Classification method at acceptance of new student at public university on the national written test

Ika S W Antari*, Ismaini Zain and Suhartono

1Statistics Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia
*Corresponding Author: windiika73@gmail.com

Abstract. Acceptance of new students at public universities through the national written test is based on the total score and the capacity of the study program. This causes the study program accepts several students who have low scores on the main subject of the study program. The purpose of this study is to find the best method in predicting the probability of being accepted on the national written test and find the minimum score for each subject that must be achieved by participants to be accepted at a public university. There are two classification methods in statistics that are studied to overcome this problem, i.e. logistic regression and random forest. The results showed that the best logistic regression model had an accuracy of 97.11 percent, whereas the random forest method had an accuracy of 96.59 percent. Furthermore, the minimum score for each subject was developed based on the univariate logistic regression model.

1. Introduction

The national written test is the pattern of acceptance based on the results of the Print-Based Written Test (PBWT) or Computer-Based Written Test (CBWT), or a combination of national written tests and skills tests, which are under the coordination of the central committee [1]. Acceptance of new students based on the total score that has been weighted and restricted by the capacity on the study program. This process ignores the score of each test subject. This causes the study program accepts several students who have low scores on the main subject of the study program. Things like this can affect the quality of students. In this study, we want to know the minimum score for each subject that must be achieved by a participant to be accepted at public university of their choice. This is important to get output better student quality.

One of the statistical methods that function to find a model or describes the data classes is a classification method. In general, the classification method is a method of grouping an observation into a particular group that is in fit with the characteristics of the group [2]. Grouping or classification carried out with separating two groups by a function. The separating function as a delimiter to determining which observation will enter into a group with the same characteristic.

In statistics, there are many methods can be grouping of observations, such as logistic regression. The simplest logistic regression is a binary logistic regression method. Binary logistic regression is used to model data with binary response variables or dichotomous [3].

Anggraeni [4] conducted a study on the national written test applicants in ITS using the variables of test subjects, high school, regional, and gender by using the logistic regression method, obtained the accuracy of cloning by 97.176 percent. Previous studies that compared logistic regression (LR), classification and regression tree (CART), and neural networks (NNs) showed that the performance of
NNs was slightly better than that of LR and CART models, however it did not demonstrate a significant improvement [5].

Another method that usually used to classification is Random Forest. This method more flexible than logistic regression because it does not require to fulfill assumptions. Random forest is a development of the CART method developed by Breiman [6]. The main idea of the random forest method is to apply the bootstrap aggregating methods (bagging) and random feature selection [6]. In a random forest, many trees (CART) are grown, so that forming a forest, then the best tree is chosen. The random forest can make a more stable model, even for big data. Random forest also has fast advantages in computational iterations [7-8].

Geneur, Poggi, Malot, and Vialaneix [9] conducted research by using random forest for big data. Delgado, Cernadas, Barro, and Amorim [10] conducted a study of 179 classification methods from 17 family classifications in 121 datasets. They showed that of all the classification methods used, the best result is random forest.

2. Theoretical Review

2.1. Logistic Regression

Logistic regression is a method used to see the relationship of the response variable (Y) that is dichotomous (data with nominal or ordinal scale with two categories) or polychotomous (data with nominal or ordinal scale with more than two categories) with one or more predictor variables (X) which is continuous or category [11]. Binary logistic regression is used to model data with binary response variables or dichotomous [3].

In binary logistic regression, the response variable is data with a nominal scale with only two categories, namely “success” (y=1) or “fail” (y=0), for example yes-no, right-wrong, live-die, attendance, men and woman, and so on. In such condition, the y variable follows the Bernoulli distribution for each single observation. The probability function for each observation

\[ f(y) = \pi^y (1-\pi)^{1-y}; \quad y = 0,1. \]  

(1)

If \( y = 0 \) then \( f(y) = 1 - \pi \), and if \( y = 1 \) then \( f(y) = \pi \), with the logistic regression function

\[ f(z) = \frac{1}{1 + e^{-z}}, \]  

(2)

where \( z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \), and \( p \) = the number of predictor variables.

The \( z \) value is between \(-\infty \) and \( +\infty \), so the value of \( f(z) \) is between 0 and 1 for each given \( z \) value. This shows that the logistic model describes the probability or risk of an object or event. The binary logistic regression model is obtained

\[ \pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}. \]  

(3)

The logistic regression model in equation (3) can be described using logit transformation so that obtained the model

\[ g(x) = log \left( \frac{\pi(x)}{1- \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \]  

(4)

In logistic regression, the response variable is expressed as \( y = \pi(x) + \varepsilon \), where \( \varepsilon \) has one of two possible values, \( \varepsilon = 1 - \pi(x) \) with probability \( \pi(x) \) if \( y = 1 \), and \( \varepsilon = -\pi(x) \) with probability \( 1 - \pi(x) \) if \( y = 0 \) and follow the binomial distribution with mean is zero and variance is \([\pi(x)][1-\pi(x)]\).
2.2. Random Forest

The random forest method was first introduced by Breiman [6], which is a development of the CART method. The main idea of the random forest method is to apply the bootstrap aggregating methods (bagging) and random feature selection [6]. In a random forest, many trees (CART) are grown, so that they forming a forest, then the best tree is chosen. Illustration Random forest shown in figure 1.

![Illustration of Random Forest](image)

**Figure 1. Illustration of Random Forest**

The following is an illustration of the random forest with $n$ observation and $p$ predictor variables [9].

1. Taking $n$ random samples with replacement from the data training.
2. In each dataset of bootstrap results, a classification tree is constructed without pruning, with the best sorting of nodes used from predictor variables taken randomly.
3. Predict the classification of sample data based on the classification tree formed.
4. Repeat steps 1 to 3 until a number of classification trees are formed.
5. Predictive classification of the final sample data based on $k$ classification tree formed by using majority vote.

However, the classification tree in the random forest has two main differences with the classification tree on CART. First, in the classification tree forming, does not use all variables, but only uses partially taken randomly. This process aims to produce a collection of single classification trees with different sizes and shapes. The expected result is to obtain a small correlation value between trees. This small correlation results in small variance from the random forest result [12]. Second, in the random forest method, without tree pruning.

In the formation of random forest, the thing that can be changed is the value of $m$ (number of predictor variables are used). The value of $m$ which is too large will result in a greater correlation between trees. Conversely, if the value of $m$ is very small, it will produce a very low accuracy. The selection of values $m$ is very important in the formation of random forest. There are several values $m$ suggested by Breiman & Cutler [13], which is $m = \left\{ \frac{1}{2} \sqrt{p}, \sqrt{p}, 2 \sqrt{p} \right\}$, with $p$ is the number of dependent variables.

2.3. Accuracy of Classification

Calculation of classification accuracy is used to see how precisely the data grouping. A good classification model has a small misclassification. There are many ways that can be done to assess the accuracy of a calcification method, one of which is to calculate the 1-APER (apparent error rate). APER shows the proportion of samples that are incorrectly classified by its classification function. Determination of misclassification can be determined using a table of accuracy classificatory [2].
Table 1. Accuracy of Classification.

| Actual | Prediction | Total |
|--------|------------|-------|
|        | Y          |       |
| 0      | n_{00}     | n_{01}|
| 1      | n_{10}     | n_{11}|
| Total  | n_{-0}     | n_{-1}|

where

\[ n_{00} : \text{Total observation from } Y = 0 \text{ which is precisely classified into } Y = 0 \]

\[ n_{01} : \text{Total observation from } Y = 0 \text{ wrongly classified to } Y = 1 \]

\[ n_{10} : \text{Total observation from } Y = 1 \text{ wrongly classified to } Y = 0 \]

\[ n_{11} : \text{Total observation from } Y = 1 \text{ which is precisely classified into } Y = 1 \]

\[ n_{0} : \text{Total actual observation from } Y = 0 \]

\[ n_{1} : \text{Total actual observation from } Y = 1 \]

\[ n_{\text{total}} : \text{Total observations predicted from } Y = 0 \]

\[ n_{\text{total}} : \text{Total observations predicted from } Y = 1 \]

\[ n : \text{Total observation} , \]

Based on table 1 can be calculated the accuracy of classification

\[ 1 - APER = \frac{n_{00} + n_{11}}{n} \times 100\% \]  \hspace{1cm} (5)

However, considering the acceptance data of new students is imbalanced data, it should be noted that accuracy is more precise and detailed as indicated by sensitivity and specificity that describes the accuracy of each class. Data prediction is imbalanced also measured by the balance of sensitivity and specificity summarized in the G-means value.

\[ Sensitivity = \frac{n_{11}}{n_{11} + n_{01}} \]  \hspace{1cm} (6)

\[ Specificity = \frac{n_{00}}{n_{00} + n_{10}} \]  \hspace{1cm} (7)

\[ G - \text{means} = \sqrt{sensitivity \times specificity}. \]  \hspace{1cm} (8)

3. Result and Discussion

Data used in this study is data of national written test on natural sciences participants in a public university in 2018 Total observations used were 27,140 observations from a public university in Indonesia. The response variable is categorical data, accepted and not accepted in the public university. The predictor variable used is the score of each subject test in the national written test. List of variables used is presented in table 2.

Table 2. Research Variables

| Variable                  | Indicator                                                                 |
|---------------------------|---------------------------------------------------------------------------|
| Y = Status                | Status of participants, accepted or not accepted                           |
| X₁ = Figural              | Figural score on the national written test                                 |
| X₂ = Numerical            | Numerical score on the national written test                               |
| X₃ = Verbal               | Verbal score on the national written test                                  |
| X₄ = Biology              | Biology score on the national written test                                 |
| X₅ = Chemistry            | Chemistry score on the national written test                              |
| X₆ = Physics              | Physics score on the national written test                                 |
| X₇ = Mathematical Science | Mathematical Science score on the national written test                   |
| X₈ = English              | English score on the national written test                                 |
| X₉ = Indonesian           | Indonesian score on the national written test                             |
| X₁₀ = Basic Mathematic    | Basic Mathematic score on the national written test                       |
3.1 Multicollinearity Checking

Multicollinearity was checked using Pearson correlation to determine the correlation between independent variables. Table 3 showed some significant variables have a correlation with a confidence level of 95 percent. Although it was significantly correlated, the correlation value was not greater than 0.95. The table 3 showed although predictor variables correlate with each other, there is no multicollinearity between predictor variables. Correlation values less than 0.95 indicate there is no multicollinearity [14].

Table 3. Correlation between Predictor Variables.

|                | Figural | Numerical | Verbal | Biology | Chemistry | Physics | Math | Science | English | Indonesian |
|----------------|---------|-----------|--------|---------|-----------|---------|------|---------|---------|------------|
| Numerical      | 0.47**  | 0.45**    | -0.01* | -0.03   | 0.05**    | 0.14**  | 0.10**| 0.20**  | -0.02** | -0.01*     |
| Verbal         | 0.36**  | 0.45**    | -0.01* |         |           |         |       |         |         |            |
| Biology        | 0.05**  | 0.14**    | 0.10** | 0.21**  | 0.19**    |         |       |         |         |            |
| Chemistry      | -0.04** | -0.02**   | -0.07**| 0.21**  | 0.09**    | 0.19**  |       |         |         |            |
| Physics        | 0.04**  | 0.00      | -0.03**| 0.13**  | 0.10**    | 0.02**  |       |         |         |            |
| Math Science   | 0.15**  | 0.20**    | 0.08** | 0.12**  | 0.22**    |         |       |         |         |            |
| English        | 0.11**  | 0.11**    | 0.06** | 0.16**  | 0.16**    |         |       |         |         |            |
| Indonesian     | 0.13**  | 0.12**    | 0.21** | 0.03**  | 0.05**    | -0.03** | 0.10**| 0.14**  |         |            |
| Basic Math     | 0.13**  | 0.20**    | 0.08** | 0.12**  | 0.22**    | 0.25**  | 0.12**| 0.19**  | 0.1**   |            |

Significant with *\(\alpha = 0.05\), **\(\alpha = 0.1\)

3.2 Logistic Regression

In this study the data is divided into two parts, training and testing data. K-fold method is used to determine training and testing data. The k used in this study are \(k = 10, 5, \) and \(4\). Based on the combination of the k values, Figure 2 results in the accuracy of the classification in the testing data.

![Figure 2. Accuracy of Classification with Logistic Regression](image)

Based on figure 2 shows that using \(k = 5\) gives the best accuracy of classification on testing data, which is 97.55 percent. Furthermore, further modeling will be carried out using binary logistic regression with \(k = 5\).

Table 4 shows that using \(\alpha = 0.05\) obtained the value of \(z_{\alpha/2} = 1.96\), with the result \(|z\text{ value}| > z_{\alpha/2}\) it can be concluded that all independent variables are significant. Besides that, it can also be seen from the results of \(p\text{ value} < \alpha\) on all variables, it resulting in the same decision. So it can be concluded
that all national written test subject scores have a significant effect on whether new students are accepted in public university or not. Based on table 4, a binary logistic regression model is,

\[
\ln \left( \frac{\pi(x)}{1-\pi(x)} \right) = -75.56 + 0.02X_1 + 0.022X_2 + 0.019X_3 + 0.015X_4 + 0.012X_5 + 0.009X_6 + 0.007X_7 + 0.005X_8 + 0.011X_9 + 0.004X_{10}
\]  

(9)

The odds ratio in Table 4 is calculated for an increase of 67 points each predictor, or equivalent to one correct answer. Table 4 shows that every increase 67 points at Figural \(X_1\) national written test score, or additions one correct number, then the probability of a participant will be accepted increase by 3.693 times. Every Numerical \(X_2\) test score increases 67 point, then the probability of a participant being accepted increase by 4.381 times, and so for other variables.

**Table 4. Binary Logistic Regression Result.**

|       | Estimate | z value | p value | Odds Ratio |
|-------|----------|---------|---------|------------|
| Intercept | -75.560 | -35.040 | 0.000 |            |
| \(X_1\)  | 0.020    | 18.830  | 0.000  | 3.693      |
| \(X_2\)  | 0.022    | 21.440  | 0.000  | 4.381      |
| \(X_3\)  | 0.019    | 20.430  | 0.000  | 3.591      |
| \(X_4\)  | 0.015    | 24.520  | 0.000  | 2.739      |
| \(X_5\)  | 0.012    | 24.980  | 0.000  | 2.184      |
| \(X_6\)  | 0.009    | 17.770  | 0.000  | 1.826      |
| \(X_7\)  | 0.007    | 13.660  | 0.000  | 1.577      |
| \(X_8\)  | 0.005    | 11.340  | 0.000  | 1.356      |
| \(X_9\)  | 0.011    | 17.400  | 0.000  | 2.107      |
| \(X_{10}\)| 0.004    | 7.980   | 0.000  | 1.282      |

After obtaining the results of parameter estimation, an examination of the classification accuracy result that produced from the model. Classification accuracy is seen for training and testing data.

**Table 5. Accuracy of Classification of Training Data with Binary Logistic Regression Observation.**

| Observation | Prediction | Sensitivity | Specificity | Accuracy |
|------------|------------|-------------|-------------|----------|
| 0          | 20.452     | 398         | 62.73%      | 99.07%   | 97.28%   |
| 1          | 192        | 670         |             |          |          |

Based on the results in table 5, the classification accuracy was 97.28 percent, while the remaining 2.72 percent of data was misclassified. The sensitivity for training data was 62.73 percent, which means that from the category \(Y = 1\) was 62.73 percent between the actual and prediction is correctly classified. The specificity value was 99.07 percent, which means that 99.07 percent of the \(Y = 0\) category between the actual and prediction was correctly classified.
Table 6. Accuracy of Classification of Testing Data with Binary Logistic Regression

| Observation | Prediction | Sensitivity | Specificity | Accuracy  |
|-------------|------------|-------------|-------------|-----------|
| 0           | 5.088      | 103         | 63.99%      | 97.11%    |
| T1          | 54         | 183         | 98.95%      |           |

Based on Table 6 shows that from 5,428 observations in testing data, there were 157 incorrect observations classified, so that the accuracy was 97.11 percent. The sensitivity for data testing was 63.99 percent, which means there was 36.01 percent of the category Y = 1 variable between the actual and prediction has misclassification. The specificity value was 98.95 percent, which means it was 1.05 percent of the Y = 0 between the actual and prediction has misclassification.

3.3. Random Forest

In this study, as many as $m$ (2, 4 and 7) predictor variables will be used randomly from a total of 10 predictor variables. Next determine the number of trees as many as $t$ (25, 50, 100, and 250). As in logistic regression analysis, the data used in this study is the same, which is divided into training and testing data, with $k$-folds. The $k$ values are 10, 5, and 4. Figure 3 results in the classification accuracy in testing data.

Figure 3 showed changes in the level of accuracy due to changing values $t$ in each taken $m$ variable. Figure 3 shows the different results for each combination of $t$ and $m$. There was no pattern of definite relationships between the values $t$, $m$, and $k$. Based on figure 3 shows that the combination produces the highest level of accuracy is using $k = 5$, $m = 4$, and $t = 250$. 

![Graph](a) $k = 4$  
![Graph](b) $k = 5$
Table 7 shows the classification results using random forest. By using training data as many as 21,712 data, there are 867 were misclassified. This means that the model can classify exactly 96.01 percent. The sensitivity was 40.61 percent, which means that from the category $Y = 1$ by 66.52 percent between the actual and prediction was correctly classified. The specificity was 96.94 percent, which means that 96.94 percent of the category $Y = 0$ between the actual and prediction was correctly classified.

Based on Table 8 shows that the model can classify observations correctly by 96.59 percent. The sensitivity for testing data was 47.76 percent, which means that there was a lot of data from the $Y = 1$ category between the actual and prediction that was misclassified. The specificity was 99.13 percent, which means that 99.13 percent of the data from $Y = 0$ between the actual and prediction were correctly classified. It can be said that the model is good at classifying observations.

### 3.4. Comparison of Logistic Regression and Random Forest

The performance of the logistic regression model and random forest model is measured by prediction accuracy. Prediction accuracy is generally measured by classification accuracy (1-APER).

| Table 9. Comparison of Classification Methods |
|-----------------------------------------------|
| **Metode**       | **Accuracy** | **Sensitivity** | **Specificity** | **G-mean** |
| Logistic Regression | 97.11%      | 63.99%          | 98.95%          | 0.804      |
| Random Forest     | 96.59%      | 47.76%          | 99.13%          | 0.688      |

Figure 3. Accuracy of Random Forest

(c) $k = 10$
Based on Table 9, it can be seen that by looking at the value of accuracy, specificity, and G-mean, logistic regression was better than random forest. Based on table 9 it appears that logistic regression had a better G-mean value than the random forest, which is 0.804. This number shows that the logistic regression method more balanced in making predictions in each class.

3.5. Determination of Minimum Score
Univariate logistic regression method is used in this section to get the minimum score. The results from univariate logistic regression shows in table 10. Based table 10, shows that all predictor variables were significant.

Based on Table 10, it can be used to calculate the probability of accepted for each subject. The minimum score that must be achieved can be seen from Figure 4. Figure 4 explains that if the probability value is greater than 0.5, than the participant is accepted, and conversely. From Figure 4, a minimum score table can be made as shown in table 11. From table 11 showed that the minimum scored for Physics, Mathematical Science, English, and Indonesian subjects is very high. On the Mathematical Science and Indonesian subjects had empirical data didn't reach acceptable probability 0.5, shown in figure 5. Then the univariate model didn't describe it well. On the Physics and English subjects in figure 5 reached 0.5, but up to 900 points accepted probability was only around 0.2 to 0.3. When the score reached 1000, there was a significant increase. That's why the subject line of physics and English in Figure 4 tend to be sloping.

![Graph of the relationship between scores for each subject test and the accepted probability based on models](image)

**Figure 4.** Graph of the relationship between scores for each subject test and the accepted probability based on models

| Variable        | Estimate | z value | p value | Odds  |
|-----------------|----------|---------|---------|-------|
| Intercept       | -12.40   | -33.950 | 0.000   |       |
| $X_1$           | 0.016    | 27.230  | 0.000   | 2.861 |
| Intercept       | -14.180  | -38.430 | 0.000   |       |
| $X_2$           | 0.018    | 32.340  | 0.000   | 3.276 |
| Intercept       | -10.410  | -35.730 | 0.000   |       |
| $X_3$           | 0.013    | 27.170  | 0.000   | 2.364 |
| Intercept | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|-----------|------:|------:|------:|------:|------:|------:|--------:|
|          -8.174 | -42.280 | 0.010 | 29.430 | 0.000 | 1.916 | 0.011 | 39.710 |
|          -9.028 | -52.550 | 0.011 | 39.710 | 0.000 | 2.074 | 0.007 | 23.360 |
|          -6.565 | -39.300 | 0.007 | 23.360 | 0.000 | 1.553 | 0.005 | 19.810 |
|          -5.370 | -40.400 | 0.005 | 19.810 | 0.000 | 1.352 | 0.006 | 22.130 |
|          -6.151 | -39.330 | 0.006 | 22.130 | 0.000 | 1.450 | 0.008 | 19.540 |
|          -7.431 | -30.840 | 0.008 | 19.540 | 0.000 | 1.690 | 0.008 | 28.640 |
|          -7.768 | -42.740 | 0.008 | 28.640 | 0.000 | 1.739 |

**Figure 5.** Graph of the relationship between scores for each subject test and the accepted probability based on empirical data.

Figure 5 showed that from the empirical data many subjects have an acceptable probability of less than 0.5 and the score didn’t reach 800. This showed that if only using one subject cannot be used to determine the probability of accepted a participant, then further research it is recommended to use multivariate models and use models in each study program.
Table 11. Minimum Score to be Accepted

| Subject / Variable          | Min Score | Number of correct answer |
|-----------------------------|-----------|--------------------------|
| Figural \((X_1)\)           | 790       | 9                        |
| Numerical \((X_2)\)         | 800       | 9                        |
| Verbal \((X_3)\)            | 813       | 10                       |
| Biology \((X_4)\)           | 843       | 10                       |
| Chemistry \((X_5)\)         | 830       | 10                       |
| Physics \((X_6)\)           | 998       | 12                       |
| Mathematical Science \((X_7)\) | 1195     | 15                       |
| English \((X_8)\)           | 1111      | 14                       |
| Indonesian \((X_9)\)        | 950       | 12                       |
| Basic Mathematics \((X_{10})\) | 943    | 12                       |

4. Conclusion

Logistic regression method gives better result in classifying the data of participant in the national written test. The results showed that the best logistic regression model had an accuracy of 97.11 percent, whereas the random forest method had an accuracy of 96.59 percent. Univariate logistic regression method has been used to obtain the minimum score of a participant that can be received for each subject. The minimum score for Figural, Numerical, Verbal, Biology, Chemistry, Physics, Mathematical Science, English, Indonesian, and Basic Math were 790, 800, 813, 843, 830, 998, 1195, 1111, 950 and 943, respectively.

References

[1] Kemenristekdikti, 16 09 2018. [Online]. Available: https://sbmptn.ac.id/?mid=13.
[2] R. A. Johnson and D. W. Wichern, 2007. Applied Multivariate Data Analysis, United State, Pearson Prentice.
[3] A. Agresti, 2007. Categorical Data Analysis Second Edition, New York, John Willey & Sons, p. 173.
[4] Y. Anggraeni, 2015. Binary Logistic Regression Modeling of ITS Enthusiasts in Selection Higher Education State Joint Entrance (SBMPTN) 2014, Final Project of ITS Statistics Department, Surabaya.
[5] Y. Y. Liu, M. Yang, M. Ramsay, X. S. Li and J. W. Coid, 2011. A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending.
[6] L. Breiman, Random Forest, Machine Learning, 45. pp. 5-32, 2001.
[7] J. C. W. Chan and D. Paelinckx, 2008. Evaluation of Random Forest and Adaboost Tree-Based Ensemble Classification and Spectral Band Selection for Ecotope Maping using Airborne Hyperspectral Imagery, Remote sensing of Environment, 112 pp. 2999-3011.
[8] P. O. Gislason, J. A. Benediktsson and J. R. Sveinsson, 2006. Random Forest for Land Cover Classification, Pattern Recognition Letters, 27, pp. 294-300.
[9] R. Genuer, J. M. Poggi, C. T. Malot and N. V. Vialaneix, 2017. Random Forest for Big Data, Big Data Research, 9 pp. 28-46.
[10] M. F. Delgado, E. Cernadas, S. Barro and D. Amorim, 2014. Do We Need Hundreds of
Classifiers to Solve Real Word Classification Problems?, *Journal of Machine Learning Research*, 15 pp. 3133-3181.

[11] D. W. Hosmer and S. Lemeshow, 2000. Applied Logistic Regression Second Edition, New York, John Willey & Sons Inc.

[12] T. Hastie, R. Tibshirani and J. Friedman, 2008. The Elements of Statistical Learning," in *Data-Mining, Inference and Prediction*, New York, Springer-Verlag.

[13] L. Breiman and A. Cutler, Manual, Setting Up, Using, and Understanding, 9 10 2018. [Online].Available: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

[14] R. Hocking, 1996. Method and Applications of Linear Models, New York: John Wiley and Sonc, Inc.

[15] R. Genuer, J.-M. Poggi, C. Tuleau-Malot and N. Villa-Vialaneix, 2017. Random Forest for Big Data, *Big Data Research*, 9 pp. 28-46.