Project 2: Survival Analysis

Decision support systems

by group 9:

Daniel Christopher Biørrith, 201909298

Jeppe Stjernholm Schildt, 201905520

Aarhus University
Department of Electrical and Computer Engineering

August 21, 2024

Contents

1 Introduction

In an era marked by the increasing importance of data-driven decision-making, survival analysis has emerged as a critical statistical tool for a wide range of applications across various fields. This report offers an introduction to survival analysis, an analytical method that examines and models the time-to-event data, particularly focusing on the occurrence of events and the time at which they transpire. Survival analysis has found its use in various disciplines such as medical research, engineering, social sciences, and economics, where understanding the relationship between time and event occurrence is of paramount importance. This report will explain the key concepts, methodologies, and applications of survival analysis, showcasing the results of key algorithms. By providing a solid foundation in survival analysis, this report aims to present readers with the necessary knowledge to interpret, conduct, and evaluate time-to-event studies.

2 Methods and materials

This section will present some key methods and terms in the field of survival analysis.

2.1 Survival function

A key part of survival analysis is the survival function. It is the probability that a lifetime exceeds a given time t, which is simply the complement of the cumulative distribution function, $F(T)$, which mathematically is:

$$S(t) = 1 - F(t) = Prob(lifetime > t)$$

Or in integral terms, defined with the probability density function (PDF) $f(x)$:

$$\int_t^\infty f(x)dx$$

We can never find out the true survival function in real-life cases, yet different estimators are used to estimate it. The one we will use is the Kaplan-Meier estimator, which will be introduced later in this section.

2.2 Hazard function

The hazard function helps answer the question: *of the people who survived until t, what is the fraction that will 'die' at t?* It is the slope of the survival curve at a given time t and might be used if one wishes to find what the highest and lowest hazard rates are and to see when the rates are increasing and decreasing. It is given by [1]:

$$HF(t) = \lim_{dt \to 0} \frac{S(t) - \frac{S(t+dt)}{dt}}{S(t)}$$

In the function, the top part of the fraction represents the PDF in an infinitesimally small timestep ($\int_t^{t+dt} f(x)\,dx$), called the instantaneous rate of event, and the bottom of the fraction is simply the survival function. With this, the hazard function is a measure of risk, where a greater value means a greater risk of failure (an event occurring).

Hazard ratio compares the rate of two different treatments, where a hazard ratio close to 1 means they are close, whereas a hazard ratio of 2, for instance, means one group's rate of 'death' is twice as fast as the other group.

2.3 Kaplan Meier

The Kaplan-Meier estimator is a great tool to estimate the survival function $S(t)$ from lifetime data. It is a great tool for estimating the fraction of events that occur after a given time t (how many will 'survive' until after time t)[2][3].

The Kaplan-Meier estimator may be particularly useful in analyzing time-to-event data, where the event of interest (e.g., death, failure, or recovery) may not have occurred for all subjects during the study period. This is referred to as censored data. The Kaplan-Meier estimator takes both uncensored (event occurred) and censored (event not observed) data into account, allowing researchers to estimate the survival function even when not all subjects experience the event.

The Kaplan-Meier estimator is calculated as follows:

$$S(t) = \prod_{(i:t_i \leq t)} (1 - \frac{d_i}{n_i})$$

Where:

- t_i represents the distinct time points at which at least one event occurred.
- n_i is the number of individuals at risk (i.e., those who have not yet experienced the event or been censored) just before time t_i
- d_i is the number of events occurring at time t_i

The Kaplan-Meier estimator calculates the survival probability at each time point t_i by multiplying the conditional probabilities of surviving beyond each previous time point. This results in a step function that provides a visual representation of the survival function over time, known as the Kaplan-Meier survival curve.

The Kaplan-Meier survival curve allows researchers to compare the survival experiences of different groups or treatment arms in a study. It's important to note that the estimator does not provide information about the factors influencing survival; to investigate this, other methods such as the Cox proportional hazards model can be used.

2.4 Cox Proportional

The Cox proportional hazards model, also known as the Cox regression model or simply the Cox model, is a widely used statistical method in survival analysis for investigating the effects of multiple predictor variables on the time-to-event data.

The model derives robust estimates of covariate effects using a hazard function, which represents the instantaneous risk of an event (e.g., death, failure, or recovery) occurring at a given time, conditional on the individual's survival up to that time.[4]

The model is a semi-parametric method, meaning that it makes certain assumptions about the relationship between the predictor variables and the hazard function, but it does not assume a specific parametric form for the baseline hazard function. This flexibility makes it particularly useful for analyzing complex survival data with various underlying distributions.

The key assumption of the model is the proportionality of hazards, which means that the hazard ratios are constant over time. Mathematically, the model can be expressed as follows:

$$h(t, X) = h_0(t) \cdot exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Where:

- $h(t, X)$ is the hazard function at time t for an individual with predictor variables $X = (X_1, X_2, ..., X_p)$
- $h_0(t)$ is the baseline hazard function at time t, representing the hazard for an individual with all predictor variables equal to zero
- $\beta_1, \beta_2, ..., \beta_p$ are the regression coefficients associated with the predictor variables $(X_1, X_2, ..., X_p)$, estimated from the data
- $exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$ is the hazard ratio, representing the effect of the predictor variables on the hazard function

The Cox model can be used to estimate the hazard ratios and their confidence intervals, as well as test the significance of the predictor variables in explaining the survival data. The results of the Cox model can provide valuable insights into the factors influencing the time-to-event and inform decision-making in various fields, including medical research, engineering, and social sciences.

2.5 Log Rank

The log-rank test is a non-parametric statistical hypothesis test used to compare the survival distributions of two or more independent samples. It focuses on the time until an event happens, such as an occurrence of a disease or the failure of a mechanical component. The purpose of this test is most often used to determine the efficacy of a given treatment or modification by comparing it with a control measurement. Examples include evaluating a medical treatment or the strengthening of a mechanical component.

Central to the log-rank test is the null hypothesis, which states that the hazard functions of the two measurements are the same. This hypothesis is tested using a statistical threshold, typically set at a p-value of less than 5%. If the hypothesis can't be confirmed to the level of the p-value, it will be rejected, which implies a difference between the two hazard functions. If this is the case, it suggests that the treatment or change may have had an effect on the event's timing, be it good or bad. It has the advantage that no need of knowledge regarding the shape of the survival curve of the distribution of survival times is needed [5].

3 Dataset and Exploratory Data Analysis (EDA)

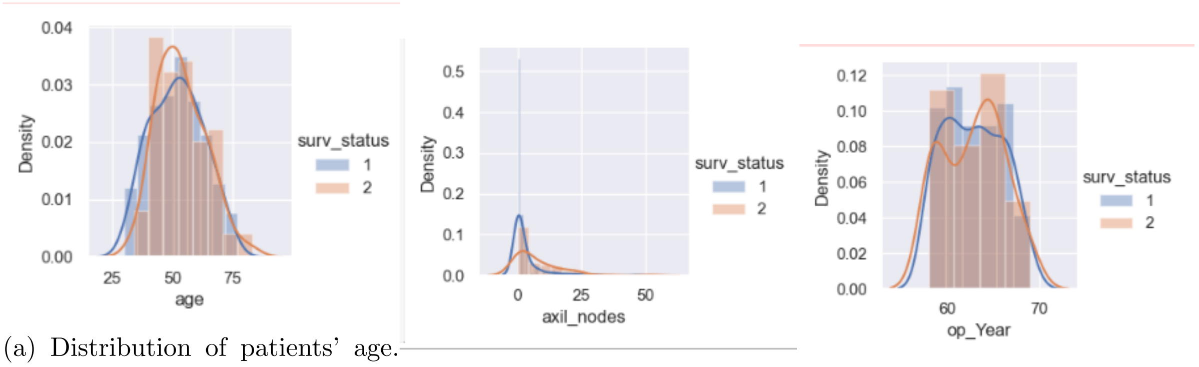
Our project examines the Habermann's Survival Dataset[6], which contains data from a 1958 to 1970 study on the survival of breast-cancer-operated patients. Initially, we examine the data and note the following important observations.

- There are no missing values in the dataset.
- The dataset is rather small with 306 observations.
- There is a rather large skew in survival vs death distribution, with about 73 % of patients still alive after 5 years.
- Survival status is in integer form: 1 is alive, 2 is deceased.

Subsequently, we performed EDA on the data-set, in an attempt to highlight important characteristics about the data. Below, in figure 1a, we see that the majority of patients unsurprisingly are elderly, with the density curve spiking higher around age 50 for survivors, and being a bit more distributed for the deceased participants.

We also see in figure 1b that the majority of survivors have 0 nodes, with deceased patients having a more distributed amount, maxing out at around 50.

Lastly, we see in figure 1c that year of operation is quite well distributed between (19)60 and (19)70, the years the data was collected. Testing how the year influenced the chance of survival is an obvious opportunity.



(a) Distribution of patients' age. A survival status of 1 means they lived 5 years or longer.

(b) Distribution of the number of positive axillary nodes detected.

(c) Distribution of patients' year of operation.

Figure 1: EDA Histograms

4 Implementation & Results

Throughout this project, any functions referenced, unless otherwise specified, is from the lifelines library, which provides a 'complete survival analysis library'[7].

4.1 Kaplan-Meier

Firstly, the Kaplan-Meier estimate was implemented. As parameters, it takes the duration function(x-axis) and the events observed. This can be seen in the following code snippet, which initializes the estimate, fits it, and plots the function:

```
km = KaplanMeierFitter()
km.fit(durations=df['age'],event_observed=df['surv_status'])
fig, ax =  plt.subplots()
km.plot_survival_function(color='C0',ax=ax)
ax.set(
    title='Kaplan-Meier survival curve',
    xlabel='Age',
    ylabel='Estimated Probability of Survival'
);
```

This gives us the image in figure 2:

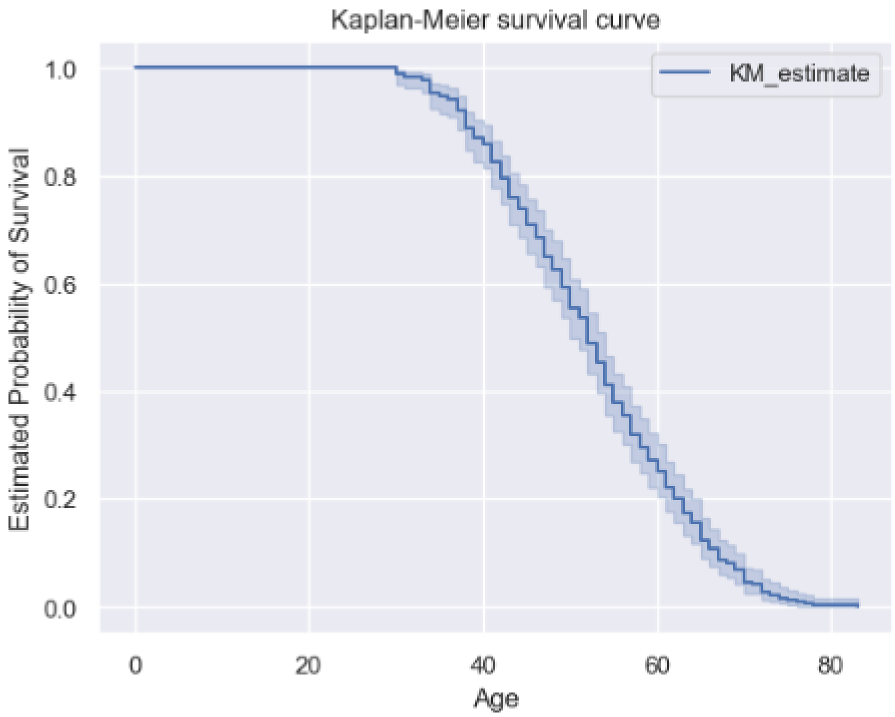


Figure 2: Kaplan-Meier curve w/ age

We subsequently fitted other KM estimates on the other variables, to gather information about the impact of those on the survivability.

Firstly, the impact of nodes may be seen in figure 3:

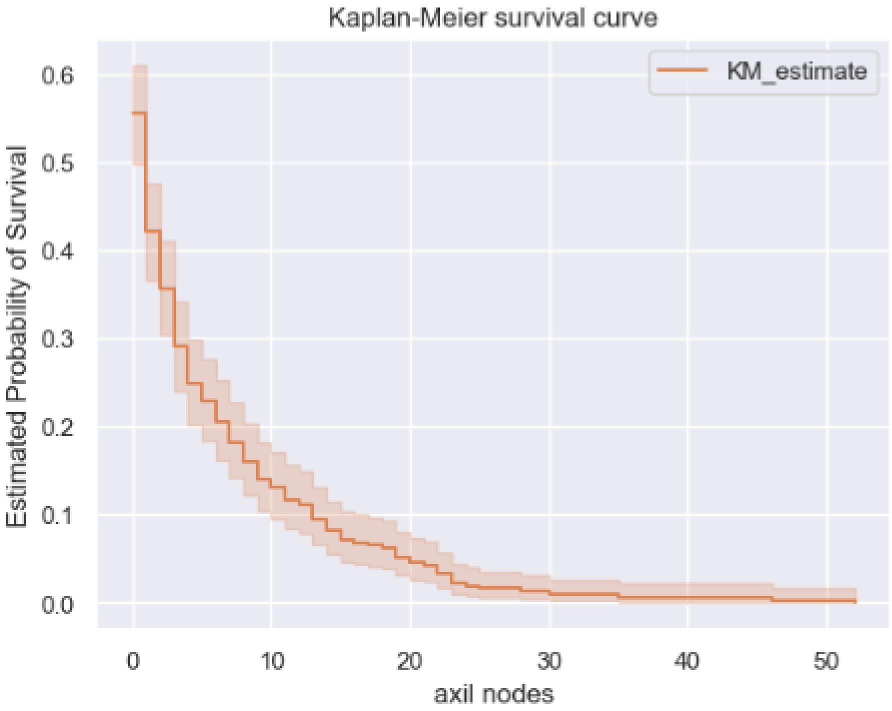


Figure 3: KM curve w/ nodes

From the figure its evident, and unsurprising, that the probability of surviving the 5 years the study ran negatively scales with the number of nodes. Having more than 25 positive axillary nodes seems to essentially be a death sentence.
Lastly, we created two groupings (cohorts) based on the amount of positive axillary nodes, to see their varying KM survival curves. This is shown below in figure 4.

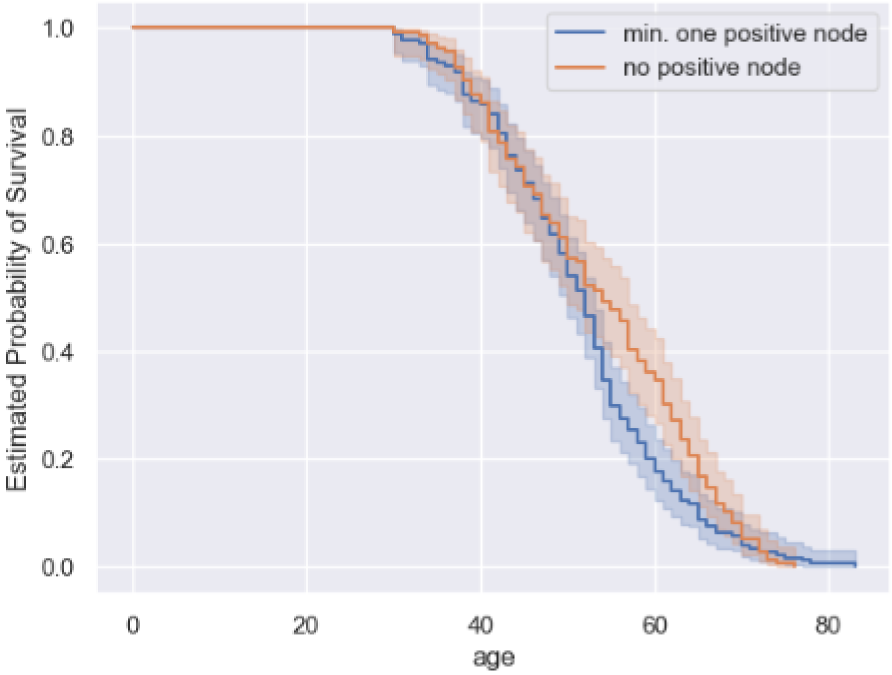


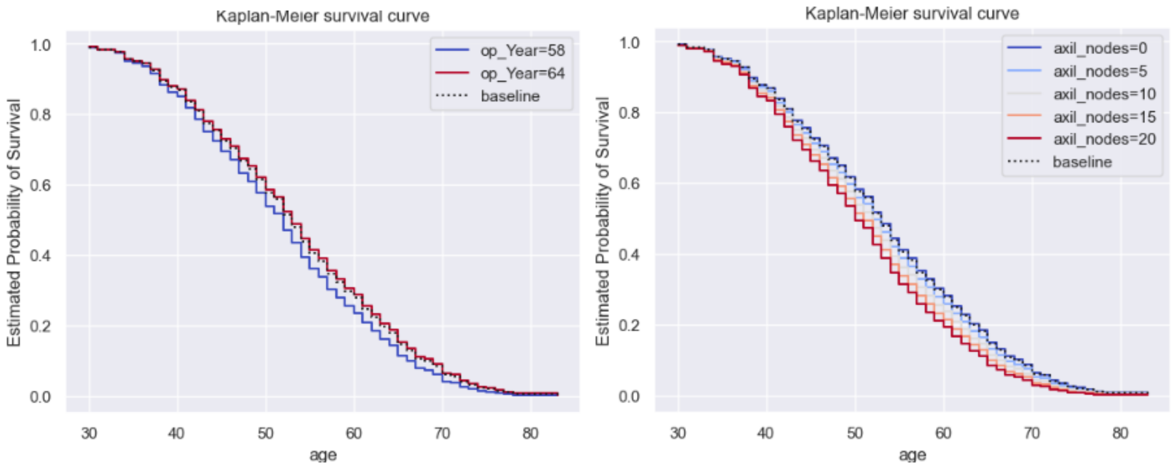
Figure 4: Kaplan-Meier curve of comparison between 2 cohorts

4.2 Cox Proportional Hazard

Next, we used the Cox Proportional Hazard function (CPH), which allows us to investigate multiple groupings in the data for their correlation of survivability. The implementation was quite simple, presented in the code-snippet below:

```
km = KaplanMeierFitter()
km.fit(durations=df['age'],event_observed=df['surv_status'])
fig, ax =  plt.subplots()
km.plot_survival_function(color='C0',ax=ax)
ax.set(
    title='Kaplan-Meier survival curve',
    xlabel='Age',
    ylabel='Estimated Probability of Survival'
);
```

Shown below, are the comparisons between the survival function when varying the covariates, using the `plot_partial_effects_on_outcome` function.

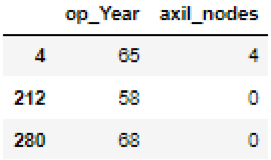


(a) Survival rate vs operation year

(b) Survival rate vs axillary nodes

Figure 5: Varying covariates impact on survival function

Lastly, three patients from the dataset were selected at random to see their individual survival functions based on the characteristics of their disease. These results are shown in figure 6.



	op_Year	axil_nodes
4	65	4
212	58	0
280	68	0

Figure 6: Characteristics of the three arbitrary patients.

A Kaplan-Meier Estimate was made for each of these, shown in figure 7.

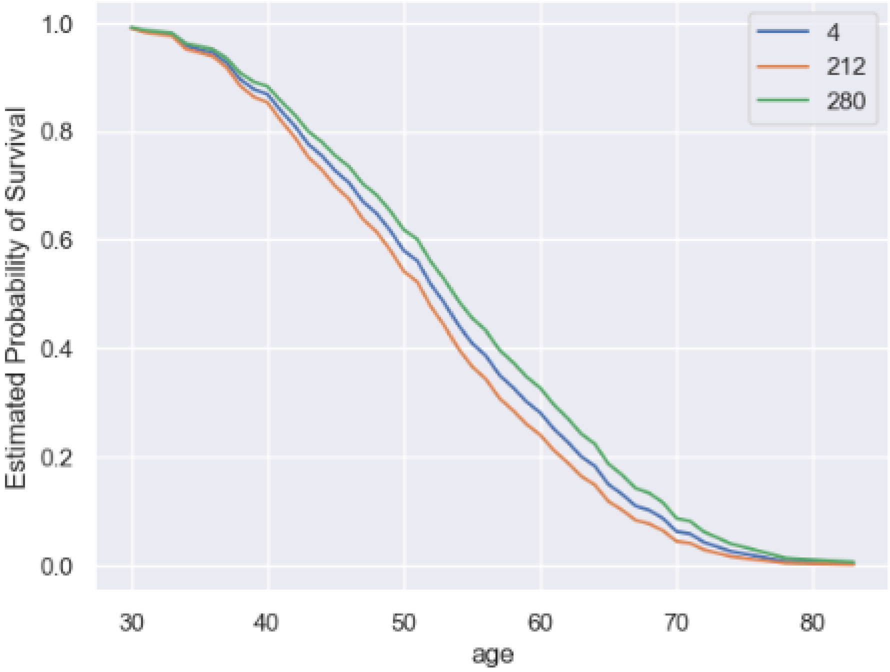
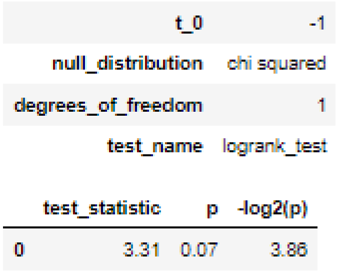


Figure 7: KM for the 3 patients

Here it's evident that patient 212 has the lowest estimated survival age while having no axillary nodes, which is likely due to his operation year being 1958, 7 years earlier than patient 4, who understandably has worse prospects than his counterpart patient 280, as patient 4 has four axillary nodes. This showcases a very tangible application for the KM estimate, with decent implicit interpretability.

4.3 Log Rank Test

Finally, we did a log-rank test, as to test the survival distribution of the group with no axillary nodes compared to a group where everyone has at least one axillary node. The hypothesis set up was that the two have the same distribution. The results are shown in figure 8.



t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	3.31	0.07	3.86

Figure 8: Log Rank Summary

From the figure, we can see that the p-value is 0.07, meaning 7%, thus resulting in the hypothesis being rejected. This implies that the two distribution curves are different, meaning people with at least one axillary node have a smaller survival chance (as expected).

5 Discussion

Survival analysis is a great statistical branch for inspecting trends and creating predictions about the time-to-event, with a very broad list of use cases. This report focused on the survival of breast cancer patients, based on their age, year of surgery, and amount of axillary (cancerous) nodes, including censored data. Our exploratory data analysis highlighted key features about the patients, and using the Kaplan-Meier estimate, as well as both The Cox Proportional Hazards model and a Log Rank test, we presented a large variety of information regarding the prediction of specific individuals. Other methods, such as using machine learning methods and Cox Regression methods, might have provided different insights. Alas, the scope of this project was not for it.

6 Conclusion

This report presented the key elements of Survival Analysis, an important statistical tool for time-to-event analysis, with a focus on the Habermann's dataset, a well-known educational collection of cancer patients from 1958 to 1970.
EDA was performed on this data, and subsequently it was examined using Kaplan-Meier estimates, the Cox Proportional Hazard function, and the Log Rank test. This showcased key features of the effect of variables on survivability. It would be interesting to see this Survival Analysis performed on a more recent dataset as well, to perhaps compare the differences in the medicinal industry over the past 50 years.

Bibliography

[1] S. Glen. "Hazard function: Simple definition." (), [Online]. Available: https://www.statisticshowto.com/hazard-function/ (visited on 02/25/2023).

[2] J. K. Manish Goel Pardeep Khanna. "Understanding survival analysis: Kaplan-meier estimate." (2010), [Online]. Available: https://www.researchgate.net/publication/50940632_Understanding_survival_analysis_Kaplan-Meier_estimate.

[3] E. Lewinson. "Introduction to survival analysis: The kaplan-meier estimator." (Sep. 17, 2020), [Online]. Available: https://towardsdatascience.com/introduction-to-survival-analysis-the-kaplan-meier-estimator-94ec5812a97a.

[4] V. S. Salil Deo Vaishali Deo. "Survival analysis-part 2: Cox proportional hazards model." (2021), [Online]. Available: doi:10.1007/s12055-020-01108-7.

[5] JM Bland and DG Altman. "The logrank test." (May 1, 2004), [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403858/.

[6] "Habermann's survival data set." (), [Online]. Available: https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set.

[7] "Lifelines library." (), [Online]. Available: https://lifelines.readthedocs.io/en/latest/.

[8] R. Raoniar. "Survival analysis in python (km estimate, cox-ph and aft model)." (Oct. 29, 2021), [Online]. Available: https://medium.com/the-researchers-guide/survival-analysis-in-python-km-estimate-cox-ph-and-aft-model-5533843c5d5d (visited on 02/25/2023).

[9] GOKUL. "Haberman's survival: Exploratory data analysis." (2018), [Online]. Available: https://www.kaggle.com/code/gokulkarthik/haberman-s-survival-exploratory-data-analysis.