International Journal Of Artificial Intelegence Research
Vol 3, No 2, December 2019, pp. 47-55 ISSN: 2579-7298

# Solution of class imbalance of k-nearest neighbor for data of new student admission selection

Siti Mutrofin[a,1,*], Ainul Mu'alif[a,2], Raden Venantius Hari Ginardi[b,3], Chastine Fatichah[b,4]

[a] *Information Systrem, Universitas Pesantren Tinggi Darul Ulum, Ponpes Darul Ulum, Jombang, 61481, Indonesia*
[b] *Informatic Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
[1] *sitimutrofin@ft.unipdu.ac.id*;* [2] *ainulmualif25@gmail.com;* [3] *hari.ginardi@gmail.com;* [4] *chastine@cs.its.ac.id*
** corresponding author*

ARTICLE INFO

*Article history:*
Received: 17 May 2019
Revised : 23 June 2019
Accepted : 30 June 2019

*Keywords:*
class balance
class imbalance
EDM
kNN
*tahfiz*

ABSTRACT

The objective of this research is to correct the inconsistencies associated with the response differences by each examiner with respect to the assessment of each *hafiz* candidate. To carry out this research, 259 students were selected within a week using 4testers. However, the examiners are also tasked with another essential mandate which must be immediately fulfilled asides testing candidates for *hafiz*. In order to overcome this problem, the Educational Data Mining (EDM) system is applied during classification. The problems associated with the use of this technique however, is the limited number of attributes and the imbalance data class. This study was proposed to apply the kNN (k-Nearest Neighbor) technique. The results obtained indicates that kNN can provide recommendations to testers who are students and it is suitable for the solving the problem associated with class imbalance as indicated by the application of Shuffled and Stratified sampling techniques which has values of accuracy, precision, recall and AUC > 0.8%.

Copyright © 2019 International Journal of Artificial Intelegence Research.
All rights reserved.

## I. Introduction

*Madrasah Tsanawiyah* Plus Darul Ulum (MTs Plus DU), is an Islamic Boarding School located in the Complex of Darul Ulum, Sub-district of Peterongan, District of Jombang, Province of East Java, with superior *Tahfiz* program. It selects prospective Holy Quran Memorizers for new students to make the Program successful.

Current problem facing MTs Plus DU is the scarce of human resources who act as examiners in the selection program of prospective Holy Quran Memorizers. Every year, 259 students are examined by only four testers using four criteria consisting of ability to read, write, and readiness to memorize the Holy Quran. The assessment duration given by MTs Plus DU is a week. Every year, 60 students are admitted into the *Tahfiz* Program, a fact which indicates examiners give different assessment results comprising of numeric, and nominal questions. Example the Holy Quran memorization examination indicated that Examiner 1 gave articles to assess students memory, the second was on its various types, while the third and fourth wrote list of articles to memorize, therefore their perceptions were different as illustrated in Table 1, Table 2, Table 3 and Table 4. In general, examiners also serve as teachers, who carry out administrative activities.

Many researches are associated with *Tahfiz* Program adopting information technology (IT), however, these studies only make applications to evaluate Holy Quran memorizers, rapport and [1], monitoring application [2] [3], cluster application used k-Means for selection of prospective MTQ competition participants [4]. There is no proposal of previous studies in solving problems associated with this case.

This study proposed the application of the classification method from data mining to predict prospective *Tahfiz* Program participants, irrespective of their educational qualifications. Therefore,

Table 1. Examples of assessment results by Examiner 1

| No | Assessment Indicator | | | |
|---|---|---|---|---|
| | **Ability to read the Qur'an** | **Ability to write the Qur'an** | **Ability to memorize the Qur'an** | **Al-Qur'an letters that have been memorized** |
| 1 | 70 | 70 | 80 | 5 |
| 2 | 70 | 80 | 50 | 8 |
| 3 | 60 | 65 | 50 | 8 |
| 4 | 70 | 80 | 75 | 9 |

Table 2. Examples of assessment results by Examiner 2

| No | Assessment Indicator | | | |
|---|---|---|---|---|
| | **Ability to read the Qur'an** | **Ability to write the Qur'an** | **Ability to memorize the Qur'an** | **Al-Qur'an letters that have been memorized** |
| 1 | 60 | 60 | ✓ | Short letter |
| 2 | 75 | 65 | × | Short letter |
| 3 | 75 | 75 | ✓ | Short letter |
| 4 | 75 | 70 | ✓ | An-Naba – An-Nazi'at |

Table 3. Examples of assessment results by Examiner 3

| No | Assessment Indicator | | | |
|---|---|---|---|---|
| | **Ability to read the Qur'an** | **Ability to write the Qur'an** | **Ability to memorize the Qur'an** | **Al-Qur'an letters that have been memorized** |
| 1 | 90 | 80 | Yes | An-Nas – Ad-Duha |
| 2 | 60 | 60 | Yes | An-Nas & Al-Ikhlas |
| 3 | 85 | 60 | Yes | An-Nas – Al-Ikhlas |
| 4 | 90 | 80 | Yes | Al-Ma'un |

Table 4. Examples of assessment results by Examiner 4

| No | Assessment Indicator | | | |
|---|---|---|---|---|
| | **Ability to read the Qur'an** | **Ability to write the Qur'an** | **Ability to memorize the Qur'an** | **Al-Qur'an letters that have been memorized** |
| 1 | 65 | 60 | Yes | اَلْمَاعُونَ |
| 2 | 78 | 60 | Yes | العصر |
| 3 | 78 | 60 | Yes | العصر |
| 4 | 78 | 60 | Yes | العصر |

final decisions made by examiners in a week could be replaced by a system within a shorter timeframe. There is no classification technique applied to predicting graduation of prospective Holy Quran memorizers, however, the classification method was used to determine students with the potential to discontinue their education with difference in the right attributes to utilize [5]. The classification algorithm used to predict this, may not be indifferent, however data characteristics must first be considered. In this study, its attributes are Class Imbalance, because the ratio between students who would be admitted (0.23%) and those not admitted (0.77%) was proportional. It is considered balance

if the ratio was 0.35%:0.65% [5]. Nevertheless, not all algorithms are consistent to solve problem of Class Imbalance, for example, algorithm of Decision Tree (DT) family, is different from k-Nearest Neighbor (kNN) which has better performance [6]. Some advantages indicate that the computation is low, simple, easy to learn, resistant to noise, and effective if training data are large [7]. In this study, the kNN application was used to select web-based *Tahfiz* Program students. The aim was to recommend examiners in making decision, thereby minimizing time taken.

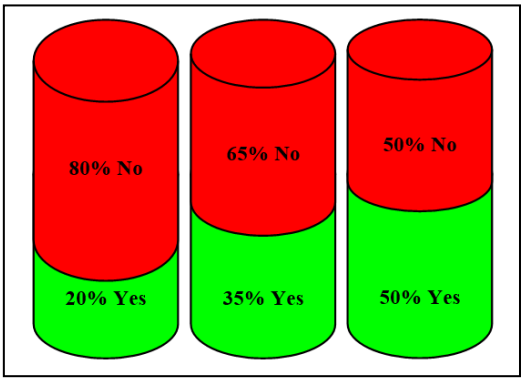


Fig. 1. Left is imbalances class, center and right is balances class

## II. Class Imbalance

Class Imbalance is also called imbalance class or imbalance data [5], as illustrated in Figure 1. Thammasiri et.al performed studies associated with it in order to predict students potential to discontinue their education during their first academic year [5]. Class distribution between those who continued their education and those who discontinued it had the following ratio: 21.3%:78.7%. It is considered balance if it has the following ratio 35%:65% [5]. A total data of 21,654 students with 34 attributes were obtained from 2005 to 2011 . The proposed algorithms are artificial Neural Network (ANN), support Vector Machines (SVM), Decision Tree (DT), and Logistic Regression (LR). The problem of class imbalance is solved by applying sampling technique consisting of Random under-sampling (RUS), Random over-sampling (ROS), and Synthetic minority over-sampling technique (SMOTE). Performance evaluation is assessed by the following nine measurements namely accuracy, sensitivity, specificity, precision+, FP-Rate, F-measure, CC, and GMEAN. Results of studies by Thammasiri et.al [5] indicate that SVM combined with SMOTE has best performance compared with other algorithm and data collection techniques.

Brown and Mues [6] conducted other studies associated with class imbalance in a financial institution having some services, such as, money loan. Ability to pay and repay for services determines the life of an organization with respect to money loan service, so that, selection of prospective debtors must be carefully conducted to avoid inconsistencies associated with decision-making. One proper way to receive loan is that it should not be rejected, and vice versa. Manual decision-making may be inconsistent, due to the numerous numbers of factors. Data mining plays a role in decision-making, due to the use of loan history whether debtor is good or bad with ratio of 70%:30%, using various characteristics. Brown and Mues [6] studied algorithm performance for class imbalance problem. The data [6] used were collected from UCI Machine Learning Repository. The proposed algorithms were Logistic Regression, Linear and Quadratic Discriminant Analysis, Neural Networks (Multi-layer perception), Least Square Support Vector Machines (LS-SVM), C4.5. Decision Trees, k-AND (memory based reasoning), Random Forests, and Gradient Boosting. Classification algorithm performance was evaluated using average rank (AR) and Area Under Curve (AUC). AR is found good if its value is low, and inversely proportional to AUC [6]. Results of study by Brown and Mues [6] are Random Forests and Gradient Boosting having best criteria for class imbalance problem, and inversely proportional to Decision Trees and Quadratic Discriminant Analysis (QDA). Meanwhile, kNN and ANN have sufficient good performance.

## III. Method

Figure 2 shows the proposed method, consisting of raw data collection, preprocessing, kNN algorithm determination, 10-fold cross validation, and evaluation model such as accuracy, precision, recall, and AUC.

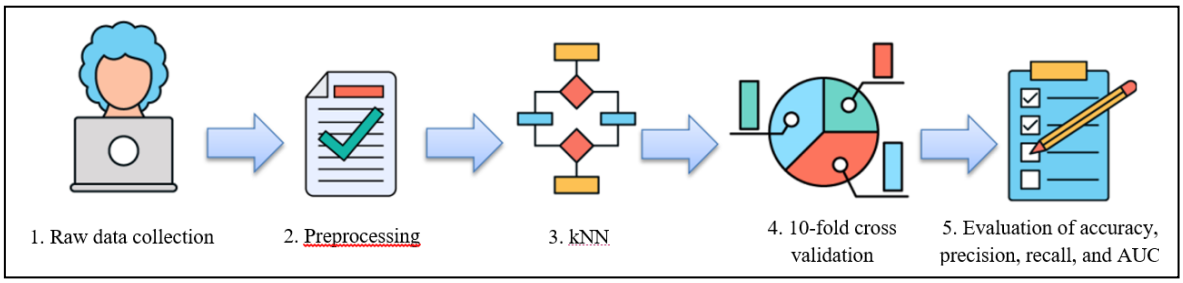


Fig. 2. The proposed method

### *A. Raw Data Collection*

Raw data were collected from the 2018 MTs Plus DU manual assessment sheet of Jombang, such as, assessment of prospective *hafiz* selection in *tahfiz* program.

### *B. Data Preprocessing*

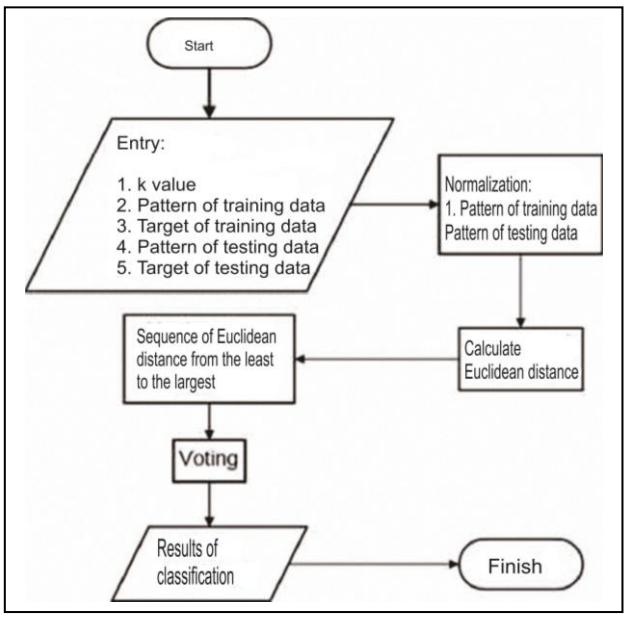


Fig. 3. kNN algorithm flowchart

This consisted of 270 assessments of students subjected to selection, though they experienced redundancy, therefore this study removed duplicate data, resulting to 259 students. The data comprises of the following four assessment criteria namely ability to read, write, memorize and readiness to take *tahfiz* program. The four predetermined attributes are based on the school policies. This result is in line with the outcome of the research [8] which stated the main determinant factor for the success of memorizing Al-Qur'an or *hafiz* as the internal factor made up of motivation and interest. This is shown in their readiness in memorizing Al-Qur'an. However, the three other criteria are based on one's ability to apply the method of memorization and everyone has the advantages and disadvantages in applying it. The Tikorul Mahfudz is a memorization method very suitable for someone with weak memory. Aside this, the Isatima'ul Mahfudz method could also be applied. It involves listening to the recitation of Al-Qur'an through tapes or from other people. The method is suitable for people with physical limitations such as blindness [9]. Also, there is another method suitable for people with strong memories. This is known as Kitabul Mahfudz [9] and it involves rewriting verses from Al-Qur'an. Furthermore, reading skills are needed such that the memorizer or

*hafiz* is able to memorize as well as read the Qur'an. Therefore, memorization by reading in advance always give good results during recitation. The use of data with only 4 attributes has also been carried out by other researchers, including the use of Iris data provided by the UCI Machine Learning. Memorization criteria assessed by examiners are different, for instance, Examiner A gave numeric values to memorization, while B gave nominal value. The problem of the different assessments was solved by making them similar (uniformity) in accordance with the standard specified by MTs Plus DU of Jombang. Assessment of current readiness to memorize is nominal "yes" and "no". In numeric, the values are converted into binary numbers 1 and 0, where 1 is "yes" and 0 is "no", furthermore, the calculation of kNN algorithm would use Euclidean distance, which is only capable of handling such data.

### *C. kNN Classification Algorithm*

kNN algorithm has the following strengths [7]:

1. simple calculation;
2. low computation;
3. easy to learn;
4. resistant to noise;
5. Effective with large training data.

kNN is also found to perform better than other algorithms such as *Decision Trees* and QDA families [6].

Figure 3 shows kNN algorithm flowchart using the following steps:

1. The minimal value of *k* is 1 and its maximal value is the total training data-1; including its training and testing pattern as well as its, training, and testing data target.
2. Normalizing all training data and testing data patterns. The aim is to make overall value interval of pattern own same interval, between 0 and 1, due to the different value interval between 0 – 1 for readiness value, but there is 0–100 interval for the three remaining assessments. Normalization calculation used was min-max method as shown in equation (1) [8],

$$Normalization = \frac{data_x - data_{\min}}{data_{\max} - data_{\min}} \quad (1)$$

Where:

$data_x$ is data in which calculation of normalization is based on column of data,

$data_{min}$ is least data in same column as data in which normalization will be calculated; and

$data_{max}$ is largest data in same column as data in which normalization will be calculated.

3. Calculating Euclidean distance symbolized as $d_{Euclidean}(x,y)$. Equation (2) shows calculation of Euclidean distance [10],

$$d_{Euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (2)$$

Where:

$i$ = the number of data, $x$ = testing data, and $y$ = training data.

4. Voting from least Euclidean distance included *k* ranking.
5. Determining classification results based on majority voice.

### *D. Validation Model*

Validation model used was 10-fold cross validation, which means that all data would be divided into ten parts, with each used to train and test it. Table 5 shows illustration of 10-fold cross validation [11], where the dark areas are testing data, and the white areas training data.

### *E. Evaluation Model*

Evaluation models used in this study to understand kNN performance in class imbalance problem were accuracy, precision, recall, and AUC. However, accuracy is the only inadequate amongst them due to the fact that it does not mean high precision and recall, while AUC is useful to understand whether classification algorithm is good or not. This evaluation calculation used confusion matrix consisting of True Positive (TP), False Matrix (FP), True Negative (TN) and False Negative (FN). FP value is negative (N), but it is misclassified as positive (P). Results of FN actually belongs to P which can be formulated as P = FN + TP, but, when misclassifying, instance will be N, which can be mathematically expressed as N = FP + TN. Confusion matrix is useful to calculate accuracy, precision, recall, and AUC.

Table 5. Stratified 10-fold cross validation [11]

| ***n*-validation** | **Partition of *dataset*** |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

## IV. Results and Discussion

The trial is conducted up to 18 times due to the weakness of the k-NN which is biased to the parameter *k* value. It is necessary to carry out many trials in order to obtain the most optimal *k* value. Therefore in this study, the *k*-test was conducted 6 times each from the random sampling method consisting of Linear, Shuffled and Stratified. The determination of random *k* values from the smallest $k = 1$ to the largest $k = 200$, as shown in Table 6, gives $k = 10$ as the optimal value for the balance data between training and test data (Shuffled and Stratified). However, for imbalance data, the optimal *k* value is $\geq 100$. Asides the value of *k*, k-NN is strongly influenced by the balance class distribution between the training and test data. It is seen in Table 6 that the Stratified sampling representing the balance class distribution has the best value compared with the imbalance by linear sampling.

It could be proven in this study that despite the datasets having the imbalanced class distribution characteristics, the classification of k-NN was successful. This is as a result of the same value existing between the training and percentage test data of each class, upon the application of both Shuffled and Stratified sampling methods. However, the Linear sampling method produces poor results for the different percentage of each class between the training and test data. The k-NN has proven to be good with an accuracy > 84%, precision> 85%, recall> 94%, and AUC> 0.78.

Also, the k-NN is not only proven to be a simple algorithm in terms of calculation and application, but also reliable in terms of its classification capability, though there exist some weaknesses of the bias value. The number of closest neighbors is not directly proportional to the results of performance of k-NN. Its general performance is influenced by the percentage of the class distribution between the training and test data, as well as the characteristics of the dataset. In Table 6, it is clear that the determination of $k = 1$ is too small to produce a poor result from the three random sampling methods

used. The k-NN with $k$ = 1 means that the test data is only based on one of the closest neighbors, which may be the one who are more like the test data.

Table 6. Results of experiment

| ***Sample*** | ***k*** | ***Accuracy* (%)** | ***Precision* (%)** | ***Recall* (%)** | **AUC** |
|---|---|---|---|---|---|
| **Linear** | **1** | **64.52** | **76.67** | **72.86** | **0.5** |
| | 10 | 78.45 | 76.67 | 93.47 | 0.044 |
| | 25 | 76.15 | 76.67 | 95.48 | 0.055 |
| | 50 | 75 | 76.67 | 96.98 | 0.06 |
| | 60 | 76.54 | 77.08 | 98.99 | 0.06 |
| | **100** | 76.92 | 76.92 | **100** | 0.06 |
| | **200** | 76.92 | 76.92 | **100** | 0.06 |
| **Shuffled** | **1** | **73.74** | 85.37 | **80.22** | **0.5** |
| | **10** | **84.2** | **85.72** | 94.57 | 0.826 |
| | 25 | 83.4 | 84.64 | 96.03 | 0.853 |
| | 50 | 83.8 | 82.84 | 99.57 | 0.869 |
| | 60 | 81.51 | 80.64 | 100 | 0.869 |
| | 100 | 76.89 | 76.89 | 100 | 0.857 |
| | 200 | 76.89 | 76.89 | 100 | 0.85 |
| Stratified | **1** | **72.98** | 84.65 | **79.34** | **0.5** |
| | **10** | **84.55** | **86.19** | 95.5 | 0.781 |
| | 25 | 83.03 | 83.93 | 96.5 | 0.853 |
| | 50 | 83.42 | 82.86 | 99 | 0.869 |
| | 60 | 82.65 | 81.66 | 100 | 0.868 |
| | 100 | 76.85 | 76.85 | 100 | 0.863 |
| | 200 | 76.85 | 76.85 | 100 | 0.855 |

Where: **Red = bad, blue = good**

The Shuffled sampling and stratified sampling methods using the nearest neighbor $k > 60$ resulted in the k-NN having a 100% recall value. This is an indication that k-NN could correctly guess which students pass and fail in each selection period (only 60 students accepted). Therefore, using the 60 closest neighbors with the balance data of each class between the test and balanced training data gives a 100% recall value. But the value of $k$ for the imbalance data is 60, and therefore could not achieve a 100% recall.. This is because the amount of passing grade data of 60 out of 259 could not be equal in terms of percentage between the trainingand test data upon encrypted.

## V. Conclusions

This study succeeded in recommending to examiners to determine the educational status of students within a limited period of time. Since all assessment data are obtained in seconds, the researchers are able to determine whether the students would graduate from the *tahfiz* program or not. In this system, there is also no significant difference in perceptions of examiners. The k-NN is proven to have good performance for imbalance class distribution when the percentage of each class between training and test data is the same. However, it gives bad results when the percentage between training and test data is not the same. Further research is expected to be able to make comparisons between the

algorithms or other methods, use of various datasets in terms of number of attributes. Then, the datasets should have the characteristics of balance class distribution, in order to have a better understanding as regards the performance of k-NN compared with other methods while solving problems related to imbalanced class distribution.

## Acknowledgment

The researcher is grateful to the Directorate of Research and Community Service, Director General of Research and Development Reinforcement, Ministry of Research, Technology and Higher Education, for financing this research on Inter-University colaboation in 2018 with titled "Classification-based Educational Data Mining to Analyze Students Having Potential to discontinue their Education in case of Imbalanced Class Distribution". Consistent algorithms with Class Imbalance problems are Random Forests, Gradient Boosting, Linear Discriminant Analysis (LDA), Neural Networks (NN), and Logistic Regression (LR). The use of Decision Tree and Quadratic Discriminant Analysis (QDA) is not recommended for Class Imbalance problems.

## References

[1] A. B. Hakim and V. Ramdhani, "Perancangan dan Pengembangan Prototipe Aplikasi Mobile Untuk Lembaga Penghafal Quran Berbasis Android Menggunakan Metode Rapid Application Development," *I-STATEMENT*, vol. 3, no. 2, pp. 74-88, 2017.

[2] R. Wulandari, "Rancang Bangun Sistem Informasi Monitoring dan Evaluasi Hafalan Al-Qur'an Program Beasiswa Santri Berprestasi (PBSB) Berbasis Web Pada Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang dengan Metode Extreme Programming (XP)," UIN Maulana Malik Ibrahim, Malang, 2018.

[3] D. Iskandar, S. D. Budiwati and R. Budiawan, "Aplikasi Penilaian dan Presensi Siswa untuk Kegiatan Pembelajaran Akademik (Studi Kasus : SD Ar-Rafi')," in *e-Proceeding of Applied Science*, 2017.

[4] A. T. R. Saragih, A. S. Sembiring and M. Sayuthi, "Penerapan Metode Clustering K-Means untuk Proses Seleksi Calon Peserta Lomba MTQ," *Jurnal Pelita Informatika,* vol. 17, no. 2, pp. 117-122, 2018.

[5] D. Thammasiri, D. Delen, P. Meesad and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Systems with Applications,* vol. 41, no. 2, pp. 321-330, 2014.

[6] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications,* vol. 39, no. 3, p. 3446–3453, 2012.

[7] S. Mutrofin, A. Izzah, A. Kurniawardhani and M. Masrur, "Optimasi teknik klasifikasi modified k nearest neighbor menggunakan algoritma genetika," *Jurnal Gamma,* vol. 10, no. 1, 2015.

[8] D. Fitriyah, "Faktor Yang Mempengaruhi Kecepatan Menghafal Al-Qur'an Antara Santri Mukim Dan Nonmukim Di Pesantren ZaIdatul Ma'aRif Kauman Parakan Temanggung," Institut Agama Islam Negeri Walisongo, Semarang, 2008.

[9] N. Aflisia, "Urgensi Bahasa Arab Bagi *Hafiz*h Al-Qur'an," *FOKUS: Jurnal Kajian Keislaman dan Kemasyarakatan,* vol. 1, no. 1, pp. 47-66, 2016.

[10] D. T. Larose and C. D. Larose, Discovering knowledge in data: an introduction to data mining, 2nd ed., Hoboken: John Wiley & Sons, 2014.

[11] R. S. Wahono, N. S. Herman and S. Ahmad, "A comparison framework of classification models for software defect prediction," *Advanced Science Letters*, vol. 20, no. 10-12, pp. 1945-1950, 2014.