Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/130102

Cluster Analysis Integration Model with Survival Analysis for Late Payment of House Ownership Loan (Case Study: House Ownership Loan Bank X Customer)

A. Zahra, A. A. R. Fernandes, and A. B. Astuti*

Department of Statistics, Faculty of Mathematics and Natural Sciences, 65145 Indonesia, Brawijaya University, Malang, Indonesia

Received: 7 Jun. 2022, Revised: 22 Sep. 2022, Accepted: 29 Dec. 2022
Published online: 1 Jan. 2024.

Abstract: This is the abstract, usually it does not have references. Usually the reader will read this part first to know what this paper is about and decide upon it to continue reading or not. The font of main text is 10 Times New Roman with single line spacing of 6 pt after and 0 pt before. The titles of sections are font 12, bold and they have single line spacing of 6pt before, 12 pt after, subsections are font 12, Italic and they have single line spacing of 6pt before, 12 pt after.
Both upper line and lower line enclosing this part is paper-specific and changes according to the paper, usually it is very similar to the journal header background color, abstract contents are Times New Roman size 10, no line spacing.

Keywords: Cluster Analysis, Survival Analysis , Extended Cox Regression, Bank, Credit Collectability.

1 Introduction

In Law No. 10 of 1998 concerning Banking, the bank is mentioned as a business entity that collects funds from the public in the form of savings and at the same time distributes them to the public in the form of credit and or other forms in order to improve people's living standards. One type of credit offered by the bank is a Home Ownership Loan (KPR). [1] House Ownership Loan (KPR) program is one of the credit services offered by the bank to customers to meet the housing needs of the community with a financing scheme up to a certain percentage of the house price. Incorrect determination of credit payment terms and incorrect assessment of the customer's ability to pay credit will cause problems for the bank. This is because credit payments made by the debtor are late or not in accordance with the stipulated time. Current research is growing along with the increasing need to solve problems in life [2]. One of the current human needs is a place to live. One of the policy schemes provided by the government to meet housing needs is through the Home Ownership Credit (KPR) program [3]. Based on [4] and [5], managing the finances and risks of an organization is very necessary. Risk mitigation that can be done by the bank is to determine policies regarding House Ownership Loan (KPR) program lending appropriately. One of the efforts that can be made to make it easier for banks to make decisions is to classify bank X House Ownership Loan (KPR) program customers based on the assessment of the 5C variables (character, capacity, capital, collateral, condition) and credit collectibility and identify factors that affect the length of time for each customer group. In paying House Ownership Loan (KPR) program installments at bank X. These two objectives can be achieved by integrating Cluster analysis with Survival analysis using the Extended Cox regression method. [2] Statistics is a science that processes data from collection to interpretation.

This study aims to classify bank X House Ownership Loan (KPR) program customers based on the assessment of the 5C variable and credit collectibility by integrating cluster analysis with survival analysis. Using Extended Cox Regression to model the probability of a customer's length of time in paying House Ownership Loan (KPR) program loan installments and to find out the factors that influence it. The order of the integration process is Cluster analysis which has produced several groups of objects followed by survival analysis using Extended Cox regression. Based on [6] and [7] in general, there are two cluster analysis methods, namely the hierarchical method and the nonhierarchical method. The two methods have differences that lie in determining the number of clusters. In the hierarchical method the number of clusters is not determined at the beginning of the analysis process, while in the nonhierarchical method the number of clusters is determined at the beginning of the analysis process. The measure of similarity between objects (distance) used is the Manhattan distance. This similarity measure is used because it has the advantage that according to the [8],Manhattan distance it can detect the presence of outliers properly. The ability to detect outliers properly is needed to be able to produce output that visualizes the data optimally [9].

Cluster analysis is exploratory in that the results cannot represent the data in the field, so further analysis is needed to find out more about the data being analyzed [10]. The model needed in this case so that the survival analysis was carried out using the Extended Cox regression method. Vice versa, if only the survival analysis method is used, then the bank cannot explore the characteristics of the customers first by grouping them so that the policy is the same for all customers. This is not a good

*Corresponding author e-mail: avidazahra1@student.ub.ac.id thing to do because the results used are too general. The output of the integration of these two methods is expected to enable the bank to work effectively and efficiently in determining customer loan policy.

Survival analysis is an analysis of the data obtained from the record time achieved by an object until the occurrence of certain events which are referred to as failure events [11]. Survival analysis consists of several approaches , namely parametric, semiparametric, and nonparametric methods. According to [12] one of the analyzes that can be used to determine the relationship between the length of time until an event occurs and other influencing variables is hazard regression analysis. Meanwhile, the semiparametric model that is often used in survival analysis research is the Cox . model proportional hazard which has a proportional hazard assumption. This assumption is an assumption in where the hazard ratio value is constant over time. Hazard ratio as an influence that can be seen in the form of a comparison of two objects with different conditions. Often time can cause changes in the hazard ratio, so it is necessary to test the proportional hazard assumption. If these assumptions are not met, an alternative method is needed to model the probability of resistance test, one of which is the Extended Cox regression which me modified from Cox model proportional hazards. Method used when there is predictor variable that depends on time so that the assumption proportional hazard is not met.

The final step of this research is to compare the integrated model of cluster analysis with Ward -based Extended Cox Regression , and Complete linkage with Manhattan distance and survival analysis model without integration where the selection of the best model can be seen from the smallest Minimum Squared Error (MSE) criteria.

2 Materials and Methods

2.1 Cluster Analysis

According to , [13] Cluster analysis is a multiple variable technique which has the main objective of grouping objects based on the similarity of their characteristics. The characteristics of objects in a cluster have a high degree of similarity, while the characteristics between objects in another cluster have a low degree of similarity [14]. Cluster analysis completion process can be done by grouping the data which can use two methods, namely hierarchical methods and non-hierarchical methods. Several hierarchical method algorithms include complete linkage and ward linkage methods. In the Complete linkage method, the distance between clusters is determined by the furthest distance between two objects in different clusters [15]. The distance from one cluster to another is calculated by equation (2.1):

$$\mathrm{d}_{(\mathrm{ij})\mathrm{k}} = \max(d_{ik}, d_{jk}) \tag{1}$$

Description :

$d_{(\mathrm{ij})\mathrm{k}}$: distances between sub samples (ij) and Cluster k

d_{ik} : distance of sub sample i and Cluster k

d_{jk} : distance of sub sample j and Cluster k

While the ward linkage method is a method of forming clusters based on the loss of information due to merging objects into clusters [15]. The error sum of squares (ESS) is used as an objective function. Two objects will be combined if they have the smallest objective function among the possibilities. The purpose of the ward method is to minimize variance in a cluster and maximize the variance between clusters (Supranto, 2004) .The ESS value is presented in equation (2).

$$ESS = \sum_{j=1}^{p}\left(\sum_{i=1}^{n} x_{ij}^{2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_{ij}^{2}\right)^{2}\right) \tag{2}$$

description:

x_{ij} : the value of the i-th object j in the j-th Cluster

p : number of variables

n : the number of objects in the cluster formed

2.2 Survival Analysis

Survival analysis is a collection of statistical procedures in which the variables of interest are time (T) until the event (events) specified occurs [16]. In survival analysis , the probability distribution of T can be expressed in three ways, namely through the survival function S (t) . P at time t = 0 then S(t) = S (0) = 1 which is interpreted as the beginning of the test where none of the objects get the specified event and the probability of survival of an object is worth one. At time t = then S(t) = S (∞) = 0, meaning that if the test period increases to infinity then in the end there will be no object that can survive so that the chance of survival of an object will approach zero. Through the probability density function , and through the hazard function . One of the analyzes that can be used to determine the relationship between the length of time until an event occurs with other influencing variables is hazard regression analysis [12]. If it is associated with the survival function , then the hazard function is as follows.

$$h(t)=\lim_{\delta t\to 0}\left\{\frac{P\left(t\le T<t+\delta t,T\ge t\right)}{P(T\ge t).\delta t}\right\}$$

$$h(t)=\lim_{\delta t\to 0}\left\{\frac{P\left(t\le T<t+\delta t\right)}{S(t).\delta t}\right\}$$

$$h(t)=\frac{1}{S(t)}\lim_{\delta t\to 0}\left\{\frac{P\left(t\le T<t+\delta t\right)}{\delta t}\right\}$$

$$h(t)=\frac{1}{S(t)}\lim_{\delta t\to 0}\left\{\frac{P\left(T<t+\delta t\right)-P\left(T<t\right)}{\delta t}\right\}$$

$$h(t)=\frac{1}{S(t)}\lim_{\delta t\to 0}\left\{\frac{F\left(t+\delta t\right)-F\left(t\right)}{\delta t}\right\}$$

$$h(t)=\frac{f(t)}{S(t)} \tag{3}$$

karena $f(t)=\frac{-d}{dt}S(t)$ then

$$h(t)=\left(\frac{-d}{dt}S(t)\right)/S(t) \tag{4}$$

d by integrating $h(t)$ the equation (5) below is obtained.

$$\int_0^t h(t)\ dt=-\int_0^t \frac{\frac{dt}{dt}S(t)}{S(t)}dt=-\log S(t) \tag{5}$$

exponentiated , then we get equation (6)

$$S(t)=\exp\left(-\int_0^t h(t)\ dt\right) \tag{6}$$

k because the cumulative hazard function is

$$H(t)=\int_0^t h(t)\ dt \tag{7}$$

Then, from equations (6) and (7), the relationship is obtained, namely:

$$H(t)=-\log S(t) \tag{8}$$

According to [17], there are 3 models to analyze the relationship of a set of predictor variables with survival time, namely the model with a parametric approach, nonparametric, and semiparametric. Model with parametric approach to survival analysis assumes that the survival time distribution follows a given probability distribution. For example: lognormal, exponential, and weibull distributions. Model with nonparametric approach using method Kaplan-Meier to estimate and graph survival probabilities as a function of time. Meanwhile, the model with a semiparametric approach uses Cox PH regression, Stratified Cox, and Extended Cox. The model can describe the effect of covariates on survival. [12], the basis of the extended cox is a regression method that can be called extended cox regression.

According to [12], when the data has certain characteristics so that it cannot use a parametric or nonparametric approach, a semiparametric approach can be used. One method which could used if there is predictor variable that depends on time (time-dependent) so that the assumption proportional hazard is not met is to use Extended Cox regression. Variables that depend on time defined as a variable whose value can change every time time depends on time [18]. On cox extended model variable which depend to time must be interacted with the time function $g_m(t)$. The Extended Cox Regression which is formed in equation 9.

$$\mathrm{h(t,x(t))}=\mathrm{h}_0\mathrm{(t)}\exp[\sum_{k=1}^{p}\beta_x x_k+\sum_{m=1}^{q}\delta_m x_m g_m\mathrm{(t)}] \tag{9}$$

Information:

$\mathrm{h(t,x(t))}$ cox extended regression function

$\mathrm{h}_0\mathrm{(t)}$: basic hazard function

β_x : variable parameter that fulfills the assumption of PH
x_k : variable that fulfills the assumption of PH
δ_m : variable parameter that does not meet the assumption of PH
x_m : variable that does not meet the assumption of PH
$g_m(\mathrm{t})$: time function

2.3 Integration of Cluster Analysis With Survival Analysis

Integration of cluster analysis with survival analysis using extended cox regression based on a dummy variable approach. Dummy variables are included in binary variables that can change qualitative variables into quantitative ones. Usually the analyzed qualitative variables assume a value of 1 or 0 [19]. In this study, a Cluster . analysis aims to group objects so that between groups are heterogeneous and between objects in each group are homogeneous. Next, a survival analysis was performed using Extended Cox regression . Then the parameter estimation on the Cluster integration model with survival analysis was performed using Extended Cox regression . The integration of cluster analysis with survival analysis using Extended Cox regression is used to determine the factors that affect the length of time a group of customers are able to pay for a House Ownership Loan (KPR) program. The dummy variable in this study is the number of Cluster formed minus one. Cluster integration model in survival analysis using Extended Cox . regression written in equation (10).

$$\begin{aligned}
&\mathrm{h(t,x(t))} = \mathrm{h}_0(\mathrm{t})\exp[\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \beta_{p+1} D_{1i} X_{1i} + \beta_{p+2} D_{1i} X_{2i} + \ldots + \\
&\beta_{p+q} D_{1i} X_{pi} + \beta_{p+q+1} D_{2i} X_{1i} + \beta_{p+q+2} D_{2i} X_{2i} + \ldots + \beta_{p+2q} D_{2i} X_{pi} + \ldots + \beta_{p+qq+1} D_{qi} X_{1i} + \\
&\beta_{p+qq+2} D_{qi} X_{2i} + \ldots + \beta_{p+2qq} D_{qi} X_{pi} + \delta_1 X_{11i} g_m(\mathrm{t}) + \delta_2 X_{12i} g_m(\mathrm{t}) + \ldots + \delta_p X_{ri} g_m(\mathrm{t}) + \\
&\delta_{p+1} D_{1i} X_{11i} g_m(\mathrm{t}) + \ldots + \delta_{p+q} D_{1i} X_{ri} g_m(\mathrm{t}) + \delta_{p+q+1} D_{2i} X_{11i} g_m(\mathrm{t}) + \ldots + \delta_{p+2q} D_{2i} X_{ri} g_m(\mathrm{t}) \\
&+ \ldots + \delta_{p+qq+1} D_{qi} X_{ri} g_m(\mathrm{t}) + \ldots + \delta_{p+2qq} D_{qi} X_{ri} g_m(\mathrm{t})]
\end{aligned} \tag{10}$$

Information:

X_{pi} : the p-explanatory variable on the i-th observation unit
β : variable parameter that fulfills the assumption of PH
D_{qi} : the qth dummy variable on the i - th observation unit
δ : variable parameter that does not meet the assumption of PH
p : v variable that satisfies the assumption of PH
r : variable that does not meet the assumption of PH
q : quantity The cluster formed is reduced by 1
i : 1,2,3,…, n

For example, if the research variables used are 3 variables with 1 variable that does not meet the PH assumption and the number of is obtained Cluster is 2 Cluster , then 1 dummy is formed . Integrated cluster model with survival analysis using Extended Cox regression can be written as in equation n (11).

Common models:

$$\mathrm{h(t,x(t))} = \mathrm{h}_0(\mathrm{t})\exp[\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 D_{1i} X_{1i} + \beta_4 D_{1i} X_{2i} + \delta_1 X_{3i} g_1(\mathrm{t}) + \delta_2 D_{1i} X_{3i} g_1(\mathrm{t})] \tag{11}$$

Cluster 1 (D = 0)

$$\mathrm{h(t,x(t))} = \mathrm{h}_0(\mathrm{t})\exp[\beta_1 X_{1i} + \beta_2 X_{2i} + \delta_1 X_{3i} g_1(\mathrm{t})] \tag{12}$$

Cluster 2 (D = 1)

$$\mathrm{h(t,x(t))} = \mathrm{h}_0(\mathrm{t})\exp[(\beta_1 + \beta_3) X_{1i} + (\beta_2 + \beta_4) X_{2i} + (\delta_1 + \delta_2) X_{3i} g_1(\mathrm{t})] \tag{13}$$

2.4 Determining The Best Model

Determination of the best integration model using Mean Square d Error (MSE) . According to [20], MSE is a method that plays a role in determining the best model by squaring each error or error . Then add up and divide by the number of observations. One of the characteristics of a good model is that it has the smallest MSE . The MSE value can be seen in equation (14) below.

$$MSE = \frac{\sum_{i=1}^{n}(\mathrm{Y}_i - \widehat{\mathrm{Y}}_i)^2}{n-k} \tag{14}$$

Where,

Y_i : actual data

$\widehat{\mathrm{Y}}_i$: forecasting data

n : the number of observational data

k : number of variables

This research uses data about 5C (Character, Capacity, Capital, Collateral and Condition), credit collectibility, and credit payment time (time). The sample used is 300 bank X customers who are Home Ownership Credit (KPR) customers. This research begins by determining the research variables used and preparing the data. Then test the proportional hazard assumption. After that, perform Cluster Analysis using complete, and ward linkage. So that a Dummy variable matrix can be formed by using the formula for the number of clusters minus one. Next, it is necessary to integrate cluster analysis with survival analysis on each linkage. The integration model of cluster analysis and survival analysis obtained using extended cox regression is as follows.

$$\begin{aligned}\mathrm{h}(\mathrm{t},\mathrm{x}(\mathrm{t})) = \mathrm{h}_0(\mathrm{t})\exp[&\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \beta_{p+1} D_{1i} X_{1i} + \beta_{p+2} D_{1i} X_{2i} + \ldots + \beta_{p+q} D_{1i} X_{pi} + \beta_{p+q+1} D_{2i} X_{1i} \\ &+ \beta_{p+q+2} D_{2i} X_{2i} + \ldots + \beta_{p+2q} D_{2i} X_{pi} + \ldots + \beta_{p+qq+1} D_{qi} X_{1i} + \beta_{p+qq+2} D_{qi} X_{2i} + \ldots + \beta_{p+2qq} D_{qi} X_{pi} + \delta_1 X_{11i} g_m(\mathrm{t}) \\ &+ \delta_2 X_{12i} g_m(\mathrm{t}) + \ldots + \delta_p X_{ri} g_m(\mathrm{t}) + \delta_{p+1} D_{1i} X_{11i} g_m(\mathrm{t}) + \ldots + \delta_{p+q} D_{1i} X_{ri} g_m(\mathrm{t}) + \delta_{p+q+1} D_{2i} X_{11i} g_m(\mathrm{t}) + \ldots \\ &+ \delta_{p+2q} D_{2i} X_{ri} g_m(\mathrm{t}) + \ldots + \delta_{p+qq+1} D_{qi} X_{ri} g_m(\mathrm{t}) + \ldots + \delta_{p+2qq} D_{qi} X_{ri} g_m(\mathrm{t})]\end{aligned}$$

Survival analysis was carried out using the Extended Cox Regression model without being integrated with Cluster analysis. The model formed is as follows.

$$\mathrm{h}(\mathrm{t},\mathrm{x}(\mathrm{t})) = \mathrm{h}_0(\mathrm{t})\exp[\sum_{k=1}^{p} \beta_x x_k + \sum_{m=1}^{q} \delta_m x_m g_m(\mathrm{t})] \; .$$

The best model is obtained by looking at the MSE value. Furthermore, the interpretation of the formed model is carried out.

3 Results and Discussion

3.1. Testing the Proportional Hazard Assumption

PH assumption testing is done by Global Test. So that the results are obtained as in Table 1. below this.

Table 1. Proportional Hazard Assumption Test

NO.	VARIABLE INDICATOR	GLOBAL TEST VALUE	CONCLUSION	NO.	VARIABLE INDICATOR	GLOBAL TEST VALUE	CONCLUSION
1	X111	0.920	Meets the assumption of PH	14	X323	0.520	Meets the assumption of PH
2	X112	0.910	Meets the assumption of PH	15	X33	0.570	Meets the assumption of PH
3	X12	0.960	Meets the assumption of PH	16	X34	0.550	Meets the assumption of PH
4	X21	0.570	Meets the assumption of PH	17	X351	0.610	Meets the assumption of PH
5	X22	0.290	Meets the assumption of PH	18	X352	0.680	Meets the assumption of PH
6	X23	0.175	Meets the assumption of PH	19	X353	0.034	Does not meet the assumption of PH

NO.	VARIABLE INDICATOR	GLOBAL TEST VALUE	CONCLUSION	NO.	VARIABLE INDICATOR	GLOBAL TEST VALUE	CONCLUSION
7	X241	0.152	Meets the assumption of PH	20	X354	0.270	Meets the assumption of PH
8	X242	0.111	Meets the assumption of PH	21	X36	0.158	Meets the assumption of PH
9	X251	0.460	Meets the assumption of PH	22	X37	0.174	Meets the assumption of PH
10	X252	0.660	Meets the assumption of PH	23	X411	0.573	Meets the assumption of PH
11	X31	0.250	Meets the assumption of PH	24	X412	0.251	Meets the assumption of PH
12	X321	0.570	Meets the assumption of PH	25	X5	0.460	Meets the assumption of PH
13	X322	0.230	Meets the assumption of PH				

All variables meet the PH assumption, except for one variable that does not meet the PH assumption, namely the X353 variable which has a global value (0.034) less than 0.05. This indicates that the X353 variable is a variable that depends on the timing of House Ownership Loan (KPR) program payments. So when making the Extended Cox regression model , the variable need to interact with time or depend on time.

3.2. Cluster Analysis

3.2.1. Determination of the Number of Clusters

The following are the results of the Cluster on the Ward Linkage Method . The dendogram formed is as shown in Figure 1. the following.

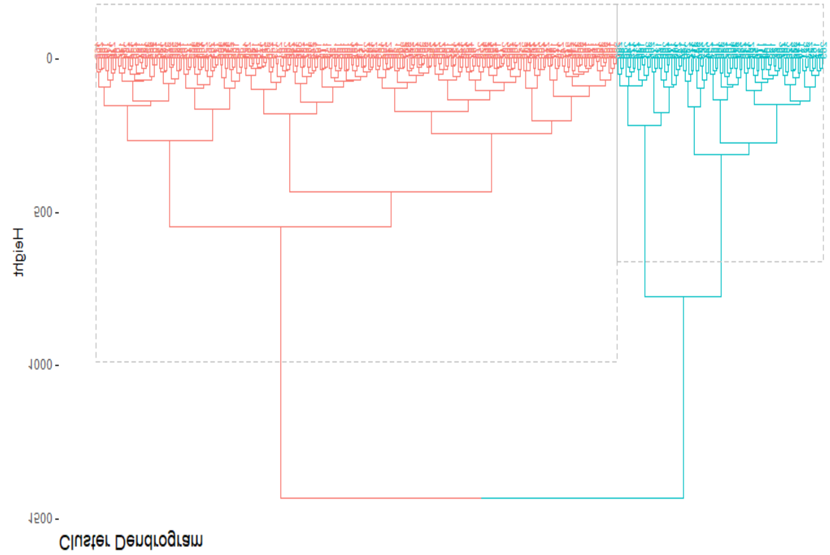

Fig.1. Dendogram from cluster analysis using the Ward linkage method

The number of members of each cluster according to Figure 1. use of Ward linkage method are 215 customers in cluster 1 and 85 customers in cluster 2. Meanwhile, the Dendogram formed is as shown in Figure 2. is cluster analysis using the

Complete linkage method.

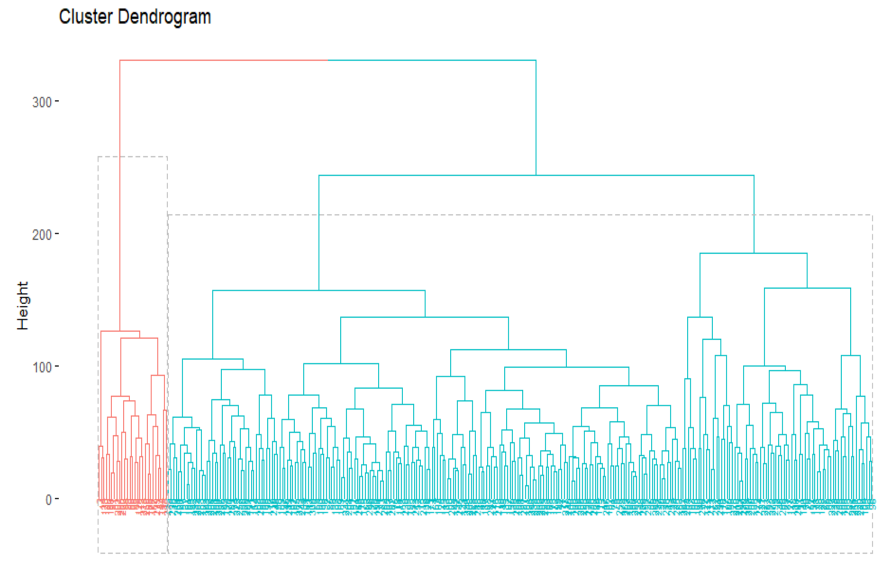

Fig2. Dendogram from cluster analysis using the Complete linkage method

The number of members of each cluster according to Figure 2. Using the Complete linkage method are 273 customers in cluster 1 and 27 customers in cluster 2.

3.3. Cox Extended Regression Model

Extended Cox Regression Model which is formed without being integrated with Cluster analysis is as follows.

$$\mathrm{h(t,x(t))}=\mathrm{h}_0\mathrm{(t)}\exp[-0,182X_{111i}+0,007X_{112i}+0,003X_{12i}+0,008X_{21i}+0,175X_{22i}-0,001X_{23i}+0,024X_{241i}+0,062X_{242i}+0,203X_{251i}$$
$$+0,192X_{252i}+0,082X_{31i}-1,011X_{321i}-0,974X_{322i}-0,974X_{323i}+0,016X_{33i}+0,086X_{34i}-0,290X_{351i}-0,313X_{352i}-0,233X_{354i}+0,002X_{36i}$$
$$-0,084X_{37i}+0,365X_{411i}+0,864X_{412i}+0,004X_{5i}-1,201X_{353i}(t)]$$

The MSE value of the extended cox regression model above is 0.330.

3.4. Cluster Integration with Dummy Variables with Survival Analysis Approach

3.4.1. Integrated cluster model in Extended Cox Regression Ward Linkage

The model formed from the Integration of Cluster Analysis using the Ward linkage method with Survival Analysis is as follows.

$$\mathrm{h(t,x(t))}=\mathrm{h}_0\mathrm{(t)}\exp[-0,115X_{111i}+0,07X_{112i}-0,005X_{12i}+0,006X_{21i}+0,174X_{22i}+0,002X_{23i}$$
$$-0,016X_{241i}+0,072X_{242i}+0,264X_{251i}+0,157X_{252i}-0,083X_{31i}-0,985X_{321i}-0,821X_{322i}$$
$$-0,957X_{323i}+0,035X_{33i}+0,145X_{34i}-0,237X_{351i}-0,398X_{352i}-0,114X_{354i}+0,006X_{36i}$$
$$-0,128X_{37i}+0,609X_{411i}+1,326X_{412i}+0,009X_{5i}-1,187X_{353i}(t)-0,122D_{1i}X_{111i}+0,006D_{1i}X_{112i}$$
$$-0,006D_{1i}X_{12i}-0,166D_{1i}X_{21i}-0,002D_{1i}X_{22i}+0,029D_{1i}X_{23i}-0,043D_{1i}X_{241i}-0,291D_{1i}X_{242i}$$
$$-0,171D_{1i}X_{251i}+0,096D_{1i}X_{252i}+5,417D_{1i}X_{31i}+5,239D_{1i}X_{321i}+5,361D_{1i}X_{322i}-0,036D_{1i}X_{323i}$$
$$-0,148D_{1i}X_{33i}+0,149D_{1i}X_{34i}+0,291D_{1i}X_{351i}-0,070D_{1i}X_{352i}-0,006D_{1i}X_{354i}+0,129D_{1i}X_{36i}$$
$$-0,550D_{1i}X_{37i}-1,158D_{1i}X_{411i}-0,009D_{1i}X_{412i}]$$

The model of Low cluster or ($D_1 = 0$) formed:

$$\mathrm{h(t,x(t))}=\mathrm{h}_0\mathrm{(t)}\exp[-0,115X_{111i}+0,07X_{112i}-0,005X_{12i}+0,006X_{21i}+0,174X_{22i}+0,002X_{23i}-0,016X_{241i}+$$
$$0,072X_{242i}+0,264X_{251i}+0,157X_{252i}-0,083X_{31i}-0,985X_{321i}-0,821X_{322i}-0,957X_{323i}+0,035X_{33i}+0,145X_{34i}$$
$$-0,237X_{351i}-0,398X_{352i}-0,114X_{354i}+0,006X_{36i}-0,128X_{37i}+0,609X_{411i}+1,326X_{412i}+0,009X_{5i}-1,187X_{353i}(t)]$$

The model of high Cluster or ($D_1 = 1$) formed:

$$
\begin{aligned}
\mathrm{h(t,x(t))} = \mathrm{h_0(t)}\exp[&-0,237X_{111i}+0,076X_{112i}-0,011X_{12i}-0,16X_{21i}+0,172X_{22i}+0,031X_{23i}-0,059X_{241i}\\
&-0,219X_{242i}+0,093X_{251i}+0,253X_{252i}+5,334X_{31i}+4,254X_{321i}+4,54X_{322i}-0,993X_{323i}-0,113X_{33i}\\
&+0,294X_{34i}-0,054X_{351i}+0,468X_{352i}-0,12X_{354i}+0,135X_{36i}-0,678X_{37i}-0,549X_{411i}+1,317X_{412i}\\
&+0,009X_{5i}-1,187X_{353i}(t)]
\end{aligned}
$$

The MSE value of this integration model is 0.265.

3.4.2. Integrated cluster model in Extended Cox Regression Complete Linkage

The model formed from the Integration of Cluster Analysis using the Complete linkage method with Survival Analysis is as follows.

$$
\begin{aligned}
\mathrm{h(t,x(t))} = \mathrm{h_0(t)}\exp[&-0,139X_{111i}+0,026X_{112i}-3,839e-04X_{12i}+0,004X_{21i}+0,159X_{22i}\\
&-0,001X_{23i}-0,043X_{241i}+0,015X_{242i}+0,169X_{251i}+0,220X_{252i}-0,006X_{31i}-0,975X_{321i}\\
&-0,905X_{322i}-0,937X_{323i}+0,021X_{33i}+0,099X_{34i}-0,288X_{351i}-0,376X_{352i}-0,233X_{354i}\\
&+0,005X_{36i}-0,086X_{37i}+0,537X_{411i}+1,173X_{412i}+0,005X_{5i}-1,203X_{353i}(t)\\
&-0,092D_{1i}X_{111i}+0,002D_{1i}X_{112i}-6,345e-04D_{1i}X_{12i}-1,047D_{1i}X_{21i}-0,003D_{1i}X_{22i}\\
&+0,044D_{1i}X_{23i}-0,059D_{1i}X_{241i}-0,263D_{1i}X_{242i}-0,293D_{1i}X_{251i}+0,041D_{1i}X_{252i}\\
&+17,850D_{1i}X_{31i}+17,740D_{1i}X_{321i}+17,770D_{1i}X_{322i}-0,021D_{1i}X_{323i}-0,107D_{1i}X_{33i}\\
&+0,165D_{1i}X_{34i}+0,195D_{1i}X_{351i}-0,119D_{1i}X_{352i}-0,004D_{1i}X_{354i}+0,114D_{1i}X_{36i}\\
&-0,402D_{1i}X_{37i}-0,005D_{1i}X_{412i}]
\end{aligned}
$$

When Low cluster ($\mathrm{D_1} = 0$) , then the model formed:

$$
\begin{aligned}
\mathrm{h(t,x(t))} = \mathrm{h_0(t)}\exp[&-0,139X_{111i}+0,026X_{112i}-3,839e-04X_{12i}+0,004X_{21i}+0,159X_{22i}\\
&-0,001X_{23i}-0,043X_{241i}+0,015X_{242i}+0,169X_{251i}+0,220X_{252i}-0,006X_{31i}-0,975X_{321i}\\
&-0,905X_{322i}-0,937X_{323i}+0,021X_{33i}+0,099X_{34i}-0,288X_{351i}-0,376X_{352i}-0,233X_{354i}\\
&+0,005X_{36i}-0,086X_{37i}+0,537X_{411i}+1,173X_{412i}+0,005X_{5i}-1,203X_{353i}(t)]
\end{aligned}
$$

When Cluster is high ($\mathrm{D_1} = 1$) , then the model formed is:

$$
\begin{aligned}
\mathrm{h(t,x(t))} = \mathrm{h_0(t)}\exp[&-0,231X_{111i}+0,028X_{112i}-0,001X_{12i}-1,043X_{21i}+0,156X_{22i}+0,043X_{23i}\\
&-0,102X_{241i}-0,248X_{242i}+0,124X_{251i}+0,261X_{252i}+17,844X_{31i}+16,765X_{321i}+16,865X_{322i}\\
&-0,958X_{323i}-0,086X_{33i}+0,264X_{34i}-0,093X_{351i}-0,495X_{352i}-0,237X_{354i}+0,119X_{36i}\\
&-0,488X_{37i}+0,537X_{411i}+1,168X_{412i}+0,005X_{5i}-1,203X_{353i}(t)]
\end{aligned}
$$

The MSE value of the model equation is 0.308.

3.5. Selection of the Best Linkage with MSE Value

The smallest MSE value is chosen to get the best model. Table 2. inform the MSE value of the three models, then compared.

Table 2. MSE Value Comparison

Model	**MSE value**
Model of Cox Extended Regression Equation	0.330
Model of Ward Linkage -based Extended Cox Regression Equation	**0.265**
Model of Complete Linkage -based Extended Cox Regression Equation	0.308

The Mean Squared Error value of Ward Linkage-based Extended Cox Regression is 0.265. This is the lowest value than others. So, the model is considered very good for describing the model. This indicates that the integration of cluster analysis and survival analysis has a better output than the model without integration.

3.6. Interpretation of the Best Model

The best model is obtained from the Ward Linkage-based Extended Cox Regression Equation Model. This is known from the lowest MSE value compared to the other 2 models. The following are the average indicators in the low and high clusters that are formed before.

Table 3. Average Score of Each Cluster

Indicator	Cluster 1	Cluster 2
	Low Cluster	High Cluster
	Average	Average
X_{12} : Length of Residence (Years)	8,635	7,700
X_{22} : Education (Years)	15,390	15,740
X_{23} : Age (Years)	39,220	39,980
X_{33} : Credit Term (Years)	10,690	12,800
X_{34} : RPA (Instalment Income Ratio)	2,218	2,676
X_{36} : Work Experience (Months)	41,990	142,200
X_{37} : Number of Family Dependents (Persons)	1,493	1,235
X_5 : Loan to Value	82.530	81,060

The model formed from the variables that affect the length of time customers pay for House Ownership Loan programs is significant in the high cluster as follows.

$$\mathrm{h}(\mathrm{t},\mathrm{x}(\mathrm{t})) = \mathrm{h}_0(\mathrm{t})\exp[-0,237X_{111i} + 0,172X_{22i} - 0,014X_{251i} + 0,253X_{252i} - 0,083X_{31i}$$
$$-0,113X_{33i} + 0,294X_{34i} + 0,135X_{36i} - 0,678X_{37i} + 0,009X_{5i}$$

Kaplan Meier Curve for Marriage Status of House Ownership Loan from customer data classified as high cluster as follows.

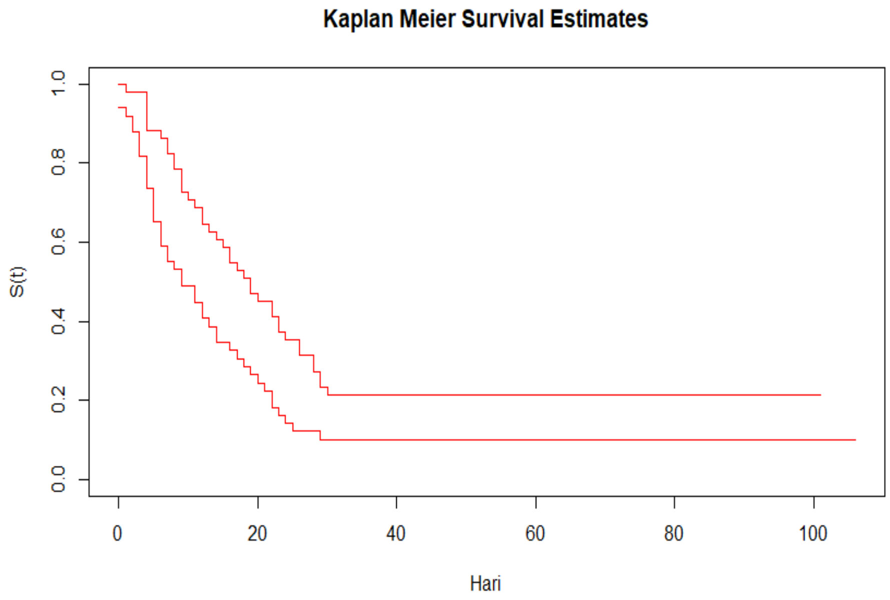

Fig. 3. Kaplan Meier Survival Curve High Cluster Customer Marriage Status

From the picture above explained that the chance of survival or customers fulfilling obligations with married status in paying House Ownership Loan program installments was 0.216 higher than the chances of survival or customers fulfilling obligations with unmarried status in paying House Ownership Loan program installments which was 0.102.

While the model in the low cluster is as follows.

$$h(t,x(t)) = h_0(t)\exp[-0,115X_{111i} + 0,174X_{22i} + 0,264X_{251i} + 0,157X_{252i}$$
$$-0,083X_{31i} + 0,035X_{33i} + 0,145X_{34i} + 0,006X_{36i} - 0,128X_{37i} + 0,009X_{5i}]$$

Kaplan Meier Curve for Marriage Status of customers from customer data classified as low cluster as follows.

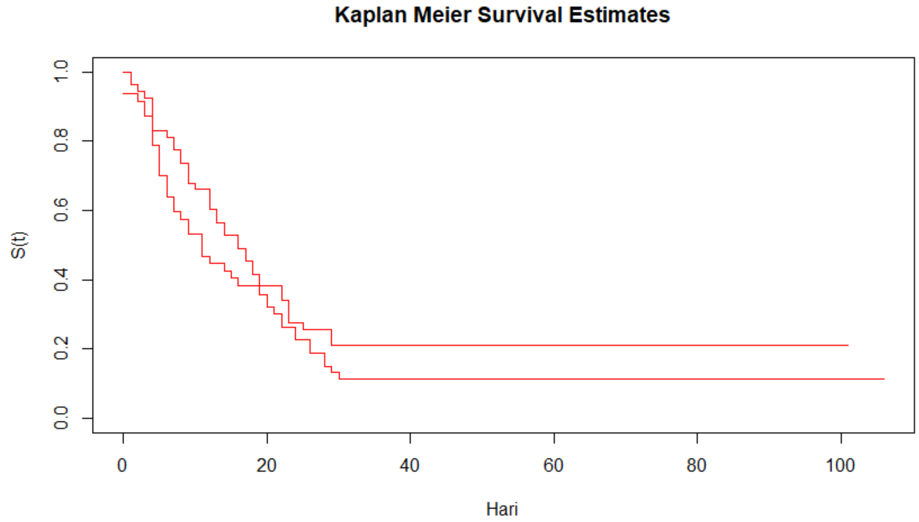

Fig 4. Kaplan Meier Survival Curve Low Cluster Customer Marriage Status

From the picture above, it is clear that the chance of survival or customers fulfilling obligations with marital status in paying House Ownership Loan program installments is 0.213 higher than the chances of survival or customers fulfilling obligations with unmarried status in paying House Ownership Loan program installments, which is 0.113.

4 Conclusions

The best model is the integrated cluster model in the ward linkage-based extended cox regression with the smallest MSE value compared to other models, which is 0.265. There are two best clusters formed, namely the low cluster and the high cluster. The best clusters formed are low clusters and high clusters. In the low cluster, the coefficient of the SHGB Guarantee Document (X_{111}) , Joint Income (X_{31}) , and Number of Family Dependents (X_{37}) are negative coefficient, so if this value increases it will increase the chances of customers in low clusters pay House Ownership Loan programs with a longer payment period than the time set by the bank (customers do not fulfill obligations). While the coefficient Education (X_{22}), Marital Status Divorced (X_{251}), Marital Status (X_{252}), Credit Term (X_{33}), Installment Income Ratio (X_{34}), Work Experience (X_{36}), and Loan to Value (X_5) are positive coefficient, meaning that the higher the coefficient value, the greater the chances of customers in low clusters pay the House Ownership Loan (KPR) program with a payment time that is faster than the time set by the bank or not past the time limit determined by the bank (the customer fulfills obligations). While in the high cluster, the coefficients of the SHGB Guarantee Document (X_{111}), Marital Status Divorced (X_{251}), Joint Income (X_{31}), Credit Term (X_{33}), and Number of Family Dependents (X_{37}) are negative coefficient. So, it can be concluded that if the coefficient value increases, it will increase the chances of customers in the cluster high paying House Ownership Loan programs with payment times longer than the time set by the bank (customers do not fulfill obligations). While the coefficients with positive values such as Education (X_{22}), Marital Status (X_{252}), Installment Income Ratio (X_{34}), Work Experience (X_{36}), and Loan to Value (X_5) have the meaning that the higher the value coefficient, it will increase the chances of customers in the cluster high paying House Ownership Loan (KPR) programs with payment times faster than the time set by the bank or not past the time limit determined by the bank (customer fulfills obligations). For high and low customer groups, married customers have a higher chance of fulfilling the obligations of paying House Ownership Loan installments.

Conflicts of Interest Statement

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript. Thank you to all parties for their support and input in the preparation of this research. All the support and input given is very useful for the perfection of this research.

References

[1] Raharjo, K., Nurjannah, N., Solimun, S., & Fernandes, A. A. R. The influence of organizational culture and job design on job commitment and human resource performance. Journal of Organizational Change Management. (2018).

[2] Fernandes, A. A. R. and Solimun. Moderating effects orientation and innovation strategy on the effect of uncertainty on the performance of business environment. *International Journal of Law and Management*, *59*(6), 1211-1219, (2017).

[3] Fernandes, A. A. R., Budiantara, I. N. I., Otok, B. W., & Suhartono. Reproducing Kernel Hilbert space for penalized regression multi-predictors: Case in longitudinal data. *International Journal of Mathematical Analysis*, *8*(40), 1951-1961, (2014).

[4] Solimun; Fernandes, Adji Achmad Rinaldo. Investigation of Instrument Validity: Investigate the Consistency between Criterion and Unidimensional in Instrument Validity (Case Study in Management Research). Int'l JL & Mgmt., 59, 1203. (2017).

[5] Sumardi, S., & Fernandes, A. A. R. The influence of quality management on organization performance: service quality and product characteristics as a medium. Property Management. (2020).

[6] Benny Hutahayan, A. A. R. F. S. N. Comparison of use of Linkage in Integrated Cluster with Discriminal Analysis Approach. International Journal of Advanced Science and Technology, 29(3), 5654 – 5668, (2020).

[7] Fernandes, A. A. R., & Taba, I. M. Welding technology as the moderation variable in the relationships between government policy and quality of human resources and workforce competitiveness. Journal of Science and Technology Policy Management. (2018).

[8] Agusta, Y. K-means–penerapan, permasalahan dan metode terkait. Jurnal Sistem dan informatika, 3(1), 47-60, (2007).

[9] Fernandes, A. A. R., Hutahayan, B., Arisoesilaningsih, E., Yanti, I., Astuti, A. B., & Amaliana, L. Comparison of Curve Estimation of the Smoothing Spline Nonparametric Function Path Based on PLS and PWLS In Various Levels of Heteroscedasticity. In *IOP Conference Series: Materials Science and Engineering* (Vol. 546, No. 5, p. 052024). IOP Publishing. (2019).

[10] Fernandes, S., & Rinaldo, A. A. R. A. A. The mediating effect of service quality and organizational commitment on the effect of management process alignment on higher education performance in Makassar, Indonesia. Journal of Organizational Change Management, 31(2), 410-425. (2018).

[11] Inayati, K. D., and Purnami, S. W. Analisis Survival Nonparametrik Pada Pasien Kanker Serviks di RSUD Dr.

Soetomo Surabaya Menggunakan Metode Kaplan Meier dan Uji Log Rank. *Jurnal Sains dan Seni ITS*. (2016).

[12] Fernandes, A. A. *Pemodelan Statistika Pada Analisis Reliabilitas dan Survival* . Malang: Brawijaya Press. (2016).

[13] Hair, J., Black, B., and Anderson, R. Multivariate Data Analysis. Upper Saddle River: Prentice Hall. (2014).

[14] Mattjik, A., and Sumertajaya, L. Sidik Peubah Ganda dengan Menggunakan SAS. Bogor: IPB Press. (2011).

[15] Johnson, R., and Winchern, D. Applied Multivariate Statistical Analysis, fifth edition. USA: Prentice Hall Englewood Cliffs, N.J. (2002).

[16] Collet, D. Modeling Survival Data in Medical Research Second Edition. London: Chapman and Hall. (2003).

[17] Purbawangsa, I. B. A., Solimun, S., Fernandes, A. A. R., & Rahayu, S. M. Corporate governance, corporate profitability toward corporate social responsibility disclosure and corporate value (comparative study in Indonesia, China and India stock exchange in 2013-2016). Social Responsibility Journal, 16(7), 983-999, (2019).

[18] Kleinbaum, D., and Klein, M. Survival Analysis, A Self-Learning Text. New York: Springer. (2005).

[19] Santoso, S. SPSS Versi 10: Mengolah Data Statistik Secara Profesional. Jakarta: PT Elex Media Komputindo. (2001).

[20] Gujarati, D. Ekonometrika Dasar. Erlangga, Jakarta. (2006).