

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**MACHINE LEARNING METHODS FOR
SURVIVAL DATA**

by
Tuğçe PAKSOY

August, 2021

İZMİR

MACHINE LEARNING METHODS FOR SURVIVAL DATA

**A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of
Science in Statistics**

**by
Tuğçe PAKSOY**

August, 2021

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**MACHINE LEARNING METHODS FOR SURVIVAL DATA**” completed by **TUĞÇE PAKSOY** under supervision of **ASSOC. PROF. DR. İDİL YAVUZ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....
Assoc. Prof. Dr. İdil YAVUZ

Supervisor

.....
Assoc. Prof. Dr.
Özlem GÜRÜNLÜ ALMA

Jury Member

.....
Assoc. Prof. Dr.
Özgül VUPA ÇİLENGİROĞLU

Jury Member

.....
Prof. Dr. Özgür ÖZÇELİK
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

Throughout my graduate education, I would like to express my sincere gratitude to my supervisor Assoc. Prof. Dr. İdil YAVUZ, for always helping me and motivating me with her smile and suggestions.

I would like to express my profound gratitude to my parents Mehmet PAKSOY and Emine PAKSOY and to my sister Ecehan PAKSOY for motivating and supporting me throughout my life.

Tuğçe PAKSOY



MACHINE LEARNING METHODS FOR SURVIVAL DATA

ABSTRACT

Survival analysis is a the statistical approach methods used in many fields. The key feature that distinguishes this method from other analysis methods is that it can be used when censored observations are present. In the literature, survival analysis can be categorized as traditional survival analysis methods and machine learning based survival analysis methods. The increasing number of data and variables, the existence of censored observations and the fact that traditional methods require some assumptions make it difficult to analyze survival data with traditional methods. In order to cope with this situation, machine learning methods specific to survival data are used. In this thesis, the traditional survival analysis methods such as the Kaplan Meier, Log rank test and the Cox regression are introduced and machine learning based Survival trees, and Random Survival Forests are studied. In application concordance statistics of Random Survival Forests and Cox regression models were obtained by using both real and simulated data in which different censoring rates were tested. Proper graphical representations were given for the survival curves and variable importance metrics when necessary. As demonstrated by the examples provided in this thesis the traditional approaches are advantageous when assumptions are met and statistical power is high however machine learning based methods work better when censoring is high an assumptions are not satisfied.

Keywords: Survival analysis, random survival forests, concordance statistics, cox regression

SAĞKALIM VERİLERİ İÇİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ

ÖZ

Sağkalım analizi birçok alanda kullanılan istatistiksel bir yaklaşım yöntemidir. Bu yöntemi diğer analiz yöntemlerinden ayıran en önemli özellik sansürlü gözlemlerin mevcut olduğu durumlarda kullanılabilmesidir. Literatürde sağkalım analizi geleneksel sağkalım analizi methodları ve makine öğrenmesi tabanlı sağkalım analizi methodları olarak ikiye ayrılmıştır. Artan veri ve değişken sayısı, sansürlü gözlemlerin varlığı ve geleneksel yöntemlerin bazı varsayımlar gerektirmesi durumları sağkalım verilerinin geleneksel yöntemler ile analiz edilmesini zorlaştırmaktadır. Bu durumla baş edebilmek için sağkalım verilerine özgü makine öğrenmesi yöntemleri kullanılmaktadır. Bu tezde Kaplan Meier, Log rank testi ve Cox regresyon gibi geleneksel sağkalım analizi yöntemleri tanıtılmakta ve makine öğrenme tabanlı sağkalım ağaçları ve rassal sağkalım ormanları çalışılmıştır. Uygulamada rassal sağkalım ormanları ve Cox regresyon modellerinin uyum istatistikleri, farklı sansür oranlarının test edildiği simülasyon sonucu elde edilen veriler ve gerçek veriler kullanılarak elde edilmiştir. Gerektiğinde sağkalım eğrileri ve değişken önem metrikleri için uygun grafiksel gösterimler verilmiştir. Bu tezde verilen örneklerde gösterildiği gibi, geleneksel yaklaşımlar, varsayımlar karşılandığında ve istatistiksel güç yüksek olduğunda avantajlıdır, ancak sansürleme yüksek olduğunda ve varsayımlar karşılanmadığında makine öğrenimi tabanlı yöntemler daha iyi çalışır.

Anahtar kelimeler: Sağkalım analizi, rassal sağkalım ağacı, concordance istatistiği, cox regresyon

CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZ.....	v
LIST OF FIGURES	viii
LIST OF TABLES.....	ix
 CHAPTER ONE – INTRODUCTION.....	 1
 CHAPTER TWO – SURVIVAL ANALYSIS	 4
2.1 Introduction	4
2.2 Censoring	5
2.2.1 Right Censoring.....	5
2.2.2 Left Censoring	7
2.2.3 Interval Censoring.....	7
2.3 Survival Models	7
2.3.1 Survival Function.....	8
2.3.2 Hazard Function	9
2.3.3 Relationship between $S(t)$ and $h(t)$	10
2.4 Parametric Methods	10
2.5 Non-parametric Methods	12
2.5.1 Kaplan-Meier Method	12
2.5.2 Log-Rank Test	13
2.6 Semi-parametric Methods	14
2.6.1 Cox Model	15
 CHAPTER THREE – MACHINE LEARNING	 17

3.1 Introduction	17
3.2 Machine Learning	18
3.3 Classification and Regression Trees	21
3.4 Random Forests	24
3.5 Machine Learning Methods for Survival Data	25
3.5.1 Survival Tree	26
3.5.1.1 Splitting Rules	28
3.5.1.2 Pruning Rules	31
3.5.2 Random Survival Forests	32
3.5.3 Performance Evaluation Metrics	36
CHAPTER FOUR – ANALYSIS	38
4.1 Data Analysis	38
4.1.1 Introduction	38
4.1.2 The Data	38
4.1.3 Kaplan Meier Curves	40
4.1.4 Cox Model Results	43
4.1.5 Survival Tree	44
4.1.6 Random Survival Forests	45
4.1.7 Concordance Index	46
4.2 Simulation	47
CHAPTER FIVE – SUMMARY AND CONCLUSION	50
REFERENCES	52
APPENDICES	59
Appendix 1: R codes for real data	59
Appendix 2: R codes for simulation data	62

LIST OF FIGURES

	Page
Figure 2.1 Right censored data.....	6
Figure 2.2 Survival Function.....	9
Figure 3.1 Regression tree for sleeping times (hour) of mammals.....	22
Figure 3.2 Survival Tree for the Lung Cancer Data	27
Figure 4.1 Kaplan Meier Curve.....	40
Figure 4.2 Kaplan Meier Curve for Gender.....	41
Figure 4.3 Curves for Each Type of Transmission	42
Figure 4.4 Survival Tree	45
Figure 4.5 Variable Importance Value for Variables	46

LIST OF TABLES

	Page
Table 2.1 Survival, hazard and probability density functions of some important distributions	11
Table 2.2 Data layout for the Kaplan-Meier method.....	12
Table 4.1 Variables of AIDS Data.....	39
Table 4.2 Log-rank Statistics for Gender	41
Table 4.3 Log-rank Statistics for Transmission	43
Table 4.4 Cox Proportional Hazard Model Assumption.....	43
Table 4.5 Gamma Values According to Different Censorship Percentages	47
Table 4.6 Cox Model Coefficients for Different Censorship Percentages	48
Table 4.7 Average Variable Importance Values of Random Survival Forest for Different Censorship Percentages.....	48
Table 4.8 Average Concordance Values of Models for Different Censorship Percentages	49

CHAPTER ONE

INTRODUCTION

Survival analysis, which has existed since the 18th century, is an important method used in statistics that enables the analysis of the time passed until the events of interest occur (Wang et al., 2019). Survival analysis has many connections in real life. For example, in the field of criminology, an application has been made to estimate the time it takes for people released from North Carolina prison to commit crime again (Chung et al., 1991). In traffic management, survival analysis has been used to estimate how the risk that pedestrians are exposed to while crossing, changes over time (Tiwari et al., 2007). In the area of tourism, how variables such as daily spending, salary and education affect the duration of summer vacation of tourists is analyzed with this type of analysis (Gokovali et al., 2007). In healthcare, it has been used to predict the chances of recurrence of people with anorexia nervosa and to find out what factors affect relapse (Carter et al., 2004). The feature that distinguishes this analysis from other types of analysis is that it is a suitable method for censored data. Censoring occurs when the event of interest is not observed during the study (Wang et al., 2019). Survival analysis helps in estimating the time to events of interest, comparing survival times between groups and determining which features affect survival times.

Survival analysis methods can be categorized as statistical methods and machine learning based methods. Statistical methods are further divided into three categories according to their estimates of survival and hazard functions. In this classification includes parametric, semiparametric and nonparametric models. In this thesis, one of the nonparametric methods (Kaplan Meier) and one of the semiparametric methods (Cox) regression will be explained. Kaplan Meier curve aims to estimate survival function using survival times (Chung et al., 1991). It is used to compare the survival probabilities of different groups visually (Behera, 2012).

Machine learning, which has been the subject of many studies in recent years, enables a model to be established by learning from data.

This method is used in various branches of science such as medicine, psychology and astronomy. There are many studies about machine learning in the literature. By making use of machine learning methods, digital records in hospitals facilitate the follow-up patients, diagnosis and treatment of patients at risk are carried out (Callahan & Shah, 2017). Ganggayah et al. (2019) examined the variables affecting the life span of women with breast cancer and used algorithms such as decision trees, support vector machines and neural networks to find which variables were effective. Pirooznia et al. (2008) have compared machine learning algorithms that aim to classify using microarray gene data. In the study conducted by Baca-Garcia et al. (2007), individuals with a family history of suicide attempts were observed and the suicidal tendency of those individuals and which variables triggered this emotion were examined by machine learning methods.

Like many other data types, survival data sizes and dimensions have increased due to technological advances in recent years. For this reason classical survival approaches need to be supplemented with machine learning methods for better inferential accuracy. However, machine learning methods have not originally been proposed to handle censored data. Recently these methods have been modified so that they can be used for survival data.

Machine learning methods have certain advantages and disadvantages compared to purely statistical approaches. They can be more flexible and can deal with data that violate common assumptions. They also generally provide more accurate predictions when trained and tuned properly. On the other hand, interpretability is usually sacrificed along the way since they focus more on prediction and less on modelling.

In this thesis, machine learning methods recommended for the analysis of survival data will be examined and their applications will be made on simulated data and real data. This thesis consists of 5 chapters. Survival analysis is described in Chapter 2. In this chapter, information about Cox regression from non parametric methods, Kaplan Meier from semi parametric methods and parametric methods are given. In Chapter 3, the definition of the concept of machine learning, why it is used in survival data, and machine learning methods that are specific to survival data are explained. Information

is given about the performance evaluation metric, which measures how well the established models fit the data. In Chapter 4, models are tested on simulated and real data. The last section contains conclusions.



CHAPTER TWO

SURVIVAL ANALYSIS

2.1 Introduction

Survival analysis is used to model and analyze the elapsed time until a specific event occurs (Wang et al., 2019). The event can be death, recurrence of the disease, bankruptcy or finding a job. The basis of this analysis goes back to 17th century. Since the reliability of war tools was an important issue during World War II, their durability had to be measured. Following the war, this method was used to scrutinize the reliability of the tools. Survival analysis is commonly used to analyze the lifetime of patients (Lee & Go, 1997). The uses of survival analysis have expanded in recent years such as economy, education, bioinformatics, marketing and forensics.

The feature that distinguishes survival data from other types of data is that they contain censored data. Censored data is the case where the duration of the event of interest specific to the individual cannot be observed during the study. This data, of which we have partial information, should not be considered as missing data. Different survival analysis methods should be used to deal with these data. Another feature of the survival data is that they have a highly skewed distribution and the event of interest takes positive values because it is a period of time (Kartsonaki, 2016).

Survival analysis helps to investigate the proportion of people who experienced the event in the period until a certain event occurs, compare the survival times of different groups and which situations are more effective in the occurrence of the event (Kim, 2012).

This section includes information on censoring, survival models, traditional survival analysis methods.

2.2 Censoring

When analyzing survival data, a problem specific to time to event data is encountered. This problem is censoring. Censoring is the situation in which the survival times of some of the observations in the study are not known exactly (Klein & Moeschberger, 2006). Observations may leave the study before the event of interest occur, or the study may be finished before the event occurs (Kleinbaum & Klein, 2012). Beadle et al.(1984a and b) conducted a study comparing two treatment methods in patients with early breast cancer. The first treatment method was applied to 46 patients, and the time to cosmetic effects of radiotherapy were examined. The second treatment method was applied to 48 patients, and the time to cosmetic effects of both radiotherapy and chemotherapy were examined. The aim of the study is to compare the time to cosmetic deterioration times of these two treatment methods. At the beginning of the study patients were observed every 4-6 months. However, as the recovery status of the patients got better, the time to come to the hospital was prolonged. The doctor recorded the measure of the breast retraction on a scale prepared (none, moderate, severe) at each visit. It has been noticed that breast retraction occurs in the time between visits, and these observations are censored. Another example of censorship can be given for a three year study that was conducted on patients with lung cancer. The event of interest in the study is the death of patients due to lung cancer. If the patient withdrew from the study after two years then the survival time was found to be 2 years or more (Lee & Go, 1997). The above two situations are examples of censored data, which is a special case of survival data.

There are some censoring types of censoring, right censoring, left censoring and interval censoring.

2.2.1 Right Censoring

For some observations, the event of interest may not occur within the time frame specified for the study. The situation where the actual survival times of these obser

uations are greater than or equal to the observed survival times is called right censoring (Kleinbaum & Klein, 2012). Let X be lifetime which is independent and identically distributed for patients in a study, probability density function $f(x)$ and survival function $S(x)$. C_r is the right censoring time. Given $T = \min(X, C_r)$, T is equal to X if lifetime is observable and if lifetime is censored T is equal to C_r (Klein & Moeschberger, 2006).



Figure 2.1 Right censored data

When the Figure 2.1 is examined, the event occurs for observation A and the survival time is 3 months. Observation A is not censored. Observation B was followed throughout the study, but no event occurred. So it is a censored observation. Survival time is at least 7 months. Observation C entered the study at the 2nd month and withdrew from the study at the 4th month. This observation is also censored. Observation D participated in the study at 4th and was observed until the study ended, but the event was not observed. So it is a censored observation. Observation E stayed in the study between the 2nd and 5th months. Then it has been lost. This observation is censored.

2.2.2 Left Censoring

Some observations are known to experience the event before the time of censorship, but the exact time of the event cannot be known. Such observations are called left censored observations. C_l is the left censoring time. If X (X is lifetime) is greater than or equal to C_l , the lifetime is known exactly (Klein & Moeschberger, 2006).

A study was conducted to determine the distribution of time until the first use of marijuana among high school boys in California. The students were asked when they first used marijuana. One student said that he used marijuana but did not remember the time. This observation is censored from the left. Because the event happened before the study was done, but it is not known exactly when the event occurred.

2.2.3 Interval Censoring

In interval censoring, which is a more general type of censorship, lifetime is observed only within a certain range. It is generally used in studies that require follow up. For example, interval censoring may occur naturally for a situation where the time until certain tools stop functioning properly is investigated and controls are made at certain intervals. Because these inspections are made at a certain intervals the working time of the tools is possibly interval censored (Klein & Moeschberger, 2006).

Interval censoring is a generalized form of censorship types. When interval censoring is $(0, C_l)$ it is left censoring and when event occurs between (C_r, ∞) interval censoring becomes right censoring (Klein & Moeschberger, 2006).

2.3 Survival Models

This section contains the basic functions used in modelling survival data. Random variable X can be defined as the time it takes until a certain event occurs. These events

may be a device malfunction, recovery of a patient taking a new medication, football teams staying in the super league or individuals finding a job. Survival models are used in modeling the distribution of the X random variable (Klein & Moeschberger, 2006). This section introduces survival function and hazard function.

2.3.1 Survival Function

Survival function is defined as $S(x)$. $S(x)$ is the probability that the random variable X will exceed a certain time. X is a random variable and x is the value that this random variable takes. x takes values between 0 and ∞ . Survival function provides probability for different values of x (Kleinbaum & Klein, 2012). Such that

$$S(x) = Pr(X > x). \quad (2.1)$$

If the random variable X is a continuous random variable then it becomes $S(x) = 1 - P(X \leq x)$. Another notation is as follows:

$$S(x) = Pr(X > x) = \int_x^{\infty} f(t) dt, \quad (2.2)$$

$$f(x) = -\frac{dS(x)}{dx}. \quad (2.3)$$

Assume that the event of interest is a person finding a job. The survival function gives the probability that the individual will not find a job until a certain time.

The features of the survival function are as follows (Kleinbaum & Klein, 2012):

- It is a non-increasing function.
- Since no event was observed at the start of the study, the survival function is 1 at $x = 0$.

- In the later time of the study, the survival function becomes 0 as x time approaches infinity.
- If the X random variable is discrete, the survival function is step function.

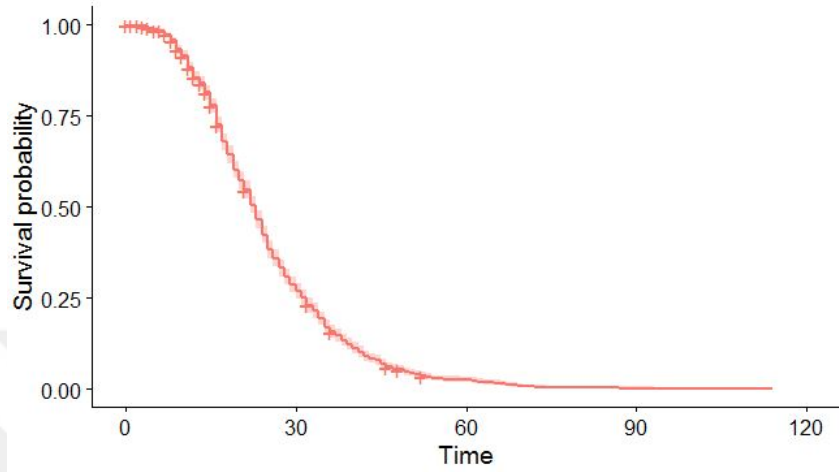


Figure 2.2 Survival Function

When Figure 2.2 is examined, while survival probability is 1 at time 0, this probability value approaches 0 as time progresses.

2.3.2 Hazard Function

Hazard function gives information about the occurrence of the event in a short interval, when it is known that the subject survives up to time x . The hazard function denoted by $h(x)$ can be written as (Klein & Moeschberger, 2006):

$$h(x) = \lim_{\Delta x \rightarrow 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}. \quad (2.4)$$

If X is a continuous random variable, the equation is:

$$h(x) = \frac{f(x)}{S(x)} = \frac{-d \ln[S(x)]}{dx}. \quad (2.5)$$

The cumulative hazard function can be written as:

$$H(x) = \int_0^x h(u) du = -\ln[S(x)], \quad (2.6)$$

where h_u is denoted by hazard function.

The properties of the hazard function can be listed as follows (Kleinbaum & Klein, 2012):

- The result of the hazard function is not probability, it is a ratio. The numerator of the function is conditional probability and its denominator is a value close to zero.
- Provides nonnegative results.

2.3.3 Relationship between $S(t)$ and $h(t)$

If either the hazard function or survival function is known, the other function can be found by mathematical operations below (Kleinbaum & Klein, 2012):

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{d}{dt}[1 - S(t)]}{S(t)} = \frac{-d}{dt} \ln[S(t)], \quad (2.7)$$

$$S(t) = \exp\left[-\int_{-\infty}^t h(s) ds\right]. \quad (2.8)$$

2.4 Parametric Methods

Non-parametric, semi-parametric and parametric methods can be used to predict survival or hazard function. Parametric methods allow estimation of survival and hazard function using various distributions (Lee & Wang, 2003).

The probability density function, hazard function and survival functions of some important distributions are shown in the table below (Wang et al., 2019).

Table 2.1 Survival, hazard and probability density functions of some important distributions

Distribution	PDF $f(t)$	Survival $S(t)$	Hazard $h(t)$
Exponential	$\lambda e^{-\lambda t}$	$e^{-\lambda t}$	λ
Weibull	$\lambda k t^{k-1} e^{-\lambda t^k}$	$e^{-\lambda t^k}$	$\lambda k t^{k-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})}$
Log-Logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$	$\frac{\lambda k t^{k-1}}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-(\frac{(t-\mu)^2}{2\sigma^2})}$	$1 - \phi((t - \mu)/\sigma)$	$\frac{1}{\sqrt{2\pi}\sigma(1-\phi((t-\mu)/\sigma))} e^{-(\frac{(t-\mu)^2}{2\sigma^2})}$
Log-Normal	$\frac{1}{\sqrt{2\pi}\sigma t} e^{-(\frac{(\log(t)-\mu)^2}{2\sigma^2})}$	$1 - \phi(\log(t) - \mu)/\sigma)$	$\frac{1}{\sqrt{2\pi}\sigma t} e^{-(\frac{(\log(t)-\mu)^2}{2\sigma^2})} \frac{1}{1-\phi(\log(t)-\mu)/\sigma)}$

Table 2.1 contains the survival and hazard functions of many probability distributions. The hazard function of the exponential distribution is equal to its parameter λ . A large lambda parameter indicates a short survival time and a high risk. The Weibull distribution is a right-skewed distribution that is consistently used in survival data (Balakrishnan & Kateri, 2008). Shape and scale parameters are included in the Weibull distribution. The shape parameter determines the shape of the hazard function (Kleinbaum & Klein, 2012). When shape parameter is equal to 1, the hazard ratio is constant, when shape parameter, k , is greater than 1, the hazard ratio increases and when it is smaller, it decreases. If the logarithm of the survival time shows a logistic distribution, the survival time has a log-logistic distribution. The hazard value of the log-logistic distribution showed an initially increasing and then decreasing structure, when λ is greater than 1 ($\lambda = (1/\sigma > 0)$) (Al-Shomrani et al., 2016). If the logarithm of the survival time shows a normal distribution, the survival time has a log-normal distribution. The hazard ratio of the log-normal distribution is 0 initially, that is, when time is 0. Then, as time progresses, it reaches its maximum value, and as time approaches infinity, the values approach 0.

As can be seen from the table, when strong assumptions about the probability distribution can be made, the survival and hazard functions can be written explicitly. In real research however, this is generally hard to justify.

2.5 Non-parametric Methods

Non-parametric methods can be used in cases where the distribution of the event time, which is the variable of interest in the study, is not known or in cases where proportional hazard assumption is not provided (Wang et al., 2019).

Non-parametric methods include Kaplan Meier, Nelson Aalen and Life Table methods and the Log-rank test. This section provides information on the Kaplan Meier method and Log-rank test.

2.5.1 Kaplan-Meier Method

Kaplan-Meier method is used to estimate the survival function (Wang et al., 2019). The curves from Kaplan Meier estimates of the survival probabilities are generally used to visualize the survival process. In the Kaplan-Meier estimator, survival estimates change at unique event times. So the survival probability between t_{i+1} and t_{i+2} is constant (Emmert-Streib & Dehmer, 2019). The data layout of this method is as follows. In Table 2.1, t_i are the distinct event times, d_i is the number of events, Y_i is the number at risk for observation i .

Table 2.2 Data layout for the Kaplan-Meier method

Time(t_i)	Number of events (d_i)	Number at risk (Y_i)	Product Limit Estimator ($\hat{S}_{t_i}$)
t_0	d_0	Y_0	$\hat{S}_{t_0}$
t_1	d_1	Y_1	$\hat{S}_{t_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_k	d_k	Y_k	$\hat{S}_{t_k}$

The explained of Table 2.1 is as follows:

- Since no event occurred initially, d_0 is 0, Y_0 is N and $\hat{S}_{t_0}$ is 1.
- Time (t_i) consists of observed times sequentially.

- Number of events (d_i) shows how many events occurred in each observed time.
- Number at risk (Y_i) contains the number of observations at risk. If there is 1 censored data before t_3 and 1 event occurred, it is expressed as $Y_3 = N - 2$.
- Product limit estimator ($\hat{S}_{t_i}$) contains estimates of the survival function for each observed time. The observations in this column form the Y-axis of the Kaplan-Meier curve.

The Kaplan Meier estimate for the survival probability at a specific time is calculated as three product of the estimated survival probabilities for all the unique event times less than or equal to that time point. This can be written as:

$$\hat{S}_t = \begin{cases} 1, & \text{if } t < t_1. \\ \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}], & \text{if } t_1 \leq t. \end{cases} \quad (2.9)$$

Kaplan-Meier formula is called the product limit formula. Because it is limited by the product of survival probabilities up to a certain survival time (Kleinbaum & Klein, 2012).

Kaplan-Meier method helps to examine the survival probabilities of different groups for different times. Kaplan-Meier curves are in the form of steps. If the steps are small, the number of observations in the data is large. If the steps are large, the number of observations is limited (Rich et al., 2010).

2.5.2 Log-Rank Test

To compare survival curves, graphs can be used as well as log-rank test statistics. The Log-Rank test is used when comparing the survival curves of two or more groups (Peto et al., 1977). The Log-Rank test provides a comparison of the effect of gender on the recovery time from cancer, the comparison of the effects of those with and without

chronic diseases on the COVID 19 recovery time, and the comparison of the factors that affect the employment time. The Log-Rank test tests the following hypothesis (Kleinbaum & Klein, 2012):

H_0 : Survival curves of two or more groups are the same.

H_a : At least one is different.

The Log-Rank test statistic for the two groups can be found as follows (Lee & Go, 1997):

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}. \quad (2.10)$$

For large samples this statistic approximates the 1 degree of freedom chi-square distribution. O_i indicates the number of failures observed for group i , and E_i the number of expected failures for group i for $i=1, 2$ (Lee & Go, 1997).

In large samples, when there are more than two groups the log-rank statistic approximates a chi square distribution with degrees of freedom $G-1$ where G is the number of groups compared (Kleinbaum & Klein, 2012).

2.6 Semi-parametric Methods

Semi-parametric methods can be thought of as a combination of parametric and non-parametric methods.

The advantage of semi-parametric methods is that there is no need for information about the distribution of survival time. This method includes Cox regression and its derivatives exist such as regularized Cox models, CoxBoost and the Time-Dependent Cox models (Wang et al., 2019).

This section provides information on the Cox model.

2.6.1 Cox Model

Survival data may contain more than one independent variable, the effects of these variables are important to the model and should be examined. The independent variables can be called covariates and factors. Covariates are continuous random variables where factors are categorical and are incorporated into the model via dummy variables. Examples of covariates are age, years of experience, incubation times of people infected with the corona virus. Examples of factors are the knowledge of a patient to take medication, the cancer cell size of a patient with cancer (squamous, small cell, adeno), the education level of a job seeker, occupational group, chronic ailment may be given. The Cox model is one of the most widely used regression methods in the literature.

The Cox model helps to model the relationship between dependent and independent variables similar to logistic and linear regression however the dependent variable it uses sets it apart (Van Dijk et al., 2008). Cox model examines the relationship between the time of the event observed and the independent variables using the hazard function. The model can be written as:

$$h(t|Z) = h_0(t) \exp(\beta^t Z) = h_0(t) \exp\left(\sum_{k=1}^p \beta_k Z_k\right). \quad (2.11)$$

When formula 2.11 is examined, the dependent variable consists of the hazard value at time t for an individual with certain independent variables. The Hazard function can take values between 0 and infinity. $h_0(t)$ is called the baseline hazard rate. It can be thought of as the constant term β_0 in regression analysis. It is a function that depends on t when there are no independent variables in the model. In the model, the independent variables can be time dependent or independent of time, that is, the value of the independent variables can remain constant as time changes. In this thesis, it is assumed that the value of the independent variables is time independent. That is, the independent variables consist of a vector of size p . The arguments can be expressed with the $Z_j = (Z_{j1}, \dots, Z_{jp})$ notation for subject j .

After the Cox model is established, various assumptions need to be provided. The assumptions of the Cox proportional hazard model are as follows (Klein & Moeschberger, 2006):

- In the proportional hazard assumption, there should not be a meaningful relationship between residuals and time.
- Continuous independent variables must be in linear form.
- $\hat{\beta}_i$ do not change with time.

Hazard Ratio

The Hazard ratio is the ratio of the hazard of one observation to the hazard of another observation defined as (Kleinbaum & Klein, 2012):

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t) \exp[\sum_{k=1}^p \beta_k Z_k]}{h_0(t) \exp[\sum_{k=1}^p \beta_k Z_k^*]} = \exp[\sum_{k=1}^p \beta_k (Z_k - Z_k^*)]. \quad (2.12)$$

Features of hazard ratio can be listed as follows (Klein & Moeschberger, 2006):

- The Hazard ratio value is fixed over time.
- In Hazard ratio, the hazards of two subjects with different independent variable values are compared.

Assume that one cancer patient received chemotherapy and not another. The aim of the study is to examine the effect of treatment and non-treatment on survival time (with treatment:1, without treatment:0). Hazard ratio is less than 1. In other words, those who are treated are more likely to survive than those who are not.

CHAPTER THREE

MACHINE LEARNING

3.1 Introduction

Artificial intelligence is the behaviour of a software or machine to imitate a human being. In order to understand the definition of artificial intelligence, it is necessary to understand 4 different categories (Russell & Norvig, 2010).

- Systems thinking humanly
- Systems acting humanly
- Systems thinking rationally
- Systems acting rationally

Acting like a human is often explained by the Turing test. The general logic of the Turing test is as follows. The interrogator poses a question to the machine and the human and tries to guess which one is the real person (Epstein et al., 2009). The aim of thinking like a human, which is a cognitive approach, is to learn how the human mind works. This can be achieved by psychological experiments or by predicting the thoughts in the mind of the human. In rational thinking, logical inference is made and demonstrable results are obtained. In acting rationally, the purpose is to construct a logical thought and make that thought and act on (Russell & Norvig, 2010).

Artificial intelligence is used in many areas in daily life: Linguistics, health, robotics, game, decision making systems, pattern recognition, security, transportation and natural disasters. Some important milestones in artificial intelligent are:

- World Chess Champion, Garry Kasparov lost to IBM's Deep Blue in 1997 (Munakata, 1996).
- Roomba robot released in 2002 and cleans the floor (Jones, 2006).

- Digital assistant Cortana, Google Assistant and Siri recognize and respond to people's voices (Hoy, 2018).

If we think of artificial intelligence as a cluster, a subset of this cluster is machine learning.

3.2 Machine Learning

Machine learning, which has been the subject of many studies in recent years, is an area of artificial intelligence that enables modeling by learning from data. The concept of machine learning was first introduced to the literature in 1959 by Arthur Samuel (Samuel, 1959). By using machine learning methods, classification, understanding the relationship between variables in the data set, image processing, clustering and estimation can be done.

Machine learning methods are used in various fields such as health, economy, security, meteorology, music and cinema. Some examples in the literature are as follows.

The multiplicity of operators has increased competition in telecommunications companies. Customers choose companies that suit their interests. Companies perform customer loss analysis to retain customers, they want to anticipate customers who want to leave and not to lose customers with appropriate campaigns. Artificial neural networks, naive bayes and support vector machines are used to predict customer loss in the field of telecommunication. As a result of the study, it was seen that artificial neural networks are more successful in predicting customer loss than other methods (Kaynar et al., 2017).

Parkinson's disease has various symptoms. One of these symptoms is speech disorders. This symptom can be used for early diagnosis. Voice records of the patient should be analyzed in order to follow the disease progress. In this study, two separate data sets

related to voice recordings of Parkinson's patients were used. Various machine learning methods such as naive bayes, support vector machine, k nearest neighbor, decision tree were used in the study. Statistically, the method that is more successful than other methods in diagnosing Parkinson's disease is the k -nearest neighbor (Badem, 2019).

Alpha Go is the first program to beat the world champion in the game of Go. Alpha Go uses 3 different algorithms. Deep neural network algorithm is used for position and movement. Neural network is trained in a supervised fashion that learns from human's previous games and plays with the reinforcement algorithm (Silver et al., 2016).

Looking at the examples in the literature, machine learning contains many methods. These methods fall under three main headings. These topics are listed as supervised, unsupervised and reinforcement learning. These 3 headings are briefly explained in the following 3 paragraphs.

Before supervised learning is briefly explained some preliminary information should be given. Before machine learning methods are used, in order to get a greater sense of the test error, the data are divided into training and test data with certain percentages. Training and test data can be composed of one independent variable and a dependent variable, as well as a matrix consisting of more than one independent variable. After this brief information, the definition of supervised learning can be made as follows. Supervised learning is basically learning from observed data. This learning is to understand the structure in the data set and to predict or classify an incoming observation in accordance with this structure (Friedman et al., 2009). In this learning, the variables and their values are determined and new observations are classified or estimated accordingly. In other words, the data structure is learned with the training data set and the accuracy of the learned structure in the test data is tested. For example, a child who learns to read and write can learn the letters one by one and recognize any letter when it is shown. Supervised learning contains many methods. Sometimes using the results of these methods alone is sufficient for the model, but sometimes combining the results of the methods used can give more accurate results (Brown 2010; Zhang 2012). This is the logic of ensemble learning. Consider the

following two examples:

- Looking at different sites to find more suitable and quality products in an online shopping
- Meeting with more than one school while making choices for the university

more accurate results can be obtained by referring to different sources. The logic of ensemble methods is similar to this thinking. More accurate results with less error can be obtained by combining the results of the methods used.

In unsupervised learning, the data set consists of independent variables and their values. The purpose of this learning is to group the values of the independent variables according to their similarities, and to reduce the dimensions if there are many variables in the data (Friedman et al., 2013).

Being a learning system in reinforcement learning, the agent acts or tests for a situation. In return, the system receives a reward that contributes to it or a penalty that contributes negatively to it and learns from its behavior (Kaelbling et al., 1996). The purpose of this learning is to be able to perform the trials that score the most reward points. Robot vacuum cleaner used in homes can be given as an example of reinforcement learning. If there is an obstacle in front of this broom, it hits it, realizes that it is an obstacle, and continues on its way from the place where there is no obstacle, and in this way, it performs the cleaning process, which is its main task, by learning from its behavior.

Before moving on to machine learning methods specific to survival data, the Classification and Regression Tree method, which forms the basis of these methods, should be briefly explained.

3.3 Classification and Regression Trees

Classification and Regression Tree is one of many the supervised learning methods. The data is generally multidimensional, that is, there are more than one independent variables. Corresponding to the dependent variable, the problem of interest may change depending on whether the dependent variable is numerical or categorical. That is, if the dependent variable takes numerical values, the method becomes a regression problem, and if the dependent variable consists of categorical data, the method turns into a classification problem (Deconinck et al., 2005).

This method can be thought of as a tree. Just as a tree has a trunk, branches, and leaves, this algorithm consists of a root node, decision nodes or splits, and leaf nodes or terminal nodes.

In this method, there is usually more than one variable in the dataset and these variables can be divided into subgroups according to different rules (entropy, information gain etc.). The tree begins to grow downward depending on whether a variable selected by the rule is less than or greater than a certain value or whether the variable belongs to that class. The growth process can continue until very few observations remain in the leaf nodes. Sometimes the tree can grow too large. This is also a problem. This problem can be solved by pruning the tree (Breiman et al., 1984; Moisen, 2008) .

The goal in the regression tree is to find splits that minimize the model error. For regression trees, the mean of the observed responses at each terminal node is used as the prediction of the dependent variable. The purpose of the classification tree is to ensure that observations in a similar class can be grouped together in each leaf node. For a new observation, the terminal leaf corresponding to that observation's independent variable values is determined and then the category that takes the majority vote in that terminal node is assigned to that observation.

An example is given to visualize the tree structure and to interpret the results.

Savage & West (2007) have studied the sleep times of mammals. The data set used in this study includes the following features. The data set consists of 83 observations and includes the variables of total sleep time, body weight of mammals, whether mammals are carnivorous or herbivorous and brain mass. The total sleep time is the response.

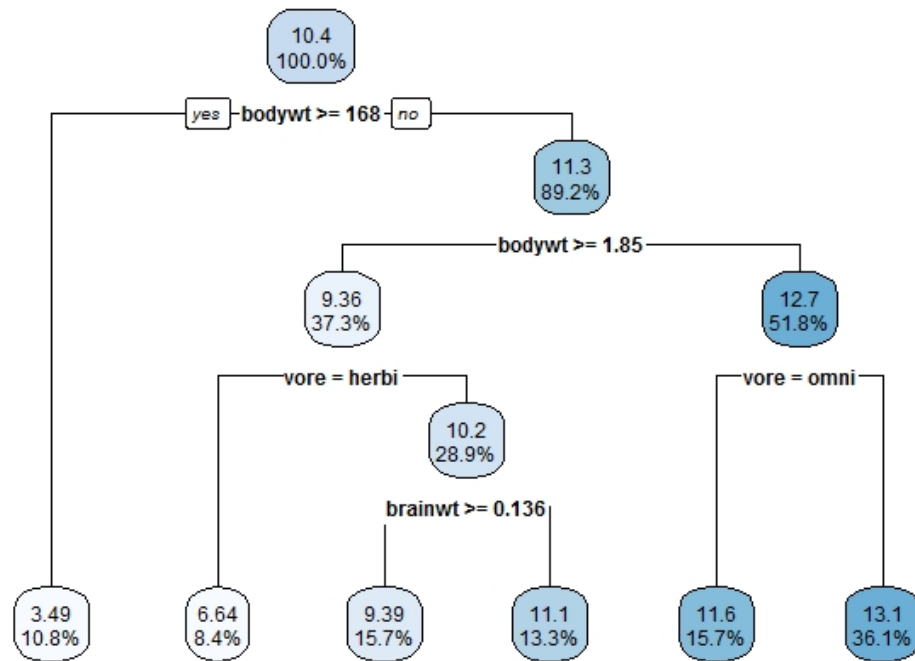


Figure 3.1 Regression tree for sleeping times (hour) of mammals

The tree began branching from mammals with a body mass of less or more than 168 kg. The place to start branching here is the root node or the stem of the tree. A decision node is where the data set is divided in a binary fashion with respect to a feature. For example, in this tree, a split is made according to body mass (less than or equal to 185 kg). Following splits use eating behaviour and brain weight. The sleeping time of the herbivorous mammals with body mass greater than or equal to 1.85 was predicted to be 6.64 hours. At the leaf nodes, the top number represents the prediction and the percentages are the percentage of observations in that node.

Gini index (I_G) or Information Gain (IG) can be used to form nodes in classification and regression trees. The purpose of the splitting process using information gain is to

minimize the entropy value and the build the tree starting from the variable with the highest information gain. The information gain is obtained by subtracting the sum of the calculated entropy values for each variable from the entropy before the tree split.

The formula is calculated as follows (Su & Zhang, 2006):

- S: Set of training instances.
- X: Feature
- S_x : Subset of S with $X=x$
- C: Number of classes
- Entropy(S): Entropy value from the all trees
- $P_s(c_i)$: Number of cases of interest in class i divided by the total number of cases
- $|\cdot|$: Number of instances

The information gain can be written as:

$$IG(S, X) = \text{Entropy}(S) - \sum_x \frac{|S_x|}{|S|} \text{Entropy}(S_x). \quad (3.1)$$

Entropy is a measure of uncertainty, variability and is calculated as follows:

$$\text{Entropy}(S) = - \sum_{i=1}^C P_S(c_i) \log P_S(c_i). \quad (3.2)$$

Gini index can also be used as a splitting criterion. It can be calculated by

$$I_G = 1 - \sum_{i=1}^k P_i^2, \quad (3.3)$$

where P_i^2 is the square of the probability values of each class if there are k classes in the variable of interest. The Gini index takes values 0 and 1, and purity increases as the value gets closer to 0.

3.4 Random Forests

Random Forests (Breiman, 2001) is an ensemble learning method formed by combining more than one tree. In this method, which is established by using more than one tree, the predictive power of the model is increased. With this method, B bootstrap samples are created from the data and a single full size tree is grown from each bootstrap samples. Trees can be classification or regression trees depending on the problem of interest.

Random Forests provide an advantage for problems with multiple independent variables by acting on the variance of the model. More precisely, for each bootstrap sample, when the tree is being grown, only a random subset of the original features are considered as candidates for a split. The randomization process is repeated for every split. Doing so, every variable is given equal chance of being included in a split. In addition, this process guarantees that the trees from the bootstrap samples will vary and be less similar to each other which in return brings the variance down and helps prevent overfitting problems when the predictions from them are averaged. The splitting rules used in the classification and regression tree can be used as the splitting rule for random forests also. Then with the estimates obtained from each tree, the final estimates are obtained and interpretations can be made. If the problem of interest is a classification problem, the predictions consist of the majority result in the leaf node of the each tree. If the problem of interest is a regression problem, the final estimate is the average of the averages from the leaf nodes of the trees.

The working logic of this method is briefly explained above. The algorithm steps of this method are as follows (Friedman et al., 2009):

- B bootstrap samples are selected from the data consisting of n observations.
- $\frac{2}{3}$ of the data is included in the bootstrap sample and this data is called in bag data. The remaining $\frac{1}{3}$ is called out of bag data. For each bootstrap sample the data set is separated into in bag and out of bag data.

- In each decision tree at every split, m independent variables are randomly selected from p independent variables. The random selection operation continues until the minimum number of observations remains in each leaf node.
- Average T_b results for the regression problem are calculated in the trees obtained from each bootstrap sample. In formula 3.4, calculation of mean predictive values for bootstrap samples is shown.

$$\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x). \quad (3.4)$$

For the classification problem, the estimate that makes up the majority of the C_b class estimates from the trees obtained from each bootstrap sample is considered the final estimate. Such that:

$$\hat{C}_{rf}^B(x) = \text{majorityvote} \hat{C}_b^B(x). \quad (3.5)$$

This method is suitable for many data types. The version applied to survival data was developed by Ishwaran et al. (2008).

3.5 Machine Learning Methods for Survival Data

Machine learning methods are suitable for use in the data with increasing number of variables and the number of observations, that is, in increasing data size and in some statistical methods, where concepts such as linearity and normality cannot be achieved (Wang et al., 2019).

Survival analysis has a unique data structure. The data used in this method includes censored observations. While traditional survival analysis methods (Kaplan Meier, Cox Regression) can cope with censored observations, machine learning methods can also be used. In order to cope with survival data, Survival tree and Random Survival Forests, which are methods specific to survival analysis data, can be used, as well

as analysis by introducing and modifying censored data for other machine learning methods.

The purpose of using machine learning methods for survival data is to make significant inferences from large-scale survival data, be useful when proportional hazard assumption is not provided, to obtain better results from traditional survival methods in terms of fit to the model in some data, and to help analyze data containing censored observations.

There are many studies in the literature using machine learning methods for survival data. Jerez et al. (2010) analyzed the time from patients to relapse of breast cancer. For this purpose, the age, hormone levels and tumor size characteristics of the patients were examined in the study. They estimated the recurrence times of the disease using artificial neural network and k nearest neighbor algorithms. In their study, Miao et al. (2018) created data with different variables that are thought to affect heart attack in order to predict the risk of dying from heart attack. In this data, where the number of variables is high, which variables affect the heart attack more were examined. They compared the model performances of the Improved Random Survival Forests where they used the weighted log-rank test as a basis for splitting and stopping and Random Survival Forests. It was found that the Improved Survival Forests method had better fit to the model.

3.5.1 Survival Tree

Tree-based methods are useful for survival data. Because these methods are non parametric, they do not require any assumptions, and since they contain more than one independent variables in the data, by dividing the subjects into subgroups using binary divisions via these variables, they are easy to interpret and can even be utilized by researchers and community not familiar with survival analysis (Härdle et al., 2006). Classification and Regression Trees are explained in section 3.3. The logic of the Survival Tree method is similar. The feature that distinguishes Survival trees from

Classification and Regression trees is that it uses different splitting and pruning rules. Generally, a full size tree is created. This tree grows under certain splitting and rules, and when the tree grows too large, it is pruned using certain pruning rules in order to avoid overfitting. In other words, the Survival tree method has emerged with the development of tree-based methods by Breiman et al. (1984) for censored data (Bou-Hamad et al., 2011). The most comprehensive and first article on this method was written by Gordon & Olshen (1985).

In order to understand the structure of the survival tree, a tree will be drawn and interpreted on a data set. In an example conducted by the US Veterans Administration, male patients with advanced inoperable lung cancer were given either a standard therapy or a chemotherapy (Prentice, 1973). The patients' survival patterns were followed along with additional independent variables such as cell-type and Karnofsky score.

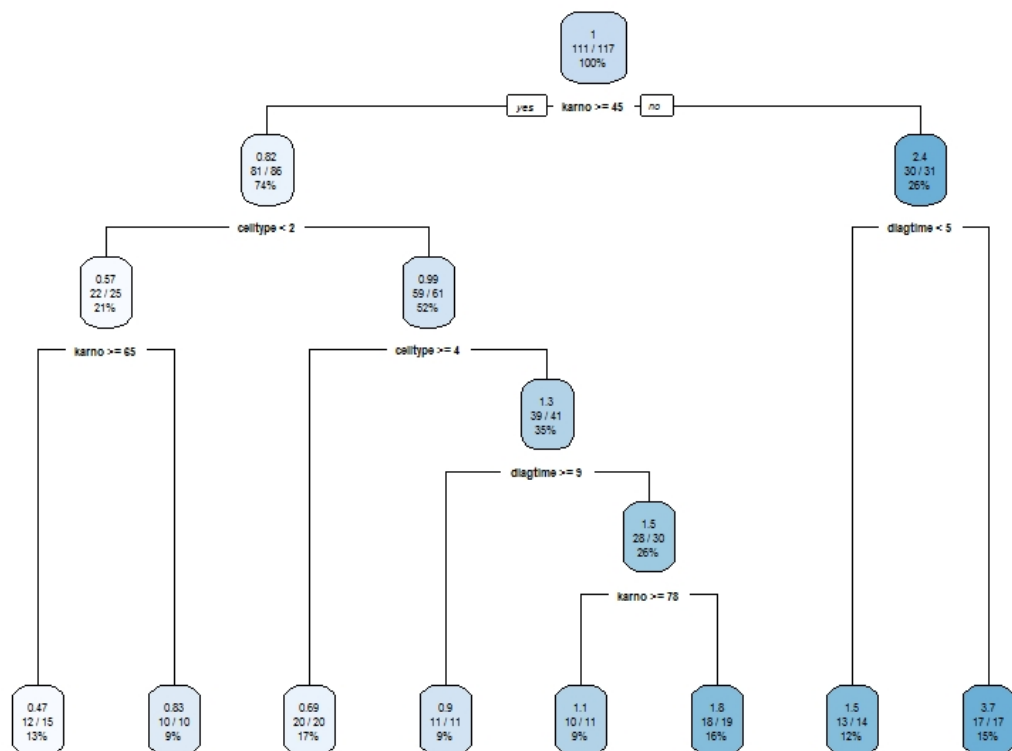


Figure 3.2 Survival Tree for the Lung Cancer Data

In Figure 3.2 Survival tree built by using the rpart package. The splitting criteria used by rpart is first step of full likelihood which will be defined in addition to other criteria in Section 3.5.1.1.. Each decision node contains the observation rate in node. When the figure is studied, it can be seen that the first split is made according to the Karnofsky score (greater than or equal to 45), then successive splits follow with respect to cell type, diagnosis time etc.

3.5.1.1 Splitting Rules

The splitting rule is important for tree based methods. Tree growth can be achieved by using splitting rules. In tree based methods without censored observations, impurity measures can be used as splitting rules. Information gain (Quinlan, 1986) and Gini index (Breiman et al., 1984) are examples of impurity measures (Shimokawa et al., 2016). If there are censored observations in the data, splitting rules such as Gini index or information gain should not be used. Different separation rules have been proposed for trees containing censored observations.

In this section, Wasserstein distance, logrank statistic, martingale residuals and first step full likelihood estimation which are among the splitting rules, are explained.

- Splitting Method: Wasserstein distance (Gordon & Olshen, 1985)

Gordon and Olshen used the Wasserstein distance as the splitting rule. This distance originally looks at the distance between the cumulative density functions of the two probability distributions. For survival analysis, the cumulative probabilities are obtained for two nodes by subtracting the Kaplan Meier survivals from one, then the Wasserstein distance is measured. The split that maximizes the distance between nodes is preferred (Shimokawa et al., 2015).

- Splitting Method: First step full likelihood estimation (LeBlanc & Crowley, 1992)

In this method the tree is grown by using the first step full likelihood estimation for proportional hazard model.

- h : Node
- $\tilde{T}$: Terminal nodes in the tree
- $\lambda_h(t)$: Hazard function for node h
- $\Lambda_h(t)$: Cumulative hazard function for node h

The likelihood can be written as:

$$L = \prod_{h \in \tilde{T}} \prod_{i \in S_h} \lambda_h(t_i)^{\delta_i} e^{-\Lambda_h(t_i)}. \quad (3.6)$$

- Splitting Method: Log-rank statistic (Ciampi et al., 1986), (Segal, 1988), (LeBlanc & Crowley, 1992), (LeBlanc & Crowley, 1993)

One of the commonly used splitting methods in the literature is Log-rank statistic. The Log-rank test is used to examine whether there is a significant difference between the two survival curves (Ishwaran & Lu, 2014).

The purpose of using log-rank statistic as the splitting rule in survival trees is to maximize the difference in survival probabilities between nodes (Segal, 1988; LeBlanc&Crowley, 1992; LeBlanc&Crowley, 1993). The notations used in formula 3.7 are given below (Ishwaran & Lu, 2014).

- h : Root node of the survival tree
- T_i : Survival time for the i^{th} observation
- X : Feature vector

- δ_i : $\delta_i=0$ (censored), $\delta_i=1$ (event occurred)
- $X \leq c$ or $X > c$: c is a specific value and allows the variable X to be split into left(L) and right(R) nodes.
- $t_1 < t_2 < \dots < t_m$: Event times
- $d_{j,L}$: Number of event at time t_j in left node
- $d_{j,R}$: Number of event at time t_j in right node
- $Y_{j,L}$: Number at risk in left node at time t_j
- $Y_{j,R}$: Number at risk in right node at time t_j
- Y_j : $Y_{j,L} + Y_{j,R}$
- d_j : $d_{j,L} + d_{j,R}$

The Log-rank statistics is defined as:

$$L(X, c) = \frac{\sum_{j=1}^m (d_{j,L} - Y_{j,L} * \frac{d_j}{Y_j})}{\sqrt{\sum_{j=1}^m \frac{Y_{j,L}}{Y_j} * (1 - \frac{Y_{j,L}}{Y_j}) * (\frac{Y_j - d_j}{Y_j - 1}) * d_j}}. \quad (3.7)$$

The $|L(X, c)|$ expression helps to calculate the distance between left and right nodes. As the distance between the left and right nodes increases, the splitting success increases.

- Splitting Method: Martingale residuals (Therneau et al., 1990)

Therneau et al. (1990) use martingale residuals as dependent variables in the Cox model and make the splitting rule according to these residuals.

- C_i : Cox-snell residual value for the i^{th} observation
- δ_i : $\delta_i=0$, if i^{th} observation is censored. $\delta_i=1$, if i^{th} observation is not censored.

The martingale residual for i^{th} observation can be calculated with the following formula:

$$M_i = \delta_i - C_i. \quad (3.8)$$

If the i^{th} observation is censored, martingale residual takes a value $[-\infty, 0]$. If the i^{th} observation is not censored, martingale residual takes a value $[-\infty, 1]$.

The purpose of the splitting rule is to ensure that the inside of the node is homogeneous and the between nodes are heterogeneous, so that the survival difference between nodes are maximum (Molinaro et al., 2004).

3.5.1.2 Pruning Rules

Decision trees continue to be divided into decision nodes and leaf nodes by applying various splitting rules. In some trees, this distinction can result in tree overfitting. The overfitting problem can be solved by pruning the tree. Even if different splitting rules are used in tree based methods used for censored data, the problem of over growth is also present in this tree structure. In this section, some pruning rules in the literature are briefly explained.

- Pruning Method: Cost complexity (Davis & Anderson, 1989), (Intrator & Kooperberg, 1995)

Cost complexity is a method used in tree based methods to reduce tree size.

- t_T : Terminal node
- $R_\alpha(t_T)$: Cost complexity metric
- α : Complexity parameter
- R_{t_T} : It allows measuring the impurity value in the terminal nodes

The cost complexity metric is then given as:

$$R_\alpha(t_T) = R(t_T) + \alpha|\tilde{t}_T|. \quad (3.9)$$

Cost complexity pruning works as follows. The cost complexity metric is composed of two terms. The impurity of the terminal nodes plus a penalty for the number of terminal nodes regularized by the tuning parameter alpha. So, a tree is chosen such that this complexity is minimized by balancing the complexity of the tree and the node impurity simultaneously.

- Pruning Method: Cross- validation and cost complexity (Gordon & Olshen, 1985), (Therneau et al., 1990), (LeBlanc & Crowley, 1992)

Some authors have suggested a single method as the pruning rule. Others have used more than one method to prune the tree and adjust the tree size. Cost complexity was explained in the previous pruning rule. Cross- validation will be explained in this section. The cross- validation method divides the data set into a certain number (5-10) roughly equally sized portions and at each turn uses one of these parts as test data while using all the others as training data. This process allows for all data points to contribute as both training and test observation therefore enables researchers to obtain a more stable idea about the test error of their models.

If the cross- validation method is used to prune the tree, the tree continues to grow with the specific splitting rule. The growth process continues until the tree remains stable. Deviance(cross-Entropy) values are then estimated for each node. Once the tree grows smoothly, the censored data in the node presents various difficulties. If all the observations of a node in the tree are censored, then the hazard of that node is estimated as 0. This is a problem. This problem was solved by changing the hazard values estimated from 0 to 0.5 (LeBlanc & Crowley, 1992).

3.5.2 Random Survival Forests

Tree-based methods can consist of a single tree such as a survival tree, or it can be one of the ensemble methods consisting of a combination of more than one tree. Random Survival Forests is one of the ensemble approaches. This method, which was developed by Ishwaran et al. (2008) to adapt the Random Forest method to survival data, is a method that enables more accurate predictions by combining the predicted values of many survival trees. In Random Survival Forests, Survival trees are created by using B randomly selected bootstrap sample from the data. For each tree, a randomly selected subset of independent variables are considered as splitting variables in each node. The number of these independent variables can be found by the square root of the number of variables in the data. By choosing a certain amount of independent variables randomly instead of all of them at each node, the chance of coexistence of the variables that are related to each other among the independent variables is reduced. These independent variables are then split into child nodes with certain splitting rules and the tree grows. The goal is to maximize the difference of the survival probabilities between nodes. After the tree has grown, the cumulative hazard function estimates are calculated for each survival tree and are averaged in the end. While applying the Random Survival Forest algorithm, the following steps are followed (Ishwaran et al., 2008):

- From the data, B bootstrap samples should be selected randomly. Approximately 63 % of the data will appear in the bootstrap samples. This is called in bag data. In bag data or bootstrap data is used to grow the tree. 37 % of the data will not appear in the bootstrap samples. This is called out of bag data. Out of bag data is not used grow the tree and is sometimes utilized to calculate out of bag errors.
- Survival trees are created for each bootstrap sample with randomly selected variables at each node. Separation is performed using variables that maximize the survival difference between child nodes. Splitting is done according to the splitting criteria in the Survival tree.

- Cumulative hazard function or survival function is calculated for each survival tree. The final estimates are obtained by averaging of all leaf nodes from the Survival trees. Cumulative hazard function can be found by Nelson- Aalen estimator. The notations in formula 3.10 are defined below.

- $t_{l,h}$: event time of l^{th} observation in node h
- $d_{l,h}$: number of events at time $t_{l,h}$
- $Y_{l,h}$: number at risk at time $t_{l,h}$

The cumulative hazard function is:

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}. \quad (3.10)$$

Ensemble cumulative hazard function for all Survival trees is obtained with the bootstrap ensemble cumulative hazard function.

- $H_b^*(t|x_i)$: conditional cumulative hazard function of the node for the b^{th} bootstrap sample. The purpose of using the (*) expression here is to obtain each Survival tree from the bootstrap data or in bag data.

Bootstrap ensemble cumulative hazard function formula is:

$$H_e^*(t|x_i) = \frac{1}{B} \sum_{b=1}^B H_b^*(t|x_i). \quad (3.11)$$

- Prediction error is calculated for the ensemble cumulative hazard function or survival function using out of bag data.
- $I_{i,b}$: The index i indicates whether the data is in bag or out of bag data ($I_{i,b}=1$ is obtained from out of bag data)

The out of bag ensemble estimator for the cumulative hazard function can then be calculated as follows:

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^B I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^B I_{i,b}}. \quad (3.12)$$

The expression (**) in formula 3.12 indicates that the estimator is an out of bag estimator. With the ensemble cumulative hazard function formula, the average of the cumulative hazard functions of the out of bag data in the B bootstrap sample is found.

When the above algorithm steps are examined, two features are encountered that reduce the error of the Random Survival Forest model, reduce the variance of the model, and give the model the randomness feature (Breiman, 1996). One is to choose randomly among many independent variables, the other is to select a sample from the data using the bootstrap method (Rezaei et al., 2020).

The Random Survival Forests method can consist of many randomly chosen independent variables. It is important to be able to interpret these variables and determine their significance for the model. Odds ratio used in classical statistical methods such as logistic regression can not be used in this method because there are censored observations in the data. Therefore, the importance rates of the variables in the method can be interpreted by looking at the variable importance. The important variables for the model are determined according to the importance index, namely VIMP. VIMP measures the total reduction in the error achieved via splits made with respect to that variable. Generally the scores are standardized such that the most important variable is scored as 100 or the total is a specific value etc.

If the variable importance value of the independent variables is large, it can be said that those variables can predict that response well. Otherwise, it can be said that those variables do not play a crucial role in explaining the variation of the do not play a response.

There are many advantages to using the random survival forests.

- The Random Survival Forests do not require any model assumptions.
- In this method, since independent variables are chosen randomly, the overall error of the model decreases and its variability decreases .

- Since it is an ensemble method, it can improve the prediction performance of the model.
- The tree is easily created by specifying a small number of parameters (Rezaei et al., 2020).
- It provides variable importance values that show which variables are more important for the model.

The model has advantages as well as shortcomings.

- As the number of trees increases, the running time of the algorithm in the computer environment increases.
- Values indicating the importance of variables such as odds ratio cannot be calculated for this method (Rezaei et al., 2020).

3.5.3 Performance Evaluation Metrics

In Chapters 2 and 3, survival analysis and machine learning methods in the literature for survival data are discussed. After applying both classical survival analysis and Survival tree and Random Survival Forests methods for the data, there is a need for criteria to compare the suitability of these methods to the data. Since survival data contains censored observations, most of the methods that measure fit of the model cannot directly be used.

In this thesis, Concordance index, which is a measure of model's appropriateness to the data, is explained to compare models applied to survival data.

The fit and estimation performance of the model are found by using the Concordance index created by (Harrell et al., 1982). The following steps are followed when calculating the Concordance index (Weathers, 2017):

- All possible couples are generated for the each observations. Couples consist of two observations.
- If there are observations with a shorter survival time among the couples, they are censored or if both of the observations within the couple have the same survival time but at least one of them has not occurred, it is necessary to exclude these couples from the cluster consisting of paired observations. The remainder of the couples are called permissible.
- Permissible couples are given the values of 1 and 0.5 for different situations. First of all, the situations where 1 is given are explained as follows.
 - In the case of $T_i \neq T_j$, the observation with assigned smaller survival.
 - In the case of $T_i = T_j$ the predicted survivals.
 - In the case of $T_i = T_j$ if the event does not occur in both observations, the observed event has lower predicted survival.

Permissible couples take the value 0.5 in this case.

- Although the event times of the observations are not equal, the predicted survivals are the same.
 - The event times are equal but the predicted survivals are not the same.
 - The survival times of the observations are equal, the event did not occur in both observations, and the predicted survival for the observed event is not smaller.
- Concordance index is found by dividing the total score to the number of permissible couples.

The error rate of the model is obtained by subtracting the concordance index value from 1. This ratio takes values between 0 and 1. And the closer the error rate is to 0, the better the model fit.

CHAPTER FOUR

ANALYSIS

4.1 Data Analysis

4.1.1 Introduction

In this thesis, Aids2 data in the MASS library in the R program was used as real data. First a brief information about AIDS is given. Then the dataset was introduced and analysis was made. In this data, Kaplan Meier curves were plotted to observe the survival curves of the variable groups. Cox model was established to look at the relationship between survival time and independent variables, but this model was not used because proportional hazard assumption could not be provided. A survival tree was created to observe the tree structure and how the independent variables were separated. Random survival forest method was used to calculate which variables are important for the model and to calculate the concordance statistics.

AIDS was first described in 1981. While the virus that caused this disease was not known at first, it was later found that HIV caused this disease (Babayigit & Bakır, 2004). This disease is contagious and can be transmitted from person to person in various ways. For example, blood contact with people with HIV, sexual contact with people who have the virus, transmission from mother to baby or through various tools such as needles can occur (Babayigit & Bakır, 2004). The disease is most common in individuals aged 15-49 years. The lack of a definitive treatment for the disease and the high prices of the drugs used by people are the reasons that make it difficult to fight this disease. In order to fight this disease, individuals should be given various trainings to raise awareness.

4.1.2 The Data

The study consists of data from 2843 AIDS patients in Australia. In this study, 1761 of 2843 patients died. The purpose of this study is to model the survival time of AIDS patients and to draw meaningful results from the independent variables that affect the survival times. Multiple variables affecting survival times were used in the study. Detailed information about these variables is as follows.

Table 4.1 Variables of AIDS Data

Variables	Class of Variables	Variable Levels
State	categorical	NSW, VIC, QLD, Other
T.categ	categorical	hs, hsid, id, het, haem, blood, mother, other
Diagnosis	date	-
Death	date	-
Status	categorical	alive, dead
Sex	categorical	male, female
Age	continious	-

- State: Regions in Australia (NSW: New South Wales, VIC: Victoria, Queensland, Other: Western Australia, South Australia, Northern Territory, Tasmania)
- T. categories: Ways of transmission to patients
 - hs: male sexual orientation (homosexual, bisexual)
 - hsid: men taking intraveneous drug
 - id: drug use by women or heterosexual men
 - het: contact with heterosexual indivuals
 - haem: blood disease
 - blood: blood draw
 - mother: HIV positive mother or contact mother

- other: other reasons
- Diagnosis: Disease onset date for individuals
- Death: Date of death of patients from AIDS
- Status: Survival or death of patients
- Sex: Gender of the patients
- Age: Age of the patients

There are 5 independent variables in the data. The variable expressing the time of survival can be found by subtracting the time of disease onset from the day the patients died of AIDS.

4.1.3 Kaplan Meier Curves

Survival analysis can be started by first plotting Kaplan Meier curves. The image of the curve containing the estimates of survival probabilities is as in Figure 4.1.

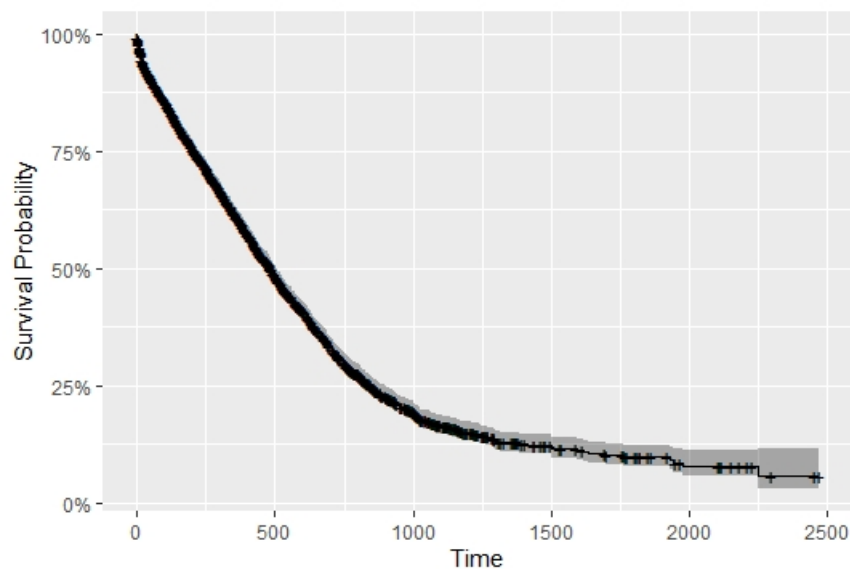


Figure 4.1 Kaplan Meier Curve

When Figure 4.1 examined, the y- axis consists of survival probabilities and the x- axis consists of the patients' survival time. Here, survival probabilities decreased with time. This situation can be explained by the characteristics of the survival function. This feature is that the probability of survival at the start of the study is 1, and the probability of survival approaches 0 as the study progresses and the risk increases.

Comparisons of Kaplan Meier curves according to variables including sex and transmission types of the disease are as follows.

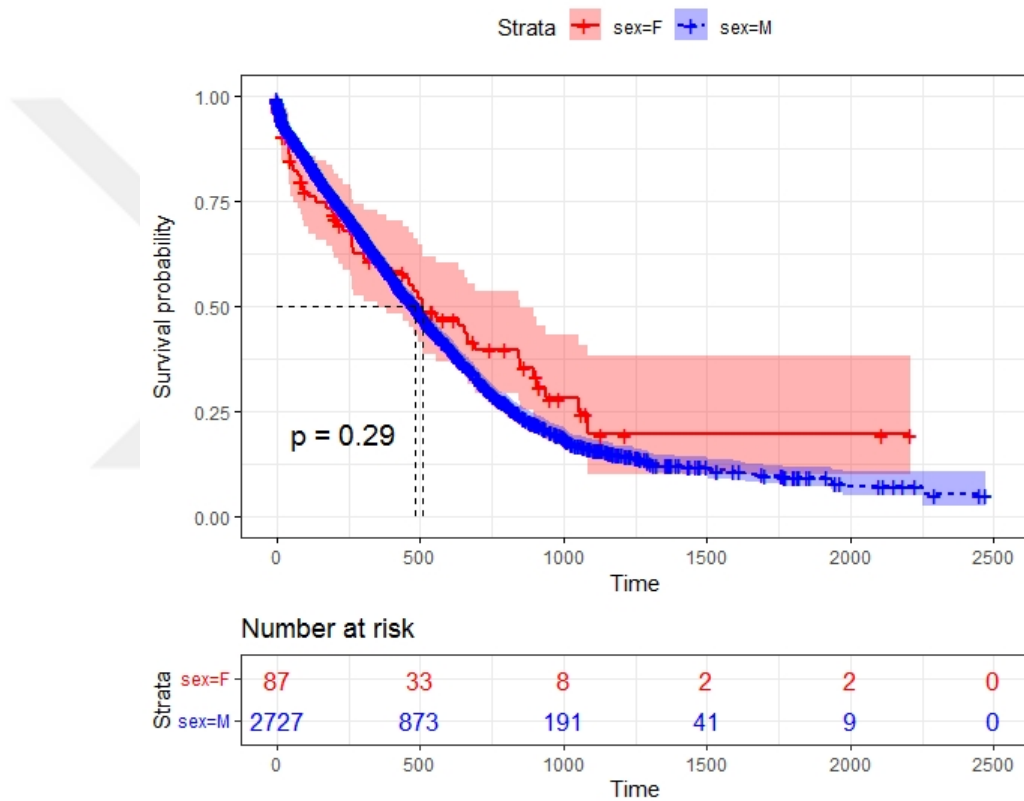


Figure 4.2 Kaplan Meier Curve for Gender

When the results are examined, women's struggle with the disease, that is, their life span is better than men. Women with this disease may live longer than men.

Table 4.2 Log-rank Statistics for Gender

	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
sex=F	87	51	59	1.0904	1.13
sex=M	2727	1682	1674	0.0384	1.13

Table 4.2 tests whether the survival probabilities of men and women are equal. Considering the p value (0.3), the null hypothesis cannot be rejected and it can be said that there is no significant difference between the survival probabilities of men and women.

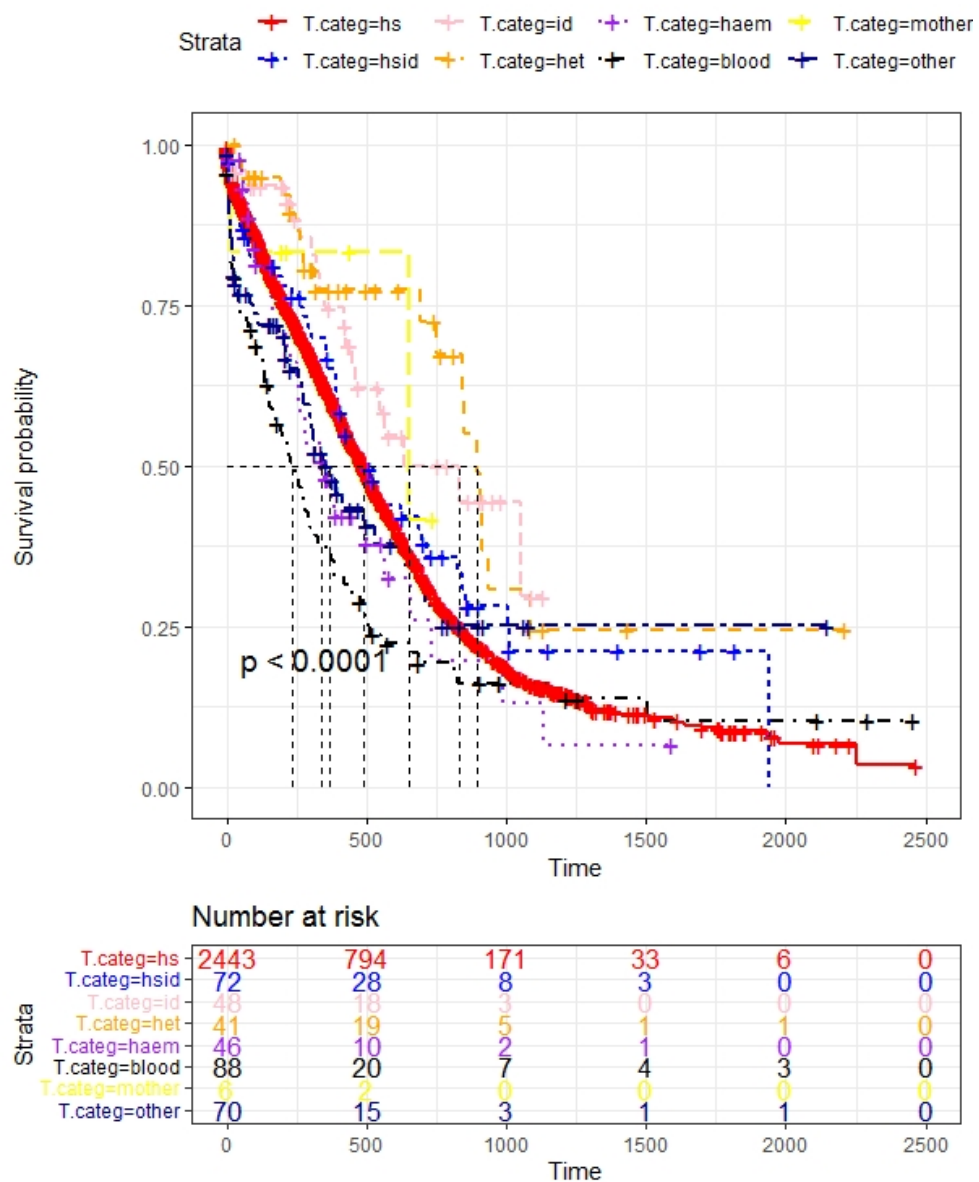


Figure 4.3 Curves for Each Type of Transmission

In Figure 4.3, the curves of the subvariables showing the ways in which the virus is transmitted to the patients are plotted. The survival rate of the male sexual orientation subvariable and patients with blood disease is low.

Table 4.3 Log-rank Statistics for Transmission

	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
T.categ=hs	2443	1511	1506.75	0.012	0.0922
T.categ=hsid	72	45	1.62	0.848	0.8765
T.categ=id	48	19	33.58	6.332	6.4722
T.categ=het	41	17	34.41	8.806	9.0136
T.categ=haem	46	29	22.55	1.847	1.8764
T.categ=blood	88	70	46.45	11.945	12.4113
T.categ=mother	6	2	3.43	0.599	0.6017
T.categ=other	70	40	34.22	0.977	0.9988

Considering the p- value (5e-05) in Table 4.3, it was seen that there was a significant difference between the transmission routes of the disease in terms of survival.

4.1.4 Cox Model Results

After examining the survival curves, it is important to find out which variables are more effective in the survival time of the patients. For this the Cox regression model was established and the results were examined.

When the Cox model is established, it should first be checked whether the assumptions are met. It is difficult to provide a Cox proportional hazard assumption.

Table 4.4 Cox Proportional Hazard Model Assumption

	chisq	df	p
age	22.70	1	1.9e-06
sex	2.47	1	0.11628
T.categ	29.21	7	0.00013
GLOBAL	43.47	9	1.8e-06

Cox proportional hazard assumptions could not be provided in this data set. In this case, interaction terms can be added to the model or variables can be stratified. The age variable has been stratified to provide the proportional hazard assumption, and the assumption is still not available. Backward elimination was applied to the model in order to provide an assumption. The concordance value is 0.55 when the assumptions are not met.

4.1.5 Survival Tree

After the models were established with traditional survival analysis methods and the results were interpreted, the analysis continued with the methods in which machine learning methods were applied to the survival data.

The `ctree` command in the `partykit` (Hothorn & Zeileis, 2015) package was used to create the tree. As a splitting rule, which is one of the important steps in the creation of the tree, node with a log rank p value less than 0.05 are divided into sub nodes.

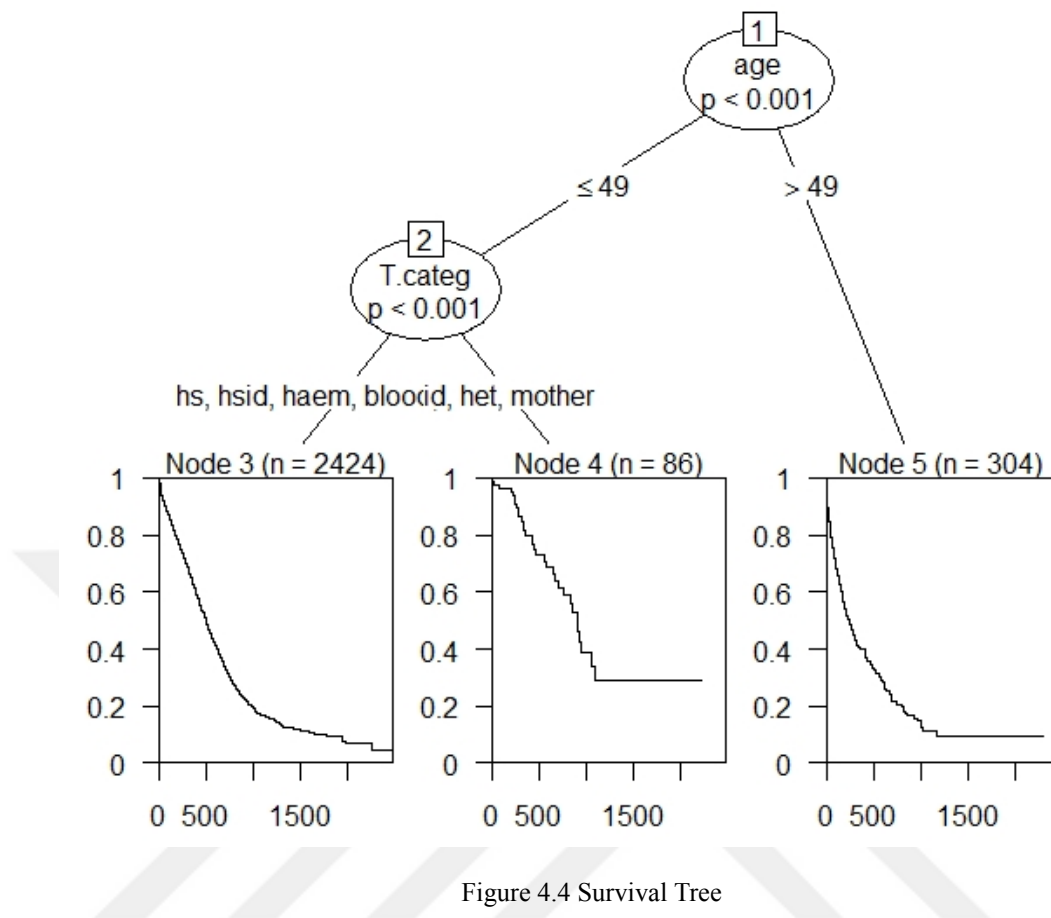


Figure 4.4 Survival Tree

The tree is first split into decision nodes based on those older or younger than 49 years old. The tree is then divided into terminal nodes according to the transmission type of the disease.

4.1.6 Random Survival Forests

After the model was established with a single tree, the random survival forests method, which averages more than one Survival tree, was tried. The RandomForest SRC package has been downloaded to implement the random survival forests method.

This package contains different splitting rules for survival data. For example, Brier score, logrank score (Hothorn & Lausen, 2003) and logrank statistics are included as separation rules in this package. Logrank statistics were used as the separation rule in

the analysis.

The Random Survival Forests consists of multiple Survival trees, and the variables in the Survival tree are chosen randomly. The number of variables used randomly for each tree in the R program is accepted as the number of variables divided 3 as a default. Since there were 2 variables in the data, the randomly selected variables was accepted 1. 500 trees were created.

Variable importance value is also important in interpreting the analysis. Looking at Figure 4.5, it was seen that the age variable was important for the model.

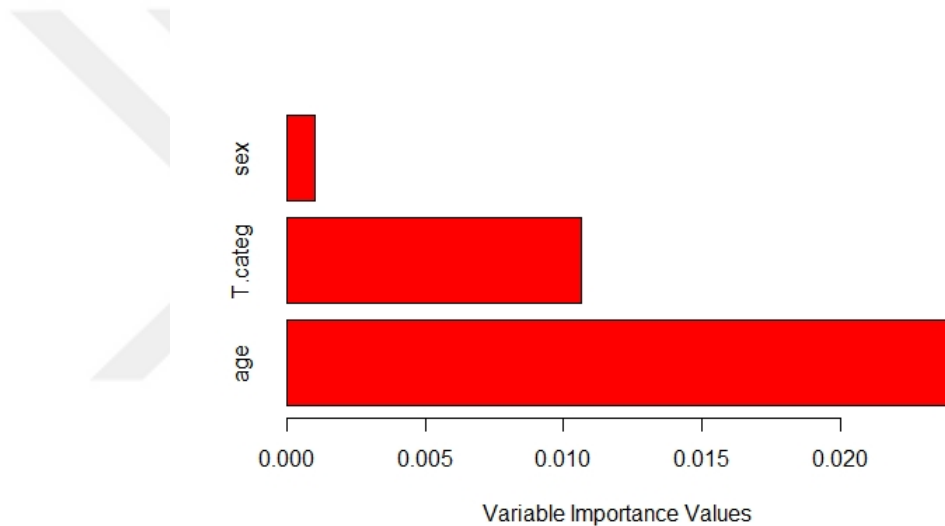


Figure 4.5 Variable Importance Value for Variables

4.1.7 Concordance Index

After the models are built, various methods are used to make comparisons. In this thesis, model comparison is made by looking at the concordance index values.

The concordance value of Cox regression was 0.55, and the concordance value of Random Survival Forests was 0.668. Since the proportional hazard assumption does not hold. As expected here, the fit of the Random Survival Forests to the model is better than the Cox regression.

4.2 Simulation

As seen from the real data analysis in the previous section, when the proportional hazard assumption is not met machine learning based survival analysis approaches can benefit the researchers. On the other hand when the assumptions are met, it is basic knowledge that statistical approaches will have more power and therefore it is suspected that when the data allows, statistical approaches may be more beneficial. To test this phenomenon, simulated data sets satisfying the proportional hazards assumption are used to compare the concordance statistics values of two models and by obtaining beta values of Cox regression and variable importance values of Random Survival Forests with different censoring degrees, the performance of the models are compared. The simulation study was inspired by the models in article (Zhou & McArdle, 2015). 400 observations and 2 independent variables were used in the study. The data is divided into 70% train and 30% test. In the data firstly, independent variables with 400 two dimensional observations uniformly distributed were randomly generated. The log hazard value in Model 4.1 was used. A variable that expresses whether the data is censored time or event time has been created. Finally, data including independent variables, event time, status and hazard values was created. There are 120 observations in the test data and 280 observations in the train data. The model used is as follows:

$$\vartheta_x : X1 + 2X2 + X1 * X2 \quad (4.1)$$

ϑ_x is defined as the logarithm of the hazard function. The model consists of $X1$, $X2$ and the interaction of both. The event times were obtained from the uniform distribution and the upper parameter of the uniform distribution was named gamma and the values of this upper parameter were found at different censoring percentages.

Table 4.5 Gamma Values According to Different Censorship Percentages

	%23	%50	%76
Gamma value	0.0795	0.275	0.93

Trees have been created for different censoring rates. Each forests consists of 500 subtrees. Since the Random Survival Forests method used randomly selected variables, not all independent variables, and the model consisted of 2 variables, the number of randomly selected variables was accepted as 1. This calculation can be obtained by taking the square root of the number of variables or dividing the number of variables by 3.

A cycle with 1000 repetitions was established and the results including the average beta and variable importance values for 2 variables and the average concordance values for both models were obtained.

Average beta values for 1000 trials in Cox regression are as in Table 4.6.

Table 4.6 Cox Model Coefficients for Different Censorship Percentages

	23% censoring	50% censoring	76% censoring
X1	1.062384	2.101149	0.9505468
X2	0.9833659	2.026415	1.036406
X1*X2	1.024052	2.036217	0.9766008

Cox could not on average define the variables contributions to the survival, since the model actually included X2 as a more dominant feature.

Table 4.7 Average Variable Importance Values of Random Survival Forest for Different Censorship Percentages

	23% censoring	50% censoring	76% censoring
X1	0.03659849	0.0362266	0.03599947
X2	0.1130446	0.1196622	0.1210422

Random Survival Forests on the other hand successfully determined X2 as the more important variable.

Average concordance values for 1000 trials in Random Survival Forests and Cox

Model are as in Table 4.8.

Table 4.8 Average Concordance Values of Models for Different Censorship Percentages

	23% censoring	50% censoring	76% censoring
Cox Regression	0.7367769	0.7231852	0.7123667
Random Survival Forests	0.6878896	0.6943654	0.6931765

Concordance values are expected to decrease as the percentage of censored observations increases. As the censoring percentages increase in the Cox regression, the concordance value decreases and increases in the Random Survival Forests. When the Table 4.8 examined, the concordance values of the Cox model are slightly better than the Random Survival Forests. But the two approaches becomes similar as censoring rate increase.

CHAPTER FIVE

SUMMARY AND CONCLUSION

Survival analysis provides the analysis of the time until the event of interest in the study and has a wide application area. This type of analysis has a unique data structure. There are censored observations in the data and it is difficult to deal with these observations. Censored observations are observed up to a certain time through out the study, but were dropped from the study for other reasons. In other words, partial information about these observations is obtained in the study. This can make the analysis difficult.

There are many studies in the literature about machine learning methods that are useful for survival data. In addition to traditional survival analysis methods, the use of machine learning methods has several advantages. Machine learning methods have been proposed because both the number of variables and rows increase in the survival data, the censored observations in it make the analysis difficult, and the models must provide various assumptions, and it is hard to justify these assumptions for real data.

In this thesis, traditional survival analysis methods and machine learning methods suitable for survival data are explained. Survival trees, random survival forests, Cox regression and Kaplan Meier methods were used in this study. Survival tree and random survival forests are tree based methods. The feature that distinguishes these two methods from other tree based methods is that they use different splitting and pruning rules.

Both real data and simulated data are used in this thesis. Kaplan Meier was used to compare survival curves and Cox regression, random survival forests and survival trees to examine variables affecting survival time. Concordance index values were used to compare the model fit of Cox regression and Random Survival Forest on real and simulated data. Model performance of Random Survival Forest is expected to be better than classical methods when the assumptions are not met. This claim has also been confirmed for AIDS data. Random Survival Forest was found to fit the model

better. Other times, when the focus is on prediction only, both approaches can benefit researchers however when the data satisfies the assumptions needed for parametric or semi-parametric statistical approaches, provide the extra benefit of interpretability and inference.



REFERENCES

- Al-Shomrani, A. A., Shawky, A., Arif, O. H., & Aslam, M. (2016). Log-logistic distribution for survival data analysis using mcmc. *SpringerPlus*, 5(1), 1–16.
- Babayigit, M. A., & Bakır, B. (2004). Hiv enfeksiyonu ve aids: Epidemiyoloji ve korunma. *TSK Koruyucu Hekimlik Bülteni*, 3(11), 280–290.
- Baca-Garcia, E., Perez-Rodriguez, M. M., Saiz-Gonzalez, D., Basurte-Villamor, I., Saiz-Ruiz, J., Leiva-Murillo, J. M., de Prado-Cumplido, M., Santiago-Mozos, R., Artés-Rodríguez, A., & De Leon, J. (2007). Variables associated with familial suicide attempts in a sample of suicide attempters. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(6), 1312–1316.
- Badem, H. (2019). Parkinson hastalığının ses sinyalleri üzerinden makine öğrenmesi teknikleri ile tanımlanması. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 8(2), 630–637.
- Balakrishnan, N., & Kateri, M. (2008). On the maximum likelihood estimation of parameters of weibull distribution based on complete and censored data. *Statistics & Probability Letters*, 78(17), 2971–2975.
- Beadle, G. F., Come, S., Henderson, I. C., Silver, B., Hellman, S., & Harris, J. R. (1984). The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *International Journal of Radiation Oncology* Biology* Physics*, 10(11), 2131–2137.
- Beadle, G. F., Silver, B., Botnick, L., Hellman, S., & Harris, J. R. (1984). Cosmetic results following primary radiation therapy for early breast cancer. *Cancer*, 54(12), 2911–2918.
- Behera, M. (2012). *A multifaceted analysis of early stage non-small cell lung cancer data*. PhD thesis, University of South Florida, Florida.
- Bou-Hamad, I., Larocque, D., Ben-Ameur, H. et al. (2011). A review of survival trees. *Statistics Surveys*, 5, 44–71.

- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton: CRC press.
- Brown, G. (2010). Ensemble learning. *Encyclopedia of Machine Learning*, 312, 15–19.
- Callahan, A., & Shah, N. H. (2017). Machine learning in healthcare. In *Key Advances in Clinical Informatics* (279–291). London: Academic Press.
- Carter, J., Blackmore, E., Sutandar-Pinnock, K., & Woodside, D. (2004). Relapse in anorexia nervosa: a survival analysis. *Psychological Medicine*, 34(4), 671.
- Chung, C.-F., Schmidt, P., & Witte, A. D. (1991). Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1), 59–98.
- Ciampi, A., Thiffault, J., Nakache, J.-P., & Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, 4(3), 185–204.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.
- Crowley, J., LeBlanc, M., Gentleman, R., & Salmon, S. (1995). Exploratory methods in survival analysis. *The Institute of Mathematical Statistics Lecture Notes–Monograph Series*, 27, 55–77.
- Davis, R. B., & Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, 8(8), 947–961.
- Deconinck, E., Hancock, T., Coomans, D., Massart, D., & Vander Heyden, Y. (2005). Classification of drugs in absorption classes using the classification and regression trees (cart) methodology. *Journal of Pharmaceutical and Biomedical Analysis*, 39(1-2), 91–103.

- Emmert-Streib, F., & Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3), 1013–1038.
- Epstein, R., Roberts, G., & Beber, G. (2009). *Parsing the Turing Test*. Netherlands: Springer.
- Friedman, J., Hastie, T., Tibshirani, R. et al. (2009). *The elements of statistical learning*. New York: Springer.
- Friedman, J., Hastie, T., Tibshirani, R. et al. (2013). *An introduction to statistical learning*. New York: Springer.
- Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*, 19(1), 1–17.
- Gokovali, U., Bahar, O., & Kozak, M. (2007). Determinants of length of stay: A practical use of survival analysis. *Tourism Management*, 28(3), 736–746.
- Gordon, L., & Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports*, 69(10), 1065–1069.
- Härdle, W., Mori, Y., & Vieu, P. (2006). *Statistical Methods for Biostatistics and Related Fields*. Berlin: Springer Science & Business Media.
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18), 2543–2546.
- Hothorn, T., & Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2), 121–137.
- Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research*, 16(1), 3905–3909.
- Hoy, M. B. (2018). Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1), 81–88.

- Intrator, O., & Kooperberg, C. (1995). Trees and splines in survival analysis. *Statistical Methods in Medical Research*, 4(3), 237–261.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. et al. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3), 841–860.
- Ishwaran, H., & Lu, M. (2014). Random survival forests. *Wiley StatsRef: Statistics Reference Online*, 2(3), 1–13.
- Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2), 105–115.
- Jones, J. L. (2006). Robots at the tipping point: the road to irobot roomba. *IEEE Robotics & Automation Magazine*, 13(1), 76–78.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237–285.
- Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology*, 22(7), 263–270.
- Kaynar, O., Tuna, M. F., Görmez, Y., & Deveci, M. A. (2017). Makine öğrenmesi yöntemleriyle müşteri kaybı analizi. *Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 18(1), 1–14.
- Kim, J. (2012). Survival analysis. *Pediatrics in Review / American Academy of Pediatrics*, 33, 172-4.
- Klein, J. P., & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. New York: Springer.
- Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis*. New York: Springer.
- LeBlanc, M., & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 48(2), 411–425.

- LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422), 457–467.
- Lee, E. T., & Go, O. T. (1997). Survival analysis in public health research. *Annual Review of Public Health*, 18(1), 105–134.
- Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis*. New Jersey: John Wiley & Sons.
- Miao, F., Cai, Y.-P., Zhang, Y.-X., Fan, X.-M., & Li, Y. (2018). Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access*, 6, 7244–7253.
- Moisen, G. (2008). Classification and regression trees. *Encyclopedia of Ecology*, 1, 582–588.
- Molinaro, A. M., Dudoit, S., & Van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1), 154–177.
- Munakata, T. (1996). Thoughts on deep blue vs. kasparov. *Communications of the ACM*, 39(7), 91–92.
- Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., & Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British Journal of Cancer*, 35(1), 1–39.
- Pirooznia, M., Yang, J. Y., Yang, M. Q., & Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(1), 1–13.
- Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika*, 60(2), 279–288.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.

- Rezaei, M., Tapak, L., Alimohammadian, M., Sadjadi, A., & Yaseri, M. (2020). Review of random survival forest method. *Journal of Biostatistics and Epidemiology*, 6(1), 59–68.
- Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3), 331–336.
- Russell, S., & Norvig, P. (2010). *Artificial intelligence: a modern approach*. New Jersey: Pearson Education.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
- Savage, V. M., & West, G. B. (2007). A quantitative, theoretical framework for understanding mammalian sleep. *Proceedings of the National Academy of Sciences*, 104(3), 1051–1056.
- Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, 44(1), 35–47.
- Shimokawa, A., Kawasaki, Y., & Miyaoka, E. (2015). Comparison of splitting methods on survival tree. *The International Journal of Biostatistics*, 11(1), 175–188.
- Shimokawa, A., Kawasaki, Y., & Miyaoka, E. (2016). A comparative study on splitting criteria of a survival tree based on the cox proportional model. *Journal of Biopharmaceutical Statistics*, 26(2), 386–401.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
- Su, J., & Zhang, H. (2006). A fast decision tree learning algorithm. *In AAAI*, 6, 500–505.
- Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1), 147–160.

- Tiwari, G., Bangdiwala, S., Saraswat, A., & Gaurav, S. (2007). Survival analysis: Pedestrian risk exposure at signalized intersections. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(2), 77–89.
- Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345), 169–173.
- Van Dijk, P. C., Jager, K. J., Zwinderman, A. H., Zoccali, C., & Dekker, F. W. (2008). The analysis of survival data in nephrology: basic concepts and methods of cox regression. *Kidney International*, 74(6), 705–709.
- Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6), 1–36.
- Weathers, B. (2017). *Comparision of survival curves between cox proportional hazards, random forests, and conditional inference forests in survival analysis*. Master's thesis, Utah State University, Utah.
- Zhang, Cha, M. Y. (2012). *Ensemble learning*. New York: Springer.
- Zhou, Y., & McArdle, J. J. (2015). Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80(3), 811–833.

APPENDICES

Appendix 1: R codes for real data

```
library(MASS)
library(survival)
library(ranger)
library(ggplot2)
library(dplyr)
library(ggfortify)
library(ranger)
library(randomForestSRC)
library(randomSurvivalForest)
library(rpart)
library(party)
library(partykit)
library(pec)
library(survminer)

Aids2
dim(Aids2)
head(Aids2)
aids2<-Aids2
aids2[,8]<-aids2$death-aids2$diag
head(aids2)
colnames(aids2)[8]<-"time"
aids<-aids2[,c("sex","status","T.categ","age","time")]
head(aids)
aids <- aids[which(aids$time > 0), ]
aids$status <- gsub("A", 0, aids$status)
aids$status <- gsub("D", 1, aids$status)
aids$status <- as.numeric(aids$status)
```

```

x=table(aids$status[aids$sex=="F"])
y=table(aids$status,aids$sex)
barplot(y)
barplot(x)
kaplan meier
sfit <- survfit(Surv(time, status) 1, data=aids)
sfit
summary(sfit)
km/sex/t.categ
sfitsex = survfit(Surv(time, status) sex, data=aids)
ggsurvplot(
sfitsex,
pval = TRUE,
conf.int = TRUE,
conf.int.style = "step",
xlab = "Time ",
break.time.by = 200,
ggtheme = theme_light(),
risk.table = "abs_pct",
risk.table.y.text.col = T,
risk.table.y.text = FALSE,
ncensor.plot = TRUE,
surv.median.line = "hv",
legend.labs =
c("Male", "Female"),
palette = c("red","blue"))
surv_diff = survdiff(Surv(time, status) sex, data = aids)
surv_diff
sfittrans = survfit(Surv(time, status) T.categ, data=aids)
ggsurvplot(sfittrans,
pval = TRUE, conf.int = F,

```

```

risk.table = TRUE,
risk.table.col =
linetype = "strata",
surv.median.line = "hv",
ggtheme = theme_bw(),
palette = c("red", "blue", "pink", "orange", "purple",
"black", "yellow", "darkblue"))
surv_diffit = survdiff(Surv(time, status) ~ T.categ, data = aids)
surv_diffit
cox model
coxm=coxph(Surv(time, status) ~ age + sex + T.categ,
data = aids)
summary(coxm)
test assumption
library(riskRegression)
test.ph =cox.zph(coxm)
test.ph
aids$age=ifelse(aids$age<mean(aids$age),"1","0")
factor(aids$age)
coxassump=coxph(Surv(time, status) ~ strata(age) + sex + T.categ,
data = aids)
cox.zph(coxassump)
fitcox3 = selectCox(formul, data = aids, rule = "aic")
survival=Surv(time=aids$time,event=aids$status)
splot=ctree(survival ~ T.categ+age+sex,data=aids)
plot(splot)
ntree =1000
set.seed(123)
rasf= rfsrc(Surv(time,status) ~ age+T.categ+sex, aids,tree.err=T,
importance="permute",
ntree=ntree, forest=T,mtry=1)

```

Appendix 2: R codes for simulation data

```
coxcon25=0
rsfimp25=matrix(0,nrow=1000,ncol=2)
coxbeta25=matrix(0,nrow=1000,ncol=3)
coxbeta5=matrix(0,nrow=1000,ncol=3)
coxbeta75=matrix(0,nrow=1000,ncol=3)
rsfcon25=0
coxcon50=0
rsfimp50=matrix(0,nrow=1000,ncol=2)
rsfcon50=0
coxcon75=0
rsfimp75=matrix(0,nrow=1000,ncol=2)
rsfcon75=0
N=400
for(i in 1:1000)
X3 =matrix(runif(N * 2), ncol = 2)
hazard3 = rep(NA,N)
hazard3 = X3[,1]+2*X3[,2]+X3[,1]*X3[,2]
haz3 = exp(hazard3)
time3 = sapply(haz3, rexp, n = 1)
Choose gamma value for a desired censoring rate
gamma=0.0795
censtime3 = runif(N, min = 0, max = gamma)
event3=as.numeric(time3<=censtime3)
sum(event3==1)/400
lasttime3=pmin(time3,censtime3)
simdat3=data.frame(lasttime=lasttime3,event=event3,x1=X3[,1],
x2=X3[,2],haz=haz3)
indices=sample(nrow(simdat3),0.7*nrow(simdat3))
train=simdat3[indices,]
```

```

test=simdat3[-indices,]
surv3=Surv(time=train$lasttime,event=train$event)
tsurv3=Surv(time=test$lasttime,event=test$event)
cox3=coxph(surv3 x1+x2+x1*x2,data=train)
tcox3=coxph(tsurv3 x1+x2+x1*x2,data=test)
testcox3=survfit(cox3,newdata=test)
t1=survConcordance(Surv(lasttime, event) predict(tcox3),
data=test)
ntree = 500
v.out = rfsrc(Surv(lasttime,event) x1+x2,
train,tree.err=T,importance="permute",
ntree=ntree, forest=T,mtry=1)
coxcon25[i]=t1$concordance
rsfimp25[i,]=v.out$importance
rsfcon25[i]=predict.rfsrc(v.out,data=test)$err.rate[500]
coxbeta25[i,]=cox3$coefficients
X4=matrix(runif(N * 2), ncol = 2)
hazard4 = rep(NA,N)
hazard4 = X4[,1]+2*X4[,2]+X4[,1]*X4[,2]
haz4 =exp(hazard4)
time4 = sapply(haz4, rexp, n = 1)
Choose gamma value for a desired censoring rate
gamma=0.275
censtime4 = runif(N, min = 0, max = gamma)
event4=as.numeric(time4<=censtime4)
lasttime4=pmin(time4,censtime4)
simdat4=data.frame(lasttime=lasttime4,event=event4,x1=X4[,1],
x2=X4[,2],haz=haz4)
indices1=sample(nrow(simdat4),0.7*nrow(simdat4))
train1=simdat4[indices1,]
test1=simdat4[-indices1,]

```

```

surv4=Surv(time=train1$lasttime,event=train1$event)
tsurv4=Surv(time=test1$lasttime,event=test1$event)
cox50=coxph(surv4 x1+x2+x1*x2,data=train1)
tcox50=coxph(tsurv4 x1+x2+x1*x2,data=test1)
t2=survConcordance(Surv(lasttime, event) predict(tcox50),
data=test1)
ntree = 500
v.out50 = rfsrc(Surv(lasttime,event) x1+x2,
train1,tree.err=T,importance="permute",
ntree=ntree, forest=T,mtry=1)
rsfimp50[i,]=v.out50$importance
coxcon50[i]=t2$concordance
rsfcon50[i]=predict.rfsrc(v.out50,data=test1)$err.rate[500]
coxbeta50[i,]=cox50$coefficients
X5 = matrix(runif(N * 2), ncol = 2)
hazard5 = rep(NA,N)
hazard5 = X5[,1]+2*X5[,2]+X5[,1]*X5[,2]
haz5 = exp(hazard5)
time5 = sapply(haz5, rexp, n = 1)
Choose gamma value for a desired censoring rate
gamma=0.93 censtime5 = runif(N, min = 0, max = gamma)
event5=as.numeric(time5<=censtime5)
lasttime5=pmin(time5,censtime5)
simdat5=data.frame(lasttime=lasttime5,event=event5,x1=X5[,1],
x2=X5[,2],haz=haz5)
indices2=sample(nrow(simdat5),0.7*nrow(simdat5))
train2=simdat5[indices2,]
test2=simdat5[-indices2,]
surv5=Surv(time=train2$lasttime,event=train2$event)
tsurv5=Surv(time=test2$lasttime,event=test2$event)
tcox75=coxph(tsurv5 x1+x2+x1*x2,data=test2)

```

```
cox75=coxph(surv5 x1+x2+x1*x2,data=train2)
t3=survConcordance(Surv(lasttime, event) predict(tcox75),
data=test2)
ntree = 500
v.out75 = rfsrc(Surv(lasttime,event) x1+x2,
train2,tree.err=T,importance="permute",
ntree=ntree, forest=T,mtry=1)
rsfimp75[i,]=v.out75$importance
coxcon75[i]=t3$concordance
rsfcon75[i]=predict.rfsrc(v.out75,data=test2)$err.rate[500]
coxbeta75[i,]=cox75$coefficients
```