Comparison of Some Statistical and Machine Learning Models for Continuous Survival Analysis

Ndou Sedzani Emanuel

(Student No: 18016736)

A research submitted in partial fulfilment of the requirements for the degree of

Masters of Science in E Science

in the

Department of Mathematical and

Computational Sciences,

Faculty of Science Engineering and Agriculture.

Supervisor: **Dr T.B Mulaudzi**

Co-supervisor: **Dr. A Bere**

Date Submitted: July 17, 2024

Abstract

While statistical models have been traditionally utilized, there is a growing interest in exploring the potential of machine learning techniques. Existing literature shows varying results on their performance which is based on the dateset employed. This study will conduct a comparative evaluation of the predictive accuracy of both statistical and machine learning models for continuous survival analysis utilizing two distinct datasets: time to first alcohol intake and North Carolina recidivism data. LassoCV was used to select variables for both datasets by encouraging limited coefficient estimates. Kaplan-Meier survival curves were utilized to compare the survival distributions among groups of variables incorporated in the model, alongside the logrank test. The proposed methods include the Cox Proportional Hazards, Lasso-regularized Cox, Survival Trees, Random Survival Forest, and Neural Networks. Model performance was evaluated using Integrated Brier score (IBS), Area Under the Curve and Concordance index. Our findings shows consistent dominance of Neural Network (NN) and Random Survival Forest (RSF) models across multiple metrics for both datasets. Specifically, Neural Network demonstrates remarkable performance, closely followed by RSF, CoxPH and CoxLasso models with slightly lower performance, and Survival Tree (ST) consistently lags behind. This study can contribute to advancing knowledge and provides practical guidance for improving survival in recidivism and alcohol intake.

Key words: Survival analysis, Statistical models, Machine Learning models, Integrated Brier score, Concordance index, Area Under the Curve.

Declaration

I, Sedzani Emanuel Ndou of Student Number: 18016736, declare that the research titled "Comparison of some statistical and machine learning models for continuous survival analysis", conducted as part of my Masters of Science in E Science program at the University of Venda, is entirely my own work. I attest that it has not been previously submitted for any degree at any other university or institution. Additionally, I confirm that the content does not include the work of others unless explicitly acknowledged and properly referenced.

Student: ..... Date: 17 July 2024

Acknowledgements

First and foremost, I would want to express my deepest gratitude and appreciation to God of Mount Zion, who supported me during the whole process by providing blessings, wisdom, and direction, enabling me to finally finish this project.

I also extend my appreciation to my supervisor, Dr T.B. Mulaudzi, and my co-supervisor, Dr A. Bere, for their dedicated efforts that contributed greatly to the success of my research project, as well as their support and patience during my research project.

Finally, I want to express my gratitude to my family for their support and love.

Dedication

This research project is dedicated to the Republic of South Africa.

Contents

List of Figures

List of Tables

Abbreviations

ANOVA	Analysis of Variance
AUC	Area Under the Curve
ANN	Artificial Neural Network
BS	Brier Score
CSF	Conditional Survival Forest
EN	Elastic Net
GB	Gradient Boost
IBS	Integrated Brier Score
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Machine Learning
NN	Neural Network
PLANN	Partial Logistic Artificial Neural Network
PH	Proportional Hazards
RSF	Random Survival Forest
RSS	Residual Sum of Square
ST	Survival Tree
TVET	Technical and Vocational Education and Training

Chapter 1

Introduction

1.1 Introduction

Statistical survival analysis is a statistical method widely used to study the duration up until the occurrence of interesting event in various fields. It can be applied to diverse scenarios such as economics, biology, engineering, and more. For example, it can be used to examine the long term survival of events of interest, such as the duration from diagnosis to death, time from treatment initiation to disease recurrence, age at first birth, age at first marriage, etc. The response also known as the failure time, survival time, or event time. The start time must be specified so that individuals are as comparable as possible. For instance, if the survival duration of subjects with a specific kind of cancer is investigated, the start time may be set to the moment of discovery of that type of cancer. Similarly, the event of interest must be adequately detailed to ensure that the time periods examined are clearly defined. In the preceding instance, this might be death from the cancer under consideration. The period of time between the start time and the end time is then calculated (Kartsonaki, 2016). Survival analysis can also be employed to make comparisons of patient survival times with different conditions, treatments, or demographic characteristics.

1.2 Special Features of Survival Analysis

The potential for some people to miss the important event is one of the reasons why survival analysis involves "special" procedures. For instance, participants may withdraw from a research or experience a separate event, such as the death caused by an accident in the scenario above, which is unrelated to the end time of interest. Another potential is that the study could end at a certain time, in which case any participants whose event has not yet occurred will not have had their event time recorded. These insufficient observations must be treated differently yet cannot be ignored. Censoring is the term for this situation. Censoring occurs when we know certain details about subject survival time but not the specific survival time. There are three different types of censorship: left, interval, and right censoring (Mills, 2010; Klein and Moeschberger, 2003). Right censoring is one of the most common kind of censoring and the easiest to deal with in the analysis. This kind of censoring happens when a person participates in the study for a while but is subsequently no longer monitored. This may occur whenever a person retreats during the investigation or whenever the investigation is conducted for an extended time but fails to monitor some of the individuals by the conclusion of the period (due to time constraints). Another kind of censoring is left censoring this is the circumstance where it is known that a particular subject has already experienced the event at a specified time, however it might have happened at any moment before the censorship time. Then lastly, we have interval censoring which happens if we are aware that the relevant event happened at some point between two specified intervals, but the exact moment of failure is uncertain. Additionally, there are Type-I and Type-II censoring schemes which are common and popular censoring schemes (Hyun et al., 2016; Klein and Moeschberger, 2003). In Type-I censoring scheme, the experimental time is fixed, but the number of observed failures is a random variable. On the other hand, in Type-II censoring scheme, number of observed failures is fixed, but the experimental time is a random variable. Hybrid censoring combines aspects of Type-I and Type-II. It has both a predefined number of failures and a predefined fixed time. The

experiment can end when either condition is met (e.g., reaching failures or reaching time), or it might require both conditions (reaching failures and finishing time) (Klein and Moeschberger, 2003).

Another concept different from censoring is truncation (Mills, 2010; Klein and Moeschberger, 2003). A situation known as truncation occurs when the event of interest is entirely disregarded for some time and not noticed at all until a certain point in time. There are three types of truncation's which are left, interval, and right truncation. Left truncation is the most encountered type of truncation, where the subject is not monitored for some time at the start of the investigation but then comes under assessment late on. Right truncation happens when the whole research population had previously experienced an event of interest. Then we have interval or double truncation this type happens when an individual is under evaluation for the duration of the experiment, pulls out for a while, and then re-joins the study.

1.3 Discrete and Continuous Time Survival Analysis

The time to a event of interest in a long time, like a month, quarter, or year, is used in discrete time estimate. An event's time is recorded using accurate and fine metric, such as week, day, or hour, in continuous time estimation.

1.3.1 Discrete time survival analysis

The purpose of discrete time survival analysis (Cox, 1972) is to examine the likelihood that an event will occur when the time variable is discretely monitored. In other words, there are limited intervals of time that can be measured, and it is unknown when the event occurred. For instance, research might interview participants once a year for six years and inquire about a specific event each time. In situations like this, we frequently

don't remember the precise moment the event happened, but we do know it happened between Years 2 and 3, Years 4 and 5, etc. Thus, there are only five conceivable time frames in which the event might have taken place. Some of disadvantages of discrete data is that additional points of data (a larger sample) are required to generate an identical inference, and there are few possibilities for analysis, with minimal indication of sources of variance.

1.3.2 Continuous time survival analysis

Kvamme and Borgan (2021) highlighted the potential of Machine Learning model Neural Network to advance the field of survival analysis by providing tools for modeling continuous-time survival data. Continuous time survival analysis deals with time as a continuous variable and therefore allow for the interesting event to occur at any time point on the axis. The most used model for survival analysis, which analyzes time to event data is the Cox PH model, sometimes referred to as the Cox regression model. Because it is semi-parametric, it does not require precise knowledge of the underlying survival time distribution, but it does make some assumptions about it. This research will include Machine Learning models in continuous time survival analysis.

According to the Cox model, the ratio of the hazard functions for any two groups remains constant across time since the hazard function is assumed to be proportionate across groups as determined by predictor variables. This makes it possible to estimate hazard ratios, which are useful for comparing the likelihood of an event occurring in various groups. In medical research, it is frequently used to analyze survival data such as the length of time to death or the progression of a disease and to pinpoint potential risk factors for the event of interest.

1.4 Machine Learning Algorithms

In order to match nonlinear and complex interaction effects between variables and achieve a more accurate forecast of individual survival probability, machine learning methods have been applied for survival analysis. In order to handle limited time to event information, a number of machine learning techniques have also been developed. These techniques allow us to build more accurate survival prediction models than standard regression models for high-dimensional, complicated data. The most popular machine learning methods now in use for survival analysis are random survival forest, boosted Cox regression, penalized Cox regression, support vector regression, artificial neural networks and survival trees, and many others (Spooner et al., 2020). All mentioned machine learning algorithms have mechanisms for variable selection. However, it's important to validate the results of variable selection using additional techniques and consider the potential impact on model performance and interpretability.

1.5 Problem Statement

In survival analysis, statistical models have been widely used, but there is a growing interest in exploring the potential of machine learning models for more accurate predictions. However, the existing literature presents varying findings regarding the performance of machine learning and statistical models in survival analysis. Moreover, it is evident that the effectiveness of these predictive models varies significantly based on the characteristics of the datasets employed (Suresh et al., 2022; Zupan et al., 2000).

1.6 Research Questions

i. What is/are the most effective models in survival analysis?

ii. Between machine learning and statistical methods of survival analysis, which methods give the most accurate predictions?

1.7 Research Aim and Objectives

1.7.1 Research aim

The main aim of this project is to compare the out of sample predictive accuracy of selected machine learning and statistical methods of survival analysis using the age at first alcohol intake data and the North Carolina recidivism data set

1.7.2 Objectives

i. To fit machine learning and statistical models to predict the time until the first occurrence of alcohol intake and the likelihood of recidivism.

ii. To assess the calibration and discrimination capabilities of the fitted models.

iii. To conduct a comparative analysis of the calibration and discrimination performance of the fitted models.

1.8 Outline of the Dissertation

In the structure of the thesis, Chapter 2 is dedicated to the review of literature, while Chapter 3 outlines the methodology. Following this, Chapter 4 presents the results and discussion, and finally, Chapter 5 provides the conclusion.

Chapter 2

Literature Review

2.1 Introduction

This section provides some related studies from various authors implementing statistical methods in survival analysis and also studies about implementing machine learning algorithms in survival analysis.

2.2 Historical Literature of the Development and Application of the Cox Model

One of the most used models for doing the event's time, is the Cox Proportional Hazards (PH) survival model. The Cox PH model expresses the association between event occurrence and a variety of factors.

Sir David Cox (1972) first presented the Cox proportional hazard model in his seminal paper. He developed the model as an extension of the traditional life table methods and showed how it can be used to examine data on time to event in the presence of censoring (Cox, 1972). In the second related paper Cox (1975) presented the technique of partial likelihood, as an effective method for analyzing Cox model data that has

been extensively used in several different domains. Assuming no tied event times exist, he demonstrated how to estimate the regression parameter by maximising the partial likelihood.

The partial likelihood technique for estimating the parameters of the Cox proportional hazard model was separately developed by Efron (1977). Without requiring the specification of the underlying hazard function, this method enables efficient parameter estimation. In order to address competing risks, non-proportional hazards, and time-dependent covariates, researchers began to build modifications of the Cox proportional hazard model. These additions made it possible to analyze complex survival data in more detail. An essentially assumption-free alternative for the Cox proportional hazard model is presented by Therneau (1997).

The Cox proportional hazard model continued to be widely used in various fields and applications, including clinical trials, health services research, and econometrics (Murtaugh et al., 1999). Advances in computing technology and statistical software made it easier to fit the model to large datasets and to perform more complex analyses. Unlike the previous studies that used Cox proportional hazard models, Stratified Cox proportional hazard models were employed by Kelly (2004), to address variable non-proportionality problems with specific variables.

Researchers continued to develop new extensions of the Cox proportional hazard model to address emerging research questions, such as the analysis of genomic data and the estimation of treatment effects in observational studies (Therneau and Grambsch, 2000; Witten and Tibshirani, 2009; Austin, 2011). They also developed new methods for model selection, variable selection, and model validation in the context of the Cox proportional hazards model. Emmert-Streib and Dehmer (2019), reviewed the theoretical foundations of survival analysis, covering hazard and survival function estimators. They discussed more about the Cox proportional hazard Model as well as methods for

evaluating the proportional hazard (PH) premise. In addition, they consider stratified Cox models for situations where the proportional hazard assumption fails.

The Cox proportional hazard model has had a significant impact on the field of survival analysis and has contributed to a better understanding of the factors that influence the timing of events in a wide range of research areas. Kalbfleisch and Schaubel (2023) examined the 1972 Cox paper to see how it influenced subsequent research and applications. This work has also sparked additional research in a variety of fields, resulting in significant theoretical and practical advances. Semiparametric models, nonparametric efficiency, and partial likelihood are examples of these.

2.3 Historical Literature of the Development of the Discrete Survival Model

The concept of discrete time survival analysis was first introduced by Andersen (1970). Andersen (1970) described a method for modeling the probability of an event occurring in a specific time interval, given that the event did not occur in the previous time interval. This approach is now known as the "discrete-time hazard model" (Andersen, 1970).

Using logistic regression, (Cox, 1975) expanded the proportional hazards model to discrete times. When the time variable is discretely monitored, discrete time survival analysis is used to analyse the likelihood that an event will occur. Thompson Jr (1977) and Prentice and Gloeckler (1978b) were reportedly the first to take part in biometrics using discrete survival data. Although Thompson Jr (1977) was concentrating on the logistic model, (Prentice and Gloeckler, 1978b) created a score test for hypothesis testing by studying grouped Cox model. Laird and Olivier (1981) and Brown (1975) were early proponents of representing hazard models as binary Bernoulli trials.

Allison (1984) extended discrete-time survival model in his book. In this book, Allison (1984) presented the idea of modeling event occurrence as a binary outcome at a specific time interval, rather than modeling the time until the event occurred. Fahrmeir (1998) also provided an overview of discrete survival models.

Researchers started to develop extensions of the discrete-time survival model to handle complex data structures, such as competing risks, time-varying covariates, and missing data. The discrete temporal hazard model's parameters are simple to read, and traditional logistic regression analysis can easily fit them (Singer and Willett, 1993). Extensions to these models were made by Mantel and Hankey (1978), who replaced the baseline hazard parameters with a polynomial. Efron (1988) suggested modeling the several baseline hazard parameters with regression splines and extensively discussed the use of the discrete hazard model's binary response model representation.

The discrete-time survival model continued to be widely used in research, particularly in the analysis of panel data and repeated events. Researchers developed new methods for model selection, variable selection, and model validation in the context of the discrete-time survival model. Muthén and Masyn (2005) proposed a general latent variable method for discrete-time survival analysis of unpredictable events. Finkelstein (1986) proposed the proportional hazards model for interval censored data, which was further developed by Cai and Betensky (2003) with the addition of smooth effects in additive models. Rabinowitz et al. (1995) and Breiman (2001) developed methods for estimating accelerated failure time models, Zeng et al. (2006) and Wang et al. (2010) studied additive hazards models.

Bayesian approaches were proposed by Fahrmeir and Knorr-Held (1997) who presented a adaptable non parametric Bayesian analysis for dynamic models with effects that change over time. Fahrmeir and Kneib (2011) investigated simulation-based full Bayesian Markov chain Monte Carlo (MCMC) inference for longitudinal and event his-

tory data.

Researchers continued to review the discrete-time survival model to address emerging research questions, such as the analysis of big data and the estimation of causal effects. Kim (2014) recommand discrete models when there are low censoring proportions, large time metrics, or limited sample numbers. Tutz et al. (2016) showed some advantages of discrete event times. One benefit of discrete time to event models is that conditional probabilities can be used to express dangers.

The discrete-time survival model has significantly influenced the study of survival analysis and helped to advance knowledge regarding the timing of events in various fields. Its flexibility and ease of interpretation have made it a popular tool for analyzing time to event data in both theoretical and applied research.

2.4 Machine Learning Methods in Survival Analysis

The community of machine learning has created numerous effective techniques in a variety of fields, including survival analysis. Researchers began exploring the application of machine learning algorithms in survival analysis, primarily focusing on decision trees, random forests, and neural networks. These algorithms showed promising results in terms of predictive accuracy but were limited in their ability to handle complex survival data structures. Machine learning techniques have made substantial progress in several practical fields over the past few years thanks to their advantages including their capacity to model non-linear interactions and the accuracy of their overall predictions. The inability to handle filtered information and the model's time estimation effectively is the fundamental obstacle that machine learning approaches face while doing survival analysis. When there are many examples in a manageable dimensional feature space, machine learning is beneficial, but this isn't always the case for survival analysis tasks (Zupan et al., 2000).

Rosenblatt (1958) released a first paper about artificial neural network. This method creates a network by connecting the straightforward artificial nodes known as "neurons" based on a weighted link. In the early 1990s, Cox and colleagues introduced the use of neural networks in survival analysis. Based on a variety of clinical and demographic characteristics, they created a neural network model to forecast the hazard function of a patient's time to event outcome. When it came to estimating the chance of death in leukemia patients, the model performed admirably (Cox et al., 1990). The non-linear ANN predictor was added to the Cox proportional hazard model, and fitting a neural network with a single logistic hidden layer and a linear output layer was proposed (Faraggi and Simon, 1995). Mariani et al. (1997) employs two models: a neural network and the conventional Cox model. Faraggi and Simon (1995) evaluate the risk variables for breast cancer recurrence. Biganzoli et al. (1998) employ the partial logistic artificial neural network approach to explore the link among the factors and the survival times in order to improve the model's prognosis. Lately, feed-forward neural networks have been utilized to generalize both discrete and continuous time models to generate a more flexible non-linear model that takes into account the filtered information in the data (Biganzoli et al., 1998). Error functions and data encoding for multilayer perceptron and radial basis function expansions of generalized linear models for survival data were introduced by Biganzoli et al. (2002). Huang et al. (2004) proposed a hybrid neural network model for predicting the survival time of patients with liver cancer. The model combined a neural network with a genetic algorithm to select the most relevant features for survival prediction. ANNs have been shown to be effective in survival analysis and have been applied to a wide range of cancers and other diseases. The use of deep learning techniques in survival analysis is an area of active research and holds promise for further improving the accuracy of survival predictions.

The survival tree is a sort of regression and classification tree designed to deal with censored data. The fundamental idea behind tree models is to split the data recur-

sively according to a specific splitting criterion, and the objects that are comparable to one another in terms of the event of interest will be located in the identical node. In the year 1981, the earliest attempt was made to use a tree structure for survival statistics (Ciampi et al., 1981). However, Gordon and Olshen (1985) is the first publication to discuss how survival trees are made. A survival tree's capacity to handle censored data using the tree structure is its key advantage over a traditional decision tree. The choice of the final tree is a crucial step in creating a survival tree. The best tree can be chosen using techniques like backward selection or forward selection (Bou-Hamad et al., 2011). But, compared to a single tree, an ensemble of trees can perform better and avoid the issue of final tree selection (Dietterich, 2000).

Breiman (2001) recently demonstrated that ensemble learning can be enhanced further by adding randomness to the initial learning process, a method known as random forests. The random survival forest developed by Ishwaran et al. (2008) and later expanded to the competing risks context by Ishwaran et al. (2014) is a totally non-parametric and does not support left-truncation. Another well-liked implementation of random forests (Jaeger et al., 2019) only accepts proportional hazards models and right-censored data. Wright and Ziegler (2015) introduce and evaluate the extension of the oblique random survival forest, but still with the same limitation of using the right-censored survival.

Cruz and Wishar (2006) and Wang et al. (2019) provided an overall overview of machine learning methods for survival analysis. Their survey showed that machine learning algorithms are rapidly evolving field and offer potential for more accurate predictions in survival analysis. Kvamme and Borgan (2021) explored the use of neural networks in survival analysis and discrete-time modeling.

2.5 Comparison of Statistical and Machine Learning Models in Survival Analysis

Spooner et al. (2020) explored the application of machine learning algorithms and feature selection methods for dementia prediction using high-dimensional, heterogeneous, and censored clinical data. Various machine learning models were compared, including penalized gradient boosting, random survival forests, and Cox regression machines. The study found that gradient boosting and random survival forests machines generally outperformed the other models in terms of prediction accuracy. Metrics like the area under the receiver operating characteristic curve and C-index were used to assess model performance.

In order to predict clinical deterioration, Churpek et al. (2017) contrasted machine learning techniques with traditional regression. The study contrasted random forest, logistic regression, and machine learning techniques with multiple regression models. Evaluation metrics used included the area under the precision-recall curve, specificity, accuracy, sensitivity. Regarding the statistical methods, logistic regression with L2 regularization and the eCART score performed similarly, but they had lower predictive accuracy and AUC-ROC compared to the machine learning method.

In order to compare Cox proportional hazards with two cutting edge machine learning methods for disease prediction in nasopharyngeal cancer (NPC) prognostication, Oei et al. (2021) used DeepSurv and conditional survival forest (CSF). Based on the outcomes of the hazard stratification, the comparison of model performance was evaluated using the Brier Score, concordance index (c-index), and logrank test. In comparison to conditional survival forest and Cox proportional hazards the results indicated that DeepSurv with C-index = 0.68, Brier score = 0.13, logrank test p = 0.02 performed the best. In their comparatively limited datasets, conditional survival forest and DeepSurv both fared better than Cox proportional hazards.

Chowdhury et al. (2023) compared different machine learning algorithms with the conventional Cox proportional hazards model for predicting hypertension incidence using survival data. The researchers developed five machine learning algorithms penalized regression Ridge, Lasso, Elastic Net (EN), random survival forest, and gradient boosting along with the conventional Cox proportional hazard model with the data from 18,322 participants in the Alberta's Tomorrow Project and considered 24 candidate features to build the prediction models. The predictive performance of the machine learning algorithms was observed to be similar to that of the conventional Cox proportional hazard model. The average C-index values for CoxPH, GB, RSF, EN, Ridge, and Lasso were 0.77, 0.76, 0.76, 0.78, 0.78, and 0.77, respectively. The study found that conventional regression based models, like Cox proportional hazard, performed well and exhibited little predictive performance difference compared to machine learning algorithms in predicting hypertension incidence in this moderate sized dataset with a reasonable number of features.

Zupan et al. (2000) compared the effectiveness of three models: the Naive Bayes Classifier, Decision Tree Induction, and Cox Proportional Hazards Model, using two datasets (preoperative and postoperative patients). Correlation between predicted probability and classification accuracy, specificity, sensitivity and Kaplan-Meier estimates, and concordance index were among the evaluation criteria used. The results indicated that, although the differences were not statistically significant, decision trees performed worse than Naive Bayes and Cox proportional hazard models. The C-index for Naïve Bayes, Decision tree, Cox proportional hazard were 0.75, 0.72 and 0.76, respectively.

Suresh et al. (2022) introduced a discrete-time survival framework using machine learning for survival predictions. The study compared the predictive performance of various models, including Cox Proportional Hazards, Elastic Net, Support Vector Machine, Gradient Boosting Machine, Random Survival Forest, Logistic Regression, Neural Network, and Conditional inference Forest, on five distinct datasets. The results revealed that certain models, such as Conditional inference Forest and Random Survival Forest, consistently demonstrated strong predictive capabilities across multiple datasets. Notably, Random Survival Forest and Conditional inference Forest consistently achieved high C-index values (above 0.8) across all datasets, indicating strong predictive ability. The Cox Proportional Hazards model also showed competitive performance. Conversely, Support Vector Machines consistently underperformed across all datasets.

2.6 Conclusion

One notable and consistent finding across the studies is that the performance of predictive models in survival analysis varies considerably with the datasets used. Different machine learning algorithms and statistical models may yield varying degrees of accuracy and effectiveness depending on the specific characteristics of the dataset. There is also a need to investigate how Machine Learning and Statistical Models differ in analyzing the types of data and how these preferences impact the performance and reliability of the models.

This study aims to conduct a comparison between machine learning and statistical models in survival analysis, using two datasets which are not yet applied in the comparison of machine learning and statistical models in survival analysis. We will employ various sets of evaluation metrics to assess the performance of selected machine learning and statistical models. The next section will provide a brief review about the datasets and the selected models to be used in this project.

Chapter 3

Research Methodology

3.1 Introduction

The models and data that will be used in our study are described in this chapter along with the research design.

The main focus of survival analysis is the period of time before an event of interest happens. It takes into consideration censoring, which is a frequent occurrence in longitudinal studies and causes some subjects' observations of the vent of interest to be missing during the study period. Because of this, the approach used here is crucial for revealing trends, connections, and elements linked to the risk of controlling survival outcomes.

3.2 The Survival and Hazard Functions

Two important concepts in survival analysis are the hazard and survival functions. The likelihood of living past a particular time point is described by the survival function, while the rate of failure at any given instant is described by the hazard function at any given time point. These two functions are strongly associated and provide important insights into the survival patterns of individuals or groups.

3.2.1 The survival function

The survival function denoted as $S(t)$ can be estimated using the Kaplan-Meier estimator or other non-parametric survival models. It stands as a fundamental concept in survival and event history models, defined as:

$$S(t) = 1 - F(t) = \Pr(T \geq t) \tag{3.1}$$

This equation represents the likelihood that a survival time T equals or exceeds a particular time t. $S(t)$ reflects the fraction of individuals who survive beyond t and $F(t)$ presenting the cumulative distribution function. At the start of time $t = 0$, $S(0) = 1$, suggesting that every individual are surviving at $t = 0$. As subsequent in the next chapters, $S(t)$ is a function that steadily lowers as time passes as surviving individuals suffer failures (Mills, 2010).

3.2.2 The hazard function

The relationship between the occurrence of an event (such as failure) and survival is summarized by the hazard rate, also known as the instantaneous transition or hazard function:

$$h(t) = \frac{f(t)}{S(t)} \tag{3.2}$$

The hazard rate denotes the rate at which individuals fail by time t, given that they have survived until t. $f(t)$ presenting the probability density function. Hazard rate represents a conditional failure rate, as demonstrated by:

$$\lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = h(t) \tag{3.3}$$

Here, the transition rate $h(t)$ signifies the instantaneous risk of the event occurring in the time interval $[t, t+\Delta t]$, given survival up to time t. Thus, the hazard rate emphasizes the probability of experiencing the event, while the survivor function concentrate on the probability of not experiencing it (Mills, 2010).

3.3 Kaplan-Meier Estimator

3.3.1 Kaplan-Meier survival estimator

A non-parametric technique for estimating the likelihood of survival over time is called the Kaplan-Meier survival estimator. It is frequently used to examine time-to-event data, like the time to death, or time to failure which is widely used in medical research, survival analysis, engineering, and other fields. Kaplan and Meier (1958) introduced the estimator. The survival function, or the likelihood of surviving past a specific time point, is the foundation of the Kaplan-Meier estimator.

The number of people who have survived up to a given period is divided by the total number of people who were at risk at that point to get the Kaplan-Meier survival estimate. Those who have not yet experienced the event of interest or have not been censored are the ones who are considered to be at risk. When a subject disappears from follow-up or the study concludes before the event of interest takes place, this is known as censoring.

The Kaplan-Meier estimator is defined as follows:

$$S(t) = \prod_{i:t_i \leq t} [1 - (d_i/n_i)], \tag{3.4}$$

where $S(t)$ is the estimated probability of survival at time t, d_i is the number of deaths at time t, and n_i is the number of individuals at risk at time t.

The estimator is calculated recursively by multiplying the probabilities of surviving each prior time point. For example, the probability of surviving at time t_1 is calculated as:

$$S(t_1) = \left[\frac{(n_1 - d_1)}{n_1})\right], \tag{3.5}$$

where n_1 is the number of subjects at risk at time t_1, and d_1 is the number of deaths at time t_1. The probability of surviving at time t_2, given that the subject survived up

to time t_1, is calculated as:

$$S(t_2) = S(t_1)\left[\frac{(n_2 - d_2)}{n_2}\right], \tag{3.6}$$

where n_2 is the number of subjects at risk at time t_2, and d_2 is the number of deaths at time t_2.

The Kaplan-Meier estimator is commonly plotted as a survival curve, which shows the estimated probability of survival over time. The survival curve is a step function that decreases as the time increases, and it can be used to compare survival between different groups of individuals.

3.3.2 Greenwood estimator

The Greenwood formula, also known as the Greenwood estimator, is a widely used non-parametric method for estimating the variance of survival probabilities in survival analysis. It was first introduced by John Greenwood in 1926 (Greenwood, 1926).

The Greenwood formula is commonly used to construct confidence intervals for the survival function estimated using the Kaplan-Meier estimator. The Greenwood formula gives an estimate of the standard error of the Kaplan-Meier estimator, which can be used to construct confidence intervals.

The Greenwood formula is as follows:

$$\sqrt{\sum_{i=1}^{k}\left[\frac{d_i}{n_i(n_i - d_i)}\right]}, \tag{3.7}$$

where: k is the number of distinct event times in the data. n_i is the number of individuals at risk just prior to time i. d_i is the number of events at time i.

Once we have the standard error of the Kaplan-Meier estimator, we can calculate confidence intervals using the following formula:

Kaplan-Meier estimate $+/-\frac{z}{2}\times$ standard error

where $\frac{z}{2}$ is the appropriate quantile of the standard normal distribution based on the desired level of confidence.

3.3.3 Logrank Test p-value

The logrank test is a statistical test used to compare survival distributions between different groups, often used in survival analysis. Logrank test assesses whether the observed number of events in each group significantly deviates from the expected number of events under the null hypothesis of identical hazard functions. In each group, J denotes the distinct times of observed events; $N_{1,j}$ and $N_{2,j}$ indicate the number of subjects "at risk" in each group at the beginning of period j; and $O_{1,j}$ and $O_{2,j}$ indicate the observed number of events in each group at time j. The logrank test statistic Z_i for each group $(i = 1, 2)$ is calculated as follows under the null hypothesis:

$$Z_i = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}}, \tag{3.8}$$

where $E_{i,j}$ is the expected number of events in group i at time j under the null hypothesis, $V_{i,j}$ represents the variance of the expected number of events in group i at time j under the null hypothesis, and $O_{i,j}$ is the observed number of events in group i at time j. This variance is calculated as follows:

$$V_{i,j} = E_{i,j}\left(\frac{N_j - O_j}{N_j}\right)\left(\frac{N_j - N_{i,j}}{N_j - 1}\right) \tag{3.9}$$

This computation considers the variability in the expected number of events, accounting for differences in sample sizes and event counts between the groups at each time point. Under the null hypothesis, the logrank test statistic Z_i has a typical nor-

mal distribution ($\mathcal{N}(0,1)$), enabling statistical inference on the differences in hazard functions between the groups (Peto and Peto, 1972).

3.4 Cox PH Model

The Cox proportional hazard model, a semi-parametric model in which the outcome distribution is still unclear, is one of the frequently used models for expressing the hazard as a function of the covariates. These models are not fully parametric since the baseline hazard (or survival rate) is not specified, even though regression coefficients for variable impacts are defined. As a result, the parametric proportional hazards model is expanded upon by the Cox model. One benefit of the Cox model is that survival times are not based on distributional assumptions (Cox, 1975).

3.4.1 Model

The Cox proportional hazards model can be represented mathematically as:

$$h(t|\boldsymbol{X}) = h_0(t) \times \exp(\boldsymbol{X} \times \boldsymbol{\beta}), \tag{3.10}$$

where $h(t|\boldsymbol{X})$, a $n \times 1$ column vector, represents the hazard at time t for a person given covariates $\boldsymbol{X}$. The baseline hazard function, or $h_0(t)$, represents the risk for a person with $X = 0$ at time t. The hazard ratio (HR), which is a $n \times 1$ column vector, is defined as $\exp(\boldsymbol{X} \times \boldsymbol{\beta})$. It is the ratio of the hazard for an individual with covariates $\boldsymbol{X}$ to the hazard for an individual with $\boldsymbol{X} = 0$. $\boldsymbol{X}$ is a $n \times p$ matrix representing a vector of covariates for an individual. $\boldsymbol{\beta}$ is a $p \times 1$ column vector of regression coefficients, where n is the number of observations in the data and p is the number of covariates, that quantify the impact of each covariate on the log-hazard.

3.4.2 Estimation and interpretation of the Cox proportional hazards model

Partial likelihood technique was employed, which just utilizes data from those who actually experience the event of interest, to estimate the Cox proportional hazards model. In particular, the regression coefficients are calculated to maximize the likelihood, function which is defined as the product of the hazard functions for each person who experiences the event of interest. In mathematical terms, the partial likelihood is expressed as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})}{\sum_{j \in R(t_i)} exp(\beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_p x_{pj})} \tag{3.11}$$

The variables in this context are as follows: x_{1i}, x_{2i},..., x_{pi} are the values of the covariates for individual i; n is the total number of events; t_i is the time of failure for individual i; and $R(t_i)$ is the set of individuals who are still at risk at time t_i.

In the Cox proportional hazards model, the regression coefficients (β) are estimated by maximizing the partial likelihood function ($L(\boldsymbol{\beta})$). This is typically done using numerical optimization techniques such as the Newton-Raphson method. The Newton-Raphson method is an iterative optimization algorithm that aims to find the maximum (or minimum) of a function. In the context of estimating regression coefficients in the Cox proportional hazards model, the algorithm iteratively updates the parameter estimates until convergence to the maximum likelihood estimates.

The iteration process of the Newton-Raphson method starts with an initial guess for the regression coefficients, denoted as $\boldsymbol{\beta}^{(0)}$. At each iteration k, the algorithm updates the parameter estimates using the following formula:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left(\mathbf{I}(\boldsymbol{\beta}^{(k)})\right)^{-1} \nabla l(\boldsymbol{\beta}^{(k)}),$$

where $\nabla l(\boldsymbol{\beta}^{(k)})$ is the gradient vector of the log partial likelihood function evaluated at $\boldsymbol{\beta}^{(k)}$, and $\mathbf{I}(\boldsymbol{\beta}^{(k)})$ is the Hessian matrix (second derivative matrix) of the log partial

likelihood function evaluated at $\boldsymbol{\beta}^{(k)}$. The iteration continues until a convergence criterion is met, such as reaching a specified number of iterations or when the parameter estimates change by a small amount between iterations. The final estimates of the regression coefficients will be obtained when the algorithm converges. The Newton-Raphson method is efficient for maximizing the likelihood function and is commonly used in practice for estimating the parameters in Cox proportional hazards models.

Once we have estimated the regression coefficients, we can interpret them as follows: A positive coefficient for a covariate indicates that an increase in that covariate is associated with an increase in the hazard rate, i.e., a higher risk of experiencing the event of interest. A negative coefficient for a covariate indicates that an increase in that covariate is associated with a decrease in the hazard rate, i.e., a lower risk of experiencing the event of interest. The hazard ratio (HR) can be calculated as the exponential of the regression coefficient. For example, if $\boldsymbol{\beta_1} = 0.5$, then the HR is $\exp(0.5) = 1.65$. This means that for every one unit increase in x_1, the hazard rate increases by a factor of 1.65, all else being equal.

3.5 Variable Selection

Variable selection is the process of identifying the subset of input variables that are most relevant for predicting the output variable in a statistical model. The goal of variable selection is to improve the accuracy, interpretability, and generalizability of the model by reducing the number of input variables. Some variable selection methods are Lasso, Cox-Boost, Elastic Net, Information creteria (e.g., AIC, BIC), Principal Component Analysis, etc.

3.5.1 Lasso Method

The Lasso which is Least Absolute Shrinkage and Selection Operator is a regularization approach that picks a selected group of the most essential covariates by shrinking the coefficients of the less important ones towards zero. In Cox regression, Lasso is used to select important covariates and estimate their coefficients.

The objective function to be minimized is:

$$L(\beta) = -\text{log likelihood} + \boldsymbol{\lambda} \sum_{j=1}^{p} |\beta_j|, \tag{3.12}$$

where the tuning parameter is $\boldsymbol{\lambda}$, and the regression coefficient vector is $\boldsymbol{\beta}$. The tuning parameter $\boldsymbol{\lambda}$, controls the level of regularization applied to the regression model. It determines the trade-off between the model's difficulty and its ability to fit the data. The tuning parameter $\boldsymbol{\lambda}$ is chosen to achieve a balance between reducing the residual sum of squares (RSS) and shrinking the regression coefficients towards zero.

The selection of the tuning parameter $\boldsymbol{\lambda}$ is typically performed using strategies such as information criteria or cross-validation. The Bayesian information criteria (BIC) and the Akaike information criteria (AIC) are two examples of information criteria, provide a quantitative measure of the model's fit and complexity. These criteria penalize the model with a large number of parameters, encouraging the selection of simpler models. The tuning parameter is chosen by minimizing the information criterion, striking a balance between model fit and complexity.

On the other hand, cross-validation provides a more data-driven approach to model selection. It assesses the model's performance on independent validation sets and can offer a more precise assessment of the model's predictive ability. Cross-validation is particularly valuable when the goal is to optimize the predictive capacity of the model or when the assumptions of the information criteria may not hold. However, cross-

validation can be computationally intensive, especially for large datasets or complex models, and it may not capture the full range of variability in the data.

3.5.2 Inference from the Model

Inference in the Cox proportional hazards model involves hypothesis testing and the confidence interval construction for the regression coefficients (parameters) of the model. The regression coefficients represent the effects of the covariates on the hazard rate. The Hessian matrix is used to estimate the variances and covariances of the parameters in the model, such as hazard rates. It provides information on the log-likelihood function's curvature, which is used to estimate the parameters of the survival model. It helps to determine the direction and magnitude of parameter changes required to achieve a better model fit. Furthermore, it is used to calculate the standard errors of the estimated parameters, which are used to develop confidence intervals and perform hypothesis tests. By using the Hessian matrix in survival analysis, we can estimate the parameters of the model with greater accuracy and make more informed decisions about the time until an event occurs.

Hypothesis testing involves testing whether the regression coefficients are equal to a specific value (often zero) or not. A commonly used null hypothesis is that the coefficient is equal to zero, revealing that the corresponding covariate has no influence on the hazard rate. A p-value is computed for each coefficient based on the Wald statistic, which is the ratio of the coefficient estimate to its standard error. The p-value is compared to a chosen significance level, and if it is smaller than the significance threshold, the alternative hypothesis is accepted and the null hypothesis is rejected that the coefficient is different from zero.

Confidence intervals are a set of values that indicate the range of values of the regression coefficient is likely to reside at a specific level of confidence. The confidence

interval is constructed using the standard error of the coefficient estimate and the t-distribution. The confidence interval is computed as the coefficient estimate plus or minus the product of the t-quantile and the standard error. The standard error of the coefficient estimate measures the variability or uncertainty associated with the estimated regression coefficient. It quantifies how much the coefficient estimate is expected to vary across different samples. A smaller standard error indicates a more precise estimate.

3.5.3 Assumption of the model

The hazard ratio's constancy over time is one of the main presumptions of the Cox PH model. This suggests that over time, the hazard rate ratio between two groups stays constant. The proportionate hazards assumption is another name for this presumption. Graphical techniques and statistical testing can be used to verify this premise. A method for verifying the proportional risks assumption involves graphing the survival function's log against time for varying covariate levels. The assumption of proportionate hazards is satisfied if the curves are parallel.

3.6 Machine Learning Algorithms

Machine learning is a sub-field of artificial intelligence that allows computers to gain information patterns from data and make predictions in the absence of detailed programming. In survival analysis, machine learning models are fitted to predict time-to-event outcomes, accommodating censored data where events may not have occurred for all individuals. Machine learning Models to be included to analyse the data in this study are Survival trees, random survival forests and Neural network.

3.6.1 Survival trees

Survival trees are a non-parametric methods for modeling the relationship between time-to-event data and covariates. They are a variation of decision trees and are a dominant machine learning technique for addressing problems in regression and classification. Survival trees are constructed by recursively dividing tree nodes, much like decision trees. If all the patients in a node of a survival tree survive for the same amount of time, the node is said to be "pure" (LeBlanc and Crowley, 1993). The most used dissimilarity metric for estimating the distinction in survival rates between two groups is the logrank test. Examine every potential split for each feature for each node, and then choose the split that increases the survival gap between two offspring nodes.

The resulting survival tree can be utilized to predict survival outcomes for new individuals depending upon their covariates values. To make a prediction, the algorithm follows the tree from the start node to the leaf node that corresponds to a subgroup of individuals with similar covariate values, and then estimates the survival probability for that subgroup using the Kaplan-Meier estimator or another survival estimator.

Survival trees can be represented mathematically as: Assume that C is the censoring variable and T is the time to event variable. With dimension $(1 \times p)$ and p representing the number of covariates, let $\boldsymbol{X}$ be a vector of covariates. The set of individuals who are at risk of experiencing the event at time t is denoted by R, which is the risk set at that point in time. The number of subjects in the risk set at time t is denoted by $N(t)$. The hazard function for an individual with covariates $\boldsymbol{X}$ at time t is denoted by $H(t|\boldsymbol{X})$. The survival function for an individual with covariates $\boldsymbol{X}$ at time t is denoted by $S(t|\boldsymbol{X})$. Let $G(\boldsymbol{X})$ represent the binary survival tree with internal nodes representing covariates and split points, and leaf nodes representing subgroups of individuals with similar covariate. Let $G(\boldsymbol{X}|\boldsymbol{x})$ be the subtree of $G(\boldsymbol{X})$ rooted at the node that corresponds to the covariate values $\boldsymbol{x}$.

Then, the survival probability for the subject with covariates $\boldsymbol{X}$ at time t is:

$$S(t|\boldsymbol{X}) = S(t|G(\boldsymbol{X}|\boldsymbol{X})), \tag{3.13}$$

where $G(\boldsymbol{X}|\boldsymbol{X})$ is the path through the tree that corresponds to the covariate values $\boldsymbol{X}$.

3.6.2 Random Survival Forest

Random survival forest is a machine learning algorithm used for survival analysis, it is based on the idea of a traditional random forest algorithm, which constructs many decision trees by randomly selecting a subset of features and samples from the dataset. However, in a random survival forest, the splitting of each tree is based on a survival criterion. It is a type of ensemble learning algorithm that integrates many decision trees to provide a more precise and reliable model. The results of a random survival forest algorithm is a survival function, which describes the probability of survival over time for each individual in the dataset. This can be utilized to estimate the likelihood of an event occurring for new individuals based on their characteristics.

Random survival forests, being an ensemble of survival trees, inherently incorporate variable selection (Ishwaran et al., 2008). Within each tree in the forest, a random subset of variables is considered for each split, reducing the chance of any single variable dominating the modeling process. The forest aggregates the predictions from multiple trees, and variable importance measures can be derived from this ensemble. These importance measures provide insights into the relative contributions of each variable and can be used for variable selection.

Random survival forests have several advantages over traditional survival analysis methods, including the ability to handle high-dimensional data, non-linear and non-monotonic relationships between covariates and survival, and interactions between covariates.

The formula for random survival forests is the same as the formula for the Cox proportional hazards model, but with an additional term that accounts for the tree-based structure of the model. The formula can be written as:

$$h(t|X) = h_0(t) \times \exp\left(\sum(w_j \times I(T_j \leq t) \times I(X_j))\right), \tag{3.14}$$

where: $\boldsymbol{X}$ is a vector with dimension $(1 \times p)$, p is the number of covariates, $h(t|\boldsymbol{X})$ is the hazard at time t for an individual with covariates $\boldsymbol{X}$, $h_0(t)$ is the baseline hazard function, which is the hazard for an individual with $\boldsymbol{X} = 0$ at time t, $\exp(\sum(w_j \times I(T_j \leq t) \times I(X_j)))$ is the hazard ratio, which present the ratio of the hazard for an subject with covariates $\boldsymbol{X}$ to the hazard for subject with $\boldsymbol{X} = 0$, where the sum is taken over all trees in the forest, w_j is the weight assigned to tree j, Tj is the survival time predicted by tree j, $I(T_j \leq t)$ is an indicator function that is 1 if T_j is less than or equal to t, and 0 otherwise, and $I(X_j)$ is an indicator function that is 1 if the covariates of the individual satisfy the conditions of the split node in tree j, and 0 otherwise.

The weights w_j are determined by the out-of-bag (OOB) samples, which are the samples that are not included in the subsample used to construct tree j. The weight wj for tree j is calculated as the OOB prediction error for the samples in the subsample used to construct tree j, divided by the sum of the OOB prediction errors for all trees in the forest.

3.6.3 Artificial Neural Network

Neural networks are a kind of machine learning algorithm that can be used for modeling time-to-event data in survival analysis. ANNs are composed of multiple layers of interconnected nodes or neurons, which begin with an input layer that represents the data features, which is then fed into one or more hidden layers before ending with an output layer that shows the outcome. More people are beginning to see neural networks as a helpful complement to conventional and cutting-edge statistical techniques

in the arsenal of statistics (Cheng and Titterington, 1994). In survival analysis neural networks falls under non-parametric machine learning approach.

Variable selection in neural networks for survival analysis is typically accomplished through a combination of network architecture design and the training process. By adjusting the number of layers, nodes, and connections in the neural network, this can control the complexity and capacity of the model, indirectly influencing the variables that contribute to the prediction. During training, the network assigns weights to each input variable, emphasizing the more influential ones. This weighting process effectively performs variable selection by assigning higher importance to relevant variables.

The formula for an ANN in survival analysis depends on the specific architecture and activation functions used, but a common approach is to use a Cox proportional hazards model as the output layer of the network. The Cox proportional hazards model assumes that the hazard function is proportional to a baseline hazard function $h_0(t)$ and a set of covariates X, with a constant coefficient β.

To incorporate an ANN into the Cox PH model, the output of the last hidden layer of the network is used as the input to the Cox model. The formula for the Cox model with an ANN output is:

$$h(t|\boldsymbol{X}) = h_0(t) \times \exp(f(\boldsymbol{X}) \times \boldsymbol{\beta}), \tag{3.15}$$

where: $f(\boldsymbol{X})$ is the output of the last hidden layer of the ANN, which maps the input covariates to a set of features that are used as input to the Cox model. The parameters of the ANN, including the weights and biases, can be learned from the data using maximum likelihood estimation or gradient descent methods.

ANNs have several advantages over traditional survival analysis methods, including the capability of capturing complicated not linear interactions between variables and

survival, and the ability to handle high-dimensional data. However, ANNs can also be prone to overfitting and can be difficult to interpret.

3.7 Performance Measures

This section focuses on evaluation of the effectiveness of machine learning techniques and statistical models on survival analysis with regards to their ability to predict the outcomes. The effectiveness of models in terms of discrimination and calibration is evaluated.

3.7.1 Discrimination measures

To compare different models, it is important to consider evaluation metrics appropriate for survival analysis, such as time-dependent area under the curve (AUC), concordance index (C-index), or logrank test p-value. In addition to evaluation metrics, other factors should also be considered when comparing models like: model interpretability, computational efficiency, and scalability should also be taken into account

Concordance Index

The Concordance Index is a suitable measure for comparing the predictive accuracy of both machine learning and statistical models. It evaluates the model's ability to accurately rank the predicted survival times. By comparing the C-index values between different models, you can find out which model performs well in regard to predictive accuracy.

$$C\text{-index} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij} \cdot I(T_i < T_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}} \tag{3.16}$$

In this context, n represents the total number of individuals or subjects, C_{ij} denotes the pairwise comparison of the predicted continuous outcomes or risk scores for individuals i and j. If $C_{ij} = 1$, it means the predicted outcome for individual i is concordant with the actual outcome for individual j. If $C_{ij} = 0$, it means they are discordant or tied.

T_i represents the survival time or time-to-event for individual i. Higher C-index values indicate better discriminatory ability, suggesting that the model has greater precision in anticipating the relative survival times of individuals.

Time-dependent Area Under the Curve (AUC)

The time-dependent area under the curve is a metric that assesses the discriminatory power of survival models over various points in time. AUC can also be used to compare the discriminatory power of ML and statistical models. It provides a measure of how well the model distinguishes between different time points in survival analysis. The AUC varies from 0.5 lacking in discrimination to 1 complete discrimination and can be calculated at specific time intervals or averaged over all time points.

$$\text{AUC}(t) = \frac{\sum[C_i(t) - D_i(t)]}{C_i(t) \cdot D_i(t) + 0.5 \cdot C_i(t) \cdot (C_i(t) - 1)} \tag{3.17}$$

$C_i(t)$ represents the number of correctly predicted pairs at time t. $D_i(t)$ represents the number of discordant pairs at time t. The numerator of the formula calculates the number of concordant pairs minus discordant pairs at the given time point, while the denominator represents the product of the number of concordant and discordant pairs. A greater AUC value implies an improved ability of the model to distinguish different time points in the event of interest at that specific time point.

3.7.2 Calibration measures

When comparing statistical models with machine learning algorithms, one approach is to use calibration models. Calibration models help assess the calibration performance of a model, which refers to how well the predicted probabilities align with the observed outcomes. Commonly used calibration model are Reality Plot, Calibration Curve, Brier Score, and Hosmer-lemeshow test.

Calibration Curve

Similar to a reliability plot, a calibration curve plots the predicted probabilities against the observed outcomes. However, instead of using intervals, it calculates the average predicted probability for different prediction quantiles and plots them. A well-calibrated model would show a calibration curve that is close to the ideal diagonal line.

Brier Score

The Brier Score is the mostly used metric for assessing the accuracy of probabilistic predictions in various fields, including forecasting, statistical modeling, and survival analysis. It measures the discrepancy between predicted probabilities and observed outcomes, providing a measure of predictive performance. The Brier Score $BS(t^*)$ at a specific time t^* is defined as follows:

$$BS(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left(I(T_i > t^*) - \hat{\pi}(t^*|\hat{X}_i)\right)^2, \tag{3.18}$$

where $BS(t^*)$ represents the BS at time t^*, which quantifies the squared variance among anticipated probability and observed results, where n is the overall amount of subjects or observations, $\sum_{i=1}^{n}$ represents the sum of all observation, $I(T_i > t^*)$ is an indicator function that returns 1 if the event for the i-th observation occurs after time t^*, and 0 otherwise, and $\hat{\pi}(t^*|\hat{X}_i)$ is the predicted probability of the event occurring at time t^* for the i-th observation, given the estimated predictors $\hat{X}_i$. The Brier Score ranges from 0 to 1, where lower scores indicate better predictive accuracy.

An overall evaluation of the model's performance over all time points, $t_1 \leq t \leq t_{max}$, is provided by the Integrated Brier Score.

$$\text{IBS} = \int_{t_1}^{t_{\max}} \text{BS}^c(t)\, dw(t) \tag{3.19}$$

is the definition of the integrated time-dependent Brier score across the interval $[t_1; t_{max}]$, where $w(t) = t/t_{\max}$ are the weighting functions.

3.7.3 Other model evaluation criteria

Model Interpretability

Statistical models like the Cox proportional hazards model are often preferred when interpretability and understanding the relationship between covariates and survival are important. Machine learning models, may provide excellent predictive performance but can be more challenging to interpret.

Computational Efficiency

Depending on the scale of the dataset and the complexity of the models, computational efficiency may be a consideration. Statistical models like the Cox model are often computationally efficient, whereas certain machine learning models may require more computational resources and time.

Scalability

If you have a large dataset with a high number of predictors, you may need to consider the scalability of the models. Some ML models, such as random survival forests or boosting models, have algorithms and implementations that can handle large-scale data efficiently.

Analysis of Variance

ANOVA is a statistical method used to contrast means between two or more groups. ANOVA allows researchers to determine whether there are differences in the means of the groups beyond what would be expected due to random sampling variation. ANOVA relies on several assumptions for valid interpretation such as independence of observations which state that each observation should be independent of others. The residuals should follow a normal distribution. When there is homogeneity of variances, the dependent variable's variance should be roughly the same for every group. The general

model for ANOVA can be represented as follows:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \tag{3.20}$$

where μ is the overall mean, τ_i is the effect of the i^{th} group, ϵ_{ij} is the error term, and Y_{ij} is the observation in the i^{th} group and j^{th} observation.

The null hypothesis (H_0) indicates no significant difference between the means of the groups and alternative hypothesis (H_1) indicate that at least one group mean is significantly different from the others. The ANOVA table presents key statistics including associated p-values, sums of squares, degrees of freedom, sources of variation, the F-ratio, and mean squares. The interpretation of results involves examining the p-value associated with the F-ratio. If the probability value is less than a predetermined significance level (e.g, 0.05), it indicates that there is sufficient evidence against the null hypothesis, suggesting that at least one group mean is significantly different from the others.

Violin Plot

Violin plots offer a visual comparison of probability distributions across different groups or categories, combining elements of box plots and kernel density plots. The box plot component provides information about the quartiles, median, and outliers, while the kernel density plot illustrates the distribution shape and density of the data. Below Figure 3.1 illustrates the common components of box plot and violin plot (WindEsco, 2024).

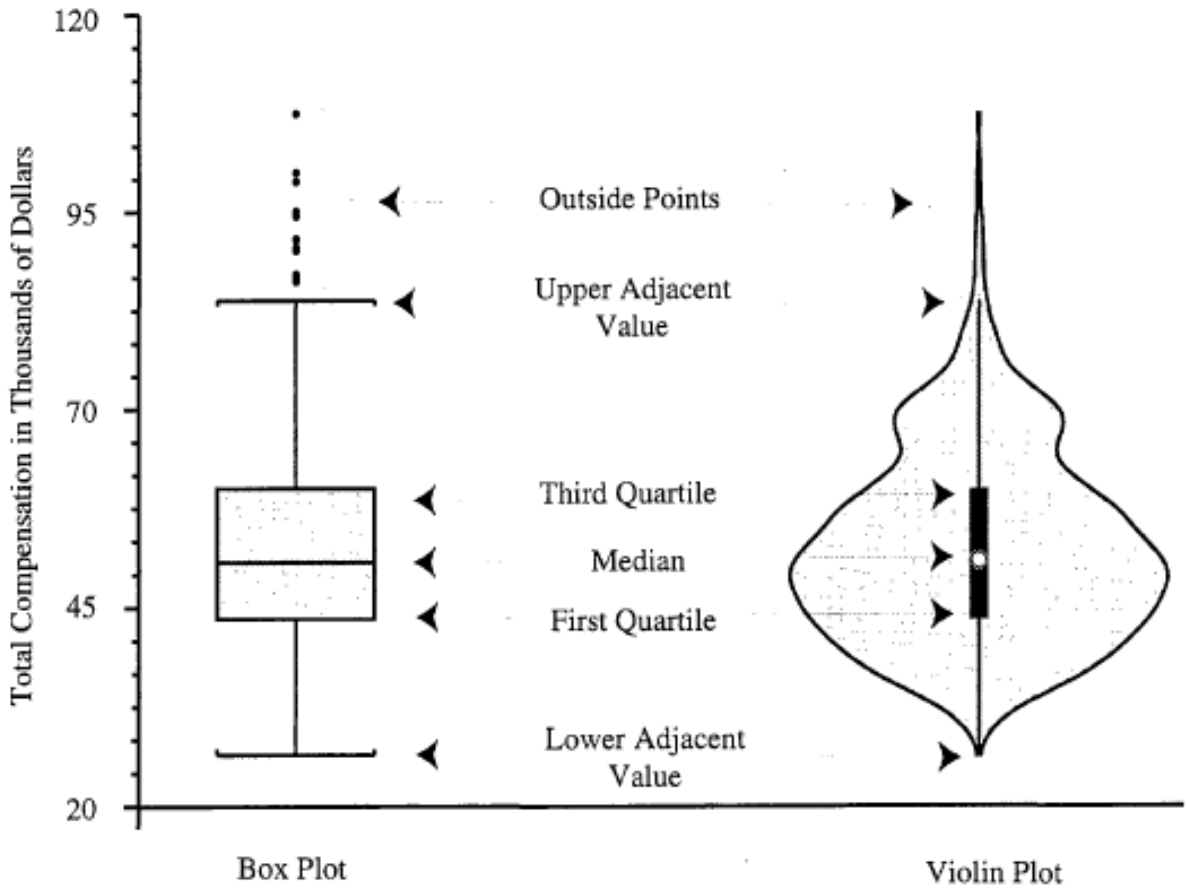


Figure 3.1: Illustration of violin plot

The violin plot has an advantage over the box plot in that it displays the whole distribution of the data as well as the above-mentioned statistics.

3.8 Data

In this section, we introduce two datasets which will be utilized in our study, which forms the keystone of our investigation into the factors influencing the survival analysis outcomes.

3.8.1 North Carolina Recidivism dataset

The study will use secondary dataset, North Carolina Recidivism dataset which is openly accessible via the Inter-University Consortium for Political and Social Research (ICPSR). For 9549 prisoners released from the North Carolina prison between 1978 and 1980 the data gives the time from release until re-offending. Prisoners who had not re-offended are censored and those who offended were observed. Table 3.1 below show data description.

3.8.2 Time to first alcohol intake dataset.

The second secondary dataset was kindly provided by Sithuba (2018) who collected the data as part of his honors project.The data provides the age at which students from both institutions, University of Venda and Vhembe TVET College, were interviewed regarding the onset of alcohol consumption. The time to first alcohol intake is the response variable. Table 3.2 shows an overview of the dataset.

Table 3.1: Recidivism data description of Variables

Variable Name	**Description**
WHITE	Ethnicity classified as Whites (1) or Blacks (0)
ALCHY	Serious alcohol problem (1) or Otherwise (0)
JUNKY	Use of hard drugs (1) or Otherwise (0)
SUPER	Supervised (e.g., parole) (1) or Otherwise (0)
MARRIED	Married (1) or Otherwise (0)
FELON	Conviction for felony (1) or Misdemeanor (0)
WORKREL	Participated in work release program (1) or Otherwise (0)
PROPTY	Conviction for crime against property (1) or Otherwise (0)
PERSON	Conviction for crime against a person (1) or Otherwise (0)
MALE	Gender: Male (1) or Female (0)
PRIORS	Number of previous incarcerations (9 if missing)
SCHOOL	Number of years of formal schooling completed
RULE	Number of violations reported during sentence
AGE	Age at release (in months)
TSERVD	Time served (in months)
FOLLOW	Follow-up time (in months)
RECID	Returned to prison (1) or Otherwise (0)
TIME	Length from release to return in months
FILE	Analysis sample (1) or Missing data sample

Table 3.2: Alcohol intake data description of Variables

Variable Name	**Description**
Gender	Gender of the individual, classified as Male or Female.
Age	Age of the individual in years.
Ethnicity	Ethnicity of the individual, classified as Black or Other race.
Monthly Income	Individual's monthly income, classified as <1000, 1000-4000, or >4000.
Family Monthly Income	Monthly family income, classified as <1000, 1000-4000, or >4000.
Siblings	Number of siblings, classified as 0-3, 4-6, or >=7.
Age at first alcohol intake	Age at first alcohol intake in years.
Parent's Qualification	Highest qualification of the parent, classified as None, Matric, or Above Matric.
Childhood Residence	Type of childhood residence, classified as Rural/farmland or Urban.
Other Drugs	Use of drugs other than alcohol, classified as Yes or No.
Drinking Peers	Peer involvement in drinking alcohol, classified as Yes or No.
Physical Abuse	Experience of physical abuse, classified as Yes or No.
Sexual Abuse	Experience of sexual abuse, classified as Yes or No.
Negative Life Events	Experience of negative life events, as Yes or No.
Stress	Experience of stress, classified as Yes or No.
Parental Drug/Alcohol Abuse	Parental history of drug or alcohol abuse, classified as Yes or No.
Subjected to Parental Discipline?	Experience of parental discipline, classified as Yes or No.
Family Dependent on Social Welfare?	Dependence on social welfare, classified as Yes or No.
Dysfunctional Family	Membership in a dysfunctional family, classified as Yes or No.
Relation with Adult	Positive relation with an adult, classified as Yes or No.

3.8.3 Flow chart of the study

Figure 3.2 shows the structural approach on how survival analysis was conducted, we will utilize two datasets: one for recidivism data and another for alcohol intake data. Both datasets was splitted into testing set of 20 percent and training set of 80 percent. Pandas library in Python was then used for data manipulation and data cleaning by managing empty values, normalizing features, and encrypting variables with categories.

Exploratory analysis and Kaplan-Meier estimation was employed to understand survival patterns in both datasets. LassoCV was the utilised for variable selection to identify important features while shrinking coefficients of less important ones. For model training, Cox Proportional Hazards (CoxPH), CoxLasso (Coxnet), Survival Tree, Random Survival Forest, and a Neural Network was employed. Then, evaluate models using Area Under the Curve, Concordance Index, and Integrated Brier Score. Finally, model performance was assessed on testing data to ensure robustness.

3.9 Ethical Consideration

Ethical approval was obtained to utilize Sithuba's recidivism dataset, which involved adhering to ethical protocols such as securing an ethical clearance certificate. Additionally, another dataset on recidivism was sourced from the Inter-University Consortium for Political and Social Research (ICPSR), where it was publicly available. The results of this project expect no harm to human health.

Figure 3.2: Structured approach

Chapter 4

Results and Discussion

4.1 Introduction

This chapter examine the application of the models and methodologies outlined in the preceding section to conduct thorough data analysis.

4.2 Experimental Framework

This study uses the Overleaf Latex and Python programming language along with libraries to explore and implement various machine learning and statistical techniques. Central to our experimentation is the Scikit-learn library (sklearn), a Python-based open-source tool. It integrates with other libraries such as Pandas for efficient data handling and manipulation, Matplotlib and Seaborn for insightful data visualization, and NumPy for numerical computations. Additionally, this study incorporates Lifelines and Scikit-Survival libraries for survival analysis, offering tools like Kaplan-Meier estimators, Cox Proportional Hazards models, Logrank-test and Random Survival Forests. Statistical analysis and visualization are facilitated through Statsmodels and Seaborn, while deep learning capabilities are explored with TensorFlow's Keras API. Techniques like Bagging Regressor and Multi-layer Perceptron Regressor are employed using Scikit-learn, along side with data splitting approach (Pedregosa et al., 2011; Pölsterl, 2020).

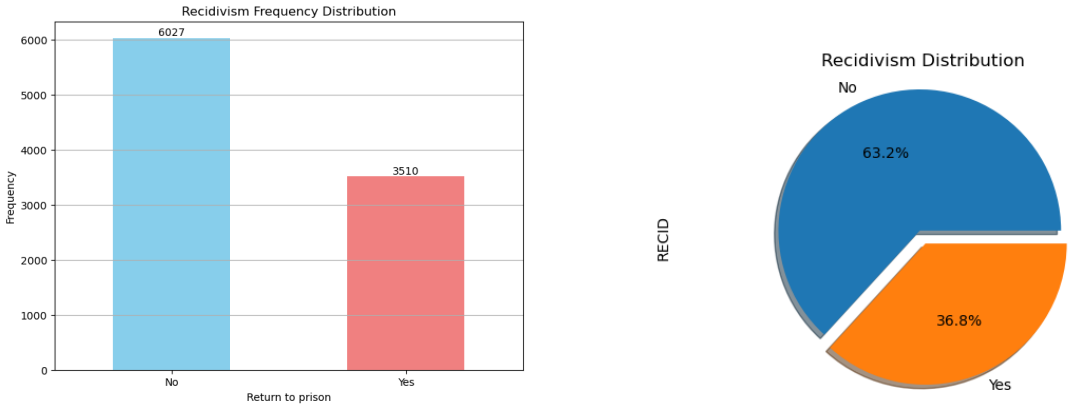
LassoCV was employed for variable selection on both the alcohol intake and recidivism datasets. After fitting the LassoCV, it extracts variables and their corresponding coefficients. We then fit our models using the selected variables from LassoCV. For each model, the code iterates over 50 iterations, separating the data into test set and training set for each iteration. It fits the models on the training data and evaluates its performance on the test data. The evaluation measures include the IBS, C-index, and AUC score. We then visualize the AUC and C-Index values across iterations for each model, allowing for a comparative assessment of their predictive capabilities. Subsequently, the plot depicting IBS across iterations provides insights into the models' calibration. Additionally, statistical analyses including analysis of variance are carried out to evaluate significant variations among the models. The violin plot further visualizes these differences, with annotations indicating significant contrasts between models.

4.3 Exploratory Data Analysis

In this section, we embark on exploration of our survival datasets, unraveling the intricate relationships between target variable and survival times. Through the examination of our data's distribution and visualization techniques for survival data.

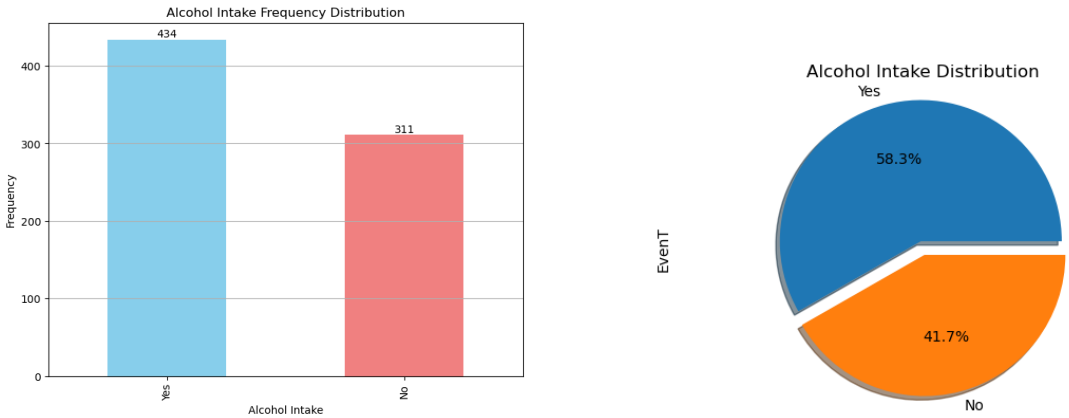
Figure 4.1 below shows the recidivism distribution where the target variable is "RECID". Among 9537 participants of which 6027 (63.2%) did not return to prison while 3510 (36.8%) returned to prison.

Figure 4.2 below shows the distribution of alcohol intake data where the target variable is "EvenT". Among 745 participants of which 434 (58.3%) drink alcohol and 311 (41.7%) don't drink alcohol.



(a) Bar chart (b) Pie chart

Figure 4.1: Recidivism Distributions



(a) Bar chart (b) Pie chart

Figure 4.2: Alcohol Intake Distributions

4.4 Kaplan-Meier

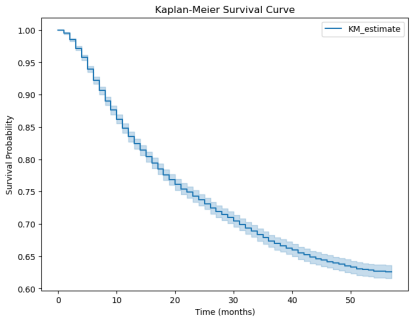
The Kaplan-Meier (K-M) survival analysis is a statistical technique employed for estimating a dataset's survival function. It is especially helpful in situations where we want to know how long it will take for an important event to happen. In our study, the event is recidivism for the first data and age at first alcohol intake for the second data.

4.4.1 Recidivism dataset

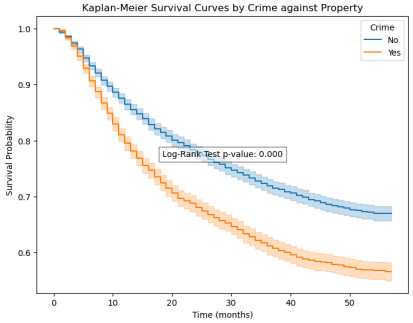
Figure 4.3(a) depicts the survival curve of the prisoners, illustrating that by month 50, approximately 65% of the prisoners have not re-offended. The subsequent plots in the figure further break down the survival probabilities based on categorical explanatory variables. Figure 4.3(b) indicates that individuals with no history of property crimes consistently exhibit higher survival probabilities compared to those who committed property crimes. The logrank test for comparing survival probabilities between these groups yields a significant p-value of of approximately 0.

Moving on to Figure 4.3(c), it becomes evident that female prisoners consistently have higher survival probabilities than their male counterparts. The logrank test confirms this observation with a p-value of 0.000, indicating a significant difference in survival probabilities between genders. Similarly, Figure 4.3 (d), 4.3(e), 4.3(f), and 4.3(g) reveal significant differences in survival probabilities based on the presence of marital status, serious alcohol problem, race, and supervision with p-values = 0.00 for all.

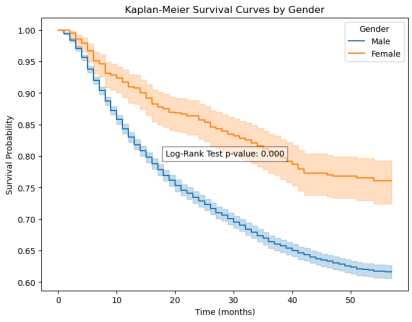
In contrast, Figure 4.3(h) shows that while initially, misdemeanors had lower survival probabilities compared to felons, the difference diminishes over time, with a p value of 0.545. Importantly, The proportionate hazards assumption is not violated.



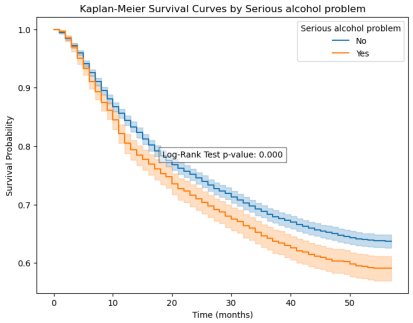
(a) Kaplan Survival Curve



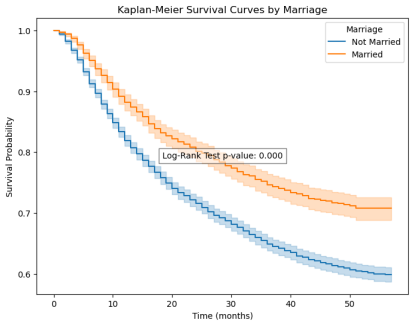
(b) K-M by crime against property



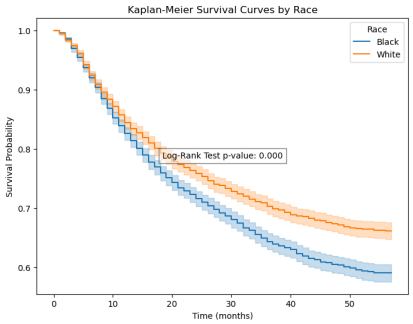
(c) K-M by Gender



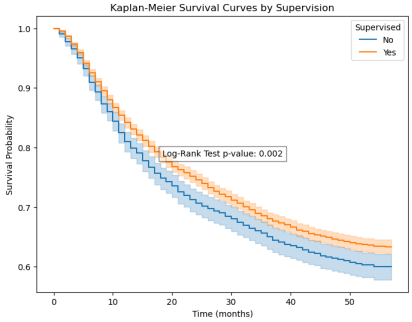
(d) K-M by serious alcohol problem



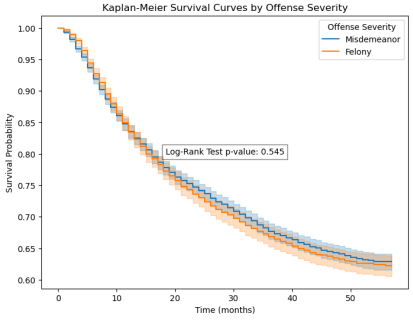
(e) K-M by Marriage



(f) K-M by Race



(g) K-M by Supervision



(h) K-M by Felony conviction

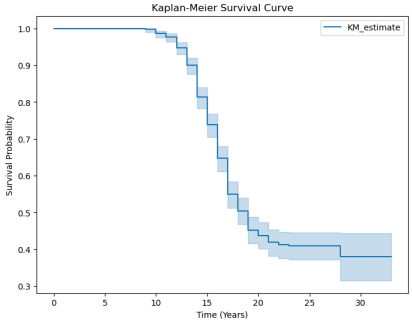
Figure 4.3: Kaplan-Meier Survival Curves

4.4.2 Alcohol intake dataset

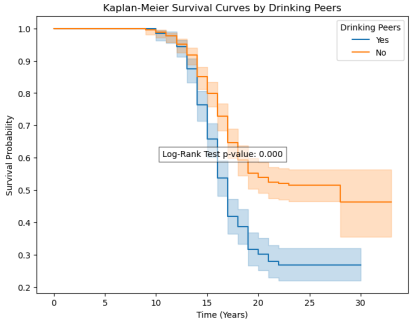
The Kaplan-Meier survival curves presented in Figure 4.4 provide insights into the age at first alcohol intake and its association with various factors. Beginning with Figure 4.4(a), we observe a consistent survival probability of 1 from ages 0 to 9 years, indicating a lack of events during this period. However, a notable decline in survival probability commences at age 9, reaching its lowest point at age 23, with a concluding survival probability of 35%, indicating a heightened risk of alcohol initiation during early age.

Figure 4.4(b) and 4.4(d), which respectively analyze age at first alcohol intake in relation to drinking peers and gender, we see distinct survival trajectories. Both panels reveal significant differences in survival probabilities. Those with drinking peers have consistently lower survival probabilities throughout, culminating with a survival probability of 25% compared to 45% for those without drinking peers at age 30. Similarly, females consistently exhibit higher survival probabilities compared to males. The associated logrank tests confirm the significance of these differences (p-value=0.000). Figure 4.4(e) Individuals with no history of other drug use have a significantly higher survival probability of 40% compared to those with such a history, who have a 15% chance by the age of 30. The logrank test validates the significance of this difference (p-value=0.000). Similarly Figure 4.4(f), 4.4(g), and 4.4(h) reveals the significance different in survival probabilities based on categorical explanatory variables (p-values<0.000) same to childhood residence ($p = 0.014$). Similarly to survival curves of negative life event and sibling with p values < 0.000 are shown in appendix A.

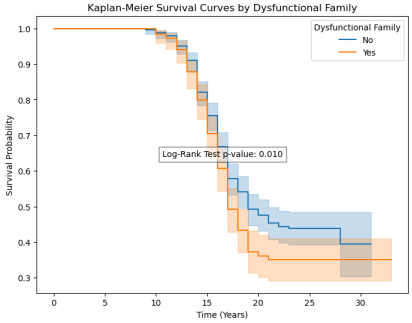
In contrast, Survival curves shown in appendix A of Parent qualification ($p = 0.07$) and individual income ($p = 0.919$) reveal that there's no significant difference. Based on the provided survival curves and logrank test results, there's no indication of significant changes in hazard ratios over time for the variables examined. Similarly to recidivism, there is no violation of the proportional hazard assumption.



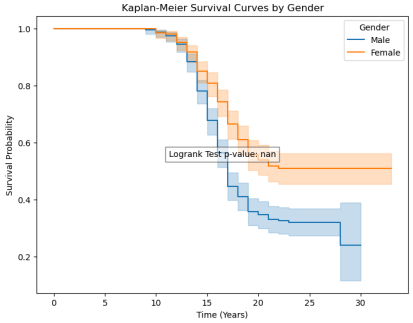
(a) K-M survival curve



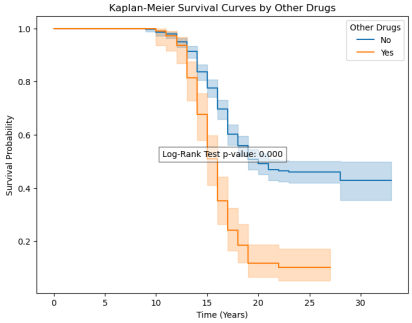
(b) K-M by drinking peers



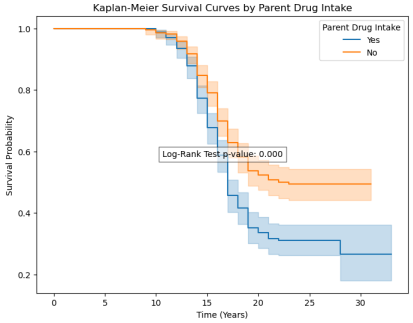
(c) K-M by dysfunctional family



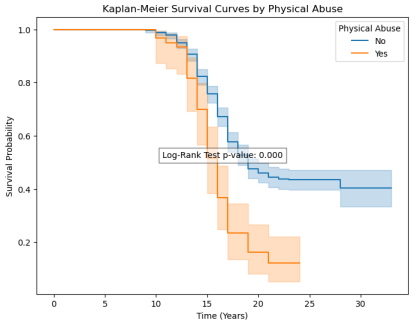
(d) K-M by gender



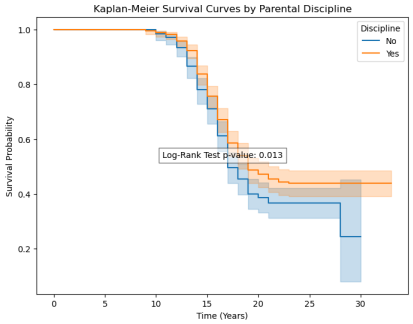
(e) K-M by other drugs



(f) K-M by parental drug intake



(g) K-M by physical abuse



(h) K-M by parental discipline

Figure 4.4: Kaplan-Meier Survival Curves

4.4.3 Conclusion

The K-M plots in Figure 4.3 and logrank tests of recidivism highlight significant associations between various factors such as gender, history of property crimes, serious alcohol problems, marital, race, and supervision all appear to influence the risk of re-offending. Understanding these associations can help policies aimed at reducing recidivism rates among prisoners.

The K-M plots in Figure 4.4 and logrank tests for age at first alcohol intake showcase a mix of factors influencing the age at first alcohol intake. Factors such as other drugs use, physical abuse, disciplinary actions, dysfunctional family environment, and drug-using peers impact survival probabilities. Knowing these associations can help interventions aimed at reducing the age of alcohol initiation.

4.5 Models Comparison

In this section, we examine the overall comparison of various machine learning and statistical models employed throughout our study. As we navigate through our models and evaluation, our aim is to find out the most effective model/s in survival analysis.

4.5.1 Variable selection

Table 4.1 shows the LassoCV selected variables with their coefficients.

Table 4.1: Selected variables

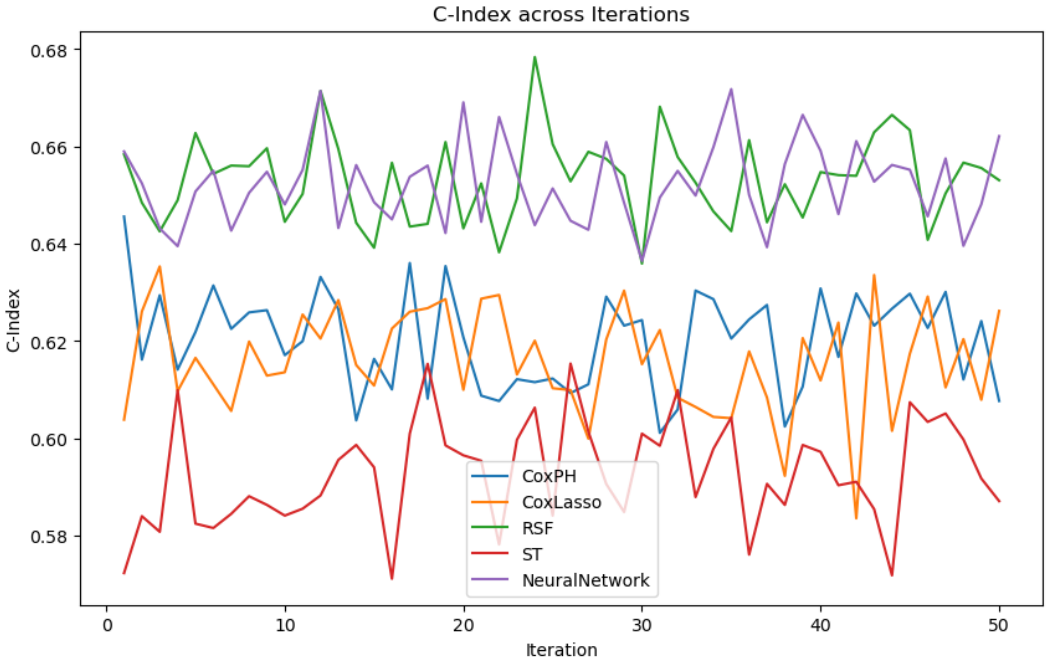
Recidivism		**Alcohol Intake**	
Variable	**Coefficient**	**Variable**	**Coefficient**
WHITE	-0.0398	OtherDr	0.1597
ALCHY	0.0294	Discipn	-0.0309
SUPER	-0.0065	DrPeers	0.0979
MARRIED	-0.0488	NegetiveLE	0.0581
FELON	-0.0221	PhysicAb	0.1105
PROPTY	0.0528	DisfunctF	0.0127
MALE	0.0392	PDrIntake	0.0982
PRIORS	0.0094	Sex	0.1028
SCHOOL	-0.0079	ChildPR	0.0254
RULE	0.0054	Siblings	0.0280
AGE	-0.0005	Income	-0.0267
TSERVD	0.0011	Pqual2	0.0335

4.6 Concordance Index

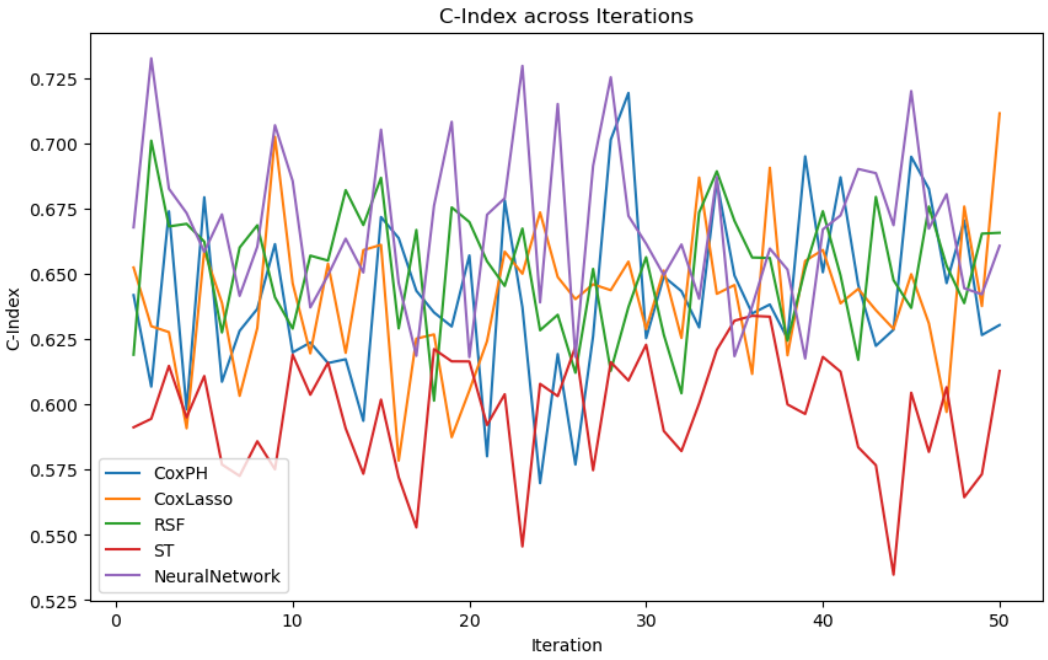
Figure 4.5(a) below portraying the C-index over 50 iterations reveals the performance of different models. The Neural Network and Random Survival Forest (RSF) shows a slight fluctuating trends, with their C-index values oscillating around 65%. In contrast, the CoxPH and Lasso-regularized Cox (CoxLasso) models demonstrate marginal fluctuation around intermediate C-index values of 62%. Notably, the Survival Tree (ST) model lags behind the others, consistently showing lower C-index values around 58% throughout the iterations.

Figure 4.5(b) below reveals a mixed notable fluctuating performance among the models over 50 iterations. However, it is evident that CoxPH, CoxLasso, NN, and RSF exhibit comparable performance, consistently around 65%, while ST consistently appear at the bottom with c-index around 58%.

Table 4.2 below compares various models based on the average concordance index (C-Index) and associated average standard errors across recidivism and alcohol intake datasets. For recidivism, the NN and RSF models demonstrate the highest average C-Index values, with the NN model at 0.6522 and the RSF model at 0.6533. Both models exhibit relatively low average standard errors (0.0017 for NN and 0.0017 for RSF), indicating their consistent and robust performance. The CoxPH model follows closely behind with a C-Index of 0.6203 and a slightly smaller standard error of 0.0016. In alcohol intake dataset, the NN model again leads with the top average C-Index of 0.6678, accompanied by a low average standard error of 0.0012, emphasizing its strong predictive performance. The RSF model maintains competitive performance with a C-Index of 0.6520 and a slightly smaller standard error of 0.0013.



(a) Recidivism plot



(b) Alcohol Intake plot

Figure 4.5: Serial plots of concodance index values obtained when the CoxPH, CoxLasso, random survival forest, survival tree and neural network are applied to the recidivism and alcohol intake data sets.

The CoxPH and CoxLasso models demonstrate comparable average C-Index values for both recidivism and alcohol intake, with slightly smaller standard errors for the CoxPH model, while ST model consistently exhibits lower average C-Index values and higher standard errors compared to other models, suggesting comparatively weaker performance.

Table 4.2: Average values and standard errors of the C-index obtained when the Cox PH, Cox-lasso, Random Survival Forest (RSF), Survival Tree (ST), and Neural Network (NN) are applied to 50 subsamples of the alcohol intake and recidivism datasets.

Recidivism			Alcohol Intake		
Models	Avg C-Index	Std Err	Models	Avg C-Index	Std Err
CoxPH	0.6203	0.0016	CoxPH	0.6416	0.0015
CoxLasso	0.6159	0.0019	CoxLasso	0.6405	0.0014
RSF	0.6533	0.0017	RSF	0.6520	0.0013
ST	0.5927	0.0018	ST	0.5972	0.0016
NN	0.6522	0.0017	NN	0.6678	0.0012

The ANOVA results for the recidivism dataset in Table 4.3(a) demonstrates significant differences among the models as evidenced by a large F-statistic of 344.2 and p-value which is < 0.05. Similarly, for alcohol intake dataset in Table 4.3(b), the ANOVA results shows a significant overall difference among the survival models with a p-value below 0.05 and F-statistic of 45.83. This underscores that at least one model significantly outperforms the others.

To further illustrate these significant differences among the models, violin plots are presented on the next page.

Table 4.3: ANOVA results obtained when the Cox PH, Cox-lasso, Random Survival Forest (RSF), Survival Tree (ST), and Neural Network are applied to 50 subsamples of the alcohol intake and recidivism datasets

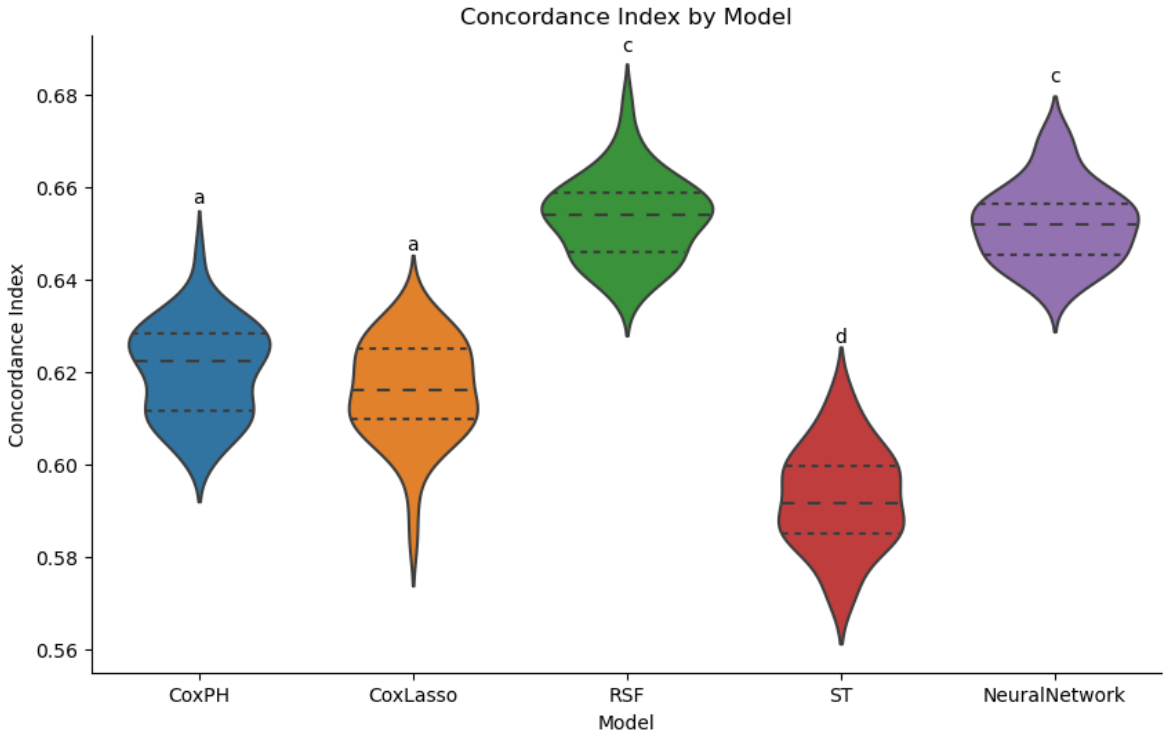
(a) Recidivism

Source	sum_sq	df	F	PR(>F)
Model	0.134	4.0	344.198	2.949×10^{-99}
Residual	0.024	245.0		

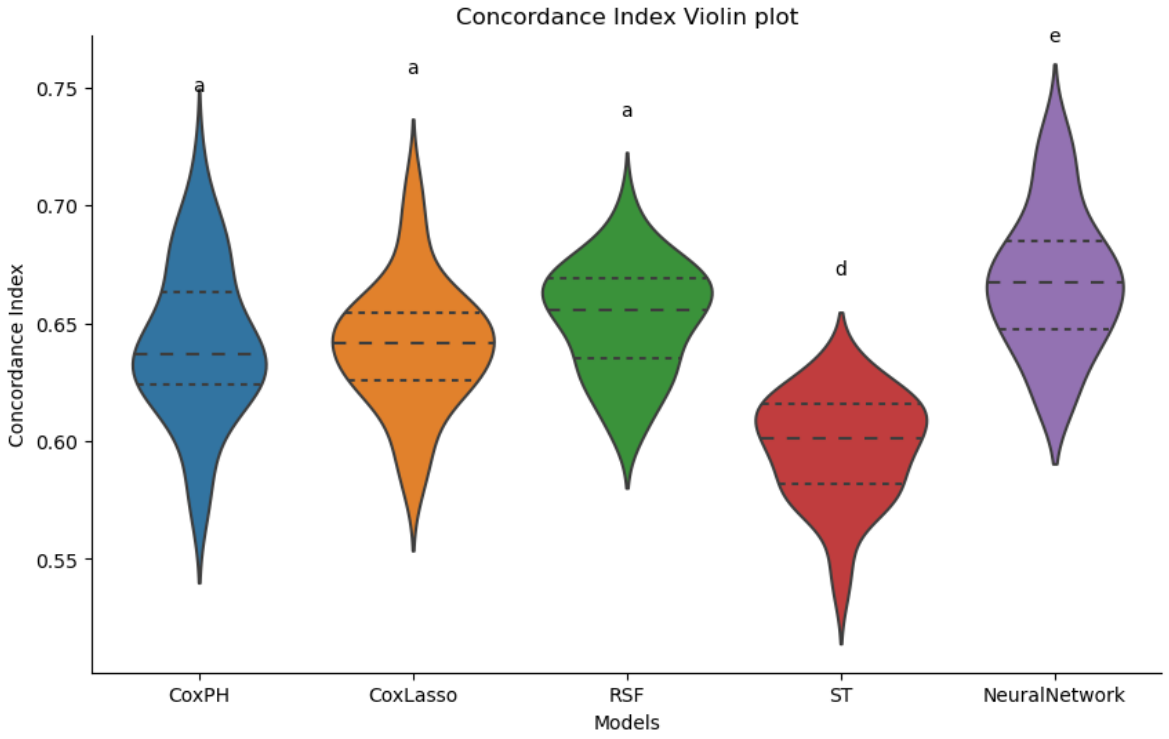
(b) Alcohol Intake

Source	sum_sq	df	F	PR(>F)
Model	0.138	4.0	45.831	1.09×10^{-28}
Residual	0.184	245.0		

Figure 4.6(a) shows the C-index violin distribution across different models for the recidivism dataset. Violin plots for CoxPH exhibit a bimodal distribution and and CoxLasso with a smooth gradually increasing curve, and fading again with long tail. These violins are annotated with the letter "a" indicating no statistically significant difference between these models. The RSF and NN models have violin plots positioned on top of the others, resembling a diamond-like shape. This violin is marked with the letter "c" to signify no difference between them but statistical distinction from other models. The ST model's violin is positioned below all others, marked with the letter "d" showing significance different among other models.



(a) Recidivism violin plot



(b) Alcohol Intake violin plot

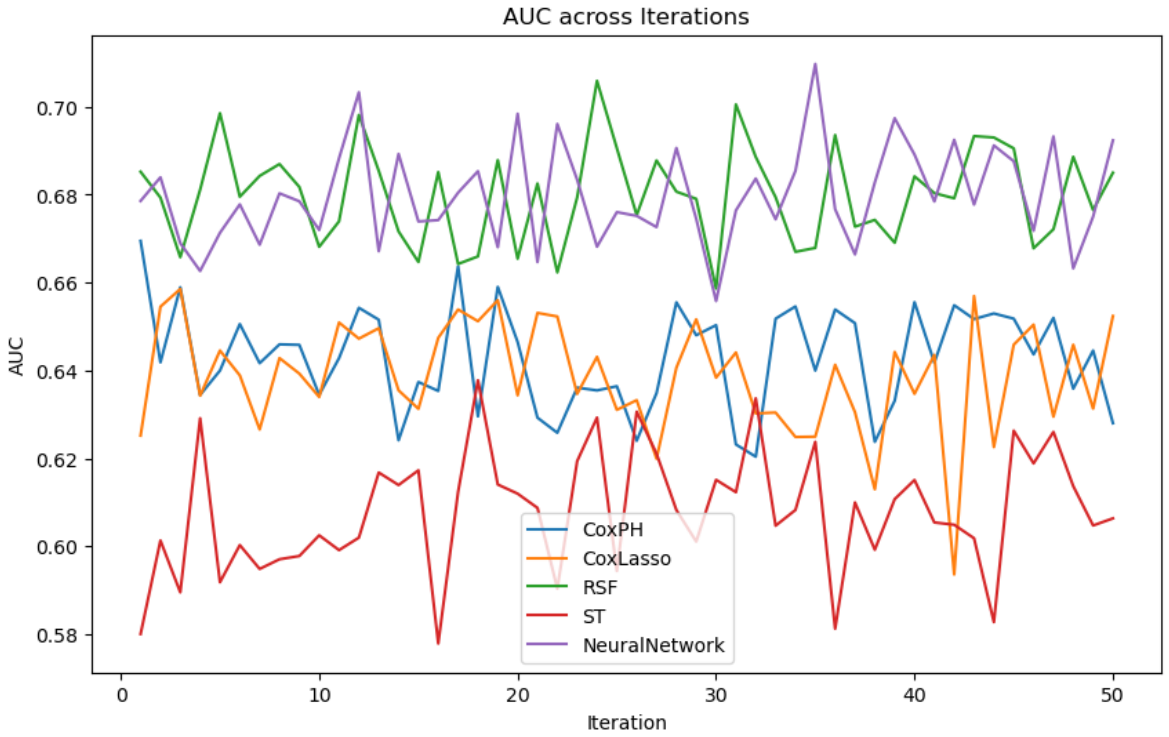
Figure 4.6: Violin plots illustrating the distribution performance measures, obtained from Tukey's HSD test

Figure 4.6(b) presents the distribution of C-index values across different models for the alcohol intake dataset. Violin plots for CoxPH, CoxLasso, and RSF are annotated with the letter "a" indicating no statistically significant difference between these models. In contrast, the ST and NN model's violin with a skewed distribution is annotated with the letter "d" and "e" respectively to signify its statistical distinction from other models.

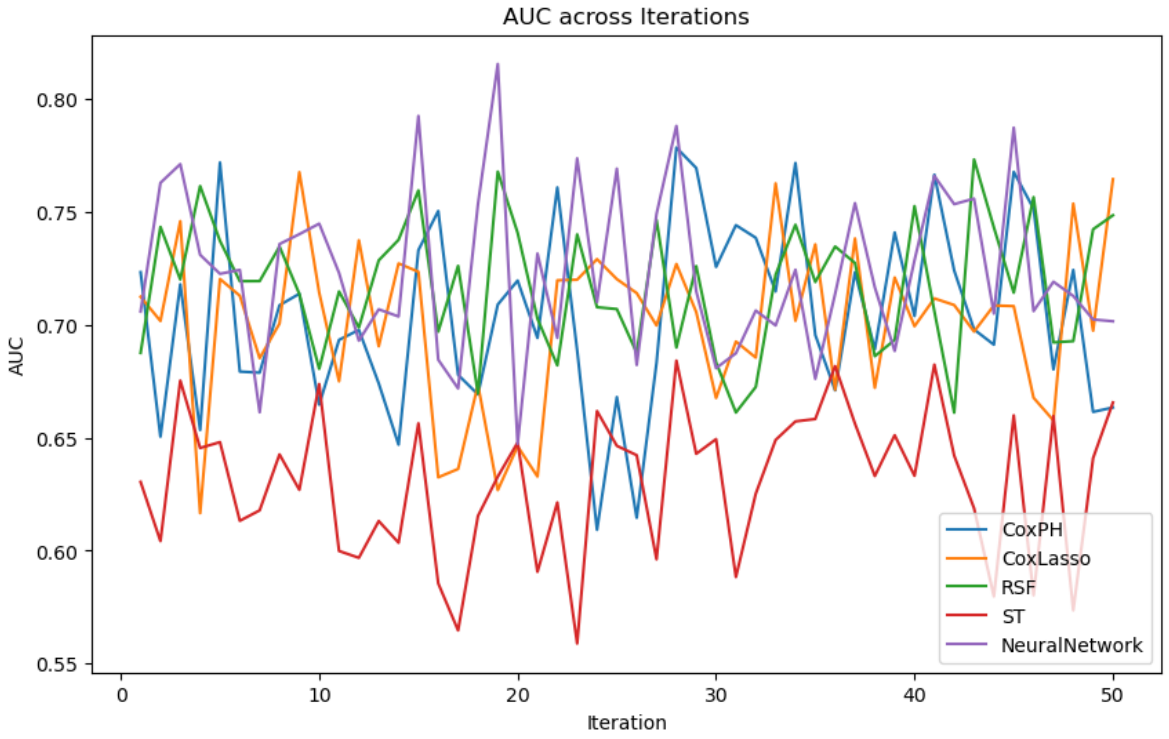
4.7 Area Under the Curve

Figure 4.7(a) illustrating the AUC across 50 iterations provides insight into the performance of various models. Both the Neural Network and Random Survival Forest (RSF) exhibit slight fluctuations, with AUC score around 0.68, presenting a poor discrimination. Moreover, the CoxPH and Lasso-regularized Cox (CoxLasso) models demonstrate relatively consistent performance, with CoxPH slightly above CoxLasso with AUC score around 0.64. The Survival Tree (ST) model consistently performs less favorably compared to the others, with AUC score consistently around 0.60 throughout the iterations.

Figure 4.7(b) reveals a similarly mixed notable fluctuation performance to those of C-index among the models over 50 iterations. However, it is evident that CoxPH, CoxLasso, NN, and RSF exhibit comparable score performance, consistently around 0.70, while ST consistently appear at the bottom with AUC score around 0.62.



(a) Recidivism plot



(b) Alcohol Intake plot

Figure 4.7: Serial plots of area under the curve values obtained when the CoxPH, CoxLasso, random survival forest, survival tree and neural network are applied to the recidivism and alcohol intake data sets

Table 4.4 below presents average values and standard errors of the AUC for both the alcohol intake and recidivism datasets. For recidivism, the Neural Network model and RSF model both demonstrate the highest average AUC values of 0.680 and 0.643, respectively. The NN model also has the lowest standard error at 0.0018, indicating its robust performance. The RSF model follows closely behind with an standard error of 0.0019. In contrast, the CoxPH and CoxLasso models exhibit slightly lower average AUC values for recidivism prediction, standing at 0.639 and 0.643, respectively, with comparable standard errors.

For alcohol intake, the Neural Network model outperforms others with the highest average AUC of 0.724 and the lowest standard error of 0.0016. The RSF model also shows competitive performance with an average AUC of 0.717 and a standard error of 0.0017. The CoxPH and CoxLasso models again demonstrate slightly lower average AUC values for alcohol intake compared to the NN and RSF models.

The Survival Tree model consistently exhibits lower average AUC values and higher standard errors compared to other models for both recidivism and alcohol intake.

Table 4.4: Average values and standard errors of the area under curve obtained when the Cox PH, Cox-lasso, Random Survival Forest (RSF), Survival Tree (ST), and Neural Network are applied to 50 subsamples of the alcohol intake and recidivism datasets.

Recidivism			Alcohol Intake		
Models	Avg AUC	Std Err	Models	Avg AUC	Std Err
CoxPH	0.643	0.0019	CoxPH	0.705	0.0019
CoxLasso	0.639	0.0020	CoxLasso	0.701	0.0017
RSF	0.680	0.0018	RSF	0.717	0.0017
ST	0.608	0.0019	ST	0.630	0.0018
NN	0.680	0.0018	NN	0.724	0.0016

The ANOVA results for the recidivism dataset, as presented in Table 4.5(a), provide statistical evidence of significant differences among the models. The "Model" factor contributes substantially to the variability in AUC values, as indicated by a large F-statistic of 314.71.

Similarly, Table 4.5(b) displays the ANOVA results for the alcohol intake dataset, demonstrating a significant overall difference among the survival models. This is evidenced by a p-value of 4.9635×10^{-34} and F-statistic of 57.190881

These results underscore that at least one model significantly outperforms the others in both recidivism and alcohol intake prediction tasks. Further illustration of these significant differences among the models is presented in the violin plots on the next page

Table 4.5: ANOVA results obtained when the Cox PH, Cox-lasso, Random Survival Forest (RSF), Survival Tree (ST), and Neural Network are applied to 50 subsamples of the alcohol intake and recidivism datasets

(a) Recidivism

Source	sum_sq	df	F	PR(>F)
Model	0.187	4.0	314.711	3.024×10^{-95}
Residual	0.036	245.0		

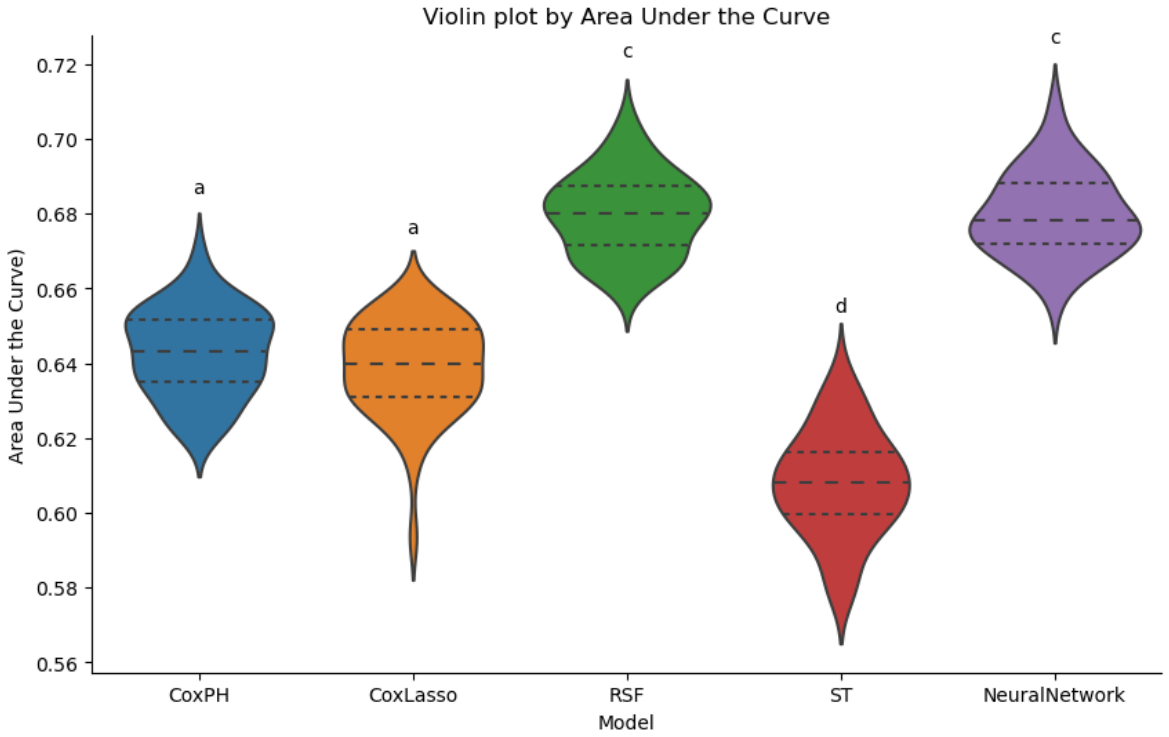
(b) Alcohol Intake

Source	sum_sq	df	F	PR(>F)
Model	0.281	4.0	57.191	4.964×10^{-34}
Residual	0.301	245.0		

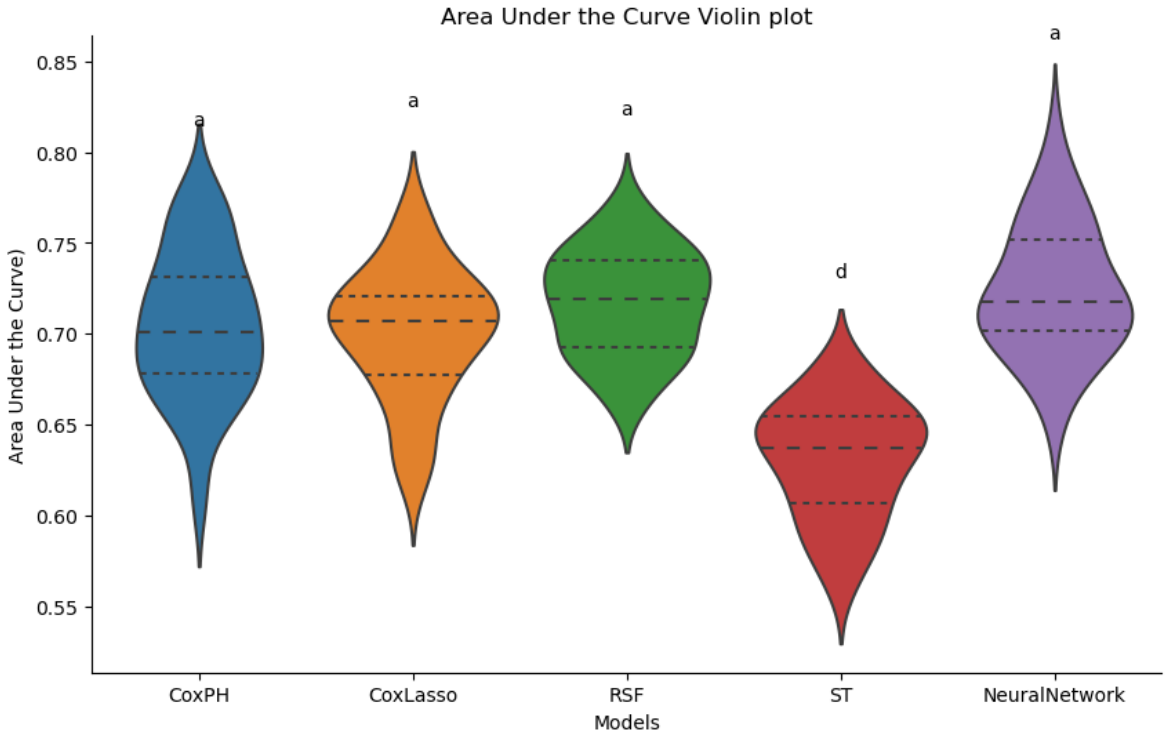
Figure 4.8(a) displays the distribution of AUC violin plots across different models for the recidivism dataset. Violin plots for CoxPH and CoxLasso exhibit similar distributions, characterized by a smooth, gradually increasing curve, and fading again. Distinctively, the CoxLasso model's violin extends downwards with a long tail and features a small downwards tail with a small curve.

These violins are annotated with the letter "a" indicating no statistically significant difference between these models. The RSF and NN models have violin plots positioned on top of the others, resembling a diamond-like shape. This violin is marked with the letter "c" to signify its statistical distinction from other models. The ST model's violin is positioned below all others and features a smooth peak, marked with the letter "d" showing significance different among other models.

Figure 4.8(b) presents the distribution of AUC values across different models for the alcohol intake dataset. Violin plots for CoxPH, CoxLasso, RSF, and NN are annotated with the letter "a" indicating no statistically significant difference between these models. In contrast, the ST model's violin with a skewed distribution is annotated with the letter "d" to signify its statistical distinction from other models.



(a) Recidivism violin plot



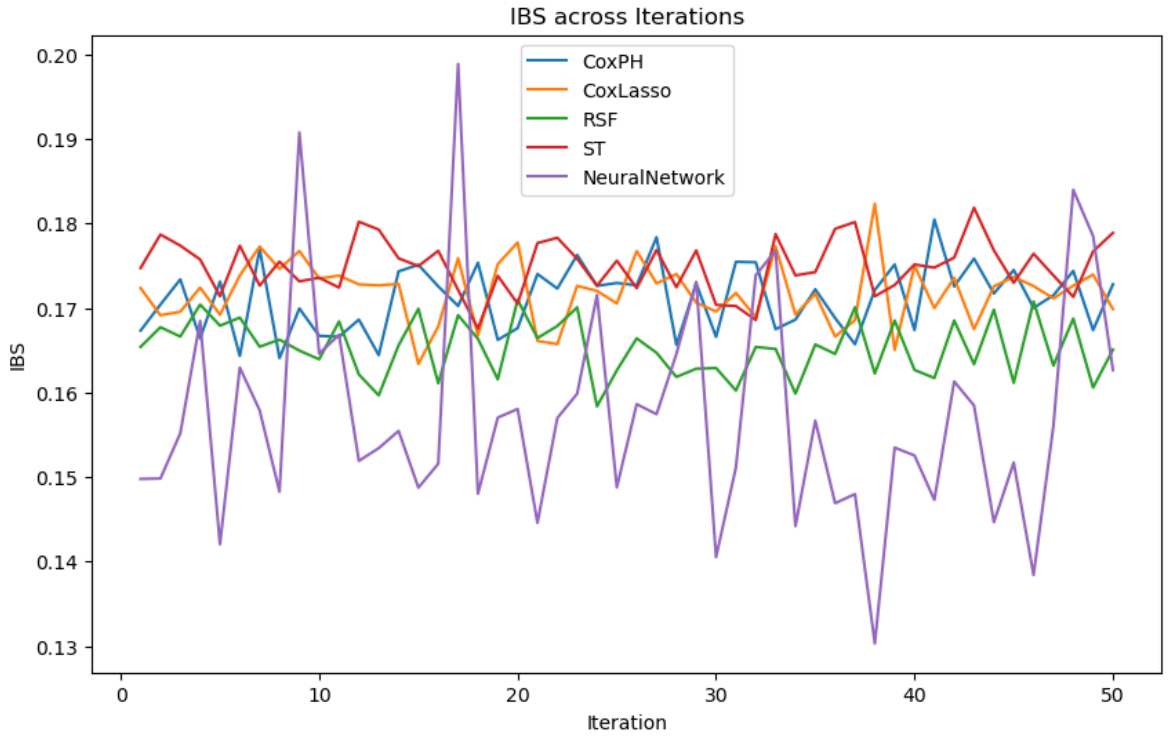
(b) Alcohol Intake violin plot

Figure 4.8: Violin plots illustrating the distribution performance measures, obtained from Tukey's HSD test

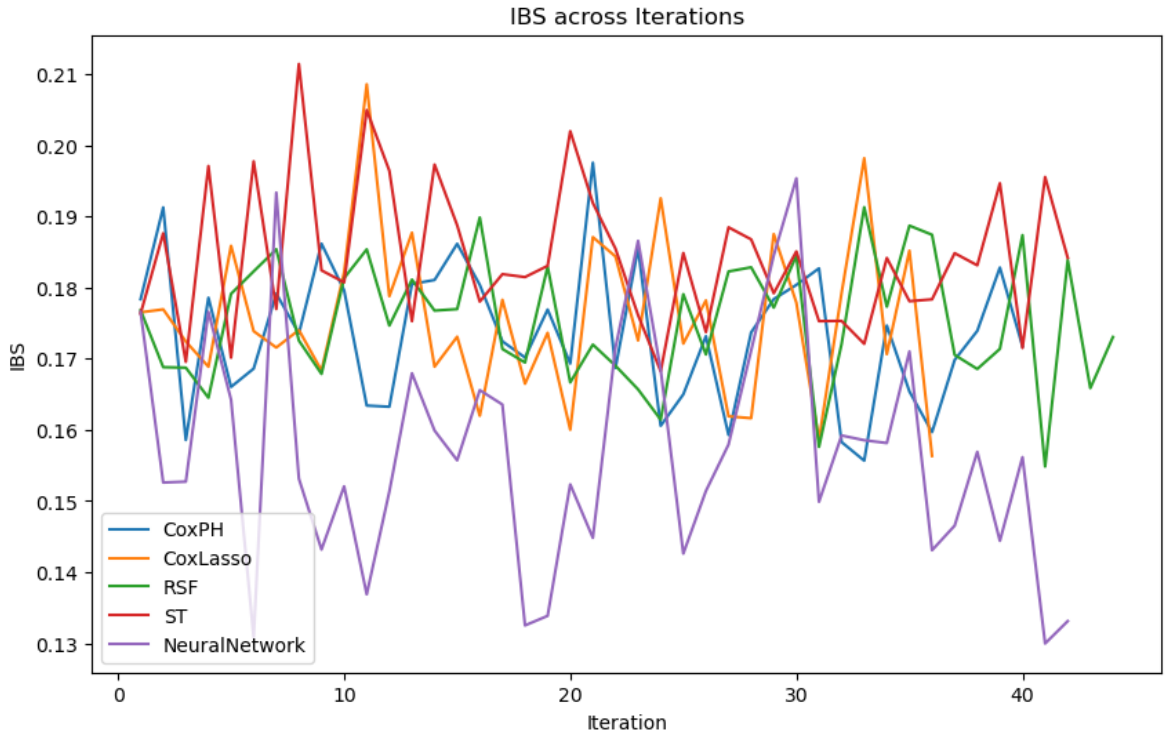
4.8 Integrated Brier Scores

Figure 4.9(a) depicting the Integrated Brier Score (IBS) of recidivism dataset across 50 iterations unveils interesting patterns in model performance with the Neural Network (NN) showing a significant variability, with lowest IBS values significant fluctuating between 0.15 throughout the iterations. Conversely, the Random Survival Forest (RSF) model demonstrates more stable performance and calibration, with moderate IBS values around 0.165. Meanwhile, the CoxPH and CoxLasso models display comparable performance and calibration, with their IBS values closely intertwined around 0.17. Interestingly, the Survival Tree (ST) model consistently maintains slightly higher IBS values than other models with a score around 0.178. Based on the IBS values across 50 iterations, the Random Survival Forest appears most stable, while the Neural Network exhibits the most variability, with other models falling in between in terms of stability and calibration.

Figure 4.9(b) further underscores the performance dynamics of the models on alcohol intake dataset. The Neural Network (NN) shows significant fluctiations, with lowest IBS values around 0.17 and 0.13 throughout the iterations. Conversely, the random survival forest (RSF), Cox PH and CoxLasso display a mixed comparable performance and calibration, with their IBS values closely intertwined around 0.16 and 0.19. Interestingly, the model ST consistently maintains slightly higher IBS reaching the score of 0.20 at some iterations. Similarly to the results of recidivism data, the Random Survival Forest appears most stable, while the Neural Network exhibits the most variability, with other models falling in between in terms of stability and calibration.



(a) Recidivism plot



(b) Alcohol Intake plot

Figure 4.9: Serial plots of integrated brier scores obtained when the CoxPH, CoxLasso, random survival forest, survival tree and neural network are applied to the recidivism and alcohol intake data sets

Table 4.6 presented below compares the predictive performance of different models for recidivism and alcohol intake, measured by the average integrated Brier score (Avg IBS) and associated standard errors. The Neural Network (NN) model outperforms others, yielding the lowest Avg IBS values for both recidivism (0.157) and alcohol intake (0.158) with small standard errors (0.00064 and 0.0004, respectively), highlighting its robust predictive accuracy. The Cox Proportional Hazards (CoxPH) model follows closely with competitive Avg IBS values of 0.171 and 0.174, accompanied by standard errors of 0.00062 and 0.0006 respectively. While Random Survival Forest (RSF) and CoxLasso models exhibit slightly higher Avg IBS values, they remain within a comparable range, and the Survival Trees (ST) model, while not the top performer, contributes to the overall comparison.

Table 4.6: Average values of the integrated Brier score and standard errors obtained when the Cox PH, Cox-lasso, Random Survival Forest (RSF), Survival Tree (ST), and Neural Network are applied to 50 subsamples of the alcohol intake and recidivism datasets.

Recidivism			Alcohol Intake		
Models	Avg IBS	Std Err	Models	Avg IBS	Std Err
CoxPH	0.171	0.00062	CoxPH	0.174	0.0006
CoxLasso	0.172	0.00067	CoxLasso	0.176	0.0005
RSF	0.165	0.00065	RSF	0.175	0.0005
ST	0.175	0.00074	ST	0.184	0.0006
NN	0.157	0.00064	NN	0.158	0.0004

The ANOVA results for recidivism in Table 4.7(a) for the IBS indicate significant differences among the models. The "Model" factor contributes significantly to the variability in IBS values, as evidenced by a p-value below 0.05 and a large F-statistic of 53.17.

Additionally for alcohol intake dataset in Table 4.7(b) evaluates the variability within the models. The results show an exceedingly small p-value (2.70044×10^{-19}), indicating strong evidence suggesting that the variability between groups is statistically significant.

Table 4.7: ANOVA results obtained when the Cox PH, Cox-lasso, Random Survival Forest (RSF), Survival Tree (ST), and Neural Network are applied to 50 subsamples of the alcohol intake and recidivism datasets

(a) Recidivism

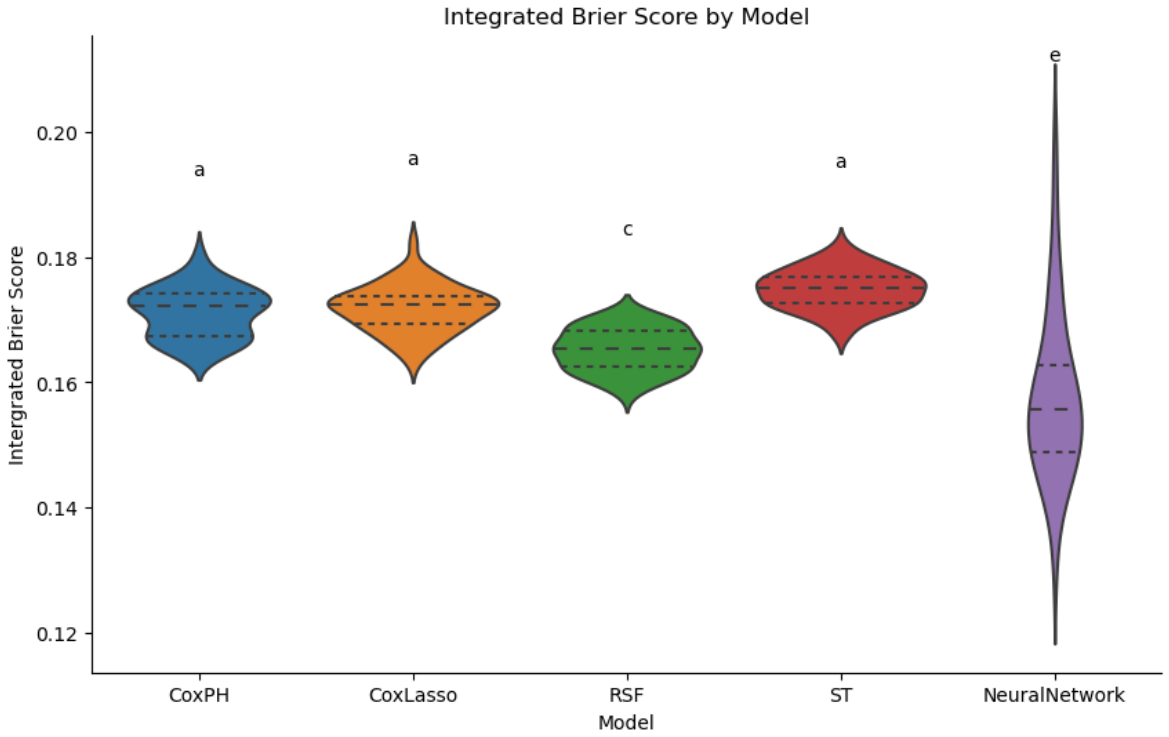
Source	sum_sq	df	F	PR(>F)
Model	0.01	4.0	53.172	3.279×10^{-32}
Residual	0.011	245.0		

(b) Alcohol Intake

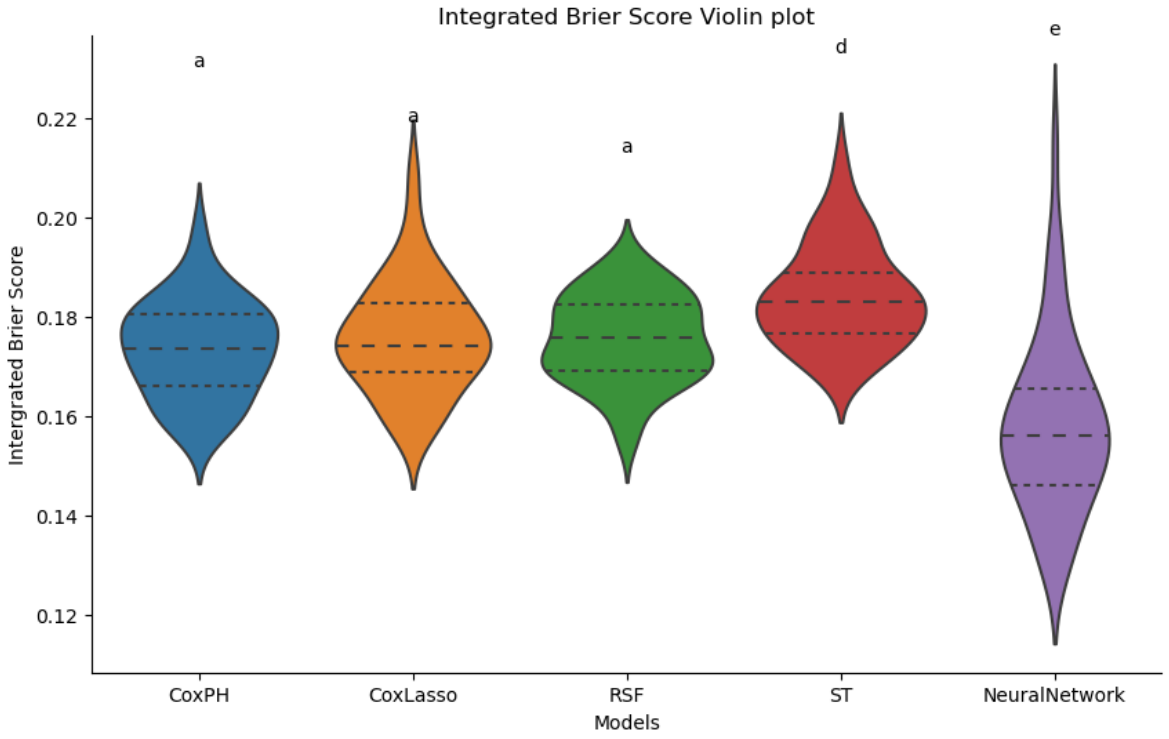
Source	sum_sq	df	F	PR(>F)
Model	0.017	4.0	29.276	2.700×10^{-19}
Residual	0.030	207.0		

The violin plots in Figure 4.10(a) illustrates the distribution of Integrated Brier Score (IBS) values across different models for the recidivism dataset. Violin plots for CoxPH, CoxLasso, and ST are annotated with the letter "a" indicating no statistically significant difference between these models. The CoxPH model's violin shows a bimodal distribution with two distinct peaks. Similarly, the CoxLasso model's violin exhibits high variability and skewness towards the extremes. The ST model's violin is small but wide. In contrast, the RSF model's violin resembles a skewed distribution with an oval shape, and is annotated with the letter "c" to signify its statistical distinction from other models. The NN model's violin suggests a distribution that is likely skewed. It is annotated as "e" to highlight its statistical differentiation from other models.

Figure 4.10(b) illustrates the distribution of Integrated Brier Score values across different models for the alcohol dataset. Violin plots for CoxPH, CoxLasso, and RSF are annotated with the letter "a" indicating no statistically significant difference between these models. The CoxPH model's violin shows a symmetrically like distribution with one peak. Similarly, the CoxLasso model's violin exhibits high variability and skewness towards the extremes. The RSF model's violin is small but wide, with bimodal distribution. In contrast, the ST model's violin resembles a skewed distribution with an oval shape, and is annotated with the letter "d" to signify its statistical distinction from other models. The NN model's violin suggests a distribution that is likely skewed. It is annotated as "e" to highlight its statistical differentiation from other models.



(a) Recidivism violin plot



(b) Alcohol Intake violin plot

Figure 4.10: Violin plots illustrating the distribution performance measures, obtained from Tukey's HSD test

4.9 Discussion of Results

The analysis of survival models applied to both the recidivism and alcohol intake datasets reveals insights into their performance. Across multiple evaluation measures, including the Area Under the Curve, Concordance Index, and Integrated Brier Score, the Neural Network (NN) consistently emerges as the top-performing model for both datasets. For the recidivism dataset, NN demonstrates a superior performance with avg C-index: 0.6522, avg AUC: 0.680 and avg IBS: 0.157, followed closely by the Random Survival Forest (RSF) model, which also displays competitive performance with average C-index: 0.6533, average AUC: 0.680 and average IBS: 0.165. In contrast, traditional Cox Proportional Hazards (CoxPH) and Lasso-regularized Cox (CoxLasso) models exhibit slightly lower but relatively stable performance with average C-index: 0.6203, average AUC: 0.643, and average IBS: 0.171, while the Survival Tree (ST) model consistently lags behind other models with average C-index: 0.5927, average AUC: 0.608, and average IBS: 0.175. ANOVA results confirm the statistical significance of these performance differences (Recidivism: F-statistic 344.198, p-value < 0.05).

Similarly, for the alcohol intake dataset, NN demonstrates the highest perfomance with avg C-index: 0.6678, avg AUC: 0.724, and avg IBS: 0.158, followed closely by RSF with avg C-index: 0.6520, avg AUC: 0.717, and avg IBS: 0.175, while CoxPH and CoxLasso models exhibit slightly lower performance with average C-index: 0.6416, average AUC: 0.705, average IBS: 0.174. Again, ST consistently performs the poorest with average C-index: 0.5972, average AUC: 0.630, and average IBS: 0.184, with ANOVA results confirming significant performance differences (Alcohol Intake: F-statistic 57.191, p-value < 0.05).

The results of this study align with the findings of previous researchers in the field of predictive modeling. For instance, Spooner et al. (2020) demonstrated the superiority of ML models, such as Neural Networks (NN), in predicting complex outcomes using heterogeneous and censored clinical data. Our results corroborate these findings, showing that Neural Network consistently outperformed traditional statistical models like Cox Proportional Hazards (CoxPH) across multiple evaluation metrics including Integrated Brier score, Concordance index, Area Under the Curve. Moreover, the analysis revealed insights into the performance of various survival models for predicting both recidivism and alcohol intake. Specifically, the Neural Network model demonstrated superior performance with higher C-index and AUC values compared to other models, consistent with findings from previous studies (Oei et al. (2021)). Furthermore, Random Survival Forest (RSF) models displayed competitive performance, closely following Neural Network in predictive accuracy.

When considered alongside Kantidakis et al. (2023) research, shows insights into the comparative performance of statistical and machine learning models. While Kantidakis et al. (2023) study primarily focused on comparing partial logistic artificial neural networks with Cox models in a clinical setting, our investigation extends this comparison to include additional ML algorithms such as Neural Networks and Random Survival Forests (RSF). Interestingly, both studies converge on the effectiveness of Neural Networks in predictive modeling tasks, with Neural Network consistently outperforming traditional Cox models across various evaluation metrics. Kantidakis et al. (2023) findings highlight the comparable predictive performance of PLANNs with Cox models, with different metrics, integrated Brier score, brier score, C-index. Similarly, our results demonstrate a better performance of Neural Networks, particularly in predicting survival outcomes in complex datasets related to recidivism and alcohol intake with diferent metrics including C-index, AUC, and IBS.

The selection of NN and RSF models for predicting recidivism and alcohol intake was based on observations gained from both the literature review and the objectives of this study. These models were chosen for their ability to deal with high-dimensional, heterogeneous data and their demonstrated effectiveness in previous research (Churpek et al. (2017)). When choosing between Machine Learning and Statistical models one should consider the specific characteristics of data type to use (Suresh et al., 2022). It's not only about predictive accuracy there are also other issues to include like type of data suitable for each model. These findings reveal that the volume and quality of data significantly impact the suitability of statistical and machine learning models. Specifically, data preferences influence the performance and reliability of these models, with statistical models typically good on smaller and well-structured datasets, while machine learning models often require large, diverse datasets to capture complex patterns.

Chapter 5

Conclusion

5.1 Summary

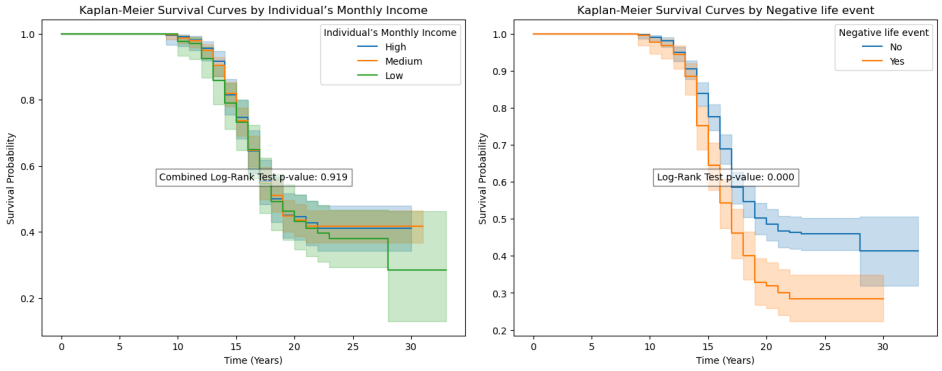
The comparison of statistical and machine learning models for continuous survival analysis in predicting recidivism and alcohol intake yields valuable insights. Across multiple evaluation metrics, including the Concordance Index, Area Under the Curve, and Integrated Brier Score, Neural Network (NN) consistently emerges as the top-performing model for both datasets, closely followed by Random Survival Forest (RSF). The traditional statistical models such as Cox Proportional Hazards (CoxPH) and Lasso-regularized Cox (Coxlasso) exhibit stable but slightly lower performance, while Survival Tree (ST) lags behind all other models. The findings underscore the dominance of advanced machine learning techniques, particularly in handling high-dimensional, heterogeneous data. These results show that the applicability of statistical and machine learning models is highly dependent on the size and quality of data (Suresh et al., 2022). Moreover, data preferences affect the performance and reliability of these models; machine learning methods tend to require vast, diverse datasets to capture complicated patterns, whereas statistical models generally function well on smaller, well-structured datasets (Suresh et al., 2022). The practical challenges associated with implementing neural networks, including data preprocessing and computational intensity, highlight

the need for careful consideration of methodology based on factors such as data characteristics, computational resources, and researcher expertise.

5.2 Conclusion

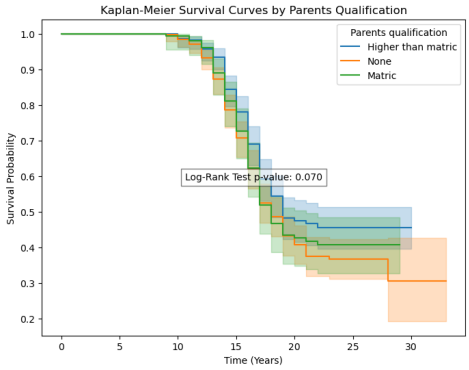
The implications of our findings are significant for decision making in real world settings. The dominant performance of machine learning models suggests that these approaches could be valuable techniques for identifying individuals at high risk of recidivism or alcohol intake, thereby informing targeted interventions and resource allocation. Moreover, the identification of specific models with better performance capabilities opens avenues for future research, such as exploring statistical and machine learning models or incorporating additional features to further enhance predictive accuracy in survival analysis. Future research should explore the integration of Machine learning and Statistical models with different types of data.

Appendix A

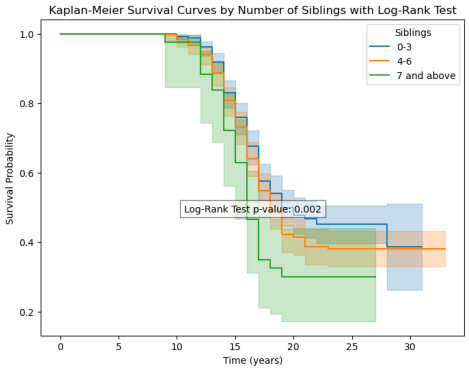


(a) K-M by individual income

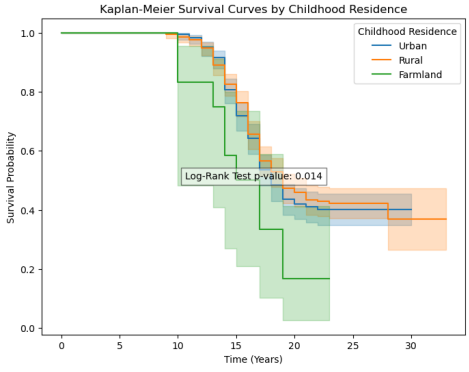
(b) K-M by negative life event



(c) K-M by parent qualification



(d) K-M by siblings



(e) K-M by childhood residence

Figure 5.1: Kaplan Meier survival curves

References

P. D. Allison. *Event history and survival analysis*. Sage Publications, Inc., 1984.

A. Andersen. Discrete models for the analysis of time series. *Journal of the American Statistical Association*, 65(332):9–33, 1970.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3): 399–424, 2011.

A. Bender, D. Rügamer, F. Scheipl, and B. Bischl. A general machine learning framework for survival analysis. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pages 158–173. Springer, 2021.

E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

E. Biganzoli, P. Boracchi, and E. Marubini. A general framework for neural network models on censored survival data. *Neural Networks*, 15(2):209–218, 2002.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5(none):44 – 71, 2011. doi: 10.1214/09-SS047. URL https://doi.org/10.1214/09-SS047.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

B. W. Brown. Conditional probability models and discrete 'time survival analysis'. *Biometrika*, 62(1):101–110, 1975. doi: 10.1093/biomet/62.1.101.

J. Cai and R. A. Betensky. Semiparametric estimation of the proportional hazards model for interval-censored data using penalized splines. *Biometrics*, 59(2):383–392, 2003. doi: 10.1111/1541-0420.00045.

J. J. Caro and J. Möller. Advantages and disadvantages of discrete-event simulation for health economic analyses. *Expert review of pharmacoeconomics & outcomes research*, 16(3):327–329, 2016.

B. Cheng and D. M. Titterington. Neural networks: A review from a statistical perspective. *Statistical science*, pages 2–30, 1994.

M. Z. I. Chowdhury, A. A. Leung, R. L. Walker, K. C. Sikdar, M. O'Beirne, H. Quan, and T. C. Turin. A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a canadian population. *Scientific Reports*, 13(1):13, 2023.

M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson. Multicentre comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Journal of biomedical informatics*, 69:112–122, 2017.

A. Ciampi, R. Bush, M. Gospodarowicz, and J. Till. An approach to classifying prognostic factors related to survival experience for non-hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975. *Cancer*, 47(3):621–627, 1981.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

D. R. Cox, H. Nguyen, and E. J. Snell. A simple neural network for survival analysis. *Statistics in Medicine*, 9(6):649–657, 1990. doi: 10.1002/sim.4780090604.

J. A. Cruz and D. S. Wishar. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2:59–78, 2006.

R. De Bin. Boosting in cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the r-packages coxboost and mboost. *Journal of Statistical Software*, (31):513–531, 2016. doi: 10.1007/s00180-015-0642-2.

T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer, 2000.

B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

B. Efron. Regression and anova with zero-one data: measures of residual variation. *Journal of the American Statistical Association*, 83(402):708–712, 1988.

A. Eleuteri, M. Aung, A. Taktak, B. Damato, and P. Lisboa. Continuous and discrete time survival analysis: neural network approaches. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5420–5423. IEEE, 2007.

F. Emmert-Streib and M. Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019.

L. Fahrmeir. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 1998. doi: 10.1007/978-1-4757-2815-3.

L. Fahrmeir and T. Kneib. Bayesian smoothing of hazard functions in proportional hazards models. *Statistics and Computing*, 21(3):387–407, 2011.

L. Fahrmeir and L. Knorr-Held. Dynamic regression models for survival data: An extension of the cox model. *Statistics in Medicine*, 16(14):1521–1538, 1997.

D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

D. M. Finkelstein. Proportional hazards models with interval-censored data. *Biometrics*, pages 427–438, 1986. doi: 10.2307/2530907.

L. Gordon and R. A. Olshen. Tree-structured survival analysis. *Cancer Treatment Reports*, 69(10):1065–1069, 1985.

M. Greenwood. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26, 1926. URL https://doi.org/10.1038/118194b0.

J. Huang, C. Ling, and P. Li. Using neural networks for conditional survival analysis. *Statistics in Medicine*, 23(4):581–596, 2004. doi: 10.1002/sim.1638.

C. Huber-Carol, S. Gross, and F. Vonta. Risk analysis: survival data analysis vs. machine learning. application to alzheimer prediction. *Comptes Rendus Mecanique*, 347(11):817–830, 2019.

S. Hyun, J. Lee, and R. Yearout. Parameter estimation of type-i and type-ii hybrid censored data from the log-logistic distribution. *Industrial and systems engineering review*, 4(1):37–44, 2016.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008. doi: 10.1214/08-AOAS169. URL https://doi.org/10.1214/08-AOAS169.

H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.

H. Iswaran, U. Kogalur, E. Blackstone, and M. Lauer. Random survival forest. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

B. C. Jaeger, D. L. Long, D. M. Long, M. Sims, J. M. Szychowski, Y.-I. Min, L. A. Mcclure, G. Howard, and N. Simon. Oblique random survival forests. *The Annals of Applied Statistics*, 13(3):1847–1883, 2019.

J. D. Kalbfleisch and D. E. Schaubel. Fifty years of the cox model. *Annual Review of Statistics and Its Application*, 10, 2023.

G. Kantidakis, E. Biganzoli, H. Putter, M. Fiocco, et al. A simulation study to compare the predictive performance of survival neural networks with cox models for clinical trial data. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

G. Kantidakis, H. Putter, S. Litière, and M. Fiocco. Statistical models versus machine learning for competing risks: development and validation of prognostic models. *BMC Medical Research Methodology*, 23(1):51, 2023.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical Association*, 53(282):457–481, 1958.

C. Kartsonaki. Survival analysis. *Diagnostic Histopathology*, 22(7):263–270, 2016.

S. Kelly. An event history analysis of teacher attrition: Salary, teacher tracking, and socially disadvantaged schools. *The Journal of Experimental Education*, 72(3):195–220, 2004.

S. Kim. *The Comparison of Discrete and Continuous Survival Analysis*. PhD thesis, Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, 2014.

J. P. Klein and M. L. Moeschberger. *Survival Analysis*. Statistics for Biology and Health. Springer New York, NY, 2 edition, 2003. ISBN 978-0-387-21645-4 (eBook). doi: https://doi.org/10.1007/b97377.

K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational*

and Structural Biotechnology Journal, 13:8–17, 2015. ISSN 2001-0370. doi: 10.1016/j.csbj.2014.11.005. URL http://dx.doi.org/10.1016/j.csbj.2014.11.005.

H. Kvamme and Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27:710–736, 2021. doi: 10.1007/s10985-021-09532-6. URL https://doi.org/10.1007/s10985-021-09532-6.

N. M. Laird and D. C. Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American statistical association*, 76(374): 231–240, 1981. doi: 10.1080/01621459.1981.10477616.

M. LeBlanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

N. Mantel and B. F. Hankey. Statistical methods for comparing survival times and distributions. *Journal of the National Cancer Institute*, 60(3):575–585, 1978. doi: 10.1093/jnci/60.3.575.

L. Mariani, D. Coradini, E. Biganzoli, P. Boracchi, E. Marubini, S. Pilotti, B. Salvadori, R. Silvestrini, U. Veronesi, R. Zucali, et al. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear cox regression model and its artificial neural network extension. *Breast cancer research and treatment*, 44:167–178, 1997.

M. Mills. *Introducing survival and event history analysis*. Sage, 2010.

P. A. Murtaugh, L. D. Burns, and J. Schuster. Predicting the retention of university students. *Research in higher education*, 40(3):355–371, 1999.

B. Muthén and K. Masyn. Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics*, 30(1):27–58, 2005.

R. W. Oei, Y. Lyu, L. Ye, F. Kong, C. Du, R. Zhai, T. Xu, C. Shen, X. He, L. Kong, C. Hu, and H. Ying. Progression-free survival prediction in patients with nasopha-

ryngeal carcinoma after intensity-modulated radiotherapy: Machine learning vs. traditional statistics. *Journal of Personalized Medicine*, 11(8), 2021. ISSN 2075-4426. doi: 10.3390/jpm11080787. URL https://www.mdpi.com/2075-4426/11/8/787.

Overleaf. Overleaf: The Online LaTeX Editor, n.d. URL https://www.overleaf.com/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A*, 135(2):185–207, 1972. doi: 10.2307/2344317. URL https://www.jstor.org/stable/2344317.

S. Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html.

R. L. Prentice and L. A. Gloeckler. Bias in linear regression due to covariate measurement error. *Biometrics*, pages 469–480, 1978a. doi: 10.2307/2529779.

R. L. Prentice and L. A. Gloeckler. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, pages 57–67, 1978b.

G. Rabinowitz, Y. Ritov, and A. Schuster. Nonparametric maximum likelihood estimation of survival function with doubly censored data. *Biometrika*, 82(2):287–298, 1995.

S. Rathod, A. Saha, and K. Sinha. Introduction to survival analysis and application in agricultural research. *Food and Scientific Reports*, 1:28–30, 2020. ISSN 2582-5437.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

J. D. Singer and J. B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics*, 18(2): 155–195, 1993.

H. G. Sithuba. The determinants of age at first alcohol use by students at tertiary institutions around thohoyandou, 2018. Department of Statistics.

A. Spooner, E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor, and H. Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, 10(1):1–10, 2020.

L. Štěpánek, F. Habarta, I. Malá, L. Štěpánek, M. Nakládalová, A. Boriková, and L. Marek. Machine learning at the service of survival analysis: Predictions using time-to-event decomposition and classification applied to a decrease of blood antibodies against covid-19. *Mathematics*, 11(4):819, 2023.

K. Suresh, C. Severn, and D. Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22(1):207, 2022.

T. M. Therneau. Extending the cox model. In *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, pages 51–84. Springer, 1997.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. CRC Press, 2000.

M. E. Thompson. Logistic regression with dependent observations. *Journal of the American statistical association*, 72(359):135–140, 1977. doi: 10.1080/01621459.1977.10480613.

W. Thompson Jr. On the treatment of grouped observations in life studies. *Biometrics*, pages 463–470, 1977.

H.-C. Thorsen-Meyer, D. Placido, B. S. Kaas-Hansen, A. P. Nielsen, T. Lange, A. B. Nielsen, P. Toft, J. Schierbeck, T. Strøm, P. J. Chmura, et al. Discrete-time survival

analysis in the critically ill: a deep learning approach using heterogeneous data. *NPJ digital medicine*, 5(1):142, 2022.

G. Tutz, M. Schmid, et al. *Modeling discrete time-to-event data*. Springer, 2016.

J. Wang, W.-Y. Chang, and Y. Ning. Additive hazards regression for interval-censored failure time data. *Journal of the American Statistical Association*, 105(490):1457–1465, 2010.

P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

WindEsco. Violin plots, 2024. URL https://www.windesco.com/knowledge/violin-plots.

D. M. Witten and R. J. Tibshirani. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 18(5):573–593, 2009.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

D. Zeng, D. Lin, and L. J. Wei. Semiparametric transformation models for survival data: a self-consistency approach. *Journal of the American Statistical Association*, 101(474):670–680, 2006.

B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75, 2000.