Demand analysis of flood insurance by using logistic regression model and genetic algorithm

P Sidi*, M B Mamat*, Sukono3, S Supian3, A S Putra4

1Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Terbuka, INDONESIA
2Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, MALAYSIA
3Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, INDONESIA
4Department of Geophysics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, INDONESIA

*Corresponding author: pram@ecampus.ut.ac.id

Abstract. Citarum River floods in the area of South Bandung Indonesia, often resulting damage to some buildings belonging to the people living in the vicinity. One effort to alleviate the risk of building damage is to have flood insurance. The main obstacle is not all people in the Citarum basin decide to buy flood insurance. In this paper, we intend to analyse the decision to buy flood insurance. It is assumed that there are eight variables that influence the decision of purchasing flood assurance, include: income level, education level, house distance with river, building election with road, flood frequency experience, flood prediction, perception on insurance company, and perception towards government effort in handling flood. The analysis was done by using logistic regression model, and to estimate model parameters, it is done with genetic algorithm. The results of the analysis shows that eight variables analysed significantly influence the demand of flood insurance. These results are expected to be considered for insurance companies, to influence the decision of the community to be willing to buy flood insurance.

Keywords: Citarum River, building damage, floods insurance, logistic regression, and genetic algorithm.

1. Introduction

Flood disasters often occur in the Citarum River Basin in the area of South Bandung, Indonesia, and often cause a lot of damage to buildings. The government and several community agencies provide compensation in the form of assistance such as building equipment and building repair costs, for houses damaged [11]. But often the amount of compensation does not fully compensate, so there is a need to find ways to reduce the losses [13; 14]. The community participation to reduce the risk of building disasters is important, by raising awareness and capacity of the community. Because the community is a party who has direct experience in the event of a disaster, so that the understanding has become the capital for the reduction of risk of loss due to flood disaster [4]. In the context of reducing the risk of damage to buildings, insurance is seen as the right solution, because insurance can minimize losses due to the disaster [2]. Because insurance is a financial protection system (or financial compensation) for life, property, health and so on, getting reimbursed from unforeseen events, which involves regular
premium payments over a period of time in exchange policy to guarantee such protection [16; 17]. Indeed, people living in areas affected by Citarum river flood have realized, how important the role of insurance to reduce the risk of damage to buildings due to flood disaster. But until now, not all the people living in areas affected by the flood of the Citarum River are willing to insure their buildings. Therefore, analysis of purchasing decision of this flood insurance is very important to do.

Based on studies from several literatures, there are some researchers who have conducted an analysis of the purchase decisions of flood insurance. Among them, Esau [9], examines the factors that influence consumer purchasing decisions on insurance products in Manado Indonesia. This research uses Classic Assumption test model and multiple regression analysis. This model is used to verify and prove the research hypothesis. The analysis results show that product, price, promotion, people, and process together influence consumer purchase decision. Atreya et al. [1], conducted an analysis of what drives households to buy flood insurance in the Georgian region of the United States. The analysis was done using econometric model. The results show that more educated individuals, older, and African-Americans, all the same, are more likely to buy flood insurance. Ulbinaite et al. [19], examines the factors affecting insurance purchase decision making in Lithuania. In this research, factor analysis and multiple regression analysis are used to determine how factors are formed and how much their relative weights are. In general, it can be concluded that insurance purchasing decisions are affected by age, income, education level, household status, and number of family members. Similar studies have also been conducted by previous researchers, among others by Desrosiers [7], Mathur & Tripathi [10], Shaw [12], and Ulbinaite & Kucinskiene [18]. They all do research on the factors that influence the purchase decisions of insurance products.

Based on the above description, this paper intends to analyse the decision to buy flood insurance by using logistic regression model and genetic algorithm. The objective is to obtain a significant logistic regression model estimator, so it can be used to determine what variables significantly influence the purchase decision of flood insurance. This analysis is very useful in order to determining the socialization strategy about flood insurance products, which can reduce the risk of damage to buildings due to flooding. Especially for people living in the Citarum River Basin, South Bandung, Indonesia.

2. Materials and methods
This section discusses the materials and methods used in this study. Discussion of the methods used, including: logistic regression model, log likelihood function, genetic algorithm, and model significance test statistic.

2.1. Materials
This sub-section discuss about the material used in the analysis. The data used in this analysis is obtained from interviews with community representatives in the Citarum River flood disaster area in South Bandung area in 2016. The data consists of 100 samples, of which \( n_1 = 18 \) belong to category 1, which suffered damage to the building and decided to buy flood insurance, and \( n_0 = 82 \) belonging to category 0, which suffered damage to the building and decided not to buy flood insurance. Factors contributing to the decision to purchase flood insurance, is assumed that there are 8 factors, including: \( Z_1 \) family income; \( Z_2 \) level of education; \( Z_3 \) the distance of the house with the source of the flood; \( Z_4 \) building elevation by road; \( Z_5 \) the frequency of floods in a year; \( Z_6 \) prediction of future floods; \( Z_7 \) perceptions of insurance company services, and \( Z_8 \) perceptions of government efforts in tackling floods.

Against observation data, normality tests need to be done, because the non-free variable is the data whose values fluctuate between high and low. This is to avoid the estimation of biased model parameters. Normality test of observational data is done using SPSS statistical software.
2.2. Methods
This subsection discuss about the methodology used for the analysis. The discussion includes: logistic regression model, log likelihood function, genetic algorithm, and statistical test of model parameter significance.

2.2.1. Logistic Regression Model
Binary logistic regression model, is a regression model that can be used to estimate the effect of some independent variables \( Z_i \) (\( i = 1, \ldots, F \)) where \( F \) many independent variables), to non-free variables \( D \), which are binary values 0 and 1 [3; 15]. The binary regression model used in this analysis is shaped as follows [5]:

\[
\pi(Z_i) = \frac{\sum_{j=0}^{F} \beta_j Z_{ij}}{1 + e^{-\sum_{j=0}^{F} \beta_j Z_{ij}}}, \quad i = 1, \ldots, F. \tag{1}
\]

When the left and right segments of equation (1), are drawn natural logarithms, then we obtain a new equation of the following form [15; 8]:

\[
g(Z_i) = \ln\left( \frac{\pi(Z_i)}{1 - \pi(Z_i)} \right) = \sum_{j=0}^{F} \beta_j Z_{ij}, \quad i = 1, \ldots, F. \tag{2}
\]

To estimate the parameters of equation (2), it can be done using the likelihood method, as described below.

2.2.2. The Likelihood Function of the Logistic Regression Model
Suppose that the vector parameter model is known as \( \mathbf{\beta} = (\beta_0, \beta_1, \ldots, \beta_F) \). According to Sukono et al. [15], the purpose of estimating the regression model parameters is to determine the values \( \beta_j \) (\( j = 0, 1, \ldots, J \)) that contribute to equation (2). Suppose there are \( J \) independent variables, \( Z_1, \ldots, Z_J \) the conditional density function of the dependent variable \( D \) with the parameter vector \( \mathbf{\beta} \), is to follow the distribution of Bernoulli in the form of the following:

\[
f(D|\mathbf{\beta}) = \prod_{i=1}^{F} \pi_i^{D_i} (1 - \pi_i)^{1-D_i}; \quad D_i = 0 \text{ or } 1. \tag{3}
\]

Maximum Likelihood Estimator (MLE), is estimating the values of the parameter vectors \( \mathbf{\beta} \) that can maximize likelihood function in (3). The dependent variable \( D_i \) is given a value of 0 or 1, for each pair \( (Z_i, D_i) \). If given \( D_i = 1 \), then contributing to the likelihood function is \( \pi(Z_i) \), and if given value \( Z_i = 0 \), then contributing to the likelihood function is \( 1 - \pi(Z_i) \). Therefore, the contribution of the pair of points \( (Z_i, D_i) \) to the likelihood function can be written as follows \[15; 3\]:

\[
L(\mathbf{\beta}) = \prod_{i=1}^{F} \pi_i^{D_i} (1 - \pi_i)^{1-D_i}; \quad D_i = 0 \text{ or } 1. \tag{4}
\]

If equation (1) is substituted into (4), then the equation is given as follows:

\[
L(\mathbf{\beta}) = \prod_{i=1}^{F} \left[ e^{\sum_{j=0}^{F} \beta_j Z_{ij}} D_i \left( 1 + e^{-\sum_{j=0}^{F} \beta_j Z_{ij}} \right)^{-1} \right]^{-1}. \tag{5}
\]

If the left and right sides of the equation (5) take the natural logarithm, then the following equations are obtained [15]:

\[
\ell(\mathbf{\beta}) = \sum_{i=1}^{F} D_i \sum_{j=0}^{F} \beta_j Z_{ij} - \ln \left[ 1 + e^{\sum_{j=0}^{F} \beta_j Z_{ij}} \right]. \tag{6}
\]
The values of the elements of the parameter vector $\beta$ in equation (6) are estimated using the genetic algorithm, as follows.

2.2.3. Estimates Using Genetic Algorithms
Genetic algorithm is a genetic mechanism-based search method, and natural generation selection [6]. This genetic mechanism represents the individual's ability to engage in marriage, and produce offspring that have characteristics similar to that of the mother. While the generation selection naturally represents that living things can survive, when able to adapt to the surrounding environment [20]. Therefore, future generations are expected to have a combination of the best characteristics of the parent, and can sustain the next generations [6].

In general, the structure of the genetic algorithm is the following stages [15]:

a) Early population generation, generated randomly to obtain the initial solution of a dummy;
b) The resurrected population includes a number of chromosomes representing the solution obtained;
c) The establishment of a new generation, a new generation formed involving three operators, namely reproduction / selection, cross over, and mutation;
d) The evaluation process, each population raised is determined by the fitness value for each chromosome, and the evaluation is performed until the stopping criterion is reached. If the stopping criterion has not been reached, then a new generation needs to be established in a way like stage b).

Based on the structure of the genetic algorithm, to maximize the log likelihood function in equation (6), a generalized genetic algorithm can be prepared as follows [6; 20]:

1) The initial population determination, determined by the number of $J$, is generated randomly. This random number of the initial population generated is converted into decimal values $\beta_j$, where $j = 1, ..., J$;

2) Chromosome evaluation, the fitness value of chromosome is the value of log likelihood function as equation (6). The resulting fitness values are the largest, for the maximization program.

3) Calculation of the convergent population percentage, convergent population $c_p$, is the percentage of the number of individuals who can produce the same fitness value and the most. The $c_p$ value is calculated using the following formula:

$$c_p = \frac{n}{s_p} \times 100\%$$

where $n$ are the number of fitness-producing individuals which is the same and the largest, and $s_p$ is the total population.

4) Evaluation of the stopping condition, the genetic algorithm process is dismissed if the generation counter has reached the number of generations $s_g$, which is in the research is set at 1000, or the percentage of conventional population $s_p$ has reached the threshold limit which in this study is determined by $\tau = 90\%$.

5) Chromosome selection, the process is based on the principle of roulette wheel selection. Since the maximization program, the evaluation of the fitness value $\text{eval}(v_i), i = 1, ..., N$ is done referring to equation (6), i.e. by using the equation:

$$\text{eval}(v_i) = f(\beta)$$

where $f(\beta)$ is the fitness value referring to equation (6), and $\beta = (\beta_0, \beta_1, ..., \beta_J)$. 

6) Cross-breeding, the new population of the selection is done by cross-breeding using the Single-Point Crossover (SPC) method.

7) Mutations, each generation is done by counting $m \times s_p \times \text{size} \times p_m$, which is $m$ the number of mutations, the size of the population $s_p \times \text{size}$, and $p_m$ the probability of mutations whose value is determined randomly.
8) Decoding, represents the process of encoding genes in a chromosome to recover its original value, i.e. converting coding into decimal values. Furthermore, this genetic algorithm is used to analyse observation data about purchasing decision of damage insurance for buildings caused by Citarum river flood.

2.2.4. Test of Significance of Estimator of Logistic Regression Model

This section discusses the significance test of estimator logistic regression model. Testing is done which includes: verification test and validation test.

Verification Test. For verification test of parameter estimator of logistic regression model, this research is done by using Wald test statistic. According to [15], Hosmer and Lameshow in 1989, parameter verifier estimation $\beta_j$ ($j = 0, 1, ..., J$) testing can be performed using Wald's individual test statistics. The Wald test statistic used is following the standard normal distribution, as follows:

$$ \hat{t}_{Stat} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}; \quad j = 0, 1, ..., J, $$

(9)

where $\hat{\beta}_j$ is an estimator of parameter $\beta_j$ ($j = 0, 1, ..., J$), and $SE(\hat{\beta}_j)$ is standard deviation of parameter estimator $\hat{\beta}_j$. Hypothesis used for this Wald test is $H_0$: $\hat{\beta}_j = 0$, with alternative $H_1$: $\hat{\beta}_j \neq 0$ ($j = 0, 1, ..., J$). The test criterion is to reject $H_0$ if $\hat{t}_{Stat} < t_{\frac{1}{2}(1-\alpha)}$ or $\hat{t}_{Stat} > t_{\frac{1}{2}(\alpha)}$, in reverse, to accept $H_0$ if $t_{\frac{1}{2}(1-\alpha)} \leq \hat{t}_{Stat} \leq t_{\frac{1}{2}(\alpha)}$, where $t_{\frac{1}{2}(\alpha)}$ is the percentile of standard normal distribution with significance level of $(1-\alpha)\%$.

Validation Test. In this research, validation test is done by using Likelihood Ratio Test, Hosmer & Lameshow Test, and Test $R^2$ Determination.

2.2.5. Likelihood Ratio Test

For the validation of estimator estimation of logistic regression model parameters, this study was conducted using likelihood ratio test statistic $\hat{G}$, as follows [6]:

$$ \hat{G} = 2 \left[ \sum_{i=1}^{F} D_i \ln x_i + \sum_{i=1}^{F} (1-D_i) \ln(1-x_i) - n_1 \ln n_1 - n_0 \ln n_0 + F \ln F \right]. $$

(10)

Hypothesis used for likelihood ratio test is $H_0$: $\hat{\beta}_0 = \hat{\beta}_1 = ... = \hat{\beta}_J = 0$ ($j = 0, 1, ..., J$), with alternative $H_1$: $\exists \hat{\beta}_0 \neq \hat{\beta}_1 \neq ... \neq \hat{\beta}_J \neq 0$ ($j = 0, 1, ..., J$). Since statistic $\hat{G}$ is distributed asymptotically with Chi-Square $\chi^2(\alpha, df)$, therefore the criterion test used is to reject $H_0$, if $\hat{G} > \chi^2(\alpha, df)$, in reverse to accept $H_0$ if $\hat{G} \leq \chi^2(\alpha, df)$, where $\alpha$ is significance level, and $df = J - 1$ with $J$ is the number of logistic regression model parameters.

2.2.6. Hosmer & Lemeshow Test

The validation test of parameter estimator of logistic regression model can be done by using Hosmer & Lemeshow test statistic $\hat{H}$, as follows [15; 8]:

$$ \hat{H} = \sum_{j=1}^{J} \left( y_j - n_j x_j \right) \text{ atau } R_{Value} = Pr(\hat{H}), $$

(11)

where $y_j = \sum_{j=1}^{J} D_j$ dan $x_j = \sum_{j=1}^{J} (m_j \pi_j / n_j)$. Hypothesis for Hosmer & Lemeshow test is $H_0$: There is no difference between the observations and the model used; with alternatives $H_0$: There is a difference between the observations and the model used. Statistic Hosmer & Lemeshow is distributed
asymptotically with Chi-Square $\chi^2_{(\alpha, df)}$, therefore the criterion test used is to reject $H_0$ if $\hat{H} > \chi^2_{(\alpha, df)}$ or $P_{\text{Value}} < (1 - \alpha)$, in reverse to accept $H_0$ if $\hat{H} \leq \chi^2_{(\alpha, df)}$ or $P_{\text{Value}} \geq (1 - \alpha)$, where $\alpha$ is significance level, and $df = g - 2$ with generally $g = 10$.

2.2.7. Determination $R^2$

According to [15], Hosmer & Lemeshow use deterministic values in logistic regression models, to show the correlation strength between independent variables and non-free variables. Statistics of deterministic $R^2$ coefficients are calculated using the following equation:

$$ R^2 = 1 - \exp \left[ -\frac{\hat{i}(\beta)}{N} \right], \quad (12) $$

where $\hat{i}(\beta)$ is the maximum likelihood log value estimator in equation (6), and $N$ the number of observation data. When the value $R^2 \rightarrow 1$, then the correlation between the independent variable and the dependent variable is strong. Conversely, when value $R^2 \rightarrow 0$, then the correlation between the independent variable and the dependent variable is not strong.

The Methodology mentioned above, and then used to analyse observation data, as follows.

3. Materials and methods

In this section, we discussed about the data conducted analysis, analysis results, and discussion of the results of analysis, based on the data processed. The data analysed are as described in section 2.1 on the material.

The first step in multivariate analysis, before the data is done the analysis needs to be tested normality. The normality test is intended to ensure that the data is normally distributed. Since the data are good for multivariate analysis, the data is normally distributed. Testing of data normality, can be done using Anderson-Darling statistic with the help of software of Minitab 16. If data is normal distribution, then data can be used to process parameter estimation from binary logistic regression model.

3.1. Results

In this section we will discuss the results of parameter estimation from binary logistic regression model. The result of normality test to the data conducted by the analysis shows that the data follows the normal distribution. Therefore, the data can then be used for the process of parameter estimation, which in this study is expressed as a vector parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$. The parameter estimate to be obtained is parameters that maximize the log likelihood function value in equation (6). In the process of parameter estimation, in this paper is done by using genetic algorithm. The estimation process using genetic algorithm is done with Matlab 7.0 software. The result of the parameter estimation process is given in Table 1.

Towards the estimators of the coefficient parameters given in Table 1, it is necessary to verify and validate. Verification and validation testing is conducted to ensure that the parameter estimators are significant, either individually or in whole against the logistic regression model estimator.

Partial verification test of parameter estimator of logistic regression model is done by using Wald statistic, referring to equation (9). The hypothesis used in the Wald test is $H_0 : \hat{\beta}_j = 0$, with an alternative $H_1 : \hat{\beta}_j \neq 0$ ($j = 0, 1, ..., 8$). If the significance level is determined by $\alpha = 0.95$, then from the standard normal distribution table, the percentile values $t_{\frac{1}{2}(1-0.95)} = -0.27$ and $t_{\frac{1}{2}(0.95)} = 0.27$ are obtained.
Looking at the results given in Table 1, column (d), because the values of $\hat{t}_j$ ($j = 0, 1, \ldots, 8$), $\hat{t}_j < t_{1-0.95}^{\frac{1}{2}}$, or $\hat{t}_j > t_{1-0.95}^{\frac{1}{2}}$, then reject the hypothesis $H_0$. This means that all parameter estimators of logistic regression models are significant.

**Table 1. Parameter estimator and standard error**

| Factors of $(Z_j)$ | Parameters Estimator $(\hat{\beta}_j)$ | Standard Error $SE(\hat{\beta}_j)$ | Ratio $\hat{t}_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$ | Prob$(\hat{t}_j)$ | Significances |
|-------------------|----------------------------------------|-----------------------------------|-----------------------------------------------|-----------------|---------------|
| $(a)$             | $(b)$                                  | $(c)$                             | $(d)$                                         | $(e)$           | $(f)$         |
| $Z_1$             | -0.202521                              | 0.084525                          | -2.395989                                     | 0.000002        | Significance  |
| $Z_2$             | 3.914214                               | 1.319371                          | 2.966727                                      | 0.000031        | Significance  |
| $Z_3$             | 0.045176                               | 0.013214                          | 3.418798                                      | 0.000000        | Significance  |
| $Z_4$             | 0.065321                               | 0.013523                          | 4.830363                                      | 0.000000        | Significance  |
| $Z_5$             | -0.018932                              | 0.003261                          | -5.805581                                     | 0.000003        | Significance  |
| $Z_6$             | -2.982538                              | 1.058321                          | -2.818179                                     | 0.000001        | Significance  |
| $Z_7$             | 1.716352                               | 0.83479                           | 2.056028                                      | 0.000000        | Significance  |
| $Z_8$             | -2.964273                              | 1.161934                          | -2.551154                                     | 0.000051        | Significance  |
| **Constant**      | -1.813250                              | 0.503184                          | -3.603553                                     | 0.000024        | Significance  |

Log likelihood statistic $\hat{G} = 24.753526$

Furthermore, the validation test is performed on the logistic regression model parameter estimators, it is intended to test that the parameters of the logistic regression model parameters together are significant influencing $\pi(Z)$. Validation test of parameter estimators of logistic regression model, in this paper is done by using Likelihood Ratio statistic referring to equation (10). The hypothesis used in this verification test is $H_0: \hat{\beta}_0 = \hat{\beta}_1 = \ldots = \hat{\beta}_8 = 0$, with an alternative hypothesis $H_1: \hat{\beta}_0 = \hat{\beta}_1 = \ldots = \hat{\beta}_8 = 0$. If determined significance level of $\alpha = 0.95$ and $df = 8$, then from the Chi-Square distribution table obtained percentile value of $\chi^2_{(1-0.95),8} = 2.7326$. While statistics $\hat{G} = 24.753526$, it is clear that $\hat{G} > \chi^2_{(1-0.95),8}$. Therefore, the decision rejects the hypothesis $H_0$. This means that the parameter estimators of the logistic regression model, are of significant influence $\pi(Z)$ together.

The next validation test is performed using the Hosmer & Lemeshow test statistic, with reference to (11). The hypotheses for Hosmer & Lemeshow test is $H_0$: There is a difference between the observation data and the estimator model, with the alternative hypothesis $H_1$: There is no difference between the observation data and the estimator model. Hosmer & Lemeshow statistical tests can also be performed using statistics $R_{value}$, with reference to (11). In this test, obtained probability $R_{value} = 0.2561325$. When the significance level is set to $\alpha = 0.95$, it is clear that $R_{value} > (1 - \alpha)$. Therefore, the decision rejects the hypothesis $H_0$, which means that there is no difference between the observational data and the logistic regression model estimator.

Furthermore, to measure the strength of the correlation between independent and non-free variables, it can be done by using the coefficient of determination $R^2$, with reference to (12). Results of data processing obtained that the value of determination $R^2 = 0.784392$. This shows that there is a strong correlation between independent variables $Z_1, \ldots, Z_8$ and non-free variable $\pi(Z)$. 


3.2. Discussion
Taking into account the logistic regression model estimators given in Table 1 shows that:
(i) The value of estimated coefficient of variables for educational level, the distance of the house with
   the source of flood, the election of the house with the road, and the public perception of the insurance
   company, are positive and significant at the significance level of $\alpha = 0.95$. This illustrates that the
   higher level of education of people living in areas affected by the Citarum river floods, is very
   influential on the awareness to buy flood insurance. The closer the house, the closer to the source of
   the flood, and the position of the house building that is lower or closer to the same level with the high
   road in front of it, the greater awareness that the residential building has the risk of damaged by the
   flood. Similarly, the better the public perception of insurance companies, the more likely they will
   not mind buying flood insurance products.
(ii) The value of estimated variable coefficient for income level, the frequency of floods in a year, the
   predicted number of floods, and the perceptions of government efforts in tackling the flood are
   negative and significant at the significance level $\alpha = 0.95$. This shows that for people with low
   income levels, frequent flood frequencies within a year, predicted floods are frequent and negative
   perceptions of government payments to cope with floods, choose not to buy flood insurance products.
   Interview results explain those who have large funds, prefer to move the location or do the
   arrangement of the location of the house by raising the building.
Furthermore, when parameter estimators are given on equation (1), we obtain a logistic regression
model which can be used to estimate the probability of damage to the building due to Citarum river
flood. The equation estimator of the logistic regression model, with the value of variable coefficient
estimators rounded up to three decimal places, is as follows:
\[
\pi(Z) = \frac{e^{-1.813-0.202Z_1+3.914Z_2+0.045Z_3+0.065Z_4-0.019Z_5-2.983Z_6+1.716Z_7-2.964Z_8}}{1+e^{-1.813-0.202Z_1+3.914Z_2+0.045Z_3+0.065Z_4-0.019Z_5-2.983Z_6+1.716Z_7-2.964Z_8}}
\]  
(13)
For example, a person living in an area affected by Citarum river floods has the following categories:
income levels that are large or in category $Z_1 = 1$; education level is high or in category $Z_2 = 1$; the
distance of the house with the source of flood is relatively close or in category $Z_3 = 1$; the building
position is classified under the road in front of the house or in category $Z_4 = 1$; the frequency of floods
in a year is classified frequently or in category $Z_5 = 1$; predicted floods in the future will be classified as
often $Z_6 = 1$; public perception of insurance company is good or in category $Z_7 = 1$; and perceptions of
the government's efforts in dealing with floods are low or in category $Z_8 = 0$. The values of this category
factor when substituted into equation (13) together can be calculated the probability of damage to
buildings due to floods, as follows:
\[
\pi(Z) = \frac{e^{-1.813-0.202(1)+3.914(1)+0.045(1)+0.065(1)-0.019(1)-2.983(1)+1.716(1)-2.964(0)}}{1+e^{-1.813-0.202(1)+3.914(1)+0.045(1)+0.065(1)-0.019(1)-2.983(1)+1.716(1)-2.964(0)}} = 0.91157
\]
Thus, a person living in a flooded area of the Citarum River has a category as mentioned above; having
a probability of buying flood insurance products is 0.91157. The example, together with nine other
peoples, based on their respective characteristics, the probability prediction of the purchase decision of
the flood insurance product is given as in Table 2.
Based on the predicted probability of purchasing decision of the flood insurance product, the
insurance company should be able to optimally use, to increase the participation of the people living in
the affected area of Citarum river flood, to be willing to buy flood insurance products. Because for
people living in areas affected by Citarum river floods, buying flood insurance products can reduce the
risk of damage to buildings due to the Citarum river flood.
Table 2. Predicted Probability of purchasing decision of flood insurance products based on their respective characteristics

| Buildings | $Z_1$ | $Z_2$ | $Z_3$ | $Z_4$ | $Z_5$ | $Z_6$ | $Z_7$ | $Z_8$ | Probability |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------------|
| 1         | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 0.91157     |
| 2         | 1     | 1     | 1     | 1     | 0     | 0     | 1     | 0     | 0.99520     |
| 3         | 0     | 1     | 0     | 1     | 0     | 1     | 0     | 1     | 0.10239     |
| 4         | 0     | 1     | 0     | 0     | 1     | 0     | 1     | 0     | 0.99554     |
| 5         | 1     | 0     | 1     | 0     | 1     | 0     | 1     | 0     | 0.79200     |
| 6         | 1     | 0     | 1     | 1     | 0     | 1     | 0     | 1     | 0.00194     |
| 7         | 1     | 1     | 1     | 0     | 1     | 0     | 0     | 1     | 0.63899     |
| 8         | 1     | 0     | 1     | 0     | 1     | 0     | 1     | 0     | 0.16424     |
| 9         | 0     | 0     | 0     | 1     | 0     | 0     | 0     | 0     | 0.46555     |
| 10        | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0.44942     |

4. Conclusion
In this paper, we have been analysis of purchasing decision of flood insurance by using logistic regression model and genetic algorithm. Based on the analysis result, it can be concluded that the analysis of purchasing decision of flood insurance can be done by using logistic regression model. For estimation of logistic regression model parameters, it can be done using genetic algorithm. The estimation of logistic regression model parameters shows that eight factors $Z_1,...,Z_8$ significantly influence the probability function $\pi(Z)$ in equation (1). Using the equation estimator (1), for a person living in the area affected by Citarum river floods that have the following categories: income levels are large; the level of education is high; the distance of houses with the source of flood is quite close; the building position belongs to the bottom of the road in front of the house; the frequency of floods in a year is frequent; predictions of future floods are often classified; public perception of insurance companies is good; and perceptions of government efforts in the prevention of floods are low, have the probability of decision to buy flood insurance products amounted to 0.91157. Predicted probability of purchasing decision of flood insurance products, by the insurance company can be used optimally to increase public participation in purchasing flood insurance products.

Acknowledgments
Authors thank Rector, Director of DRPMI, and Dean of FMIPA, Universitas Padjadjaran, which provides a grant program of the Academic Leadership Grant (ALG) and a grant program of Competence Research of Lecturer of Unpad (Riset Kompetensi Dosen Unpad/RKDU).

References
[1] Atreya, A., S. Ferreira, and E. Michel-Kerjan, 2014, What Drives Households to Buy Flood Insurance Evidence from Georgia, Working Paper # 2014-08, Risk Management and Decision Processes Center, The Wharton School, University of Pennsylvania.
[2] Aliagha, A.G., T. E. Jin, W. W. Choong, M. Nadzri Jaafar, and H. M. Ali, 2014, Factors affecting flood insurance purchase in residential properties in Johor, Malaysia, Nat. Hazards Earth Syst. Sci., 14, 3297–3310, 2014, http://www.nat-hazards-earth-syst-sci.net/14/3297/2014/, doi:10.5194/nhess-14-3297-2014.
[3] Bayaga, A., 2010, Multinomial Logistic Regression: Usage and Application in Risk Analysis. Journal of Applied Quantitative Methods. Vol. 5, No. 2, Summer 2010, pp. 288-297.
[4] Ciumas, C. and R.A. Coca, 2015, An Analysis of the Factors Influencing the Demand for Catastrophe Insurance, Journal of Public Administration, Finance and Law, Special Issue 2/2015, pp. 69-78.
[5] Cramer, J.S., 2002, The Origins of Logistic Regression. Tinbergen Institute Discussion Paper, T1.2002-119/4. Faculty of Economics and Econometrics University of Amsterdam and Tinbergen
Institute, pp. 1-15.

[6] Demir, E. And O. Akkus, 2015, An Introductory Study on “How the Genetic Algorithm Works in the Parameter Estimation of Binary Logit Model?”. International Journal of Sciences: Basic and Applied Research (IJSBAR), Volume 19, No 2, pp 162-180.

[7] Desrosiers, M.A., 2012, How Individuals Purchase Insurance: Going Beyond Expected Utility Theory, Casualty Actuarial Society E-Forum, Winter 2012-Volume 2, pp. 1-18.

[8] Eshima, N., M. Tabata, and G. Zhi, 2001, Path Analysis with Logistic Regression Models Effect Analysis of Fully Recursive Causal Systems of Categorical Variables. J. Japan Statist. Soc. Vol 31, No. 1, pp. 1-14.

[9] Esau, E.Y.S, 2015, Factors Affecting Consumer Purchase Decision on Insurance Product in PT. Prudential Life Assurance Manado, Jurnal EMBA, Vol.3 No.3 Sept. 2015, pp.491-501, ISSN 2303-11.

[10] Mathur, A. and A. Tripathi, 2014, Factors Influencing Customer’s Choice for Insurance Companies-A Study of Ajmer City, IOSR Journal of Business and Management (IOSR-JBM), e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 16, Issue 2. Ver. 1 (Feb. 2014), pp. 35-43.

[11] Setiadi, H.A. 2013, Identification of Building Damage and Infrastructure Function due to Citarum Flood in Bandung Regency, Working Paper, Pusat Litbang Perumahan, Balitbang, Kementerian Pekerjaan Umum Jalan Panyawangan – Cileunyi Wetan, Kabupaten Bandung.

[12] Shaw, M.M., 2004, Group Flood Insurance Program and Flood Insurance Purchase Decisions, International Journal of Mass Emergencies and Disasters, August 2004, Vol. 22, No. 2, pp. 59–75.

[13] Sidi, P., M. Mamat, Sukono, and S. Supian, 2017a, Supply and demand analysis for flood insurance by using logistic regression model: case study at Citarum watershed in South Bandung, West Java, Indonesia, IOP Conf. Series: Materials Science and Engineering 166 (2017) 012002 doi:10.1088/1757-899X/166/1/012002

[14] Sidi, P., M. Mamat, Sukono, and S. Supian, 2017b, Insurance Premium Risk Evaluation Model for Building Damage Caused by Flood of the River: Case Study on Citarum Watershed in South Bandung Indonesia, Journal of Engineering and Applied Sciences 12 (17): 4420-4425.

[15] Sukono, A. Sholahuddin, M. Mamat, and K. Prafidya, 2014, Credit Scoring for Cooperative of Financial Services Using Logistic Regression Estimated by Genetic Algorithm. Applied Mathematical Sciences, Vol. 8, 2014, no. 1, pp. 45-57. http://dx.doi.org/10.12988/ams.2014.310600

[16] Sukono, Nahar, J., Putri, F. T., Subiyanto, Mamat, M., & Supian, S. 2017a. Estimation of outstanding claims reserving based on inflation risk on car insurance companies by using the bootstrap method. Far East Journal of Mathematical Sciences, 102(4), 687-706. doi:10.17654/MS102040687.

[17] Sukono, Suyudi, M., Islamiyati, F., & Supian, S. 2017b. Estimation model of life insurance claims risk for cancer patients by using bayesian method. Paper presented at the IOP Conference Series: Materials Science and Engineering, 166(1) doi:10.1088/1757-899X/166/1/012022

[18] Ulbinaite, A. and M. Kucinskiene, 2013a, Insurance Service Purchase Decision-Making Rationale: Expert-Based Evidence From Lithuania, Journal of EKONOMIKA 2013 Vol. 92(2), pp. 137-155.

[19] Ulbinaite, A., M. Kucinskiene, and Y. L. Moullac, 2013b, Determinants of Insurance Purchase Decision Making in Lithuania, Inzinerinė Ekonomika-Engineering Economics, 2013, 24(2), 144-159. http://dx.doi.org/10.5755/j01.ee.24.2.3439

[20] Yan, X., W. Luo, W.Li, W.Chen, C. Zhang, and H. Liu, 2013, An Improved Genetic Algorithm and Its Application in Classification. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 1, January 2013, pp. 337-346.