BAREKENG: Journal of Mathematics and Its Applications
September 2025 Volume 19 Issue 3 Page 2111-2126
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss3pp2111-2126

SURVIVAL ANALYSIS ON DATA OF STUDENTS NOT GRADUATING ON TIME USING WEIBULL REGRESSION, COX PROPORTIONAL HAZARDS REGRESSION, AND RANDOM SURVIVAL FOREST METHODS

Ramya Rachmawati [1*], Nur Afandi[2], Muhammad Arib Alwansyah[3]

[1,2]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Bengkulu*
Jln. WR. Supratman Kandang Limun, Muara Bangkahulu, Bengkulu, 38122, Indonesia

[3]*Mathematics Education, Faculty of Mathematics and Natural Sciences, University of Jakarta*
Jln. Rawamangun Muka, Jatinegara Kaum, Pulo Gadung, ADM. East Jakarta, 13220, Indonesia
*Corresponding author's e-mail: * **ramya.rachmawati@unib.ac.id***

Article History:

Received: 9th February 2024
Revised: 20th March 2025
Accepted: 13th April 2025
Published: 1st July 2025

Keywords:

Cox Proportional Hazards Regression;
Length of Study;
Random Survival Forest;
Survival Analysis;
Weibull Regression.

ABSTRACT

This article presents a comprehensive study of the factors that influence the length of study data of undergraduate students at FMIPA UNIB class 2018 and 2019. This study is essential because observations show that many students study for more than 8 semesters. The purpose of this study is to determine the factors that significantly influence the length of study of undergraduate students. These factors can be internal and external. Survival analysis is the right method to identify these factors because ordinary regression analysis is unable to estimate survival data. Therefore, methods such as Weibull regression, Cox Proportional Hazards regression, and Random Survival Forest are used. This study does not compare the methods used because these methods are independent of each other, but have the same goal, namely, to determine the factors that influence the length of study of students. The data used in this study are data on the length of study of students from the 2018 and 2019 cohorts sourced from the academic subsection of FMIPA UNIB, with variables of GPA, gender, region of origin, university entry route, parents' occupation, type of study program, and length of study. The results showed that GPA and the type of study program significantly influenced the length of study in Weibull regression analysis. In Cox proportional hazard regression, the GPA variable is an influential factor, while using the Random Survival Forest method, all factors significantly influenced the length of study, with their respective levels of importance.

How to cite this article:

R. Rachmawati, N. Afandi and M. A. Alwansyah., "SURVIVAL ANALYSIS ON DATA OF STUDENTS NOT GRADUATING ON TIME USING WEIBULL REGRESSION, COX PROPORTIONAL HAZARDS REGRESSION, AND RANDOM SURVIVAL FOREST METHODS," *BAREKENG: J. Math. & App.*, vol. 19, no. 3, pp. 2111-2126, September, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Higher education serves as a place to educate the next generation of the nation in academic and non-academic aspects. According to [1], every university makes maximum efforts to increase the graduation rate of its students, both in terms of quantity and quality. Bengkulu University, as one of the leading state universities in Bengkulu Province, has a vision to become a world-class university. To achieve this vision, hard work and dedication from the entire academic community are required. The Faculty of Mathematics and Natural Sciences is one part of Bengkulu University, which has 14 study programs.

The quality of university graduates can be influenced by internal and external factors. Internal factors include intelligence, learning ability, and family background. Meanwhile, external factors include the learning environment, socialization, and available facilities. These factors are believed to have an impact on the duration of student studies [1]. Survival analysis is the method utilized to look at the variables that affect the students' study duration in this study. A statistical technique called survival analysis is applied when the data set involves the amount of time before a specific event takes place. According to [2], in analyzing survival data, ordinary linear regression cannot be used because it is unable to handle the presence of censored data. If an individual or observation has not gone through a certain occurrence, the data is considered censored. It is referred to be uncensored data if the subject had an experience prior to the conclusion of the observation.

Previous research on the length of study conducted by [3] using Cox proportional hazard regression showed that the significant factors influencing the length of study of ULM FMIPA undergraduate students were gender, grade point average (GPA > 3.50), and residence status. Further research on the comparison of Cox proportional hazard regression and random survival forest was carried out by [4]. The results obtained were that the factors that significantly influenced the length of study using Cox proportional hazard regression were GPA, while using random survival forest, the factors that significantly influenced were GPA, gender, and part-time work.

There are several survival analysis methods, such as the Kaplan-Meier method, Accelerated Failure Time, Parametric Proportional Hazards Model, Stratified Cox Model, Weibull regression, Cox Proportional Hazards regression, and Random Survival Forest. This study uses the Weibull regression method, Cox Proportional Hazards regression, and Random Survival Forest because these methods can handle censored data and can also be used for large amounts of data. This study is the latest research from previous research, namely, using an additional method, the Weibull regression method. These three survival analysis methods are used to see based on the three methods whether there are variables that have a consistent influence on the three methods used and use more data. This study does not compare the methods used because these methods are independent of each other, but have the same goal, namely, to determine the factors that influence the length of study of students. Random Survival Forest is a collection of random tree methods used for right-censored survival time data. Survival time is divided into 2 types, namely non-coincidence time and co-occurrence time. Co-occurrence is a condition where there are two or more individuals who experience an event at the same time. Two or more students who are doing thesis attempts in the same month are considered as a joint event in this study. If the students' thesis attempts last more than 48 months, then the data is considered censored.

2. RESEARCH METHODS

2.1 Survival Analysis

Survival analysis is a time-related statistical technique that begins at the beginning of a certain occurrence. One survival analysis that is used to examine data with survival time as the dependent variable is Cox regression. Survival time is the time from the beginning of the study to the time of occurrence of an event or events [5].

2.2 Censored Data

If a certain event has not occurred in the observation, the data is considered censored. It is referred to be uncensored data if the subject had an experience prior to the conclusion of the observation [2]. According to [6], there are four types of censoring in survival analysis, namely right censoring, left censoring, interval censoring, and random censoring.

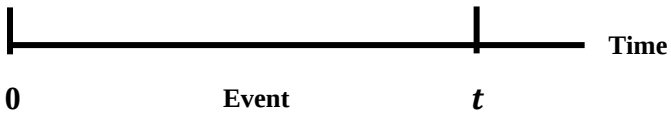

Figure 1. Censorship Illustration

Suppose the case in Figure 1 is that if individuals experience events up to the time limit $\boldsymbol{t}$, then the data is said to be uncensored. If individuals experience events or do not experience events beyond the time limit $\boldsymbol{t}$, then the data is said to be censored.

2.3 Weibull Distribution

The Weibull distribution is a generalization of the Exponential distribution, which was originally used in examining the durability of materials. The Weibull distribution has a shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$.

2.3.1 Weibull Regression

The Weibull model is a survival model with survival time that follows the Weibull distribution, having scale parameters (β) and shape parameters (γ) with the assumption of Accelerated Failure Time (AFT). The AFT formula of the Weibull distribution is as follows [5]:

$$S(t) = 1 - F(t) = 1 - \left(1 - \exp\left(-\left(\frac{t}{\beta}\right)^{\gamma}\right)\right)$$

$$S(t) = \exp(-\beta t^{\gamma}) \tag{1}$$

The AFT assumption is that the explanatory variables are independent of time. This can be seen by looking at the plot of $\ln_e[-\ln_e S_t]$ against survival time (t) for each independent variable, forming a parallel pattern. Here is the hazard function in Weibull regression.

$$h(t) = \beta\gamma t^{\gamma-1}$$

Where $\beta = \exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_j X_j)$ with t is the survival time, X is the independent variable.

2.3.2 Weibull Regression Parameter Estimation

Parameter estimation can be done using the maximum likelihood estimation method. This parameter estimation method is to determine the parameter that maximizes the probability of the sample data. If $f(t;\theta)$ is a joint probability function where t is a realization of T, then the likelihood function of θ is defined as follows:

$$L(\theta|t) = f(t;\theta) \tag{2}$$

For survival data that is assumed to be independent and identical and complete, when there are $t_1, t_2, t_3, \cdots, t_n$ observations, the likelihood function can be written:

$$L(\theta|t) = \prod_{i=1}^{n} f(t_i;\theta)$$

For incomplete survival data with the possibility of right censoring, it can be represented as a pair of survival observation values with their censored status, namely $(t_i, \delta_i),\ i = 1, 2, 3, \ldots, n$ with

$$\delta_i = \begin{cases} 0\ ;\ If\ i \text{ is censored} \\ 1\ ;\ If\ i \text{ is not censored} \end{cases}$$

Assuming each (T_i, δ_i) is independent of the other, the likelihood function for right censored data is:

$$L(\theta) \propto \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}$$

With $\theta = (\theta_1, \cdots, \theta_p)$ are the p parameters to be estimated, $f(t_i;\theta)$ is the density function for i who got the event and $S(t_i;\theta)$ is the survival function for i who did not get the event. The log-likelihood function for right-censored data from the likelihood function is:

$$\ell(\theta) \propto \sum_{i=1}^{n} (\delta_i)\ \log\big(f(t_i;\theta)\big) + \sum_{i=1}^{n} (1-\delta_i) \log\big(S(t_i;\theta)\big)$$

Furthermore, to obtain an estimate using the maximum likelihood estimation method is the result of the first and second partial derivatives after obtaining the log of the likelihood function. The results obtained using the method are implicit, so that parameter estimates are obtained computationally with the help of software using the Newton-Raphson iteration method.

2.3.3 Multicollinearity Detection

Multicollinearity is a condition in which two or more independent variables in the regression model are strongly correlated. This can cause problems in regression analysis, such as difficulty in determining the effect of variables on other variables, making coefficient estimates unstable, or reducing the accuracy of model predictions **[7]**. According to **[8]**, multicollinearity is caused by several factors, namely the application of data collection, the limitations contained in the model, or differences in the population being sampled.

This study uses the Variance Inflation Factor (VIF) value to detect multicollinearity. The following is the VIF **Equation (3) [9]**:

$$VIF = \frac{1}{\left(1 - R_j^2\right)} \tag{3}$$

The VIF value is theoretically impossible to be negative because it is calculated from the square of the correlation, where R_j^2 is the coefficient of determination of the regression of the variable X_j against the other independent variables. Since R_j^2 has a value between 0 and 1, the VIF is always positive and has a minimum value of 1. The VIF also cannot be zero, because a value of 0 will appear if $R_j^2 = 1$, which mathematically will cause division by zero (undefined). VIF values below 10 are considered to indicate no serious multicollinearity, while values above 10 indicate high multicollinearity that needs to be addressed.

2.3.4 Significant Testing of Weibull Regression Parameters

Knowing whether the independent variable really influences the model is the goal of parameter significance testing. In this research, the Wald test is used for partial tests and the partial likelihood ratio test for simultaneous tests. The Wald test is used because it is more efficient, namely, by using the coefficient estimate $(\hat{\beta})$ and its standard error, the Wald test is practical for evaluating the influence of one variable at a time on survival time in the model. The partial likelihood ratio test compares models with all variables and models without the tested variables, thus providing a more comprehensive picture of the overall contribution of the variables

a. Simultaneous Test

1. Hypothesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

H_1 : at least one $\beta_j \neq 0$ where $j = 1, 2, \cdots, p$

2. Significance level $\alpha = 5\%$

3. Test Statistics

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

4. Rejection Criteria

Reject H_0 if $G_{hit} > \mathcal{X}^2_{(df=p;\alpha)}$ or $P\text{-}value < \alpha$

5. Conclusion

Because the $P\text{-}value < \alpha$ so H_0 is rejected, meaning that at least one independent variable has an effect on the dependent variable, and the model is feasible.

b. Partial Test

1. Hypothesis

$H_0 : \beta_j = 0$

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

2. Significance level $\alpha = 5\%$

3. Test Statistics

$$Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{4}$$

Rejection Criteria

Reject H_0 if $|Z| > Z_{\alpha/2}$ or $P\text{-}value < \alpha$

4. Conclusion

Because the $P\text{-}value < \alpha$ so H_0 is rejected, the independent variable affects the dependent variable.

2.4 Cox Proportional Hazards Regression (CPH)

The Cox model is a semiparametric distributed model because in Cox, the estimation of regression parameters of the Cox model does not have to determine the basic hazard function; besides that, the Cox model does not require information about the underlying distribution of survival time [10].

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

$$\boldsymbol{h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) = h_0(t) e^{\sum_{j=1}^{p} \beta_j X_j}} \tag{5}$$

2.4.1 Co-Occurrence Data

Co-occurrence data is often found in survival analysis. Co-occurrence is an event where two or more individuals experience an event at the same time. The following is an example of co-occurrence data.

Table 1. Example of Co-Occurrence Survival Data

i	1	2	3	4	5	6
t_i	4	3	5	7	6	5
c	0	1	1	0	0	1

Suppose, i is the i-th individual, c is the censored data status and t_i is the event time for the i-th individual. Suppose the event times $t_1 < t_2 < t_3 < t_4 < t_5 < t_6$ are the observed times and have been sorted. At time t_3 and t_4 there are two individuals who experienced the event, and it is not known which individual experienced the event first.

2.4.2 Parameter Estimation of the Cox Proportional Hazards Model

The Breslow partial likelihood approach is an alternate technique for handling co-occurrence data that is provided by [6]. The partial likelihood equation for the Breslow approach is as follows:

$$L(\beta)_{Breslow} = \prod_{i=1}^{r} \frac{e^{\left(\sum_{j=1}^{p} \beta_j S_k\right)}}{\left(\sum_{i \in R(t_j)} e^{\left(\sum_{j=1}^{p} \beta_j X_{ij}\right)}\right)^{d_i}} \quad (6)$$

Where S_k is the amount of covariance X on the joint event and d_i is the number of cases of the joint event at time t_i.

2.4.3 Cox Proportional Hazards Regression Parameter Testing

The purpose of parameter significance testing is to determine if the independent variables actually affect the model [11]. The ratio partial likelihood test is used in this study's simultaneous test, while the score test is used in the partial test.

Simultaneous test steps using the ratio partial likelihood test

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 value (α)

3. Test statistic:

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

   Where $\ln L_R$ is the Log likelihood value for the Cox model without independent variables, while $\ln L_F$ is the Log likelihood value for the Cox model with independent variables.

4. Rejection Criteria:

   Reject H_0 if $G \geq \chi^2_{(\alpha:df=p)}$ or $P\text{-}value < \alpha$

5. Conclusion:

   Reject H_0 if $G \geq \chi^2_{(\alpha:df=p)}$ or $P\text{-}value \leq \alpha$, meaning that at least one independent variable has an effect in the model.

The test statistic in the score test follows the chi-square distribution with a free degree of p. Here are the score test steps [12]:

1. Hypothesis:

   $H_0: \beta_j = 0$ (variable X_j has no effect in the model)

   $H_1: \beta_j \neq 0, j = 1, 2, \ldots, p$ (variable X_j has an effect on the model)

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

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

4. Rejection criteria:

   H_0 is rejected if $z \geq \chi^2_{(\alpha:df=p)}$ or $P\text{-}value \leq \alpha$

5. Conclusion:

   Reject H_0 if $z \geq \chi^2_{(\alpha:df=p)}$ or $P\text{-}value < \alpha$, means that variable X_j has an effect on the model

2.4.4 Cox Proportional Hazard Model Assumptions

The Schoenfeld residual plot is one way to test the proportional hazard assumption in a Cox model [13]. According to [2], Schoenfeld residuals are defined as residuals for each individual and each independent variable based on the first derivative of the log-likelihood function. The following is the Schoenfeld Residual equation:

$$R_{ji} = \delta_i \left(X_{ji} - \frac{\sum_{l \in R(t_j)} X_{jl} e^{(\widehat{\boldsymbol{\beta}}' \boldsymbol{X_l})}}{\sum_{l \in R(t_j)} e^{(\widehat{\boldsymbol{\beta}}' \boldsymbol{X_l})}} \right), j = 1, 2, \ldots, p \tag{7}$$

With $\widehat{\boldsymbol{\beta}}$ being the maximum partial likelihood estimator of $\boldsymbol{\beta}$. When the sample size is large, the expected value of R_{ji} is zero, so the Schoenfeld residuals are uncorrelated with each other.

2.5 Random Survival Forests (RSF)

Random Survival Forest (RSF) is a nonparametric ensemble method for analyzing right-censored survival data, built as a time-to-event extension of random forests for classification. This method can handle multiple covariates, noise covariates, and complex nonlinear relationships between covariates without the need for prior specification [14].

2.5.1 Splitting

Let's say we wish to divide a node h into two child nodes throughout the tree construction process. If n observations with survival times and censoring indicators $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ are present at the node h, then observation I is considered censored at time T_i if $\delta_i = 0$, and uncensored at time T_i if $\delta_i = 1$.

The log-rank test statistic for splitting according to the independent variable X at a value of c is as follows:

$$|L(X,c)| = \frac{\sum_{j=1}^{m} \left(d_{j,L} - Y_{j,L} \frac{d_j}{Y_j} \right)}{\sqrt{\sum_{j=1}^{m} \frac{Y_{j,L}}{Y_j} \left(1 - \frac{Y_{j,L}}{Y_j}\right) \left(\frac{Y_j - d_j}{Y_j - 1}\right) d_j}} \tag{8}$$

The value of $|L(X,c)|$ is a measure of the node separation which is directly proportional to the performance of the node separation. The greater the value of $|L(X,c)|$ the better the separation and the greater the difference between the two groups. [15].

2.5.2 Bootstrap

Since the original sample will be used as the population, bootstrap is a nonparametric resampling approach that may function without any distribution assumptions. The bootstrap method's steps include resampling the original dataset to produce fresh data [4].

2.5.3 Variable Importance Selection

The permutation significance approach is a sampling technique for selecting variables based on their importance. Permutation importance is a technique that determines variable X significance in the t-t tree average by comparing the prediction error of variable X before and after permutation [16]. A positive significance value for a variable indicates that it has excellent predictive power [17]. The variable is not predictive even if the importance value is zero or negative. Let $\mathcal{L}^*(\Theta_m)$ be the mth bootstrap sample and $\mathcal{L}^{**}(\Theta_m) = \mathcal{L}\backslash\mathcal{L}^*(\Theta_m)$ be the OOB data. Given $\boldsymbol{X} = (X_1, X_2, .., X_p)$ where X_j is the th-j feature coordinate. The permutation value of the th-j coordinate in $\boldsymbol{X}$ is denoted by $\tilde{\boldsymbol{X}}^j$.

$$\tilde{\boldsymbol{X}}^j = (X_1, \dots, X_{j-1}, X_j, X_{j+1}, , .., X_p)$$

Variable importance is calculated by looking at the difference between the permutation value of $\tilde{\boldsymbol{X}}^j$ and the original value of $\boldsymbol{X}$ in the OOB data. In other words, let $I(X_j, \Theta_m, \mathcal{L})$ be the variable importance for $X^{(j)}$ in the th-m tree. The permutation variable importance equation is as follows:

$$I(X^j, \Theta_m, \mathcal{L}) = \frac{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} \ell\left(Y_i, h\left(\tilde{\boldsymbol{X}}_i^{\,j}, \Theta_m, \mathcal{L}\right)\right)}{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} 1} - \frac{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} \ell\left(Y_i, h(\boldsymbol{X}_i, \Theta_m, \mathcal{L})\right)}{\sum_{i\epsilon\mathcal{L}^{**}(\Theta_m)} 1} \quad (9)$$

2.6 Duration of Study

According to [4], the length of study is the time required for students to complete their education according to their respective education levels; for example, for the undergraduate level is 4 years, the Diploma level is 3 years, and the Magister is 2 years.

3. RESULTS AND DISCUSSION

The types of data in this study are nominal data and numeric data. Nominal data in this study are data on factors that affect the length of study of students, such as gender, parents' occupation, region of origin, university entry path, and study program. While the numeric data in this study is the GPA variable. The data used in this study is secondary data, namely data on the length of study of students at FMIPA UNIB, batches of 2018 and 2019. Data is said to be censored if an individual or observation has not experienced a certain event. If an individual experiences an event before the end of the observation, it is called uncensored data.

3.1 Weibull Regression Modeling

3.1.1 Parameter Estimation

Weibull regression parameter estimation for each variable in the length of student study data.

Table 2. Parameter Estimation of the Weibull Regression Model

Variable	β_j	Standard Error
Intercept	5.233	0.104
GPA	-0.392	0.030
Gender	0.022	0.016

Variable	β_j	Standard Error
Origin	-0.001	0.012
Selection	-0.001	0.009
Parents Occupation	-0.006	0.013
Study Program	0.028	0.004
Scale = 0.0939	Shape=1/Scale = 10.650	
Chisq = 197.97	$P\text{-}Value = 5.1 \times 10^{-40}, df = 6$	

Based on the output results of the Weibull regression model on survival data, an intercept value of 5.233 was obtained with a standard error of 0.104. The coefficient of the GPA variable of -0.392 with a standard error of 0.030 indicates that the higher the GPA value, the shorter the study period tends to be. This means that students with high GPAs tend to complete their studies faster. The Gender variable has a positive coefficient of 0.022 with a standard error of 0.016, indicating that gender has a small and positive effect on survival time, although the effect is relatively small. The Origin and Selection variables each have a coefficient of -0.001 with a standard error of 0.012 and 0.009, respectively, meaning that their effect on survival time is very small and tends to be insignificant. Likewise, Parents' Occupation has a coefficient of -0.006 with a standard error of 0.013, indicating a very weak negative effect. Meanwhile, the Study Program has the most prominent influence with a positive coefficient of 0.028 and a standard error of 0.004, indicating that the type of study program plays a fairly important role in determining the length of student study. The scale model parameter of 0.0939 and the shape of 10.650 indicate that the Weibull distribution shape approaches a hazard distribution that increases over time, indicating that the risk of graduation increases as the study period increases. So that the Weibull regression model estimation is obtained as follows:

$$S(t|X) = \exp\left(-\left(\exp\begin{pmatrix}5.233 - 0.392X_1 + 0.022X_2 - 0.001X_3 - 0.001X_4 - 0.006X_5 \\ +0.028X_6\end{pmatrix} t\right)^{10.650}\right) \quad (10)$$

3.1.2 Multicollinearity Detection

Multicollinearity is a condition in which two or more independent variables in the regression model are strongly correlated. A regression model is considered good if there is no multicollinearity.

Table 3. Multicollinearity Detection Results

Variable	VIF Value
GPA	1.143
Gender	1.018
Origin	1.048
Selection	1.079
Parents Occupation	1.072
Study Program	1.130

It is known in **Table 3** that the data has a VIF value <10, meaning that all variables in the data do not have multicollinearity.

3.1.3 Parameter Testing

The partial likelihood ratio test is used for parameter testing to see if each variable in **Equation (10)** affects the model:

1. Hypothesis

   H_0: $\beta_1 = \beta_2 = \cdots = \beta_6 = 0$

   H_1: At least one $\beta_j \neq 0, j = 1, 2, 3, \ldots, 6$

2. Level of significance: $\alpha = 5\%$

3. Test statistic:

$$G = -2[\ln L_R - \ln L_F] = -2[-1018.7 - (-919.7)] = 197.97$$

4. Rejection criteria:

Reject H_0 if $G \geq \chi^2_{(\alpha:df=p)}$ or $P - value < \alpha$

Since $G = 197.97 \geq \chi^2_{(0.05:6)} = 12.5916$ or $P - value = 5.1 \times 10^{-40} < \alpha = 0.05$, H_0 is rejected

5. Conclusion:

There is at least one independent variable that has an effect on the model.

The following is a partial parameter significance test using the Wald test.

Table 4. Partial Parameter Testing Results with Wald Test

Variable	z	$P - Value$	Decision
Intercept	50.20	2×10^{-16}	H_0 is rejected
GPA	-13.21	2×10^{-16}	H_0 is rejected
Gender	1.38	0.17	H_0 is accepted
Origin	-0.11	0.91	H_0 is accepted
Selection	-0.11	0.91	H_0 is accepted
Parents Occupation	-0.43	0.66	H_0 is accepted
Study Program	8.04	9.1×10^{-16}	H_0 is rejected

Based on **Table 4**, the GPA and Study Program variables have an effect on the model, this can be seen from the P-value which is smaller than the 5% alpha value. While the Selection, Origin, Selection, and Parents Occupation variables have no effect in the model because the P-value is greater than the 5% alpha value. It can be concluded that the final Weibull regression model is as follows:

Table 5. Weibull Regression Final Model Results

Variable	β_j	Standard Error	Z	P-Value
Intercept	5.247	0.101	52.01	2×10^{-16}
GPA	-0.396	0.029	-13.56	2×10^{-16}
Study Program	0.028	0.004	8.09	5.9×10^{-16}
Scale = 0.0943		Shape=1/Scale = 10.604		
Chisq = 195.76		$P\text{-}Value = 3.1 \times 10^{-43}$		$df = 2$

Based on the final results of the Weibull regression model in **Table 5**, it is obtained that the Intercept is 5.247 with a standard error of 0.101, which indicates that the intercept is significant. The GPA variable has a regression coefficient of -0.396 with a standard error of 0.029, which is also very significant. This coefficient indicates that the higher the GPA, the shorter the time to graduation, or in other words, students with higher GPAs tend to complete their studies faster. Furthermore, the Study Program variable has a coefficient of 0.028 with a standard error of 0.004, indicating a significant effect on survival time. A positive coefficient indicates that there is a difference in the length of study between study programs. The scale parameter in the model is 0.0943, and the shape is 10.604, which indicates a hazard distribution shape that increases over time. From the estimation results of the final Weibull regression model, the Weibull regression model equation is obtained:

$$S(t|X) = \exp(-(\exp(5.247 - 0.396X_1 + 0.028X_6)t)^{10.604}) \quad (11)$$

Based on **Equation (11)**, it is interpreted that the intercept of 5.247 indicates the basic hazard risk without being influenced by other variables. The coefficient of -0.396 for GPA indicates that the higher the GPA value, the lower the risk of occurrence, meaning that the higher the GPA value obtained by students, the more likely the students are to graduate on time in 8 semesters. The coefficient of 0.028 for the study program indicates that the higher the value of the study program variable, the higher the risk of occurrence, meaning that the higher the value of the study program variable, the more the study program has the incentive to encourage students to graduate on time. The shape parameter of 10.604 indicates that the data has a distribution form with decreasing risk over time, meaning that the hazard decreases over time.

3.2 Cox Proportional Hazard Regression Modeling

3.2.1 Parameter Estimation

Cox proportional hazard regression parameter estimation for each variable in the student length of study data.

Table 6. CPH Model Parameter Estimation Using the Breslow Method

Variable	β_j	e^{β_j}	Standard Error	Lower Limit	Upper Limit
GPA	3.2793	26.5578	0.3432	13.5538	52.0384
Gender	-0.0269	0.9735	0.1728	0.6938	1.3659
Origin	-0.0340	0.9665	0.1314	0.7471	1.2505
Selection	0.1159	1.1228	0.0919	0.9378	1.3444
Parents Occupation	-0.1019	0.9031	0.1472	0.6768	1.2050
Study Program	-0.2577	0.7728	0.0378	0.7176	0.8323

Based on the results of the Cox Proportional Hazards (CPH) model parameter estimation using the Breslow method in **Table 6**, it can be seen that the GPA variable has a coefficient of 3.2793. The exponential value of the coefficient of 3.2793 indicates that a one-unit increase in GPA is associated with a 26.56-fold increase in the chance of an event occurring, meaning that GPA has a very large and significant influence on the acceleration of graduation time; the higher the GPA, the more likely students are to graduate faster. Conversely, the Gender variable has a coefficient of -0.0269, indicating that gender has a small and insignificant influence on graduation time. The Origin and Parents' Occupation variables that have an influence on survival time are considered insignificant. Overall, the most significant variables in this model are GPA and Study Program, while other variables tend not to have a significant influence on survival time. So that the Cox proportional hazard model estimate is obtained as follows:

$$\boldsymbol{h(t,X) = h_0(t)\exp(3.2793X_1 - 0.0269X_2 - 0.0340X_3 + 0.1159X_4 - 0.1019X_5 - 0.2577X_6)} \quad \textbf{(12)}$$

3.2.2 Parameter Testing

To find out whether all variables in **Equation (12)** affect the model, parameter testing is carried out with the partial likelihood ratio test:

1. Hypothesis

   H_0: $\beta_1 = \beta_2 = \cdots = \beta_6 = 0$

   H_1: At least one $\beta_j \neq 0, j = 1, 2, 3, \ldots, 6$

2. Level of significance: $\alpha = 5\%$

3. Test statistic:

$$G = -2[\ln L_R - \ln L_F] = -2[-1506.215 - (-1448.317)] = 115.796$$

4. Rejection region:

   Reject H_0 if $G \geq \chi^2_{(\alpha:db=p)}$ or $P - value < \alpha$

   Since $G = 115.796 \geq \chi^2_{(0.05:6)} = 12.5916$ or $P - value = 2 \times 10^{-16} < \alpha = 0.05$ H_0 is rejected

5. Conclusion:

   There is at least one independent variable that has an effect on the model.

The following is a partial parameter significance test using the Score test:

Table 7. Partial Parameter Testing Results with Score Test

Variable	z	$\chi^2_{(0,05:1)}$	$P - Value$	Decision
GPA	9.555	3.841	2×10^{-16}	H_0 is rejected
Gender	0.156	3.841	0.876	H_0 is accepted
Origin	0.259	3.841	0.796	H_0 is accepted
Selection	1.261	3.841	0.207	H_0 is accepted
Parents Occupation	0.693	3.841	0.489	H_0 is accepted
Study Program	6.813	3.841	9.6×10^{-12}	H_0 is rejected

Based on **Table 7**, the GPA and Study Program variables have an effect on the model with a P-value that is smaller than the 5% alpha value. The Gender, Origin, Selection, and Parents Occupation variables

have no effect in the model because the P-value is greater than the 5% alpha value. It can be concluded that the final Cox proportional hazard model is as follows:

$$h(t,X) = h_0(t)\exp(3.2638X_1 - 0.2558X_6)$$

3.2.3 Proportional Hazard Assumption Testing

The slope curve of the Schoenfeld Residual Plot suggests that the coefficient of X_i is constant if it is near zero. Thus, the proportionate hazard assumption can be said to be met. The Schoenfeld residual plot of the GPA variable is shown below.

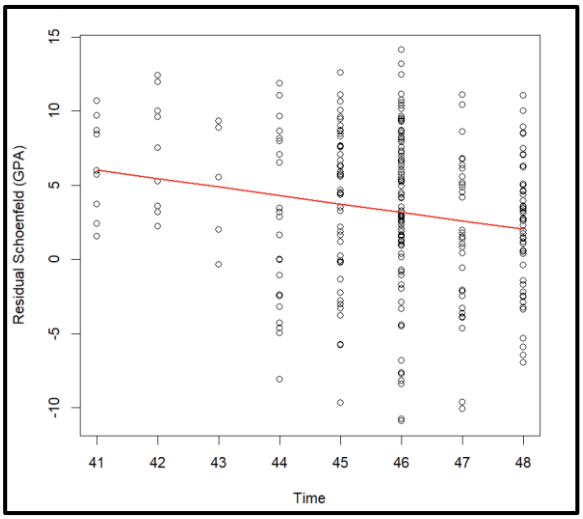

Figure 2. Schoenfeld Residual Plot for GPA Variable

Based on **Figure 2**, it can be seen that the Schoenfeld residual plot against the survival time of the GPA variable has a slope close to zero, so it can be said that the proportional hazard assumption is met.

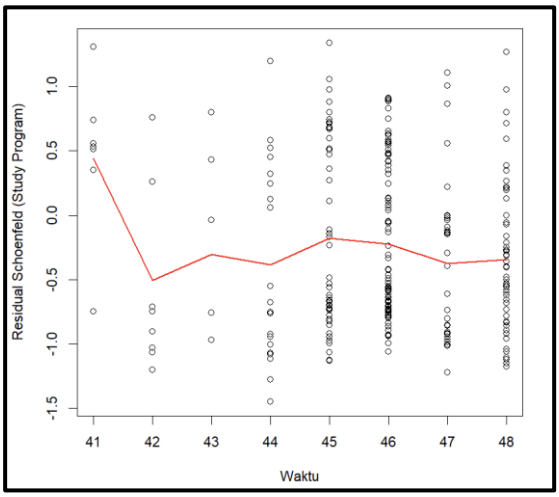

Figure 3. Schoenfeld Residual Plot for the Study Program Variable

Based on **Figure 3**, it can be seen that the Schoenfeld residual plot against the survival time of the Study Program variable does not have a slope close to zero, so it is said that the proportional hazard assumption is not met. According to **[18]**, if the proportional hazard assumption is not met, then remove the independent variables that do not meet the assumptions from the model, and the Cox proportional hazard model can still be used.

3.2.4 Interpretation of the Cox Proportional Hazard Model

Based on parameter testing and proportional hazard assumptions, it is concluded that the final Cox proportional hazard model is:

$$h(t,X) = h_0(t)\exp(2.4322X_1) \quad (13)$$

Equation (13) illustrates the impact of independent variables on the hazard function by displaying the value of e^{β_1}. Since the GPA variable in this study is numerical, the hazard ratio is calculated by selecting a value that falls between the GPA ranges. The hazard ratio, assuming that this study compares cumlaude GPA values (GPA 3.8) with those that are not cumlaude (GPA 3.28), is as follows:

$$Hazard\ Ratio = \frac{\exp\left(coef(3.8)\right)}{\exp\left(coef(3.28)\right)} = \frac{\exp(2.4322(3.8))}{\exp(2.4322(3.28))} = 3.5$$

Based on the calculation results, it can be said that students with a GPA of 3.8 are 3.5 months faster to graduate on time than students who have a GPA of 3.28.

3.3 Random Survival Forest *(*RSF*)*

3.3.1 RSF Method Data Processing

Separating the data into 85% training and 15% testing is the first stage in the random survival forest technique. The outcomes of using the random survival forest approach to data processing are as follows:

Table 8. RSF Method Data Processing Results

Sample size	:	352
Number of deaths	:	226
Number of trees	:	1000
Forest terminal node size	:	15
Average no. of terminal nodes	:	13.539
No. of variables tried at each split	:	3
Total no. of variables	:	6
Resampling used to grow trees	:	Swor
Resample size used to grow trees	:	222
Analysis	:	RSF
Family	:	Surv
Splitting rule	:	logrank *random
Number of random split points	:	3
(OOB) CRPS	:	0.01507168
(OOB) Requested performance error	:	0.24172093

Based on **Table 8**, it can be seen that the sample size indicates the number of samples used as bootstrap data to build a survival tree. In this study, the bootstrap data sample was 352 samples. Then the number of students graduating on time was 226, stating that there were 226 students who failed (in the case of this study, students graduating on time) from 352 students who were randomly selected as bootstrap samples. The number of trees indicates the number of trees built to construct the forest, and it is 1000 trees. The forest terminal node size is the number of training samples in the terminal node. This means that the survival tree is built until the terminal node has a size of 15, or in other words, the survival tree cannot grow anymore at a terminal node size of less than 15 nodes. While the average no. of terminal nodes shows the average number at the terminal node is 13.539. In addition, the no. of variables tried at each split shows the number of candidate variables in the node separation obtained from the root of the number of independent variables, namely $\sqrt{6} = 2.5$, into 3 variables. Resampling for bootstrap samples is done by swor (sampling with replacement). Then "analysis: RSF and family: surv" shows that data processing is done using the random survival forest method, which is one of the methods in survival analysis. The splitting rule or splitting rule used in this analysis is the random log-rank rule. The number of random split points shows the number of random split points is 3. From **Table 8**, it can also be seen that the prediction error results obtained are 1.52%.

3.3.2 Variable Importance

The selection of important variables is done to find out which variables affect the length of the study. The representative method used for variable importance selection is the permutation importance method. Permutation importance is a technique that determines a variable X degree of significance in the tree average by comparing the prediction error before and after permutation.

Table 9. Importance Value of Free Variables

Variable	Importance
GPA (X_1)	0.1343
Study Program (X_6)	0.0931
Parents Occupation (X_3)	0.0043
Origin (X_4)	0.0042
Gender (X_2)	0.0023
Selection (X_5)	0.0021

Based on Table 9, it can be seen that the importance value for variables X_1, X_2, X_3, X_4, X_5 and X_6 is positive. This shows that all variables are predictive variables, or in other words, these variables are variables that can significantly predict whether students graduate on time. When viewed from the importance value, variable X_1 is a variable with a greater importance value among other variables, meaning that variable X_1 is the most predictive variable compared to other variables.

This research shows that the GPA has a significant influence on students' graduation time at the FMIPA UNIB. These results align with previous research findings showing that GPA is an important indicator in determining students' academic success and graduation time. For example, research conducted by [4] shows that students with higher GPAs tend to graduate more quickly than those with lower GPAs. This research makes a significant contribution to the field of educational continuity analysis because it confirms that academic quality, reflected in GPA, is closely related to the effectiveness of the learning process and timely graduation. These findings also support educational policies that emphasize strengthening students' academic quality to encourage accelerated graduation times. Policies such as academic mentoring programs, curriculum improvements, and monitoring academic progress can be more focused on increasing students' GPAs, which in turn is expected to shorten the duration of their studies. In other words, the results of this research are in line with previous studies that show the importance of academic factors in higher education success and provide a strong basis for formulating policies that can help improve overall student graduation rates.

4. CONCLUSIONS

Based on the results and discussion of Weibull regression, Cox proportional hazard regression, and random survival forest, it can be concluded that the resulting Weibull regression model is $S(t|X) = \exp(-(\exp(5.247 - 0.396X_1 + 0.028X_6)t)^{10.604})$, while the resulting Cox proportional hazards model with the Breslow approach is $h(t,X) = h_0(t)\exp(2.4322X_1)$. A significant factor influencing the length of study for FMIPA UNIB undergraduate students in the class of 2018 and 2019, based on these three methods, is the cumulative achievement index. Students are advised to achieve and maintain a high GPA so as not to repeat courses. If this is consistently done, then students have a high probability of graduating on time, namely 8 semesters.

REFERENCES

[1] R. Widyastuti, D. Wulandari, and D. Prasetyowati, "PENERAPAN REGRESI COX UNTUK MENGANALISIS VARIABEL YANG BERPENGARUH TERHADAP DURASI STUDI MAHASISWA," vol. 13, pp. 88–98, 2024. doi: https://doi.org/10.14710/j.gauss.13.1.88-98

[2] E. T. Lee and J. W. Wang, *STATISTICAL METHODS FOR SURVIVAL DATA ANALYSIS*, Third Edit. A John Wiley & Sons, INC., Publication, 2015.

[3] U. L. Mangkurat, "ANALISIS REGRESI COX UNTUK MENENTUKAN FAKTOR-FAKTOR YANG MEMPENGARUHI

LAMA STUDI MAHASISWA S1 FMIPA UNIVERSITAS LAMBUNG MANGKURAT 1,2,3," vol. 13, pp. 1–12, 2024. doi: https://doi.org/10.14710/j.gauss.13.1.1-12

[4] M. A. Alwansyah, "SURVIVAL ANALYSIS OF STUDENTS NOT GRADUATED ON TIME USING COX PROPORTIONAL HAZARD REGRESSION METHOD AND RANDOM SURVIVAL FOREST METHOD," *J. Stat. Data Sci.*, vol. 2, no. 1, pp. 13–21, 2023. doi: https://doi.org/10.33369/jsds.v2i1.24312

[5] D. A. I. Maruddani, Tarno, A. Hoyyi, R. Rahmawati, and Y. Wilandari, *SURVIVAL ANALYSIS*. UNDIP Press Semarang, 2021.

[6] J. P. Klein and M. L. Moeschberger, *SURVIVAL ANALYSIS TECHNIQUES FOR CENSORED AND TRUNCATED DATA SECOND EDITION*, vol. 19, no. 5. John Wiley. New Jersey., 2003. doi: https://doi.org/10.1007/b97377

[7] Roger Koenker, V. Chernozhukov, X. He, and L. Peng, *HANDBOOK OF QUANTILE REGRESSION*, First Edit. CRC Press Taylor & Francis Group, 2018. doi: https://doi.org/10.1201/9781315120256

[8] D. C. Montgomery, E. A. Peck, and G. G. Vining, *INTRODUCTION TO LINEAR REGRESSION ANALYSIS*, Sixth Edit. Wiley Series in Probability And Statistics, 2021.

[9] M. Sriningsih, D. Hatidja, and J. D. Prang, "PENANGANAN MULTIKOLINEARITAS DENGAN MENGGUNAKAN ANALISIS REGRESI KOMPONEN UTAMA PADA KASUS IMPOR BERAS DI PROVINSI SULUt," *J. Ilm. Sains*, vol. 18, no. 1, p. 18, 2018. doi: https://doi.org/10.35799/jis.18.1.2018.19396

[10] C. Kartsonaki, "SURVIVAL ANALYSIS," *Diagnostic Histopathol.*, vol. 22, no. 7, pp. 263–270, 2016. doi: https://doi.org/10.1016/j.mpdhp.2016.06.005

[11] D. F. Moore, *APPLIED SURVIVAL ANALYSIS USING R*. New York:Springer, 2016. doi: https://doi.org/10.1007/978-3-319-31245-3

[12] N. Eliyati, S. I. Maiyanti, O. Dwipurwani, and S. W. Hamidah, "Model Regresi Cox Untuk Menganalisis Pengaruh Faktor Asupan Makanan Terhadap Risiko Kekambuhan Endometriosis," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 15, no. 1, pp. 103–114, 2021. doi: https://doi.org/10.30598/barekengvol15iss1pp103-114

[13] D. Collet, *MODELLING SURVIVAL DATA IN MEDICAL RESEARCH*, Fourth edi. CRC texts in statistical science, 2023.

[14] K. L. Pickett, K. Suresh, K. R. Campbell, S. Davis, and E. Juarez-Colunga, "RANDOM SURVIVAL FORESTS FOR DYNAMIC PREDICTIONS OF A TIME-TO-EVENT OUTCOME USING A LONGITUDINAL BIOMARKER," *BMC Med. Res. Methodol.*, vol. 21, no. 1, pp. 1–14, 2021. doi: https://doi.org/10.1186/s12874-021-01375-x

[15] M. Rezaei, L. Tapak, M. Alimohammadian, A. Sadjadi, and M. Yaseri, "JOURNAL OF BIOSTATISTICS AND EPIDEMIOLOGY," *J Biostat Epidemiol.*, vol. 1, no. Iran, pp. 37–44, 2020.

[16] S. Nurhaliza, K. Sadik, and A. Saefuddin, "A COMPARISON OF COX PROPORTIONAL HAZARD AND RANDOM SURVIVAL FOREST MODELS IN PREDICTING CHURN OF THE TELECOMMUNICATION INDUSTRY CUSTOMER," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 4, pp. 1433–1440, 2022. doi: https://doi.org/10.30598/barekengvol16iss4pp1433-1440

[17] M. Mohammed, I. B. Mboya, H. Mwambi, M. K. Elbashir, and B. Omolo, "PREDICTORS OF COLORECTAL CANCER SURVIVAL USING COX REGRESSION AND RANDOM SURVIVAL FORESTS MODELS BASED ON GENE EXPRESSION DATA," *PLoS One*, vol. 16, no. 12 December, pp. 1–22, 2021. doi: https://doi.org/10.1371/journal.pone.0261625

[18] B. M. Iskandar, "MODEL COX PROPORTIONAL HAZARD PADA KEJADIAN BERSAMA," Universitas Negeri Yogyakarta, 2015.