Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 16(1) 2023: 37-46
http://ejournal.undip.ac.id/index.php/media_statistika

KAPLAN-MEIER AND NELSON-AALEN ESTIMATORS FOR CREDIT SCORING

Tatik Widiharih, Agus Rusgiyono, Sudarno, Bagus Arya Saputra
Department of Statistics, Diponegoro University, Indonesia

e-mail: widiharih@gmail.com

DOI: 10.14710/medstat.16.1.37-46

Article Info:
Received: 18 December 2022
Accepted: 24 July 2023
Available Online: 4 October 2023

Keywords:
Kaplan-Meier; Nelson-Aalen; Survival Function; Cumulative Hazard Function.
.

Abstract: Financial institutions use credit scoring analysis to predict the probability that a customer will default. In this paper, we determine the probability of default using nonparametric survival analysis that are Kaplan-Meier and Nelson-Aalen. The analysis is based on survival function curves, cumulative hazard function curves, mean survival time, and standard error of estimators. Based on the curves of survival function for both Kaplan Meier and Nelson Aalen estimators relatively the same. Based on the curves of cumulative hazard function, mean survival time, and standard error the Nelson-Aalen estimators are slightly higher than Kaplan-Meier.

1. INTRODUCTION

A default occurs when a borrower stops making the required payments on a debt. Defaults can occur on secured debt, such as a mortgage loan secured by a house, or unsecured debt, such as credit cards or a student loan. Default is a big problem in the financial industry, another term for default is also called bad credit. The probability of bad credit can be determined using credit scoring analysis with parametric, semiparametric, and nonparametric methods. Nonparametric methods for credit scoring analysis according to bad and good credit have been investigated. Mukid et al. (2018) used the Weighted K Nearest Neighbor (WKNN) method by some kernels and applied it to private bank in central of Java. Gaussian and rectangular kernels are better performance than triangular, triweight, epanichov, and inversion kernels. Widiharih & Mukid (2018) did credit scoring using Method of Local Means based K Harmonic Nearest Neighbor (MLMKHNN), and applied it to motorcycle loans in Central Java, at the same level MLMKHNN need more neighbor than Local Means K Nearest Neighbor (LMKNN). Widiharih et al. (2018) did credit scoring using kernel discriminant, normal kernel is relevant to be selected for credit scoring model. Pratiwi et al. (2019) used the Pseudo Nearest Neighbor (PNN) method, applied to national bank in central of Java, K nearest neighbor (KNN) is better than PNN. Mukid et al. (2019), compared several nonparametric methods based on K Nearest Neighbor (KNN), Distance Weighted KNN (DWKNN), PNN, LMKNN and applied it to customers' data in micro credit a government bank in Wonogiri district. LMKNN is better than the other methods.

The parametric method used to credit scoring by classification is discriminant analysis. Mukid & Widiharih (2016) conducted a credit assessment using discriminant analysis with mixed of independent variables binary and continuous variables, applied to banks in Lampung City. Classification errors in the training data were 0.1970 and 0.3753 in the test data. Another

parametric model is binary logistic regression, Sa'idah et al. (2021) use this method to assess motorcycle loans with an accuracy of 76.38%.

Credit scoring analysis by taking into account the time until the occurrence of default has also been studied. Credit scoring can be analyzed using survival analysis (Narain, 1992). Cox regression model is a semiparametric method in survival analysis (Collet, 2003), it does not need to follow a certain distribution but must meet the proportional hazard assumption. If this assumption is not met, then the model is said non-proportional hazard (Ata and Sozer, 2007). Kurniawan et al. (2015) use Extended Cox model for motorcycle financing data. Gupta (2017) has applied the Cox proportional hazard model for impact of financial ratios and capital market ratios. Dirick et al. (2017) used Cox and accelerate failure test models for Belgian and UK financial institutions.

The nonparametric survival methods are the Kaplan-Meier and Nelson-Aalen estimators. This analysis is used in the case of small data and does not follow a particular distribution. The researchers who have done this method including Njomen & Wandji (2014) work in theory on some asymptotic properties of nonparametric estimator and applied it in completing risks. Jaber et al. (2017) used the Kaplan-Meier and Nelson-Aalen estimators which were applied to banking data in Jordan, Nelson-Aaleen was better than Kaplan-Meier. Besides that, they also do parametric models including Exponential, Log-normal, Log-logistic, Gamma, Weibull, and Gompertz. The results Gompertz model is the best model. Sukono et al. (2018) did the application of Kaplan-Meier and Nelson-Aalen estimators for estimating mean and variance of infected duration of dengue fever. Hikmah and Ekawati (2021), the results are not significant differences between recovery time for husband and wife sufferers of hypertension. Obed et al (2021) did the application of Kaplan-Meier and Nelson-Aalen estimators to COVID-19 cases in Kurdikistan. The result, Nelson-Aalen estimator demonstrated that the chances of surviving were higher during a short period after being exposed to the virus.

The purpose of this research is to determine the Kaplan-Meier and Nelson-Aalen estimators that are applied to data of part private banks in the Aceh region. The first step is goodness of fit test of the distribution of survival time data, which shows that there is no suitable distribution with the data so using Kaplan-Meier and Nelson-Aalen is appropriate. We determine survival probability and cumulative hazard. The Kaplan-Meier estimator is based on gender and the adequacy of the customer's balance is also done.

2. LITERATURE REVIEW.

Kaplan-Meier and Nelson-Aalen are used in the case of small data and do not follow a particular distribution. One of the tests used to test the suitability of data following a certain distribution is the Anderson-Darling test (AD-Test). The Anderson-Darling Goodness of Fit Test (AD-Test) is a measure of how well your data fits a specified distribution**.**

2.1. AD-Test

The hypotheses for the AD-test are:
H_0: The data comes from a specified distribution.
H_1: The data does not come from a specified distribution.

The formula of AD is:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln F(X_i)\right) + \ln\left(1 - F(X_{n-i+1})\right) \quad (1)$$

where n: the sample size; $F(X_i)$: Cumulative Distribution Function for the specified distribution; i: the *i*-th sample, calculated when the data is sorted in ascending order. Reject H_0 if p-value < α where p-value depends on the specified distribution

2.2. Survival Analysis

Survival analysis is a statistical method to analyze data of time until an event occurs (Kleinbaum & Klein, 2012). The survival time of an individual is denoted by T with the probability density function $f(t)$. There are three functions in survival analysis. The functions are the probability density function $f(t)$, survival function $S(t)$, and the failure function $h(t)$.

Survival function is the probability of an individual surviving more than time t. It can be written as follows:

$$S(t) = P(T > t) \tag{2}$$

The function $S(t)$ is a non-increasing function with respect to time t with the properties $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$.

The hazard function is denoted by h(t) and is defined as the probability of an object failing in the interval time $t + \Delta t$ if the object has been survived in time t.

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \le T \le t + \Delta t | T \ge t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln S(t)) \tag{3}$$

The cumulative hazard function is defined as:

$$H(t) = \int_0^t h(x)\, dx \tag{4}$$

The relationship between the survival function and cumulative hazard function is:

$$S(t) = \exp(-H(t)) \tag{5}$$

$$H(t) = -\ln S(t) \tag{6}$$

2.3. Kaplan-Meier Estimator

The Kaplan-Meier method is a non-parametric survival method that does not require assumptions to follow a certain distribution (Kleinbaum & Klein, 2012). In this case, we use right censored data, so it is commonly called time to event data. Data is denoted by $\{(t_i, d_i) i = 1,2, \dots . n\}$ where t_i is time to even dan d_i status of censored. Let the event occurs at D distinct times $t_1 < t_2 < \cdots < t_D$, at time t_i there are d_i events, n_i the number of individuals at risk at time t_i. The survival function S(t) of the Kaplan Meier estimator is:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_i \\ \prod_{t_i \le t} \left(\frac{1 - d_i}{n_i}\right) & \text{if } t_i \le t \end{cases} \tag{7}$$

Cumulative hazard $\hat{H}(t)$ is determine by $\hat{H}(t) = -\ln\left(\hat{S}(t)\right)$

The estimate of Var $\left(\hat{S}(t)\right)$ is:

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{8}$$

The $(1 - \alpha)100\%$ pointwise confidence interval for survival function $S(t)$ at $t = t_k$ is:

$$\hat{S}(t_k) \pm Z_{\alpha/2}\sqrt{\hat{V}\left(\hat{S}(t_k)\right)}\ \ \hat{S}(t_k) \tag{9}$$

Where $Z_{\alpha/2}$ is upper-tail $\alpha/2$ percentile of standard normal distribution

2.4. Nelson-Aalen Estimator

Nelson-Aalen estimator is used to determine the cumulative hazard function and it is defined as follows:

$$\widehat{H}(t) = \begin{cases} 0 & \text{if } t < t_i \\ \sum_{t_i \le t} \dfrac{d_i}{n_i} & \text{if } t_i \le t \end{cases} \tag{10}$$

Survival function $\hat{S}(t)$ is determine by $\hat{S}(t) = \exp(-\widehat{H}(t))$

The estimate of variance $\left(\widehat{H}(t)\right)$ is:

$$\hat{V}\left(\widehat{H}(t)\right) = \sum_{t_i \le t} \frac{d_i}{{n_i}^2} \tag{11}$$

The $(1-\alpha)100\%$ pointwise confidence interval for cumulative hazard:

$$\widehat{H}(t_i) \pm Z_{\alpha/2}\sqrt{\hat{V}\left(\widehat{H}(t_k)\right)} \tag{12}$$

2.5. Mean Survival Time

Estimator of mean survival time is:

$$\hat{\mu} = \int_0^{\vartheta} \hat{S}(t)\, dt \tag{13}$$

where ϑ is either the highest time to event or preassigned by the researcher.

The alternative to determine $\hat{\mu}$ using: $\hat{\mu} = \sum_{i=1}^{D} \hat{S}\,(t_{i-1})(t_i - t_{i-1})$ where $\boldsymbol{t}_D$ is the highest time to event and $t_0 = 0$ (Zeng, 2017).

The estimated variance of $\hat{\mu}$ is:

$$\hat{V}(\hat{\mu}) = \sum_{i=1}^{D} \left(\int_{t_i}^{\vartheta} \hat{S}(t)\right)^2 \frac{d_i}{n_i(n_i - d_i)} \tag{14}$$

Standard error of $\hat{\mu}$ is $\sqrt{\hat{V}(\hat{\mu})}$

3. MATERIAL AND METHOD

Credit scoring analysis is applied to small branch offices of private banks in the Aceh region. Data used up to 31 July 2021, we obtain the data from bank employees. Research variables are presented in Table 1.

The steps of the analysis are: (1) Goodness of fit test of the distribution of survival time data with the Anderson Darling test to conclude that the nonparametric approach is suitable for the data; (2) Determine estimator Kaplan-Meier includes survival probability and their cumulative hazard; (3) Plotting the survival probability of Kaplan-Meier; (4) Determine

estimator Nelson-Aalen includes cumulative hazard and their survival probability; (5) Plotting the survival probability of Nelson-Aalen; (6) Plotting the survival probability of Kaplan-Meier and Nelson-Aalen in the same graph; (7) Plotting the cumulative hazard of Kaplan-Meier and Nelson-Aalen in the same graph; (8) Plotting the survival probability of Kaplan-Meier based on gender; (9) Plotting the survival probability of Kaplan-Meier based on balance adequacy, (10) Determine mean survival time and standard error of mean survival for Kaplan-Meier and Nelson-Aalen estimator.

Table 1. Research Variables

No	Variable	Description
1	Instalment periode (T)	Duration time of payment (in months)
2	Status of Credit	1: bad credit 0: good credit
3	Gender (X_1)	1: male 0: female
4	Balance adequacy (X_2)	1: Sufficient balance 2: Insufficient balance

4. RESULTS AND DISCUSSION

Based on 56 data customers, there are 11 customers with bad status, and 45 customers with good status. Customers with bad status are 6 females and 5 males, but based on balance adequacy there are 10 sufficient balance and 1 insufficient balance. Goodness of fit by Anderson-Darling method is presented in Table 2.

Table 2. Goodness of Fit for Survival Time

Distribution	AD	P
Normal	6.691	<0.005
Lognormal	2.311	<0.005
Exponential	3.551	<0.003
Weibull	3.540	<0.001
Gamma	3.699	<0.005
Logistic	5.594	<0.005
Loglogistic	1.956	<0.005

Table 3. Survival Probability and Cumulative Hazard for Kaplan-Meier Estimator

Time Event	Number at risk	Number failed	Survival probability	Cumulative hazard
0≤ t<7	56	1	0.982143	0.018018
7≤t <8	51	1	0.962885	0.037821
8≤t<12	34	1	0.934565	0.067674
12≤t<3	31	1	0.904418	0.100464
13≤t<17	19	1	0.856817	0.154531
17≤t<44	10	1	0.771135	0.259892
44≤t<65	7	1	0.660973	0.414042
65≤t<73	6	2	0.440649	0.819507
73≤t<78	3	1	0.293766	1.224972
78≤t<83	2	1	0.146883	1.918119

Based on Table 2 it can be concluded that there is no suitable distribution for survival time, so a nonparametric approach can be used. The Kaplan-Meier estimator is obtained in Table 3, the graph of survival probability is presented in Figure 1. Based on Table 3 shows that time event 73 months the survival probability less than 0.5, it indicated that mean survival time between 65 to 73 months.

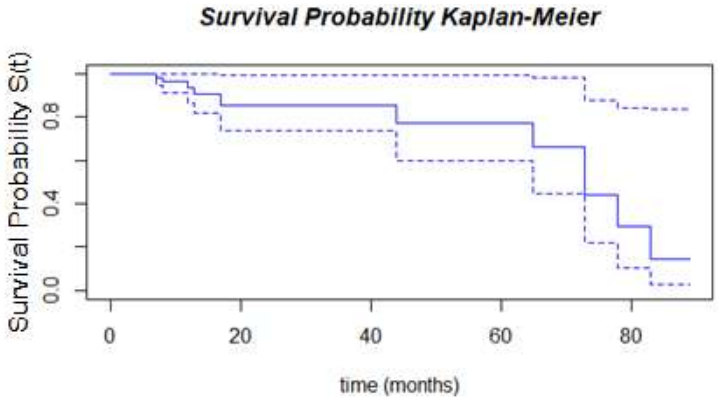

Figure 1. Curve of Survival Probability Kaplan-Meier Estimator

The Nelson-Aalen estimator is obtained in Table 4, the graph of survival probability is presented in Figure 2. Based on Table 4, it relatively same with Table 3, shows that time event 73 months the survival probability less than 0.5, it indicated that mean survival time between 65 to 73 months. Plot survival probability Kaplan-Meier versus Nelson Aalen is presented in Figure 3.

Table 4. Survival Probability and Cumulative Hazard for Nelson-Aalen Estimator

Time Event	Number at risk	Number failed	Survival probability	Cumulative hazard
$0 \leq t < 7$	56	1	0.9820	0.01816
$7 \leq t < 8$	51	1	0.9630	0.03770
$8 \leq t < 12$	34	1	0.9350	0.06721
$12 \leq t < 13$	31	1	0.9060	0.10093
$13 \leq t < 17$	19	1	0.8590	0.15432
$17 \leq t < 44$	10	1	0.7770	0.26007
$44 \leq t < 65$	7	1	0.6740	0.41400
$65 \leq t < 73$	6	1	0.4830	0.81871
$73 \leq t < 78$	3	2	0.3460	1.22418
$78 \leq t < 83$	2	1	0.2100	1.91732

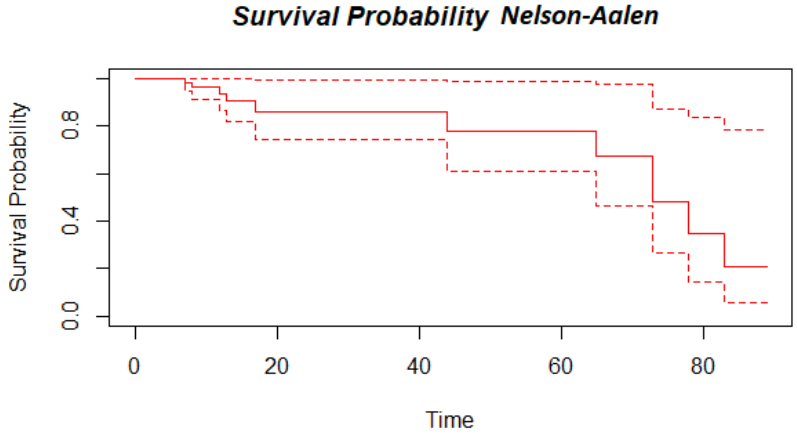

Figure 2. Curve of Survival Probability Nelson-Aalen Estimator

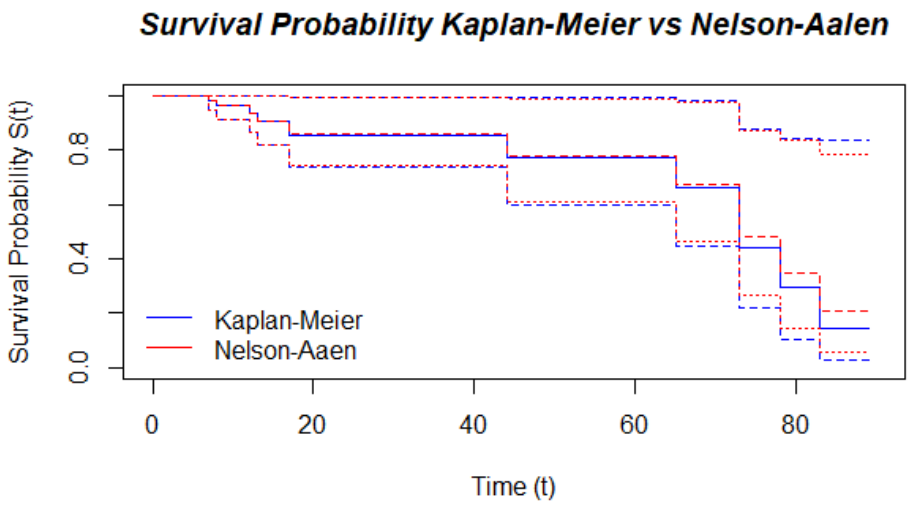

Figure 3. Curve of Survival Probability for Kaplan-Meier versus Nelson-Aalen

Based on Figure 3 shows that the survival probability of the two estimators is relatively the same. The plot of the cumulative hazard function of Kaplan-Meier versus Nelson-Aalen is presented in Figure 4.

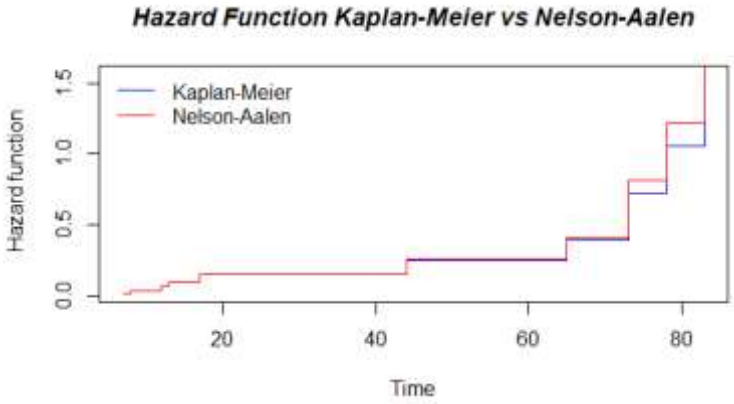

Figure 4. Curve of Cumulative Hazard for Kaplan-Meier versus Nelson-Aalen

Based on Figure 4 it can be seen that the cumulative hazard function of Nelson-Aalen with a life time more than 65 months is slightly higher. The survival probability plot of Kaplan- Meier based on gender is presented in Figure 5 and survival probability plot based on balance adequacy is presented in Figure 6.

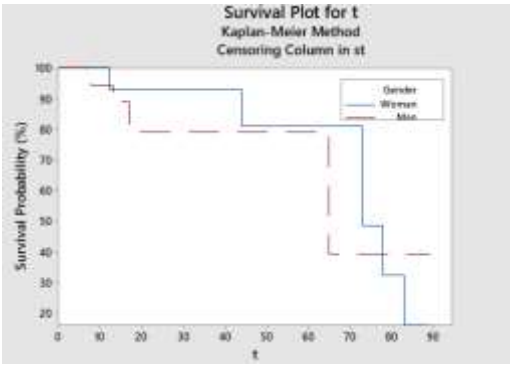

Figure 5. Curve of Survival Probability of Kaplan-Meier Based on Gender

Based on Figure 5 it can be seen that the survival probability of women is higher than men. Based on Figure 6 it can be seen that the survival probability of insufficient balance is higher than sufficient balance.

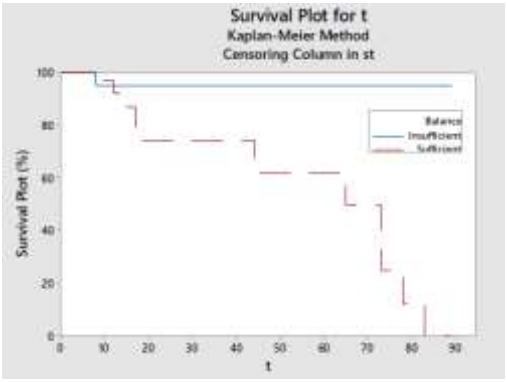

Figure 6. Curve of Survival Probability of Kaplan-Meier Based on Balance Adequacy

Mean survival time and standard error for Kaplan-Meier and Nelson-Aalen estimator is presented in Table 5.

Table 5. Mean Survival Time and Standard Error for Kaplan-Meier and Nelson-Aalen Estimator

Estimator	Mean time	Standard Error
Kaplan-Meier	65.519	5.51932
Nelson-Aalen	66.700	5.60000

Based on Table 5 the mean time of the Nelson-Aalen estimator is higher , this result is the same with Obed (2021). The standard error of the Nelson-Aalen is slightly higher than the Kaplan-Meier estimator, this result is different with Jaber (2017).

5. CONCLUSION

Based on Table 2 and Table 4, survival probability of Kaplan-Meier and Nelson-Aaleen are relatively the same until time event 13. Based on mean time Nelson-Aalen longer than Kaplan-Meier. Based on standard error, Kaplan-Meier is better than Nelson-Aalen. Based on gender, survival probability of men is smaller than women. Based on balance adequacy, survival probability of sufficient balance adequacy is smaller than insufficient balance adequacy.

ACKNOWLEDGMENTS

This research was supported by the Faculty of Science and Mathematics UNDIP based on contract number 23.B/UN7.F.8/PP/II/2023.

REFERENCES

Ata, N. & Sozer, M.T. (2007). Cox Regression Model with Nonproportional Hazard Applied to Lung Cancer Survival Data. *Hacettepe Journal of Mathematics and Statistics*, 36(2),157-167.

Collett, D. (2015). *Modelling Survival Data in Medical Research, second edition*. New York: Chapman and Hall/CRC.

Dirick, L. (2017). Time to Default in Credit Scoring Using Survival Analysis: a Benchmark Study. *Journal of the Operational Research Society,* 68, 652-665.

Gupta, V. (2017). A Survival Approach to Prediction of Default Drivers for Indian Listed Companies. *Theoretical Economics Letters,* 07(02),116-138.

Hikmah & Ekawati, D. (2021) Analisis Data Tersensor Berpasangan dengan Estimasi Kaplan Meier dan Nelson Aalen. *Jurnal Matematika, Sains dan Pembelajarannya,* 7(2),133-138.

Jaber, J.J., Ismail, N., & Ramli, S.N.M. (2017). Credit Risk Assessment Using Survival Analysis for Progressive Right-Censored Data: A Case Study in Jordan. *Journal of Internet Banking and Commerce,* 22(1), 1-18.

Kleinbaum, D.G. & Klein, M. (2012). *Survival Analysis A Self-Learning Text*. New York: Springer Science Business Media, Inc.

Kurniawan, I., Kurnia, A., & Sartono, B. (2015). Survival Analysis with Extended Cox Model About Durability Debtor efforts on Credit Risk. *Forum Statistika dan Komputasi*: *Indonesian Journal of Statistics,* 20(2), 85-95.

Narain, B. (1992). Survival Analysis and The Credit Granting Decision. In: Man, R. 2014. *Survival Analysis in Credit Scoring: A Framework for PD Estimation*. Rabobank International – Quantitative Risk Analytics & University of Twente.

Njomen, D.A.N. & Wandji, J.N. (2014). Nelson-Aalen and Kaplan-Meier Estimators in Completing Risks. *Applied Mathematics*, 5, 755-766

Mukid, M.A., Widiharih, T., & Mustafid. (2019). An Empirical Comparison of Some Modified Nearest Neighbor Rule for Credit Scoring Analysis: Case Study in Indonesia. *Journal of Theoretical and Applied Information Technology,* 97(5), 1644-1654

Mukid, M.A., Widiharih, T., Rusgiyono A., & Prahutama A. (2018). Credit Scoring Analysis Using K Nearest Neighbor. *Journal of Physics Conference Series,* 1025012114, doi:10.1088/1724-6596/1025/1/012114

Mukid, M.A. & Widiharih, T. (2016). Model Penilaian Kredit Menggunakan Analisis Diskriminan Dengan Variable Bebas Campuran Biner dan Kontinu, *Media Statistika* 9(2),107-118

Obed, S.A., Mohammed, P.A., & Kadir, D.H. (2021). The Estimations of (COVID-19) Cases in Kurdikistan Region Using Nelson Aalen Estimator. *CUESJ*, 5(2), 24-31

Pratiwi, H., Mukid M.A., Hoyyi, A., & Widiharih, T. (2019). Credit Scoring Analysis using Pseudo Nearest Neighbor. *Journal of Physics Conference Series,* 1217012100, doi:10.1088/1724-6596/1217/1/012100

Sa'diah, C., Widiharih, T., & Hakim, A.R. (2021). Klasifikasi Pemberian Kredit Sepeda Motor Menggunakan Metode Regresi Logistic Biner dan Chi-Square Automatic Interaction Detection (CHAID) dengan GUI R. *Gaussian* 10(2),159-169

Sukono, Susanti, D., Kartiwa, A., Rani, Q.F., Hidayat, Y., & Bon, A.T. (2018). Application of Product Limit and Nelson Aalen Methods in Health Insurance for Estimating Mean and Variance of Infection Duration of Dengue Fever. *Proceeding of the International Conference on Industrial Engineering and Operations Management Bandung:* 2718-2725.

Widiharih, T. & Mukid, M.A. (2018). Credit Scoring Menggunakan Metode Multi Local Means Based K Harmonic Nearest Neighbor (MLMKHNN). *Media Statistika*, 11(2), 107 – 117

Widiharih, T., Mukid, M.A., & Mustafid (2018). Credit Scoring Analysis Using Kernel Discriminant. *Journal of Physics Conference Series*, 1025012124, doi:10.1088/1742-6569/1025/1/012124

Zeng, P. (2017). *Survival Analysis: Nonparametric Estimation*. Department of Mathematics and Statistics, Auburn University, Alabama US.