UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 (4), 2023 Page 353 - 360.
ISSN(*Online*): 2985-475X

Emprical Study for Algorithms Comparison of Classification and Regression Tree and Logistic RegressionUsing Combined 5×2cvF Test

Fayza Annisa Febrianti, Dodi Vionanda*, Yenni Kurniawati, Fadhilah Fitri
Department of Statistics, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : July 17th, 2023
Revised : August 16th, 2023
Accepted : August 21st, 2023

ABSTRACT

A statistical method to estimate the class of an object based on its characteristics is classification. Several learning algorithms can be applied in classification, such as Classification and Regression Tree (CART) and logistic regression. The main goal of classification is to find the best learning algorithm that can be applied to get the best classifier. In comparing two learning algorithms, a direct comparison by seeing the smaller prediction error rate may be possible when the difference is very clear. In this case, direct comparison is misleading and resulting in inadequate conclusions. Therefore, a statistical test is needed to determine whether the difference is real or random. The results of the 5×2cv paired t-test sometimes reject and sometimes fail to reject the hypothesis. It is distracting because the change of the $p_i^{(j)}$ *should not affect the test result. Meanwhile, the overall results of the combined 5×2cv F test show that the tests fail to reject the hypothesis. This indicates that CART and logistic regression perform identically in this case.*

Keywords: *5×2cv Paired t-test, CART, Combined 5×2cv F Test, K-fold Cross Validation, Logistic Regression.*

I. INTRODUCTION

Classification is one of the data analysis methods used to estimate the class of an object based on its characteristics. It classifies objects based on their characteristics into predetermined classes. To determine or estimate the class of the object, a model or classifier is needed. A classifier is a function that maps or determines the class of a given input (Dietterich, 1998). Classifiers need a learning algorithm that is constructed from a set of labeled data (Kohavi, 1995). Several learning algorithms can be applied in classification, such as Classification and Regression Tree (CART) and logistic regression. CART is one of the decision tree methods. The main idea of CART is to split all observations into two parts and repeat them until the minimum number of observations is reached or homogeneous in each part of the observations (Timofeev, 2004). Logistic regression is a part of regression analysis that the dependent variable is categorical (Rogel-Salazar, 2017).

In some cases, the main goal of classification is to find the best learning algorithm that can be applied to get the best classifier. The selection of the best algorithm can be seen through a smaller error rate. Error measures the performance of the model by calculating all of the prediction errors in the model. In classification, errors can be identified by checking whether the prediction result is the same as the actual value (Wood, 2017). In comparing two learning algorithms, a direct comparison by seeing the greater accuracy or the smaller prediction error rate may be possible when the difference is very clear. However, in most cases, this direct comparison may be misleading, resulting in inadequate conclusions (Stapor, 2017). Therefore, a statistical test is needed to determine whether the difference is real or random.

A reliable statistical test is a test that can control all sources of variation. This source of variation is due to the selection of training data and test data. The difference in the selection of training data and testing data may lead to different results. The sources of variation from the training data can be addressed by running the learning algorithm multiple times and measuring the variation in the accuracy or the error rate of the resulting model. Testing data variation can be addressed by considering the size of the testing data (Dietterich, 1998). The difference in the selection of training data and testing data may lead to different results.

One of the statistical tests that is used to compare two classification learning algorithms is the combined 5×2cv F test. The hypothesis of this test is whether the two algorithms have the same error rate or not. In this test, 5 replications of 2-fold cross validation are performed. This test was developed to overcome the shortcomings of the 5×2cv paired t-test by combining all of the error rates in each fold and replication. The 5×2cv paired t-test

ISSN(*Online*): 2985-475X

may produce different results due to the change in the order of replication or fold that is used, while it should not affect the results. In other words, it is giving the same test results.

Based on the previous description, research will be carried out to compare the CART algorithm and logistic regression using the Combined 5×2cv F test. This research set out to determine whether the CART algorithm and logistic regression have the same error rate or not.

II. RESEARCH METHODS

Direct comparison and statistical tests will be used in comparing those algorithms. It will see whether CART and logistic regression have the same error rate or not (H_0). The direct comparison will be used in this research is 10-fold cross validation. In direct comparison, the absolute error rate difference between CART and logistic regression is observed. The statistical tests that will be used in this research are the 5×2cv paired t-test and the combined 5×2cv F test. Both of the tests will perform 10 differences in the error rate between the two algorithms ($p_i^{(j)}$). The 5×2cv paired t-test uses one of that differences to measure the t-statistic. Changing $p_i^{(j)}$ should not affect the test result. Hence, it will be observed whether changing $p_i^{(j)}$ will affect the test result or not by using all of 10 differences to measure the t-statistic and then totaling the rejects out of 10. The combined 5×2cv F test combines all of the differences ($p_i^{(j)}$). It makes this test more robust. In the combined 5×2cv F test, it will be observed the result of the test.

A. Data Source

This research uses simulated data that is generated using the data generation process in R Studio software. The dataset consists of binary categorical dependent variables (Y) and numerical independent variables (X) with 100 observations. Variable X is generated by normal distribution. Dataset generation is generated under several conditions, as follows.

1. Univariate Dataset: $N(\mu, \sigma)$
   a. Univariate 1 ($\mu_1 = 0,\ \mu_2 = 0.5,\ \sigma = 1$)
   b. Univariate 2 ($\mu_1 = 0,\ \mu_2 = 1,\ \sigma = 1$)
   c. Univariate 1 ($\mu_1 = 0,\ \mu_2 = 2,\ \sigma = 1$)
2. Bivariate Dataset: $N(\boldsymbol{\mu}, \Sigma)$
   a. Bivariate 1 $\left(\boldsymbol{\mu_1} = \begin{bmatrix}0\\0\end{bmatrix},\ \boldsymbol{\mu_2} = \begin{bmatrix}0.5\\0.5\end{bmatrix},\ \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$
   b. Bivariate 2 $\left(\boldsymbol{\mu_1} = \begin{bmatrix}0\\0\end{bmatrix},\ \boldsymbol{\mu_2} = \begin{bmatrix}1\\1\end{bmatrix},\ \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$
   c. Bivariate 3 $\left(\boldsymbol{\mu_1} = \begin{bmatrix}0\\0\end{bmatrix},\ \boldsymbol{\mu_2} = \begin{bmatrix}2\\2\end{bmatrix},\ \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$

B. Classification and Regression Tree (CART)

Classification and Regression Tree (CART) is a decision tree method. A decision tree is a flow chart that looks like a tree structure that will be used to classify a set of objects. CART is a statistical method that uses a binary recursive partitioning algorithm (Lewis, 2000). Binary means that in each split the parent node will only produce two child nodes. Recursive means that the process of node splitting is repeated over and over again. Partitioning refers to the rule that the data is divided into separate parts so that each part consists of the same class as possible.

There are three steps in constructing a classification tree using the CART algorithm, which is the selection of the splits, the terminal node determination, and the labeling. The selection of the split in each node is to find a split that can produce a node with the highest level of homogeneity of the dependent variable. The level of homogeneity of the node can be measured by its impurity. The impurity will be higher if all classes are fairly mixed and lower if it contains only one class. The impurity function used in this method is the Gini index.

$$i(t) = 1 - \sum_{j=1}^{n} P^2(j|t) \tag{1}$$

$$P(j|t) = \frac{n_j(t)}{n(t)} \tag{2}$$

Where $i(t)$is the Gini index, $P(j|t)$ is the proportion of class j at node t, $n_j(t)$ is the number of observations of class j at node t, and $n(t)$ is the number of observations at node t.

The attributes from the splitting selection will build a class group called a node. Split s partitions observation within node t, so both the left and the right node are homogeneous as possible. These nodes will recursively split until they become terminal nodes. To evaluate the split s at node t, goodness of split is required.

$$\phi(s,t) = \Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \tag{3}$$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/85

UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 (4), 2023 Page 353 - 360.
ISSN(*Online*): 2985-475X

Where t_R is the right node, t_L is the left node, P_R is the proportion of object in t_R, and P_L is the proportion of object in t_L.

The split that gives the highest goodness of split is the best one. The tree construction process is repeated until a terminal node is formed. A node can be defined as a terminal node if the node stop splitting and the tree construction is stopped. The splitting will stop if there are less than or equal to 5 ($n \leq 5$) observations in the daughter node (Breiman, 1993).The next step is labeling. Labeling the class is the process of identifying each node in a particular class. This labeling is needed at each terminal node so that the process of predicting an object of a certain class will be on the identified terminal. The class label on the terminal node t is determined through the majority vote.

C. Logistic Regression

Logistic regression is applied to analyze a categorical dependent variable based on one or more independent variables. It is used to determine the relationship between a categorical dependent variable and a categorical or continuous independent variable. If the dependent variable is nominal, nominal logistic regression is used, while for ordinal dependent variables, ordinal logistic regression is used. The nominal logistic regression model is used when there is no order between the dependent categories. Based on the number of categories in the dependent variable, nominal logistic regression is divided into two, which are binary logistic regression and multinomial logistic regression.

Binary logistic regression is used to determine the relationship between a dichotomous (two categories) dependent variable and one or more independent variables that are categorical or continuous. A Binary logistic regression model with p variables as.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \quad (4)$$

Where β_p is the variable coefficient X_p and $\pi(x)$ is the probability of a success event within $0 < \pi(x) < 1$.

Logistic regression does not assume a linear relationship between the independent and dependent variables. π(x) is a non-linear function, so it needs to be transformed into a logit form to get the linear function. The binary logistic regression model in logit form as.

$$Logit\left(\pi(x)\right) = \left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (5)$$

D. Error Rate Estimation

The application of classifier methods for prediction is related to their performance. The performance is based on the ability of the model to predict. The error can measure model performance by calculating all forms of prediction error rates in the model. Error is the difference between the predicted value and the actual value. In classification, errors can be determined by checking whether the predicted results are the same as the actual ones (Wood, 2017). If the prediction result is the same as the actual, thus there is no error and vice versa.

Let (x_0^T, y_0) be a new test point that is drawn from the same population and distribution as the data that used in model construction. The predicted result of x_0 is define by $\hat{y}_0$, so the difference between the predicted and actual can be defined as.

$$Q(y_0, \hat{y}_0) = \begin{cases} 0, & y_0 = \hat{y}_0 \\ 1, & y_0 \neq \hat{y}_0 \end{cases} \quad (6)$$

From that function, the true error rate in the classification using a test point (x_0^T, y_0) is defined as.

$$Err = E_{0F} Q(y_0, \hat{y}_0) \quad (7)$$

Err can be viewed as the probability of a test point (x_0^T, y_0) being misclassified. To estimate the true error rate, it can be defined as.

$$\overline{err} = E_{0\hat{F}} Q(y_0, \hat{y}_0) = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \hat{y}_i) \quad (8)$$

E. K-fold Cross Validation

K-fold cross validation randomly divides data into k groups that are then split into training and testing data. This is performed k times by leaving one group as testing data in each fold. The proportion for training data in k-fold cross validation is $k-1/k$ and $1/k$ for testing data (Raschka, 2018). The error rate estimation of each fold can be estimated using the following formula.

$$\widehat{Err}_j^{CV} = \frac{1}{n_j^*}\sum_{i=1}^{n_j^*} Q\left(y_{ij}^*, \hat{y}_{ij}^*\right) \quad (9)$$

ISSN(*Online*): 2985-475X

where n_j^* is the number of test data in the j iteration, y_{ij}^* is the i actual value of y in the j-th iteration in the test data, and $\hat{y}_{ij}^*$ is the i predicted value of y in the j-th iteration in the test data. So that the error rate estimaion with the k-fold cross validation method can be estimated as follows.

$$\widehat{Err}^{CV} = \frac{1}{k}\sum_{i=1}^{k}(\widehat{Err}_j^{CV}) \tag{10}$$

F. The 5×2cv Paired t-Test

The 5×2cv paired t-test is a test for comparing two algorithms that was proposed by Dietterich in 1998. This test is used to determine whether the two algorithms have the same error rate or not (Dietterich, 1998).In this test, 5 replications of 2-fold cross validation are performed. The data is split randomly into two equal parts S_1 and S_2 (50% training data and 50% testing data) five times (Raschka, 2018). In each replication, the split data is used in two algorithms (A and B) to estimate the errors. Then, rotate the training and testing data (training data becomes testing data and vice versa) to be used in the two algorithms again. This will give four error estimates, which are $p_A^{(1)}$ and $p_B^{(1)}$ (training data in $S_1$and testing data in S_2), $p_A^{(2)}$ and $p_B^{(2)}$ (training data in S_2 and testing data S_1). The difference of the error rate between two algorithms at replication i=1, 2, …, 5 and fold j=1, 2 is denoted by $p_i^{(j)}$. Thus, each replication results in two differences.

$$p^{(1)} = p_A^{(1)} - p_B^{(1)}$$
$$p^{(2)} = p_A^{(2)} - p_B^{(2)} \tag{11}$$

The result of the difference between these errors is used to estimate the mean and variance of each replication.

$$\bar{p}_i = \frac{p_i^{(1)} + p_i^{(2)}}{2} \tag{12}$$

$$s_i^2 = \left(p_i^{(1)} - \bar{p}_i\right)^2 + \left(p_i^{(2)} - \bar{p}_i\right)^2 \tag{13}$$

The variance of the difference is measured for 5 iterations and then used to measure the t statistic as follows.

$$t_i^{(j)} = \frac{p_i^{(j)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \tag{14}$$

The hypothesis that two algorithms have the same error rate (H_0) is rejected at 95% confidence level if the statistics t is greater than 2.571.

G. The Combined 5×2cv F Test

The combined 5×2cv F test was proposed by Alpaydin in 1999 to address the shortcomings of the 5×2cv paired t-test. The 5×2cv paired t-test has the drawback that changing the order of replication or fold ($p_i^{(j)}$) may give different results. The combined 5×2cv F test is a more robust test than the 5×2cv paired t-test because the test is performed with the combination of 10 results from $p_i^{(j)}$(Alpaydin, 1999). The combined 5×2cv F test is similar to the 5×2cv paired t-test. This test also performs 5 replications of 2-fold cross validation. The difference between these tests is in the statistical test that is used. The combined 5×2cv F test use F test as follows.

$$f = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}(p_i^{(j)})^2}{2\sum_{i=1}^{5} s_i^2} \tag{15}$$

The hypothesis that two algorithms have the same error rate (H_0) is rejected at 95% confidence level if the statistics f is greater than 4.74.

III. RESULTS AND DISCUSSION

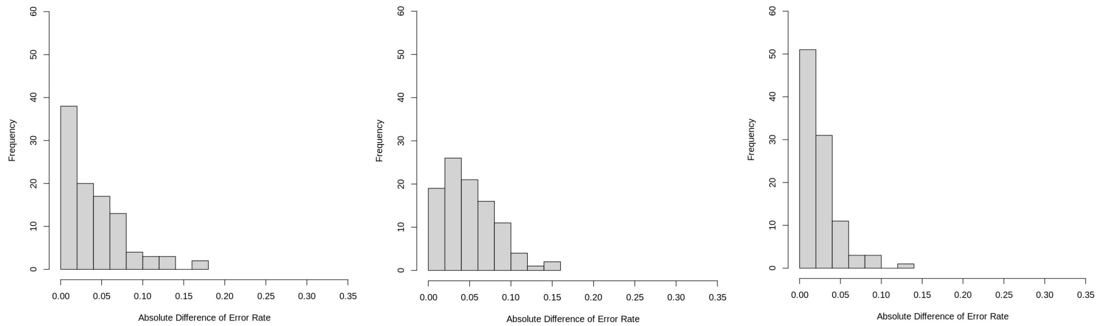
The datasets are generated in univariate and bivariate, whereas each of them is generated in 3 different conditions or populations. For each population, the dataset is generated 100 times. The generated datasets both in univariate and bivariate are used in comparing CART and logistic regression. Here are the result of comparing CART and logistic regression using k-fold cross validation, the 5×2cv paired t-test, and the combined 5×2cv F test for univariate dataset.

Figure 1 shows the histogram of the absolute error rate difference between CART and logistic regression by using k-fold cross validation. As shown in Figure 1, the histograms of univariate 1, univariate 2, and univariate 3 have one central tendency. This indicates that most of the datasets in univariate 1, univariate 2, and univariate 3 have the same absolute error rate difference. However, there is some dataset with a higher absolute error rate difference than most of the dataset. The histograms above are right-skewed, which implies most of the datasets in univariate 1, univariate 2, and univariate 3 have low absolute difference of error rates. The absolute error rates

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/85

ISSN(*Online*): 2985-475X

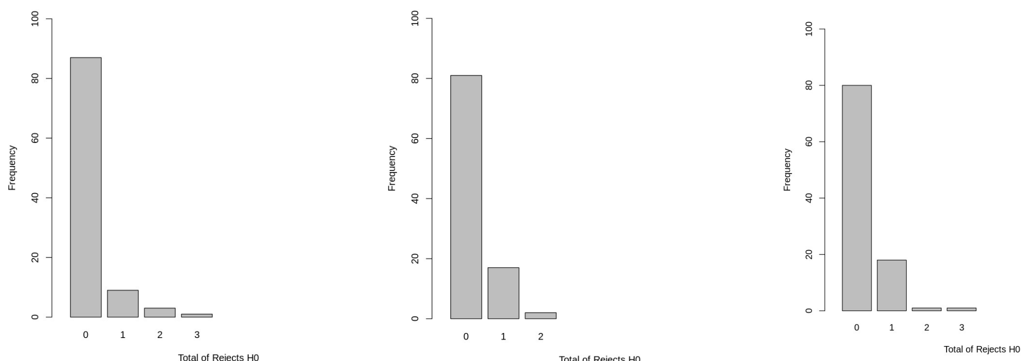
difference between CART and logistic regression in univariate 1, univariate 2, and univariate 3 are quite low. Most all of the data have a low difference. It may be possible to perform a direct comparison when the difference is very clear. But in this case, the difference is not clear at all.The direct comparison may be misleading and result in inadequate conclusions.Thus, a statistical test is required to determine whether the difference is real or random.



(a) (b) (c)

Figure 1. (a) Histogram of error rate difference using k-fold cross validation univariate 1, (b) Histogram of error rate difference using k-fold cross validation univariate 2, and (c) Histogram of error rate difference using k-fold cross validation univariate 3

The results of the 5×2cv paired t-test are shown in Figure 2. Figure 2 displays the totals of rejects H_0 for univariate 1, univariate 2, and univariate 3 using different $p_i^{(j)}$ in the 5×2cv paired t-test giving various results. In multiple data, the totals of rejects H_0 in univariate 1, univariate 2, and univariate 3 are not 0 or 10. The total of rejects H_0 should be 0 if the result fails to reject the hypothesis or 10 if the result rejects the hypothesis. This indicates that changing $p_i^{(j)}$ affects the test result. The test sometimes fails to reject and sometimes rejects the hypothesis by using of different $p_i^{(j)}$. The test should give the same result for all of $p_i^{(j)}$ that used in this test. This is disturbing because the changing of the $p_i^{(j)}$ should not affect the test result. To overcome that problem, the combined 5×2cv F test is used in this case.



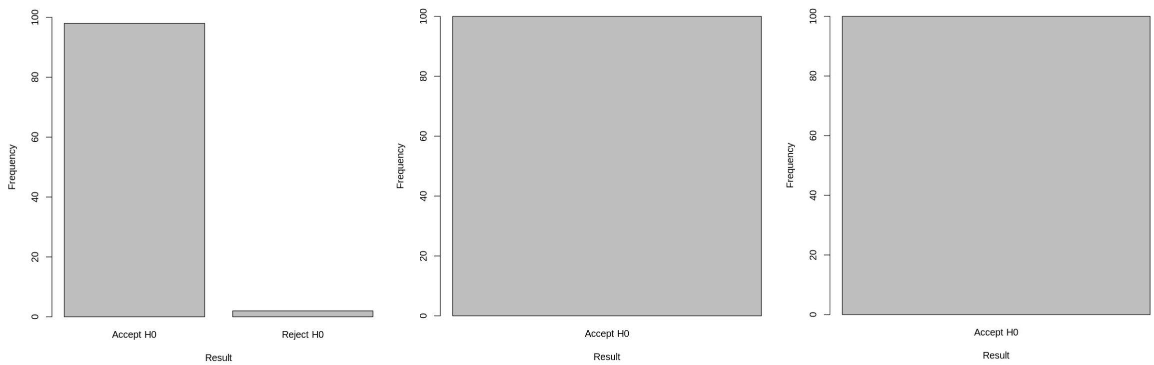
(a) (b) (c)

Figure 2. (a) Barplot of 5×2cv paired t-test univariate 1, (b) Barplot of 5×2cv paired t-test univariate 2, and (c) Barplot of 5×2cv paired t-test univariate 3

The results of the combined 5×2cv F test are shown in Figure 3. As can be seen from Figure 3, the combined 5×2cv F test for univariate 2 and univariate 3 give the same results. All of the datasets in univariate 2 and univariate 3 fails to reject the hypothesis, which means that both of the algorithms has the same error rate. Some of the dataset in univariate 1 rejects the hypothesis, while most of the rest accepts the hypothesis. Overall, these results in univariate 1, univariate 2, and univariate 3 indicates that the algorithms CART and logistic regression have the same error rate for each population of univariate dataset. It implies that CART and logistic regression perform identically in this univariate data.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/85

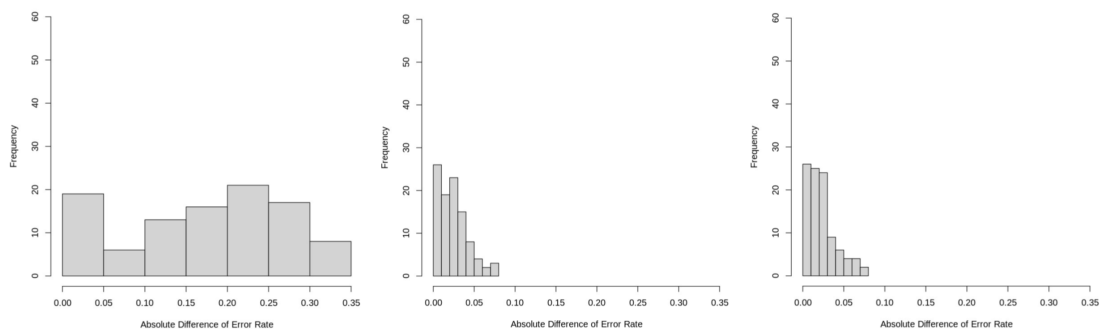
UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 (4), 2023 Page 353 - 360. ISSN(*Online*): 2985-475X



(a) (b) (c)

Figure 3. (a) Barplot of combined 5×2cv F test univariate 1, (b) Barplot of combined 5×2cv F test univariate 2, and (c) Barplot of combined 5×2cv F test univariate 3

A comparison of univariate dataset reveals that CART and logistic regression have the same error rate. Now turn to bivariate dataset. Looking at Figure 4, the histogram of bivariate 1 has two central tendencies and the highest maximum value of the absolute error rate difference. Figure 4 shown the histogram of bivariate 2 and bivariate 3 have one central tendency and right-skewed. This indicates that most of the datasets in bivariate 2 and bivariate 3 have the same absolute error rate difference and have low absolute error rate difference. However, there is some dataset with a higher absolute error rate difference than most of the dataset. The absolute error rates difference between CART and logistic regression are quite low. Most of the data has a low difference. It may be possible to perform a direct comparison when the difference is very clear. But in this case, the difference is not clear at all and the direct comparison may be misleading and result in inadequate conclusions.Thus, a statistical test is required to determine whether the difference is real or random.



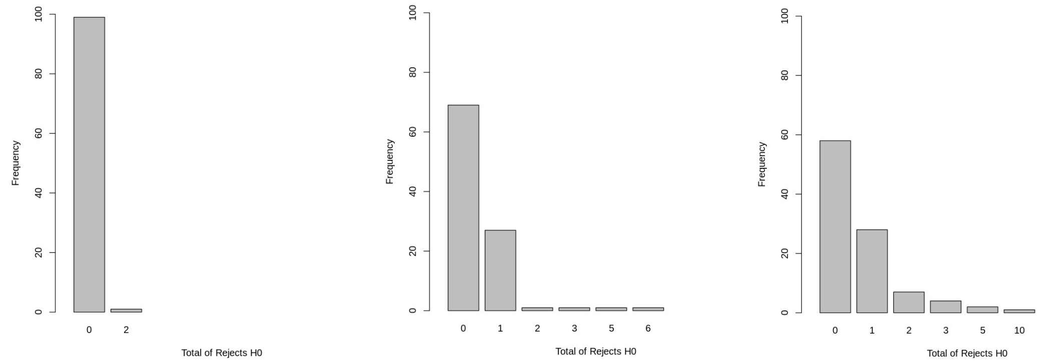
(a) (b) (c)

Figure 4. (a) Histogram of error rate difference using k-fold cross validation bivariate 1, (b) Histogram of error rate difference using k-fold cross validation bivariate 2, and (c) Histogram of error rate difference using k-fold cross validation bivariate 3

Figure 5 shows the results of the 5×2cv paired t-test. As can be seen from Figure 5, the 5×2cv paired t-test giving various results of the totals of rejects H_0 in using different $p_i^{(j)}$ for bivariate 1, bivariate 2, and bivariate 3 in. In multiple data, the totals of rejects H_0 for bivariate 1, bivariate 2, and bivariate 3 are not 0 or 10. The total of rejects H_0 should be 0 if the result fails to reject the hypothesis or 10 if the result rejects the hypothesis. This indicates that the test sometimes rejects and fails to reject the hypothesis by using of different $p_i^{(j)}$. The test should give the same result for all of $p_i^{(j)}$ that used in this test. This is disturbing because the changing of the $p_i^{(j)}$ should not affect the test result. To overcome that problem, the combined 5×2cv F test is used in this case.

 DOI: https://doi.org/10.24036/ujsds/vol1-iss4/85

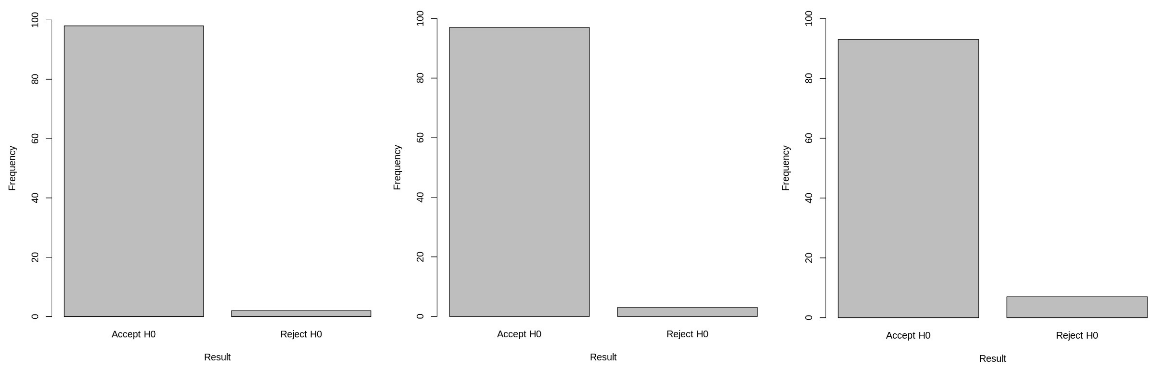
UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 (4), 2023 Page 353 - 360.
ISSN(*Online*): 2985-475X



(a) (b) (c)

Figure 5. (a) Barplot of 5×2cv paired t-test bivariate 1, (b) Barplot of 5×2cv paired t-test bivariate 2, and (c) Barplot of 5×2cv paired t-test bivariate 3

The results of the combined 5×2cv F test are shown in Figure 6. As shown in Figure 6, the results of the combined 5×2cv F test for bivariate 1, bivariate 2, and bivariate 3 same. A small subset of the dataset in bivariate 1, bivariate 2, and bivariate 3 rejects the hypothesis, whereas most of it fails to reject the hypothesis. The rejected hypothesis is a small subset, thus it can be summarized that CART and logistic regression have the same error rate for each population of bivariate dataset. In other words, that CART and logistic regression perform identically on this bivariate data.



(a) (b) (c)

Figure 6. (a) Barplot of combined 5×2cv F test bivariate 1, (b) Barplot of combined 5×2cv F test bivariate 2, and (c) Barplot of combined 5×2cv F test bivariate 3

The result of comparing CART and logistic regression univariate and bivariate dataset give the same results. Based on the test above, the error rate of CART and logistic regression are same. This indicates that both CART and logistic regression perform identically in this case.

IV. CONCLUSION

The direct comparison is not applicable in this case because the error rates difference between CART and logistic regression are quite low. The use of direct comparison in comparing CART and logistic regression will lead to misleading and give inadequate conclusions. Futhermore, the results of the 5×2cv paired t-test are different for using different $p_i^{(j)}$ in multiple dataset. As a whole results of the combined 5×2cv F test show that tests accepts the hypothesis, thus between CART and logistic regression have the same error rate. CART and logistic regression perform identically in this case, both on univariate dataset and bivariate dataset. However, more research on this topic needs to be undertaken because these results may not be applicable to all types of dataset. The use of different condition and type of data may perform different result. The combined 5×2cv F test is more robust in comparing the two algorithms because its combines the results of ten possible statistics and gives a robust decision.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/85

UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 (4), 2023 Page 353 - 360.
ISSN(*Online*): 2985-475X

REFERENCES

Alpaydin, Ethem. (1999). Combined 5×2cv F Test for Comparing Supervised Classification Learning Algorithm, *Neural Computation*, 11(8):1885–1892.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1993). *Classification And Regression Trees*. New York: Chapman and Hall

Dietterich, Thomas G. (1998). Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms, *Neural computation*, 10(7):1895–1923.

Kohavi, Ron. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Eds: C.S. Mellish, hal. 1137–1143.

Lewis, R., (2000), "An Introduction to Classification and Regression Tree (CART) Analysis", *Presented at Annual Meeting of the society for Academic Emergency Medicine*, San Fransisco, California, Januari 2000

Raschka, Sebastian. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, *arXiv*:1811.12808.

Rogel-Salazar, J. (2017). *Data Science and Analytics with Python* (Vol. 21, Issue 1). Boca Raton: Taylor & Francis Group.

Stapor, Katarzyna. (2017). Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations. In *International Conference on Computer Recognition System*; Springer: Cham, Switzerland, 2017; pp. 12-21.

Timoveef, Roman. (2004). "Classification and Regression Trees (CART) Theory and Applications", *Tesis*, Humboldt University, Berlin, 20 December 2004.

Wood, I., Vixxcher, P., & Mengersen, K. (2007). Classification Based Upon Gene Expression Data: Bias and Precision of Error Rates. *Bioinformatics*, 23(11), 1363-1370.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/85