JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 2, No 1, March 2023
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Survival Analysis of Students Not Graduated on Time Using Cox Proportional Hazard Regression Method and Random Survival Forest Method

Muhammad Arib Alwansyah [1*]

[1] S1 Statistics Study Program, Bengkulu University, Bengkulu

* Corresponding Author: muhammadaribalwansyah232@gmail.com

Article Info

Article History:
Received: October 23, 2022
Accepted: April 12 2023
Available Online: May 02, 2023

Key Words:
Survival Analysis
Cox Proportional Hazard
Random Survival Forest
Harrell's Concordance Index
Study Length.

Abstract

Higher education is a place to educate the next generation of the nation in terms of academic and non-academic. Basically every college tries to maximize the graduation of its students, both in quantity and quality. The undergraduate education program is targeted to complete 8 semesters of study or can also be taken in less than 8 semesters and a maximum of 14 semesters. Many factors are thought to affect the length of student study, both internal and external factors. Based on the factors that are thought to affect the length of study of the student, it is necessary to conduct research to determine what factors have a significant effect on the length of study of the student. The method that can be used to determine these factors is survival analysis using cox proportional hazard regression and random survival forest. Factors that affect the length of study using cox proportional hazard regression is GPA, while by using the random survival forest method, the factors that influence the length of study of students are GPA, gender, and part time. Based on the comparison using the C-Index method, random survival forest is a suitable method to use in the data because the C-Index error value is 26.9% which is smaller than the cox proportional hazard which is 27.8%.

1. INTRODUCTION

Higher education is a place to educate the next generation of the nation in terms of academic and non-academic. Every university tries to maximize the graduation of its students, both in quantity and quality. The undergraduate education program is targeted to complete 8 semesters of study or can also be taken in less than 8 semesters and a maximum of 14 semesters. Students are said to have graduated from college if they have completed all courses and academic programs required by each study program. The quality of graduates from universities can be influenced by several factors, both internal and external. These internal and external factors are thought to affect the length of student study in completing the education being pursued [1]. Therefore, researchers are interested in conducting research to determine the factors that affect the length of study for undergraduate students at FMIPA UNIB class 2017.

The analysis used to examine the factors that affect the length of study of students in this study is survival analysis. Survival analysis is a statistical method used when the data case is related to the time until a certain event occurs. This study used the Cox proportional hazard and random survival forest methods because they were able to handle censored data. Random survival forest is a collection of random tree methods used for right-censored survival time data. The joint incident in this study was 2 or more students who had a thesis trial in the same month. The data is said to be censored if the thesis trial student is more than 48 months. Therefore, the author wants to conduct a study entitled " Survival Analysis of Students Not Graduated on Time Using Cox Proportional Hazard Regression Method and Random Survival Forest Method ".

1.1 Survival Analysis

Survival analysis is a statistical method related to time, which starts from the time origin or start point to a special event (failure event or end point). Cox regression is a survival analysis used to analyze data with the dependent variable in the form of survival time. Survival time is the time from the start of the study to the time of the occurrence of an event or events [2].

1.1.1Censored Data

Data is said to be censored if the individual or observation has not experienced a certain event. If the individual experiences an event before the end of the observation, it is called uncensored data [3].

1.1.2 Opportunity Density Function

If T is a random variable of the lifetime of an individual in the interval $[0, \infty)$, then the probability density function is $f(t)$ and the cumulative distribution function is $F(t)$. The survival time T has a probability density function which is defined as the individual probability of failure in the time interval t to $t + \Delta t$ or the probability of failure in the interval per unit time. This can be expressed as [4].

$$f(t) = \frac{\lim\limits_{\Delta t \to 0} P[failure\ in\ (t, t + \Delta t)]}{\Delta t} = \frac{\lim\limits_{\Delta t \to 0} P(t < T < t + \Delta t)}{\Delta t}$$

While the cumulative distribution function is:

$$F(t) = P(T \leq t) = \int_0^t f(x)dx$$

$$f(t) \ = \frac{dF(t)}{dt} = F'(t)$$

1.1.3 Survival Function

If T is a random variable in the interval $[0, \infty)$ which indicates the time an individual experiences an event in the population, $f(t)$ is a function of the probability density of t, then the probability of an individual not experiencing an event until time t is expressed by the survival function $S(t)$ [4].

$$S(t) = P(T \geq t) = \int_t^\infty f(x)dx$$

The following is a survival function from the definition of the cumulative distribution function T

$$S(t) = P(T \geq t) = 1 - P(T \leq t)$$

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

The relationship between the probability density, the cumulative distribution function of T and the survival function is

$$f(t) = -\frac{d(S(t))}{dt} = -S'(t) \tag{1}$$

$$f(t) = F'(t) = -S'(t)$$

1.1.4 Hazard Function

Suppose T is a random variable in the interval $[0, \infty)$ which indicates the time an individual experiences an event in a population, then the probability that an individual experiences an event in the interval $(t, t + \Delta t)$ is expressed by the hazard function $h(t)$ [4].

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t . S(t)}$$

$$h(t) \ = \frac{f(t)}{S(t)} \tag{2}$$

1.1.5 Cumulative Hazard Function

The following is the result of substituting Equation (1) into Equation (2) [4]:

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

$$S(t) = e^{\left[-\int_0^t h(x)dx\right]} = e^{[-H(t)]}$$

1.2 Cox Proportional Hazards Regression

Cox proportional hazards regression or known as the cox regression model is used to determine the relationship between the dependent variable and the independent variable, where the data used in the cox proportional hazards regression is data on the survival time of an individual

Cox proportional hazards regression model is as follows [5]:

$$h(t, X) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p\right) = h_0(t)e^{\sum_{j=1}^{p} \beta_j X_j} \quad (3)$$

1.2.1 Shared Event Data

Occurrence data are often found in survival analyses. A joint event is an event where two or more individuals experience an event at the same time.

1.2.2 CPH Model Parameter Estimation for occurrence

The alternative methods offered by [5] to handle co-occurrence data are the *Breslow* partial likelihood method, the *Efron* partial likelihood method, and the *Exact* partial likelihood method. The following is the partial likelihood equation for the *Breslow* method:

$$L(\beta)_{Breslow} = \prod_{i=1}^{r} \frac{e^{\left(\sum_{j=1}^{p} \beta_j s_k\right)}}{\left(\sum_{i \in R(t_j)} e^{\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}\right)^{d_i}}$$

1.2.3 Parameter Test

There are three ways to test the significance of the parameters, namely the partial likelihood ratio test, the Score test, and the Wald test. Parameter significance testing aims to check whether the independent variables have a real influence in the model. The simultaneous test in this study used a partial likelihood ratio test, while the partial test used a score test [6].

Simultaneous test steps using the partial likelihood ratio test:

1. Hypothesis

   $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

   H_1: there must be at least one $\beta_j \neq 0, j = 1, 2, 3, \ldots, p$

2. Significance level: alpha value (α)
3. Test statistics:

$$G = -2[\ln L_R - ln\, L_F]$$

4. Rejection area:

   Reject H_0if $G \geq \chi^2_{(\alpha:db=p)}$ or $p - value < \alpha$

5. Conclusion:

   Reject H_0if $G \geq \chi^2_{(\alpha:db=p)}$ or $p - value < \alpha$, meaning that there is at least one independent variable that has an effect on the model.

The test statistic on the test score follows a chi-square distribution with degrees of freedom p. Here are the steps for the score test:

1. Hypothesis:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0, j = 1, 2, \dots , p$

2. Significance level: (α)
3. Test statistics:

$$z = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)$$

4. Rejection area:
   H_0 is rejected if $\geq \chi^2_{(\alpha:db=p)}$ or $p - value \leq \alpha$
5. Conclusion:
   Reject H_0 if $z \geq \chi^2_{(\alpha:db=p)}$ or $p - value < \alpha$, it means that the variable X_j has an effect on the model.

1.2.3 Assumptions of the Cox Proportional Hazard Model

There are two ways to test the proportional hazard assumption in a Cox proportional hazard model. The two methods are a graphical approach using a log-minus-log survival plot and using a Schoenfeld residual plot. Schoenfeld residuals are defined as residuals in which each individual and each independent variable is based on the first derivative of the log likelihood function [3]. The Schoenfeld residual for the i-th individual on the j-th independent variable is as follows:

$$R_{ji} = \delta_i \left(X_{ji} - \frac{\sum_{l \in R(t_j)} X_{jl} e^{(\widehat{\beta}' X_l)}}{\sum_{l \in R(t_j)} e^{(\widehat{\beta}' X_l)}} \right), j = 1,2, \dots , p$$

1.3 Random Survival Forests

Random Survival Forest is a machine learning method in survival analysis that can be used to make predictions involving many independent variables and is also used for large amounts of data. Random survival forest itself is a collection of random tree methods used for right-censored survival data. This method only relies on data and does not rely on model assumptions so that it is considered a method that can predict survival and selection of variables better [7].

1.3.1 *Splitting*

A log-rank split rule that divides the vertices by maximizing the log-rank test statistic. Suppose at a node h in the process of tree formation we want to split into two child nodes. Suppose that at node h there are n observations with survival times along with censorship indicators $(T_1, \delta_1), \dots , (T_n, \delta_n)$ where observation i is said to be censored at time T_i if $\delta_i = 0$, otherwise it is said to be uncensored at time T_i if $\delta_i = 1$.
The log-rank test statistics for splitting based on the independent variable X at the value of c are:

$$|L(X, c)| = \frac{\sum_{j=1}^{m} \left(d_{j,L} - Y_{j,L} \frac{d_j}{Y_j} \right)}{\sqrt{\sum_{j=1}^{m} \frac{Y_{j,L}}{Y_j} \left(1 - \frac{Y_{j,L}}{Y_j} \right) \left(\frac{Y_j - d_j}{Y_j - 1} \right) d_j}}$$

1.3.2 *Bootstrap*

Bootstrap is a nonparametric resampling technique that can work without the need for distribution assumptions because the original sample will be used as the population. The steps in the bootstrap method are sampling back from the initial dataset to get new data. The sampling technique was returned from the original data with the same size, and was returned B times. The original sample is the initial sample generated from observations that are considered as a population [8].

1.3.3 *Variable Importance*

The independent variable is selected by filtering based on the importance of the variable. The representative method used for the selection of the importance variable is the importance permutation method. Permutation importance is a method to calculate the level of importance of variable X in the average of the t-tree from the difference between prediction errors after permutation and before permutation of variable X. A variable with a positive importance value indicates that the variable has good predictive ability. Meanwhile, if the importance value is zero or negative, then the variable is non-predictive [9].

$$I\left(X^j, \Theta_m, \mathcal{L}\right) = \frac{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} \ell\left(Y_i, h\left(\widetilde{\boldsymbol{X}}_i^{\ j}, \Theta_m, \mathcal{L}\right)\right)}{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} 1} - \frac{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} \ell\left(Y_i, h(\boldsymbol{X}_i, \Theta_m, \mathcal{L})\right)}{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} 1}$$

1.3.4 Comparing the Cox Proportional Hazard Method with the Random Survival Forest Method

Comparison of the Cox proportional hazard and random survival forest models or commonly referred to as prediction errors can be calculated using the Harrell's concordance index (C-Index) approach. The C-Index method is a tool to assess the accuracy of prediction performance in survival analysis which is quite popular. The comparison is made by comparing the error values of the two methods with the C-Index, where the error value is 1 – C-Index, so it can be said that a larger C-Index value produces smaller errors and provides better prediction accuracy [7].

The following is the Harrell's concordance index equation to assess the accuracy of prediction performance in survival analysis

$$C = \frac{\sum_{i\neq j} 1\{\eta_i < \eta_j\} 1\{T_i > T_j\}\, d_j}{\sum_{i\neq j} 1\{T_i > T_j\}\, d_j}$$

1.4 Length of Study

The length of study is the time required for students to complete their education from the time they enter until they graduate. The quality of a student's graduation is expected to be used as capital to find work according to the expertise possessed. The quality of student graduates is influenced by several factors, namely internal factors and external factors [10]. The existence of internal factors and external factors is very influential for a student in taking his education. Internal factors are factors that come from within the student, such as intelligence, emotion, level of intelligence, psychological state, and others. On the other hand, external factors are factors that come from outside the individual, such as the family environment, community environment, campus environment, educational infrastructure provided by the campus, and also the learning motivation given to them [11].

2. METHOD

The types of data in this study are nominal data and numerical data. Nominal data is categorical data. The nominal data in this study are data on factors that affect the length of study for students, gender (x_2), parents' occupation (x_3), regional origin (x_4), entrance to university (x_5), scholarships (x_6) and Part time (x_7). While the numerical data in this study is the GPA (x_1). The data used in this study are primary data and secondary data. Primary data was obtained by conducting direct interviews with electronic media (telephone and whatsapp) and distributing google forms. The secondary data referred to in this study is data on the length of study of students at FMIPA UNIB class 2017.

The stages of the research carried out are as follows:

1. Data exploration
2. Cox proportional hazard regression analysis
    a. Parameter estimation
    b. Parameter testing with simultaneous test and partial test
    c. Testing the proportional hazard assumption with the Schoenfeld residual residual plot
    d. Cox proportional hazard model interpretation
3. Analysis of random survival forest
    a. Take bootstrap samples from real survival data

b. Form a survival tree from each bootstrap sample
c. Choose the independent variable randomly at each tree node
d. Split (split) tree nodes using independent variables
e. Calculate CHF ensemble value
f. Selection of variable importance

4. Comparing the CPH method with the RSF method using the C-Index with the smallest error value
5. Get the best method

3. RESULTS AND DISCUSSION

3.1 Parameter Estimation

There are three ways to test the significance of the parameters, namely the partial likelihood ratio test, the Score test, and the Wald test. Parameter significance testing aims to check whether the independent variables have a real influence in the model. The simultaneous test in this study used a partial likelihood ratio test, while the partial test used a score test [6].

Table 1. Parameter Estimation of CPH Model with Breslow Approach

Variable	β_j	e^{β_j}	*Standar Error*	Lower limit	Upper limit
X_1	3.8307	46.0958	0.8422	8.8473	240.167
X_2	0.3033	1.3544	0.3185	0.7254	2.529
X_3	-0.1640	0.8487	0.3491	0.4282	1.682
X_4	-0.1436	0.8663	0.3625	0.4257	1.763
X_5	0.1431	1.1538	0.4092	0.5174	2.573
X_6	0.0019	1.0019	0.3284	0.5264	1.907
X_7	0.4900	1.6323	0.2998	0.9070	2.938

So that the estimation of the Cox proportional hazard model with the Breslow approach is obtained as follows:

$$h(t,X) = h_0(t)exp(3.8307X_1 + 0.3033X_2 - 0.1640X_3 - 0.1436X_4 + 0.1431X_5 + 0.0019X_6 + 0.49X_7) \quad (4)$$

3.2 Parameter Test

To find out whether all the variables in equation (4) have an effect on the model, then the parameter testing is carried out with the partial likelihood ratio test as follows:

1. Hypothesis
   $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
   H_1: At least one $\beta_j \neq 0, j = 1, 2,3, \ldots, p$
2. Significance level: $\alpha = 5\%$
3. Test statistics:

$$G = -2[\ln L_R - ln\, L_F]$$

$$= -2[-228.1884 - (-215.3723)] = 25.3723$$

4. Rejection area:
   Reject H_0 if $G \geq \chi^2_{(\alpha:db=p)}$ or $p - value < \alpha$
   Because $= 25.3723 \geq \chi^2_{(0.05:7)}14.0671$ or $-value = 0.001214 < \alpha = 0.05$ so H_0 is rejected
5. Conclusion:
   There is at least one independent variable that has an effect on the model.

The results of the partial parameter testing using the score test, as follows:

Table 2. Partial Parameter Test Results with Score Test

Variable	z	$\chi^2_{(0,05:1)}$	$p-value$	Decision
X_1	4.549	3.841	5.4e-06	H_0 rejected
X_2	0.952	3.841	0.341	H_0 received
X_3	-0.470	3.841	0.638	H_0 received
X_4	-0.396	3.841	0.692	H_0 received
X_5	0.350	3.841	0.727	H_0 received
X_6	0.006	3.841	0.995	H_0 received
X_7	1.634	3.841	0.102	H_0 received

3.3 Testing the Proportional Hazard Assumption

In this Schoenfeld residual, if the slope curve is close to zero, then the curve indicates that the coefficient of X_1 is constant. So it can be interpreted that the proportional hazard assumption is met. The following is a plot of the Schoenfeld residual of the GPA variable.

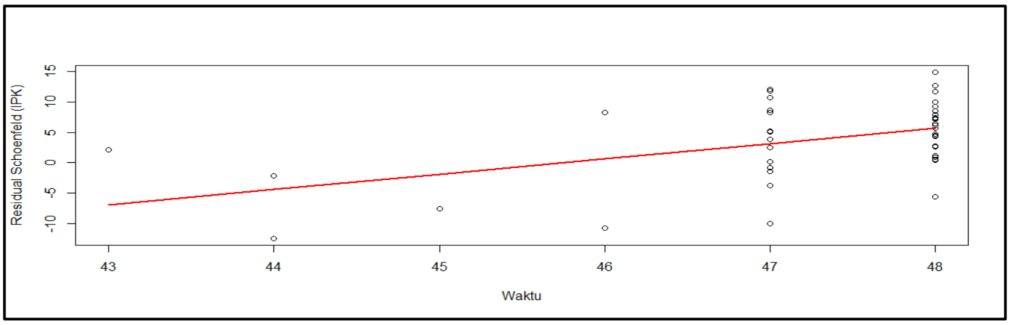

Figure 1. Schoenfeld Residual Plot for GPA Variables

3.4 Interpretation of the Cox Proportional Hazard Model

The final cox proportional hazard model is as follows:

$$h(t,X) = h_0(t)exp(3.8307X_1) \tag{5}$$

Equality (5) shows the value of $e^{\beta 1}$which shows the effect of the independent variable on the hazard function. In this study, the GPA variable is a numerical variable, so the hazard ratio is obtained by taking a number that ranges between the GPA values. It is assumed that in this study comparing the cum laude GPA (GPA 3.7) with the non- cumlaude (GPA 3.32), it is obtained

$$Hazard\ Ratio = \frac{\exp\ (coef(3.7))}{\exp\ (coef(3.32))} = \frac{exp\ (3.8307(3.7))}{exp\ (3.8307(3.32))} = 4.287$$

Based on the calculation results, it can be said that students with a GPA of 3.7 are 4,287 times faster to graduate on time than students who have a GPA of 3.32.

3.5 Random Survival Forest

The first step in the random survival forest method is dividing the data into 75% training data and 25% testing data. The following are the results of data processing using the random survival forest method:

Table 3. Results of RSF Method Data Processing

Sample size	:	90
Number of deaths	:	37
Number of trees	:	1000
Forest terminal node size	:	15
Average no. of terminal nodes	:	3.488
No. of variables tried at each split	:	3
Total no. of variables	:	7
Resampling used to grow trees	:	Swor

Resample size used to grow trees	:	57
Analysis	:	RSF
Family	:	Surv
Splitting rule	:	logrank *random
Number of random split points	:	3
(OOB) CRPS	:	0.0078016
(OOB) Requested performance error	:	0.34716407

3.6 Importance Variable

The representative method is used for the selection of the importance variable is the importance permutation method. Permutation importance is a method for calculating the level of importance of variable X in the mean of the t-th tree from the difference between the prediction error after the permutation and before the permutation of the variable X.

Table 4. Free Variable Importance Value

Variabel	*Importance*
X_1	0.1168
X_2	0.0141
X_3	0.0082
X_4	-0.0030
X_5	-0.0032
X_6	-0.0034
X_7	-0.007

A variable with a positive importance value indicates that the variable has a good predictive ability. Meanwhile, if the value of importance is zero or negative, then the variable is non-predictive. Based on **Table 4** shows that the variables $X_1, X_2, X_7$are predictive variables. Meanwhile, the variables X_3, X_4, X_5, X_6 are considered not predictive because the importance value is negative [9].

3.7 Comparing the Cox Proportional Hazard Method with the Random Survival Forest Method

The following table compares the Cox proportional hazard and the random survival forest method:

Table 5. Harrell's Concordance Index

	CPH	RSF
C-Index	0.722	0.731
Error C-Index	0.278	0.269

Based on **Table 5**, the C-Index error in the random survival forest method is 26.9%, smaller than the cox proportional hazard method, which is 27.8% so that the most suitable method used for data on the length of study for undergraduate students at FMIPA UNIB class of 2017 is the 2017 method. random survival forest (RSF).

4. CONCLUSION

Based on the results and discussion of cox proportional hazard regression and random forest survival, the following conclusions can be drawn:

1. Significant factor affecting the length of study for undergraduate students at FMIPA UNIB class of 2017 using cox proportional hazard regression is the cumulative achievement index. Meanwhile, using the random survival forest method is the cumulative achievement index, gender, and part time.
2. The final cox proportional hazard model is as follows:

$$h(t,X) = h_0(t)exp(3.8307X_1) \tag{6}$$

Equality (6) shows the value of $e^{\beta 1}$ which shows the effect of the independent variable on the hazard function. In this study, the GPA variable is a numerical variable, so the hazard ratio is obtained by taking numbers that range between the GPA values. It is assumed that in this study comparing the cum laude GPA (GPA 3.7) with the non-cum laude (GPA 3.32), it is obtained

$$Hazard\ Ratio = \frac{\exp\ (coef(3.7))}{\exp\ (coef(3.32))} = \frac{exp\ (3.8307(3.7))}{exp\ (3.8307(3.32))} = 4.287$$

Based on the calculation results, it can be said that students with a GPA of 3.7 are 4,287 times faster to graduate on time than students who have a GPA of 3.32.

3. The results of the comparison of methods using the harrell's concordance index state that the random survival forest method is more suitable for use in data on the length of study for undergraduate students at FMIPA UNIB class of 2017 because the resulting C-Index error value is 26.9%, smaller than the cox proportional hazard method. that is 27.8%.

REFERENCES

[1] Fitriana, R. (2016). “Analisis Survival Faktor-Faktor yang Mempengaruhi Lama Studi Mahasiswa Pendidikan Matematika Angkatan 2010 dengan Metode Regresi Cox Proportional Hazard”. Tugas Akhir Program Studi Statistika Terapan Dan Komputasi, Jurusan Matematika FMIPA UNNES. https://lib.unnes.ac.id/25050/

[2] Kleinbaum, D. G., & Klein, M. (2005). “Survival Analysis”: A Self-Learning Text Second Edition. In Media. Springer Science+Business Media. Inc. USA

[3] Lee, E. T., & Wang, J. W. (2003). “Statistical Methods For Survival Data Analysis (Third Edit)”. Oklahoma City, Oklahoma.

[4] Lawless, J. F. (2003). “Statistical Models and Methods for Lifetime Data”. In Biometrics (Vol. 39, Issue 3). https://doi.org/10.2307/2531129

[5] Klein, J. P., & Moeschberger, M. L. (2003). “Survival Analysis Techniques for Censored and Truncated Data Second Edition” (Vol. 19, Issue 5).

[6] Hosmer., D. W., & Lemeshow, S. (2008). “Applied Survival Analysis: Regression Modeling of Time to Event data”. New Jersey : John Wiley.

[7] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). “Random survival forests”. Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

[8] Sungkono, J. (2015). “Bootstrap Resampling Observasi Pada Estimasi Parameter Regresi Menggunakan Software R”. 92, 101–106.

[9] Myte, R. (2013). “Covariate Selection for Colorectal Cancer Survival Data”. Umea Universitet.

[10] Putri, Y. L. D., & Bustami. (2021). “Mengidentifikasi Faktor-Faktor Yang Mempengaruhi Lama Masa Studi Mahasiswa Menggunakan Regresi Cox Proportional Hazard”.

[11] Putra, N. A. J., Nitiasih, P. K., Adil, N., & Gede Gunatama. (2014). “Identifikasi Faktor-Faktor Yang Mempengaruhi Lama Masa Studi Mahasiswa Di Fakultas Bahasa Dan Seni Undiksha”.