Heart Failure Death Prediction using Machine Learning

Math 569 [Statistical Learning]

Submitted on Dec 6, 2020

Submitted By:

- **Ayush Dadhich [A20449379]**
  adadhich@hawk.iit.edu *(contribution-33.34%)*
- **Parth Gupta [A20449774]**
  pgupta31@hawk.iit.edu *(contribution-33.33%)*
- **Arinjay Jain [A20447307]**
  ajain80@hawk.iit.edu *(contribution-33.33%)*

Abstract:

We explore building generative predictive models of popular machine learning environments. Our heart failure death prediction model can be trained quickly in a supervised manner to learn clinical trial data. There are some factors that affects Death Event. Our dataset contains person's information like age, sex, blood pressure, smoke, diabetes, ejection fraction, creatinine phosphokinase, serum creatinine, serum sodium, time and we have to predict their DEATH EVENT. By calculating a HF DEATH EVENT based on patient-specific characteristics from Electronic Health Records (EHRs), we can identify high-risk patients and apply individualized treatment and healthy living choices to potentially reduce their mortality risk. Our results show the models which were built using EHR data are more accurate with the convenience of being more readily applicable in routine clinical care.

Introduction:

Heart failure (HF) is primarily caused by the inability of the heart to supply sufficient blood flow to the body. It has become one of the most deadly cardiovascular diseases in the 21st century [1]. Therefore, it is important to identify patients who are at a higher risk of mortality due to HF and assess the impact of HF therapy on their outcomes. Several studies have developed prognostic tools for HF, and one of the most commonly used tools is the Seattle Heart Failure Model (SHFM) [1]. SHFM was based on the PRAISE I clinical trial database and validated in five other cohorts. Heart failure (HF) is a complex clinical syndrome and not a disease. It prevents the heart from fulfilling the circulatory demands of the body, since it impairs the ability of the ventricle to fill or eject blood. In this project we are trying to build machine learning algorithm that can lean and predict Heart failure.

On researching this topic, we found that most papers have worked on analyzing the Heart failure and their risk score to compare their previous trial runs. Our main focus was to analyze every available clinical trial feature and see which ones have the most impact on the Heart failure cause. To do so, we had to find large enough datasets of Heart failure clinical data and a technology to handle this big data, analyzed it and give better prediction result.

When it came to choosing the Machine Learning Algorithms for this project, we decided to go with the most popular and power full classifications Machine Learning algorithms like SVM, Decision Tree, KNN, Logistic regression and Random Forest for our prediction model implementation.

An Open-Heart model:

Photo by Robina Weermeijer on Unsplash

Problem Statement or Data Sources:

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 worldwide. 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 behavioral 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, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Data Source:
https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1

We analyzed a dataset containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All 299 patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure.

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
(CPK)			
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving	Percentage	[14,..., 80]
	the heart at each contraction		
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/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]
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

Table 1

The dataset contains 13 features, which report clinical, body, and lifestyle information (Table 1), that we briefly describe here. Some features are binary: anaemia, high blood pressure, diabetes, sex, and smoking (Table 1). The hospital physician considered a patient having anaemia if haematocrit levels were lower than 36%. Unfortunately, the original dataset manuscript provides no definition of high blood pressure.

Regarding the features, the creatinine phosphokinase (CPK) states the level of the CPK enzyme in blood. When a muscle tissue gets damaged, CPK flows into the blood. Therefore, high levels of CPK in the blood of a patient might indicate a heart failure or injury. The ejection fraction states the percentage of how much blood the left ventricle pumps out with each contraction. The serum creatinine is a waste product generated by creatine when a muscle breaks down. Especially, doctors focus on serum creatinine in blood to check kidney function. If a patient has high levels of serum creatinine, it may indicate renal dysfunction. Sodium is a mineral that serves for the correct functioning of muscles and nerves. The serum sodium test is a routine blood exam that indicates if a patient has normal levels of sodium in the blood. An abnormally low level of sodium in the blood might be caused by heart failure. The death event feature, that we use as the target in our binary classification study, states if the patient died or survived before the end of the follow-up period, that was 130 days on average. The original dataset article unfortunately does not indicate if any patient had primary kidney disease, and provides no additional information

about what type of follow-up was carried out. Regarding the dataset imbalance, the survived patients (death event = 0) are 203, while the dead patients (death event = 1) are 96. In statistical terms, there are 32.11% positives and 67.89% negatives.

As done by the original data curators, we represented this dataset as a table having 299 rows (patients) and 13 columns (features). For clarification purposes, we slightly changed the names of some features of the original dataset (Additional file 1). We report the quantitative characteristics of the dataset in Table 2 and Table 3. Additional information about this dataset can be found in the original dataset curators' publication.

	Full sample		Dead patients		Survived patients	
Category feature	**#**	**%**	**#**	**%**	**#**	**%**
Anaemia (0: false)	170	56.86	50	52.08	120	59.11
Anaemia (1: true)	129	43.14	46	47.92	3	40.89
High blood pressure (0: false)	194	64.88	57	59.38	137	67.49
High blood pressure (1: true)	105	35.12	39	40.62	66	32.51
Diabetes (0: false)	174	58.19	56	58.33	118	58.13
Diabetes (1: true)	125	41.81	40	41.67	85	41.87
Sex (0: woman)	105	35.12	34	35.42	71	34.98
Sex (1: man)	194	64.88	62	64.58	132	65.02
Smoking (0: false)	203	67.89	66	68.75	137	67.49
Smoking (1: true)	96	32.11	30	31.25	66	32.51

#: number of patients. %: percentage of patients. Full sample: 299 individuals. Dead patients: 96 individuals. Survived patients: 203 individuals.

Table 2

	Full sample			Dead patients			Survived patients		
Numeric feature	**Median**	**Mean**	**σ**	**Median**	**Mean**	**σ**	**Median**	**Mean**	**σ**
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

Full sample: 299 individuals. Dead patients: 96 individuals. Survived patients: 203 individuals. σ: standard deviation

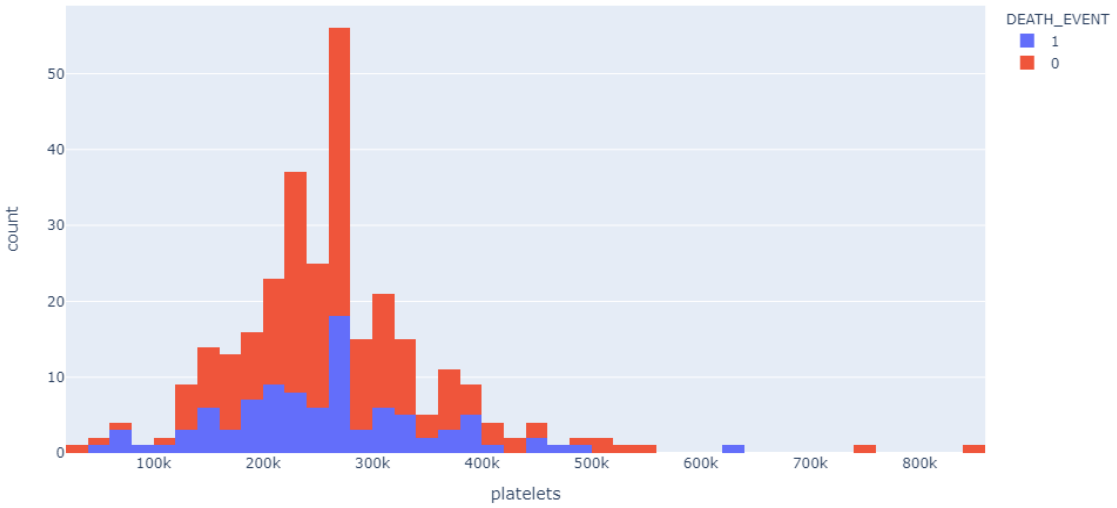
Table 3

Analysis:

We did some data analysis that are counted in our analysis. We have two class for Death Event.

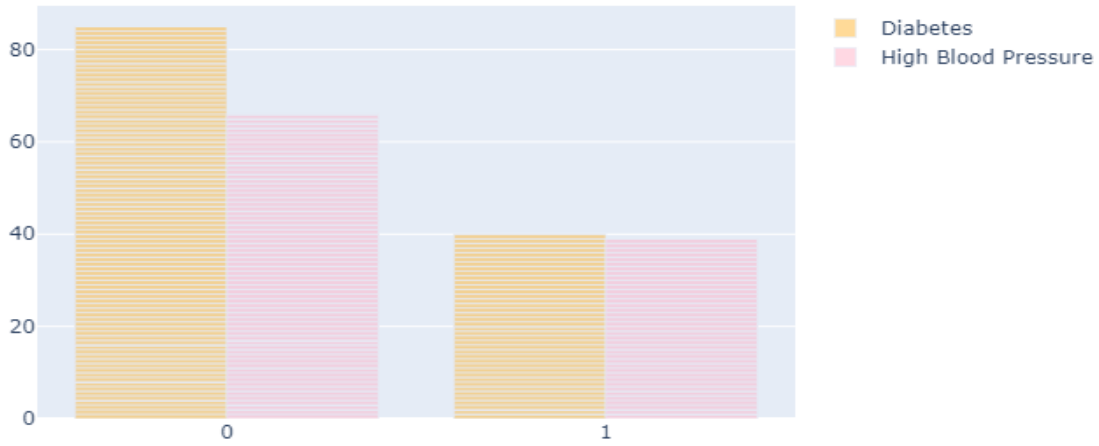
- **Survival class represent to 0.**
- **Not Survival class represent to 1.**

1. Platelets analysis



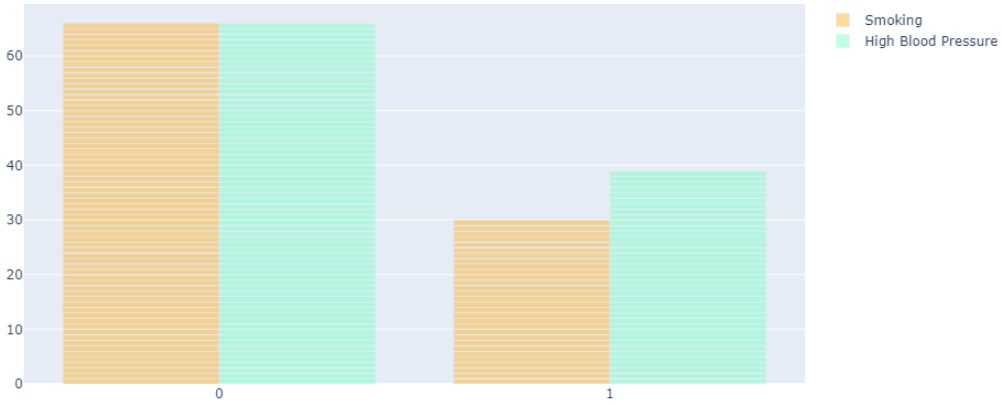
Platelets, or thrombocytes, are small, colorless cell fragments in our blood that form clots and stop or prevent bleeding. During our analysis on this feature variable, we deduced that a higher count of platelets is directly correlated with a lower chance of death rate, as can be seen from the above plot. This coincides with our domain knowledge as platelets are the first responders of the body in case a foreign bacterium enters our blood.

2. Diabetes and Blood Pressure:



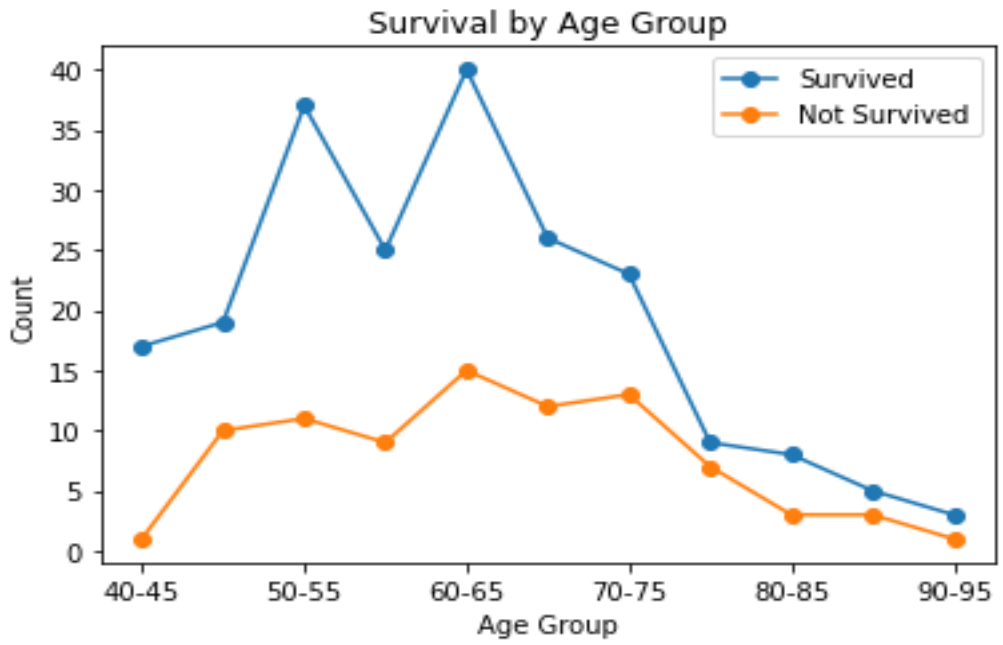
In the above plot, we analyze the feature variables 'Diabetes' and 'Blood Pressure' using a histogram to identify the correlation of each with our response – Death or No death. High blood pressure means a person is experiencing hypertension, while Diabetes may or may not be self-induced and can be a genetic Passover. We analyzed that both the features are highly correlated with our response and thus, we included them for further modelling.

3. Smoking and high Blood Pressure:



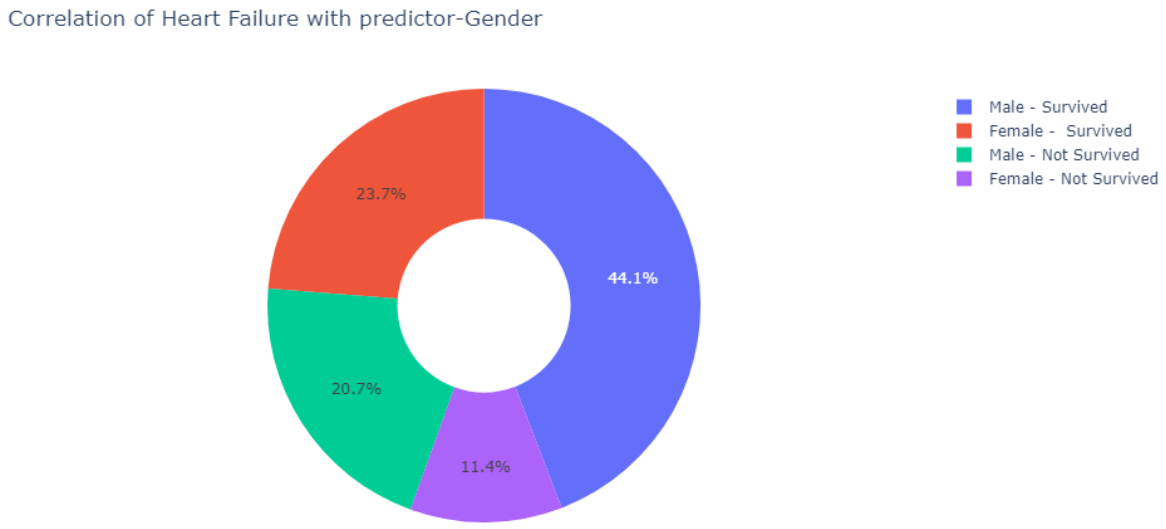
Again, we wanted to analyze the predictors “Blood pressure” and “Smoking”, with the number of deaths vs non-deaths to investigate their validity. We plotted the above histogram and found more people are susceptible to death by High BP rather than smoking and therefore we used these features in our modelling

4. Survival by age-group:



The predictor “Age” was a very important feature in our dataset. On performing data analyses with this variable, we found Cardiovascular diseases are indeed less frequent in people who are between the age range 40 – 75. This, again coincides with our domain knowledge that older people are at a higher risk of contracting cardiovascular diseases and are more prone to dying.

5. Correlation of heart-failure with predictor gender:



In the above pie-plot, we utilized with predictor "Gender" and found the survivability rate of males is dominant, compared to females by approximately 20%. While, it is also interesting to note, the death rate of Males is also higher than that of females by approximately 9%. Thus, this predictor became an important aspect in our modelling.

Variable selection:

Feature selection is a technique where we choose those features in our data that contribute most to the target variable. In other words, we choose the best predictors for the target variable.

The classes in the **sklearn.feature_selection** module can be used for feature selection/dimensionality reduction on sample sets, either to improve estimators' accuracy scores or to boost their performance on very high-dimensional datasets.

Advantage:

1. Reduces Overfitting: Less redundant data means less possibility of making decisions based on redundant data/noise.
2. Improves Accuracy: Less misleading data means modeling accuracy improves.
3. Reduces Training Time: Less data means that algorithms train faster.

After applying the sklearn we got four best predictors are Age, Time, Ejection fraction and Serum creatinine for our predictive machine learning models.

Proposed Methodology:

In this section, we first list the machine learning methods we used for the binary classification of the survival and machine learning methods we employed for the feature ranking, discarding each patient's follow-up time. We then describe the logistic regression algorithm we employed to predict survival and to perform the feature ranking as a function of the follow-up. We implemented all the methods with the open source Python programming language and made it available in the attached files.

To predict patient's survival, we employed ten different methods from different machine learning areas. The classifiers include one linear statistical method (Linear Regression), two tree-based methods (Random Forests, Decision Tree), two Support Vector Machines (linear, and with Gaussian radial kernel), one instance-based learning model (k-Nearest Neighbors).

For the feature ranking and the classification made on the top three features, we employed different sets of the machine learning methods than the ones we used for the survival prediction on the complete dataset Random Forests, Regression Trees Support Vector Machines with linear kernel, for the feature ranking, and Random Forests, and SVM with radial kernel. We decided to use three different sets of methods because we aimed to demonstrate the generalizability of our approach, by showing that our computational solution is not only valid with few machine learning classifiers, but rather works for several groups of methods.

Machine Learning Algorithms-

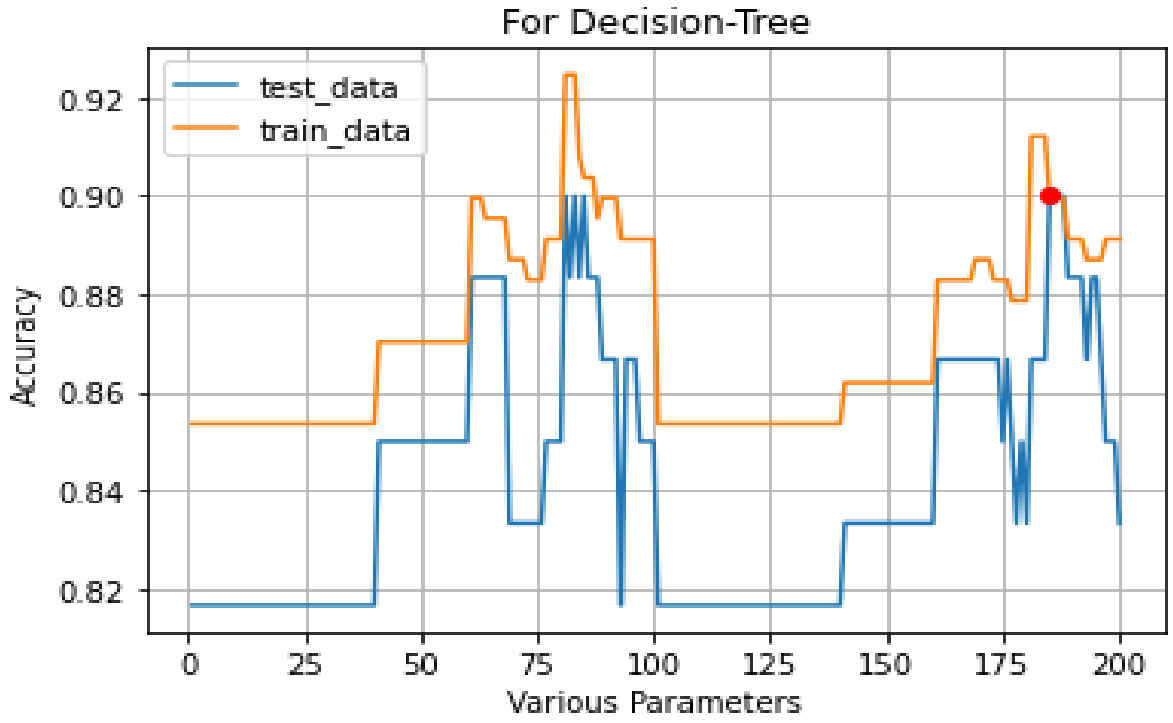
Decision Tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

We used decision tree machine learning algorithm on our data sets to classify the death event feature variable.

We did Hyper-tunning on 'criterion', 'max_depth', 'min_sample_leaf' and 'min_sample_split' parameter and best parameter values:

[criterion = entropy, max_depth = 5, min_sample_leaf = 2, min_sample_split = 2]

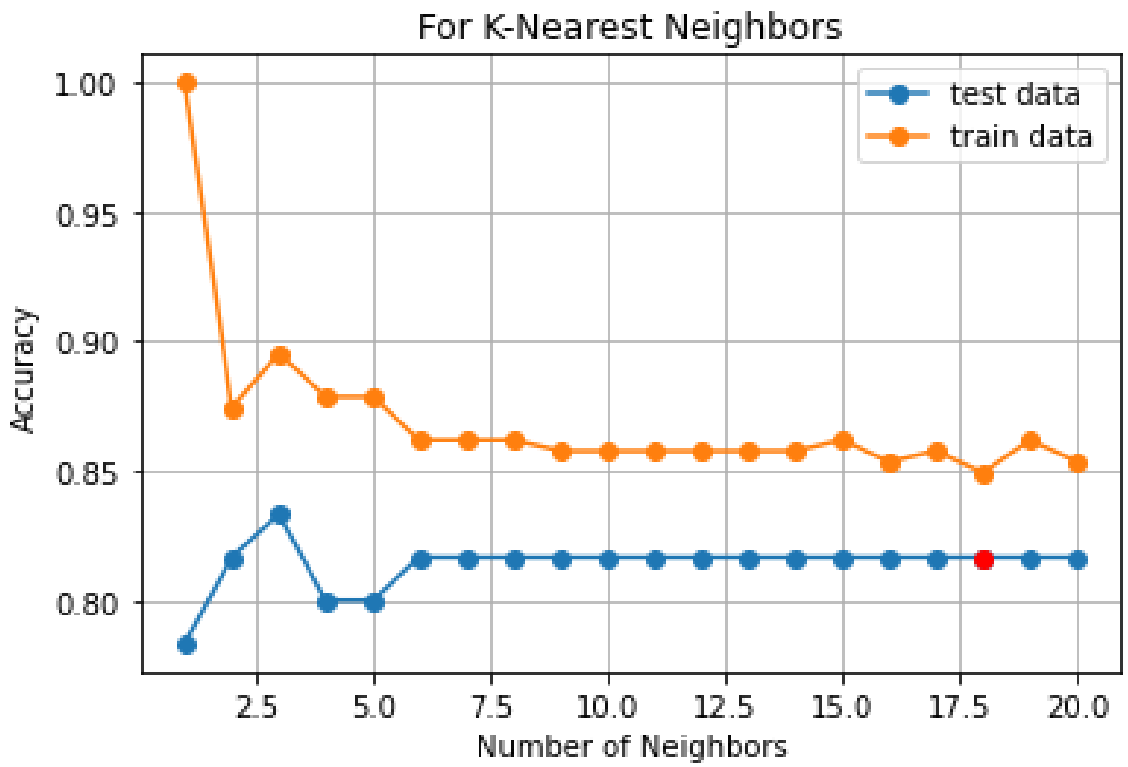


K-Nearest Neighbors Algorithm is a simple and east to implement supervised machine learning algorithm that can be used to solve both classification and regression problem. Our focus will be primarily on how does the algorithm work and how does the input parameter affect the output/prediction.

We used K-Nearest Neighbors Algorithm machine learning algorithm on our data sets to classify the death event feature variable.

We did Hyper-tunning on number of neighbors (K) parameter and best parameter values:

Best Parameter k = 18



Support Vector Machine (SVM):

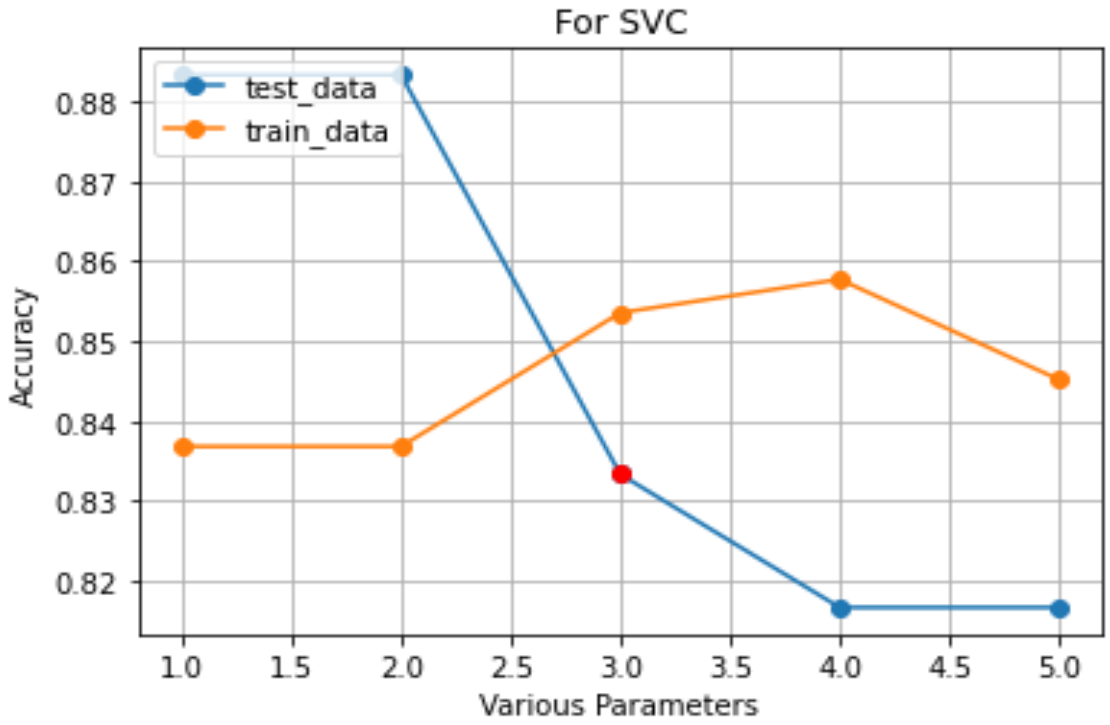
In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Developed at AT&T Bell Laboratories by Vapnik with colleagues (Boser et al., 1992, Guyon et al., 1993, Vapnik et al., 1997), it presents one of the most robust prediction methods, based on the statistical learning framework or VC theory proposed by Vapnik and Chervonenkis (1974) and Vapnik (1982, 1995). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

We used SVM Algorithm machine learning algorithm on our data sets to classify the death event feature variable.

We did Hyper-tunning on kernel and gamma and best parameter values:

kernel = scale, gamma = poly



Logistic Regression is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1".

Let us try to understand logistic regression by considering a logistic model with given parameters, then seeing how the coefficients can be estimated from data. Consider a model with two predictors x1 and x2, and one binary (Bernoulli) response variable Y. which we denote p=P(Y=1). We assume a linear relationship between the predictor variables and the log-odds (also called logit) of the event that Y=1. This linear relationship can be written in the following mathematical form (where ℓ is the log-odds, b is the base of the logarithm, and Bi are parameters of the model)

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

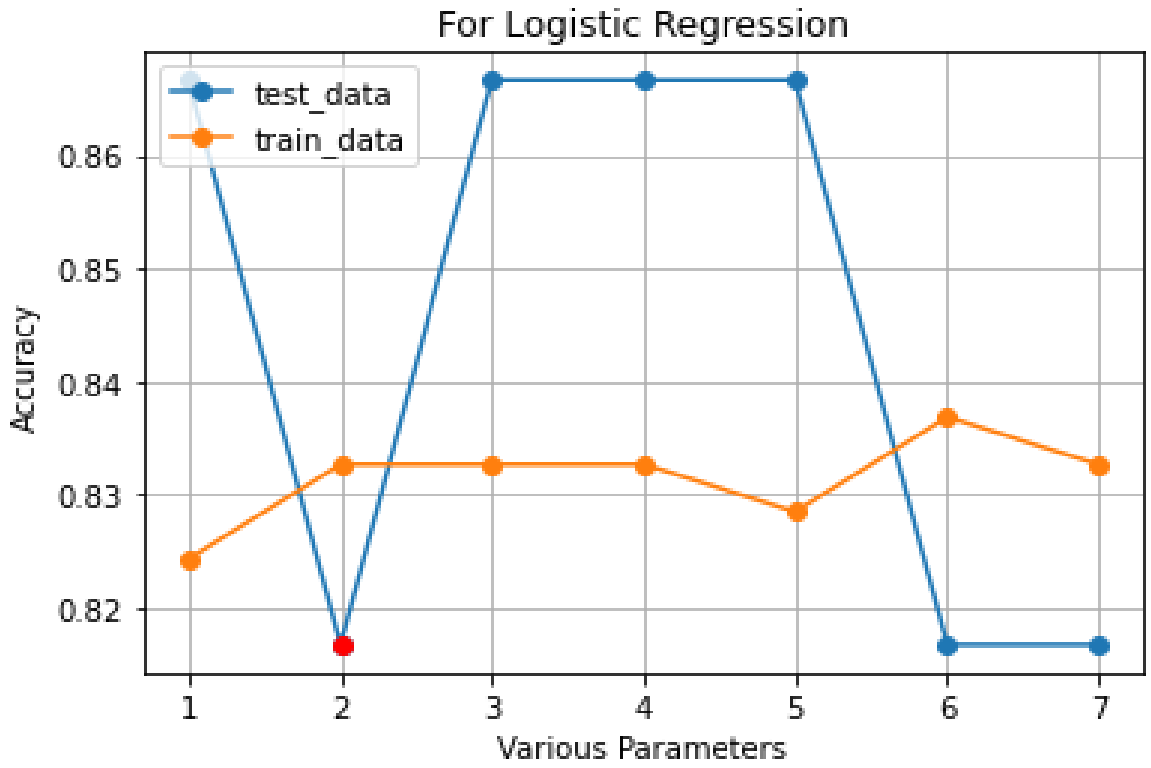
We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}.$$

We used logistic regression machine learning algorithm on our data sets to classify the death event feature variable.

We did Hyper-tunning on penalty and solver and best parameter values:

Best Parameter = [penalty = l1, solver = saga]



Random Forest or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a

small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

We used random forest machine learning algorithm on our data sets to classify the death event feature variable.

Since random forest is an ensemble learning so it is self-capable to tune its own hyper parameters.

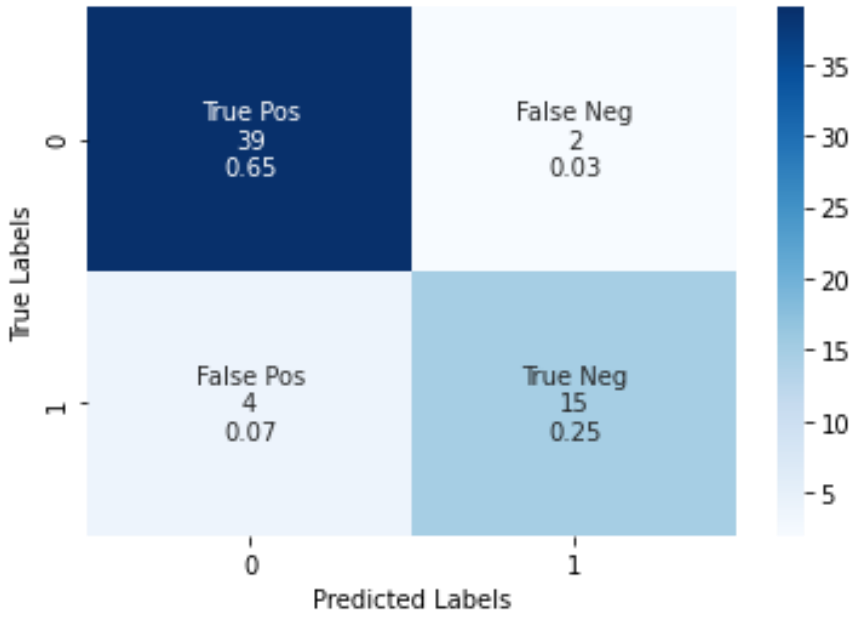
Results:

1. **Accuracy of Models:**

	Models	Accuracy
0	Decision-Tree	0.900000
1	KNN	0.816667
2	Logistic	0.816667
3	Random-Forest	0.850000
4	SVM	0.833333

Here we represent accuracy as our final metric of evaluation of our model(s). As highlighted, we achieved the best accuracy of 90% using Decision-Tree algorithm. We utilized Machine Learning algorithms to draw the best accuracy.

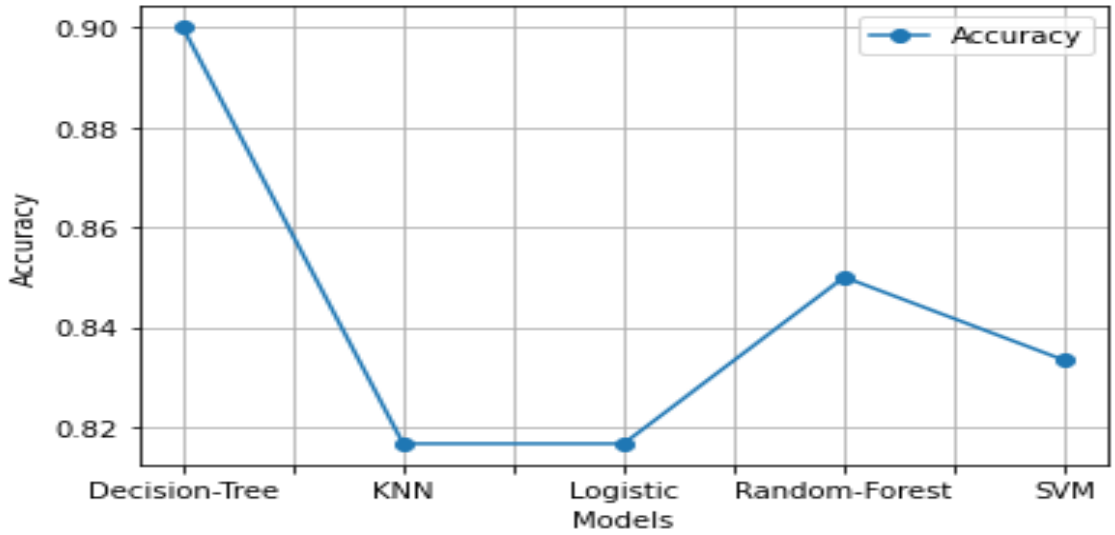
2. Confusion Matrix:



Our prediction results showed that Decision Trees outperformed all the other methods, by obtaining the top accuracy (0.900). The Decision Trees obtained the top results on the true positives (sensitivity = 0.95) and on the F1 score (0.924) and was the only classifier able to correctly predict the majority of deceased patients. The linear Support Vector Machines achieved an almost perfect prediction score with accuracy (0.833).

Basis visualization for final models:

3. Model's Accuracy Plots:



The above plot represents 'Accuracy' of various Machine Learning algorithms we used to draw the best possible model. While, KNN and Logistic models gave us approximately an accuracy of 82%, using Decision-Tree we achieved an accuracy of 90%.

4. Best Parameter for Decision Tree model:

	Parameters	test_Accuracy	train_Accuracy	difference	count
184	[entropy, 5, 2, 2]	0.9	0.899582	0.000418	185

Further, we performed Hyper-parameter tuning to perfect our model using Decision-Tree. On our initial runs, we got an accuracy above 80% We tweaked and experimented with different hyper-parameters to further reduce over-fitting and draw a better accuracy and achieved the best accuracy of 90%.

5. Predicted values for Test dataset:

	age	ejection_fraction	serum_creatinine	time	Predicted_DEATH_EVENT
273	42.0	40	0.7	245	0
151	62.0	60	0.9	117	0
116	60.0	60	0.7	94	0
282	42.0	30	3.8	250	0
27	70.0	45	1.3	26	1

As part of results, the above table gives us the predicted response on our test dataset. Some samples our shown with the predicted outcome.

Conclusions:

In our work, the fact that our traditional statistics analysis selected ejection fraction and serum creatinine as the two most relevant features confirmed the relevance of the feature ranking executed with machine learning. Moreover, our approach showed that machine learning can be used effectively for binary classification of electronic health records of patients with cardiovascular hearth diseases.

As a limitation of the present study, we have to report the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases. Also, if an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation cohort to verify our findings.

Regarding future developments, we plan to apply our machine learning approach to alternative datasets of cardiovascular heart diseases and other illnesses (cervical cancer, neuroblastoma, breast cancer, and amyotrophic lateral sclerosis).

Abbreviations:

CPK: creatinine phosphokinase

CVDs: cardiovascular diseases

EF: ejection fraction

EHR: electronic health records

GLM: generalized linear model

HF: heart failure

MCC: Matthews correlation coefficient

PCC: Pearson correlation coefficient

PR: precision-recall

RF: Random Forests

SC: serum creatinine

SVM: Support Vector Machine

TN rate: true negative rate

TP rate: true positive rate

References:

[1] Alba AC, Agoritsas T, Jankowski M, et al. Risk Prediction Models for Mortality in Ambulatory Patients with Heart Failure: A Systematic Review. Circulation. Heart failure. 2013 Sep.6(5):881–9.

[2] World Health Organization, World Heart Day.

https://www.who.int/cardiovascular_diseases/world-heart-day/en/

[3] National Heart Lung and Blood Institute (NHLBI). Heart failure.

https://www.nhlbi.nih.gov/health-topics/heart-failure

[4] Feature selection for supervised models using SelectKBest

https://www.kaggle.com/jepsds/feature-selection-using-selectkbest?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com

[5] Interpreting Random Forest Classification Models Using a Feature Contribution Method

https://link.springer.com/chapter/10.1007/978-3-319-04717-1_9

[6] Study and Analysis of Decision Tree Based Classification Algorithms

https://www.researchgate.net/profile/Purvi_Prajapati/publication/330138092_Study_and_Analysis_of_Decision_Tree_Based_Classification_Algorithms/links/5d2c4a91458515c11c3166b3/Study-and-Analysis-of-Decision-Tree-Based-Classification-Algorithms.pdf