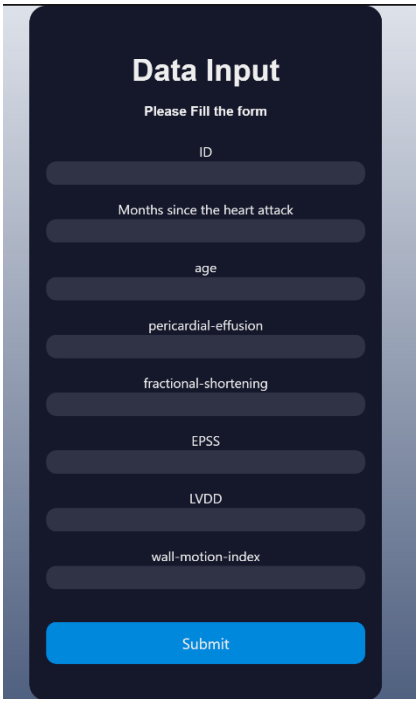
Communication of Results

https://heart-beat-app.herokuapp.com

https://github.com/Amine-Miadi/heartbeat-app



The results



Data entry form

The site provides information about us, privacy and an entry form. After entering the data, the user is shown the output information, as well as charts describing what specific range the patient's medical results are in the overall distribution of values in our dataset (using percentage ratio).