Bayesian Network for Censored Survival Data using Inverse Probability Censored Weighting (IPCW)

Vivek Kumar, Bachelor of Engineering

A Dissertation

Presented to the university of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Bahman Honari

September 2021

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Vivek Kumar

September 03, 2021

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Vivek Kumar

September 03, 2021

Bayesian Network for Censored Survival Data using Inverse Probability Censored Weighting (IPCW)

Vivek Kumar, Master of Science in Computer Science

University of Dublin, Trinity College Dublin, 2021

Supervisor: Dr. Bahman Honari

Estimating the chances of experiencing life threatening health outcomes over a specific time interval is an important tool in medical field to take adequate action much before the patient would encounter it and avoid the risk of happening. Moreover, if the tool predicts it based on few patients' medical history record, it becomes a highly significant. Since these health records are derived from administrative and clinical database, some records will not have information for entire time frame. Few patients might have disenrolled from the system and hence loss to follow up. This scenario is very common in any health system, still the requirement of such tool is very essential. In statistical terms, the event time is considered right censored when the observations do not have complete follow up and if they experienced the event or not is unknown to us. This dissertation is addressing the problem of such censored survival data of heart transplant by exploring the weight assignment technique for censored observation. We used Inverse Probability Censoring Weight (IPCW) approach of weight allocation and incorporated it with Bayesian network model to show the causal dependency between the features and find out the probability of feature to be present or not based on event status.

Also, we applied learning algorithms for structure and parameter learning of Bayesian Network. We have further evaluated the model based on different evaluation metrics and reported the result. We have developed the model on heart transplant data where Age, prior Surgery and transplant indicator had come out as key factor to affect the risk outcome. Further, we can use this model as general purpose and also can compare it with other machine learning classifiers to predict the outcome.

Keywords Bayesian networks · Kaplan – Meier , Heart Transplant data · Survival analysis · Inverse probability of censoring weights · Risk prediction

Acknowledgments

First and foremost, I would like to thank my supervisor, Dr. Bahman Honari, for his unwavering support throughout the duration of my time as a postgraduate student at Trinity College Dublin. His positivity and encouragement have been the key factor which motivated me and giving me the space to discover my interests and passion towards data science and data driven projects. I sincerely thank him for his patient support and inspirational advice. His guidance was always able to keep me on track when I got carried away, and without his precious feedback at all stages of this work, it would not have been possible.

Also, I owe special thanks to Course Director Professor Dr. John Dingiliana, for his unparalleled support and guidance throughout my study.

I would like to thank my helpful, thoughtful, and wonderful fellow postgraduate student: Sneha, Salil for their academic help and to have wonderful friendship.

VIVEK KUMAR

University of Dublin, Trinity College

September 2021

Contents

Appendix

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The probability of risk predictions of a patient with different health condition and unfavourable events like death are very crucial in healthcare practice. The information on risk probability and proper classification will help medical field practitioners to device specific strategies, better allocation of resources to mitigate the high-risk chances of adverse events and also could provide a plan to patients to stick with such strategy. Survival analysis is one such mechanism with various analysis procedures where the dependent variable of "interest "is the time till an event occurs. In the field of medical and healthcare, these events are death, reoccurrence of disease, sign of new disease or reaction of a treatment. We have multiple machine learning techniques and statistical methods to infer survival models from the data which can predict the forward graph of the patient based on the known variables. Even though machine learning models can handle nonlinear and complex data structures still, the conventional ML models are not suitable for survival analysis because of inability to predict time to event occurrences and most of the time observation time gets ignored [1].

Bayesian Networks [2] are excellent model to handle such scenarios and it is best for knowledge representation. They can express causal dependencies in covariates and represent them probabilistically. Also, they can learn the parameters and structure from the data which corresponds to causality and human reasoning. In this thesis, we will explore how to use Bayesian networks for censored survival data with conditional probability within covariates. Furthermore, we will also check the classification along with causal dependency.

1.2 Research Question

This research lies on two foundations: Survival Analysis of a clinical dataset where the length of time is highly variable among each subject. Due to that, most of the observations doesn't have sufficient information in terms of follow up time to predict if they have experienced the event of interest (death, in this case) within a given time period. Such data also termed as right censored.

The data which we are dealing with is incomplete follow up and censoring.

Secondly, to represent the information in a proper probabilistic graphical network to analyse the risk predictions and causal dependency of the event with multiple covariates.

This thesis argues that Bayesian network integrated with "Inverse probability censoring weights" could handle the censored data and provide the realistic risk predictions for the patients and answer the relationship between the variables.

1.3 Motivation

The principal challenge which we need to discover here is the pattern in the data for the length or the duration of time when the event will occur. For that, we have survival analysis which looks for the data distribution that measures the duration of time till the occurrence of an event. Consequently, the basic objective of these longitudinal studies is to find out the probability of a certain event to happen in a specific future time. Still, it could not answer "how to predict If a patient/subject will face a certain event by the end of observation period/study time having information of occurrence of event at the early stage of the data.

This problem manifests two major challenges:

1) Incomplete information about the occurrence of events or also called as censored data.

2) Information with few subjects have experienced the events at initial period of study.

Let us discuss below real-world applications which lead us to this thesis [3]:

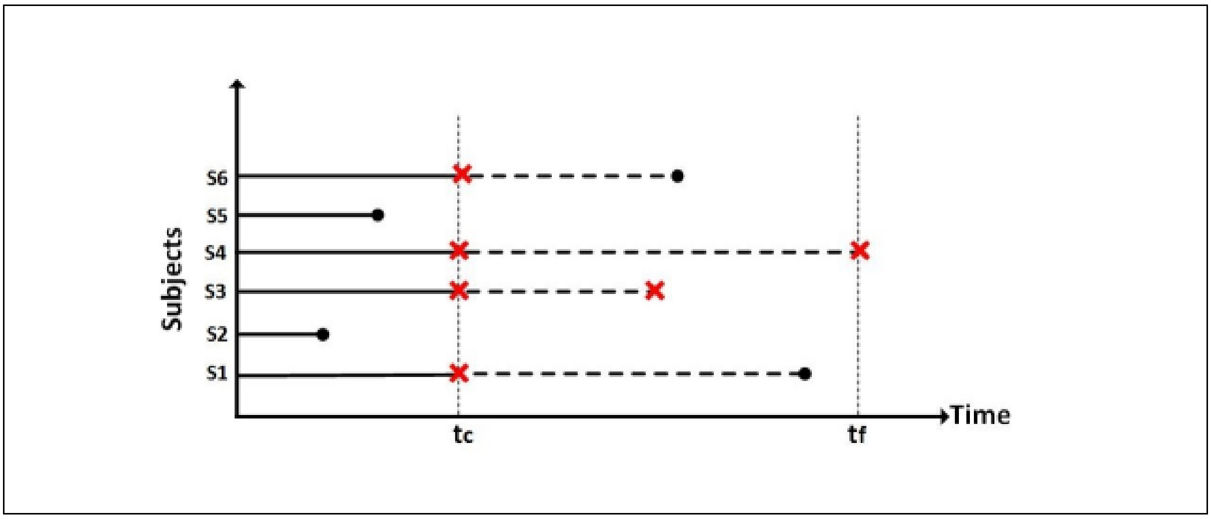
a. In the healthcare domain, when we would like to study the effect of a new treatment option on a certain group of patients in order to understand the efficacy of the treatment. The patients have been monitored for particular period of time and in this case, event refers to patients needed to be hospitalized. So, the effectiveness of treatment must be estimated as early as possible.

b. Prediction of newly launched product's reliability is very common scenario in industry where event corresponds to time taken for a device to fail. If such models can be learned using information from a very few device failures, the early alerts of failure can be given about the failures in near future.

c. In credit score modelling, it is important to estimate whether a customer will default and when this is going to happen. If a model can be developed based on few default cases only the better precaution measures can be taken against those who are more likely to default in the future.

All the above scenarios need such models which can predict beyond time and with very few trainings set. Therefore, this dissertation aims to build a method which uses very limited and incomplete information while learning but predict the events to occur beyond time accurately.

For a better understanding of the complexities and concerns related to this problem, let us consider an example of a time dependent study on below 6 subjects and the information for event occurrence until time tc is recorded. Subjects 2 and 5 had only experienced the event till tc. Our thesis aim to estimate the event occurrence by time tf with the given information.



[3] Fig 1.1: Sample to show the event forecasting at time tf having information till tc

The above example quite emphasises to develop algorithms that can predicts the occurrence of the events with the data at time tc where only few events had occurred. This is the case of longitudinal studies because the only way to have complete and trustworthy data is to wait for that period till we get the information about all the event's occurrences.

There are various ML (Machine Learning) techniques like classification, semi-supervised learning, transfer learning, imbalance learning etc are not sufficient and fit enough to handle such scenarios because the training dataset is incomplete and is only available till tc. Moreover, there are some advanced statistical methods like survival analysis which can be able to find the probability of the survival, but they couldn't predict the event for a time later that study time / observation time. It is also because survival model is only valid for the given observation time and not beyond that. It is very important to understand here that this scenario is not "time series forecasting "because in this problem we need to find the probability of event occurrence for each observation /object for future time beyond the observation time whereas in time series analysis models it predicts the next time step value. Additionally, these survival data are censored and has incomplete information of events even in the observation time which makes it further complex and difficult for any conventional machine learning algorithms to model such data. Moreover, these censored data cannot be ignored as it will create a substandard and insignificant model and if censoring time would be considered as event time it again leads to compromised model with actual performance.

1.4 Thesis Contribution & Overview

An extended technique to analyse longitudinal survival study and a general-purpose method for mining right-censored time-to-event data. Specifically, introducing a pre-processing step to modify the data by assigning weights to each observation using inverse probability of censoring weights (IPCW). This weighted data would be analysed considering the weightage of each subject and modelled with any machine learning algorithms which can integrate these observation weights. In the process, zero weights are assigned to the subjects where event status is unknown and the subjects with a known event status are given weights to shadow those subjects who were censored early and would have had the same event time. There are some subjects having larger event times are given higher weights because there is high probability for them to be censored prior to facing the event. This paper is discussing the mathematical proofs and relevant functions to support the above concept in later section.

There are some research on the clinical censored data analysis which used IPCW in different machine learning methods. We are using this approach and integrating it to learn Bayesian Network with various methods and analysing the conditional probability between the different factors affecting the event. Bandyopadhyay et al. [4] , IvanŠtajduhara et al [5] discusses the similar approaches and generalizes the IPCW technique to handle censored survival data.

IPCW is one of the efficient methods to calculate weights and assign for censoring and later integrated with various machine learning techniques like in classifiers, class probability estimation, sampling, or generation of similar kind of censored data.

PROPOSED BAYESIAN APPROACH

As discussed in research question, predicting events in censored survival data is a complex to analyse. It can be categorised as the conventional machine learning problem like classification and standard regression problems or a time series forecasting problem as labels for the data is provided in these cases. Also, statistical methods like survival analysis are there but it has its own limitation with ability to analyse the problem within observation time and could not predict beyond the observation time. Thus, for these longitudinal studies training data must be obtained only by waiting for the occurrence of sufficient number of events.

Therefore, in this paper, we are discussing a framework which predicts the event to occur in future for the subject even with partial information on the event for few subjects. We will discuss all the methods and its results in details in the later section. Before that, the similar work in the area of censored survival data and its analysis, and mechanism to handle it will be presented in next section.

The rest of the paper is organised as follows. In Section 2, we describe the data , its characteristics and find the key insights of it. Next section 3 discusses in detail about survival analysis and the key parameters of it. Also introduces survival analysis of censored data. Section 4 and 5 explains about modelling aspects and the designing and learning of Bayesian Network. In section 6, we will give the details of implementation and methodology. how IPCW has been implemented and integrated with BN. Also, we will evaluate our model using various performance metrics and discuss about the outcome of conditional probability table. Section 7 concludes our thesis and future scope of it.

1.5 RELATED LITERATURE

There have been wide range of methods and techniques put forward to handle right censored, time-to-event data [7]. Such data are also called censored data as the dependent variable is subject to censoring like failure, death, admission to hospital, emergence of disease etc. [8]. There are prior work done on similar approach where it has been proposed how to tackle such event status which are right censored with no information about the event status. Few have proposed some steps to impute the missing data or exclude from the study. In some research, they assert to adapt some specific machine learning techniques to censored data[adaptive]. OLS (Ordinary Least Squares) method has always been used to solve regression problem with the fact of minimizing sum of square error, but it does not work in case of censored data because when data records are censored it is almost impossible to find the error between actual response and predicted response which is coming from regression model [9]. Even so, the likelihood method which estimates the probability can be useful in solving the censored regression problem when we don't have information about the censored observation [10].

There are various statistical methods are also there which can handle survival data but doesn't work well when required to predict on the future time points. Since, there must be few subjects who will survive by the end of the observation period or study and do not experience the event or failure.It means some subjects have certainly survived beyond the time t, observation time in the study. Therefore, we can say in that study of survival, two-time classifications are involved. One, the subject experiences the event /failure during the observation time and has the death or failure within time t. Second, the subject survives till the time point t before the end of the study and then lost to further follow up or left the study group or the observation period/study is completed before the subject fails. So, this complex and incomplete observation of the event or failure time t is considered as "censored observation".

To handle such censored scenarios and overcome the problem of such data, various techniques have been proposed which mainly focuses on Maximum Likelihood Estimation (MLE) [11, 12]. Mostly, different machine learning methods are suggested to adapt survival data [13]. However, as described above such longitudinal study is not possible to be modelled only by conventional regression or classification approaches as it has mixture of information about the event status for each observation. Also, there is no specific pattern of missing information. For a particular period of study time few subjects have the event status and rest have no information. It might look like that the censored observations in survival data are similar as unlabelled data in classification of unsupervised learning or unknown data point in regression assuming event status is not known for some subjects. However, it is not correct analogy as time frame and observation period is not available in these cases. For accurate predictions of such censored data, it is required to have a proper machine learning technique which can handles this complexity properly. Moreover, in survival analysis for censored data the information is present up to a certain time point before censoring happens. To achieve better results from the model this partial information should also be passed to the model. Therefore, the conventional semi-supervised ML techniques [14, 15] are not applicable in this scenario.

[16] discussed the complexity of such analysis in their work. [17] These analyses are complex in nature because of uncooperative behaviour of the subjects under study who leaves the study group and refuses to remain in the study either till the end of the observation period or till when they experience the event. Also, they might have experienced the event/death, but we do not know about it as we have lost the connection with them in midway of the study. Important point is we do not want to ignore such observations because they have some or limited information about the survival too which is an important factor. Although , [18] for these subjects we have partial information, we know that the if event did not occur , it will occur sometime after the date of last follow-up. Even though, we assume they definitely survived beyond a certain time point but we cannot say exactly the date of the event. There is one more scenario which makes the study difficult is that we have few subjects that enters into the study after a significant time has elapsed. Hence, we will have a shorter observation period for such subjects, and they may or may not experience the death/event in that period of stipulated time. Still, we cannot exclude these subjects from the study because in real world scenario chances of such cases will be large and avoiding these observations will make our sample size too small to come to any conclusion. One such method to handle such

impartial information is The Kaplan-Meier estimate. It provides a computation method for survival over time with the challenges in the censored data.

The Kaplan-Meier provides us a survival curve that defines the probability of surviving in a specific period of observation while considering time as small intervals of period [17]. It uses three assumptions while doing calculations. First as, at any time point, subjects who are censored would have the same survival likelihood compared to those who are continuing in the observation period.

Second, the probability of surviving is same for patients joined early or late in the study. Third, events will happen at the specific time [19].

Inverse Probability Censoring Weights

The techniques in Survival analysis have been grown and extended to incorporate different kind of data which can be analysed. Even though, various methods have been developed to analyse data in case of informative censoring, still for censoring having causal dependency or dependent censoring, the Inverse Probability of Censoring Weighted (IPCW) estimator has only been developed.(Robins and Finkelstein 2000[65]) found out that Kaplan – Meier estimator though effective in censored survival data analysis but in case where covariates are associated with both lifetime and censoring mechanism gives biased results. They have explored the applicability of inverse probability censoring weights and explained how it is effective in removing the bias. [Rotnitzky and Robins2004][66] used AIPW, Inverse probability weighted augmented in survival analysis and handled the analytical challenge created by such high dimensional data in this domain. They generalized and made IPCW more efficient. Bang and Tsiatis 2000[68] have introduced a class of weighted estimators for censored data and found that this weight estimators are consistent and asymptotically normal with nominal variances. They studied the efficiency of the method for estimation of medical cost which is a challenge where follow up data is mostly incomplete. Their results shows that even estimation of mean medical cost is plausible with weight estimator technique.

There are some prior works where they have used inverse probability of censoring weighting (IPCW) in machine learning techniques. Authors like [4] have used IPCW with Bayesian network modelling to handle right censored survival data whereas [6] used it as more general-purpose technique and proposed to use it with multiple machine learning methods. This technique checks the dependent censoring in censored data especially in right censored data by assigning additional weight to subjects for whom we have event information and are not censored. Using IPCW weights the survival function can be estimated even in the absence of censoring. IPCW can be applied in survey analysis, when a survey sample is representing the entire population and only few individuals with rare characteristics will be included. It creates a problem when only few rare cases are included in the study. It can be resolved by sampling probability for these individuals which may include few more in the sample. However, the survey sample is no longer representative of the entire population. Therefore, in such cases to estimate the parameters for entire population, the sample subjects need to be weighted based on their inverse sampling probability.

The author from [5] have suggested a method for handling censored survival data by dividing it into three groups, primarily proposed by Zupan et.al[45]. One, as those instances for which event occurred at any time labelled positive, other where censoring happens after a certain critical point T are considered negative, or event free and third, observations censored before point T are split into both possible outcomes and hence get doubled. Then an estimated probability outcome will be assigned using Kaplan-Meier method. Each observation which are doubled gets a new weight based on its observation time. Sum of weights for both positive and negative part of the instance will be 1. This method includes all the instances even if they were censored much before the observation time. After that, data pre-processing and weight assignment, the modified set of observation will be used to for learning Bayesian network structure.

There are few authors other tha Zupan[45] and Bandyopadhyay et al. [4] , who applied the Bayesian network for censored survival data . One, Sierra and Larranaga [25] studied the implementation of genetic algorithms to learning BNs from data. Marshall et al. [47], tried Dynamic Bayesian Network to handle time variable in censored survival data. Along with it , they used latent Markov Model and by combining these two the have handled both causal representation and survival events. Also, [1] properly describes the impact of censoring on learning parameters of Bayesian network and suggested the potential usefulness of the learnt Bayesian network for predicting the probability of the event.

Critiques on Bayesian network Application

Also , there are some section of researchers who considers neural networks to handle censored survival data with an assumption that the possible censoring and event times are few in number [21, 22]. Additionally, many of them several have pondered to adapt different machine learning technique like SVM(support vector machines) for the outcomes by changing the loss function to require for censoring [23, 24]. These methods can work but reduces the generalizability of the approach to handle right censored survival data results. Many authors have adapted other technique using tree-based models like Decision Tree and Random Forrest to censored outcomes by modifying the splitting criterion to accommodate the changes. They choose the method of splitting which maximizes the log-rank statistics which compares the difference in the survival curves between two different groups in lieu of splitting the data to minimize the node impurity. However, the application of Bayesian Network is fundamentally different from the recursive partitioning approach of decision therefore, using splitting criteria to handle censoring does not apply to the other approaches. [Kraisangka, Jidapa, and Marek J Druzdzel [48] discusses on alternative approach for Bayesian Network for survival analysis and suggests a different mechanism altogether to handle censored survival data. This paper argues on Bayesian Network application alone is not sufficient and asserts building BN based purely on expert knowledge which can be time consuming and costly. It suggests Cox Proportional hazards (CPH) model is most used method in survival analysis and it explores the dependency between covariates and hazard in the same way as multiple linear regression. Also, CPH provides a parameter, hazard ratio to measure the impact of risk factor. It proposes Bayesian Network interpretation of CPH , (BN-Cox) to handle censored survival data.

On the other hand, there are various other ad hoc techniques for handling censored observations where status of events is unknown. It includes 1) discarding those observations [25, 26], 2) treating them as non-events [27, 28], or 3) repeating those observations twice in the dataset, one as experiencing the event and one event-free. In all such cases of subject’s, each observation is allocated with a newly generated weights depending on the marginal probability of facing the event between the event time and censor time [5]. The above assumptions and simple techniques increase the chances of bias in the estimation of the risk. On discarding observations with unknown events or considering them as non-events will definitely overestimate or underestimate the risk [27, 28]. Though the third approach is more sophisticated, still these weights weaken the relationship between the features and the target and hence resulted in poor calibration of risk estimates.

Referring the discussion above, there have been multiple proposal and critiques on application of Bayesian network for censored data. Since, our thesis is dealing with censored data in biomedicine where Bayesian Network has been used extensively and efficiently. [Lucas et al. 2004][13].

(Lucas et al. 2004)[13]. Zupan et al. (1999)[45] and Stajduhar and Dalbelo-Basic (2010)[5] have also worked on clinical data and proposed that the weight assignment technique where censored observations are repeated twice and weight allocation is based on marginal probability of experiencing an even between observation time and censoring time. This technique is actually intuitive but proven biased and inconsistent because the weight assignment depends on the marginal probability for each repeated observation and does not take covariates into the account [4]. In 2012 , [Stajduhar and Dalbelo-Basic 2012][49] have used the likelihood approach to impute event time , but again there is chances of inconsistencies and technique may perform poor if the parametric distribution of even time is incorrectly assumed. Other proposal of choosing other machine learning techniques like SVM or multiple adaptive regression splines (Therneau et al. 1990; Kattan et al. 1998;Kattan 2003) [50][51][52]are required to mine the data permit a continuous outcome and not very compliant with the approach considered here.[4] have applied the IPCW with Bayesian network and explained the model averaging strategy to control the complexity of the model and risk predictions outcome for small groups of patients and found the result as an improved classification model.

A Bayesian Network demonstrates the underlying probability distribution in the domain. It explains this distribution with its two attributes, a directed acyclic graph (DAG) and a set of conditional probability tables (CPTs). A BN has a qualitative part, Network structure (DAG) because it shows the relationship between the covariate with directed arcs in a way of causal dependency or cause-and-effect-like manner and provides the indication about which covariates are dependent to each other [2]. On the other hand, Conditional Probability Tables (CPTs) depicts the quantitative part of Bayesian Network as it represents with what probabilities these covariates are dependent to each other which are connected. This structure and probability represent as a knowledge graph and causality can be interpreted using the directions of the arc. The inference of the BN is probabilistic in nature, and it gives the output as a class probability which is useful for classifications. Because each node in the network is independent of its non-descendants, it helps in fast computation. From these discussions and reason, it can be considered that Bayesian Networks are best mechanism to showcase information graphs and for the requirement in clinical trials, it is an excellent instrument for knowledge representation in medicine. They represent causal influences and dependencies in covariate and also demonstrates these interactions in terms of numerical probability. These representations correlates to human reasoning with respect to causal dependency and prediction probability[2]. Above it, they can be learnt from data. That is the reason for BNs usefulness in the field of biomedicine and health care [20] for diagnosis, treatment, prognosis, and discovery of functional interactions. Our thesis deals with heart transplant data where BN is considered as most useful tool [29]. There are many methods where BNs learn from the data have been emerged where it efficiently handles the learning of parameter and structure from the complete data [30], [31], [32] and censored data (where data has missing instances) [33]. While learning parameter and the structure of the network, we can get new information about the causal dependence in that domain [2]. The structure, which is obtained from easily computed quantities, can then be further used to predict those parameters which are difficult to compute or measure.

Because of so many features which BN provides, it has become one of the best models in the predictive modelling literature and have been used predominantly for classification and successfully applied in other domains. As discussed in support, The authors Lucas et al. [20] and

Bandyopadhyay et al. [4] applied Bayesian network approach to model censored data where they allocated weight to censored instances in order to learn Bayesian networks from survival data using weight censoring technique. In this thesis, we will take this approach and try to handle the problem of censoring using Kaplan-Meier method [20] to estimate the probability of event and probability of censoring for each censored subject. Along with it, We will try to build Bayesian networks which accounts for right-censored event indicators using inverse probability of censoring weights (IPCW) [65][66] and the design we will follow is as suggested in [4].

Bayesian Network with EM algorithm

Since we are going to use BN with IPCW technique, an important aspect of BN model to work efficiently is its learning parameters. In Bayesian Network field literature, many techniques have been proposed to learn the parameters especially when data are missing. Two important methods for learning algorithms are Gibb's sampling [43] and Expectation Maximisation algorithm [41]. Gibbs sampling as per [55]is the basic method of simulation and can be used in any directed or undirected graphical model or even can be applied for both continuous and discreet variables.[44]. Though it's a strong intuitive method which completes the samples from the available information however it suffers from the convergence problems. Also, it is not efficient when data are missing completely random. The EM algorithm which can be considered as a deterministic type of Gibbs sampling which is used to find Maximum Likelihood estimates for model parameters [44]. Although, when there is large number of missing values or presence of multiple hidden variables, EM algorithm gets trapped in local maximum. Other than that, EM algorithms experiences problem if the initial starting points are quite away from the optimal solution and learned parameters are unpredictable [42].

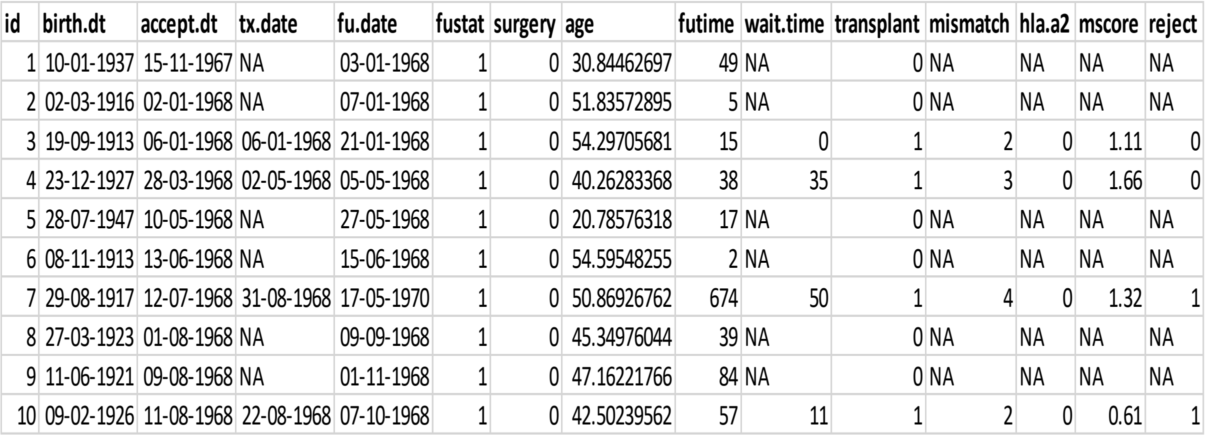
Chapter 2

Data & Data Analysis

2.1 Data

The dataset which we selected for our thesis is “Stanford heart transplant study, published by Clark et al. in the Annals of Internal Medicine in 1971[53]”. It has information related to the survival of the patients who were enrolled into the transplant program. The Stanford University Heart Transplant Study was conducted to find the results if a heart transplant program has increased the lifespan of the patients. Each patient entering the program was considered ill and would be most likely to get benefitted from the heart transplant.

A data consists of 103 observations with 8 variables. A snapshot of the data with 10 observations. The description of each variable are mentioned below.



id	birth.dt	accept.dt	tx.date	fu.date	fustat	surgery	age	futime	wait.time	transplant	mismatch	hla.a2	mscore	reject
1	10-01-1937	15-11-1967	NA	03-01-1968	1	0	30.84462697	49	NA	0	NA	NA	NA	NA
2	02-03-1916	02-01-1968	NA	07-01-1968	1	0	51.83572895	5	NA	0	NA	NA	NA	NA
3	19-09-1913	06-01-1968	06-01-1968	21-01-1968	1	0	54.29705681	15	0	1	2	0	1.11	0
4	23-12-1927	28-03-1968	02-05-1968	05-05-1968	1	0	40.26283368	38	35	1	3	0	1.66	0
5	28-07-1947	10-05-1968	NA	27-05-1968	1	0	20.78576318	17	NA	0	NA	NA	NA	NA
6	08-11-1913	13-06-1968	NA	15-06-1968	1	0	54.59548255	2	NA	0	NA	NA	NA	NA
7	29-08-1917	12-07-1968	31-08-1968	17-05-1970	1	0	50.86926762	674	50	1	4	0	1.32	1
8	27-03-1923	01-08-1968	NA	09-09-1968	1	0	45.34976044	39	NA	0	NA	NA	NA	NA
9	11-06-1921	09-08-1968	NA	01-11-1968	1	0	47.16221766	84	NA	0	NA	NA	NA	NA
10	09-02-1926	11-08-1968	22-08-1968	07-10-1968	1	0	42.50239562	57	11	1	2	0	0.61	1

Table 2.1 : 10 Records of Dataset used in thesis

The data is available in the “survival” package in R with data set named “jasa”

Survival of patients on the waiting list for the Stanford heart transplant program.

Id	Patient id
birth. dt:	birth date
accept. dt:	acceptance into program
tx. date:	transplant date
fu. date:	end of follow up
fustat:	dead or alive
surgery:	prior bypass surgery
age:	age (in years)
futime:	followup time
wait. time:	time before transplant
transplant:	transplant indicator
mismatch:	mismatch score
hla.a2:	particular type of mismatch

mscore:	another mismatch score
reject:	rejection occurred

2.2 Analysis of the Data

2.2.1 Time dependent Covariates

The variables or data sets used in survival analysis follows a special characteristic with some limitations and caveats. An important caveat is that the values of the covariates must be available at time t= 0, the time when the patient comes under the observation the study and remains there till the study completes. This problem comes with survival data because such data depends on time variable and evolve with time, and it would be improper to use the value a covariate to model survival information that is observed before the covariate's value is known. For these "time dependent covariates" to analyse , a proper mechanism is required to obtain valid parameters.
Below are some simple analysis of the data :

Results from the Data Analysis – R script

n= 103, number of events= 75

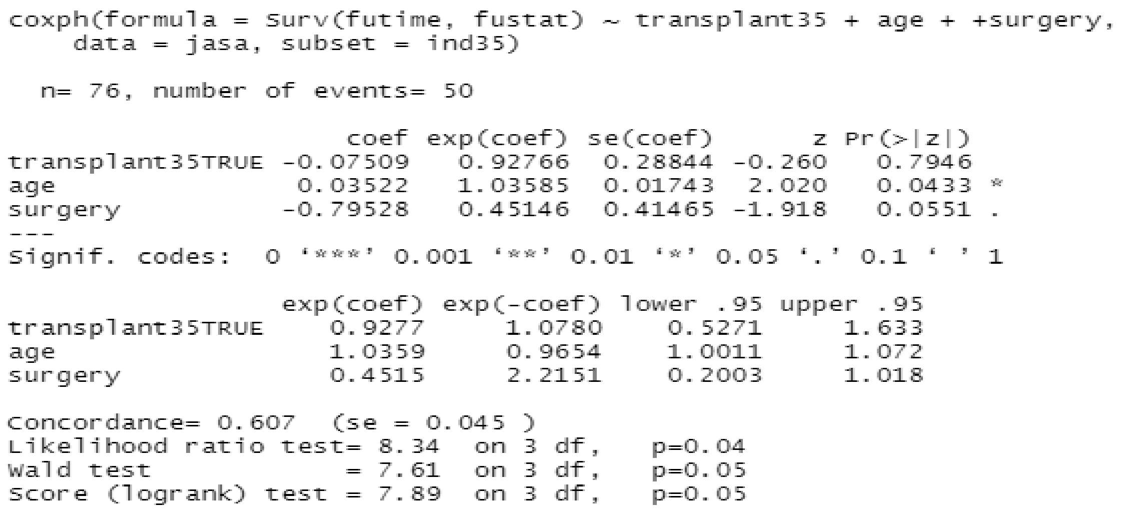
	coef	exp(coef)	se(coef)	z	Pr(>|z|)
transplant	-1.71711	0.17958	0.27853	-6.165	7.05e-10 ***
age	0.05889	1.06065	0.01505	3.913	9.12e-05 ***
surgery	-0.41902	0.65769	0.37118	-1.129	0.259

Table 2.2 : Summary of coefficients for key features in dataset

On initial analysis, we can infer that covariate "transplant" is one of the key variables and has value 1 for those who received a heart transplant and 0 for those who did not. The coefficient value for transplant is 1.717 which gives the estimate of the transplant coefficient, and the p-value is very small. This result can be considered as an indicator which shows that transplants are extremely effective in increasing the lifespan of the recipients based on the coefficients value. The problem in understanding could be that the transplant is a time dependent covariate; patients who received a transplant had to live long enough to receive that transplant. Primarily, the above analysis only infers that the patients who live longer to receive a transplant have longer lives than patients who did not live as long.
Another way to understand this by defining an indicator period/ time to split the patients into two groups. In this process, patients who receive the intervention prior to the indicator go into the one group and those who did not are placed in the comparison group. Key requirements of this approach are that
(a) patients who survive up to the indicator time are only included in the study, and
(b) all patients who are in the comparison group will keep in their original group irrespective to what happens in the future after the indicator time.
Consider, we set the indicator at 35 days. We would select the number of patients who lived at least 35 days. It comes out as 76 out of 103 patients. Refer below R output.

Again, out of these 76 patients, 32 had a transplant within 35 days, and 44 did not. Of these 44, 30 subsequently had a heart transplant, but we will consider them in the "no transplant within 35 days" group.

Results from R : Data Analysis – R script (Appendix C)



```
coxph(formula = Surv(futime, fustat) ~ transplant35 + age + +surgery,
    data = jasa, subset = ind35)

  n= 76, number of events= 50

                      coef exp(coef) se(coef)      z Pr(>|z|)
transplant35TRUE -0.07509   0.92766  0.28844 -0.260   0.7946
age               0.03522   1.03585  0.01743  2.020   0.0433 *
surgery          -0.79528   0.45146  0.41465 -1.918   0.0551 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
transplant35TRUE    0.9277     1.0780    0.5271     1.633
age                 1.0359     0.9654    1.0011     1.072
surgery             0.4515     2.2151    0.2003     1.018

Concordance= 0.607  (se = 0.045 )
Likelihood ratio test= 8.34  on 3 df,   p=0.04
Wald test            = 7.61  on 3 df,   p=0.05
Score (logrank) test = 7.89  on 3 df,   p=0.05
```

Table 2.3 : Summary result of survival fitted model using indicator time

The coefficient for transplant35 comes out as 0:075 with the p-value as 0.79, which is not at all statistically significant. This indicator method has some new information which indicates that there is little or no difference in survival between those who got a transplant and those who did not. This above-mentioned method has mostly discarded more than 25% of the patients from the analysis. Also, there is no scientific approach towards keeping the indicator time as 35 days. We need to have a algorithm which models the "transplant" as time dependent variable. This can be achieved using Cox proportional hazards framework model. We can pick few patients with all features and model the data based on "transplant". The output is :

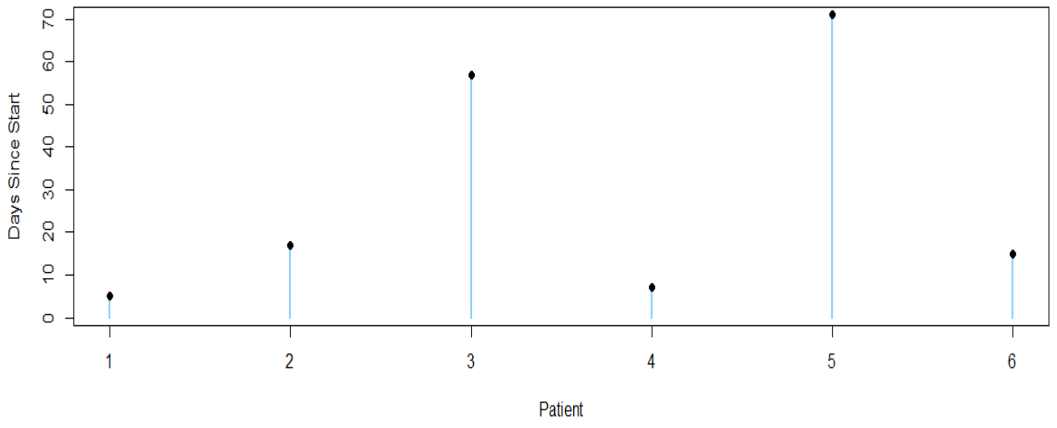


Fig 2.1 : Sample records of patients showing days since start when censoring occurred

This is the sample of six patients from the Stanford heart transplant data set. In the plot, the timelines of patients who received a transplant from the days since start.

2.2.2 Data Distributions

The data distribution of Age covariate

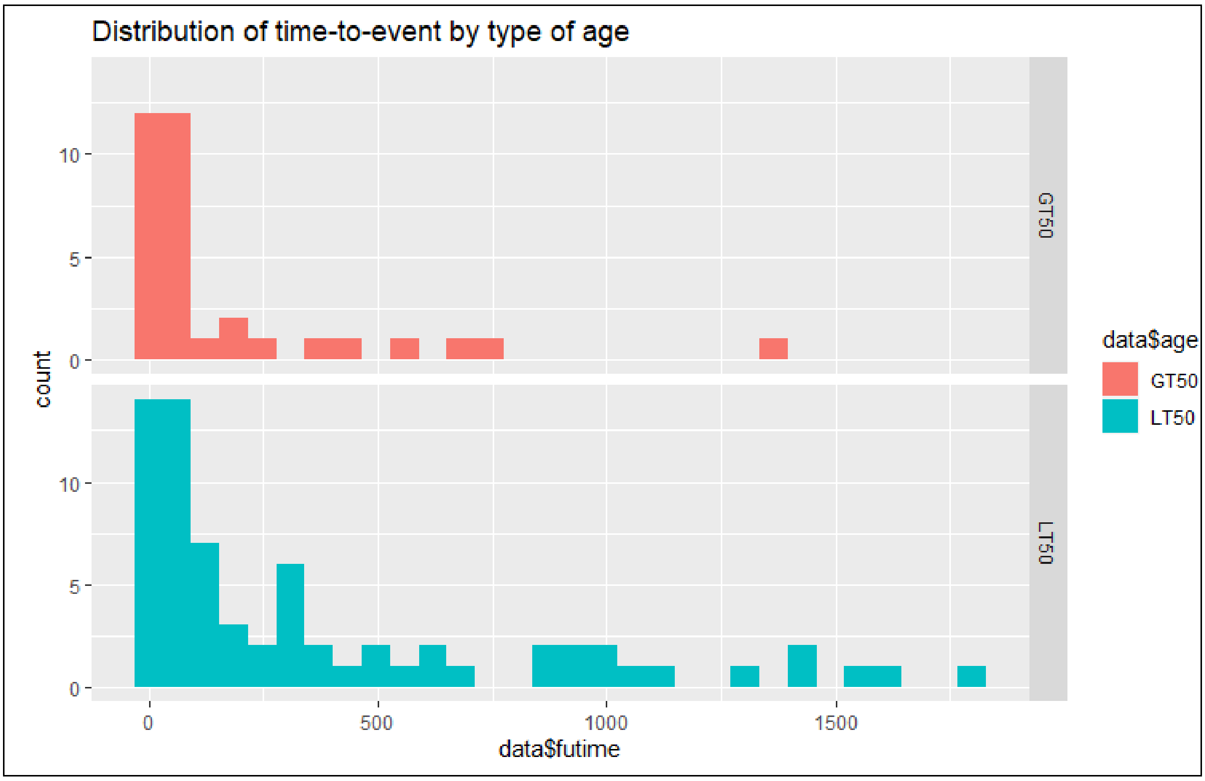


Fig 2.2 Distribution of Variable Age against follow-up time

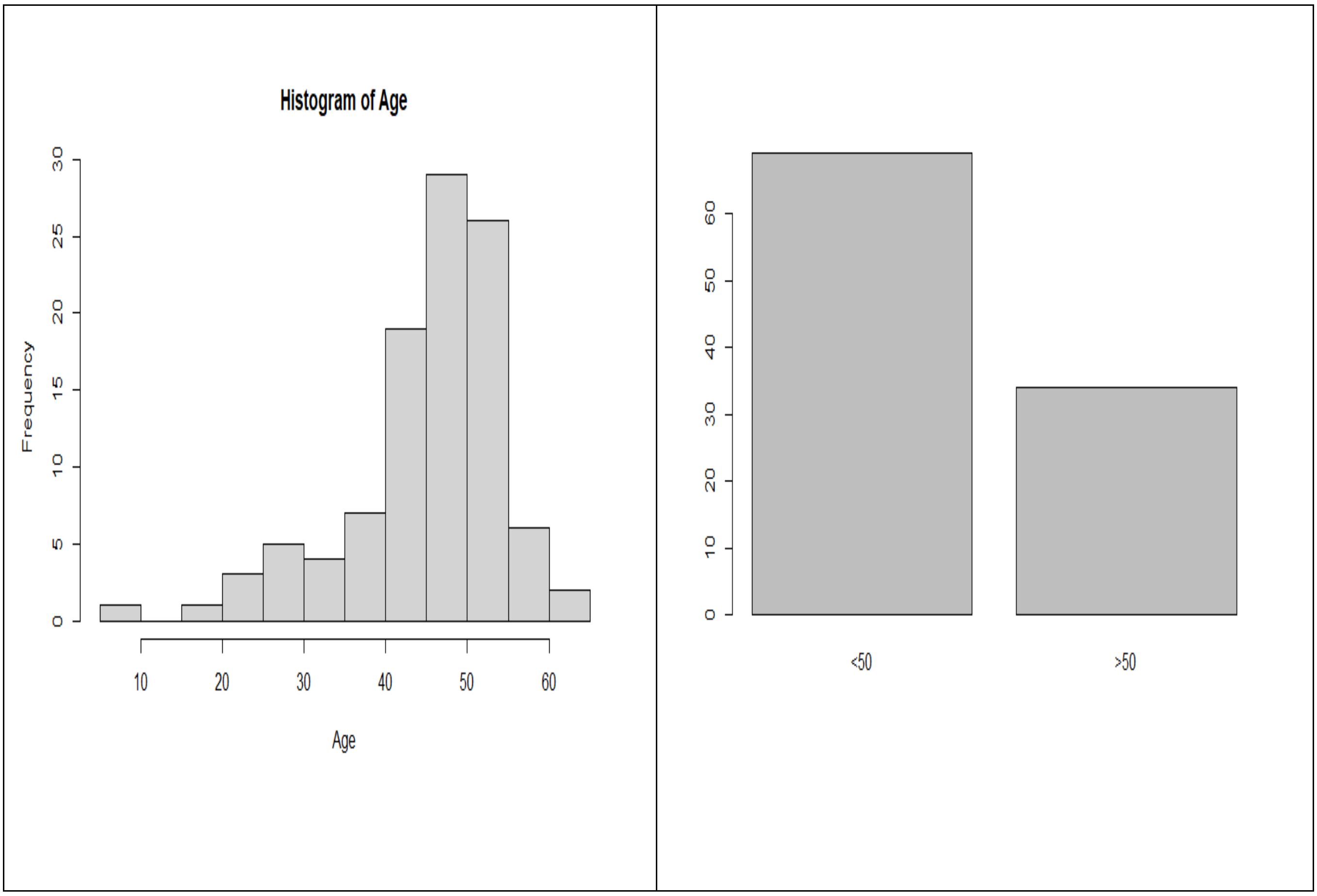


Fig 2.3 : Histogram plot of Age against no of patients

Age is a continuous variable, and it shows from the above plot that the dataset contains patient who are mostly above 40 and peak is between 45-50. We will try to divide the age into groups of two suppose > 50 and >50 , It will clearly picturize the patients between these 2 groups. Patients below 50 are more in numbers comparing to patients with age greater than 50.

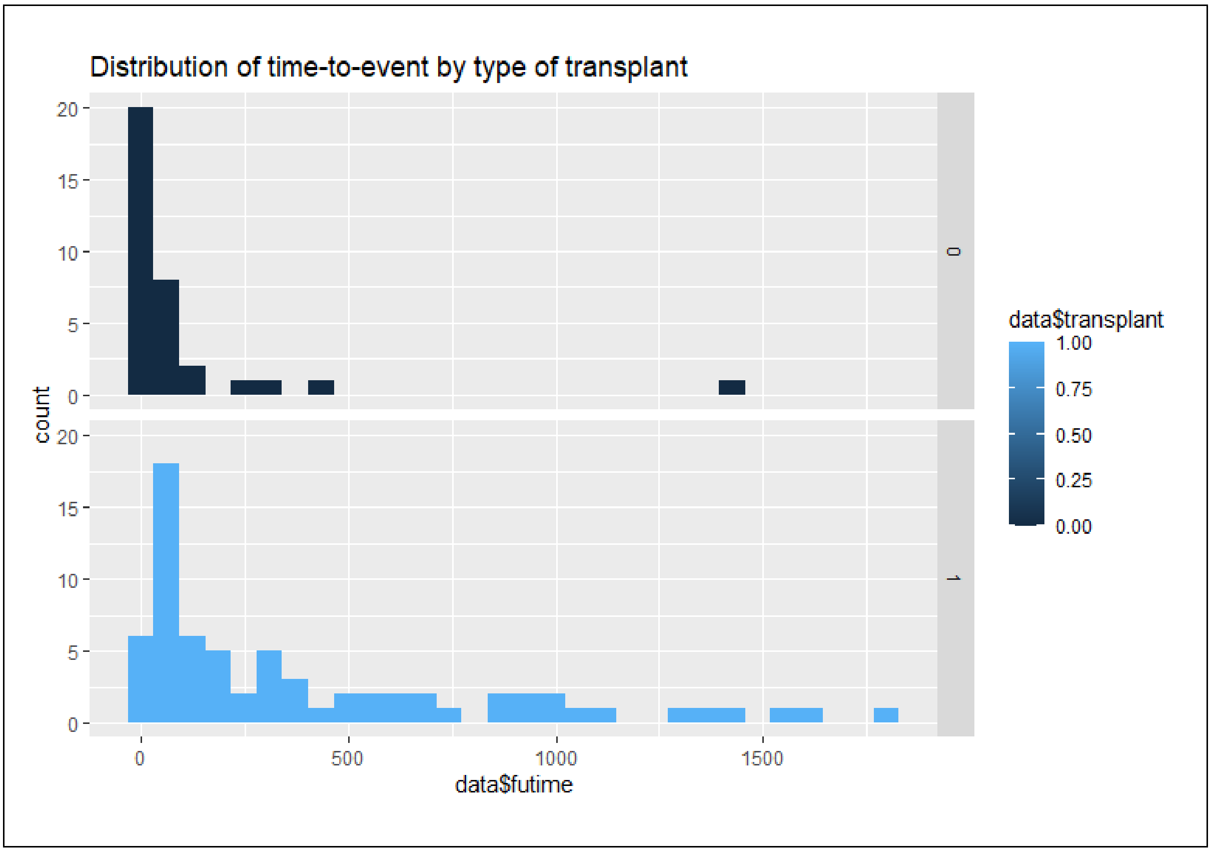


Fig 2.4 : Distribution of transplant indicator over follow up time

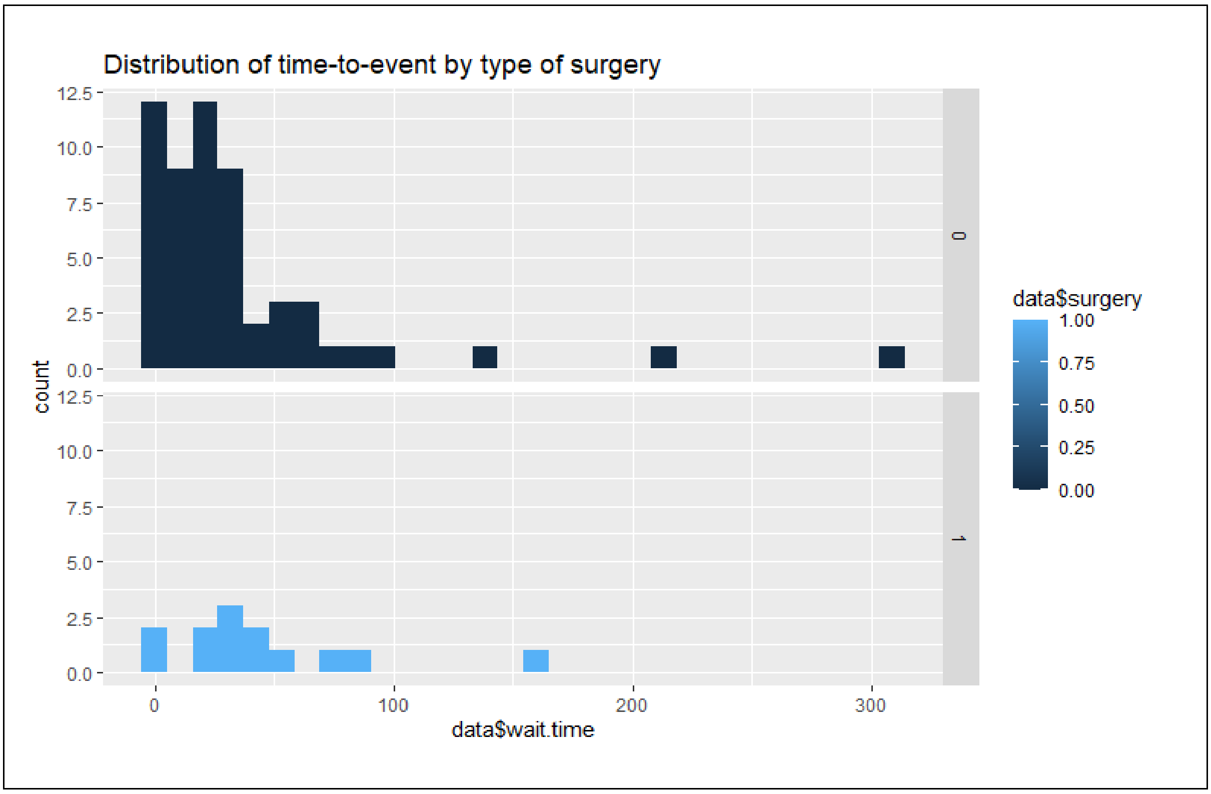


Fig 2.5 : Distribution of prior surgery indicator over follow up time

Correlation Plot

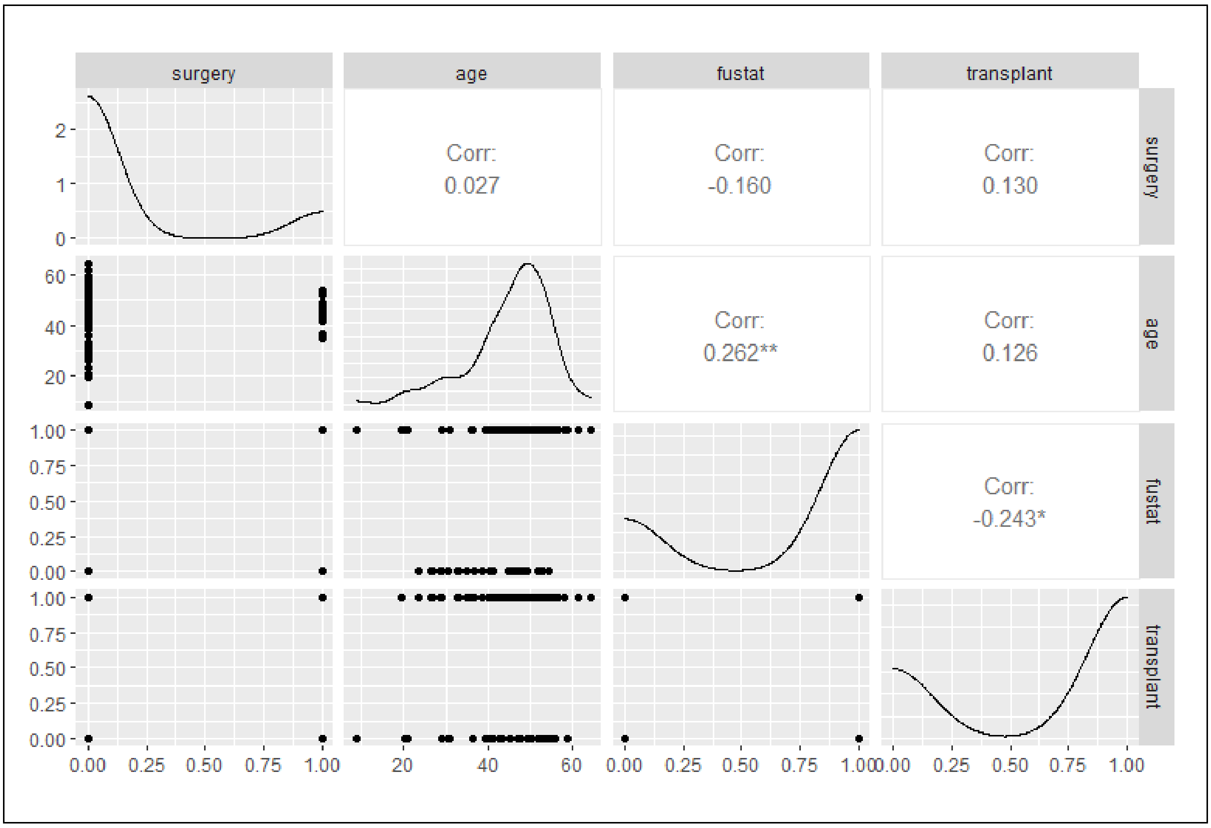


Fig 2.6 : Correlation plot between all covariates

Correlation: This is an interesting plot to discuss. As visible from the above plots , fustat is highly correlated with that transplant , surgery and age. fustat is negative correlated with prior surgery, which means if there is any prior surgery of the patient chances of experiencing the event is more. Similarly with transplant its very intuitive. Percentage of patients alive are more when they received the transplant and can survive more. This information is very useful to create the structure of the Bayesian Network for the dependency.

2.2.3 Censored Data Analysis

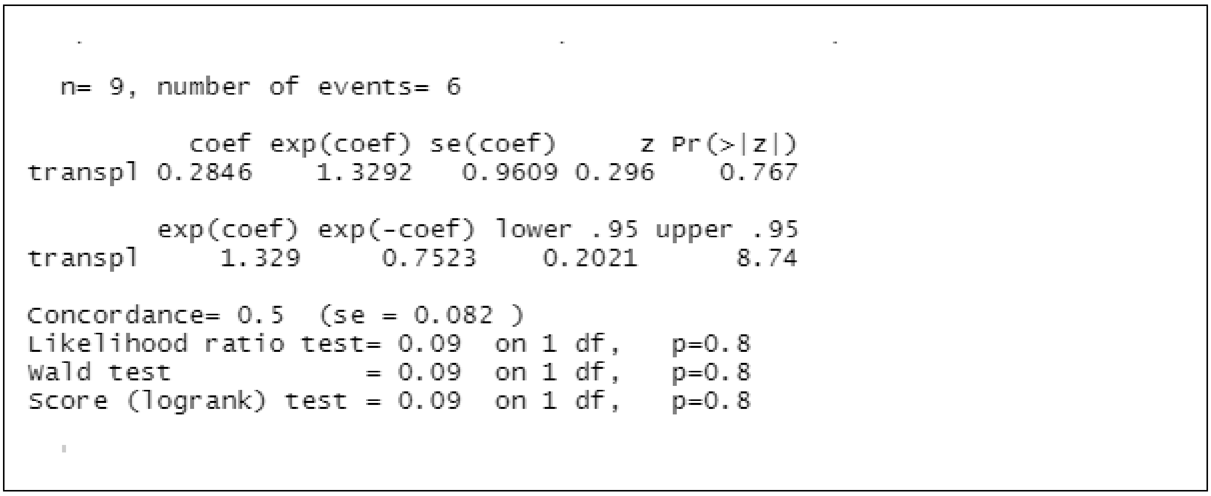
As we have seen the "coxph" function able to model the time dependent variables. However, our thesis deals with censored data where some information is partially available. Let's try to analyse it again by coxph but by first pre-process the data in the form of start-stop. The significance of this method will be based on partial likelihood theory [56]. Basically, this approach divides the time for patients as before transplant and after transplant. Consider any Patient #10 was a non-transplant patient from entry until day 11. Now , since that patient received a transplant at that time, the future for that patient, had he or she not received a transplant, is not known. Hence, we censor that portion of the patient's timeline at t = 11. Following the transplant, we start a new record for Patient #10. This second piece of the record is left-truncated at time t = 11, and a death is recorded at time t = 57. It is left-truncated because that patient's survival experience with the transplant starts at that point. For the first part of this patient's experience, the "start" time is 0, and the "stop" time is 11, which is recorded as a censored observation. For the second piece of that patient's experience, the start time is 11 and the stop time is 57. Thus, to put the data in start-stop format, the record of every patient with no transplant is carried forward as is, whereas the record of each patient who received a transplant is split into pre-transplant and post-transplant records.

This way we can analyse on small group of patients to check how coxph infers on censored data.

id	tstart	tstop	death	transpl
2	0	5	1	0
5	0	17	1	0
10	0	11	0	0
10	11	57	1	1
12	0	7	1	0
28	0	70	0	0
28	70	71	1	1
95	0	1	0	0
95	1	15	1	1

Table 2.4 : time split to show start stop

Coxph model on censored data created above



```
  n= 9, number of events= 6

           coef exp(coef) se(coef)     z Pr(>|z|)
transpl 0.2846    1.3292   0.9609 0.296    0.767

        exp(coef) exp(-coef) lower .95 upper .95
transpl     1.329     0.7523    0.2021      8.74

Concordance= 0.5  (se = 0.082 )
Likelihood ratio test= 0.09  on 1 df,   p=0.8
Wald test            = 0.09  on 1 df,   p=0.8
Score (logrank) test = 0.09  on 1 df,   p=0.8
```

Table 2.5 : Summary output after start stop splitting

2.3 Challenges in the heart transplant dataset

Missing Features:

As observed the dataset we have is censored survival. Such data have common pattern of unmeasured attributes for certain observations/patients .

One of the advantages of Bayesian networks is that we can still obtain predictions for subjects in the validation set with incomplete features; we can also use information on subjects in the training set to learn parameters in the Bayesian network without having to impute the missing covariate values[4].

Chapter 3

Survival Analysis

3.1 Survival Analysis

Survival analysis study the survival period and all the affecting variables. It includes time when the object enters into a clinical trial till the death, from the time of development and the progression of the disease. In other words, from the time of beginning till its survival or death. In some cases, event can also be considered as time of entry till how the tumor responds in a clinical trial.

The longitudinal survival studies involve the assessment of survival distributions, their comparison between different survival distributions and also analysis of the factors which may affect the survival times.

3.2 Principle of Survival Analysis[54]

Survival analysis algorithms is subjected to survival distribution, and most common ways to explain it are hazard and survival functions. The survival function defines the probability of surviving up to a point t. Mathematically it can be represented as ,

$$S(t) = pr(T > t) , 0 < t < \infty$$

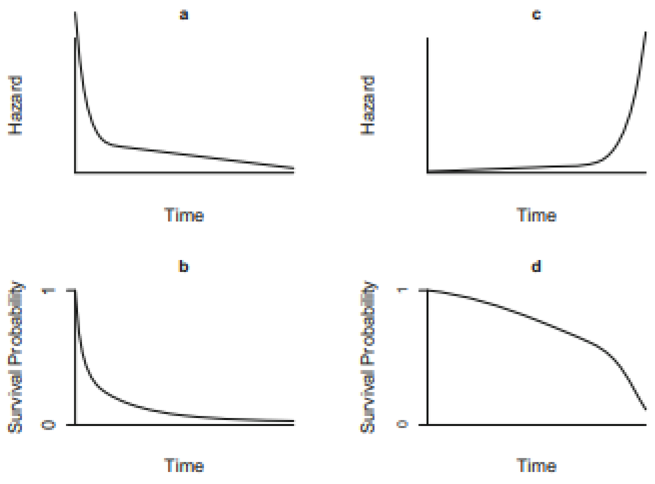
Survival function, S(t) takes the value 1, maximum at initial time point at t =0 and henceforth decreases or may remain constant over time but definitely never reduced to below 0. It is always right continuous.

The survival function can also be defined in terms of hazard function which can be interpreted as rate of failure, or the force of mortality and it is the age specific failure rate or instantaneous death rate. Hazard function, h(t) gives the probability that a subject which has survived up to time t, the object fails in the next small interval of time, divided by the length of that interval. In mathematical equation form, it can be expressed as below:

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

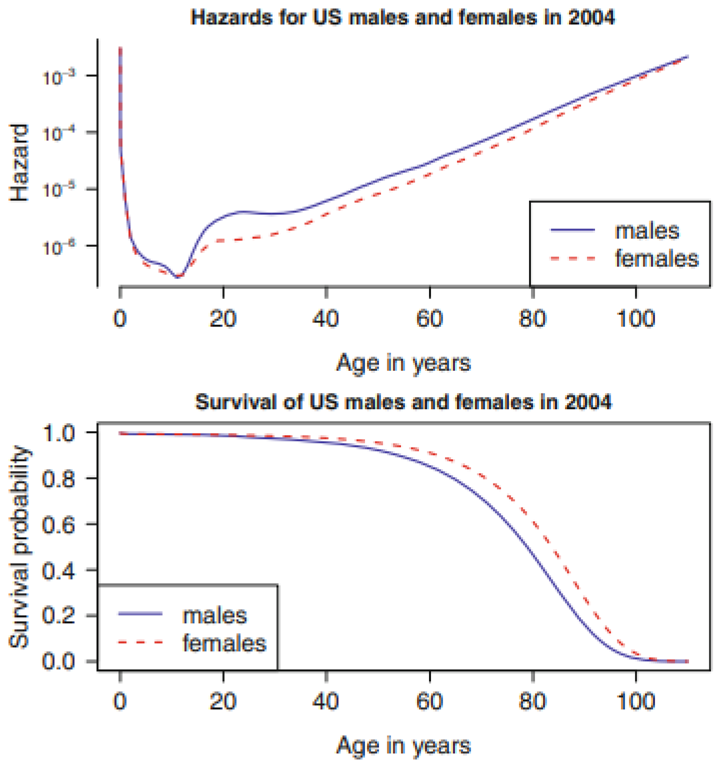
and it tells the intensity function or failure rate.

These two functions are the methods of specifying a survival distribution. Below Charts explains the relationship between these two functions. In first case, hazard is initially very high. These suggests the example of lifetime of objects which has high mortality in early stage of life.



[54]Fig 3.1 Hazard and Survival Functions Demonstration

(a)displays the hazard and corresponding survival plot is shown by (b). In second case (c), where hazard is low initially and increases subsequently later in life, and corresponding survival in (d).



[54]Fig 3.2 Example of Hazard and Survival of US males & females in 2004:

3.3 Different Representations of Survival Distributions

Other than survival and hazard functions, it can be expressed in terms of cumulative hazard functions:

$$H(t) \ = \ \int_0^t h(u)du$$

Mathematically, survival function can also be expressed in terms of hazard function as

$$S(t) \ = \ exp(-\int_0^t h(u)du)$$

This representation of the survival function shows that cumulative hazard function is key entity in survival analysis. Simply, we say:

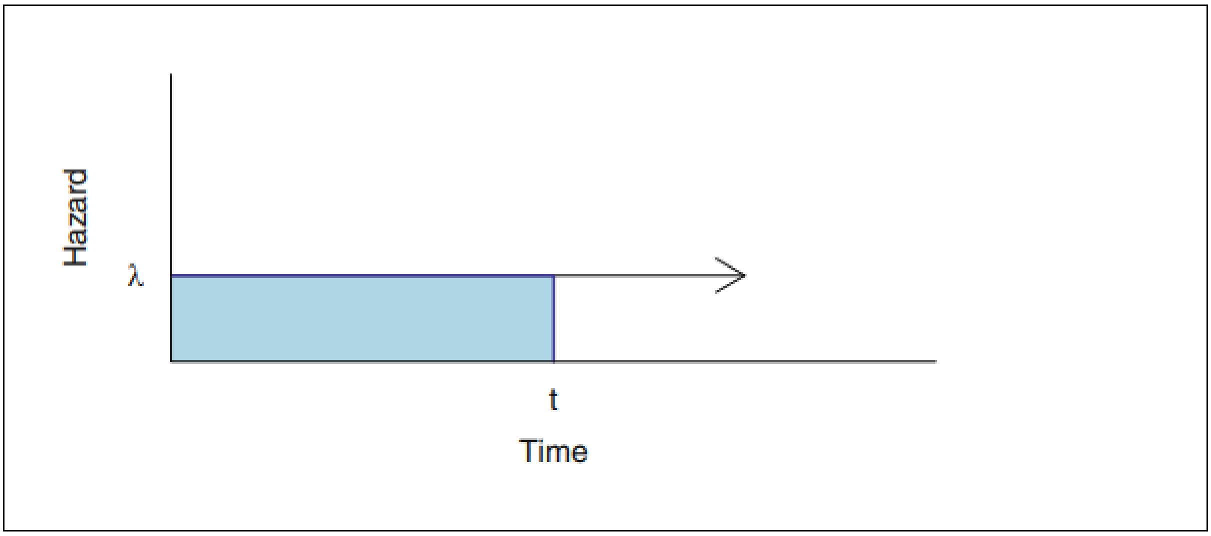
$$= exp(-H(t))$$

3.3.1 Parametric Survival Distributions

There are other survival distributions also available to model the survival data, one of it is exponential distribution. It has a constant hazard, h(t) = ƛ. Using the above equations, we can derive the relationship with cumulative hazard function as :

$$H(t) \ = \ \int_0^t h(u)du = \int_0^t ƛdu = ƛt|_0^t$$

We can clearly infer it as the cumulative hazard at time t is the area ƛ.t of the shaded rectangle.



[54]Fig 3.3 Hazard function distribution

The exponential distribution is easy to work with, but the constant hazard assumption not practical for describing the lifetimes in real world.

The Weibull distribution provides more flexibility as compared to exponential distribution in modelling survival data analysis. Its hazard function can be defined as

$$h(t) \ = αƛ(ƛt)^{α-1} = αƛ^{α}t^{α-1}$$

Similarly , Survival and cumulative hazard functions for the Weibull distribution are , respectively :

$$H(t) \ = (\lambda t)^{\alpha}$$

And,

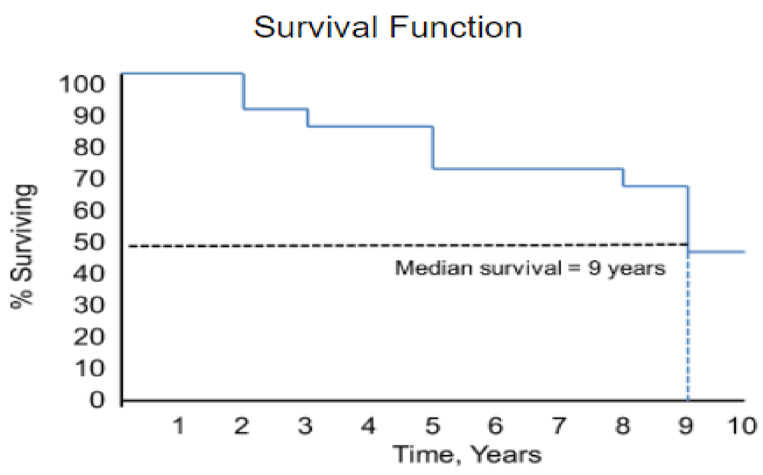
$$S(t) \ = exp[-(\lambda t)^{\alpha} \,]$$

3.3.2 Nonparametric Distribution of the Survival Function

There are various types of parametric model available to handle survival data using hazard functions. However, when dealing with clinical trials or human survival scenarios it is very difficult to choose which parametric model will be used because they don't provide enough flexibility or approximates accurately with actual shape of survival distribution. Hence, in such cases, we opt for non-parametric method. The most widely used of these is the product limit estimator, also known as the Kaplan-Meier estimator.

3.4 Estimating the Survival Function : Calculation

As discussed in above section, there are several different functions to project a survival curve. There are parametric methods like exponential, Weibull, Gompertz and log-normal distributions though differs in assumptions that are made about the distribution of survival times in the population. Among them, the most popular is the exponential distribution, which assumes that a participant's likelihood of suffering the event of interest is independent of how long that person has been event-free[3]. Other distributions make different assumptions about the probability of an individual developing an event i.e., it may increase, decrease or change over time.
However , as we understood non parametric methods which doesn't assume about how the probability that a person develops the event changes with time , are quite flexible and more useful while handling human clinical data. Using nonparametric methods, we estimate and plot the survival distribution or the survival curve. Mostly , the survival curves represented as step functions where X -axis is Time and survival probability on Y.[55]



[55]Fig 3.4 : Survival Probability with time period

Survival probability on Y axis doesn't only shows the percentage of people who are surviving but sometime also represent the people who are free of another event / death / disease or can also shows who do not experienced a event.
But the most complex part of survival analysis comes with censoring where we do not have observation of events for every subject. While performing our analysis, some may be still surviving.
Example: Let's consider a small example to understand how the survival probability is going to be calculated. Suppose there is a small study designed to study the time to death. It involves 20 participants who are 65 years of age or older and they are enrolled over a 5-year period and are followed for up to 24 years until they die, the study ends, or they drop out of the study (lost to follow-up.

For simplicity, We have aggregated the data against the time intervals for easy computation. [estimation of survival]

Interval in Years	Number Alive at Beginning of Interval	Number of Deaths During Interval	Number Censored
0-4	20	2	1
5 to 9	17	1	2
10 to 14	14	1	4
15 to 19	9	1	3
20 to 24	5	1	4

[55] Table 3.1 : Survival Probability estimation

Below notations are used for computation:

Nt = count of subjects who are event free and at potential risk during interval t
Dt = count of subjects who experience the event during interval t
Ct = count of subjects who are censored in interval t
Nt* = the average count of subjects at risk during interval t
Nt* = the average count of subjects at risk during interval t
Nt* =Nt-Ct/2 , qt = proportion facing event during interval t, qt = Dt/Nt*
pt = proportion who are surviving during interval t, pt = 1-qt
St, cumulative survival probability and we can compute it as :
First, the probability of surviving at t=0 is 1.in simple terms,
the probability that a subject survives past interval 1 is S1 = p1 and past interval 2 means that they had to survive past interval 1 and through interval 2: S2 = P(survive past interval 2) = P(survive through interval 2)*P(survive past interval 1), or S2 = p2*S1. In general, St+1 = pt+1*St.

Based on above formula, we will compute the estimation of surviving for each interval[55]
Interval 0-4 : At time 0, there are 20 surviving or at risk. Two participants die in the interval and 1 is censored. We apply the correction for the number of participants censored during that interval to produce Nt* =Nt-Ct/2 = 20-(1/2) = 19.5. Stepwise calculation is present in the below table. The probability that a participant survives past 4 years, or past the first interval (using the upper limit of the interval to define the time) is S4 = p4 = 0.897.

Similarly, for the interval, 5-9: The number of patients or subjects at risk is the number at risk in the previous interval (0-4 years) less those who die and are censored (i.e., Nt = Nt-1-Dt-1-Ct-1 = 20-2-1 = 17). And same goes to other intervals.

Interval in Years	Number At Risk During Interval, N_t	Average Number At Risk During Interval, N_{t^*}	Number of Deaths During Interval, D_t	Lost to Follow-Up, C_t	Proportion Dying During Interval, q_t	Among Those at Risk, Proportion Surviving Interval, p_t	Survival Probability S_t
0-4	20	20-(1/2) = 19.5	2	1	2/19.5 = 0.103	1-0.103 = 0.897	1(0.897) = 0.897
05 to 9	17	17-(2/2) = 16.0	1	2	1/16 = 0.063	1-0.063 = 0.937	(0.897) (0.937)=0.840

[55]Table 3.2: Survival Probability estimation - calculation

3.5 Handling Censored Survival Data

The major difference between survival analysis and standard classification is censoring. If our event of interest like death is not observed for any particular observation, we say it censoring. In a scenario like ours of “Stanford heart transplant dataset”, it corresponds to the early withdrawal from the study group of a patient for any reason and we don’t have any information about it any further on the timeline. For example, censoring in a heart transplant study occurs for a patient who died in a car accident or moved to another country during the trial before the event could be observed [1].

Kleinbaum [56] gives three reasons for the occurrence of censoring: (1) a person does not experience the event before the end of the study; (2) a person is lost to follow-up during the study; (3) a person is withdrawn from the study because of death unrelated to the event observed. These kinds of censoring are often referred to as right censoring.

This is the primary reason behind unable to use supervised ML methods to predict the events and are not used for fitting models in clinical trials. Hence, we cannot clearly assert if the final outcome for a patient about the occurrence of the event or if the subject’s observation was censored. In that case, if we need to model the final outcome is of interest, it is quite difficult to learn the models from censored data. From conventional statistical modelling point of view, it can be considered as noise in the dependent variable. If such censored observations are comparatively low in dataset, we can go ahead with standard classification models and data can be handled normally without affecting the model’s performance however, if it is high in numbers , it will definitely affects the performance and leads to completely inaccurate modelling .

In this paper, we are treating all censored cases as event-free, irrespective of observation time. So, if the subject has left the study group early, or survived till the observation period would be considered Censored. Doing this enables us to assess the influence of censoring in the data on BN learning methods and how to learn correct BN structures and good classifiers. However, for the cases those censored early creates a problematic scenario as we have no clue for the event occurrence. Their probability of survival is close to the a priori probability of survival for the whole sample. This leads to bias of the model [Kattan et al. [51]]. Few have addressed this by learning models from different learning sets for several time intervals.

3.6 Survival Plots

Overall Survival Analysis

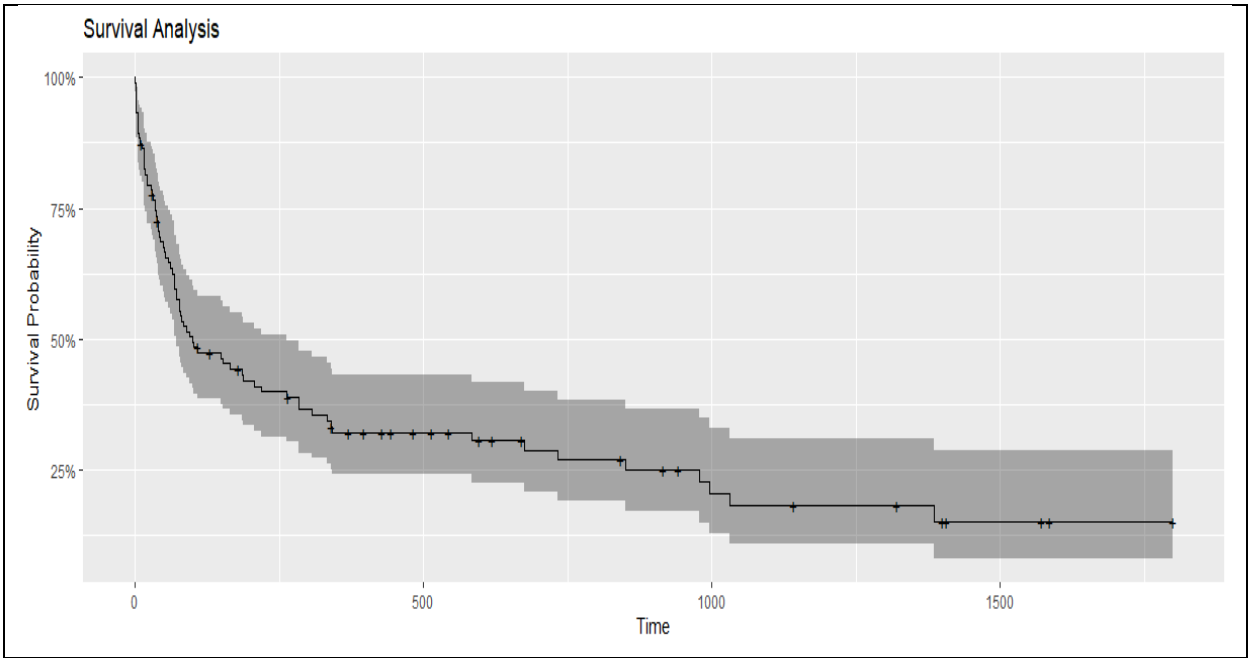


Fig 3.5 : Overall survival plot of the dataset

The plot gives the survival probability across the time. As can be inferred from the plot, the survival probability decreases with time. It decreases from 100% to nearly 25% within time interval of 500 unit. The plot gives the Kaplan Meier Survival curve in which the survival probability is plotted across the time.

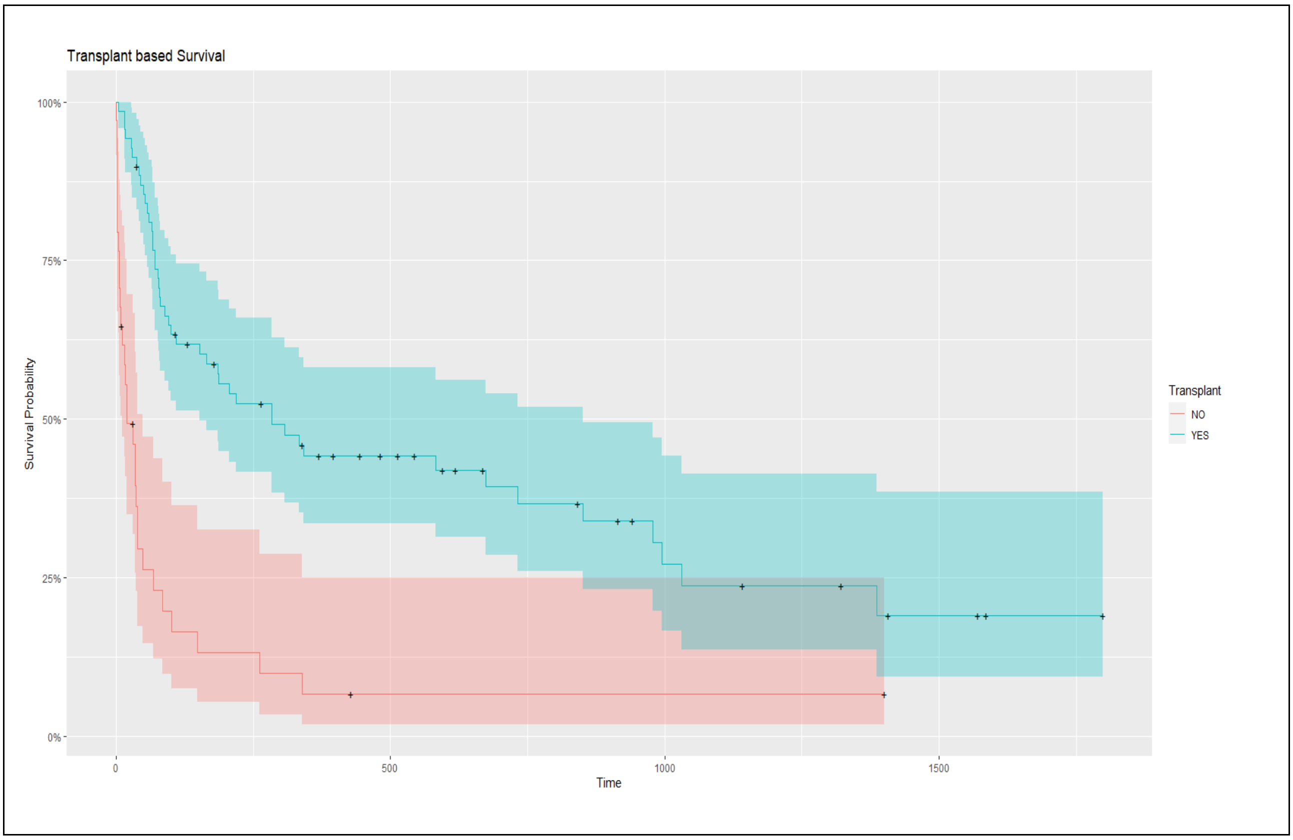


Fig 3.6 : Plot shows the difference in the survival probability for transplant indicator.

Above plot explains, how the survival probability of the patients is higher and having less steepness who have transplant indicator “Yes” . Also , it infers that probability of surviving increases when undergoning with heart transplant.
Below plot also compares the probability of surviving if the subjects has underwent into the transplant program. It also plot the p-value of log rank test as well. The p-values suggests the hihl signigficant results if we consider p< 0.05 as statistically significant value.

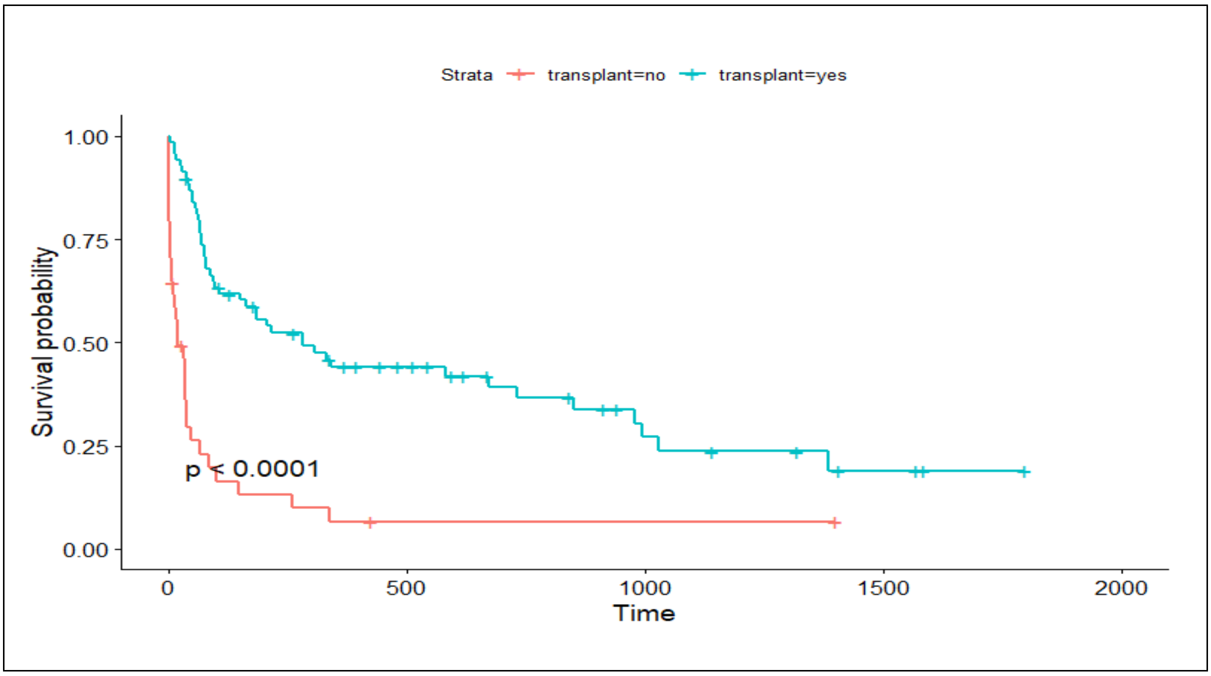


Fig 3.7 : comparative plot of the survival probability for transplant indicator with log rank test

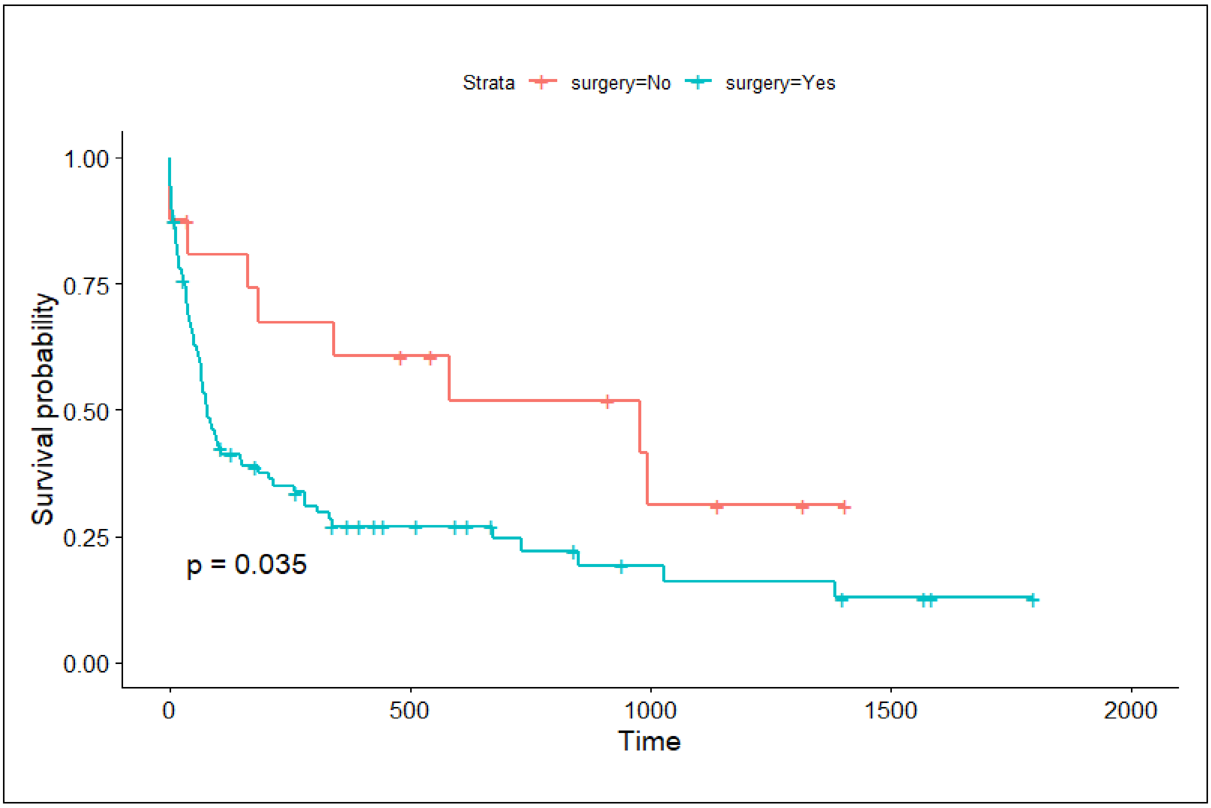


Fig 3.8 : comparative plot of the survival probability for prior surgery with log rank test

The above plot shows the survival probability based on the if the patient had the prior surgery. It is a clearly visible with the plot that patients who had any kind of prior surgery has low surviving probability thorough out the time. The decrease in survival probability is follows a very different path and this result is significant as well.

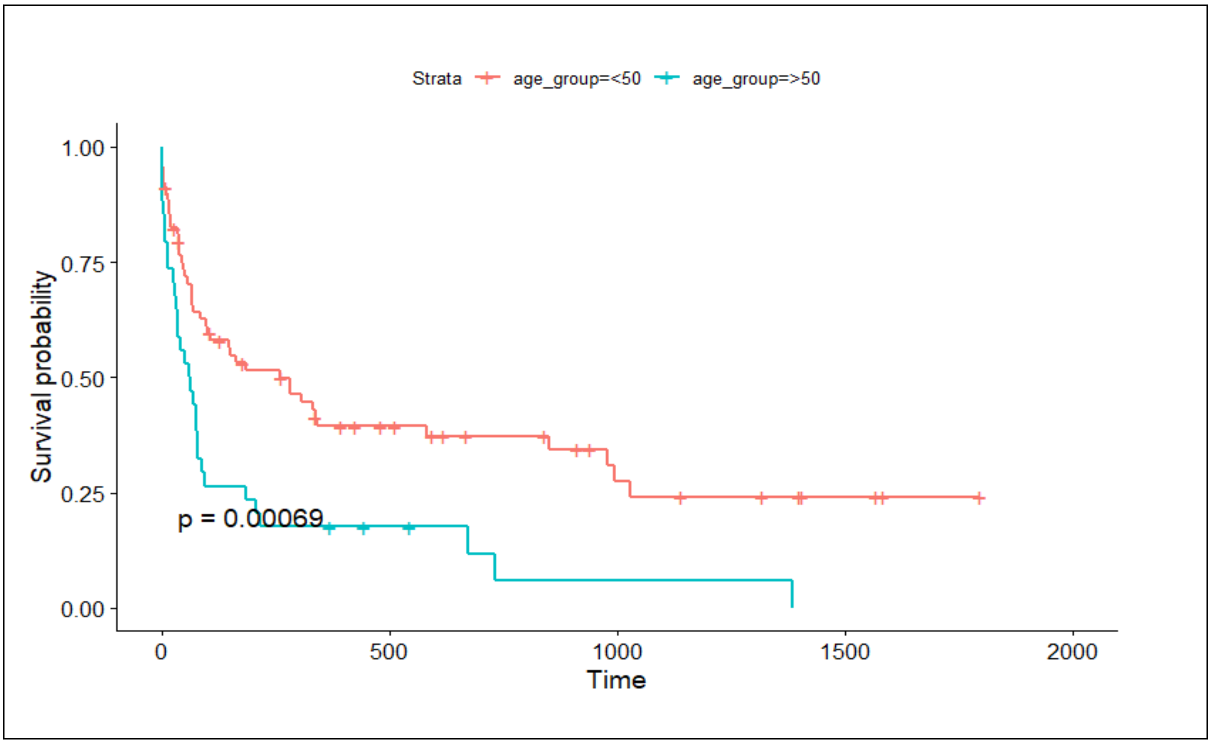


Fig 3.9 : Comparative plot of the survival probability for different age group with log rank test

This is very crucial plot to understand, and it is very intuitive too. First of all, the long rank test p value is very low which means it is highly significant results. Now, the patient who belongs to the category of age group above 50 and enrolled in the study have higher chance of dying early compared to patients who belong to age group below 50. However, only the age factor is not the only perspective to see it but it is an import key feature having dependency with event status.

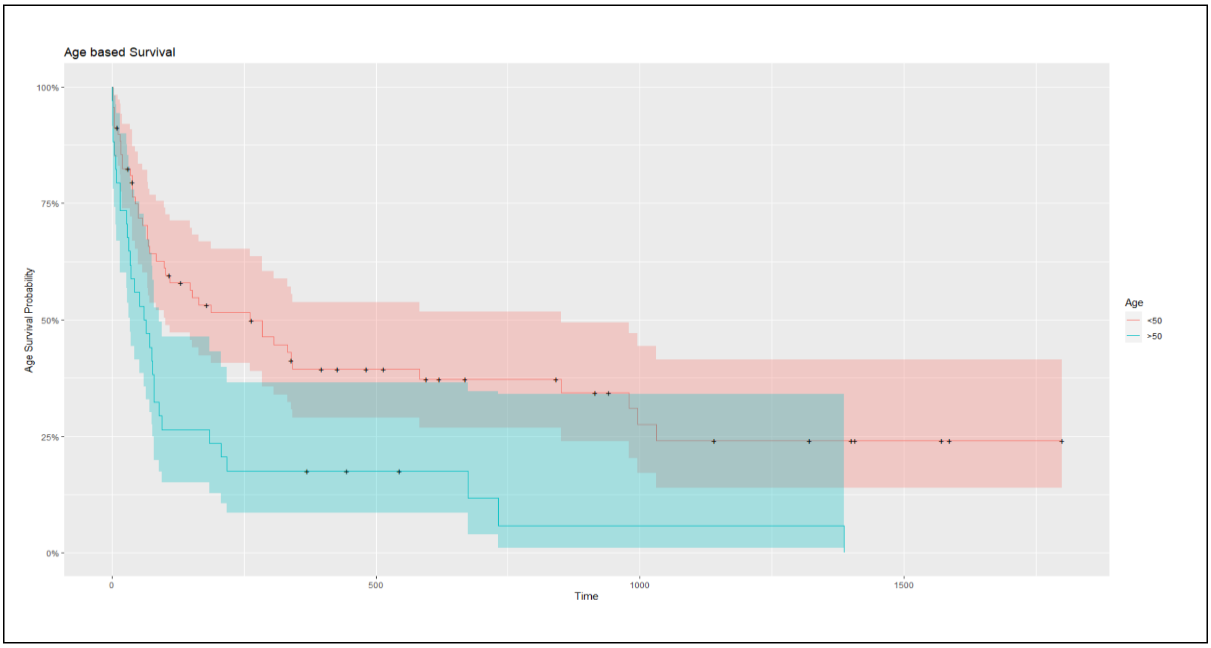


Fig 3.10 : Kaplan Meier survival probability for age group >50 and <50

This plot for the same Age group. But it shows superposing region where the probability of surviving is similar still there is clear distinction between the expected value of 2 trend. It can be observed that the decrease in survival probability is more prominent in age over 50.

There are some better systematic methods of analysing the survival data in terms of hazard function. Cox hazard models gives us the framework to look in the covariates in terms of rate of failure also called Hazard Ratio. Below plot is a forest plot which shows Hazard ratio (HR) which is being derived from the model. As per hazard function definition, h(t) , HR > 1 indicates increased risk of death , specifically to patient condition. Similarly, HR <1 shows decreased risk.

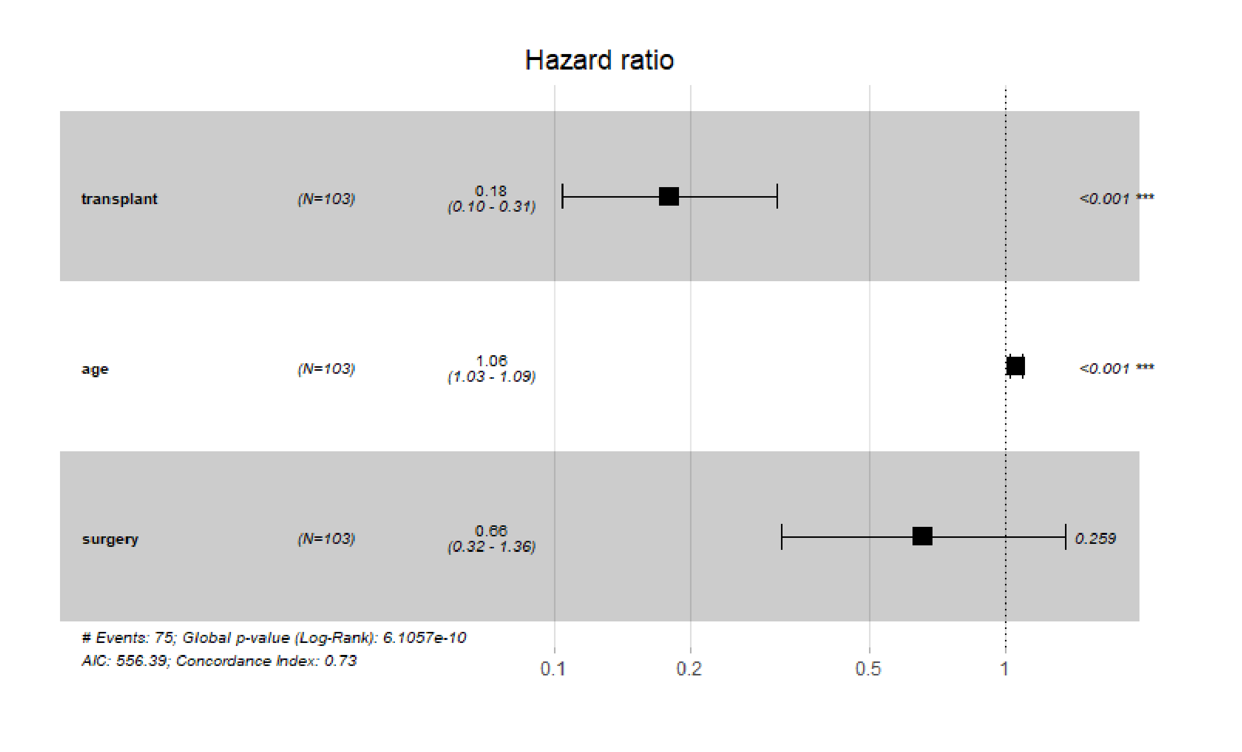


Fig 3.11 : Hazard Ratio Plot for Key covariates

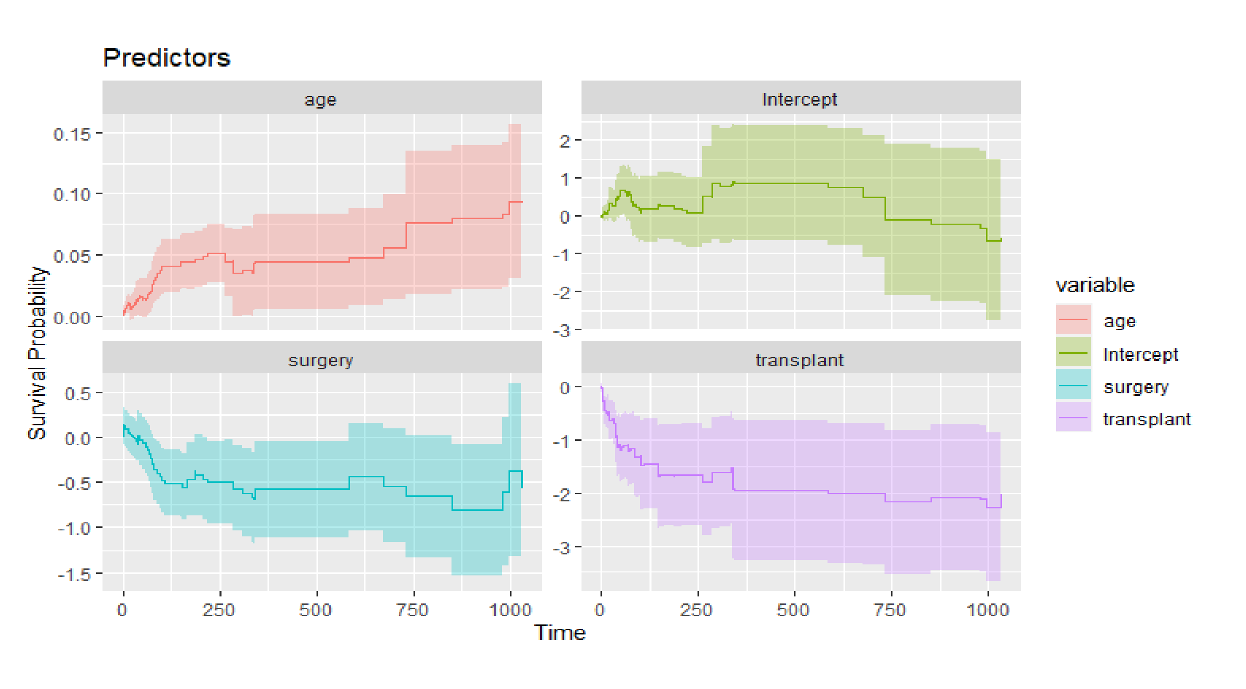


Fig 3.12 : Survival Probability of Predictors with its significance

Random Forrest plot: Below plot shows the plot for each observation which could be helpful in certain scenario .It plots 25 random patients from the dataset and the dark black line shows the global average for all the subjects .

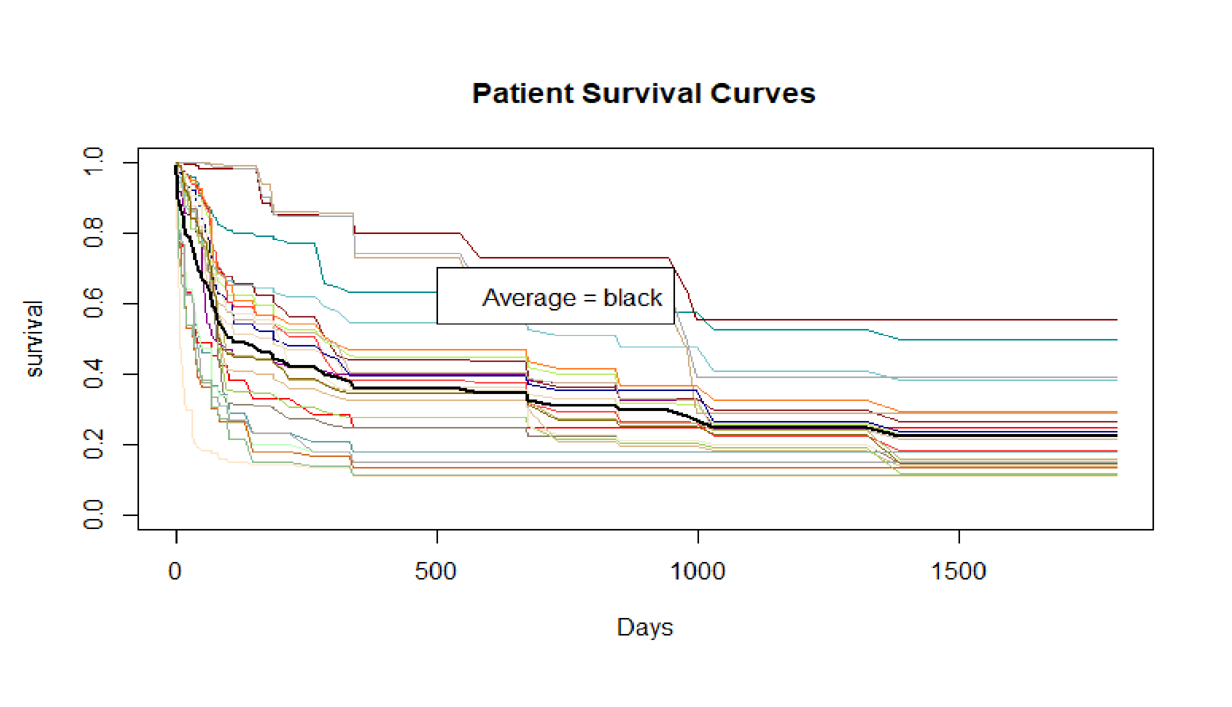


Fig 3.13 : Individual Patient survival Curve

Now , we will check how the random forest provides variable importance among the transplant , surgery and age .

Variable	Variable_importance
transplant	0.1535
surgery	0.0260
age	0.0223

Table 3.3 – Variable Importance by Random forrest

Above results shows that transplant is most important factor for the study and survival probability also. It lies with our belief in considering the patients who underwent through transplant program will enhance the probability of survival or reduces the risk of death.

Chapter 4

Modelling Technique

4.1 Bayesian Network

We need to represent the covariates, its causal dependencies, and risk prediction in a probabilistic graphical structure. One such model is Bayesian Network which displays the structure to construct our prediction model. A Bayesian network B [9] is formally defined as a pair B=(G,Pr), where G is a DAG G=(V(G),A(G)) with a set of vertices V(G)={V1,V2,…,VN}, representing stochastic covariates, and a set of arcs A(G)⊆V(G)×V(G), representing conditional and unconditional stochastic independencies among the covariates[1]. On the set of covariates V, a joint probability distribution Pr is defined that respects the (in)dependencies represented in the graph:

$$\mathrm{Pr}(\mathrm{V1},\ldots,\mathrm{VN})=\prod_{i=1}^{N} Pr\big(Vi\big|\pi(Vi)\big)$$

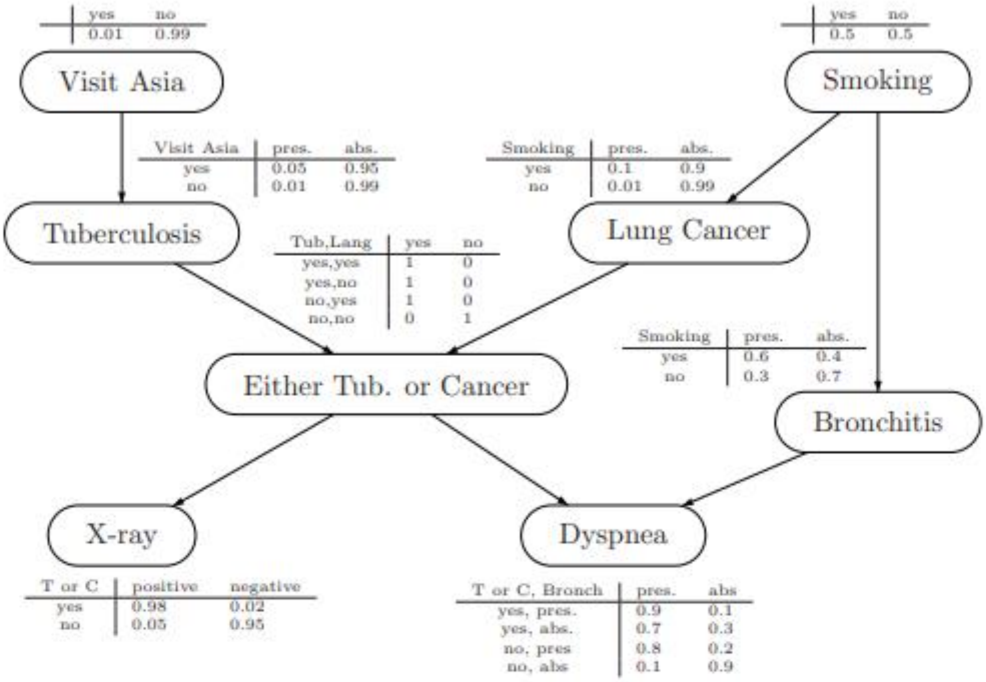
where π(Vi) stands for covariates corresponding to the parents of vertex Vi.

The causal nature of Bayesian networks [57], where nodes represent states and arrows represent causal influence, is probably too simple to be assigned a single inventor. In statistics the redecessors of these kinds of models are usually stated to be path diagrams [58,59] and structural equation models [58].

Bayesian networks are broadly applicable and significant because of it is both intuitive with respect to the domain of interest and, also itemize the objects that which provides equations to predict. Mathematically, it explains the generalization, and planning. These important attributes of BN make them widely useful in clinical datasets. The way how BN automatically learns the structure from the data supplements its prospects for global usage. Academically, BN’s intuitive nature makes it a very contemporary and interesting scenarios in developing complex structures by principled ways of using data. One important thing to understand here is, because of its robustness to handle the causal dependency it is more as dependence networks instead of Bayesian Network as there is no Bayesian in BN. This is a part of probabilistic graph models and viewed as one member in a much larger family of graphical models [68, 37].

4.1.1 Bayesian Network as Knowledge Representation

As probability theory has been used for knowledge graph because of its ability to answer the correlated evidence, still it has its own set of issues which makes it infeasible. It requires all possible permutations and combinations of events which may happen to model. Bayesian networks comes up with a solution by developing the structures using joint distribution of the domain into smaller chunks of interconnected sections (Fig 4.1). It acknowledges much efficient inference and calculate conditional probabilities which measures the probabilities or chances of unknown events given the data observations. Its efficiency improves with compactness of the graph. The Bayesian network structure contains causal links which represents causal mechanism of the domain generating graph connecting these links, which makes the probabilities easy to access which is required to measure the dependencies between the covariates and target variables.



[57] Fig 4.1 Bayesian Network as interconnected sections

Bayesian networks causal inference not only explains the joint probability distribution but also demonstrates how the real world responds on changes occurred by additional external forces [62]. That’s the reason why Bayesian networks are useful for planning and explanation.

4.1.2 Bayesian networks as joint probability distributions

A Bayesian network is a representation of a joint probability distribution for multivariate random variable. In this paper, we are considering Xs as multivariate features with finite domains where each coordinate Xi of an n-dimensional random vector X = (X1, . . . , Xn) contains finite values.[57]

Basically, a Bayesian network has 2 parts:

Qualitative part, also can be seen as structure that can be represented as a DAG (Directed Acyclic Graph) and

Quantitative part, also considered as parameters that further demonstrates the causal dependence and relationship between the covariates, defined by the structure.

The structure G for a multidimensional random variables X’s, has one node per X. It means a BN will have number of nodes equal to selected features or variables. In particular, “node Xi” corresponds to the variable Xi. The structure of a BN can be denoted as a vector G = (G1, . . . , Gn) where each Gi refers a set of those nodes from which there are arcs to node Xi . Therefore, The set Gi is called as the parents of node Xi and the set Gi ∪ {Xi} the family of Xi []. It is to be noted that all vector of nodes is not mandatorily to be valid, it can be empty set if Xi doesn’t have any parent. This fulfils the acyclicity.

Considering the joint probability, a Bayesian network B = (G, θ) the probability of a vector X = (X1,.., Xn) can be defined as

$$\Pr(X|B) = \prod_{i=1}^{N} Pr(Xi|Gi \ = X\mathrm{Gi}\,, \Theta\mathrm{i})$$
$$= \prod_{i=1}^{N} \Theta\mathrm{ijiki}$$

4.1.3 Bayesian Network as generative Models

Bayesian network can also be used as generating device. When it is required to sampling the data from the parameters, BN can be used to sample a data vector and generates the values of its coordinates in topological order. To maintain the relevance of parent child structure, it is required to be generated in an order that confirms that the parents' variables are sampled before child variable. Therefore, it is easy to generate the Xi by the probability distribution P(Xi | Gi = XGi , θi) that is readily available in a network.

[57]Algorithm

```
Pseudo Code to use BN as generative machine

Gendata (B, topolorder):
input:   Bayesian network B = (G, θ),
         topological ordering topolorder of indices {1 ... n} by G
output: data vector X
n ← length(G)
X ← vector of n numbers all -1
for i in topolorder do
         j ← $X_{Gi}$
         Xi ← random sample by ($θ_{ij}$)
end
return X
```

4.2 Bayesian Network Requirement for this Dissertation

This dissertation is handling the clinical data with censored information and based on the property of Bayesian Network, it is very well suited to tackle the intricacies of risk prediction in Stanford heart transplant dataset. The BN have a clear upper hand over other ML classifiers because of its interpretability in the healthcare domain. We are adopting the BN framework which handles the missing information efficiently.

As we are estimating the risk parameter for random ith observation considering the data with features represented by a p-dimensional vector X =(X1, . . . , X p) where Xi is the ith risk factor, To be noted , values for some of the factors could be missing for certain subjects. Consider, event E = 1 refers to an event occurred for a particular observation within study period τ years, and E = 0 shows no such events in same time frame. Just for simplicity for now, considering the that at least

τ years of follow-up is available on each patient so that E is fully observed. We can refer Bayes theorem to estimate

PE|X(e|x), the conditional probability that E = e given the features x of a particular patient.
The conditional probability of an event as

$$PE|X(e = 1|x) = \frac{PX|E(x|e=1)\ PE(e=1)}{\sum_{e\ \epsilon\{0,1\}} PX|E(x|e)\ PE(e)}$$

so that the focus is shifted to estimation of the conditional density/probability PX|E (x|e) and the probability PE (e) for e = 0, 1.
To simplify the joint modelling task, one can represent the joint distributions of X|E = e using a directed acyclic graph (DAG), i.e., a Bayesian network. The DAG encodes conditional independence relationships between variables, allowing the joint distribution to be decomposed into a product of individual terms conditioned on their parent variables

$$PX|E(x|e) = \prod_{i=1}^{p} PX\mathrm{i}|P\mathrm{a}(X\mathrm{i}), \mathrm{E}\{\mathrm{xi}|\mathrm{Pa(xi)}, \mathrm{e}\}$$

where Pa(Xi) are the parents of Xi .
Based on above derivation, we would look for individual conditional probabilities of covariates present in heart transplant dataset. Hence the above equation would turn into P(Age|E =0,1).
Similarly, for multiple features and depends on parent-child relationship between these covariates.
Bayesian network provides a graphical structure with conditional probability. Hence , BN modelling is very much suitable for the kind of problem we have.

Chapter 5

Learning Bayesian Network (Parameters and Structure)

Learning the structure of the Network is probably the most challenging task while dealing with Bayesian Network. Like any other machine learning model, we do not need just a suitable model but to have an optimized version of it. It requires training and subsequently learned parameters of the model. It is very much required to achieve generalization for any kind of similar problem. In the context of Bayesian Network, it has a structure and parameter to learn.

Suppose we have a dataset D and we need to find a model B that is a best fit for D. Usually, we introduce a scoring function or a common metric which evaluates all the possible structures with respect to given dataset and then compare these scores against each network structure and finds the best one. So, this is a common approach to tackle this problem to evaluate networks according to the score based metrics [60]. We generally use belief scoring function and minimum description length-based scoring functions which is same as BIC.

There is another way to learn network especially in the field of Bayesian Network is constraint-based learning. Constraints are typically conditional independence statements that are determined by statistical tests on the data. Once the structure of Bayesian Network is completed depends on the dataset D , we can estimate conditional probability tables(CPTs) directly from the data using frequency distributions over conditional spaces.

Now from the constructed network, we can make predictions about the target variable of "interest" by processing data evidence. There are various methods for learning Bayesian Network from the data, but most of them belongs to constraint-based or score-based as described above.

In the later section of Learning BN from the data, we would try to explain one algorithm from each method, conditional independence algorithm for constraint-based methods and a hill-climbing algorithm for score-based methods.

It is majorly categorized as learning the parameter of BNs when structure is given, and other is learning the structure itself. This chapter will discuss the techniques that are used for learning Bayesian Network parameters and structures.

5.1 Learning Parameters for BN

Bayesian Network parameter learning is one of the categories in learning BNs. We learn the parameters of BN with the structure is given. In the Bayesian network application, where Bayesian methods are predominantly used, the major problem arises with conjugate prior. We need to estimate a distribution family which has posterior over the parameters belongs to same family as the prior. When learning the parameters for a Bayesian network, the common assumption is to have a complete data and structure G of the network which is already been generated in previous section. However, the major challenge is to find which parameter values actually generated the data.

While performing parameter learning, if data is sufficient enough Bayesian Networks can be easily constructed using conventional methods like maximum likelihood (MLH) approach. But in most clinical survey analysis, data is mostly insufficient, and which tends MLH to overfit. In cases like ours of rare disease or heart transplant, collecting sufficient data is not an option which leads to overfitted BN model.[2].

[54] explained it very well by demonstrating a basic theory .

Suppose, there is a BN (G) with nodes X1, ..., Xn. If there is a directed arc from Xi to Xj , Xi is called a parent of Xj , pa(Xj). Given its parent nodes, a node is conditionally independent from all the other nodes. Thus the joint distribution of all the nodes can be written as

$$p(X_{1,}X_{2},,,,X_n = \prod_{i=1}^{n} p(X_i \,|pa(X_i)$$

Now in BN , each node is linked with multiple parameters to describe the conditional probability distribution of the random variable given its parent node. Suppose θ is a vector of parameter value θ_{ijk} , such that

$$\theta_{ijk} = p(x_i^k|pa_i^j) \ ,$$

Where i=(1 ,,,, n) ranges from all the variables in BN , j(j = 1,,,,,,,q_i) is the all possible parent configuration of X_i , and k , k(1,,,,,, r_i)ranges over all possible values of X_i.

Now the goal is to find the most probable value $\hat{\theta}$ for θ which is best fit for dataset D . This $\hat{\theta}$ is mostly quantified by the log-likelihood function, log (p(D| θ) , can be denoted as $L_D(\theta)$. In case when data D is complete, Maximum likelihood estimation method can be easily applied.

However, in case when data D is incomplete, we cannot apply MLE and instead have to use EM algorithm. Now, consider $Y = \{Y_{1,}Y_{2,},,,,Y_N\}$ is observed data, and $Z = \{Z_{1,}Z_{2,},,,,Z_N\}$ is missing data , $D_l = Y_l \; U \; Z_l$. EM algorithm starts at any initial guess at the MLE, $\theta^{(0)}$ and then iteratively generates successive estimates like $\theta^{(1)}$, $\theta^{(2)}$,,,, by repeatedly applying Expectation step followed by Maximisation step.

In this process, E step finds the conditional expectation of log-likelihood function

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}[\log p(D|\theta)|\theta^{(t)}, Y]$$

And M step finds a new parameter $\theta^{(t)}$ maximises the expected log likelihood with a assumption that distribution found in E step is correct.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

For each iteration, we can be assured to get the increased likelihood and finally the algorithm converges to a local maximum of the likelihood function.

To control overfitting, one could consider M to be a tunable parameter and select the number of mixture components using the Bayes information criteria (BIC) or some other goodness-of-fit measure.

5.2 Learning Structure of BN

The score-based approach is the most popular methods of building Bayesian Network from the data specifically when pdf (probability distribution function) estimation is required. This method assigns a score to each possible vector set of BN, which measures how well the BN fits the dataset D. Suppose a BN structure G, on given dataset D . Then score can be defined as :

$$S(G, D) = Pr(G|D)$$

It can also be inferred as posterior probability of G given dataset D.

A score based algorithm tries to maximize the score . the above equation can be represented in other form using Bayes' law:

$$S(G, D) = Pr(G|D) = \frac{Pr(D|G)Pr(G)}{Pr(D)}$$

In order to maximize the score, we need to maximize the numerator since the denominator is not a function of G. We can analyse Pr(G) from the view of prior information (Heckerman (1995)). To understand the structure learning, we can ignore the calculation of $Pr(G)$ or equivalently assume uniform prior over structures. However, to calculate Pr(D|G), the Bayesian approach averages over all possible parameters, and weighing each of it by its posterior probability:

$$Pr(D|G) = \int Pr(D|G,p)Pr(p|G)dp$$

For multinominal local pdfs (Cooper and Herskovits (1992)) , we will get

$$Pr(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_{ij}} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$$

where α_{ijk} and $N_{ijk}$are the hyperparameters and counts for the pdf of Xi for parent configuration j. In large sample limit, the terms $Pr(D|G,p)$, $Pr(p|G)$ in above equation would be approximated as multivariate Gaussian (Kass et al., 1988; Kass and Raftery, 1995). Further, approximating the mean of the gaussian with maximum likelihood value $\hat{p}$, we will have the BIC score as :

$$\text{BICscore(G,D)} = \log\ Pr(D|\hat{p}, G) - \frac{d}{2}\log N$$

first given by Schwartz (1978). The term $\hat{p}$ is the set of maximum-likelihood estimates of the parameters p of the BN, while d is the number of free parameters of the multivariate Gaussian, i.e., its number of dimensions, which coincides with the number of free parameters of the multinomial local pdfs. It also has the intuitive interpretation of the data likelihood minus a "penalty term"

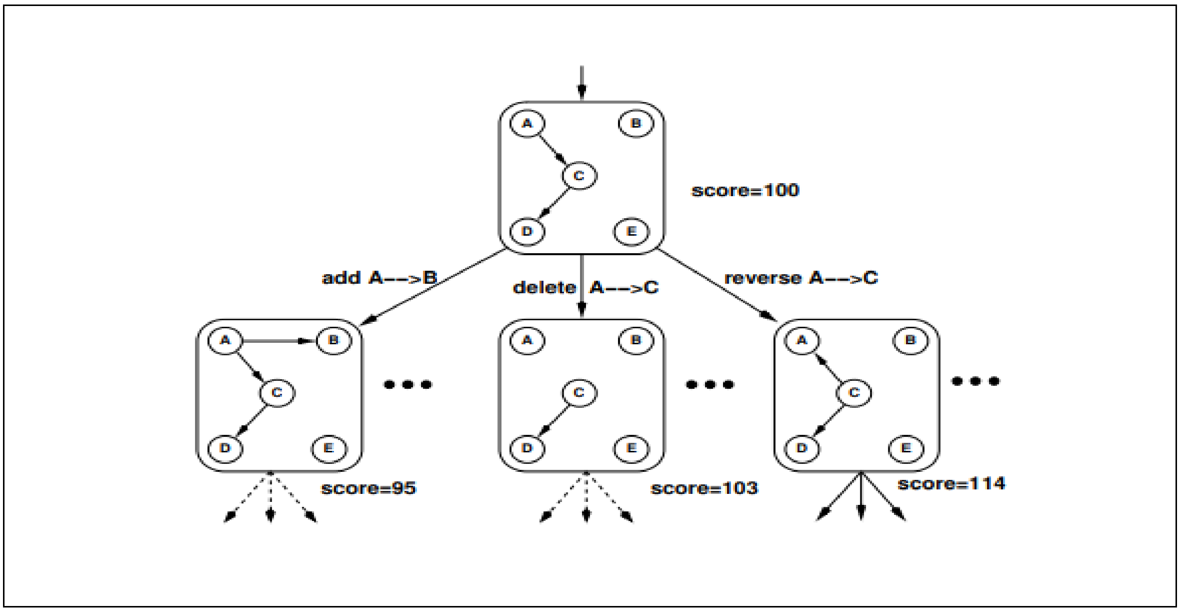
($\frac{d}{2}\log N$) which has the effect of discouraging overly complicated structures and acting to automatically protect from overfitting. The BIC score has been proved as equal to minus the MDL (Minimum Description Length) score (described by Rissanen (1987)).

While optimizing the score with returning structure which tries to maximize it brings up problem as the space of all possible structure is exponential in terms of nodes. There are $\frac{n(n-1)}{2}$ possible undirected edges and $2^{\frac{n(n-1)}{2}}$structures for every subset of these edges. Additionally, there can be more than one orientation of the edges for each choice and hence brute force approach cannot be applied. We will see the possible approaches in next section.

5.3 Approach

5.3.1 Score-based approach: Hill Climbing Algorithm

As discussed above, score-based methods assign a score to each candidate BN which basically measures how well a BN describes the dataset D. To prevent overfitting, the score is modified by adding a factor to penalise overly complex structures. We have found out how the BIC score is the criterion function to be minimised. Since the space of possible structure is in exponential terms , we cannot choose a brute-force approach for computation of scores for each BN structure and hence we choose to pick a heuristic search algorithms also called as hill-climbing [62][1].
The algorithm of HC starts with an empty graph. For each pair of nodes, algorithm will try to add, remove, or reverse an arc. For each iteration of the network which minimizes the score becomes the current structure and then further process goes on and check for next structure. The process will stop when there is no single-arc change can further lower the score. However, there is no definite assurance that the algorithm will find the global minimum. Still, some minor changes as hyperparameter tuning may increase the chance to reach global minimum. Of course, there is no guarantee that this algorithm will find the global minimum.



[61] Fig 5.1 : Hill Climbing Algorithm

5.3.2. Constraint based methods

The constraint-based methods are another way of learning BN structure. These are typically conditional independence statements. However, non-independence constraints can also be used in scenarios where latent variables exist (Verma and Pearl (1990). In this paper, we are only discussing the conditional independence.

It uses conditional independence tests to find a suitable structure of the network, find V-nodes and applies some set of rules to find the directions of the remaining arcs.

It starts with a complete undirected graph, the algorithm tries to find conditional independencies [x,y|Z] in the data. For each pair of [x,y] it sets Z ranging from 0 to total no of covariates minus two.

Next, the set Z becomes a subset of covariates that are next to both x and y. The arc between x and y will be removed from the structure if the algorithm finds any independency.

To test conditionally independence between the pair of covariates, a network structure with arcs For all z belongs to Z: z →y is compared with one with arcs {x → y} U For all z belongs to Z : z →y . This is done using Bayesian metric [73].

The algorithm works on assumption of having data with a perfect map. Any graph G can be called as perfect map with a set of dependency Σ if it follows below conditions:

(1) every dependency logically implied by Σ can be deduced from G and

(2) every dependency inferred from G is logically implied by Σ.

If the above condition does not satisfy then the algorithm could not assign directions for all the deleted arcs [1]. Since, we are not using this method in the paper, we are not going in detail about undirected arcs and its application in Bayesian network.

5.3.3 SoftEM Algorithm

Incomplete data also called as censored data in survival analysis are common scenarios in clinical trials. Bayesian Network is most useful tool in such cases because of its graphical and causal interpretation and representation. As discussed above, we have to optimize the structure and parameter learning of BN using different approaches. Expectation – Maximisation algorithm (EM) is one such method which does both learning using belief propagation and computes the necessary statistics[64].

In the context of Bayesian Networks, structure and parameter learning can be achieved using EM, specifically “SEM” efficiently via below process:

For structure learning, the SEM algorithm [63] can be implemented as:

1. E-step complete the data by computing the necessary expected statistics using the current network structure.
2. M-step finds the best network structure which maximises the expected score and function for the completed data in E step.

For the parameter learning, the E-step and M-step become:

1. The Expectation (E) step computes the expected values of the sufficient statistics (the counts {nijk}) taking inferences from the complete and incomplete samples.
2. The Maximisation (M) step estimates the parameters of the network by taking sufficient statistics computed in E step.

SEM algorithm is realistic and computationally feasible as it finds the best network structure inside of the EM rather than embedding EM inside the network structure learning algorithm. Also, the most important key factor in SEM is it guarantees the convergence of the original both in its maximum likelihood and Bayesian formulations. However, because of the dimensionality of the sufficient statistics computed in each iteration and large number makes it quite expensive to operate. It is different from the parametric EM as in parametric one, we already which expected statistics are going to be used whereas in SEM , we cannot determine it in advance and have to handle each query separately. Hence, most of the execution time spends in the computation of expected statistics.

Chapter 6

Implementation

6.1 Implementation – Theory

Ideally, the development of model for censored data assumes presence or absence of an event as completely observed for all the subjects in the dataset, however, it will lead us to inaccurate modelling and also in real world scenario it is not possible. In our design, if the subject leaves the health system or the observation period ends and he/she is surviving, their status and event history will not be recorded in our dataset. In this paper, we have assumed observation period as τ. If the subject's follow-up ends prior to τ, then their event status at τ is unknown and their event indicator is said to be right-censored.

This chapter will explain to build a probabilistic graphical model, a Bayesian network to predict the risk of an event "fustat" (dead/alive) in τ years when the event status at τ years is right censored. To establishing notation as per the standards in statistical literature, we have T as the time between the beginning of the follow-up period and a event, and define C as the time between the beginning of the follow-up period and disenrollment or the end of the study period. We would calculate, $V = \min(T, C)$ and $\delta = I(T < C)$, the indicator for whether or not a event occurs. If $\delta = 0$, the subject's event time is right-censored. We can only ascertain the value of E either if $\delta = 1$, or if $\delta = 0$ and $V > \tau$; in other words, the value of E is only known if $\min(T, \tau) < C$.

As explained above, one of the naïve approaches to handle earl censored subjects where we don't know about the event E to exclude from the training set or make $E = 0$, but in both the cases, our predictions would be biased towards events to occur. Therefore, in this paper, we are proposing to use inverse probability of censoring weighting (IPCW) approach to handle right censoring subjects. This approach handles censored event times by assuming the censoring time C is independent of the event time T and all features X.

Suppose, $G(t) = P(C > t)$ be the probability that the censoring time is greater than t. As mentioned in previous chapters, it can be estimated $G(t) = P(C > t)$, using the Kaplan–Meier estimator of the survival distribution (i.e., 1 minus the cumulative distribution function) of the censoring times. The Kaplan–Meier estimator (Kalbfleisch and Prentice 2002)[69] of the censoring process is given by i

$$\widehat{G}(t) = \prod_{i:ti<t} \frac{(ni - di*)}{ni}$$

di* is the number of subjects who were censored at time ti , and ni is the number of subjects "at risk" for censoring ,who do not previously censored or experiencing a CV event) at time ti. Unlike other ad hoc approaches to handling censored observations, the Kaplan–Meier estimator is a consistent estimator of G (Kalbfleisch and Prentice 2002)[69].We note that, for IPCW, Kaplan–Meier is applied to estimate the distribution of censoring times, whereas it is much more commonly used to estimate the distribution of event times. Standard software functions for computing the Kaplan–Meier estimator of events times can be used to estimate G by setting the "event" indicators to $\delta{*}i = 1-\delta i$.

Once ˆG, survival probability of each observation j is computed, we will calculate IPCW wj as below:

$$\mathrm{wj} = \begin{cases} \frac{1}{\widehat{\mathrm{G}}(\min(\mathrm{Vj},\tau)}, & \text{if } \min(\mathrm{Tj},\tau) < \mathrm{Cj} \\ 0, \text{otherwise} \end{cases}$$

While fitting the Bayesian Network with above weights, weighted maximum likelihood would be used to estimate the parameters in PE(e) and PGi|E (gi,|e) where the contribution of the jth subject to the likelihood is weighted by ωj. Hence the probability equations become:

$$\hat{P}(E(e) = \frac{1}{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \Vert[\mathrm{Ej} = \mathrm{e}]\mathrm{wj}$$

$$\hat{P} \text{ Zi|E (Zi,|e)} = \frac{\frac{1}{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \Vert[\mathrm{Zij=zij}\ ,\mathrm{Ej=e}]\mathrm{wj}}{\frac{1}{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \Vert[\mathrm{Ej=e}]\mathrm{wj}}$$

We can clearly infer from the above equations, IPCW estimator of PE(e = 1) is the Kaplan–Meier event probability. To estimate the parameters in PYi |Zi ,E (yi |zi , e) we use a weighted EM algorithm where the contribution of each subject j is weighted by ωj .

In this approach of IPCW, the subjects for whom we can assess E would only add up to the data to be analysed, but they are weighted in such a way that they will overshadow the patients who were censored before the observation time τ.

We can understand it as, the subjects or patients who has comparatively longer time to event are more chances to have smaller survival probability, G. Hence, it will receive larger weights. Also to be noted, the patients with E = 0 (and V > τ) the weights for all individuals are

1/ $\widehat{\mathrm{G}}$ (τ), so the maximum likelihood estimators for PGi|E (gi,|e) are the same as in the unweighted analysis.

In the later section, we will describe how inverse probability of censoring weightings results in consistent estimators

6.1.1 Inverse Probability of Censoring Weighting (IPCW)

As explained in [6],the IPCW – Inverse Probability of Censoring Weighting, oversample subjects with E = 1 if we exclude patients for whom E is unknown. Post that, we can apply any machine learning for prediction of risk considering the calculated weights as weight parameters. The general-purpose IPCW method proceeds as follows:

1. Using the training data, evaluate the function G(t) = P(Ci > t), the probability that the censoring time is greater than t, using the Kaplan-Meier estimator of the survival distribution of the censoring times.

2. For each patient i in the dataset, calculate an inverse probability of censoring weight. Patients whose event status is unknown at τ (i.e., are censored prior to τ and therefore have Ci ≤ min (Ti, τ)) are assigned weight ωi = 0. The remaining patients are assigned weights inversely proportional to the estimated probability of being censored after their observed follow-up time.

3. Apply an existing prediction method to a weighted version of the training set where each member i of the training set is weighted by a factor of ωi. In other words, if ωi = 3, it is as if the observation appeared three times in the data set.
   Using the above method of weight assignment, it can be used to get modelled using conventional ML classifiers.

6.1.2 Interpretation IPCW

In this section, we will explain and interpret the process of IPCW [6], inverse probability of censoring weighting and understand it with simple example that how it handles censoring and leads to accurate risk prediction across a variety of machine learning techniques. Example quoted from [6].Suppose we estimate that 1/3 of subjects have censoring times greater than 2.5 years (i.e., Gˆ(2.5) = 1/3), and that the ith subject is observed in our study to experience an event at t = 2.5 years (i.e., δi = 1 and Vi = 2.5). For this subject, the event status is known (E = 1) and her/his IPC weight is ωi = 3. This subject is weighted by a factor of 3 because she/he can be thought of as representing 3 individuals: 2 similar or "shadow" subjects censored prior to their event time at t = 2.5 for whom E is unknown, plus themselves (recall that on average 2/3 of subjects in this example with event times equal to 2.5 are censored prior to experiencing the event). Thus, subjects with known event status E and a longer time-to-event receive larger weights as they represent a greater number of "shadow" subjects whose event status is unknown due to censoring. IPCW is conceptually equivalent to creating a new dataset where each subject is replicated ωi times. However, creating such an expanded dataset is often not advisable, both for reasons of practicality (memory/storage limitations) and mathematical precision (ωi may not be an integer or simple fraction). A full justification of the use.

Lets take a below simple dataset with single binary covariate with 50 data sample provided in [6] to understand the weight calculation.

Xi	Vi	δi	Ei	min(Vi,τ)	Gˆ{min(Vi,τ)}	ωi
1	0	1	1	0	1	1
0	0.2	1	1	0.2	1	1
0	0.4	0	?	0.4	0.98	0
0	0.7	0	?	0.7	0.96	0
1	0.8	0	?	0.8	0.94	0
1	0.8	1	1	0.8	0.94	1.07
0	0.9	1	1	0.9	0.94	1.07
1	0.9	1	1	0.9	0.94	1.07
0	1	1	1	1	0.94	1.07
0	1	1	1	1	0.94	1.07
1	1.3	0	?	1.3	0.89	0
0	1.3	0	?	1.3	0.89	0
1	1.4	1	1	1.4	0.89	1.12
0	1.6	1	1	1.6	0.89	1.12
1	1.6	1	1	1.6	0.89	1.12
0	2.1	0	?	2.1	0.87	0
0	2.3	0	?	2.3	0.84	0
1	2.3	1	1	2.3	0.84	1.19
1	2.4	0	?	2.4	0.81	0
0	2.5	1	1	2.5	0.81	1.23
1	2.6	1	1	2.6	0.81	1.23
0	2.8	0	?	2.8	0.79	0
1	3.2	0	?	3.2	0.73	0
0	3.2	1	1	3.2	0.73	1.37
1	3.2	0	?	3.2	0.73	0
0	3.3	0	?	3.3	0.7	0
0	3.4	0	?	3.4	0.67	0
1	3.4	1	1	3.4	0.67	1.49
0	3.5	1	1	3.5	0.67	1.49
0	3.7	1	1	3.7	0.67	1.49
0	3.7	1	1	3.7	0.67	1.49
0	3.8	1	1	3.8	0.67	1.49
0	3.9	0	?	3.9	0.63	0
1	4.2	1	1	4.2	0.63	1.58
1	4.3	1	1	4.3	0.63	1.58
0	4.9	1	1	4.9	0.63	1.58
1	5.3	0	0	5	0.63	1.58
1	5.7	1	0	5	0.63	1.58
0	5.8	0	0	5	0.63	1.58
0	6.1	0	0	5	0.63	1.58
1	6.4	1	0	5	0.63	1.58
0	6.5	1	0	5	0.63	1.58
0	6.6	1	0	5	0.63	1.58
0	6.6	1	0	5	0.63	1.58
1	6.6	1	0	5	0.63	1.58
0	6.8	0	0	5	0.63	1.58
1	6.8	0	0	5	0.63	1.58
1	7.6	1	0	5	0.63	1.58
1	7.8	1	0	5	0.63	1.58
0	8.7	0	0	5	0.63	1.58

Suppose that we wish to estimate the probability of having an adverse event within 5 years within each level of the covariate (i.e., τ = 5). So , in normal case where Ei is available for each observation , we would just take the average of Ei with each level of the covariate. But, as we can see there are many instances where Ei is not available (denoted by question mark in above table) , such subjects do not have complete information and did not experience an event during its observation period.

Now, as discussed in theoretical part IPCW is a best technique to handle such data. We will calculate the survival distribution of censoring as G(t) using a Kaplan-Meier estimator and compute Gˆ{min(Vi, τ)}. The weight, ωi for each subject is given by Equation 1. Now to estimate probability of having an adverse event within 5 years within each level of the covariate we take a weighted average of Ei within each level of the covariate:

$$\hat{P}(E=1|X=1) = \sum_{t}^{Xt=1} \frac{EtWt}{Wt} = 0.58 \quad \text{and}$$

$$\hat{P}\ (E=1|X=0) = \prod_{t}^{Xt=0} \frac{EtWt}{Wt} = 0.53$$

And subjects for whom we don’t know about Ei will have the weights equal to 0. Hence, Eiωi = 0. There are more sophisticated machine learning methods but the conceptually the idea of IPCW is applicable.

6.1.3 Important Assumptions while implementing IPCW

The IPCW method depends on the below assumptions [6]:

- There are no unmeasured cofounders for censoring.
- If the hazard of censoring is conditioned on the recorded history, it does not further depend on X (sequential ignorability of censoring).
- The data is coarsened at random (CAR), i.e. the censoring mechanism does not depend on the outcome, but may depend on the covariates.

If all these assumptions are satisfied and all prognostic factors are recorded , IPCW estimators will correct the bias due to dependent censoring completely.

Methods to handle observations in which event status is unknown[6]

As explained and understood, Subjects who experienced an event are considered E =1 and those with event free are E = 0. And those E is unknown for them who were event-free but censored before accruing 5 years of follow-up. Below 4 ML strategies could be applied to handle instances of unknown E:

1. Set E = 0 if E is unknown. Techniques using this strategy are denoted with the suffix -Zero.

2. Discard observations with E unknown. Techniques using this strategy are given the suffix -Discard.

3. Use IPCW on observations with E known. The resulting techniques, as described in Section 3, have the suffix -IPCW.

4. “Split” observations with E unknown into two observations with E = 1 and E = 0 with weights based on marginal survival probability. The resulting techniques, as described subsequently, have the suffix -Split.

The split technique of splitting observations with E unknown was described by [5]. For each observation i in the training set for which Ei is unknown, we create two observations, one with E = 1 and the other with E = 0, but with the same features Xi. Suppose , Fˆ(t) is the (KM) Kaplan-Meier estimator of the survival probability at time t, then as per survival function :

$$\hat{F}(t) = \prod_{j:Vj<t} \frac{(nj - dj*)}{nj}$$

where dj is the number of subjects who are observed to experience the event at time Vj, and nj is the number of subjects “at risk” for the event (i.e., not yet censored or experienced an event) at time Vj.

If E is unknown for ith observation in the training data, the weight for the imputed observation with E = 0 is $\hat{F}$ (τ)/ $\hat{F}$ (Vi) and the weight for the imputed observation with E = 1 is 1 − $\hat{F}$ (τ)/ $\hat{F}$ (Vi). The implementation of weights will be similar like how we can apply IPCW. It provides weights to all observations and hence can be better interpreted in analysis.

6.2 Consider it as survey weights

IPCW method modifies our dataset with additional weights and also increases the number of records based on event time and censor time. This data will now act as weighted survey data. As an ideal case assumption, each subject of the target population has the same chances of getting included in the sample but in real world scenario, participation in survey is mostly voluntary and some might refuse as well. In such case, weighting the survey data is one of the best practical way to deal with imbalanced dataset however it is not the best recommended solution. We are following similar approach while using IPCW method to handle censored data.

6.3 High-level overview of the approach

This is a basic algorithm to be followed in our implementation which was mentioned in [4]. Theoretically, we are approaching the below process from weight assignment to building Bayesian network.

Pseudo Algorithm

```
  Input:
    Graphical structure of the probabilistic relationships (edges) between input features (nodes)
    Training dataset (each record consists of input values, follow-up time, and an event indicator)
  Output:
    Function for estimating conditional probability of the event given the input values

1: Estimate survival distribution of the censoring times using Kaplan–Meier estimator
2: For each subject (each record in a dataset):
3:    Compute the inverse probability of censoring weight using the distribution from Line
4: For each node Xi in the graphical model:
5:    Identify the set of parent nodes of Xi , i.e., Pa(Xi )
6:    Model the conditional joint distribution of Xi and Pa(Xi ) given event status as follows:
7:       Let Gi = {Xi , Pa(Xi )}
8:       Partition Gi into continuous and discrete features
9:       For the set of discrete features of Gi :
10:         Compute the IPCW estimates of the conditional probability given event status
11:      For the set of continuous features of Gi :
12:        For each distinct state of discrete features of Gi and event status:
13:           For each number of multivariate Gaussian mixtures (vary from 1 to 4):
14:              Create 40 bootstrap samples of the training data
15:              For each bootstrap sample:
16:                 Use IPCW EM algorithm to estimate parameters in multivariate
                          normal mixture distribution using the IPCW weights
17:                 Calculate the BIC values for the current model
18:                Compute the average model BIC value across 40 bootstrap samples
                    From individual BIC values
19:           Find the model complexity weight for each number of mixtures using
                    the average model BIC values
20:        Use the model complexity weights
21:        Obtain conditional joint distribution of Gi given event status by multiplying
           discrete probability estimates
22:      Derive conditional distribution of node Xi given parents Pa(Xi ) and event status
                    using the conditional joint distribution of
23: Estimate the marginal probability of event using IPCWs
24: Derive conditional probability estimation function of event given input features
           using (a) conditional distribution of nodes given their parents and event status
           and (b) the marginal probability of event.
```

6.4 Methodology
Bayesian Network generation in R /Python with weight integration

6.4.1 Development process

1. Dataset : Stanford heart transplant
   We are using dataset Stanford heart transplant data "jasa" . From initial validation we found that transplant is a key variable which has a causal dependency on "fustat" which is a dependent or target variable. Other important covariates are age and prior surgery.

2. Creating a function which compares the event time, censor time and observation period. Basically, it creates an indicator against each observation or for each patient id if the function finds out the minimum of event time and observation period is less than censor time. In case of true, it generates 1 else 0.

3. Kaplan Meier Survival Function to find the survival probability.
   This is a core function of entire thesis where we populate weights for each observation/patient as per its importance. The value of its importance is measured by survival probability of that observation which is being calculated by Kaplan Meier survival function.

4. Inverse Probability Censoring Weights
   Once get the Kaplan Meier Survival factor from above, we will inverse it to assign the weights to that observation where v is non- zero. Hence, for the observations having censor time less than event time or observation period weights will be zero. Also, we are fitting the survival model object with "v" no of times. This will increase the number of records with v times to overshadow those patients which are left in middle of the study or before the observation time.

5. Now, we have the IPCW weighted data with each observation have its own weights. However, the subjects which left the study group before the total observation period, they will get 0 weights. Still, we have not excluded it from the modelling. They do contribute in weights calculation. Hence , we have the modified dataset with new columns like weights , surviving probability of each subject. This is a weighted dataset which we are going to save it for later use.

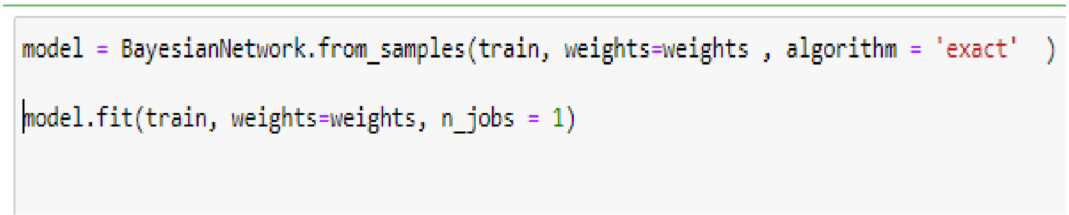
6. Survey weight – we will integrate this new generated data with weights (conceptually as weighted survey data) with our Bayesian network model to learn the parameter and structure. Now , this dataset is a weighted survey dataset having individual weights for each observation. We need to check the continuous and discreet variables and factorize it based on its structure and property.

7. We have used pomegranate framework in python which is specialised in probabilistic graphical model. The we will do some pre processing steps and create our X and y as no of features which we want to create for structure like transplant , age , surgery etc and y as our target variable as event status "fustat".

8. The dataset will be split into train and test dataframe in 80:20 ratio respectively. Also it is important to note that weights needs to be used very carefully. We have to reshape the weights as in numerical array to be passed as "weight parameter" in the model object.

9. Model object gets generated from "BayesianNetwork.from_samples() method(available in pomegranate package) with weights array as parameter as below . we will then fit the model on training set.

   Code Snippet

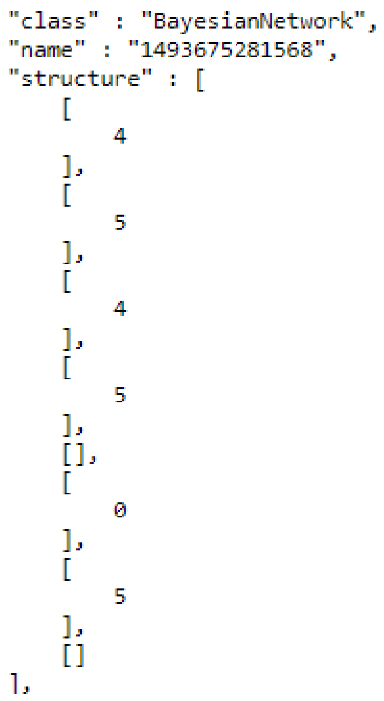


```
model = BayesianNetwork.from_samples(train, weights=weights , algorithm = 'exact'  )

model.fit(train, weights=weights, n_jobs = 1)
```

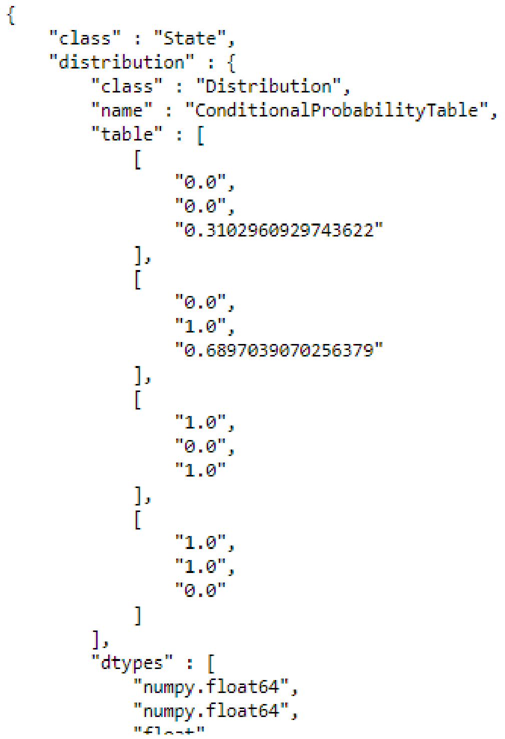
10. The Bayesian model object created will generate a structure and conditional probability table in below format . we have to execute the model.plot() using pygraphiz for graphical structure ;

Structure



```
"class" : "BayesianNetwork",
"name" : "1493675281568",
"structure" : [
    [
        4
    ],
    [
        5
    ],
    [
        4
    ],
    [
        5
    ],
    [],
    [
        0
    ],
    [
        5
    ],
    []
],
```

Conditional Probability Table :



```
{
    "class" : "State",
    "distribution" : {
        "class" : "Distribution",
        "name" : "ConditionalProbabilityTable",
        "table" : [
            [
                "0.0",
                "0.0",
                "0.3102960929743622"
            ],
            [
                "0.0",
                "1.0",
                "0.6897039070256379"
            ],
            [
                "1.0",
                "0.0",
                "1.0"
            ],
            [
                "1.0",
                "1.0",
                "0.0"
            ]
        ],
        "dtypes" : [
            "numpy.float64",
            "numpy.float64",
            "float"
```

11. We will discuss the different Bayesian network structure and conditional probability table created using different algorithms in next section. We have performed prediction analysis also using this model as classifier . It resulted in 70% accuracy. Other metrics are discussed in detail in Performance Evaluation".

12. To generate Bayesian Network in R :
    We have generated BNs in R also using “bnstruct” package which provides a framework to generate DAG “Bayesian Network and connects the variable as per its conditional dependencies. It learns the parameter and structure based on algorithm and scoring function provided in the framework.
    Requirement for Bayesian network to build using “bnstruct”:
    It requires a data object in the format of “BNDataset” which checks the discreetness, node sizes and weighted data to create a new object “BNDataset Object”. It later passed in learn.network() with algorithm and scoring function to generate Bayesian Network.

13. Algorithms and scoring Function
    Here ,we are using multiple algorithms to generate structure and scoring function as BIC to learn the parameters.

14. To handle missing data and imbalanced dataset, we can try for bootstrapping with imputation on BNDataset object created above . However, in case of using SEM algorithm, it automatically do the sampling while execution.

After completion of all the above steps, we have got number of Bayesian network based on algorithm and framework used. We are going to discuss about the structure and conditional probability table results while executing the above steps.
Also we have developed a BN classifier to predict the event status for test set . we will discuss on it and evaluate our model based on some key metrics.

Chapter 7

Results & Evaluation

7.1 Bayesian Network building With Weight Integration

7.1.1 Results

Our implementation consists of learning parameter and structure of Bayesian network considering the IPCW integrated dataset. We have used multiple algorithms to generate the structure and along with it we will have adjacency matrix and conditional probability table for each algorithm. We are using scoring function "BIC", also called as MDL as common method to reduce the risk of overfitting.

Although, we have tried to implement with various algorithms, our main focus will be along "SEM" and "Hill-Climbing "approach. Additionally, we will explain the network learning with all features present in data and also only with those features which are strongly correlated.

1. Algorithm : " SoftEM" , Scoring Function : BIC

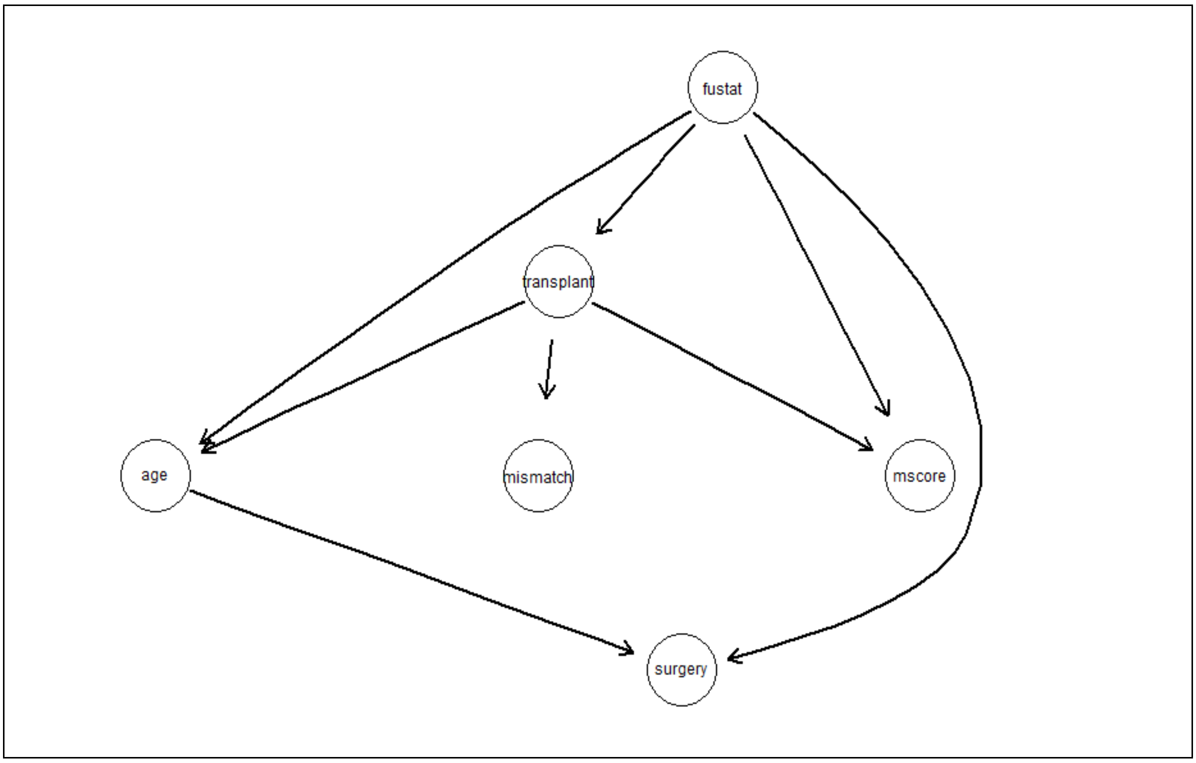


Fig 7.1 : Bayesian Network using "SEM" Algorithm

Above structure represents a DAG which demonstrates the relationship between different covariates. It is easy to interpret as graphical representation. It infers that age has a causal dependency with transplant indicator, and event status “fustat” . Similarly, “prior surgery” has interactions with age and event status.

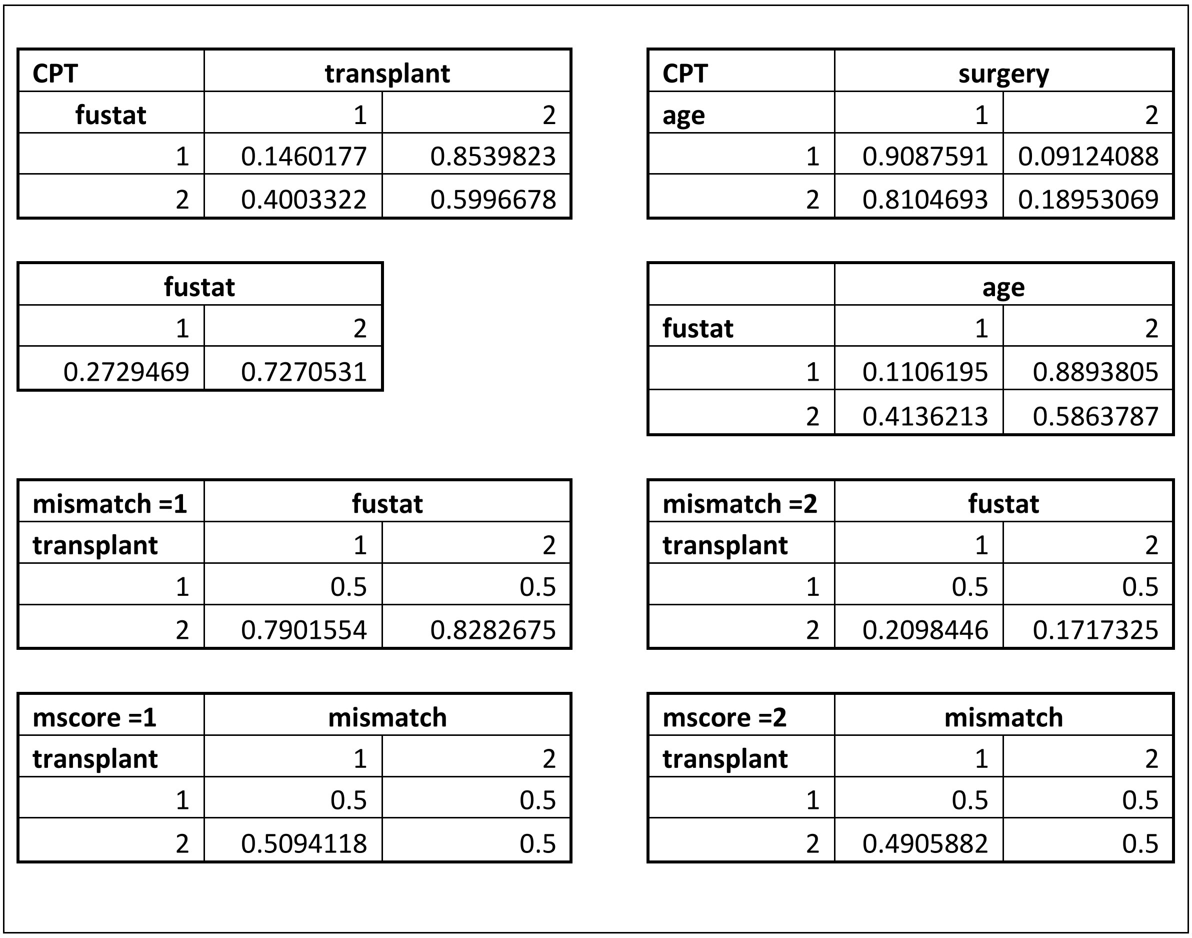
Next, we will check the conditional probability table information from the below table which shows how significantly these covariates affects each other and with what probability.

It also generates the adjacency matrix:

	transplant	fustat	surgery	age	mismatch	mscore
transplant	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
fustat	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE
surgery	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
age	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
mismatch	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
mscore	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table 7.1 Adjacency Matrix for Sem Algorithm

Conditional Probability Table



CPT	transplant	
fustat	1	2
1	0.1460177	0.8539823
2	0.4003322	0.5996678

CPT	surgery	
age	1	2
1	0.9087591	0.09124088
2	0.8104693	0.18953069

fustat	
1	2
0.2729469	0.7270531

	age	
fustat	1	2
1	0.1106195	0.8893805
2	0.4136213	0.5863787

mismatch =1	fustat	
transplant	1	2
1	0.5	0.5
2	0.7901554	0.8282675

mismatch =2	fustat	
transplant	1	2
1	0.5	0.5
2	0.2098446	0.1717325

mscore =1	mismatch	
transplant	1	2
1	0.5	0.5
2	0.5094118	0.5

mscore =2	mismatch	
transplant	1	2
1	0.5	0.5
2	0.4905882	0.5

Table 7.2 : Conditional Probability Table for “SEM” BN

Fustat 2 – Dead, transplant 2 = yes , surgery 2 = yes

CPT Inference

1. CPT for Transplant: It suggests the probability of event not to occur(1=Alive) given that transplant has been done (2) is 85%, and 40% probability of event happening if transplant has not been done.
2. CPT for fustat (event): Overall probability of event to be occurred is 72% against 27% not to occur.
   Let’s check the conditional probability which involves more than 2 nodes
3. CPT for transplant given fustat and mismatch
   a. It suggests if mismatch =1, then probability of transplant to happen is going to be same irrespective of event status. However, the probability of transplant not done in case of patient is dead is slightly more compared to alive.
   b. Whereas for mismatch =2 , probability of transplant to happen is same again given any event , and also its very low probability for transplant to occur with any status of event.
4. CPT for age: It is interesting and intuitive also. Since age =1 is age >50. The probability of age is greater than 50 given that surgery is true i.e its high probability to patient to have age greater than 50 with any prior surgery and its ~90% . It has a lot of information in it.
5. Predict event given age: It infers that probability of being alive (fustat 1) when patient is less than 50 is almost 88% compared to 11% for patient with age >50.
6. Similarly, as mismatch, now we have 2 condition for mscore:
   a. For mscore =1
      Probability of transplant occurred given that mismatch is 1 or 2 is almost similar as 50%.
   b. For mscore =2 , CPTs are similar as transplant, and mismatch condition as a.

2.Algorithm: “hc “(Hill Climbing) and Scoring Function : “BIC”

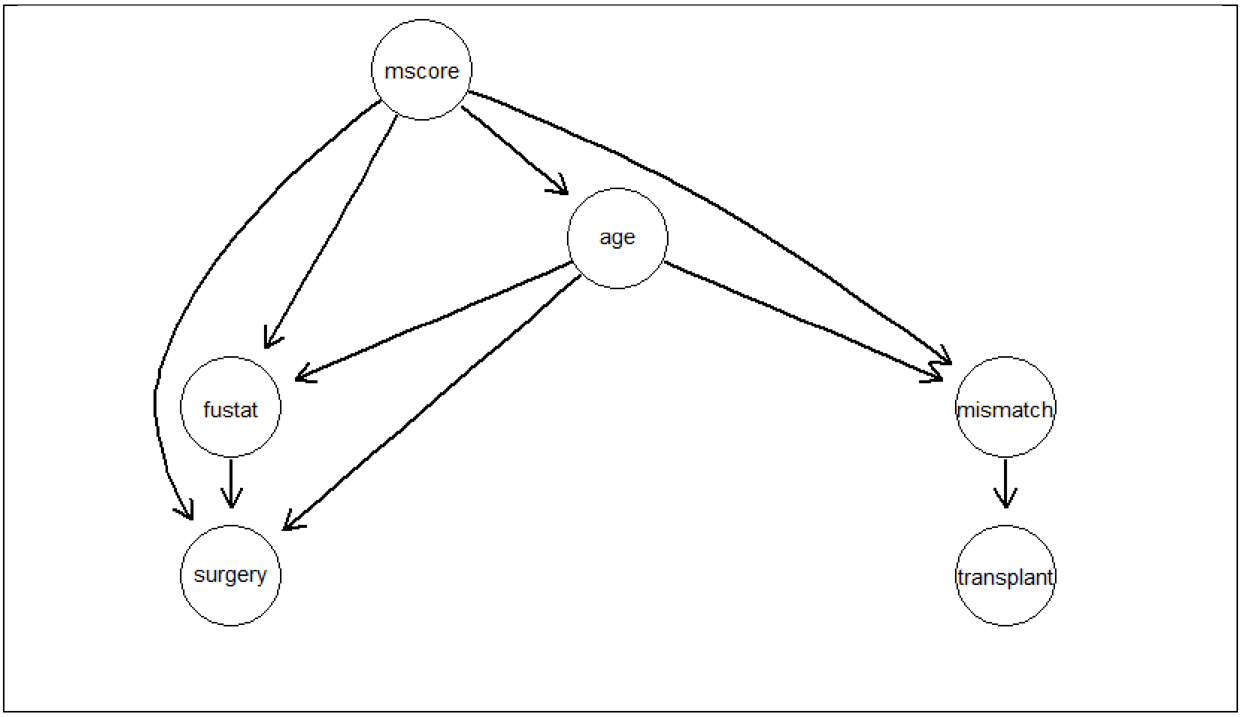


Fig 7.2 : Bayesian Network using “Hill Climbing”

Hill climbing has generated a different structure with slight changes in dependencies also. As in this network , event status “fustat” has a dependency with age , mscore as parent node . Also, surgery is dependent on fustat. We will find out the probability of these dependency using conditional probability table mentioned below.Also , it generates an adjacency matrix , which gives us the idea about causal dependency among the covariates.

Adjacency Matrix : It shows the causal dependency between multiple covariates .

	transplant	**fustat**	**surgery**	**age**	**mismatch**	**mscore**
transplant	0	0	0	0	0	0
fustat	0	0	1	0	0	0
surgery	0	0	0	0	0	0
age	0	1	1	0	1	0
mismatch	1	0	0	0	0	0
mscore	0	1	1	1	1	0

Table 7.3 : Adjacency Matrix HC Algorithm

Conditional Probability Table

	transplant	
mismatch	1	2
1	0.002347	0.997653
2	0.010204	0.989796

	age	
mscore	1	2
1	0.2443609	0.7556391
2	0.4379845	0.5620155

fustat = 1	**mscore**	
age	1	2
1	0.130769	0.146018
2	0.559702	0.389655

fustat = 2	**mscore**	
age	1	2
1	0.8692308	0.8539823
2	0.4402985	0.6103448

mscore	
1	2
0.5076336	0.492366

mismatch =1	**mscore**	
age	1	2
1	0.869231	0.712389
2	0.798508	0.886207

mismatch =2	**mscore**	
age	1	2
1	0.1307692	0.2876106
2	0.2014925	0.1137931

mscore = 1, surgery = 1

	age	
fustat	1	2
1	0.970588	0.713333
2	0.853982	0.816384

mscore = 1, surgery = 2

	age	
fustat	1	2
1	0.02941176	0.2866667
2	0.1460177	0.1836158

mscore = 2, surgery = 2

	age	
fustat	1	2
1	0.5	0.712389
2	0.997409	0.816384

mscore = 2, surgery = 1

	age	
fustat	1	2
1	0.5	0.2876106
2	0.00259067	0.1836158

Table 7.4 : Conditional Probability table HC Algorithm

CPT Inferences:

It shows mscore as root node and having the probability of mscore for being 1 given all the features is 50.7% and 49% for being 2.

1. CPT for mismatch: It suggests very low probability of mismatch to attain value 1 given that transplant has been done (1) is just 0.2%, and even same with to attain the probability of mismatch to be 2. However, approx. 99% probability that mismatch will take any value 1 or 2 when transplant is not done.
2. CPT for fustat (event) given mscore, surgery and age: This is quite intuitive and complex to understand as it involves 4 covariates and their relationship. We have 4 different probability table for 4 different conditions.

a. mscore = 1, surgery = 1

It suggests the probability of patients to be alive event to take place given the patient is greater than 50 years and he/she had some prior surgery and mscore value is 1, even for younger patient also it is high at 70+ % of probability to event to take place if there is history of prior surgery.

b. mscore = 1, surgery = 2

As above, It says the probability of event to take place given the patient is greater than 50 ears and he/she had some prior surgery and mscore value is 1.

c. mscore = 2, surgery = 1

we can infer from this table that the if there is no prior surgery, the chance of having death is anyway too less. The probability of event to take place means its high probability to be alive given the age is less than 50.

d. mscore = 2, surgery = 2

This is challenging, lets leave mscore, as it is not giving much information. If there is prior surgery, the probability of experiencing death is very high , more than 95% if they belong to age greater than 50. And even, for age less than 50 , probability of experiencing death is more than 80%.

We have to explore more on CPTs.

Other network structure generated during the experiments

We have got different structures by changing the parameters and when selecting the few features or important features based on the above results. Also explored to build network without IPCW. However , it didn’t create any intuitive results to draw any insights.
Below are some BNs generated while exploring :

1. With 3 key features “Transplant indicator, Prior Surgery and “Age” :

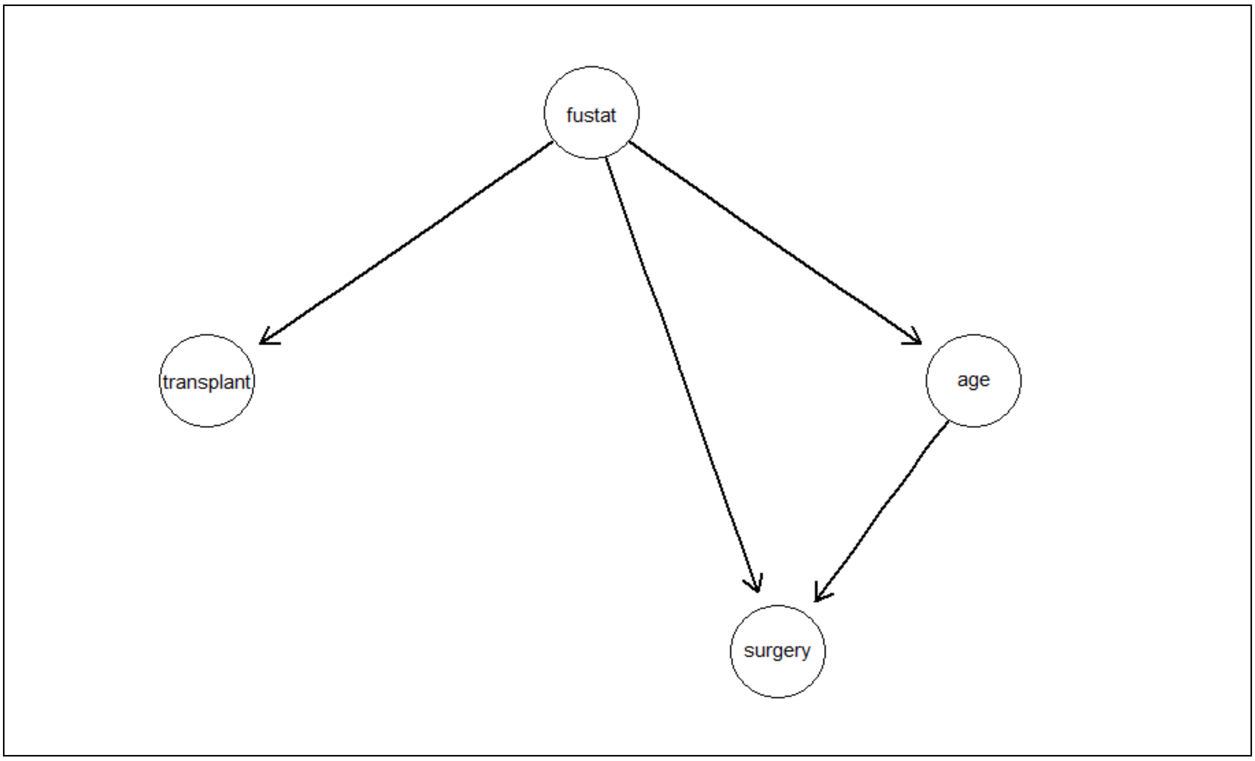


Fig 7.3 : Bayesian Network for 3 Key features(Transplant, Age,Surgery)

2. Using Algorithm : mmhc

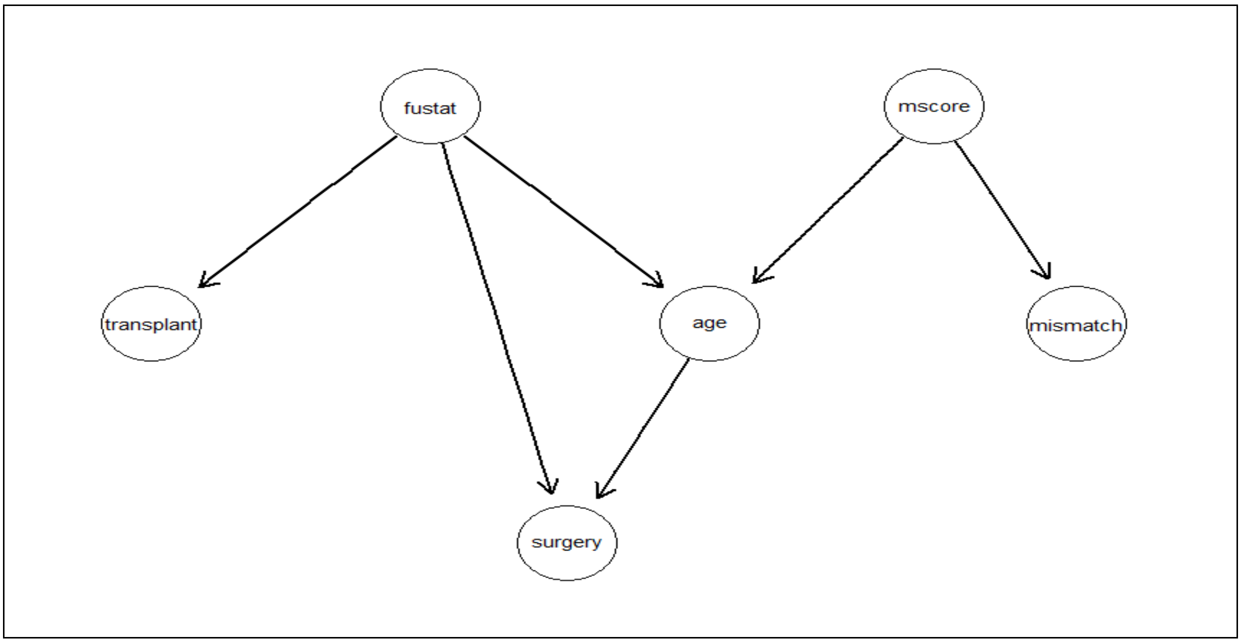


Fig 7.4 Bayesian network using mmhc algorithm

3. With bootstrapping with weight integration

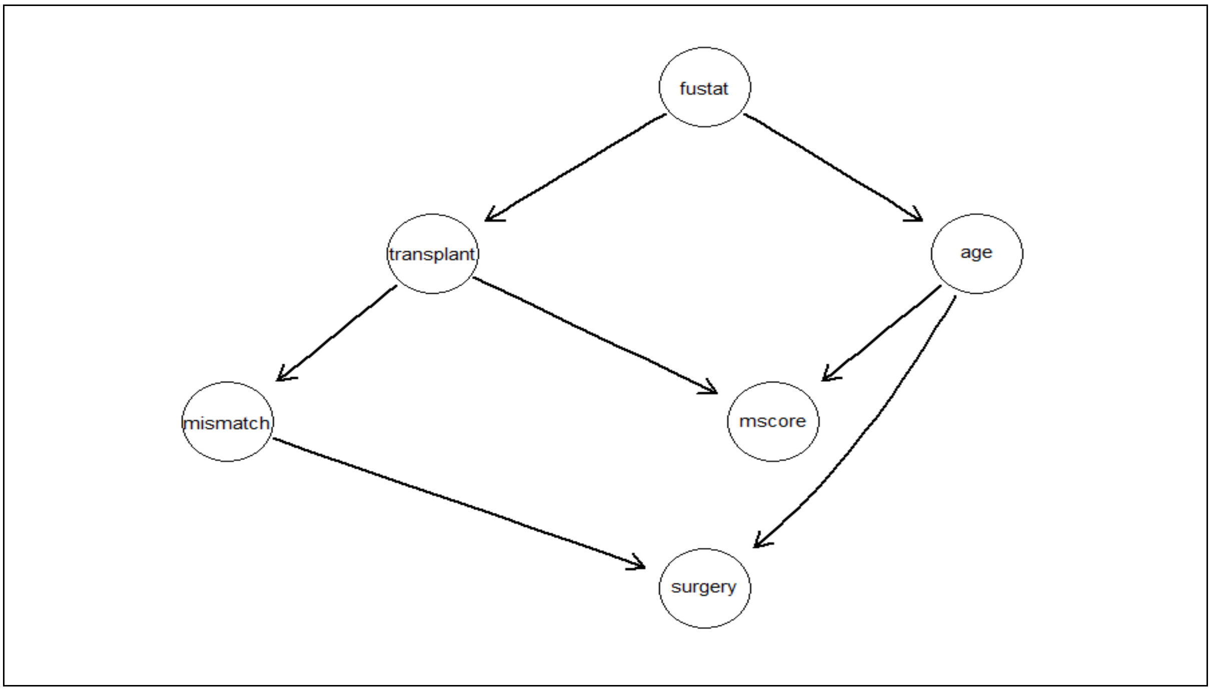


Fig 7.5 Bayesian Network using bootstrapping

4.Without weight Integration:

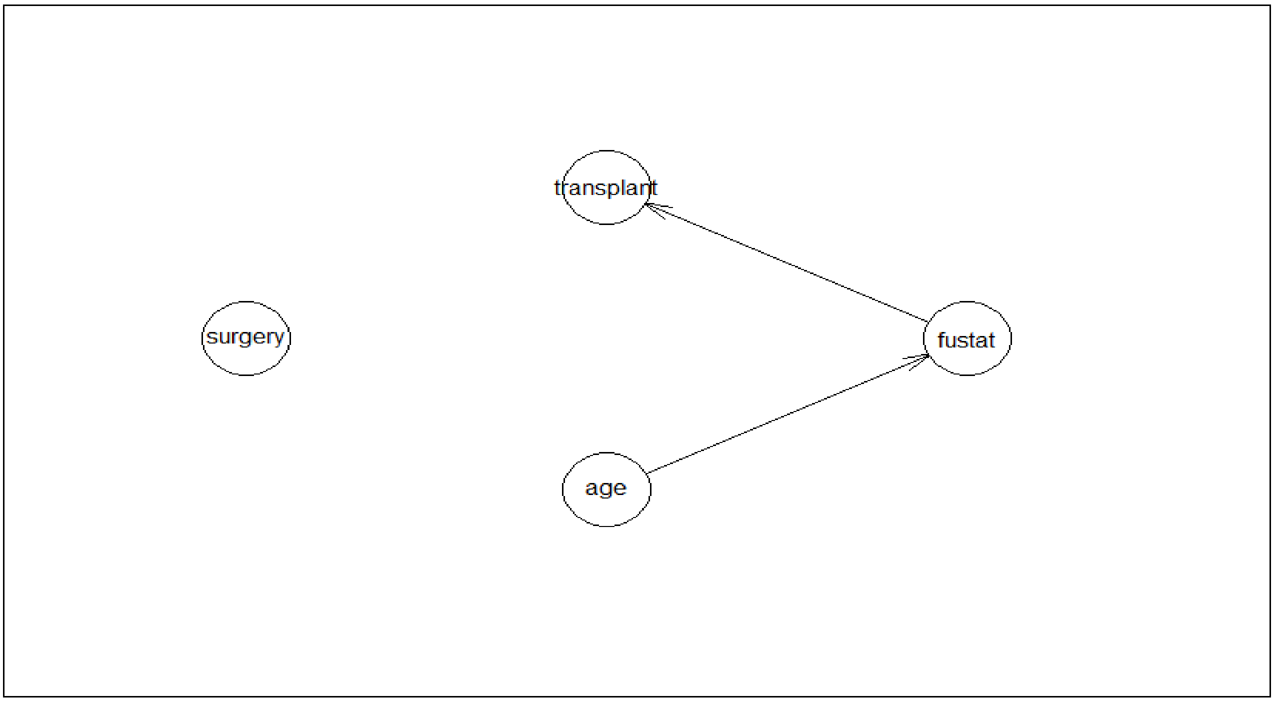


Fig 7.6 Bayesian Network without using weight integration

7.2 Performance Evaluation

As we have designed a Bayesian network with learned parameters. Also, we have interpreted the causal dependency between the covariates based on different algorithm. We have tried to the risk of getting overfitted using scoring function “BIC”. In this section, we will evaluate the performance of the proposed models measured using following metrics:

1. **Accuracy** of the model is one of the key parameters to evaluate the model. However, in case of imbalanced dataset where presence of any one class(majority) is very high as compared to minority class. Hence, we have assigned weights to the observation and try to nullify the effect of imbalance. So instead of plain Accuracy we have measured weighted classification accuracy which is expressed in the percentage of subjects in the test set.
   Weighted Classification Accuracy: The accuracy of the trained model with weights is 70.3.

2. Another metric is **F-measure** which is defined as a harmonic mean of precision and recall. A high value of F-measure indicates that both precision and recall are reasonably high.
   F − measure = 2 × Precision × Recall / (Recall + Precision)
   F1 score for event status, 0 is 0.61
   F1 score for event status, 1 is 0.76

3. **Classification Report**

	precision	recall	f1-score	support
Fustat 0	0.87	0.48	0.61	82
Fustat 1	0.64	0.93	0.76	83
accuracy			0.7	165
macroavg	0.75	0.7	0.69	165
weightedavg	0.75	0.7	0.69	165

Table 7.5 Evaluation Metrics for BN model

We can see from above table with multiple metrics for each class of events. Overall accuracy shows 70%. It looks good as compared to sample size and presence of imbalance nature present between the classes.

4. **Confusion Matrix**: This is a very strong table which contains lots of information about how the model is predicting against the actuals.
In below confusion matrix, our BN model with weighted data is predicting 77 true positive and 39 false positive. However, there are 49 cases where model is not predicting accurately which is required to be analysed.
Intuitively, at 6 instances where our model is predicting is to be present when its not.

CM		Predicted	
		0	1
ACTUAL 1	0	39	6
	1	43	77

Table 7.6: Confusion Matrix for Predicted result on test data

7.3 Overall Results

We can draw below conclusion from the results and evaluations:

1. We have successfully implemented the Bayesian Network model using IPCW method for censored survival data. We have generated multiple networks using different framework and discussed the conditional probability and structure generated in the process.

2. As few covariates in the data has large number of null/NA values. To overcome this issue we have tried imputation methods and bootstrapping for resampling. We have also tried to overcome the chances of overfitting using scoring function “BIC” which uses penalty factor to regularize the function.

3. We were able to develop a Bayesian model as a classifier to predict the risk of the event. Also, evaluated the performance of the model using different metrics and achieved to have 70% accuracy when predicting the results with test set.

4. We can further analyse to tune the model and also can check the impact of imbalanced data set .

Chapter 8

Discussion & Future scope

8.1 Conclusion and Discussion

The clinical health records have crucial information where large number of patients seeking care and hence such efficient risk prediction mechanism is demand of time. However, such sources contain lots of data infrequency as many patients may be lost to follow up after enrolling into the system. Most of the conventional machine learning techniques doesn't take observation time into the account and could not predict the event in future time frame. This paper has proposed a general purpose "Inverse probability of censoring weights "a weight assigning technique to the observation based on its survival probability. This weight allocation seems to improve the prediction probability when incorporated with Bayesian network. Using IPCW along with Bayesian network model leading the weighted accuracy to 70+ percentage.

This approach is a multipurpose with easy computations and can be used directly with advanced classification or regression technique like Neural network or other statistical models as well. However, we have not implemented it with other models but the mathematics behind it is simple and easily approachable which we have shown in this paper. Also, the principles outlined this algorithm is such that, it can be adapted widely on various tools. Because of its approach and mechanism, it is very much possible to develop it with ensemble – based model for risk predictions to apply to censored survival data.

This extended Bayesian Network incorporate with IPCW is easy to implement and have better approximation with other machine learning techniques. We can do a comparative analysis using other machine learning technique as future scope.

Finally, a properly treated censored outcomes with the help of IPCW emphasises to acknowledge that if all the patients had complete follow up information or not. This technique gives us the opportunity to include all the records even though having partial information so as to not leading to inaccurate predictions which is very critical in field of healthcare, as we cannot afford large false negatives.

8.2 Advantages

This paper focused on the application of a machine learning approach to risk prediction using heart transplant data with censored information due to unequal follow up of each subject. The traditional statistical models for censored survival data are well developed within the observation time but they are less flexible than efficient machine learning classifiers. However, machine learning classification techniques have its limitation in handling censored data as it assumes labels in the training data should be fully observed. The method which we have applied in this paper combines both features from statistical modelling to handle censored data and classifier techniques for prediction and flexibility by using inverse probability weighting to extend the Bayesian network technique for censored event data. Although we apply our approach to a Bayesian network, IPCW can be extended to other machine learning classifiers. Moreover, we have implemented the Bayesian network model using weights as a parameter to predict the event status without letting network to overfit using proper scoring functions.

Other than providing modelling flexibility and statistical validity, this technique handled the missing data efficiently and provides the opportunity to get the insights in sometime very complex clinical data without a domain expert as well.

From the entire process we can suggest that:

1. Excluding censored observations from the study results in very poor classification results and affects the relationship between the dependent covariates.

2. From the different studies, Bayesian network performs better in learning non-linear relationship between the features and outcome result compared to conventional hazards regression models while dealing with censored data.

3. We also suggest to choose proper evaluation metrics while assessing the performance of the model in censored data as some common metrics can give the misleading results because of imbalanced dataset.

8.3 Limitations

Although this method is very intuitive in nature, ease with usability and can handle censored data still there are some shortcomings in it. As we noted from the beginning, there are some assumptions on which this technique has been based on. The applicability of IPCW (inverse probability of censoring weighting) assumes that the censoring time is independent of the event time. Additionally, we assume heuristically, that patients more likely to have an event are not more or less likely to leave from the study group. We could relax this assumption considering the censoring time related to risk as a function of it. Though we have multiple techniques used for searching across the different Bayesian network structures still, we focussed on methods for learning the parameters for a given network structure. The reason behind this was the Bayesian Network are very much prone to get overfitted so to avoid it we chose for it. And we are handling the risk of overfitting by using scoring function "BIC" and bootstrapping. Though, this method is not sensitive to tuning parameters and number of bootstrap samples if set within nominal ranges.

Additionally, major drawback in IPCW is assigning weights to zero to those subjects for whom event status are unknown and hence, this method can be inefficient as these subjects do not contribute directly to the estimation. However, they do have a role in computation of weights.

8.4 Future Scope

Our work was completed with a reference of heart transplant data, still the technique which we have approached is globally applicable to all those scenario of clinical censored survival data and even beyond the field of medicine as well. In case of prediction of machine reliability system or in economics where we want to estimate the probability of re-hiring of recently unemployed candidates within a fixed time interval. Also, it can be applied in the scenarios where an result is likely to be censored. We can explore this method to be integrated with other machine learning techniques and statistical models also where target is to estimate the risk prediction for subjects with medical history given.

By applying some tuning parameters and with different dataset to improve the overall performance of the model and also comparative analysis can be done in order to compare the performance of the technique with other models as well.

We can even apply Weibull and Log-Logistic distributions to verify the results.

The IPCW technique has a limitation that it assigns zero weights where the even status is unknown. To overcome this, we can try positive and negative instances to each observation and get the weights for each observation.

As suggested by Kraisangka, Jidapa, and Marek J Druzdzel [48] , the alternative approach to use BN-Cox which we can be applied in the light of Cox Haphazard model with BN to get the best of both worlds in terms of getting the best statistical approach to handle the survival data by Cox and classifier flexibility with BN.

References

1. Ivan Štajduhar, Bojana Dalbelo-Bašić, Nikola Bogunović,
   Impact of censoring on learning Bayesian networks in survival modelling,
   Artificial Intelligence in Medicine, Volume 47, Issue 3,2009,Pages 199-217,

2. J. Pearl Probabilistic reasoning in intelligent systems: networks of plausible inference Morgan Kaufmann, San Francisco, CA, USA (1988)
3. Mahtab Jahanbani Fard “Bayesian Approach For Early Stage Event Prediction In Survival Data” , Wayne State University , 1-1-2015
4. Bandyopadhyay et al. [18] Under 4 77. Robins and Finkelstein 2000; Bang and Tsiatis 2000, 2002; Rotnitzky and Robins 2004; Tsiatis 2006

5. IvanŠtajduhara et al štajduhar I, Dalbelo-Bašić B. Learning Bayesian networks from survival data using weighting censored instances. *Journal of Biomedical Informatics.* 2010.

6. Vock, David M et al. “Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting.” Journal of biomedical informatics vol. 61 (2016): 119-31.

7. Miller, R.G., Halpern, J.: Regression with Censored Data. Biometrika Trust 69(3), 521–531 (1982)
8. Lee, E.T., Wang, J.: Statistical methods for survival data analysis, vol. 476. John Wiley & Sons (2003)
9.] Miller, R.G.: Least squares Regression with Censored Data. Biometrics Trust 63(3), 449–464 (1976)
10. Buckley, J., James, I.: Linear Regression with Censored Data. Biometrics Trust 66(3), 429–436 (1979)
11. Cox, D.R.: Regression Models and Life-Tables. Journal of the Royal Statistical Society 34(2), 187–220 (1972)
12. Koul, H., Susarla, V., Van Ryzin, J.: Regression Analysis with Randomly RigheCensored. The Annals of Statistics 9(6), 1276–1288 (1981)
13. Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine. 2004;30(3):201–214
14. Chapelle, O., Sch¨olkopf, B., Zien, A., et al.: Semi-supervised learning, vol. 2. MIT press Cambridge (2006)
15. Zhou, Z.H., Li, M.: Semi-supervised regression with co-training. In: IJCAI, pp. 908–916 (2005)
    [https://digitalcommons.unf.edu/cgi/viewcontent.cgi?article=1149&context=etd
16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/]
    17 – 16
17. 3. Altman DG. London (UK): Chapman and Hall; 1992. Analysis of Survival times.In:Practical statistics for Medical research; pp. 365–93.
18. 2. Berwick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Crit Care. 2004;8:389–94.

19. Kitty J. Jager, Paul C. van Dijk, Carmine Zoccali, Friedo W. Dekker,

The analysis of survival data: the Kaplan–Meier method, Kidney International,Volume 74, Issue 5,2008,

20. Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine. 2004;30(3):201–214

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

22. Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. Clinical Applications of Artificial Neural Networks. 2001:237–255

23. Shivaswamy PK, Chu W, Jansche M. A support vector approach to censored targets. Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007.; IEEE; pp. 655–660. 2007.

24. Khan FM, Zubek VB. Support vector regression for censored data (SVRc): a novel tool for survival analysis. Eighth IEEE International Conference on Data Mining (ICDM 2008), IEEE.2008. pp. 863–868

25. Sierra B, Larranaga P. Predicting survival in malignant skill melanoma using Bayesian networks automatically induced by genetic algorithms: An empirical comparison between different approaches. Artificial Intelligence in Medicine. 1998;14(1-2):215–230

26. Blanco R, Inza I, Merino M, Quiroga J, Larrañaga P. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. Journal of Biomedical Informatics. 2005;38(5):376–388.

27. Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. Computers and Biomedical Research. 1998;31(5):363–373.
28. štajduhar I, Dalbelo-Bašić B, Bogunović N. Impact of censoring on learning Bayesian networks in survival modelling. *Artificial Intelligence in Medicine.* 2009;47(3):199–217.

Under 1
29. N. Hoot, D. Aronsky
Using Bayesian networks to predict survival of liver transplant patients
C.P. Friedman, J. Ash, P. Tarczy-Hornoch (Eds.), AMIA annual symposium proceedings, vol. 2005, American Medical Informatics Association, Bethesda, MD, USA (2005), p. 345

30. G.F. Cooper, E. Herskovits
A Bayesian method for the induction of probabilistic networks from data

Mach Learn, 9 (4) (1992), pp. 309-347

31. W. Lam, F. Bacchus
Learning Bayesian belief networks: an approach based on the MDL principle
Comput Intell, 10 (4) (1994), pp. 269-293

32. D. Heckerman, D. Geiger, D.M. Chickering
Learning Bayesian networks: the combination of knowledge and statistical data
Mach Learn, 20 (3) (1995), pp. 197-243

33. N. Friedman
The Bayesian structural EM algorithm
G.F. Cooper, S. Moral (Eds.), Proceedings of the 14th annual conference on uncertainty in artificial intelligence, Morgan Kaufmann, San Fransisco, CA, USA (1998), pp. 129-138

34. Van Belle, V., Pelckmans, K., Van Huffel, S., Suykens, J.a.K.: Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial intelligence in medicine 53(2), 107–18 (2011)

35. Bennett, K.P., Demiriz, A.: Semi-supervised support vector machines. In: Advances in Neural Information Processing Systems, pp. 368–374. MIT Press (1998)

36. Cordon-cardo, C., Kotsianti, A., Verbel, D.A., et. al.: Improved prediction of prostate cancer recurrence through systems pathology. Journal of clinical investigation 117(7), 1876–1883 (2007).

37. Evers, L., Messow, C.M.: Sparse kernel methods for high-dimensional survival data. Bioinformatics (Oxford, England) 24(14), 1632–8 (2008)

38. Shiao, H.T., Cherkassky, V.: Learning using privileged information (LUPI) for modeling survival data. 2014 International Joint Conference on Neural Networks (IJCNN) pp. 1042–1049 (2014)

39. Robins and Finkelstein 2000; Bang and Tsiatis 2000, 2002;

40. Rotnitzky and Robins 2004; Tsiatis 2006

41. Dempster, A. P., et al. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, 1977, pp. 1–38. JSTOR, www.jstor.org/stable/2984875. Accessed 1 Sept. 2021.

42. Liao, W. and Ji, Q., 2009. Learning Bayesian network parameters under incomplete data with domain knowledge. Pattern Recognition, 42(11), pp.3046-3056. https://www.ecse.rpi.edu/~qji/Papers/PR_BNlearning_revision_v2.pdf

43. S. Geman and D. Geman, “Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, pp. 721–741, 1984

44. W. L. Buntine, “Operations for learning with graphical models,” Artificial Intelligence Research, vol. 2, pp. 159–225, 1994.

45. Zupan B, Demsar J, Kattan MW, Beck JR, Bratko I (1999) Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif Intell Med 1620:346–355.

46. Sierra B, Larranaga P. Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches. Artif Intell Med 1998;14(1–2):215–30.

47. Marshall A, McClean S, Shapcott M, Millard P. Learning dynamic Bayesian belief networks using conditional phase-type distributions. Lect Notes Comput Sci 2000:516–23.

48. Kraisangka, Jidapa, and Marek J Druzdzel. “A Bayesian Network Interpretation of the Cox's Proportional Hazard Model.” International journal of approximate reasoning : official publication of the North American Fuzzy Information Processing Society vol. 103 (2018): 195-211.

49. Stajduhar I, Dalbelo-Basic B (2012) Uncensoring censored data for machine learning: a likelihood-based approach. Expert Syst Appl 39(8):7226–7234

50. Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-based residuals for survival models. Biometrika
77(1):147–160

51. KattanMW, Hess KR, Beck JR (1998) Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression.
Comput Biomed Res 31(5):363–373

52. Kattan MW (2003) Comparison of Cox regression with other methods for determining prediction models and nomograms. J Urol 170(6):S6–S9

53. Data reference :
References
Turnbull B, Brown B, and Hu M (1974). "Survivorship of heart transplant data." Journal of the American Statistical Association, vol. 69, pp. 74-80.

54. Moore, Dirk F. Applied Survival Analysis Using R. Springer Science+Business Media, 2016.

55. Sphweb.bumc.bu.edu. 2021. Survival Analysis. [online] Available at: <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival_print.html>

56. D.G. Kleinbaum
Survival analysis: a self-learning text
(2nd ed.), Springer-Verlag, New York, NY, USA (2005)

57. https://core.ac.uk/download/pdf/14916911.pdf

58. S. Wright. Correlation and causation. Journal of Agricultural Research, 20:557–585, 1921.

59. S. Wright. The method of path coefficients. Annals of Mathematical Statistics, 5:161–215, 1934.

60. G.F. Cooper, E. Herskovits
A Bayesian method for the induction of probabilistic networks from data
Mach Learn, 9 (4) (1992), pp. 309-347

61. https://www.cs.cmu.edu/~dmarg/Papers/PhD-Thesis-Margaritis.pdf

62. S.J. Russell, P. Norvig Artificial intelligence: a modern approach (2nd ed.), Prentice Hall, Upper Saddle River, NJ, USA (2002)

63. Friedman, N. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 125–133.

64. Ruggieri, A., Stranieri, F., Stella, F. and Scutari, M., 2020. Hard and Soft EM in Bayesian Network Learning from Incomplete Data. Algorithms, 13(12), p.329.

65. Robins JM, FinkelsteinDM(2000) Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 56(3):779– 788

66. Rotnitzky AG, Robins JM (2004) Inverse probability weighted estimation in survival analysis. In: Armitage P, Colton T (eds) The encyclopedia of biostatistics, 2nd edn. Wiley, Hoboken, NJ

67. Tsiatis AA (2006) Semiparametric theory and missing data. Springer, New York

68. H Bang, AA Tsiatis, Estimating medical costs with censored data, Biometrika, Volume 87, Issue 2, June 2000, Pages 329–343,

69. Kalbfleisch, J.D. and Prentice, R.L. (2002) The Statistical Analysis of Failure Time Data. 2nd Edition, John Wiley and Sons, New York.

APPENDIX

1. Data Analysis , Survival analysis and Survival Plots are generated in R scripts
2. IPCW mechanism has been developed in R using Surv function
3. Generated weights has been integrated with BN in python