UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 (4) 2023, Page 321 - 328.
ISSN(*Online*): 2985-475X

Comparison of Error Rate Prediction Methods in Classification Modeling with the CHAID Method for Imbalanced Data

Seif Adil El-Muslih, Dodi Vionanda*, Nonong Amalita, and Admi Salma
Department of Statistics, Padang State University, Padang, Indonesia
*Corresponding author:dodi_vionanda@fmipa.unp.ac.id

Submitted : July 12nd, 2023
Revised : August 16th, 2023
Accepted : August 21st, 2023

ABSTRACT

CHAID (Chi-Square Automatic Interaction Detection) is one of the classification algorithms in the decision tree method. The classification results are displayed in the form of a tree diagram model. After the model is formed, it is necessary to calculate the accuracy of the model. The aim is to see the performance of the model. The accuracy of this model can be determined by calculating the predicted error rate in the model. There are three methods, such as Leave one out cross-validation (LOOCV), Hold-out, and K-fold cross-validation. These methods have different performances in dividing data into training and testing data, so each method has advantages and disadvantages. Imbalanced data is data that has a different number of class observations. In the CHAID method, imbalanced data affects the prediction results. When the data is increasingly imbalanced the prediction result will approach the number of minority classes. Therefore, a comparison was made between the three error rate prediction methods to determine the appropriate method for the CHAID method in imbalanced data. This research is included in experimental research and uses simulated data from the results of generating data in RStudio. This comparison was made by considering several factors, including the marginal opportunity matrix, different correlations, and several observation ratios. The results of the comparison will be observed using a boxplot by looking at the median error rate and the lowest variance. This research finds that K-fold cross-validation is the most suitable error rate prediction method applied to the CHAID method for imbalanced data.

Keywords:*CHAID, Hold-out, Imbalanced Data, K-fold, LOOCV*

I. INTRODUCTION

The CHAID is an iterative technique that sytematically tests each independent variable used in classification, organizing them based on the statictical significance level of the *chi-square* test against the dependent variable. The classification model is displayed in the form of a tree diagram and needs to be assessed for accuracy. The method that can be used to assess accuracy is the error rate prediction method. In error rate prediction, the smallest error value can be used to measure model accuracy. These are two methods that can be used to calculate the predicted error rate, namely *training error rate* and *test error rate*. In the *training error rate*, a set of data is used to build a tree and then reused to make predictions, this causes *overfitting* which can reduce model performance and lead to unreliable results. To avoid *overfitting*, the *error rate test* method can be used, where some of the data is used to build a tree, and the rest is used to predict (James *et al*, 2013).

The most frequently used error rate prediction method is *cross-validation*. *Cross-validation* produces data that shows the accuracy or suitability of measurement with real data, enabling accuracy-based testing using test data parameters and training data. The most important factor in implementing *cross-validation* is choosing the separation ratio between training data and testing data (Yadav and Shukla, 2016).

There are three methods for determining the ratio of the separation for comparison of error rate prediction methods in cross-validation, namely LOOCV, *Hold-out*, and *K-fold cross-validation*. In LOOCV, testing data is a single observation while the rest according to the number of observations is used as training data, this is done repeatedly for the number of observations made. In *Hold-out cross-validation*, this method divides the data in half with a ratio of 2/3 for training data and 1/3 for testing data. In *K-fold cross-validation*, this method divides the data

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81

into *k* subsets with the same sample size, *k*-1 samples are used for training data and the remaining 1 sample is used for testing data (James *et al*, 2013).

Imbalanced data is data that has a different number of data class distributions. Data imbalance can affect the performance of the classification model and unstable prediction results. Almost all classification algorithms will provide higher accuracy for more dominant classes when dealing with imbalanced data (Ali *et al*, 2015). The CHAID method is used for categorical data. In categorical data, it is very susceptible to data imbalance between the majority class and the minority class (Chawla *et al*, 2002).

The aim of this research was to assess how various error rate prediction methods perform on imbalanced data with different class comparisons. Additionally, the study aimed to determined the most suitable error prediction method for the CHAID technique in imbalanced data scenarios. The benefits of this research are as additional knowledge for comparison of the error rate prediction method and the CHAID method for imbalanced data and are expected to be a source of reference for further studies.

II. RESEARCH METHODS

The type of research used is experimental research. Experimental research aims to make a comparison of the effect of a particular treatment with another different treatment or to test the effect of a causal relationship between a variable and other variables. This study aims to compare the error rate prediction algorithms in classification modeling with the CHAID method for imbalanced data. The error rate predictions being compared are LOOCV, *hold-out*, and *k-fold cross-validation*. In this study, each error rate prediction method produces one hundred error values obtained from the iteration of one hundred observations. It is this predicted value that will be used in forming the graphical boxplot and after that comparison is made with it. The things to see in that comparison are the median value and the value of the variation or diversity of the error values obtained.

A. Data Source

The data is simulated data derived from the results of generating data in the *RStudio software*. The generated simulation data consists of the dependent variable (Y) and the independent variable (X). Both variables are categorical with a ratio or interval data scale. The number of observations generated is 100 samples. The Y variable is ordinal with values 1 and 2, while the X variable is generated based on the number of variables consisting of one variable (univariate) and two variables (bivariate). The data generated uses the *MultiOrd* package (Demirtas, 2006) which divides the data into two groups and is differentiated based on the marginal opportunity matrix. In addition, the data is also differentiated using correlation (*binObj* package on *RStudio*). Where the correlation contained in this package ranges from -0.0642 to 0.275 according to the conditions and capabilities of the computer. In the bivariate case, a different correlation structure is used, while the difference in the univariate case marginal opportunity matrix is generated from the binomial distribution described in Table 1 below.

Table 1. Data Generating Provisions For Univariate Data

Variable Number	Setting	Marginal Opportunity Matrix	
		$m^{(1)}$	$m^{(2)}$
Univariate	1	0.9	0.1
	2	0.8	0.2

Table 1 is the provision for generated data rules for univariate data types with several trials of 1. Meanwhile, for the case of bivariate data generated based on the marginal opportunity matrix with the *OrdPmat* conditions presented in Table 2 below.

Table 2. Data Generating Provisions For Bivariate Data

Variable Number	Setting	Marginal Opportunity Matrix
		$m^{(1)}$
Bivariate	*OrdPmat*1	$\begin{bmatrix} 0.9 & 0.8 \\ 0.1 & 0.2 \end{bmatrix}$
		$m^{(2)}$
	*OrdPmat*2	$\begin{bmatrix} 0.1 & 0.2 \\ 0.9 & 0.8 \end{bmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81

In Table 2 the bivariate data uses the marginal opportunity matrix for the *MultiOrd* package with the *OrdPmat* conditions divided into two, namely *OrdPmat*1 and *OrdPmat*2. Furthermore, the provisions for the correlation structure in bivariate data using the *binObj* package can be seen in Table 3 below.

Table 3. Correlation Setting for Bivariate Data

Setting	Correlation Structure
1	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$
2	$\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}$
3	$\begin{bmatrix}1 & 0.6\\ 0.6 & 1\end{bmatrix}$

Table 3 shows the correlation setting applied to the bivariate case. Setting 1 shows data cases with no correlation, Setting 2 shows data cases with moderate correlation, and Setting 3 shows data cases with high correlation.

This study uses imbalanced data. Imbalanced data is data that has a different number of observation class distributions and affects the resulting error rate. Setting the proportion of imbalanced data using the ratio of observations described in Table 4 below.

Table 4. The Ratio of Observations for Imbalanced Data

Setting	Ratio
1	50:50
2	60:40
3	70:30
4	80:20
5	90:10

B. CHAID Method

CHAID uses *chi-square* (X^2) statistics in two ways, the first is used to determine whether the categories in an independent variable are uniform and can be combined into one, and then to determine which independent variables are most significant to divide or differentiate the categories in the dependent variable (Gallagher, 2000).

C*hi-square* test is a statistical test to compare two unpaired categorical variables. Basically, the *chi-square* test aims to determine the independence between two variables at each level, this is useful for determining whether or not the paired categorical variables are significant.

The hypothesis for the *chi-square* test is as follows:

H_0: Independent variables (there is no relationship between dependent variable and independent variable)

H_1: Two variables are not mutually independent (there is a relationship between dependent variable and independent variable.

Cross classification of data with categorical variables is usually presented in contingency tables. If there are two categorical variables, the data is presented in a two-way contingency table which can be seen in Table 5.

Table 5. Chi-Square Test Data Structure

Factor I	Factor II				Number of Lines
	B_1	B_2	...	B_k	
A_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
.	.	.	...	.	.
.	.	.	...	.	.
A_b	n_{b1}	n_{b2}	...	n_{bk}	$n_{b.}$
Number of Columns	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81

Where,
A_b : Category of the dependent variable to-b
B_k : Category of the independent variable to-k
n_b : Number of lines to-b
$n_{.k}$: Number of columns to-k
n : Lines and columns total
While the test statistic is

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Where, to calculate the expected frequency (E_{ij}) of each cell the formula is used

$$E_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Where,
n_{ij} : Number of observations included in the i-category of the first variable and the j-category of the second variable
E_{ij} : Expected frequency of observations that are included in the i-category of the first variable and the j-category of the second variable
r : Number of categories in the first variable
c : Number of categories in the second variable
$n_{i.}$: Number of observations included in the i-category of the first variable
$n_{.j}$: Number of observations included in the j-category of the second variable
n : Total number of observations

In the decision taken from this *chi-square* test, H_0 is rejected if the calculated X^2 value is $>X^2$ table or the *p-value*$<\alpha$ which states that there is a significant relationship between variables.

According to Bagozzi (1994), explains that the steps of CHAID analysis are broadly divided into three stages, namely: Merging, in the first stage the significance of each independent variable category will be examined against the dependent variable. Splitting, in this second stage choosing the independent variable which will be used as the best-split node, the selection is done by comparing the *p-value* (from the merging stage) on each independent variable. Stopping, in this third stage, is done if the tree growth process must be stopped if there are no more significant independent variables showing differences from the dependent variable.

CHAID will produce a classification tree diagram that describes the formation of segments. The CHAID diagram consists of a tree trunk divided into smaller branches starting from the root node through three stages at each node formed and repeating. In general, the tree diagram of CHAID is as follows: (Lehmann and Eherler, 2001).

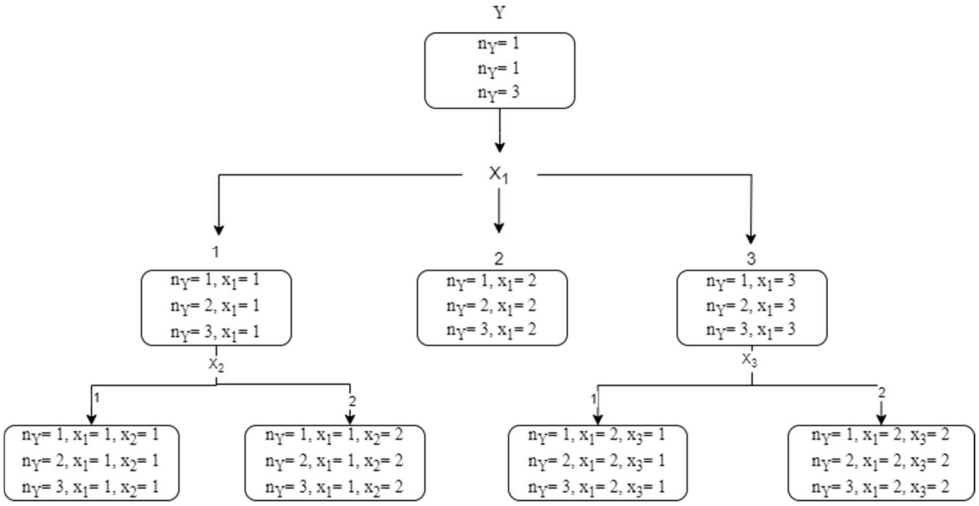


Figure 1. CHAID Classification Tree Diagram

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81

Figure 1 is a flow chart of the CHAID method classification tree. In the CHAID decision tree, there is an actual value *(y_i)* and an expected value *($\hat{y}_i$)*. The actual value is the actual value of the dependent variable that the model wants to predict, while the expected value is the value predicted by the model for the dependent variable based on the conditions at the decision tree node.

C. Error Rate Prediction Method

Error rate prediction for the model is to measure all forms of prediction error in the model. *Cross-validation* is a method that aims to obtain maximum accuracy results. This method will test the effectiveness of a model that has been formed by dividing the data into two parts, namely training data and testing data. Training data is used to train the model so that the model obtained can understand the data pattern and can be validated, while data testing is used to carry out tests (Refaeilzadeh *et al*, 2016).

Error is a measure used to assess the accuracy of the classification tree using various indicators with the following formula (James *et al*, 2013):

$$Err_i = I(y_i \neq \hat{y}_i)$$

Where,

$$I = \begin{cases} 1, y_i \neq \hat{y}_i \\ 0, y_i = \hat{y}_i \end{cases}$$

The purpose of *cross-validation* is to test a model from the available data with the algorithm contained in it and then compare the results obtained for each to get the best algorithm. *Cross-validation* has several learning algorithms and the most commonly used are *LOO*, *Hold-out*, and *K-fold* (James *et al*, 2013).

D. LOOCV

LOOCV is one of the best model validation methods by dividing data into two parts, namely training data and testing data based on the number of observations. Where the training data contains N-1 observations and 1 observation becomes testing data. The calculation process can be repeated *k* times which always leaves 1 observation for data testing (James *et al*, 2013). The level of accuracy obtained is almost unbiased but the resulting variance is higher (Refaeilzadeh *et al*, 2016). To calculate the error value in the LOOCV algorithm, the following formula can be used (James *et al*, 2013):

$$\widehat{Err}^{LOO} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

E. *Hold-Out Cross-Validation*

Hold-out is the best model validation method by dividing the data into two parts randomly with the number of observations divided by 2/3 for training data and 1/3 for testing data, whereas in *hold-out* there is always more training data compared to testing data. This aims to maintain the ratio between classes. In the *hold-out*, not all data have the opportunity to become data testing so it is possible that not all data that becomes data testing can predict well or that become training data can predict well so that the results obtained are very dependent on data separation, but because the distribution is random there still a possibility data has never been data testing (Refaeilzadeh *et al*, 2016). To calculate the error value for the *hold-out*, the following formula can be used (James *et al*, 2013):

$$\widehat{Err}^{Ho} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} I(y_i \neq \hat{y}_i)$$

F. *K-Fold Cross-Validation*

K-fold is used to estimate prediction error in evaluating model performance and is one of the best model validation models by dividing data into *k* subsets with the same sample size for each subset. Many observations commonly used are 5 or 10 groups (*fold*) (James *et al*, 2013). Four factors affect the estimated accuracy obtained by using the *k-fold*, namely the number of *folds*, the number of observations in one *fold*, the average level, and the repetition of the *k-fold*. To determine the best model is to look at the model that has the smallest average value. To calculate the error value in the *k-fold* with 5 groups can be used the following formula (James *et al*, 2013):

$$\widehat{Err}^{CV} = \frac{1}{5}\sum_{k=1}^{5}\frac{1}{n_k}\sum_{i=1}^{nk} I(y_i \neq \hat{y}_i)$$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81

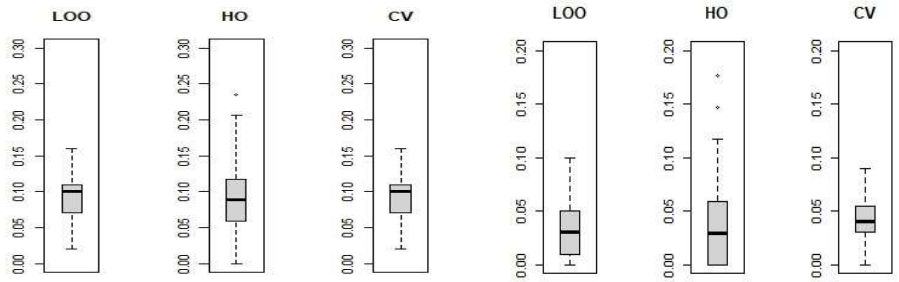
G. Imbalanced Data

So far, several studies have found due to the unequal distribution of data, the number of one data class is less or more than the number of other data classes (Ali *et al*, 2015).This condition is called imbalanced data. The application of the CHAID method uses categorical data. In categorical data, it is very susceptible to an imbalance in the amount of data between the majority class and the minority class. The majority class is a class that has a larger amount of data than the minority class which has a smaller amount of data (Chawla *et al*, 2002). One of the main problems with class unequal distribution of data is that most standard algorithms are accuracy driven. However, in the imbalanced data class, classification accuracy shows little in the minority class.

III. RESULTS AND DISCUSSION

A. Analysis Results

This study aims to obtain the best error rate prediction algorithm that is suitable for use in the CHAID method for imbalanced data. The error rate prediction performance can be seen from the error rate variations displayed by the boxplot. Then you can also see the effect of differences in the marginal opportunity matrix, correlation, and data imbalances on the prediction of the error rate. The following is a comparison boxplot for the error rate prediction method.



(a) **(b)**

Figure 2. Error Rate Prediction Results on Data (a) Univariate (b) Bivariate

Figure 2 is the result of the error rate prediction from two different types of data, namely univariate and bivariate using setting 1. In Figure 2, it can be seen that the *Hold-out* algorithm has the highest error rate variation value among the three error rate prediction algorithms. However, the resulting median error rate tends to be smaller than other algorithms. The LOOCV and*K-fold* algorithms have comparable performance in predicting error rates with almost the same variation values, but if looked closely, the error rate variations for the *K-fold* algorithm are smaller than LOOCV. Based on this explanation, it can be concluded that the *K-fold cross-validation* method is a better error rate prediction method and is suitable for use in classification modeling with the CHAID method for imbalanced data. After that, it turns out that based on correlation on bivariate data, the *K-fold* algorithm is also better. The following is the result of a comparison of the correlation to the error rate for bivariate data settings.

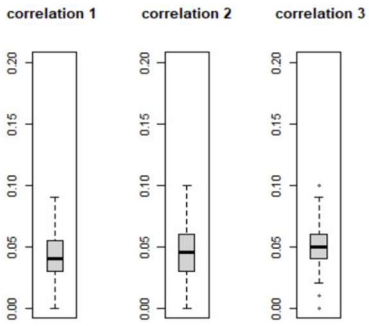


Figure 3. Comparison of *K-fold* Boxplot Based on Correlations in Bivariate Data

Figure 3 is a correlation comparison in the *K-fold cross-validation*. It can be seen that the difference in the predicted value of the error rate for different cases. Correlation 1 is the k-fold algorithm boxplot on bivariate data

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81

without a correlation value, correlation 2 is a moderate correlation value, and correlation 3 is a high correlation value. Based on the comparison results, it appears that error rate results have a tendency for the median value to increase, and in the end, the variation values get smaller.

The data generated with different data class proportions produces different error rate prediction values, as can be seen in Figure 4.

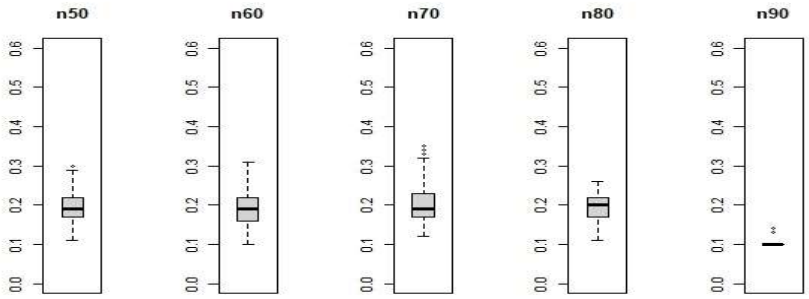


Figure 4. Comparison of Error Rate Results with Different Class Ratios of Observations on *K-fold* univariate data

In Figure 4 it can be seen that more the imbalanced data, the smaller the predicted value will be. Nonetheless, this does not indicate that as imbalanced data increases, it is considered better. On the contrary, such an assumption is incorrect. In imbalanced data, prediction errors often occur in the minority class so the resulting error rate tends to be smaller. This is related to Figure 4, when the comparison of the proportion of class data is balanced with n=50:50, the error rate results spread between 0.1 to 0.3, but for data with class proportions that are very imbalanced with n=90:10 the error rate results only is around 0.1.

B. Discussion

From the results of the analysis that has been described, the settings on the simulation data affect the predicted values of different error rates. The predicted value of the error rate will be directly proportional to the correlation value, the predicted value of the error rate will be even greater if the correlation value used is also greater. The *Hold-out* algorithm has the greatest error rate variation and the smaller median value, this is because this algorithm tends to be unstable in predicting the error rate because the amount of training data is less than the other algorithms. The LOOCV and *K-fold* algorithms show nearly the same performance in predicting the error rate. However, the *K-fold* algorithm is superior with a smaller error rate variation. So it can be said that the best error rate prediction method is the *K-fold cross-validation*.

The ratio of the number of different observations also influences the results of the error rate. Simulation data with a balanced ratio of the number of observations with n=50:50 will produce a larger error rate variation, conversely, if the data ratio is very imbalanced with n=90:10 it tends to produce a small error rate variation. This is because the more imbalanced data, the resulting error rate will approach the ratio of the smallest number of observations resulting in a smaller error rate.

IV. CONCLUSION

Based on the results of the three algorithms of the error rate prediction method with the CHAID method for imbalanced data, it was found that the error rate prediction method that has the best error rate value is the *K-fold cross-validation* method. The influence of the marginal opportunity matrix, the correlation structure, and the ratio of the number of observations that differ gives different error rate results. When the correlation value will be greater, the error value will be greater and the more unequal the ratio of the number of observations, the resulting error rate tends to be smaller. It is hoped that future research can develop new research by adding data imbalance handling to get better results.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis* (Third). John Wiley & Sons. Inc

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem: A review Classification Int. J. Advance Soft Compu. Appl, Vol. 5 No. 3.

Bagozzi, R. P. (1994). *Advanced Methods of Marketing Research*. Blackwell Publisher Ltd, Oxford.

Baron, S., & Phillips, D. (1994). Attitude Survey Data Reduction Using CHAID: An Example in Shopping Centre Market Research. *Journal of Marketing Management*, Vol. 10 No. 1–3, 75–88.

Chawla, N. v, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Demirtas, H. (2006). A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, Vol. 76 No. 11, 1017–1025.

Eherler, D., & Lehmann, T. (2001). *Responder Profiling with CHAID and Dependency Analysis*.

Gallagher, C. A., Monroe, H. M., & Fish, J. L. (2000). An Iterative Approach to Classification Analysis. *Journal of Applied Statistics*, 238–280.

James, G. Gareth M., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Source: Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 29. No. 2, 119–127.

Kohavi, Ron. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of 14th International Joint Conference on AI; pp.* 1137–45.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons. Inc.

Reddy, U. S., & Somasundaram, A. (2016). Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data. *Proc. of 1st International Conference on Research in Engineering, Computers and Technology*. https://www.researchgate.net/publication/320895020

Refaeilzadeh, P., Tang, L., Liu, H., 2016. Cross Validation. *Encyclopedia of Database Systems*.

Vluymans, Sarah. 2019. *Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods*. Belgium: Springer.

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *Proceedings- 6th International Advanced Computing Conference*.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/81