**Survival Prediction of Patients with Heart Disease using Machine Learning Techniques**

**Abstract:**

Heart Failure is one of the common diseases that can create a dangerous situation. It is extraordinarily complex to detect heart failure at the right time. That is where data comes into the picture. We now live in a time where we have a vast collection of data with the rising of technology. By this data, we can know many factors which affect the health of a heart like blood pressure, heart rate, chest pain, pulse rate, etc. In this paper, we will use these data for different classification techniques like Support Vector Machine, decision tree, and Random Forest for predicting the survival of patients with heart disease. We have taken the 13 attributes that are present in the dataset as an input and applying on the dataset which is taken from the UCI repository to develop an accurate model of predicting the survival. It is found that the Decision Tree provides the most accuracy. The ability of the machine learning model presented in this study with medical info systems would be useful to predict the survival of patient using the data collected from hospitals.

**Key Words: classification, regression, support vector machine, decision tree, random forest, Uci, accuracy, survival, machine learning**

**Introduction:**

Heart diseases are most common in the modern world with increasing stress, and irregular lifestyle. The heart is one of the vital devices of the human body responsible for transporting the blood and oxygen throughout the body. If any irregularity in heart functioning leads to different functional disorders and it'll paralyze the traditional flow of life. According to the Centers for disease control and prevention, about 6,10,000 people die in the United States due to heart diseases and that is 1 in every 4 deaths that occur. Heart diseases are responsible for killing a huge population annually. Due to uncontrolled high blood pressure and low-density lipoprotein cholesterol or current smoking increases the risk of heart disease in about 47% of adults.

Machine learning has become one of the foremost powerful technology in giving predictive results. As formally defined, machine learning is the ability of the machines to learn from their past experiences. The past experiences here denote the dataset that is used to train the model. Machine learning is broadly classified as supervised learning and unsupervised learning. Survival prediction, comes under the supervised learning. In this paper, we discuss about the algorithms that can be used for the prediction of the deaths in a heart patient. The prediction model gets the data from the patients’ health check reports and provides a suggestion to the user about the health of their heart. It will help to predict the occurrence of the disease accurately, eliminating the chances for human errors.

**Literature Review:**

Davide Chicco et.al used the dataset of 299 patients with heart failure collected in 2015 applied stratified logistic regression to predict the patient’s survival, and highlight the features corresponding to the most important risk factors. In their result they found that two features, serum creatinine and ejection fraction are sufficient to predict survival of patients.

Tanvir Ahmed et.al used the dataset of 45+ age of 105 women and 194 men, performed cox regression to develop a model that can link the hazards of death of an individual with one or more explanatory variables. In the result they found that 32% patients died due to CHD.

**Problem and Data Set Description:**

The main explanation for heart stroke is thanks to blockage in arteries. It has many other names like disorder and arterial hypertension. There are 26 million people around the world affecting by heart disease. Most of the time patients choose several tests, which may overburden them with extra physical activities, time, and additional finance charges. It is extraordinarily complex to detect heart failure at the right time

This paper has taken a small step towards saving the lives of heart failure patients and describes a way to improve the performance of diagnosing the patients based on their medical history.

The heart failure clinical records Data Set contains the medical records of 299 patients who had heart failure.

The dataset contains 13 clinical features (some of them are binary, others are numerical), the follow-up period, and the label **DEATH\_EVENT** that indicates whether the patient has died.

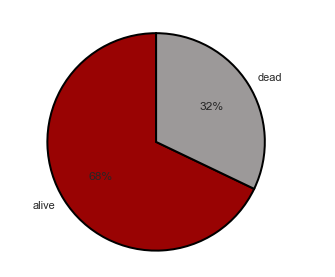
We can find some features strictly associated with medical aspects like levels of enzymes, sodium, creatinine, and platelets within the blood et al. that are more common like age, sex, or smoking.

The dataset is collected in 2015.

The data set is complete in every manner. So, as we thought use **rapid miner** tool for data pre-processing but due to the complete data, we must skip this step and focus on **data validation and finding data co-relation**

Most cardiovascular diseases are often prevented by addressing behavioural risk factors like tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol using population-wide strategies. People with the disorder or who are at high cardiovascular risk (due to the presence of 1 or more risk factors like hypertension, diabetes, hyperlipidaemia, or already established disease) need early detection and management wherein a machine learning model is often of significant help. The prediction model is trained with the training data and tested with the testing data. 80% of the data in the dataset is used as training data and the remaining 20% as test data is the testing data. 13 attributes are used as input and the “target‟ class is predicted by the machine learning model.

One thing to consider is the possible **class imbalance**. We can plot it with the help of **matplotlib** and get the better visuals of **data imbalance**



**Table: Used attributes from the dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Explanation** | **Measurement** | **Range** |
| **Age** | **Age of the patient** | **Years** | **[40, 95]** |
| **Anaemia** | **Decrease of red blood cells or haemoglobin** | **Boolean** | **0,1** |
| **High blood pressure** | **If a patient has hypertension** | **Boolean** | **0,1** |
| **Creatinine phosphokinase** **(CPK)** | **Level of the CPK enzyme** **in the blood** | **mcg/L** | **[23, 7861]** |
| **Diabetes** | **If the patient has diabetes** | **Boolean** | **0, 1** |
| **Ejection fraction** | **Percentage of blood leaving** **the heart at each contraction** | **Percentage** | **[14, 80]** |
| **Sex** | **Woman or man** | **Binary** | **0, 1** |
| **Platelets** | **Platelets in the blood** | **kilo platelets/mL** | **[25.01, 850.00]** |
| **Serum creatinine** | **Level of creatinine in the blood** | **mg/dL** | **[0.50, 9.40]** |
| **Serum sodium** | **Level of sodium in the blood** | **mEq/L** | **[114, 148]** |
| **Smoking** | **If the patient smokes** | **Boolean** | **0, 1** |
| **Time** | **Follow-up period** | **Days** | **[4,285]** |
| **DEATH EVENT** **(TARGET)** | **If the patient died during the follow-up period** | **Boolean** | **0, 1** |

**Methods:**

In this paper, we used various machine learning algorithm methods for predicting deaths in heart disease patients using medical dataset. applied the different machine learning techniques and found out the algorithm that gives the best accuracy and our target value contains 0 and 1 which tells us that if 0 means the patient is safe and 1 indicates the patient is dead.

We have used 3 different machine learning algorithms namely Decision Tree, Random Forest, and Support Vector Machine. We will discuss the working of each algorithm separately.

**Very first method is to visualize the data in every manner and plotting the data features with target and after doing so, we can better know the relationship between the features and target. This visualization will give us clear view about the uniqueness of features and their count.**

**1.Decision Tree:**

A decision tree is a graphical representation of all the practical solutions to a decision based on certain condition.

Decision tree may be a supervised learning method used for classification and regression.

Regression tree is used when dependent variable is continuous.

In case of regression, the tree will make splits such that each group has the lowest mean squared error.

Classification trees are used when variable is categorical

They are non-parametric models that learn by recursively split the predictor space (and therefore the train samples) consistent with the simplest feature (greedy approach) until the tree reaches a constrained depth, the subsets contain elements of only one class or it meets another stopping criterion (e.g., less than 5 samples in the subset).

The best feature is that the feature that, if want to discriminate samples, allows us to get the simplest possible split consistent with a measure.

Such measure in this case is the Gini Index (in this Sklenar implementation the unnormalized Gini index is used)

**𝐺=1−∑𝑗𝑝2𝑗**

where 𝑝 is the ratio between number of samples of class 𝑗

and total number of samples.

Gini index is evaluated on both the splits, weighted by the number of samples in each split and therefore the feature that provides us the lower overall Gini index is chosen.

Due to the nature of this model, the explain ability is quite high. Is in fact possible to plot the tree to see how the dataset is split at every step and it is easy to describe how the prediction works.

The main drawback of decision tree is the fact that sometimes is a too simple model, that provides lower values of accuracy and it is easy to overfit.

To tackle this problem, it is possible to train an ensemble of decision trees, called random forest.

**2.Support Vector Machine:**

Support vector machine may be a powerful model used for both classification and regression.

It consists in trying to fit a hyperplane that best divides the dataset into the two classes by maximizing the margin (the distance between the hyperplane and the closest points).

Hard margin

The simplest implementation is that the hard margin SVM during which data must be linearly separable to permit the algorithm to converge.

**𝐿= {𝑣: ⟨𝑤, 𝑣⟩+𝑏=0}, ‖𝑤‖=1**

the distance of a point form the hyperplane L can be evaluated in that way:

𝑑(𝑥,𝐿)=|⟨𝑤,𝑥⟩+𝑏|

while the space between two points of two different classes on the margin is:

‖𝑊‖= (+ 𝑏) −(+ 𝑏) ‖𝑊‖=2‖𝑊‖

Equivalently:

**𝑚𝑖𝑛𝑤, 𝑏12‖𝑤‖2 𝑠.𝑡. ∀𝑖, 𝑦𝑖 (⟨𝑤, 𝑥𝑖⟩+𝑏)>1**

where 𝑦𝑖

is the true label and therefore the prediction is that the evaluated distance of the sample from the hyperplane? If the label and therefore the prediction have an equivalent sign it means the prediction is on the right side of the margin.

When the algorithm has converged, then the model is often described using only the points on the margin (called support vectors). This means that SVM can scale well.

The main drawback is that, in the real world, most of the problems are not linearly separable, and an algorithm like this one would not converge.

For this reason, we can add a term on the constraints to relax them.

Soft margin.

**𝑚𝑖𝑛𝑤, 𝑏(12‖𝑊‖2+𝐶∑𝑖=1𝑚𝜉𝑖) 𝑠.𝑡. ∀𝑖, 𝑦𝑖 (⟨𝑤, 𝑥𝑖⟩+𝑏) ≥1−𝜉𝑖 𝑎𝑛𝑑 𝜉𝑖≥0**

is called slack variable, it is the distance of 𝑥𝑖 from the corresponding class's margin if 𝑥𝑖

is on other side of the margin and 0 otherwise.

Smaller values of C allow more errors in exchange of a bigger margin, while higher values can be used where it is needed to be less permissive regarding misclassifications, with a higher risk of overfitting.

In some cases, problems are not well separable on the first feature space but they're separable on another space.

Therefore, it is possible to map our features in another higher dimensional space in which hopefully it will be easier to learn.

One simple example is the polynomial mapping:

A degree k polynomial is 𝑝(𝑥)=∑𝑘𝑗=0𝑤𝑗𝑥𝑗

that can also be seen as: ⟨𝑤, 𝜓(𝑥)⟩ 𝑤ℎ𝑒𝑟𝑒 𝜓(𝑥)= (1, 𝑥, 𝑥2, 𝑥𝑘)

𝜓(𝑥)

is the polynomial mapping function

The problems arise within the case of multiple predictors.

In fact, in a polynomial mapping on a IR2

space:

**𝜓(𝐱)= (𝑥21, 𝑥22,2⎯⎯√𝑥1𝑥2)**

(Considering only the quadratic features)

The solution is using kernel functions.

Kernel trick

For solving the support vector machine optimization, computing the inner products, it's needed.

In fact, according to representer theorem, 𝑤

can be written as:

**𝑤=∑𝑖𝛼𝑖𝜓(𝑥𝑖)**

so, it follows that:

**‖𝑤‖2=⟨∑𝑗𝛼𝑗𝜓(𝑥𝑗), ∑𝑗𝛼𝑗𝜓(𝑥𝑗)⟩=∑𝑚𝑖, 𝑗=1𝛼𝑖𝛼𝑗⟨𝜓(𝑥𝑖), 𝜓(𝑥𝑗)⟩**

and the problem now consists in minimizing the margin over 𝛼.

Once 𝛼

Considering the mapping, this operation could be unfeasible.

A **kernel function** is a function that implements the inner product in the new feature space (given *𝜓*):

*𝐾*(**𝐱**,**𝐱**′)=⟨*𝜓*(**𝐱**),*𝜓*(**𝐱**′)⟩

In this way we do not have to explicitly apply *𝜓*

on our data and then compute the inner product.

A symmetric function *𝐾: 𝑋*×*𝑋*→IR

can be a kernel function if and only if it respects the **Mercer theorem** that says that the Gram matrix, that is the matrix such that: *𝐺𝑖, 𝑗*=*𝐾 (***𝐱𝑖, 𝐱𝑗**)

needs to be **positive semidefinite**.

A matrix is semidefinite if and only if:

**𝐱***𝑇𝐺***𝐱**>0 ∀**𝐱**∈*𝑅𝑛*∖0

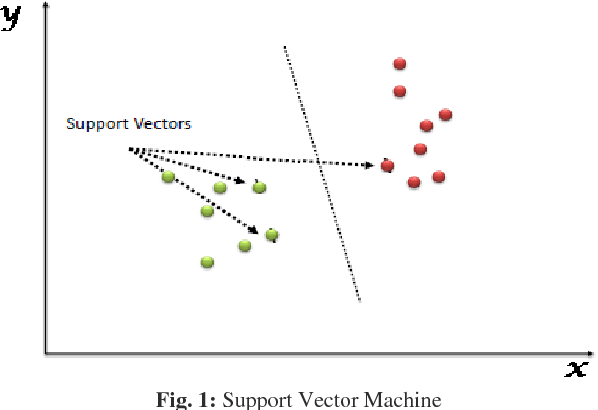
and the eigen values are **non-negative**.

The most used kernels are:

* the **polynomial kernel** (degree k)
* (⟨**𝐱, 𝐱**′⟩+1) 𝑘
* the **radial basis function** (gaussian kernel)
* *𝑒*−*𝛾*‖**𝐱**−**𝐱**′‖2

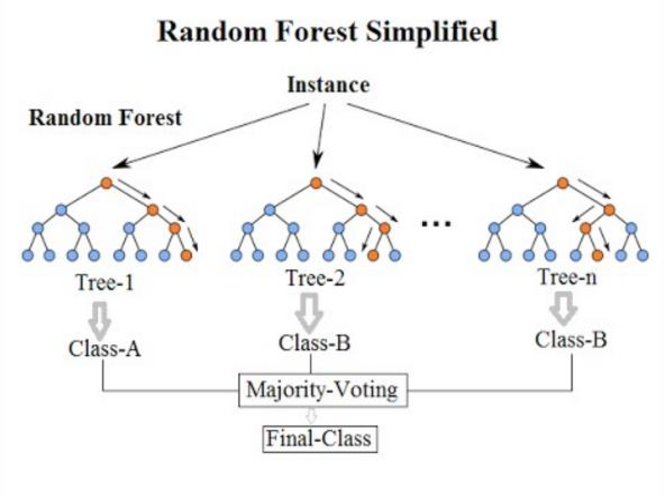
Linear kernel means that no mapping is done and the kernel is simply ⟨**𝐱, 𝐱**′⟩

**An SVM (Support Vector Machine) model is a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).**



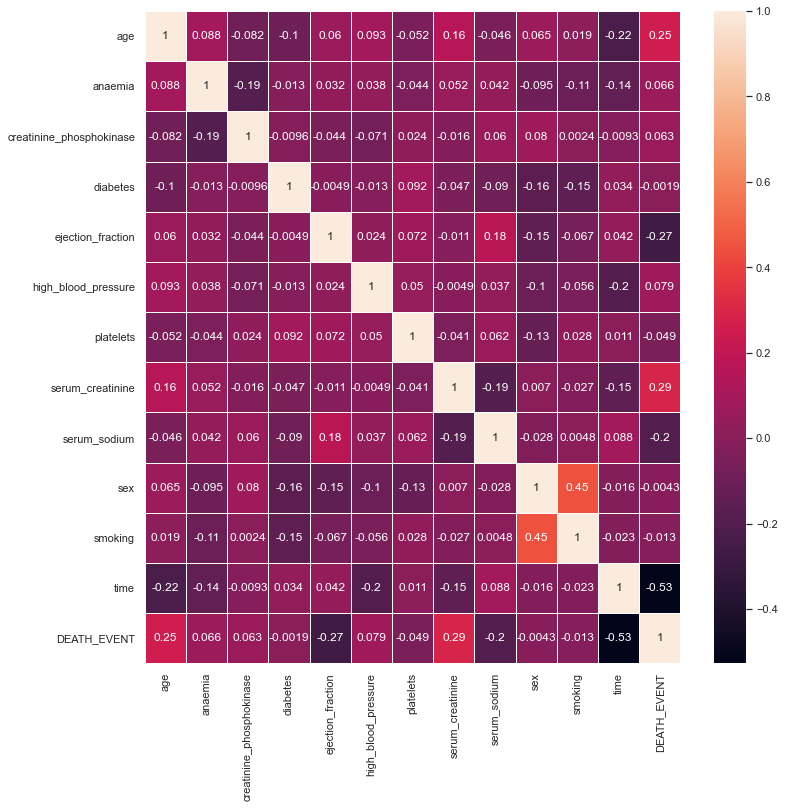
**3.Random Forest:**

Random forest algorithm may be a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a category expectation and therefore the class with most votes turns into a model's forecast. In the random forest classifier, the more the number of trees higher is that the accuracy. It is used for classification also as regression task, but can had best with classification task, and may overcome missing values. Besides, being slow to get predictions because it requires large data sets and more trees, results are unaccountable.



**Experimental Setup:**

The dataset was already pre-processed, so that was not required that much. We found the correlations of the attributes by a correlation matrix



**1. The matrix shows the correlation among the features and their correlation with the DEATH\_EVENT.**

**2. Five features - 'age', 'ejection fraction', 'serum creatinine, 'serum sodium', 'creatinine phosphokinase', 'time' seem to be the most correlated to the death event when compared to the other features.**

**Data Preparation:**

As the ‘platelets’ and ‘creatinine phosphokinase’ features are in quite a different range. So, we need to perform data normalization on these features to change the values to a common range.

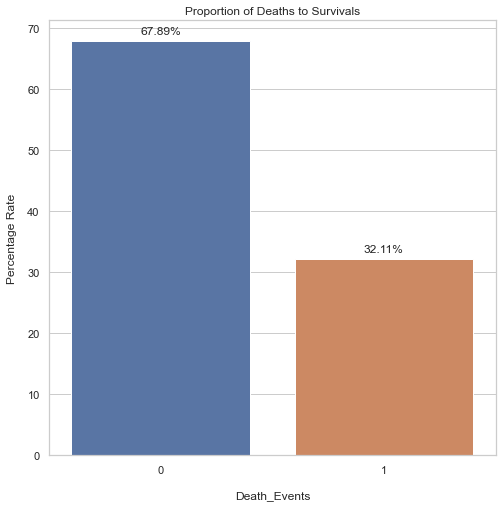
To deal with imbalanced data, resampling data is the most common approach. Synthetic Minority Oversampling Technique is used for oversampling where the synthetic samples are generated for the minority class. This algorithm helps to beat the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interruption between the positive instances that lie together.

We visualized the target column of the dataset, in which we found that

**1. within the population being studied for the condition of coronary failure out of the 299 cases, 96 have succumbed to the condition while 203 cases survived.**

**2. In percentages, 32.11% of the cases succumbed (positive examples) while 67.89% of cases survived (negative examples) the condition.**

.



We visualized all the other columns of the dataset analysed the results

**Results:**

The aim of this project was to predict heart diseases using various classification algorithms. By using techniques of Decision Tree, Support Vector Machine and Random Forest algorithm on the UCI dataset, we found the accuracies of them, whether the patient will survive or not. The dataset was split into training and testing data and the accuracies were noted.

The comparison of accuracies of the algorithms is given below

**All models are evaluated considering the following metrics**

* ***𝑎𝑐𝑐𝑢𝑟𝑎𝑐𝑦*=*𝑇𝑃*+*𝑇𝑁𝑇𝑃*+*𝑇𝑁*+*𝐹𝑃*+*𝐹𝑁***
* ***𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛*=*𝑇𝑃𝑇𝑃*+*𝐹𝑃***
* ***𝑟𝑒𝑐𝑎𝑙𝑙*=*𝑇𝑃𝑇𝑃*+*𝐹𝑁***
* **${F\_1 = 2 \times \frac {precision \times recall} {precision+recall}**

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| **Decision Tree** | **81.67%** |
| **Support Vector Machine** | **80.00%** |
| **Random Forest** | **78.33%** |

**Conclusion:**

The project discusses the use of various machine algorithm like Naïve Bayes, Support Vector Machine and Random Forest for predicting whether the patient will survive or not. Decision Tree has given the highest accuracy of 81.67%. It will help the health professionals in early prediction of heart disease and take preventative measures. Machine Learning models can prove to be a time saving factor in predicting the event of death and can help the doctors take additional precautionary measures with critical care**.** Many lives can be saved in early detection of these serious heart disease.The accuracy and model performance can be increased if more data is made available for the model to train on. This data limitation, limits the model performance. As more data become available, the prediction accuracy and model's ability to act on other similar real-time data could be enhanced.

**1. 'age', 'ejection\_fraction', 'serum\_creatinine', 'serum\_sodium', 'time' seem to be the features that are more closely related to the death event.**

**2. The event of a heart failure in the old age could be fatal.**

**3. The levels of 'ejection\_fraction', 'serum\_creatinine', 'serum\_sodium' play a significant role in the event of heart failure or the abnormalities in the levels of these become a key factor in triggerring heart failure condition.**

**4. Machine Learning models can prove to be a time saving factor in predicting the event of death and can help the doctors take additional precautionary measures with critical care.**

**5. The imbalance in the dataset had been dealt using the SMOTE oversampling of the minority class. SMOTE balanced the no. of instances for each class but the amount of data is still limited.**

**6. The accuracy and model performance can be increased if more data is made available for the model to train on. This data limitation, limits the model performance. As increased data become available, the prediction accuracy and model's ability to act on other similar real-time data could be enhanced.**

**Team Contribution:**

**References:**

1. [**Heart Disease Prediction using Machine Learning Techniques (ijert.org)**](https://www.ijert.org/research/heart-disease-prediction-using-machine-learning-techniques-IJERTV9IS110259.pdf)
2. [**International Journal of Recent Technology and Engineering (IJRTE)**](https://www.ijrte.org/wp-content/uploads/papers/v8i4/D4537118419.pdf)
3. [**10.pdf (ijiet.com)**](http://ijiet.com/wp-content/uploads/2019/04/10.pdf)
4. [**Figure 3 from Heart disease prediction using machine learning techniques : a survey | Semantic Scholar**](https://www.semanticscholar.org/paper/Heart-disease-prediction-using-machine-learning-%3A-a-Ramalingam-Dandapath/fd328a4c3f414e87c64eafbd5b3625042adb4b78/figure/2)
5. [**https://www.kaggle.com/andrewmvd/heart-failure-clinical-data**](https://www.kaggle.com/andrewmvd/heart-failure-clinical-data)
6. [**https://www.researchgate.net/publication/321259715\_Chronic\_Heart\_Failure\_Detection\_from\_Heart\_Sounds\_Using\_a\_Stack\_of\_Machine-Learning\_Classifiers**](https://www.researchgate.net/publication/321259715_Chronic_Heart_Failure_Detection_from_Heart_Sounds_Using_a_Stack_of_Machine-Learning_Classifiers)
7. [**https://www.tutorialspoint.com/machine\_learning\_with\_python/classification\_algorithms\_support\_vector\_machine.htm**](https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_support_vector_machine.htm)
8. [**https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001**](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001)
9. [**https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5**](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5)