# A Study on Survival Experiments with Machine Learning Techniques

*Project report submitted to*

**UNIVERSITY OF MADRAS**

*In partial fulfilment of the requirement*
*For the award of the degree of*

**MASTER OF SCIENCE**
**IN**
**STATISTICS**

*by*

**ARCHANA S N (32821003)**
**BHUVANESWARI B (32821004)**
**MAHARAJA G (32821025)**

*Under the guidance of*

**Dr. M. RAMADURAI**
*Assistant Professor*

**DEPARTMENT OF STATISTICS**
**UNIVERSITY OF MADRAS**
**CHENNAI-600 005**

**APRIL 2023**

DEPARTMENT OF STATISTICS
UNIVERSITY OF MADRAS
CHENNAI-600 005

**CERTIFICATE**

This is to certify that the Project Report entitled **"A STUDY ON SURVIVAL EXPERIMENTS WITH MACHINE LEARNING TECHNIQUES"** submitted in partial fulfilment of the requirement of the award of degree of MASTER OF SCIENCE in STATISTICS is a bonafide record of work done by **ARCHANA S N (32821003), BHUVANESWARI B (32821004)** and **MAHARAJA G (32821025)** under the guidance of **Dr.M.RAMADURAI** during the academic year of 2022-2023, in the Department of Statistics, University of Madras, Chennai - 600 005.

**Dr.M.RAMADURAI**
Assistant Professor
Department of Statistics,
University of Madras

**Dr.M.R.SINDHUMOL**
Associate Professor and Head(i/c)
Department of Statistics
University of Madras

Place : Chennai

Date :

# ACKNOWLEDGEMENT

We are profoundly grateful and sincerely thankful to our guide **Dr.M. RAMADURAI,** Assistant Professor, Department of Statistics, University of Madras for his unstinted support and valuable guidance throughout this project and encouraging us in times of discouragement and failures for the successful completion of this project.

We express our gratitude to **Dr. M. R. SINDHUMOL,** Associate Professor and Head(i/c), Department of Statistics, for providing us with all necessary facilities.

We wish to thank **Dr. S. SURESH,** Assistant Professor, Department of Statistics, University of Madras, for his assistance during the time of our project work.

We thank all the non-teaching staff of Department of Statistics for their timely help during our course.

We would also like to thank all the guest faculties and non-teaching staffs for their help and support.

APRIL 2023

ARCHANA S.N
BHUVANESWARI B
MAHARAJA G

## ABSTRACT

Survival analysis is concerned with studying the time between entry to a study and a subsequent event. Originally the analysis was concerned with time from treatment until death, hence the name, but survival analysis is applicable to many areas as well as mortality. The popular survival methods include Kaplan Meier, Nelson Aalen, Log Rank, Cox Proportional Hazard Model, and so on.

This article deals about analyzing the dataset containing cardiovascular medical records taken from 299 patients. All patients in the cohort were diagnosed with the systolic dysfunction of the left ventricle and had previous history of heart failures. Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. In India, the estimates range between 1.3 and 4.6 million, which translates to a prevalence of 0.12–0.44 %, although this may be underestimated. Kaplan Meier Estimate and Log Rank Test helps to estimate the survival probability of risk patients. We found that Serum Sodium, Serum Creatinine, Creatinine Phosphokinase have major impact on the survival probability of the heart failure patients.

The Cox Proportional Hazard Model acts as a link to the survival time of the individuals with covariates and it is used to predict the hazard function based on the previous records of chronic heart failure patients. The model is convenient for its flexibility and simplicity; however, it has been criticized for its restrictive proportional hazard assumption which is often violated. In order to overcome this, stratification or transformation of variables are used. We have stratified the variables based on medical criteria and the results were evaluated by using concordance index (C-index) but, this technique limit the effects of stratified variables. Machine learning methods like random survival forest, gradient boosting model pose as a solution for this problem. Also, we have divided the entire dataset into 80% for training set and the remaining 20% is considered for test set, further the evaluation was made by using C-index and found that the results of gradient boosting performs better than the random survival forest methodology.

# TABLE OF CONTENTS

# Chapter 1

# INTRODUCTION

## 1.1 SURVIVAL ANALYSIS:

This study will provide an overview for analysing the multivariate data by using survival and machine learning survival techniques. **Survival analysis** is a field of statistics that focuses on analysing the expected time until a certain event happens. Survival analysis is concerned with studying the time between entry into a study and a subsequent event. Originally the analysis was concerned with time from treatment until death, hence the name, but survival analysis is applicable to many areas as well as mortality. Survival analysis is a model for time until a certain "event." The event is sometimes, but not always, death.

The process of survival analytics can be explored through various techniques such as:

- Life tables
- Kaplan-Meier analysis
- Survivor and hazard function rates
- Cox proportional hazards regression analysis
- Parametric survival analytic models
- Survival trees
- Random survival forest
- Gradient Boosting.

In this project, we shall explore more on Kaplan-Meier analysis, Cox proportional hazard function, Survival trees, Random survival forest and Gradient Boosting.

There are three primary goals of survival analysis,

- to estimate and interpret survival and / or hazard functions from the survival data;
- to compare survival and / or hazard functions,
- to assess the relationship of explanatory variables to survival time.

### 1.1.1 Basic Terminology in Survival analysis

**Time:** By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs.

**Event:** By event, we mean variable of interest to be observed or studied.

In Survival Analysis, data can be of either exact or censored data:

I. **Exact data**: Exact data is also known as uncensored or complete data. It occurs when the precise time taken until the occurrence of the event of interest is known exactly.

II. **Censored data**: In most of the situations, survival data or survival times are frequently censored. It means the survival time of an individual is said to be censored when the end point or the variable of interest couldn't be observed for that individual is called Censored data.

**Note:**

1. In most situations, survival data are collected over a finite period of time due to practical reasons.
2. The observed time-to-event data are always non-negative and may contain either censored or uncensored observations.

**Survival Time:** The period of time taken by an individual or a patient until the occurrence of the event of interest from the start of the study is known to be the survival time of that individual.

**Study Time:** In most of the study, all the individuals are patients or not recruited or entered exactly at the same time. Normally the calendar time period fixed by the experimenter is known to be the study time.

**Patient Time:** The period of time, that a patient or an individual spends in the study, measured from that particular patients or individuals time origin is reported as patient time.

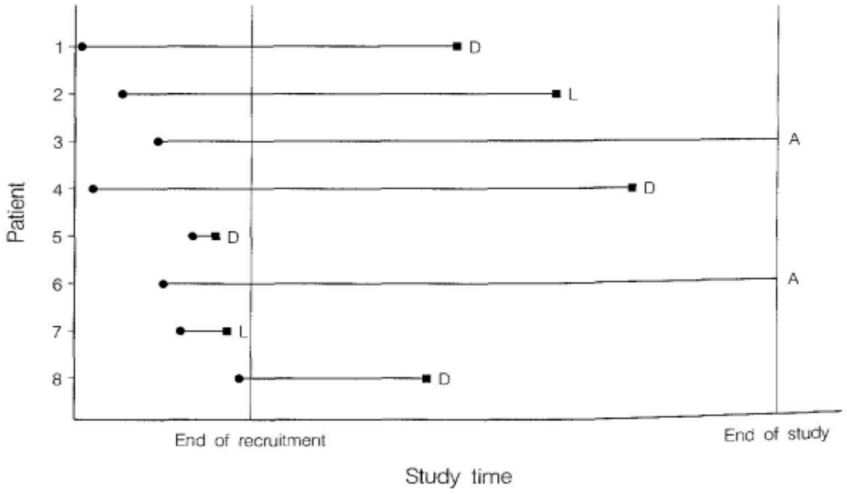



FIGURE 1.1**:** Pictorial representation of Survival time

In this study, the patients 1,4,5, and 8 are died during the course of the study, the individuals 2 and 6 are loss to follow up and the individuals 3 and 6 are still alive at the end of the observation period.

## 1.1.2 Censoring

The survival time of an individual is said to be censored, when the end point or the variable of interest could not be observed for that individual.

**Reasons for Censoring**

1. **Loss to follow up**: The patient may decide to move elsewhere, due to many reasons, such as job transfer, change of resistance and so on.
2. **Drop outs**: In clinical setup, the therapy which they are receiving may have side effects and this may lead to discontinuation from the treatment or the patient may still be alive or in contact but they may refuse to continue in the treatment.

3. **Termination of the study**: Due to some practical reasons, for example, it may be realised that the present study may be irrelevant or non-availability of fund or well experienced doctors or technicians or equipment etc.

### 1.1.3 Types of Censoring

Censored observations contain only partial information about the random variable of interest.

a) **Type I censoring / Fixed time Censoring**

Consider n items and observe their failure time up to a fixed time say $t_c$ .Suppose $T_1, T_2, \dots, T_n$are the failure times of these n units respectively. Then the observed values are

$$Y_i = \begin{cases} T_i & ; T_i < t_c \\ t_c & ; T_i > t_c \end{cases} \qquad i = 1,2, \dots, n$$

**b) Type II censoring / Fixed number of Censoring**

Let $T_{(1)}, T_{(2)}, \dots, T(n)$ be the order statistic of $T_1, T_2, \dots, T_n$.In this Type II censoring the process of observations ceases immediately after the occurrence of the $r^{th}$ failure, where $r \leq n$. so that we can observe $T_{(1)}, T_{(2)}, \dots, T(n)$. If $Y_1, Y_2, \dots, Y_n$ are the observations then

$$Y_i = \begin{cases} T_{(i)} & ; i = 1,2, \dots, r \\ T_{(r)} & ; r = r+1, r+2, \dots, n \end{cases}$$

c) **Type III censoring / Random censoring**

Let $T_1, T_2, \dots, T_n$ be identically and independently distributed random variable with distribution function F and let $C_1, C_2, \dots, C_n$ be identically and independently distributed random variable with distribution function G, where $T_i$, Survival time of the $i^{th}$ unit, where $i = 1,2, \dots, n$ and $C_i$ be the censoring time of the $i^{th}$ unit, where $i = 1,2, \dots, n$ . Assume that $T_i's$ and $C_i's$ are independent. In this process, we are observing $(Y_i; \delta_i);\ i = 1,2, \dots, n$

Where, $Y_i = Min(T_i, C_i)$ and $\delta_i = \begin{cases} 1 & ; if T_i \leq C_i \\ 0 & ; if T_i > C_i \end{cases}$

**Other types of Censoring**

a. **Left Censoring:** Left censoring is not common in clinical trials; an observation is left censored if the event of interest has already occurred when observation of time begins or left censored data can occur when a subject's survival time is incomplete at the beginning of the follow up period.
b. **Right Censoring:** Right censoring occurs when the subject leaves the study before an event occurs or the study ends before the event has occurred.

## 1.2 SURVIVAL AND HAZARD FUNCTIONS:

The following are the two functions used for summarizing the survival data and they are survival function and the hazard rate/function.

### 1.2.1 Survival function

The survival function is a function that gives the probability that a patient, device, or other object of interest will survive from past till a certain time. In other words, the survival

probability, also known as the survivor function S(t), is the probability that an individual survives from the time origin (e.g., diagnosis of heart failure) to a specified future time t.
Let the lifetime T be a continuous random variable with cumulative distribution function F(t) on the interval [0, ∞). Its survival function or reliability function is:

$$\mathrm{S(t)} = \Pr(\mathrm{T} > \mathrm{t}) = 1 - \mathrm{F(t)}.$$

### 1.2.2 Hazard function

The hazard function $h(t)$, is the probability that an individual who is under observation at a time, t has an event at that time and the probability that the random variable associated with an individual's survival time T lies between t and $t + \delta t$, conditional on $T \geq t$ written $P(t \leq T < t + \delta t | T \geq t)$, as $\delta t \to 0$.

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}$$

i.e., $h(t)$ represents the probability that an individual dies immediately after time t conditional on his/her survived till time t.

Then the hazard function (also known as the failure rate, hazard rate, or force of mortality) and is denoted *h(t)* is defined as the ratio of the probability density function f(t) to the survival function S(t), given by

$$\mathrm{h(t)} = \left( \frac{\mathrm{f(t)}}{\mathrm{S(t)}} \right) = \left( \frac{\mathrm{f(t)}}{1 - \mathrm{F(t)}} \right)$$

where F(t) is the distribution function.

### 1.2.3 Relationship between survival and hazard functions

The survival function is related to the hazard function as follows:

- $S(t) = e^{\{-H(t)\}}$

  where H(t) is the cumulative hazard function; $H(t) = \int_0^t h(u)du$

- $h(t) = -\frac{S'(t)}{S(t)}$
- $h(t) = -\frac{d}{dt}\{logS(t)\}$

## 1.3 NON-PARAMETRIC TECHNIQUE IN SURVIVAL ANALYSIS

We know that non-parametric methods do not require any specific assumptions about the underlying distributions of the survival time. Here, in general there are three non-parametric procedures namely,

1. Life table estimate
2. Nelson-Aalen estimate
3. Kaplan-Meier estimate and
   the general non-parametric testing procedure namely Log rank test.

### 1.3.1 Life Table Estimate

The life table method also known as the actuarial estimate is one of the basic tools in the description of the mortality experience of the population and the method is due to E. Halley (1693). Life tables are particularly suited for analysing the large dataset or group who have survived to the end of each time interval.

### 1.3.2 Nelson Aalen Estimate

The Nelson-Aalen analysis method belongs to the descriptive methods for survival analysis such as life table analysis and Kaplan-Meier analysis. The Nelson-Aalen approach can quickly give you a curve of cumulative hazard and estimate the hazard functions based on irregular time intervals.

In this study we are going to discuss about Kaplan-Meier Estimate and Log rank test.

### 1.3.3 Kaplan Meier Estimate

The non-parametric estimation of the survival function from incomplete or censored data was first considered by Kaplan Meier (1958) and this estimator is also known as Product limit estimator. Kaplan Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. Kaplan-Meier estimate is the simplest way of computing the survival over time in spite of all these difficulties associated with subjects or situations. The survival curve can be created assuming various situations. It involves computing of probabilities of occurrence of event at a certain point of time and multiplying these successive probabilities by the earlier computed probabilities to get the final estimate.

**Definition:**

The Kaplan–Meier estimator, also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. The survival function of the Kaplan-Meier estimate is given by,

$$\hat{S}(t) = \prod_{j=1}^{k} \left( \frac{n_j - d_j}{n_j} \right)$$

for $t_{(k)} \leq t \leq t_{(k+1)}; k = 1,2, \dots, r$ with $\hat{S}(t) = 1$ for $\hat{S}(t_{r+1}) = \infty$

where $n_j = n_{j-1} - \left[c_{j-1} + d_{j-1}\right]$. $t_1, t_2, \dots, t_n$ be the survival time of the individuals under study and r death times where $r \leq n$, the order death times are $t_{(1)}, t_{(2)}, \dots, t_{(r)}$. $n_j; j = 1,2, \dots, r$ be the number of individuals who are alive just before the time $t_{(j)}$ and $d_j$ be the number of individuals who die at time $t_{(j)}$ . For each time interval, the survival probability is calculated as the number of subjects surviving divided by the number of patients at risk.

Subjects who have died, dropped out, or move out are not counted as “at risk” i.e., subjects who have lost are considered “censored” and are not counted in the denominator.

The following are some examples of survival studies that Kaplan-Meier estimate can be applicable, which include death times of kidney transplant patients, times to infection for burn patients and times to death for a breast-cancer trial.

There are three assumptions used in Kaplan Meier analysis.

- At any time, patients who are censored have the same survival prospects as those who continue to be followed.
- The survival probabilities are the same for subjects recruited early and late in the study.
- The event happens at the time specified.

### 1.3.4 Log Rank Test

The log rank test is a popular test to test the null hypothesis of no difference in survival between two or more independent groups. The test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical (overlapping) or not.

$$H_0: S_1(t) = S_2(t)$$

$$H_1: S_1(t) \neq S_2(t)$$

The log rank test is based on the same assumptions as the Kaplan-Meier survival curve namely, that censoring is unrelated to prognosis, the survival probabilities are the same for subjects recruited early and late in the study, and the events happened at the times specified. Deviations from these assumptions matter most if they are satisfied differently in the groups being compared, for example if censoring is more likely in one group than another.

The log rank test is most likely to detect a difference between groups, when the risk of an event is consistently greater for one group than another. It is unlikely to detect a difference when survival curves cross, as can happen when comparing a medical with a surgical intervention. When analysing the survival data, the survival curves should always be plotted. Because the log rank test is purely a test of significance it cannot provide an estimate of the size of the difference between the groups or a confidence interval.

### 1.3.3 What the Kaplan–Meier method and the Log-Rank Test can and cannot do?

The Kaplan–Meier method is the most popular method used for survival analysis. Together with, the log-rank test, it may provide us with an opportunity to estimate survival probabilities and to compare survival between groups. Most of the time, however, one would like to do more than that.

Let us consider an example for this case, where the survival of heart failure patient with the various other factor groups i.e., categorical in nature, one would have liked to be informed on the size of potential difference. In addition, to make a fairer comparison between the groups about the survival probabilities.

Some of the drawbacks of using the Kaplan-Meier and log rank test is as follows. The log-rank test is purely a significance test, it cannot provide an estimate of the size of the difference between groups and a related confidence interval. Secondly, the Kaplan–Meier method and the log-rank test can only study the effect of one factor at the time, and therefore they cannot be used for multivariate analysis. For these purposes, one may use a regression technique like the Cox proportional hazards model, which will be described in the next section.

## 1.4 SEMI-PARAMETRIC TECHNIQUE FOR SURVIVAL ANALYSIS

In most of the studies, the supplementary information will also be recorded for each individual in the study and the supplementary information is also called as the covariates or explanatory variables in one of the objectives of survival analysis is to describe the survival experiences of the individual and possibly to access whether survival is associated with explanatory variables. The statistical modelling approach is used to explore the relationship between the survival experience of a subject and explanatory variables.

Hence, it is used to work with the survival function for descriptive analyses and hazard function for accessing the relationship between explanatory variable and survival time. The survival methods for modelling may be divided into two broad categories

1. Proportional Hazard Model Approach
2. Accelerated Failure Time Model Approach

In Cox Proportional Hazard Model, we can identify the relationship between selected explanatory variables and the survival time, it is one of the most popular regression techniques for survival analysis, which is used to relate several risk factor's or exposures considered simultaneously to survival time. An Accelerated Failure Time (AFT) model is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant.

### 1.4.1 Cox Proportional Hazard Model

The principle of the Cox proportional hazards model is to link the survival time of an individual to covariates. For example, in the medical domain, we are seeking to find out which covariate has the major impact on the survival time of a patient.

The Cox model is expressed by the hazard function denoted by *h(t)*. Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$$h_i(t) = h_0(t) * e^{(b_1x_1+b_2x_2+\ldots+b_px_p)}$$

where,

- $t$ - survival time
- $h_i$(t) - hazard function determined by a set of p covariates $(x_1, x_2, \ldots, x_p)$
- $(b_1, b_2, \ldots, b_p)$ - measures the impact (i.e., the effect size) of covariates.

- $h_0$ - is called the baseline hazard. It corresponds to the value of the hazard if all the xi are equal to zero (the quantity $e^0$ equals 1). The 't' in h(t) reminds us that the hazard may vary over time.

### 1.4.2 What is Hazard?

Hazard is essentially the inverse of survival, or the probability of failure (death event). It basically represents the slope of the survival curve — a measure of how rapidly subjects is dying.

### Hazard Ratio

Hazard Ratio (HR) is the probability of an event in a treatment group relative to the control group over a unit of time. This ratio is an effect size measure for time-to-event data. The hazard ratio can be used to estimate the treatment effect in clinical trials when we want to assess time-to-event data. For example, HRs can determine whether a medical treatment reduces the duration of symptoms or prolongs survival in cancer patients.

- **Hazard Ratio = 1**: An HR equals one when the numerator and denominator are equal. This equivalence occurs when both groups experience the same number of events in a period.

- **Hazard Ratio > 1**: The numerator is greater than the denominator in the hazard ratio. Therefore, the treatment group experiences a higher event probability within any given period than the control group.

- **Hazard Ratio < 1**: The numerator is less than the denominator in the HR. Consequently, the treatment group experiences a lower event probability during a unit of time than the control group.

### 1.4.3 Proportional Hazard Assumptions

The Cox model assumes the following assumptions:

- All individuals or things in the data set experience the same baseline hazard rate.
- The regression variables X do not change with time.
- The regression coefficients β do not change with time.

### What to do if proportional hazard assumption fails?

- Stratification
- Modify the functional form
- Bin variable and stratify on it
- Introduce time-varying covariates

## 1.5 MACHINE LEARNING SURVIVAL MODELS

Machine learning techniques that inherently handle high-dimensional data have been adapted to handle censored data, allowing machine learning to offer more flexible alternatives for analyzing high-dimensional, right-censored, heterogeneous data. This flexibility is expected to lead to more accurate predictions. The algorithms used in the survival analysis are modifications of algorithms known from classification or regression, appropriately adapted to the censored data.

## 1.6 RANDOM SURVIVAL FOREST

### Why Random Survival Forest Model?

The Cox-proportional hazards model is a popular choice for analysis of right censored time-to-event data. The model is convenient for its flexibility and simplicity; however, it has been criticised for its restrictive proportional hazards (PH) assumption which is often violated. A number of extensions to the Cox proportional hazards model to handle time-to-event data where the PH assumption is not met have been suggested and implemented. Other approaches to handle non-proportional hazards include methods such as stratification, but these limit the ability to estimate the effect(s) of the stratification variable(s). Thus, an alternative method shall be used for analysis i.e., Random Survival Forest and conditional inference survival forest model. In this study we are discussing the Random Survival Forest.

### 1.6.1 Tree based Model – Random Survival Forest

Survival trees and Random Survival Forests (RSF) are an alternative approach to the Cox proportional hazards models when the PH assumption is violated. These methods are the extensions of classification and regression trees and Random Forests (RF) for time-to-event data.

Survival tree methods are fully non-parametric, flexible, and can easily handle high dimensional covariate data. Drawbacks of random survival forest includes a bias towards inclusion of variables with many split points. This effect leads to a bias in resulting summary estimates such as variable importance.

A Random Survival Forest (RSF) is an assemble of trees method for analysis of right censored time-to-event data and an extension of Brineman's random forest method. Survival trees and forests are popular non-parametric alternatives to (semi) parametric models for time-to-event analysis. They offer great flexibility and can automatically detect certain types of interactions without the need to specify them beforehand. A survival tree is built with the idea of partitioning the covariate space recursively to form groups of subjects who are similar according to the time-to-event outcome. Homogeneity at a node is achieved by minimizing a given impurity measure. The basic approach for building a survival tree is by using a binary split on a single predictor.

For a categorical covariate $X$, a single split is defined as $X \leq c$ where $c$ is some constant and with many split-points, the potential split is $X \in \{c_1, \dots, c_k\}$ where $c_1, \dots, c_k$ are potential split values of a predictor variable $X$. The goal in survival tree building is to identify prognostic factors that are predictive of the time-to-event outcome. In tree building, a binary split is such that the two daughter nodes obtained from the parent node are dissimilar and several split-rules

(different impurity measure) for time-to-event data have been suggested over the years. The impurity measure or the split-rule of the algorithm is very important in survival tree building. In this article, we used the log-rank and the log-rank score split-rules.

### 1.6.2 The log-rank split-rule

Suppose a node $h$ can be split into two daughter nodes $\alpha$ and $\beta$. The best split at a node $h$, on a covariate $x$ at a split point $s^*$ is the one that gives the largest log-rank statistic between the two daughter nodes. The algorithm for building a survival tree using the split-rule based on the log-rank statistic is given in Algorithm 1 below.

**Algorithm 1:** The Log-rank Survival Tree Algorithm

1. At each node randomly select $\sqrt{p}$ covariates from p covariates as candidates for splitting the node into two daughter nodes.
2. At a node h, compute the log-rank statistic impurity measure defined above for daughter nodes $\alpha \; and \; \beta$ formed by all possible splits on all covariates considered for splitting at the node.
3. Choose the covariate that has the largest significant log rank statistics calculated from one of the daughter nodes created by the splits. Partition the node into two daughter nodes based on the values of the covariate obtained from the split with the largest statistic.
4. Recursively repeat steps 2 and 3 by treating each daughter node as a root node.
5. The node is terminal if it has no less than $d_0 > 0$ unique observed events.

### 1.6.3 The log-rank score split-rule

The log-rank score split-rule is a modification of the log-rank split-rule mentioned above. It uses the log-rank scores. Given $r = (r_1, r_2, \dots, r_N)$,the rank vector of survival times with their indicator variable $(T, \delta) = ((T_1, \delta_1), (T_2, \delta_2), \dots, (T_N, \delta_N))$ and that $a = a(T, \delta) = (a_1(r), a_2(r), \dots, a_N(r))$ denotes the score vector depending on ranks in vector $r$. Assume that the ranks order the predictor variables in such a way that $x_1 < x_2 < \cdots x_N$. The log-rank scores for an observation at $T_l$ is given by:

$$a_l = a_l(T, \delta) = \delta_1 - \sum_{k=1}^{\gamma_t(T)} \frac{\delta_k}{N - \gamma_K(T) + 1}$$

Where, $\gamma_k(T) = \sum_{l=1}^{N} X\{T_l \leq T_k\}$ is the number of individuals that have had the event of interest or were censored before or at time $T_k$.

$$i(x, s^*) = \frac{\sum_{x_j \leq s*}(a_j - R_1 \, \overline{a})}{\sqrt{R_1 \left(1 - \frac{R_1}{N}\right) S_a^2}}$$

Where, $\bar{a} \; and \; S_a^2$ are the mean and sample variance of the scores $a_j; j = 1,2, \dots, n$. The best split is the one that maximizes $|i(x, s*)|$ over all $x_j's$ and possible splits s*. Trees are generally unstable and hence researchers have recommended the growing of a collection of trees, commonly referred to as random survival forests.

### 1.6.4 Random Survival Forests (RSF) algorithm

The random survival forests algorithm implementation is shown in Algorithm 2.

**Algorithm 2:** Random Survival Forest Algorithm

Draw B bootstrap samples from the original dataset. Each bootstrap samples excludes about 30% of the data and this is called out-of-bag (OOB) data.

1. Grow a survival tree for each bootstrap sample. At each node randomly select $\sqrt{p}$ variables. Split the node by selecting the variable that maximizes the difference between daughter nodes using a predetermined split rule.
2. Grow the tree to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique events.
3. Calculate the Cumulative Hazard (CH) for each tree. Average to obtain the ensemble prediction
4. Using OOB data, calculate prediction error curves for the ensemble cumulative hazard.

For this study, we used the log-rank split-rules in Step 2 of the algorithm. The random survival forests algorithm, has been criticised for having a bias towards selecting variables with many split points and the conditional inference forest algorithm has been identified as a method to reduce this selection bias. Conditional inference forests are formulated in such a way that it separates the algorithm for selecting the best splitting covariate is separated from the algorithm for selecting the best split point. To illustrate this, consider a dataset with a time-to-event outcome variable $T$ and two explanatory variables $x_1$ and $x_2$ with $k_1$ and $k_2$ possible split-points, respectively. Furthermore, consider that $T$ is independent of $x_1$ and $x_2$, and that $k_1 < k_2$. In the random survival forests algorithm, the search for the best covariate to split on and the best split-point by comparing the effect for both the covariates on $T$, gives $x_2$ the highest probability of being selected just by chance.

Survival trees is similar to decision tree which is built by recursive splitting of tree nodes. A node of a survival tree is considered "pure" if all the patients in the node survive for an identical span of time. The log rank test is most commonly used dissimilarity measure that estimates the survival difference between two groups. For each node, examine every possible split on each feature, and then select the best split, which maximizes the survival difference between two children's nodes.

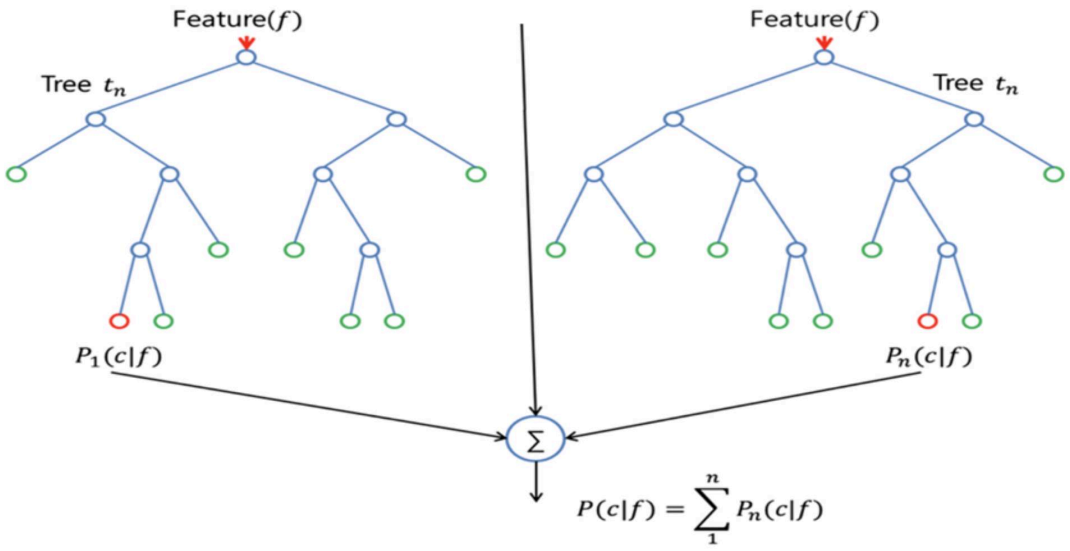



Figure 1.2: Splitting of tree nodes

## 1.7 GRADIENT BOOSTING

### Why Gradient Boosting Models (GBM)?

Many statistical methods have been developed for survival analysis. In particular, Cox proportional hazards model, one of the most prevalent models in survival analysis, assumes that different covariates contribute multiplicatively to the hazard function. But not always those assumptions posed under CPH model will be satisfied, so we relax the proportional hazards assumption and allow for more complicated relationships between covariates, parametric models based on artificial neural networks (ANN) and ensembles of tree models based on boosting. In order to handle the censored data, all these models use an approximation of the likelihood function, called the Cox partial likelihood, to train the predictive model. The partial likelihood function is computationally convenient to use.

### 1.7.1 Gradient Boosting Machine

### Definition

The gradient boosting machine (GBM) is an ensemble learning method, which constructs a predictive model by additive expansion of sequentially fitted weak learners. The general problem is to learn a functional mapping $y = F(x;\beta)$ from data $\{x_i, y_i\}_{i=1}^n$. Gradient Boosting Models does not assume any functional form of $F$ but uses additive expansion to build up the model. This nonparametric approach gives more freedom to researchers. GBM combines predictions from the ensemble of weak learners and so tends to yield more robust results than the single learner.

### 1.7.2 Loss Function

### Cox's Partial Likelihood

The loss function can be specified via the loss argument loss; the default loss function is the partial likelihood loss of Cox's proportional hazards model. Therefore, the objective is

to maximize the log partial likelihood function, but replacing the traditional linear model $X^T\beta$, with the additive model $f(x)$:

$$\arg\min_{f} \quad \sum_{i=1}^{n} \delta_i \left[ f(\mathbf{x}_i) - \log \left( \sum_{j \in \mathcal{R}_i} \exp(f(\mathbf{x}_j)) \right) \right].$$

## OBJECTIVE OF THIS PROJECT WORK

To analyse and interpret the survival data using the suitable survival technique and to explore the use of machine learning algorithms.

In this dissertation, Chapter 2 deals with the brief description of the data considered in our study. Chapter 3 deals with the analysis of the models in chapter 1 and its interpretation. Chapter 4 deals with the overall conclusion of our study.

# Chapter 2

# HEART FAILURE DATASET

## 2.1 HEART FAILURE

Heart Failure (HF) is the state in which muscles in the heart wall fades and enlarges, limit the heart's tendency of pumping blood. The ventricles of heart can get inflexible and do not fill properly between beats. With the passage of time, heart fails in fulfilling the proper demand of blood in body and as a consequence person starts feeling difficulty in breathing. The main reason behind heart failure includes coronary heart disease, diabetes, high blood pressure and other diseases like HIV, alcohol abuse or cocaine, radiation or chemotherapy, etc. **As stated by WHO Cardiovascular Heart Disease (CHD) is now top reason causing 31% of deaths globally.**

## 2.2 HEART FAILURE AROUND THE WORLD:

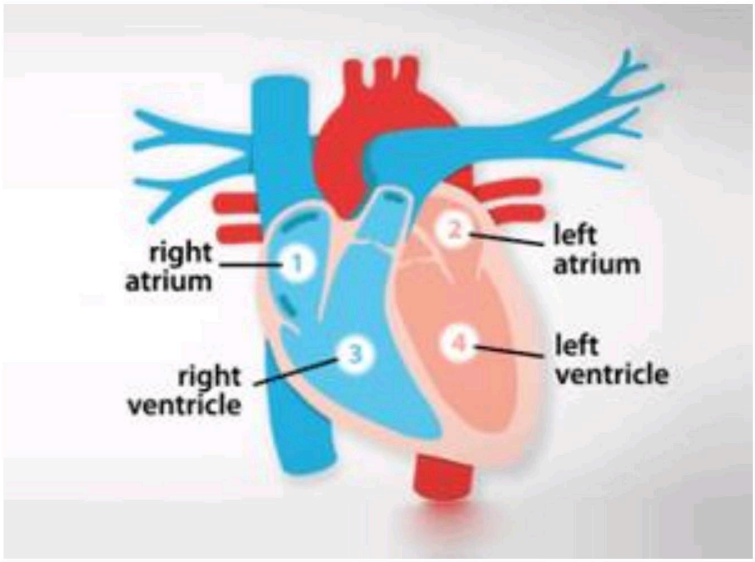
Heart Failure (HF) is a global pandemic affecting at least 26 million people worldwide and is increasing in prevalence. HF health expenditures are considerable and will increase dramatically with an ageing population. Despite the significant advances in therapies and prevention, mortality and morbidity are still high and the quality of life is poor. The prevalence, incidence, mortality and morbidity rates reported show geographic variations, depending on the different aetiologies and clinical characteristics observed among patients with HF. Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. **In India the estimates range between 1.3 and 4.6 million, which translates to a prevalence of 0.12–0.44 %, although this may be underestimated.** According to the last searchable Global Health Data Exchange (GHDx) period (i.e., year 2017), the current worldwide prevalence of HF is estimated at 64.34 million cases (8.52 per 1,000 inhabitants, 29% of which mild, 19% moderate and 51% severe HF), accounting for 9.91 million YLDs (years of years of healthy life lost due to disability) (11.61 per 1,000 YLDs).

## 2.3 NORMAL HEART FUNCTION:

The normal healthy heart is a strong, muscular pump, a little larger than a fist. It pumps blood continuously through the circulatory system. The heart has four chambers, two on the right and two on the left:

- Two upper chambers called atria (one is called an atrium)
- Two lower chambers called ventricles

The right atrium takes in oxygen-depleted blood from the rest of the body and sends it through the right ventricle where the blood becomes oxygenated in the lungs. Oxygen-rich blood travels from the lungs to the left atrium, then on to the left ventricle, which pumps it to the rest of the body. For the heart to function properly, the four chambers must beat in an organized way, if this rhythm gets disturbed then many complications will start to occur.




## 2.4 HEART FAILURE CONDITION:

Heart failure is a chronic, progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen. Heart failure can involve the heart's left side, right side or both sides. However, it usually affects the left side first. Basically, the heart can't keep up with its workload. This results in fatigue and shortness of breath and some people have coughing. Everyday activities such as walking, climbing stairs or carrying groceries can become very difficult. Our body takes temporary measures to mask the problem of heart failure, but they don't solve it completely.

### Symptoms of Heart Failure

- Shortness of breath with activity or when lying down.
- Fatigue and weakness.
- Swelling in the legs, ankles and feet.
- Rapid or irregular heartbeat.
- Reduced ability to exercise.
- Persistent cough or wheezing with white or pink blood-tinged mucus.
- Swelling of the belly area (abdomen)

## 2.5 DATASET DESCRIPTION

The dataset used for this study is taken from the Faisalabad Institute of Cardiology. The dataset contains cardiovascular medical records taken from 299 patients. The patient cohort comprised of 105 women and 194 men between 40 and 95 years in age. All patients in the cohort were diagnosed with the systolic dysfunction of the left ventricle and had previous history of heart failures. As a result of their previous history every patient was classified into either class III or class IV of New York Heart Association (NYHA) classification. The data contains no missing values and the datatypes are integer and float values.

### 2.5.1 Variable Description:

The dataset contains 13 features namely, age, anaemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, sex, platelets, serum creatinine, serum sodium, smoking, time, death event. Out of the above mentioned variables, 6 variables such as anaemia, high blood pressure, diabetes, sex, smoking, death event are qualitative variables. Similarly, the rest of the variables like, age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, time are quantitative variables.

Table 2.1 Feature Description

| Feature | Explanation | Measurement | Range |
|---|---|---|---|
| Age | Age of the patient | Years | [40 -- 95] |
| Anaemia | Decrease of red blood cells or haemoglobin level in the blood. | Binary | 0, 1 |
| High Blood Pressure | Indicative of whether the patient has hypertension. | Binary | 0, 1 |
| Creatinine Phosphokinase | Level of the CPK enzyme in the blood | mcg/L | [23 -- 7861] |
| Diabetes | If the patient is diabetic | Binary | 0, 1 |
| Ejection Fraction | Percentage of blood leaving the heart at each contraction | Percentage | [14 -- 80] |
| Sex | Gender of the patient. | 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 | Indicates if the patient has smoking habit. | Binary | 0, 1 |
| Time | Follow-up period for the next doctor visit. | Days | [4 -- 285] |
| Death Event | If the patient died during the follow-up period. | Binary | 0, 1 |

### 2.5.2 About the dataset

The dataset containing the medical records of 299 heart failure patients.

Table 2.2 Dataset

| Si.No | age | anaemia | creatinine phosphokinase | diabetes | ejection fraction | high blood pressure | platelets | serum creatinine | serum sodium | sex | smoking | time | DEATH EVENT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 75 | 0 | 582 | 0 | 20 | 1 | 265000 | 1.9 | 130 | 1 | 0 | 4 | 1 |
| 2 | 55 | 0 | 7861 | 0 | 38 | 0 | 263358.03 | 1.1 | 136 | 1 | 0 | 6 | 1 |
| 3 | 65 | 0 | 146 | 0 | 20 | 0 | 162000 | 1.3 | 129 | 1 | 1 | 7 | 1 |
| 4 | 50 | 1 | 111 | 0 | 20 | 0 | 210000 | 1.9 | 137 | 1 | 0 | 7 | 1 |
| 5 | 65 | 1 | 160 | 1 | 20 | 0 | 327000 | 2.7 | 116 | 0 | 0 | 8 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 295 | 62 | 0 | 61 | 1 | 38 | 1 | 155000 | 1.1 | 143 | 1 | 1 | 270 | 0 |
| 296 | 55 | 0 | 1820 | 0 | 38 | 0 | 270000 | 1.2 | 139 | 0 | 0 | 271 | 0 |
| 297 | 45 | 0 | 2060 | 1 | 60 | 0 | 742000 | 0.8 | 138 | 0 | 0 | 278 | 0 |
| 298 | 45 | 0 | 2413 | 0 | 38 | 0 | 140000 | 1.4 | 140 | 1 | 1 | 280 | 0 |
| 299 | 50 | 0 | 196 | 0 | 45 | 0 | 395000 | 1.6 | 136 | 1 | 1 | 285 | 0 |

### 2.5.3 Data source

The dataset is downloaded from Kaggle.https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?resource=download&select=heart_failure_clinical_records_dataset.csv

### 2.5.4 Summary of the data:

Table 2.3 Quantitative Variables

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| age | 299.0 | 60.833893 | 11.894809 | 40.0 | 51.0 | 60.0 | 70.0 | 95.0 |
| creatinine_phosphokinase | 299.0 | 581.839465 | 970.287881 | 23.0 | 116.5 | 250.0 | 582.0 | 7861.0 |
| ejection_fraction | 299.0 | 38.083612 | 11.834841 | 14.0 | 30.0 | 38.0 | 45.0 | 80.0 |
| platelets | 299.0 | 263358.029264 | 97804.236869 | 25100.0 | 212500.0 | 262000.0 | 303500.0 | 850000.0 |
| serum_creatinine | 299.0 | 1.393880 | 1.034510 | 0.5 | 0.9 | 1.1 | 1.4 | 9.4 |
| serum_sodium | 299.0 | 136.625418 | 4.412477 | 113.0 | 134.0 | 137.0 | 140.0 | 148.0 |
| time | 299.0 | 130.260870 | 77.614208 | 4.0 | 73.0 | 115.0 | 203.0 | 285.0 |

Table 2.4 Qualitative Variables

| Variables | Division | Counts | Percentage |
|---|---|---|---|
| Anaemia | 1 (Yes) | 129 | 43.1% |
| | 0 (No) | 170 | 56.9% |
| Diabetes | 1 (Yes) | 125 | 41.8% |
| | 0 (No) | 174 | 58.2% |
| High Blood Pressure (HBP) | 1 (Yes) | 105 | 35.1% |
| | 0 (No) | 194 | 64.9% |
| Sex | 1 (Men) | 194 | 64.9% |
| | 0 (Women) | 105 | 35.1% |
| Smoking | 1 (Yes) | 96 | 32.1% |
| | 0 (No) | 203 | 67.9% |
| Death Event | 1 (Yes) | 96 | 32.1% |
| | 0 (No) | 203 | 67.9% |

## 2.6 Detail description of variables:

1) **Age:** Aging can weaken and stiffen once heart**.** People 65 years or older have a higher risk of heart failure. **Older adults are also more likely to have other health conditions that cause heart failure**. Age is considered to be one of the most important factors that contribute greatly to heart failure.
2) **Anaemia:** A condition in which the number of red blood cells or the haemoglobin concentration within them is lower than normal and is often considered as a comorbidity with heart failure. Haemoglobin acts as the oxygen carrier and if one shows a decrease in red blood cell count causing a decrease haemoglobin level, there will be a decreased capacity of the blood to carry oxygen to the tissues and organs. Anaemic condition is common in cases of heart failure and the common triggers of anaemia involve age,

gender, nutritional iron deficiency, deficits in folate, vitamin B12, & vitamin A, chronic kidney disease, and cytokine production as noted in the studies by Shah et. al.,

3) **Creatinine Phosphokinase (CPK):** An enzyme (protein that helps to elicit chemical changes in the body) found in the heart, brain, and skeletal muscles. Any kind of damage to the muscle tissue causes the enzyme to leak into the blood stream. Consequently, high levels of CPK typically indicate a sort of elevated stress to the heart or other muscles. **The normal range of CPK in Mens is between 39 –308 U/L and 26 –192 U/L in Women.** Identifying the specific type of CPK helps determine what kind of a tissue could be damaged. In the condition of a heart failure the levels of CPK2 (CK-MD) are at elevated levels and could point to a myocardial muscle damage, electrical injury, or heart attack.

4) **Diabetes:** A chronic disease that occurs when the blood glucose levels are too high. The pancreas is no longer able to prepare the required amount of insulin. Insulin acts as a bridge to let the glucose from food flow from the blood stream into the body cells to generate energy. **Patients with history of diabetes are at a higher risk of developing heart failure.** High levels of blood sugar can potentially damage the blood vessels and nerves that control the heart.

5) **Ejection fraction (EF):** The measurement how much blood is pumped out of the left ventricle with each contraction. **The ideal range for EF may lie somewhere between 50 to 70 percent.** An EF of 60 indicates that 60% of the total blood volume in the left ventricle is pushed out with each heartbeat. An EF of less than 40% is indicative of heart failure or cardiomyopathy usually categorized as "systolic" heart failure. Heart failure triggered due to EF can be categorized into

   (i) Heart failure with reduced ejection fraction: EF<=40%,

   (ii) Heart failure with preserved EF: EF >=50%,

   (iii) Heart failure with mid-range EF: EF between 41–49(both inclusive) percent range

**6)** **Sex:** Although the overall lifetime risk of developing a heart failure for both men and women stands similar, there are striking differences in both the genders about the nature of heart failure. **Men are known to have a higher incidence rate for heart failure than women.**

7) **Platelets(thrombocytes):** Platelets are cells that circulate within the blood and play a major role in blood clotting mechanism by binding together when some kind of a damage or injury to a blood vessel is recognized. The normal range for platelet counts in the body ranges between 150,000 to 450,000 per μL of blood. The condition of having greater than 450,000 platelets is known as thrombocytosis whereas a count less than 150,000 is known as thrombocytopenia.

8) **Serum Creatinine:** Creatinine, a chemical waste formed as a by-product of normal muscle functioning present in the blood stream which is filtered in the kidney and eliminated through urine. **The normal creatinine levels in men and women are 0.6 to 1.2 milligrams/decilitres (mg/dL) and 0.5 to 1.1 mg/dL respectively**. **Men usually have higher creatinine levels compared to women since men, on an average, have more muscle mass.** The renal dysfunction and heart failure are closely related and associated with a high mortality rate.

9) **Serum Sodium:** An essential electrolyte, sodium helps in maintaining the balance of water level in and around the cells. Maintaining proper sodium levels is important for proper

functioning of muscles, nerves, and maintain stable blood pressure levels. **The normal sodium level is between 135-145 milliequivalents per litre. The condition of sodium level less than 135mEq/Lis known as hyponatremia.** Abebe el. al., [41] discuss the impact of sodium levels in prognosis of a heart failure condition and mention that hyponatremia is one of the vital factors in the prognosis of heart failure condition.

10) **Smoking:** People with habit of smoking have a major risk of developing ischemic heart disease. Ischemic disease happens due to building up of plaque within the coronary artery. Plaque can choke the arteries by forming blood clots, thereby limiting the flow of blood to the heart muscles. In the event of heart not receiving enough blood, depletes it from getting the adequate amount of oxygen and nutrients for appropriate functioning. This condition is called ischemia. Insufficient blood supply to the heart muscles puts the person at risk for a heart attack. The chemicals that go in with smoke triggers the builds up of plaque in the arteries, damaging the blood vessels, and altering the way they work by disturbing the normal heart rhythm.

11) **High Blood Pressure** (Hypertension): Termed as a **silent killer**, hypertension is a condition which occurs when the force with which the blood pushes the walls of the blood vessels in always on the higher side. When heart beats, blood is pumped out of the heart with some amount of force or pressure to circulate the blood throughout the body via the circulatory system. This pressure is made of two force components – (1) Systolic pressure –the force with which blood is pumped out of the heart into the circulatory system, and (2) Diastolic pressure –the force generated when the heart rests between heart beats. High blood pressure triggers harm by increasing the workload of the heart and blood vessels and forcing them work harder and inadequately.

12) **Time:** The time period where a heart failure patient transits from the in-patient setting to the out-patient setting is considered critical in managing heart failure condition. Mueller et. al., discuss the importance of a well-designed and structured follow-up program for refining the treatment outcomes. The authors highlight the importance of patient education during the follow-up visits for self-monitoring the signs& symptoms of any kind of decline in the heart health. Similarly, Agostinhoet.al., mention how a well-structured and protocol-based follow-up program can lead to reduction in hospital re-admission and mortality rates.

13) **Death Event:** Heart disease or cardiovascular disorders are one of the leading causes of death globally. A trigger for heart failure could be plaque building up within the blood vessels which reduces or blocks the flow of blood, dysfunction of the renal system causing high levels of creatinine, low sodium levels, fluctuating ejection fraction, or other cardiac abnormalities. The severity of the above mentioned factors could determine the criticality of the condition and death can happen due to acute myocardial infarction, progressive heart failure, sudden death, or other cardiovascular irregularities. The event of death may vary depending upon the gender, race, and ethnicity.

# Chapter 3

# STATISTICAL ANALYSIS AND INTERPRETATION OF DATA

## 3.1 KAPLAN MEIER ESTIMATE:

The study dataset contains 6 qualitative and 7 quantitative variables. “Time” variable is survival time of the patients under study and the Death event is the event of interest. Hence for the other variables survival probability curve is plotted by using Lifelines library in python under which KaplanMeierFitter() is used to plot the Kaplan Meier curve.

The Kaplan Meier survival curve for all the variables is plotted to study how these variables influence the survival rate at various level over time and to find the effect of survival probability based on the levels between groups. But our dataset contains only five Qualitative variables they can be studied directly. In order to study remaining variables under consideration, we need to follow few steps to convert Quantitative variables into Qualitative variables.

Table 3.1 Ordinal conversion of Medical Terms

| Si. No. | Features and the levels | Measurement | Ordinal values |
|---|---|---|---|
| 1 | Ejection Fraction | | |
| | - Reduced | <=40% | 1 |
| | - Mid-range | 41-49% | 2 |
| | - Preserved | >=50% | 3 |
| 2 | Platelets | | |
| | - Low | <1,50,000 | 1 |
| | - Normal | 1,50,000-4,50,000 | 2 |
| | - High | >4,50,000 | 3 |
| 3 | Serum Sodium | | |
| | - Low | <135 mEq/L | 1 |
| | - Normal | 135-145 mEq/L | 2 |
| | - High | >145mEq/L | 3 |
| 4 | Serum Creatinine | | |
| | Men: | | |
| | - Normal | 0.6-1.2 mg/dl | 2 |
| | - High | >1.2 | 3 |
| | Women: | | |
| | - Normal | 0.5-1.1 mg/dl | 2 |
| | - High | >1.1 mg/dl | 3 |
| 5 | Creatinine Phosphokinase | | |
| | Men: | | |
| | - Low | < 39 U/L | 1 |
| | - Normal | 39-308 U/L | 2 |
| | - High | >308 U/L | 3 |
| | Women: | | |
| | - Normal | 26-192 U/L | 2 |
| | - High | >192 U/L | 3 |

By converting these variables, we can able to find at what levels the survival rate of the patients increase / decreases over time. In order to find survival estimate for the Quantitative variables in the study, we have stratified the available data based on the medical criteria mentioned above, where each Quantitative variables in this data are stratified as follows, except age. Since, we know that "As age INCREASES survival rate DECREASES over time".

### 3.1.1 Kaplan-Meier Curve for the Qualitative variables

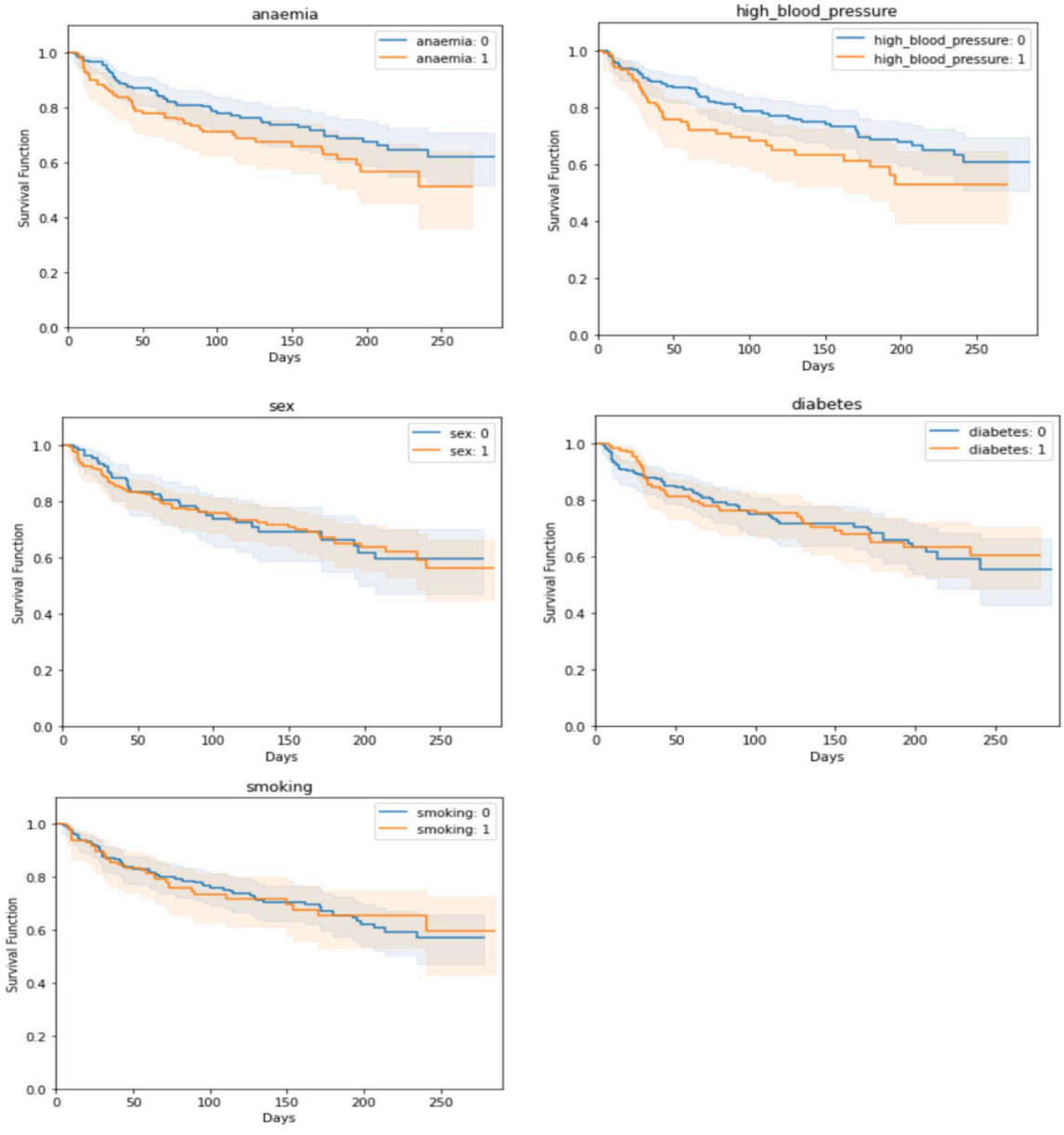



Figure 3.1-Kapalan Meier Estimates for the qualitative variables

**Qualitative insights:**

1. **Anaemia:** The KM estimate plot for anaemia shows a decreased survival probability if the person is known to be anaemic. In other words, the person who is not anaemic will

have higher survival probability of more than 70%. And hence Anaemia is a significant factor.

2. **Diabetes:** However, the KM estimate curve and survival probability for diabetes shows almost comparable trend for both diabetic and non-diabetic patients making diabetes a non-significant factor.
3. **High Blood Pressure:** The KM estimate curve also shows a similar trend as anaemia where patients with high blood pressure (hypertension) are at an increased risk of survival due to heart failure with significant lower survival probability.
4. **Sex:** The incidence and prevalence of heart failure is lower in women than in men at all ages. However, due to the steep increase in incidence with age, and the proportionally larger number of elderly women in the populations of the developed world, the total number of men and women living with heart failure is similar. The KM estimate curve and the survival probability curve shows similar trend for both men and women since, the population under consideration are at an advanced stage of heart failure. Here also the sex becomes a non-significant factor.
5. **Smoking:** Smoking is a major risk factor for developing initial stages of heart disease, People who smoke are at an increased risk of developing heart failure condition. However, the KM estimate curve for smoking and survival probability shows almost comparable trend for both smokers and non-smokers making smoking a non-significant factor.

Thus, from the KM estimates for Quantitative variables, **Anaemia and High Blood Pressure** are known to have a major impact on the survival probability of patients.

### 3.1.2 Kaplan Meier Curve for the Converted Quantitative Variables:

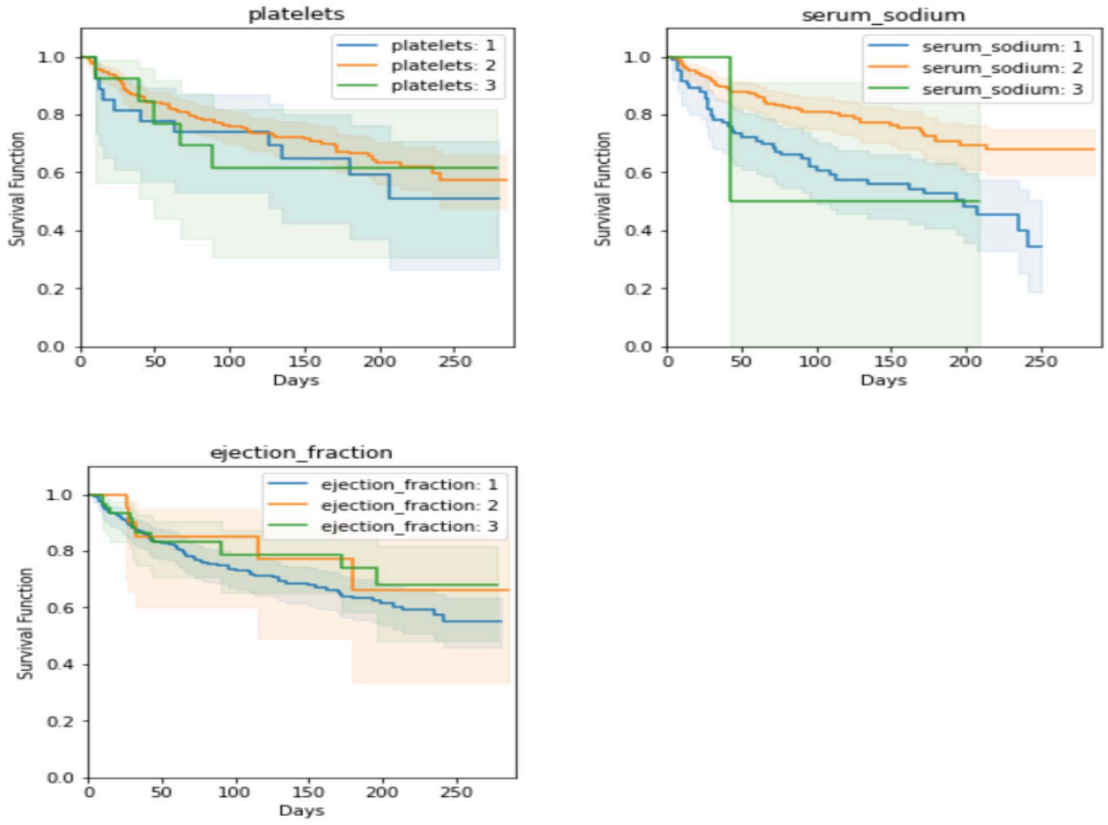


Figure 3.2: Kaplan Meier survival curve for EF, Platelets and Serum Sodium

Insights:

1. **Ejection Fraction (EF):** EF is a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction. Individuals who has low ejection fraction level will have lower survival rate than individuals having normal and high level of ejection fraction. Hence there is not much variations in survival rate due to Normal and High level of survival curve.
2. **Platelets:** Heart failure is associated with increased risk of venous thromboembolism, stroke, and sudden death. A normal platelet count ranges from 150,000 to 450,000 platelets per microliter of blood. According to medical conditions, the platelets counts does not differ much significantly based on survival probability. This is indicative that platelets could be an insignificant feature for predicting the survival probability and death prediction. But the person who has lower platelets counts will have a considerably lower survival probability when compared to other two cases.
3. **Serum Sodium:** Hyponatremia or low serum sodium level is one of the crucial factors in the clinical prognosis and a common biochemical disorder featured in heart failure patients. A normal blood sodium level is between 135 - 145 milliequivalents per litre. From the KM plots, it is observed that survival probability is least for population with extremely low sodium levels and the population come under the normal serum sodium showed better probability and best chances of surviving a heart failure condition.

### 3.1.3 Kaplan Meier Curve for the Sex based category variables:

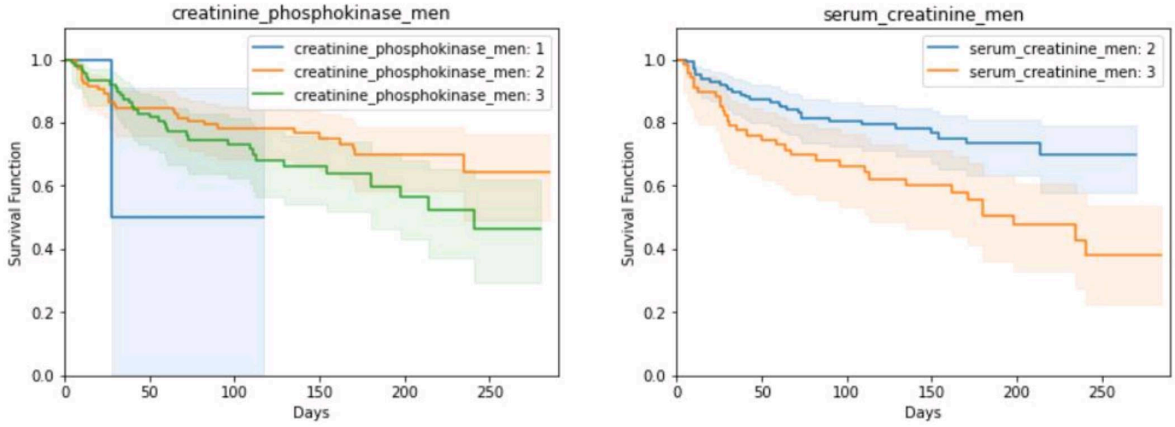



Figure 3.3: Kaplan Meier Survival curve for CPK and Serum Creatinine for Men

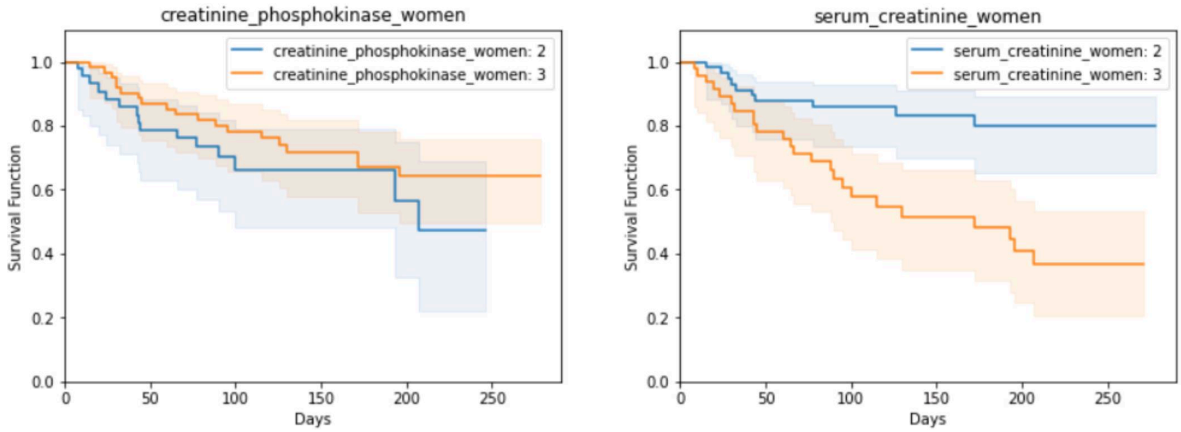



Figure 3.4: Kaplan Meier Survival curve for CPK and Serum Creatinine for Women

**Insights:**

1. **Creatinine Phosphokinase (CPK):** High levels of CPK typically indicate a sort of elevated stress to the heart muscles indicating either myocardial muscle damage, electrical injury, or heart attack. Creatinine phosphokinase values differ based on the Sex and hence it can be divided into CPK due to Men and Women. In Men category, the population falls under the HIGH CPK level will have a lower probability of survival, while the normal CPK population will have better chance of survival. But controversially in the Women category, the people having normal CPK value will have lower survival rate than the higher value of CPK. This makes CPK levels somewhat insignificant.
2. **Serum Creatinine:** Creatinine is a chemical waste product in the blood that passes through the kidneys to be filtered and eliminated in urine. Heart failure condition usually report an increase in serum creatinine levels in the scale of $\geq 0.3$ mg/dL. The range of serum creatinine levels for the population ranges from [0.5 - 9.4] which was divided based on the clinical values as mentioned above, for KM estimate curves. Since Serum Creatinine also accounts for Men and Women separately, in both the cases the increased level of serum creatinine will have lower probability of survival rate when compared to the normal level of Serum Creatinine. Hence it is not significant.

These variations in between the variables could be due to the fact that the population considered as part of this study (dataset) 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.

Thus, the KM estimate for the converted variables, we infer that if a person maintains the normal level will have better chance of survival even though they are at the extreme stages of heart failure.

## 3.2 LOG RANK TEST

In order to check that there is a significant difference between the groups we use the log rank test for all the variables, let us use the same data for checking the significance between groups in heart failure dataset. The log rank test is carried out by using the SPSS software.

Table 3.2
Classification on Significance between groups

| Si. No. | Variable Name | P Value | Significant or NOT |
|---|---|---|---|
| 1 | Age | 0.000 | Significant |
| 2 | Anaemia | 0.099 | NOT Significant |
| 3 | Diabetes | 0.840 | NOT Significant |
| 4 | High Blood pressure | 0.036 | Significant |
| 5 | Sex | 0.950 | NOT Significant |
| 6 | Smoking | 0.964 | NOT Significant |
| 7 | Ejection Fraction | 0.433 | NOT Significant |
| 8 | Platelets | 0.716 | NOT Significant |
| 9 | Serum Sodium | 0.000 | Significant |
| 10 | Serum creatinine -Men | 0.002 | Significant |
| 11 | Serum creatinine – Women | 0.000 | Significant |
| 12 | Creatinine Phosphokinase – Men | 0.235 | NOT Significant |
| 13 | Creatinine Phosphokinase – Women | 0.246 | NOT Significant |

From both the Kaplan Meier curve and the Log rank test we shall conclude that AGE, High Blood Pressure, Serum Sodium, Serum Creatinine have a major impact on the survival probability of patients who are at an advanced stages of heart failure condition.

## 3.3 EVALUATION METRICS

In our study we have consider two Evaluation Metrics P-value and C-index.

### 3.3.1 Probability Value (P-value)

The P-value is known as the probability value. It is defined as the probability of getting a result that is either the same or more extreme than the actual observations. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event. The P-value is used as an alternative to the rejection point to provide the least significance at which the null hypothesis would be rejected. If the P-value is small, then there is stronger evidence in favour of the alternative hypothesis.
P-value Table

The P-value table shows the hypothesis interpretations:

Table 3.3

| P-value | Decision |
|---|---|
| P-value > 0.05 | The result is not statistically significant and hence don't reject the null hypothesis. |
| P-value < 0.05 | The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis. |
| P-value < 0.01 | The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis. |

Generally, the level of statistical significance is often expressed in p-value and the range between 0 and 1. The smaller the p-value, the stronger the evidence and hence, the result should be statistically significant. Hence, the rejection of the null hypothesis is highly possible, as the p-value becomes smaller.

### 3.3.2 Concordance Index (C – Index):

The C-statistic gives the probability a randomly selected patient who experienced an event (e.g. a disease or condition) had a higher risk score than a patient who had not experienced the event.

Concordance index is a measure of how discriminant the model is. For survival analysis, say you have a covariate $X$ and a survival time $T$. Assume that higher values of $X$ imply shorter value for $T$ (thus $X$ has a deleterious effect on $T$). Discrimination means that you are able to say, with high reliability, that between two patients which one will have a shorter survival time.

It is equal to the area under the ROC curve and ranges from [0.5 - 1].

- A value below 0.5 indicates a very poor model. Such model is no better than predicting an outcome than random chance.
- Values over 0.7 indicate a good model.
- Values over 0.8 indicate a strong model.
- A value of 1 means that the model perfectly predicts those group members who will experience a certain outcome and those who will not.

## 3.4 COX PROPORTIONAL HAZARD MODEL

### 3.4.1 Training and Test Data Split

We have chosen 20% of our data to be as test dataset and the remaining 80% of the data is considered as training dataset using **sklearn**(python library).After fitting the data using **COXPH FITTER**(),there are 200 right-censored observations who have not yet had the event (or may have left the study).

### 3.4.2 Analysis

Among our variables, ejection fraction fails the proportional hazard assumption. So, we bin the ejection fraction variable based on the medical instruction specified especially for

ejection fraction of a human heart. The proportional hazard function gets satisfied as modification to ejection fraction was done.

Table 3.4

| | |
|---|---|
| **model** | lifelines.CoxPHFitter |
| **duration col** | 'time' |
| **event col** | 'DEATH_EVENT' |
| **baseline estimation** | breslow |
| **number of observations** | 299 |
| **number of events observed** | 96 |
| **partial log-likelihood** | -478.686 |

| | coef | exp(coef) | se(coef) | coef lower 95% | coef upper 95% | exp(coef) lower 95% | exp(coef) upper 95% | cmp to | z | p | -log2(p) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **age** | 0.044 | 1.045 | 0.009 | 0.026 | 0.062 | 1.026 | 1.064 | 0.000 | 4.815 | <0.0005 | 19.375 |
| **anaemia** | 0.412 | 1.510 | 0.217 | -0.014 | 0.838 | 0.986 | 2.311 | 0.000 | 1.895 | 0.058 | 4.107 |
| **creatinine_phosphokinase** | 0.000 | 1.000 | 0.000 | -0.000 | 0.000 | 1.000 | 1.000 | 0.000 | 1.847 | 0.065 | 3.950 |
| **diabetes** | 0.119 | 1.127 | 0.223 | -0.318 | 0.557 | 0.728 | 1.745 | 0.000 | 0.535 | 0.592 | 0.755 |
| **ejection_fraction** | -0.267 | 0.766 | 0.152 | -0.565 | 0.031 | 0.568 | 1.032 | 0.000 | -1.755 | 0.079 | 3.657 |
| **high_blood_pressure** | 0.501 | 1.651 | 0.217 | 0.076 | 0.927 | 1.079 | 2.526 | 0.000 | 2.311 | 0.021 | 5.583 |
| **platelets** | -0.000 | 1.000 | 0.000 | -0.000 | 0.000 | 1.000 | 1.000 | 0.000 | -0.258 | 0.796 | 0.329 |
| **serum_creatinine** | 0.278 | 1.321 | 0.064 | 0.153 | 0.404 | 1.165 | 1.497 | 0.000 | 4.342 | <0.0005 | 16.110 |
| **serum_sodium** | -0.059 | 0.942 | 0.022 | -0.102 | -0.017 | 0.903 | 0.983 | 0.000 | -2.730 | 0.006 | 7.302 |
| **sex** | -0.067 | 0.935 | 0.247 | -0.551 | 0.417 | 0.576 | 1.517 | 0.000 | -0.273 | 0.785 | 0.350 |
| **smoking** | 0.139 | 1.149 | 0.251 | -0.354 | 0.631 | 0.702 | 1.880 | 0.000 | 0.552 | 0.581 | 0.783 |

| | |
|---|---|
| **Concordance** | 0.711 |
| **Partial AIC** | 979.372 |
| **log-likelihood ratio test** | 61.038 on 11 df |
| **-log2(p) of ll-ratio test** | 27.326 |

### 3.4.3 Statistical significance using the p-value

The p value from the summary tells us that **Age, Anaemia, High blood pressure, Serum Creatinine, Serum Sodium** are highly significant. Their p-value is less than 0.05, implying a statistical significance at a (100 – 0.05) = **99.95%** or higher confidence level. These features are highly correlated to the death event.

### 3.4.4 Interpretation of the Coefficients

## Anaemia:

Anaemia is a binary variable, 0 means the particular patient has no anaemia and 1 means the patient has anaemia. Its coefficient is 0.412.

$$Hazard\ Ratio = \frac{Hazard\ rate\ for\ patients\ having\ anaemia}{Hazard\ rate\ for\ patients\ not\ having\ anaemia}$$

$$= \frac{e^{(0.412*1)}}{e^{(0.412*0)}}$$

$$= 1.510$$

1.510 is the Hazard Ratio associated with anemia. So, if the patient has anemia the instantaneous hazard of death at any given time t, increases by: (1.510 – 1) * 100 = 51%.

## High Blood Pressure (HBP):

High BP is a binary variable taking either 1, meaning the patient has High BP or 0, that a patient does not have high BP. Its coefficient 0.501 is interpreted as:

$$Hazard\ Ratio = \frac{Hazard\ rate\ for\ patients\ having\ High\ BP}{Hazard\ rate\ for\ patients\ not\ having\ High\ BP}$$

$$= \frac{e^{(0.501*1)}}{e^{(0.501*0)}}$$

$$= 1.650$$

1.650 is the Hazard Ratio associated with High BP. So, if the patient has High BP the instantaneous hazard of death at any given time t, increases by: (1.650 – 1) * 100 = 65%.

## Serum Creatinine

Creatinine is the waste product in the body which is formed when creatine breaks down. If high levels of creatinine are found in a body then that can mean your kidneys are either damaged or not working properly.

$$Hazard\ Ratio = 1.320$$

1.320 is the Hazard Ratio associated with Serum Creatinine. The instantaneous hazard of death at any given time t, increases by: (1.320 – 1) * 100 = 32%.

## Serum Sodium

Sodium is key to helping send electrical signals between cells and controlling the amount of fluid in your body. Your body needs it for your cells to work the right way. A person having too little or too much can cause problems.

$$Hazard\ Ratio = 0.943$$

0.943 is the Hazard Ratio associated with Serum Sodium. The instantaneous hazard of death at any given time t, increases by: (0.943 – 1) * 100 = -5.7%.

## 3.4.5 Partial Effects of Covariates

Construct plots comparing the baseline curve of the model versus what happens when a covariate(s) is varied over values in a group. This is useful to compare subjects' survival as we vary covariate(s), all else being held equal. The baseline curve is equal to the predicted curve at all average values (median for ordinal, and mode for categorical) in the original dataset.

## Age

With increasing age, the survival probability decreases. Increasing age has a deleterious effect on the survival chances.

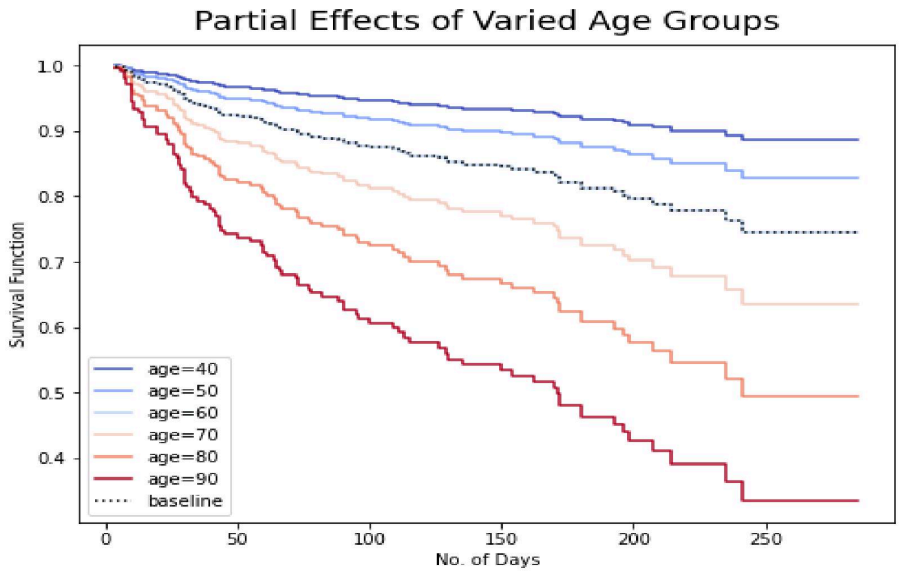



Figure 3.5

## Anaemia

Anaemic patients show a greater probability of encountering a hazard due to heart failure condition. Baseline and non-anaemic curves are overlapped.

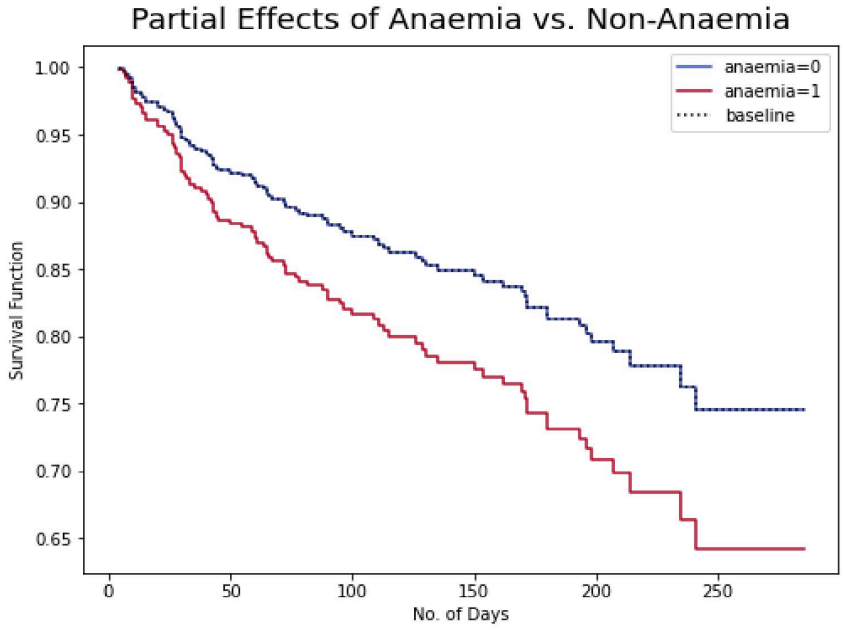



Figure 3.6

## Serum Creatinine

With increasing levels of serum creatinine in the blood, the survival probability decreases for to any complication arising out of a heart failure condition. Increasing creatinine levels have a decreasing effect on the survival chances. Baseline and serum_creatinine=1curves are almost overlapped.

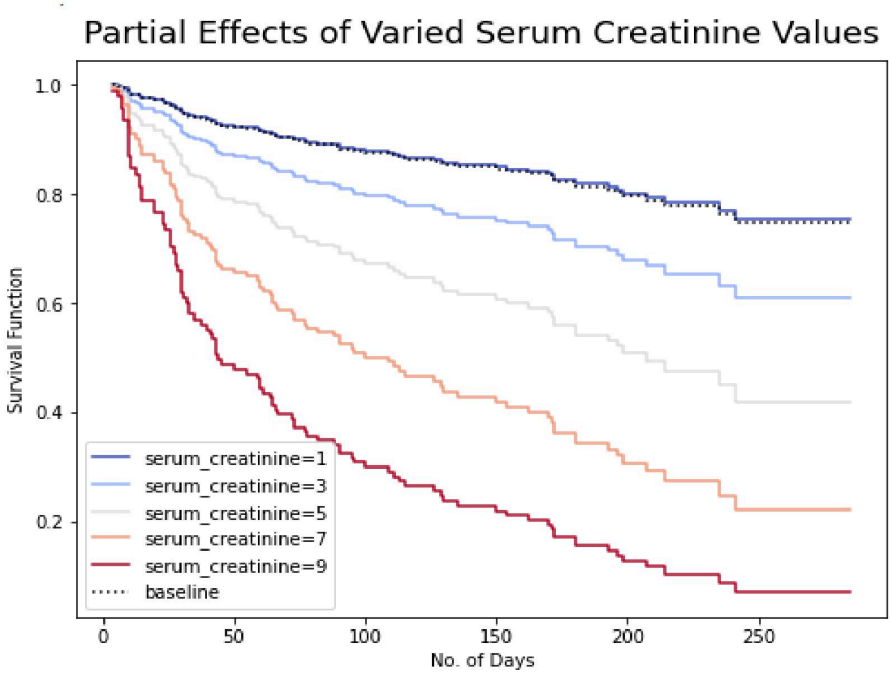



Figure 3.7

## High Blood Pressure

Patients with high blood pressure has decreasing survival rate when compared to patients not having high blood pressure. Baseline and no high blood pressure curves are overlapped.

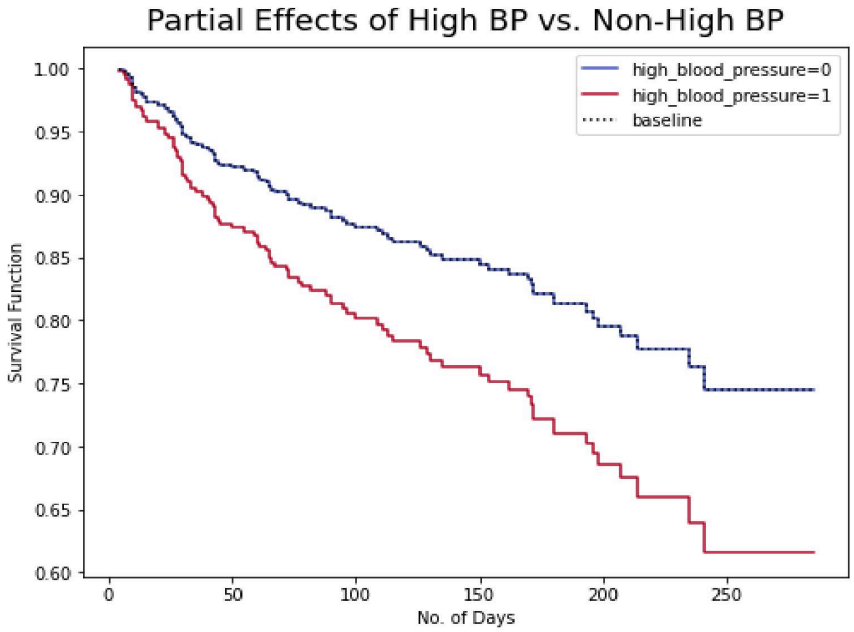



Figure 3.8

## 3.5 RANDOM SURVIVAL FOREST - INTERPRETATION

Since our original dataset, did not satisfy the Proportional hazard assumption we can use the Random survival forest as an alternative to Cox proportional hazard. By using the python software, we have trained, validate our results with the test data. The samples are selected randomly with 80% data as training data remaining 20% data are used for testing.

The analysis are made using scikit-survival and scikit-learn library in python, the model is fitted **from sksurv.ensemble import RandomSurvivalForest** and predicted for the test results. Then the validation of the data is done by using the C-index value, in and the result obtained as 0.6978 i.e., 69.78%.

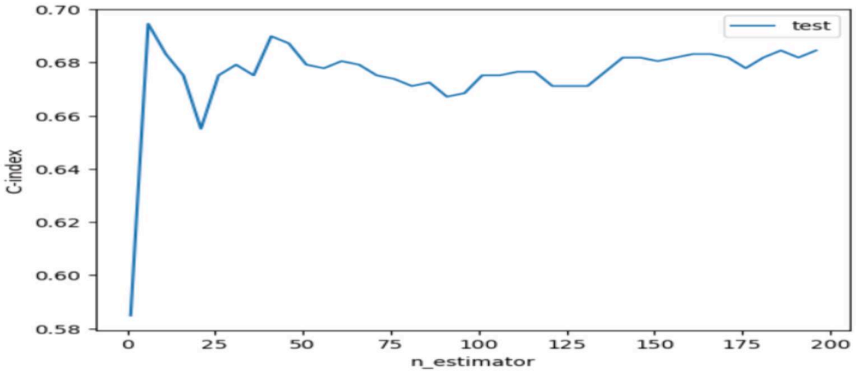



Figure 3.10 n_estimator vs C-index value -RSF

This C-index value is attained by fixing the suitable n_estimators (Number of trees) for this particular model.
Let's see how the test performance changes with the ensemble size (n_estimators). By fixing the appropriate n_estimator we obtain this concordance value.

## 3.6 GRADIENT BOOSTING - INTERPRETATION

We are using gradient boosting on Cox's partial likelihood with regression trees base learners, which we restrict to using only a single split (so-called stumps). This model achieves a concordance index of 0.72 on the test data. Let's see how the test performance changes with the ensemble size (n_estimators). We can see that the performance quickly improves, but also that the performance starts to decrease if the ensemble becomes too big.

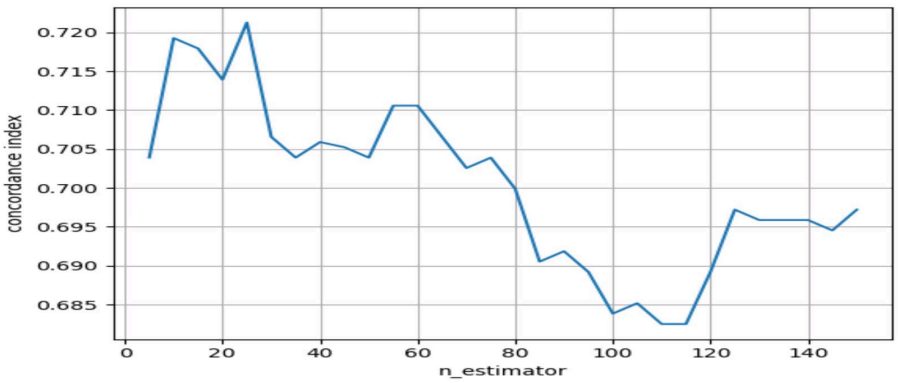



Figure 3.11 n_estimator vs C-index value -Gradient Boosting

## 3.7 COMPARISON BETWEEN THE RANDOM SURVIVAL FOREST AND GRADIENT BOOSTING

Table 3.5

| | Random Survival Forest | Gradient Boosting |
|---|---|---|
| **Concordance Index** | 0.697 | 0.72 |

By comparing these two models, Gradient boosting model performs better and the model with greater Concordance index will be a best fitted model.

# Chapter 4

# CONCLUSION

Heart failure dataset considered for the study involves the classes III or IV heart failure patients. Survival techniques like Kaplan Meier, Log rank test, Cox proportional Hazard Model were performed in order to estimate survival and hazard rate of the patients with respect to the variables considered under this study. Kaplan Meier estimate gives the survival probability for all the variables under study and we infer that if the patients maintain the normal level and not anaemic, HBP even at the last stage of heart failure will have a better chance of survival irrespective of smoking, diabetes and sex. Log rank test is performed in order to check the significance between the variable. From both the Kaplan Meier curve and the Log rank test - Age, High Blood Pressure, Serum Sodium, Serum Creatinine are statistically significant, but in clinical trails it is not true always, there will be variation in the significance. Usually the traditional model - cox proportional hazard model will give the hazard ratio based on the covariate and we can predict the results, but in our study the proportional hazard assumption fails. In order to overcome this, we follow stratification technique for fitting the model and results were predicted, we obtain the accuracy as 0.71 which is best fit. But we lose some information about the variable, in order to resolve the problem of proportionality assumption and losing of information there is a need of machine learning models such as random survival forest and gradient boosting. These models are especially for the survival time to event data, both the models are fitted and predicted for the test data. The random survival forest has an accuracy of 0.697 and the gradient boosting model has an accuracy of 0.704. When both models are compared, the random survival forest showed greater accuracy than the gradient boosting model. Our dataset only has limited number of records of heart failure patients. But, by collecting more data and we can enhance the accuracy scores of the survival models and its predictions in a real time healthcare setting. We can use the advanced technology and data to design a efficient care plan which will indeed improve the overall health of patients. Furthermore, awareness about having a healthy diet plan and physically active lifestyle could eventually help in reducing the mortality rate due to heart failure conditions.

## REFERENCE


1. Collett David – Modelling survival data in medical research (Chapman&Hall-CRC,2004)(ISBN 1584883251).(N.D.).
2. Kleinbaum, D. G., & Klein, Mitchel. (2005). *Survival analysis : a self-learning text*. Springer.
3. Mishra, S. (2022). A Comparative Study for Time-to-Event Analysis and Survival Prediction for Heart Failure Condition using Machine Learning Techniques. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *4*(3), 115–134. https://doi.org/10.35882/jeeemi.v4i3.225
4. Heidenreich, P. (n.d.). *Heart Failure Patients Need More Than Heart Failure Care**. http://www.sf-36.org/tools/SF36.shtml.
5. Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017b). Survival analysis of heart failure patients: A case study. *PLoS ONE*, *12*(7). https://doi.org/10.1371/journal.pone.0181001
6. Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine*, *2013*. https://doi.org/10.1155/2013/873595
7. Chicco, D., & Jurman, G. (2020). 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). https://doi.org/10.1186/s12911-020-1023-5
8. Nasejje, J. B., Mwambi, H., Dheda, K., & Lesosky, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology*, *17*(1). https://doi.org/10.1186/s12874-017-0383-8
9. Pickett, K. L., Suresh, K., Campbell, K. R., Davis, S., & Juarez-Colunga, E. (2021). Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01375-x
10. Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L., & Lu, J. J. (2020). A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.551420