JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Okwonu, F. Z., Ahad, N. A., Hamid, H., Muda, N., & Sharipov, O. S. (2023). Enhanced robust univariate classification methods for solving outliers and overfitting problems. *Journal of Information and Communication Technology, 22*(1), 1-30. https://doi.org/10.32890/jict2023.22.1.1

Enhanced Robust Univariate Classification Methods for Solving Outliers and Overfitting Problems

[1]Friday Zinzendoff Okwonu, [2]Nor Aishah Ahad,
[2]Hashibah Hamid, [3]Nora Muda &
[4]Olimjon Shukurovich Sharipov
[1]Department of Mathematics, Faculty of Science,
Delta State University, Nigeria
[2]School of Quantitative Sciences,
College of Arts and Sciences,
Universiti Utara Malaysia, Malaysia
[3]School of Mathematical Sciences,
Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, Malaysia
[4]Department of Statistics,
School of Mathematics, National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan

*okwonufz@delsu.edu.ng; aishah@uum.edu.my; hashibah@uum.edu.my; noramuda@ukm.edu.my; osharipov@mail.ru
*Corresponding author

Received: 16/2/2022 Revised: 22/9/2022 Accepted: 30/10/2022 Published: 19/1/2023

ABSTRACT

The robustness of some classical univariate classifiers is hampered if the data are contaminated. Overfitting is another hiccup when the data sets are uncontaminated with a considerable sample size. The performance of the classification models can be easily biased by

the outliers' problems, of which the constructed model tends to be overfitted. Previous studies often used the Bayes Classifier (BC) and the Predictive Classifier (PC) to address two groups of univariate classification problems. Unfortunately for substantial large sample sizes and uncontaminated data, the BC method overfits when the Optimal Probability of Exact Classification (OPEC) is used as an evaluation benchmark. Meanwhile, for small sample sizes, the BC and PC methods are extremely susceptible to outliers. To overcome these two problems, we proposed two methods: the Smart Univariate Classifier (SUC) and the hybrid classifier. The latter is a combination of the SUC and the BC methods, known as the Smart Univariate Bayes Classifier (SUBC). The performance of the new classification methods was evaluated and compared with the conventional BC and PC methods using the OPEC as a benchmark value. To validate the performance of these classification methods, the Probability of Exact Classification (PEC) was compared with the OPEC value. The results showed that the proposed methods outperformed the conventional BC and PC methods based on the real data sets applied. Numerical results also revealed that the SUC method could solve the overfitting problem. The results further indicated that the two proposed methods were robust against outliers. Therefore, these new methods are helpful when practitioners are confronted with overfitting and data contamination problems.

Keywords: Bayes, Predictive, Outliers, Overfitting, Classification.

INTRODUCTION

In practice, we often have difficulty allocating or classifying an object into one of two groups based on the object score. In many instances, a comparison between the object score and the benchmark value is made to assign an object to the correct group (Wald, 1944; Song et al., 2020). Classification methods can be applied to many studies to identify unique membership (Pang et al., 2019; Hamid et al., 2018; Okwonu et al., 2012). For instance, El Abbassi et al. (2021) applied a univariate classifier to nanoelectronics and spectroscopy to classify relevant information from the data set. Classification can be applied to determine ICT knowledge awareness (Dávideková et al., 2019). Jimoh et al. (2022) applied classification methods to classify malaria infection. The classification method can also be used to classify students as first-class or second-class upper based on their final CGPA.

However, the classification benchmark value must be well defined and compared before assigning the object to the respective groups (Hamid et al., 2016). Thus, this benchmark formulation criterion is crucial in classification tasks.

Classification methods often require groups and variables of interest to be defined prior to model construction. In some cases, two or more variables could be of interest to the researcher to obtain group information. In contrast, other researchers may be interested in a single variable in order to ascertain actual or predicted group membership.

The classical classification methods often show a minimum misclassification rate if the data set follows a normal distribution pattern. On the other hand, if this pattern is violated, the inference breaks down as we might draw erroneous inferences and wrong conclusions (Das & Imon, 2016). Therefore, normality should be taken seriously, because if this assumption does not hold, it is impossible to draw accurate and reliable conclusions (Field, 2009; Oztuna et al., 2006). Departure from the normality for any samples demonstrated that the Type I error rate is affected (Blanca et al., 2017; Cain, Zhang & Yuan, 2017). The group variables of the study determine the importance of the research inference. In this case, the dimension of the variables may be categorised into univariate or multivariate. The other aspect is the homoscedasticity of the variance when discussing the robustness of these techniques. Besides multivariate classification, univariate classification is essential and widely applied in different fields of study.

In recent times, the paradigm of univariate classification focused on different techniques that have been proposed using time series data (Thet Zan & Yamana, 2016; Sun et al., 2014). For example, the Bayes rule is a unique classifier based on group membership probabilities (Taheri & Mammadov, 2013). The predictive classifier (Huberty & Holmes, 1983) relies on the average between group means. These techniques have been applied to classification problems in different fields of studie (Chattopadhyay et al., 2012; Wei, 2018; Banik, 2019; Barbini et al., 2013; Aborisade & Anwar, 2018; Bharadwaj & Shao, 2019; Trovato, 2016; Berry, 2004; Ma et al., 2011; Theodoridis & Koutroumbas, 2009).

Conventionally, the univariate classification for two groups is performed using two groups: students' -test and power analysis. The

procedures of this test and analysis often rely on the probability value, power values, and sample size to make an inference. Donoho and Jin (2008) devised other procedures to classify dimensional data sets without utilising the covariance matrix. Relying on this complicated classification method, Huberty and Holmes (1983) proposed a univariate variable classifier for two groups by comparing the univariate variable's value with the average of the two group means. This technique was proposed as an alternative to power analysis.

The Bayes Classifier is assumed to be robust because it uses the posterior probability to assign an object to the actual group. The robustness of this method can be attributed to the non-application of any measure of central tendencies, such as the mean, which is easily influenced by the outliers. Intuitively, outliers in one group may increase the probability of that group over the other and may influence the correct classification of the object into its actual group. The evaluation of this technique is conventionally optimal if the Probability of Exact Classification (PEC) is measured by $\sum_{i=1}^{2} P_i = 1$, where P_i is the group membership probabilities. The pitfall with evaluating the Bayes procedure using the Optimal Probability of Exact Classification (OPEC) is overfitting, especially when the data set satisfies the normality assumption. Outliers often influence the univariate predictive and the Bayes classifiers. These methods are sensitive to overfitting when the data sets are normally distributed.

In this paper, we proposed two methods, namely the Smart Univariate Classifier (SUC) and the hybrid classifier, which was a combination of the Smart Univariate Classifier (SUC) and the Bayes Classifier (BC). This newly constructed hybrid method was called the Smart Univariate Bayes Classifier (SUBC). The two proposed methods were designed to address the weaknesses of the conventional univariate methods concerning the sensitivity to outliers and overfitting problems. The SUC method mimicked Fisher's linear classifier, while the SUBC method depended on data transformation and plug-in using the Bayes classifier. First, these two methods extracted outliers by applying the F-weight to determine the inlier of each data set. Then, the second phase of the proposed methods transformed the outlier data points into inlier data points before the classifier was applied to determine group membership. Finally, the hybrid method applied the Bayes rule to perform group classification on the transformed F-Weighted data set. The main contribution of this paper is the robustness of

the two proposed methods to solve the overfitting problem and the transformation of outliers to inliers. Furthermore, we investigated the classification performance of the proposed methods and the conventional classification methods using continuous and discrete variables.

The next part of this paper discusses the Predictive Classification method, followed by the Bayes Classifier. The proposed Smart Univariate Classifier and Smart Univariate Bayes Classifier are outlined in the next section, followed by data collection and comparative performance analysis. The performance comparison through the probability of exact classification (hit ratio) and analysis is then presented, and finally the conclusion is conferred in the last section.

METHODOLOGY

In this section, we discuss the conventional methods, Predictive Classifier (PC) and Bayes Classifier (BC), with the proposed methods, Smart Univariate Classifier (SUC) and Smart Univariate Bayes Classifier (SUBC). The different classifiers are based on step computational procedures defined in the following sections.

Predictive Classifier

We describe the predictive classifier according to Huberty and Holmes (1983). This technique classifies an object X_i to Group 1 (Δ_1) according to Equation 1.

$$X_i < \frac{(\bar{X}_1+\bar{X}_2)}{2} \tag{1}$$

where $\bar{X}_i$ is the mean vectors of Group Δ_i as shown in Equation 2.

$$\bar{X}_i = \frac{\sum_{k=1}^{n_i} x_k}{n_i}, \ \ i = 1,2 \tag{2}$$

Otherwise classify X_i to Group 2 (Δ_2). Equation 1 is very sensitive to outliers as the mean is easily perturbed with a slight change in the data set. The basic idea of Equation 1 is variable swap, in which the independent variable is swapped with the dependent variable and vice versa. This procedure mimics the point biserial correlation coefficient in which a variable swap is applicable as shown in Equation 3.

$$r_{p,b} = \frac{(\bar{x}_2 - \bar{x}_1)\sqrt{p \times q}}{\sqrt{s_z^2}} \quad (3)$$

where $n = n_1 + n_2$, $p = \frac{n_1}{n}$ and $q = \frac{n_2}{n}$ are the group probabilities of Group 1 and Group 2, and s_z^2 is the pooled covariance of Groups 1 and 2. To avoid having a negative value of the correlation coefficient, the mean of the first group is always greater than the mean of the second group for efficiency and better classification purposes. In this case, n_1 refers to the smaller group and n_2 refers to the larger group. The concept of variable swap can be reversed by changing the classification "inequality". Based on this statement, Equation 1 can be written as the following Equation 4, which implies that the conventional variable position remains.

$$X_i > \frac{(\bar{X}_1 + \bar{X}_2)}{2} \quad (4)$$

Equation 4 implies that an object X_i is assigned to Δ_1, otherwise, assign X_i to Δ_2. Unfortunately, Equation 1 and Equation 4 give the same classification result. The biserial point correlation that corresponds to Equation 4 is given in Equation 5.

$$r_{p,b} = \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{p \times q}}{\sqrt{s_z^2}} \quad (5)$$

Taking the absolute value of Equation 5 yields the corresponding value of Equation 3. At this point, we have addressed what we may consider the weakness of variable interchange in Equation 1. The remaining part of this paper considers other univariate classifiers, such as the Bayes classifier and the proposed univariate classifier methods, SUC and SUBC.

Bayes Classifier

Given that $g_1(x)$ and $g_2(x)$ are the probability density functions and $X_{d\times 1}$ is a random variable for the two groups denoted as Δ_1 and Δ_2, any value of $X_{d\times 1}$ can be assigned to any of the groups. In the following, we assume that the priori probabilities and the cost of inaccurate classification for the two groups are equal. Let $C_i = \{x_{i,d\times 1}, d = 1\}$ be the sample spaces of the univariate random variables. Suppose ∇_1 is the values of $x_{1,d\times 1}$ to be classified into Group Δ_1, and ∇_2 denotes the values of $x_{2,d\times 1}$ to be classified into Group Δ_2. The aim is that each random value or object must only be classified into one group (Johnson & Wichern, 1992). This univariate approach is based on the conditional probability concept. Let us describe the process of classifying the object as Equation 6 and Equation 7.

$$P(1|2) = P\left(x_{1,d\times1}\epsilon\nabla_1|\Delta_2\right) = \int_{\nabla_1} g_2(x)dx \quad (6)$$

$$P(2|1) = P\left(x_{2,d\times1}\epsilon\nabla_2|\Delta_1\right) = \int_{\nabla_2} g_1(x)dx \quad (7)$$

Equation 6 implies the conditional probability $P(1|2)$ of assigning an object to Group Δ_1 when it is actually from Group Δ_2, while Equation 7 implies the conditional $P(2|1)$ of assigning an object to Group Δ_2 when it is from Group Δ_1. Let us denote the priori probability of Δ_1 as w_1 that is $P(\Delta_1) = w_1$, and Δ_2 as w_2 implying that $P(\Delta_2) = w_2$ such that adding this two probabilities will produce a total of 1 as shown in Equation 8.

$$P(\Delta_1) + P(\Delta_2) = \sum_{i=1}^{2} w_i = 1 \quad (8)$$

Let A_C denotes the correct allocation and A_M is the misallocation. Therefore, the probability of correct allocation or the probability of misallocation for the two groups can be expressed as Equations 9 to 12.

$$P(A_C\ to\ \Delta_1) = P\left(x_{i,d\times1}\epsilon\nabla_1|\Delta_1\right)P(\Delta_1) = P(1|1)w_1 \quad (9)$$

$$P(A_M\ to\ \Delta_1) = P\left(x_{i,d\times1}\epsilon\nabla_1|\Delta_2\right)P(\Delta_2) = P(1|2)w_2 \quad (10)$$

$$P(A_C\ to\ \Delta_2) = P\left(x_{i,d\times1}\epsilon\nabla_2|\Delta_2\right)P(\Delta_2) = P(2|2)w_2 \quad (11)$$

$$P(A_M\ to\ \Delta_2) = P\left(x_{i,d\times1}\epsilon\nabla_2|\Delta_1\right)P(\Delta_1) = P(2|1)w_1 \quad (12)$$

Equation 9 and Equation 11 give the probabilities of correct allocation, while Equation 10 and Equation 12 give the probabilities of misallocation, respectively. In other words, Equation 9 to Equation 12 summarize the confusion matrix at a glance such that Equation 9 and Equation 11 is the diagonal of the confusion matrix whereas Equation 10 and Equation 12 is the off diagonal of the confusion matrix. Due to equal probability and equal cost of misallocation assumptions (Johnson & Wichern, 1992; Johnson, 1987), it is sometimes easy to implement this procedure. Nevertheless, it is advisable to consider alternative allocation methods when this assumption fails. We shall consider the Bayes posterior probability rule for allocating an object to the desired group. The Bayes procedure (Sainin et al., 2021; Ma et al., 2011; Theodoridis & Koutroumbas, 2009) can be stated as Equation 13.

$$P\left(\Delta_i|x_{i,d\times1}\right) = \frac{P(\Delta_i)P\left(x_{i,d\times1}|i\right)}{P\left(x_{i,d\times1}\right)} = \frac{w_i g_i\left(x_{i,d\times1}\right)}{\sum_{i=1}^{2} w_i g_i\left(x_{i,d\times1}\right)} = \left(\sum_{i=1}^{2} w_i g_i\left(x_{i,d\times1}\right)\right)^{-1} w_i g_i\left(x_{i,d\times1}\right) \quad (13)$$

From Equation 13 and for the two groups' univariate allocator, we can separate to Equation 14 and Equation 15.

$$P(\Delta_1|x_{i,d\times1}) = \frac{w_1 g_1(x_{i,d\times1})}{\sum_{i=1}^{2} w_i g_i(x_{i,d\times1})} = \left(\sum_{i=1}^{2} w_i g_i(x_{i,d\times1})\right)^{-1} w_1 g_1(x_{i,d\times1}) \quad (14)$$

$$P(\Delta_2|x_{i,d\times1}) = \frac{w_2 g_2(x_{i,d\times1})}{\sum_{i=1}^{2} w_k g_k(x_{i,d\times1})} = 1 - \left(\left(\sum_{i=1}^{2} w_i g_i(x_{i,d\times1})\right)^{-1} w_1 g_1(x_{i,d\times1})\right) \quad (15)$$

Therefore, allocate x_i to Δ_1 if $P(\Delta_1|x_{i,d\times1}) > P(\Delta_2|x_{i,d\times1})$, otherwise, allocate x_i to Δ_2 if $P(\Delta_1|x_{i,d\times1}) < P(\Delta_2|x_{i,d\times1})$. This decision criteria can be compactly written as shown in Equation 16 below.

$$\pi = \begin{cases} x_i \text{ to } \Delta_1 \text{ if } P(\Delta_1|x_{i,d\times1}) > P(\Delta_2|x_{i,d\times1}) \\ x_i \text{ to } \Delta_2 \text{ if } P(\Delta_1|x_{i,d\times1}) < P(\Delta_2|x_{i,d\times1}) \end{cases} \quad (16)$$

Smart Univariate Classifier

The SUC method applied the F-weight to extract the inlier from the data matrix and detected the outliers. The outliers detected were subjected to F-weight treatment, which transformed the outliers into inliers. Subsequently, the classifier was applied to the transform data set to build the classification model and the benchmark value. This method is called smart because of its sensitivity in identifying inliers from the data sets and subsequent conversion of outliers to inliers. The formulation of the SUC method (Okwonu, Ahad, Okoloko et al., 2022) is as follows.

Let x_1 and x_2 be univariate random observations from Δ_1 and Δ_2 such that $n-2 \geq p,\ p=1$. We assume that x_1 and x_2 are normally distributed with mean and variance, which is $x_i \sim N(\mu, \sigma^2)$. Let W denote the F-weight as shown in Equation 17.

$$w_i = \frac{x_i}{Z},\ \ w_i \in W,\ i = 1,2 \quad (17)$$

where $Z = X_1 \grave{} X_2,\ x_1 \in X_1,\ x_2 \in X_2$, and w_i are the weight associated with each group. Applying Equation 17, the F-weight mean vector is given as Equation 18.

$$\bar{x}_i = \frac{\sum_{x_i=1}^{n_i} w_i x_i}{w_i} \quad (18)$$

We determine the inliers from Equation 17 and Equation 18 as shown in Equation 19.

$$I = \begin{cases} 1\ if\ \ w_i < \bar{x}_i, \bar{x}_i = \bar{x}_1, \bar{x}_2 \\ 0\ if\ \ w_i > \bar{x}_i \end{cases} \quad (19)$$

If no outlier exists, we proceed as follows by computing the sample variance s given in Equation 20.

$$S_i^2 = \frac{\sum_{i=1}^{n_i}(w_i x_i - \bar{x}_i)^2}{n_i - 1} \quad (20)$$

From the sample variance in Equation 20, we compute the pooled F-weight variance as shown in Equation 21.

$$S_{pool}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2) - 2} \quad (21)$$

Based on Equations 18 to 21, we obtain the coefficient of the model as can be referred to Equations 22 and 23.

$$w_{x_1} = \frac{(\bar{x}_1 - \bar{x}_2)`}{S_{pool}^2} w_1 x_1 = (\bar{x}_1 - \bar{x}_2)` S_{pool}^{-2}(w_1 x_1) = \omega w_1 x_1 \quad (22)$$

$$w_{x_2} = \frac{(\bar{x}_1 - \bar{x}_2)`}{S_{pool}^2} w_2 x_2 = (\bar{x}_1 - \bar{x}_2)` S_{pool}^{-2}(w_2 x_2) = \omega w_2 x_2 \quad (23)$$

On the other hand, if an outlier exists, we repeat Equations 17 and 18 on the outliers only and apply Equation 19 to detect whether the outlier still exist. This process continues until all the outliers are transformed into inliers. The inliers are merged to form the reweighted data set $\breve{w}_i$. The mean vectors, sample variance of $\breve{w}_i$, are computed similarly as Equations 18 to 20 and substituted into Equations 21 to 23, respectively.

The coefficient (ω) in Equations 22 and 23 is similar to $\left(\sum_{i=1}^{2} w_i g_i\left(x_{i,d\times 1}\right)\right)^{-1}$ in Equation 13, and $w_1 x_1$ and $w_2 x_2$ in Equations 22 and 23 is similar to $w_i g_i\left(x_{i,d\times 1}\right)$ in Equation 13. Therefore, Equations 22 and 23 mimic Equations 14 and 15. Hence, Equations 22 and 23 are simply a linear combination that allocates $w_1 x_1$ to Δ_1 or otherwise to Δ_2 if it is true. This procedure mimics the Fisher linear classification method (Sheth, 2019; Fisher, 1936). The insensitivity of the SUC method towards outliers and overfitting is due to the F-weight application to the data sets.

To evaluate the performance of the SUC method, the group evaluation criteria are obtained as in Equations 24 to 26.

$$T_{x_1} = (\bar{x}_1 - \bar{x}_2)` S_{pool}^{-2} \bar{x}_1 = \omega \bar{x}_1 \tag{24}$$

$$T_{x_2} = (\bar{x}_1 - \bar{x}_2)` S_{pool}^{-2} \bar{x}_2 = \omega \bar{x}_2 \tag{25}$$

$$T_{x_1 x_2} = \frac{T_{x_1} + T_{x_2}}{2d},\ d = 1 \tag{26}$$

Thus, Equation 26 is the benchmark value used for classification tasks. To assign an object to any of the groups correctly, the following decision criterion was adopted based on Equations 24 to 26 as presented in Equation 27.

Classify $w_1 x_1$ to Δ_1 if $w_{x_1} > T_{x_1 x_2}$, otherwise assign $w_1 x_1$ to Δ_2 if $w_{x_1} < T_{x_1 x_2}$. In a simpler form, this decision criteria can be written as in Equation 27.

$$\Delta = \begin{cases} w_{x_1} > T_{x_1 x_2} & \text{assign object to } \Delta_1 \\ w_{x_1} < T_{x_1 x_2} & \text{assign object to } \Delta_2 \end{cases} \tag{27}$$

Smart Univariate Bayes Classifier

The hybrid method (SUBC) is a combination of the SUC and BC procedures with two phases. The first phase utilises the SUC procedures (see Equations 17-23), while in the second phase, we apply the BC procedures as described in Equations 6 to 16 to build the SUBC model and assign the group membership. It implies that the proposed SUBC method mimics and retains the basic characteristic of the BC decision criteria. Therefore, the algorithm for SUBC method can be explained in the following Algorithm 1.

Algorithm 1: Algorithm for SUBC method

Phase 1 - Utilise the SUC procedure as follows

Step 1: Determine the inliers or outliers (apply Equation 19).
Step 2: If outliers exist, transform the outliers into inliers.
Step 3: Repeat Step 1 and Step 2 to obtain a reweighted F-weight.
Step 4: Merged the inliers to form the reweighted data set $\breve{w}_i$.
Step 5: Compute the mean vectors and sample variance of $\breve{w}_i$ in a similar way as in Equations 18 to 20.
Step 6: Substitute all values obtained from Step 4 into Equations 21 to 23 to obtain the model's coefficient.

Phase 2 - Build the SUBC model as follows

Step 1: Apply the BC procedures as described in Equations 6 to 15 to build the SUBC model.
Step 2: Assign the group membership using the decision criteria in Equation 16.

Data Collection

This study aims to investigate the comparative classification performance of the proposed methods (SUC and SUBC) against the conventional methods (BC and PC). In addition, we examined the breakdown of the methods based on random outliers. Finally, we also studied the relationship to determine whether the strong, moderate, or weak correlations correspond to a minimum or maximum misclassification rate. To achieve these objectives, we applied two types of data sets (continuous and discrete). Part A consists of continuous data set, while Part B consists of discrete data set.

Part A: This section consists of four data sets. The first data set was based on the effect of quality of sleep; short sleep (1-4 hours), and long sleep (5-8 hours) in relation to undergraduate academic performance with an emphasis on grade point average (GPA) categorisation using Pittsburgh Sleep Quality Index (PSQI) Scale and Perceived Stress Scale (PSS) (Lok, 2018). Then, using the methods discussed, which were BC, PC, SUC and SUBC, we applied the PSQI and PSS scales to classify students into graduate classes based on their corresponding GPA using the methods discussed. The second data set consisted of the air quality index for three locations in Malaysia (Putrajaya, Kuala Lumpur and Petaling Jaya) before the outbreak of the Covid-19 pandemic in 2019 (14/10/2019-10/11/2019) and during the same period with the Covid-19 pandemic in 2020 (14/10/2020-10/11/2020) (Department of Environment, 2020). The data set was collected with two conditions: without movement control order (2019) and with movement control order (pandemic period 2020). The third data comprised the PH water level in Ampang Pecah and Kg. Timah (Department of Environment, 2020), while the fourth data consisted of body weight measurement (mg) of wide and laboratory-bred female and male Aedes albopictus mosquitoes and body size measurements of Aedes albopictus mosquitoes (Okwonu et al., 2012).

Part B: This section covered discrete data and consisted of two data sets. The first data set in this section covered the Covid-19 daily report. The data set was paired as follows: confirmed and discharged, confirmed and dead, discharged and dead patients. We used the Covid-19 data set from Malaysia (Kementerian Kesihatan Malaysia, 2020) and Nigeria (NCDCgov, 2020). The second data set in this section comprised road traffic accidents via car and motorcycle.

These data were categorised based on severe injury, slight injury, and the annual summary of car/motorcycle road traffic accidents (Jabatan Siasatan dan Penguatkuasaan Trafik, 2020).

Therefore, the performance of the classification methods for continuous and discrete data was investigated to infer information to enable us to distinguish the performance of these methods based on classification accuracy and correlation value.

Comparative Performance Analysis

In this section, the comparative performance analysis was carried out using the four methods. The performance of the proposed methods was evaluated using the optimum probability of exact classification (OPEC). The computed probabilities of exact classifications (PEC) (Okwonu, Ahad, Ogini et al., 2022) of these methods were based on the hit ratio compared to the OPEC value. The misclassification error rate (ε) can be obtained by computing the difference between OPEC and PEC, which is $\varepsilon = OPEC - PEC$. The robustness of these methods can be inferred by the value of ε. A minimal value associated with ε implies that the methods are robust. The breakdown of these methods was investigated by introducing random outliers to the PSQI and PSS data sets. Therefore, the robustness and the breakdown capability of the methods discussed above can be determined by comparing the difference between OPEC and PEC.

The evaluation benchmark for this study was designed using the acceptable benchmark approach called the Optimum Probability of Exact Classification (OPEC) as shown in Equation 28.

$$\begin{aligned} \nabla &= \frac{(\bar{x}_1 - \bar{x}_2)}{s} \\ \delta &= \theta(\nabla) \end{aligned} \tag{28}$$

where represents the standard normal distribution function. Equation 28 is used to obtain the OPEC value, while Equation 29 is applied to compute the probability of misclassification.

$$\begin{aligned} \varepsilon &= \theta(\nabla) - PEC \\ &= \delta - PEC \end{aligned} \tag{29}$$

Equation 30 below describe the optimal misclassification rule to determine how robust the classifiers is. However, Equation 30 often lead to high misclassification rate hence we developed an alternative optimal evaluation criteria for classification performance.

$$\varepsilon A = 1 - PEC \tag{30}$$

Equation 29 is also used to detect the overfitting of these methods. If has a negative value, it implies that overfitting has occurred. Equations 28 and 29 are applied to evaluate the performance of the classification methods by focusing on the proportion of correct group membership prediction (Jimoh et al., 2022; Huberty & Holmes, 1983; Alf & Abrahams, 1968; Levy, 1967). Another method similar to Equation 29 to analyse the performance of classifiers was discussed in Mohd Noor et al. (2020). The data sets used in this paper are applied to Equation 28 to obtain the evaluation benchmark.

RESULTS AND DISCUSSION

Part A

Table 1 shows the performance analysis of the four methods using the first continuous data set. The values reported in Table 1 are the PEC values. In Table 1, the outliers are randomly introduced to the original data to determine the robustness and the breakdown of the methods. We observe that the BC and the SUBC are more robust than the other methods because the PEC values of these two methods (0.9933) are equal to the OPEC value () for the uncontaminated data set. In the second row of Table 1, we introduce ten random outliers (RO) into the original data set. Then, in the third row, we add eight outliers in addition to the ten outliers in Row 2, making it 18. Finally, we introduce outliers for the other rows in a similar procedure. As more outliers are introduced, the proposed SUC method shows outstanding performance over the other three methods since 0.8600 is the highest among the four methods.

Table 1

Performance Analysis of PSQI/PSS Data for Graduate Categories and the Effect of Outliers (n=150)

BC	PC	SUBC	SUC	Random Outliers (RO)
0.9933	0.9133	**0.9933**	0.9300	-----
0.9667	0.9067	**0.9667**	0.9233	16=1,20=2,21=41,17=57,12=32,9=19 ,5=15,2=12,1=11,3=13 (RO=10)

(continued)

BC	PC	SUBC	SUC	Random Outliers (RO)
0.9267	0.8833	**0.9267**	0.9033	17=5,30=3,17=7,20=1,4=14,11=31,5 =15,7=27 (RO=18)
0.8867	0.8633	**0.8867**	0.8833	23=3,27=2,19=1,5=25,2=22,6=26 (RO=24)
0.86	0.8533	0.86	**0.8667**	32=2,22=2,9=29,16=46 (RO=28)
0.8467	0.8467	0.8467	**0.8600**	23=3,9=69 (RO=30)

$\delta = 0.999, r = 0.283,\ r^2 = 0.08$

Figure 1 displays the breakdown of the different methods based on the PEC and the number of random outliers introduced. The performance analysis in Figure 1 shows that SUBC performs as good as BC, and it remains consistent regardless of the number of outliers in the data set. However, when the number of random outliers increases, the SUC outperforms the other methods. Thus, it can be concluded that the SUC method is more robust when the data set comprise more outliers.

Figure 1

Comparative Breakdown Analysis of the Classification Methods

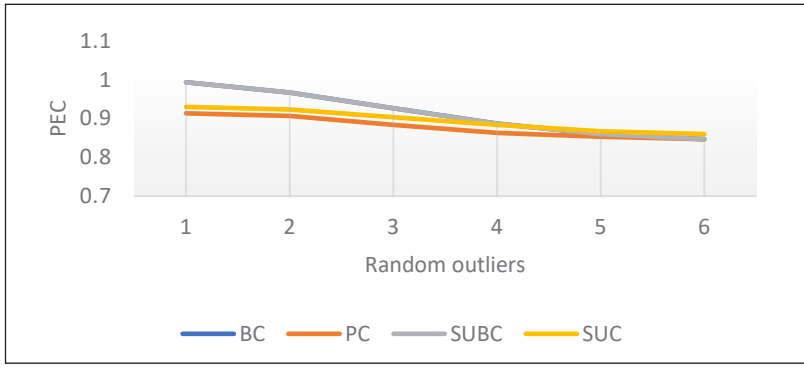


Table 2

Probability of Correct Classifications

	BC	PC	SUBC	SUC
PCC	0.9933	0.9167	0.9933	0.9300
1-PCC	0.0067	0.0833	0.0067	0.0700

Table 2 describes the probability of correct classification (PCC) for each method. The probability shows the number of students is correctly classified to their respective graduate classes. Most of the methods show that the majority of the objects belong to the actual groups, whereas minorities are misclassified to another group. From the analysis, there are misclassified individuals who fail to fit the classification of their graduate classes. Table 2 shows the observations that are correctly classified and misclassified observations based on the hit ratio. The hit ratio values for each method are discussed as follows. The BC and SUBC achieved the same classification accuracy of 99.33 percent, followed by the SUC with 93.0 percent correct classification. The PC reaches 91.67 percent of prediction accuracy. The performance analysis (see Table 1) reveals that the correlation value is positively weak. The weak correlation value indicates that the minimum misclassification rate is associated with a weak positive correlation value. It can be further explained that the exact group classification does not translate to a strong relationship of a group membership. It means that the classification performance cannot be compared with the level of association of the data set. This is a new concept of relating the classification performance with the strength of the correlation value. The rest of the analysis for the other data sets is based on the comparison between these two criteria: the proportion of correct classification and the correlation among the methods.

Table 3

Performance analysis of cumulative air quality index before (2019) and during lockdown (2020) (n=84)

BC	PC	SUBC	SUC
0.5833	0.5833	**0.61**	0.593

$\delta = 0.702,\ r = -0.1481,\ r^2 = 0.0219$

Table 3 to Table 6 contain the classification of the air quality index (AQI) before and during the Covid-19 pandemic in Malaysia. In Table 3, the SUBC method (86.9%) is the best classifier, followed by SUC (84.5%), while both BC and PC account for 83.1%. However, this data set is weakly negatively correlated (); therefore, the minimum misclassification error rates relate to a very weak and negative relationship.

Table 4

Performance Analysis of Putrajaya AQI before (2019) and during Lockdown (2020)

	BC	PC	SUBC	SUC
	0.607	0.518	**0.571**	0.518
εA	*0.393*	*0.482*	*0.429*	*0.482*
$\boldsymbol{\varepsilon = \delta - PEC}$	***-0.001***	0.088	**0.035**	0.088

$\delta =$***0.606***, $r = -0.063,\ r^2 = 0.0004$

As shown in Table 4, an overfitting problem occurs for the BC method ($\varepsilon = -0.001$). It is considered overfitting in this aspect because the OPEC (δ) is used as the performance benchmark. Nevertheless, the BC is in order by applying the basic axiom of probability. The following criterion implies that the misclassification error for BC (refer to Equation 30) is equal to

$$\varepsilon A = 1 - PEC = 1 - 0.607 = 0.393$$

Applying this criterion to the classification problem suggests that the BC method performs poorly, with a significant misclassification error rate of 39.3 percent. Therefore, the justification for applying the OPEC as an evaluation and validation benchmark is adopted and justified. The SUBC classifier correctly predicts a group membership of 94.2 percent, followed by PC and SUC with 85.5 percent, respectively, based on the OPEC criterion. The correlation value obtained is negative and very weak (-0.063). The effect of outliers is generally negative in all statistical and probability models. That is why the proposed robust classification methods are essential.

Table 5

Performance Analysis of Kuala Lumpur AQI before (2019) and during Lockdown (2020) (n=28)

BC	PC	SUBC	SUC
0.679	0.643	0.679	**0.696**

$\delta =$***0.905***, $r = -0.251,\ r^2 = 0.063$

In Table 5, the SUC classifier accurately predicts the group membership with 76.9 percent, BC and SUBC with 75 percent, and PC with 71 percent. The relationship shows a weak and negative correlation.

Table 6

Performance Analysis of Petaling Jaya AQI before (2019) and during Lockdown (2020) (n=28)

BC	PC	SUBC	SUC
0.536	0.518	0.500	0.519

$\boldsymbol{\delta}$ **=*0.539***, $r = -0.1960$, $r^2 = 0.0384$

The performance analysis in Table 6 demonstrates that the BC achieves the highest accuracy of 99.4 percent in classifying the group membership, followed by SUC (96.3%) and PC (96.1%). Meanwhile, SUBC accounts for 92.8 percent, with a weakly negative relationship.

Table 7

Performance Analysis of PH Level of Water in Ampang Pecah and Kg. Timah (Mm) (n=39)

	BC	PC	SUBC	SUC
	0.8611	0.875	0.8611	**0.8333**
ε	-0.001	-0.015	-0.001	

$\delta = 0.8599, r = 0.0793$, $r^2 = 0.0063$

The analysis shows that accurate group membership prediction does not reflect highly correlated data sets. Instead "the better the prediction power, the weaker the correlation value". This finding is relatively possible for continuous data sets. Table 7 contains the classification analysis of the PH level of water in two different locations. This was applied to observe if the PH level of water quality in these locations is unique or varies. It was verified in Table 7 that the BC, PC and SUBC methods overfit ($\varepsilon = -0.001, -0.015, -0.001$) based on the OPEC benchmark. On the other hand, SUC shows a 96.9 percent accurate prediction of a group membership. The correlation value displayed for this data is positive and very weak, 0.08.

Table 8

Performance Analysis of Wide Female and Male Aedes Albopictus Mosquitoes' Weight (mm) (n=96)

BC	PC	SUBC	SUC
0.4688	0.4896	**0.5100**	0.4896

δ=0.536, $r = 0.218$, $r^2 = 0.047$

Based on Table 8, the proposed SUBC achieves the highest prediction accuracy with 95.2 percent, followed by PC and SUC with 91.3 percent prediction accuracy, respectively. Meanwhile, the BC records the lowest prediction accuracy at 87.5 percent. The relationship for the Aedes Albopictus mosquito data is also positively weak.

Table 9

Performance Analysis of Laboratory-reared Female and Male Aedes Albopictus Mosquitoes Body Size (mm) (n=30)

	BC	PC	SUBC	SUC
	1.00	**0.950**	1.00	**0.950**
ε	-0.002		-0.002	

δ =0.998, $r = -0.095$, $r^2 = 0.009$

In Table 9, the BC and SUBC methods predict the group membership with 100% accuracy. However, both methods overfitted ($\varepsilon = -0.002$). Meanwhile, the PC and SUC methods achieve 95.2% of the group membership prediction. Therefore, the association shown has a weak negative correlation.

Table 10

Performance Analysis of Wide and Laboratory-reared Female and Male Aedes Albopictus Mosquitoes Wing Length (mm) (n=20)

	BC	PC	SUBC	SUC
	1.00	**0.925**	1.00	0.90
ε	-0.007		-0.007	

=0.993, $r = -0.038$, $r^2 = 0.001$

Similarly, the data set in Table 10 demonstrates that the BC and SUBC methods obtain 100 percent accuracy in predicting the group membership, and there is also overfitting in both methods ($\varepsilon = -0.007$). The accurate group predictions for the PC and SUC methods are 93.2 percent and 90.6 percent, respectively. Even though the BC and SUBC show overfitting in Table 9 and Table 10, the methods can be accepted because the OPEC value is approximately one. Therefore, the correlation for this data set is weak and negatively correlated.

Part B

The results in this section demonstrate the performance analysis of the four methods on the two discrete data related to Covid-19 data sets. Table 11 to Table 13 demonstrate the performance analysis based on the Malaysian Covid-19 data sets for 189 days, while Tables 14 to 16 show the performance analysis of the Covid-19 data sets for 163 days in Nigeria.

Table 11

Performance Analysis of Confirmed and Discharged Cases for Covid-19 Virus, Malaysia data (n=189)

BC	PC	SUBC	SUC
0.466	0.471	0.466	**0.492**

δ=0.512, $r = 0.379$, $r^2 = 0.144$

As shown in Table 11, the SUC predicts 96.1 percent correct group membership, and this is followed by the PC method with 92 percent accuracy, while both BC and SUBC achieve 91 percent correct classification. The correlation value is positive and relatively weak ($r = 0.379$).

Table 12

Performance analysis of confirmed and death cases for Covid-19 virus, Malaysia data (n=189)

	BC	PC	SUBC	SUC
	0.889	**0.725**	0.889	0.646
ε	-0.03		-0.03	

δ =0.864, $r = 0.659$, $r^2 = 0.434$

Table 12 shows that the BC and SUBC methods are overfitted (ε = -0.03), while both PC and SUC methods predict the group membership correctly at 83.9 percent and 74.8 percent, respectively. Dissimilar to other data sets, the correlation of this data is positive and moderately strong (r = 0.659). This makes sense as it reveals a considerably strong relationship between confirmed and death cases, implying that the number of deaths would also increase if the confirmed cases increases.

Table 13

Performance Analysis of Discharged and Death Cases for Covid-19 Virus, Malaysia Data (n=189)

BC	PC	SUBC	SUC
0.825	0.741	**0.825**	0.685

δ= 0.867, $r = 0.305$, $r^2 = 0.093$

Table 13 also reveals that both the BC and SUBC methods achieve the highest correct predictions at 95.2 percent, while PC and SUC correctly predict the group memberships with 85.5 percent and 79 percent accuracy, respectively. This data set shows a weak positive relationship. The implication is that the more people are discharged, the lower the death rate. Hence the weak correlation value is justified.

Table 14

Performance Analysis of Confirmed and Discharged Cases for Covid-19 Virus, Nigeria Data (n=163)

	BC	PC	SUBC	SUC
	0.798	0.647	0.798	**0.497**
ε	-0.244	-0.099	-0.244	

δ= 0.548, $r = 0.153$, $r^2 = 0.023$

Table 14 reveals that BC, PC, and SUBC methods are overfitted (ε = -0.244, -0.099, -0.244), based on the OPEC benchmark values. In contrast, the SUC method attains 90.7 percent accuracy in predicting the group membership, with a very weak positive correlation for this data set.

Table 15

Performance Analysis of Confirmed and Death Cases for Covid-19 Virus, Nigeria Data (n=163)

BC	PC	SUBC	SUC
0.834	0.804	**0.834**	0.733

δ =0.939, $r = 0.65$, $r^2 = 0.423$

Table 15 illustrates that the BC and SUBC can predict the group membership correctly with 88.8 percent, while the PC and SUC have 85.6 percent and 78.1 percent accuracy, respectively. Similar to the Malaysian data set, a moderately strong correlation can also be found in this Nigeria data set for confirmed and death cases.

Table 16

Performance Analysis of Discharged and Death Cases for Covid-19 Virus, Nigeria Data (n=163)

	BC	PC	SUBC	SUC
	0.767	0.73	0.761	**0.509**
ε	-0.145	-0.108	-0.139	

δ =0.622, $r = 0.198$, $r^2 = 0.039$

Again, the data set in Table 16 also reveals that BC, PC and SUBC methods are overfitted (ε = -0.145, -0.108, -0.139), while the SUC method predicts 81.8 percent group membership, and the data show an extremely weak and positive correlation.

Tables 17 to 19 display the annual road traffic accident data categorised based on the degrees of fatalities. Based on the results in Table 17 to Table 19, the probability-based methods (BC and SUBC) overfit the OPEC value. However, the PC method outperforms the SUC method with a moderate to very strong relationship. Based on the adopted benchmark value, we may conclude that these two methods (PC and SUC) provide more accurate predictions for classification than the probability methods (BC and SUBC). The Pearson correlation of this data set is very strong and positively associated, mainly for data sets in Tables 17 and 19.

Table 17

Performance Analysis of Car and Motorbike Road Traffic Accident (n=30)

	BC	PC	SUBC	SUC
	1.00	**0.83**	1.00	0.82
ε	-0.022		-0.022	

δ =0.978, $r = 0.946$, $r^2 = 0.895$

Table 17 reveals that the PC and SUC achieve 85.3 percent and 83.4 percent correct group prediction, with a very strong positive correlation value of 0.946. Meanwhile, the probability-based methods (BC and SUBC) in Table 18 show overfitted values (ε = -0.007), in contrast to the PC and SUC methods which accurately predict the group membership at 86.6 percent and 85.6 percent, respectively, with a moderately positive correlation of 0.613 for this data set.

Table 18

Performance Analysis of Car Severe and Motorbike Severe Road Traffic Accident Injuries (n=25)

	BC	PC	SUBC	SUC
	1.00	**0.86**	1.00	0.84
ε	-0.007		-0.007	

δ =0.993, $r = 0.613$, $r^2 = 0.376$

Table 19

Performance Analysis of Car Slight and Motorbike Slight Road Traffic Accident Injuries (n=25)

	BC	PC	SUBC	SUC
	1.00	**0.76**	1.00	0.70
ε	-0.117		-0.117	

δ =0.883, $r = 0.968$, $r^2 = 0.937$

Similar to the results in Table 19, the probability-based classifiers are overfitted, while the classification using PC and SUC methods attain 86.1 percent and 79.3 percent accuracy in predicting the group membership. This data set reveals a very strong positive correlation which is 0.968.

Table 20 summarises the best-performed method for the continuous data sets based on the minimum misclassification rate (ε). The results show that the proposed methods outperformed the conventional methods. Table 20 also includes the comparative analysis of the performance of different methods with respect to the Pearson correlation $(r)$and the coefficient of determination(r^2)values. Based on the results in this table, we may conclude that a weak positive

or a very weak negative correlation is associated with the minimum misclassification rate for a continuous data set. We also observe that very small r^2 values correspond to the minimum misclassification rate. The results of the data sets in Table 20 are depicted in Figure 2.

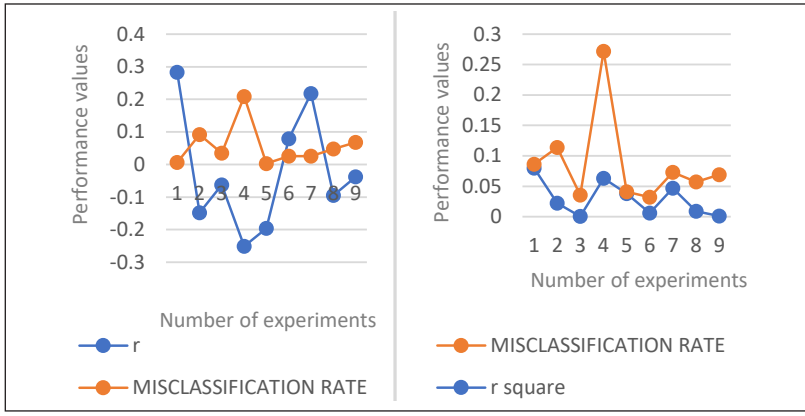
Table 20

Comparative Analysis of the Best Methods for Continuous Data

r	r^2	PEC	$OPEC(\delta)$	$\varepsilon = OPEC - PEC$	Best Methods
0.283	0.080	0.993	0.999	0.006	BC & SUBC
-0.148	0.022	0.61	0.702	0.092	SUBC
-0.063	0.0004	0.571	0.606	0.035	SUBC
-0.251	0.063	0.696	0.905	0.209	SUC
-0.196	0.038	0.536	0.539	0.003	BC
0.079	0.006	0.833	0.859	0.026	SUC
0.218	0.047	0.51	0.536	0.026	SUBC
-0.095	0.009	0.95	0.998	0.048	PC & SUC
-0.038	0.001	0.925	0.993	0.068	PC

Figure 2

Comparative Analysis of Continuous Data



The comparative analysis in Table 21 shows that the proposed methods are more robust than the conventional methods based on the misclassification error for discrete data cases. Furthermore, the analysis reveals that a weak to a strong positive correlation (r) are associated with the minimum misclassification rate. Moreover, the r^2 shows very small to very large values associated with a relatively minimum misclassification rate. These results are illustrated in Figure 3.

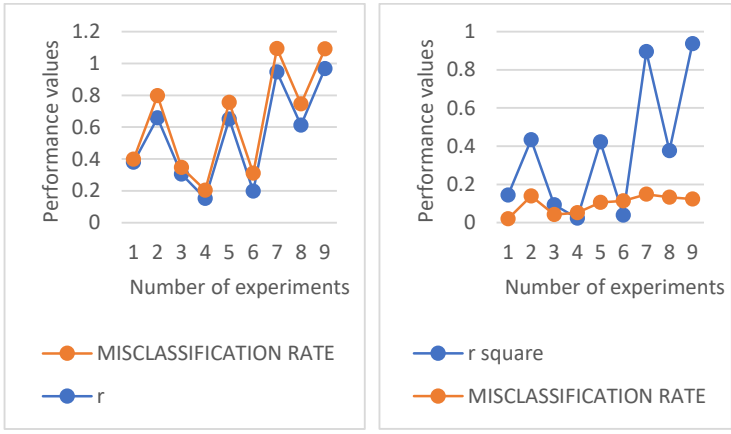
Table 21

Comparative Analysis of the Best Methods for Discrete Data

r	r^2	PEC	$OPEC(\delta)$	$\varepsilon = OPEC - PEC$	Best Methods
0.379	0.144	0.492	0.512	0.02	SUC
0.659	0.434	0.725	0.864	0.139	PC
0.305	0.093	0.825	0.867	0.042	BC & SUBC
0.153	0.023	0.497	0.548	0.051	SUC
0.65	0.423	0.834	0.939	0.105	BC & SUBC
0.198	0.039	0.509	0.622	0.113	SUC
0.946	0.895	0.83	0.978	0.148	PC
0.613	0.376	0.86	0.993	0.133	PC
0.968	0.937	0.76	0.883	0.123	PC

Figure 3

Comparative Analysis of Discrete Data



(a) (b)

DISCUSSION

Based on the comparative performance analysis, we observed that different methods performed differently based on the characteristics of the data sets. We had shown that the SUC method did not overfit the data when the OPEC benchmark evaluation was applied. We have also revealed that all the methods performed differently depending on the data types, as summarised in Tables 20 and 21. This study has shown that for continuous data, as reported in Table 20, a relatively minimum misclassification rate could be associated with very weak negative and weak positive correlation values. This analysis was also applicable to the r^2 Meanwhile, for a discrete data set (Table 21), it indicated that a weak to a strong positive correlation is related to a minimum misclassification rate. A similar performance analysis is portrayed by r^2 The results showed overfitting for the discrete data sets was more than for the continuous ones. Thus, it can be inferred that the proposed SUC and SUBC methods performed better than the conventional methods (BC and PC) for both types of data sets. Based on the comparative analysis, the proposed methods were found to be more robust than the conventional methods.

The proposed method was able to penalise the outlier based on the following. The proposed methods transformed the outliers to inliers using the F-weight, and if any outliers still existed, we reweighted the F-weight until the outliers were transformed into inliers. This study also showed that the proposed SUC method solved the overfitting problem associated with the BC method when OPEC was used as a performance benchmark. We may technically generalise that using continuous data for classification problems yielded a weak negative and weak positive correlation. Meanwhile, the discrete data used for classification yielded a relatively weak to a strong positive correlation. The strength and the direction of the correlation values do not imply robust or poor classification performance of the methods. This suggested that the classification performance is independent of the correlation values.

CONCLUSION

The analysis reveals that the performance of the investigated methods is data-dependent. The results show that the proposed Smart Univariate

Classifier (SUC) is robust and has solved the problem of overfitting associated with the conventional methods when OPEC is used as a performance benchmark. The proposed SUBC method is robust but still affected by the overfitting problem. In addition, both proposed methods achieve high classification accuracy. The performance of the proposed methods based on the introduction of random outliers indicates that they are more robust and capable of resisting influential observations than conventional methods. The findings also show that a minimum misclassification error is independent of the strength of the correlation values. Conclusively, the proposed methods are robust and capable of overcoming the overfitting problem, which often occurs in conventional methods.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Aborisade, O. M., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. *IEEE International Conference on Information Reuse and Integration (IRI)*. https://doi.org/ 10.1109/IRI.2018.00049.

Alf, E., & Abrahams, N. M. (1968). Relationship between per cent overlap and measures of correlation. *Educational and Psychological Measurement*, *28*(3). https://doi.org/ 10.1177/001316446802800307.

Banik, D., Ekbal, A., Bhattacharyya, P., Bhattacharyya, S., & Platos, J. (2019). Statistical-based system combination approach to gain advantages over different machine translation systems. *Heliyon*, *5(9)*. https://doi.org/ 10.1016/j.heliyon.2019.e02504.

Barbini, E., Manzi, P., & Barbini, P. (2013). Bayesian approach in medicine and health management. *Current Topics in Public Health*.

Berry, D. A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science, 19*(1). https://doi.org/10.1214/088342304000000044.

Bharadwaj, P., & Shao, Z. (2019). Fake news detection with semantic features and text mining. *International Journal on Natural Language Computing, 8*(3). https://doi: 10.5121/ijnlc.2019.8302.

Blanca, M. J., Alarcon, R., Arnua, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*(4), 552-557. https://dx.doi.org/ 10.7334/psicothema2016.383.

Cain, M. K., Zhang, Z., & Yuan, K. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: prevalence, influence and estimation. *Behav Res, 49*, 1716-1735. https://dx.doi.org/0.3758/s13428-016-0814-1.

Chattopadhyay, S., Davis, R. M., Menezes, D. D., Singh, G., Acharya, R. U., & Tamura, T. (2012). Application of bayesian classifier for the diagnosis of dental pain. *Journal of Medical Systems*, *36*(3). https://doi.org/10.1007/s10916-010-9604-y.

Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, *5*(1), 5-12. https://doi.org/ 10.11648/j.ajtas.20160501.12.

Dávideková, M., Michal Greguš, M. L., & Bureš, V. (2019). Yet another classification of ICT in knowledge management initiatives: Synchronicity and interaction perspective. *Journal of Engineering and Applied Sciences, 14*(Special Issue 9), pp. 10549-10554.

Department of Environment (Dec. 2020). Official Portal of Department of Environment. https:// www.doe.gov.my/portalv1/en/

Donoho, D., & Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences of the United States of America, 105*(39). https://doi.org/ 10.1073/pnas.0807471105.

El Abbassi, M., Overbeck, J., Braun, O., Calame, M., van der Zant, H. S., & Perrin, M. L. (2021). Benchmark and application of unsupervised classification approaches for univariate data. *Communications Physics*, *4*(1), pp. 1-9.

Field, A. (2009). *Discovering Statistics using SPSS* (3rd ed). London: SAGE Publications Ltd, p. 822.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2). https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Hamid, H., Ngu, P. A. H., & Alipiah, F. M. (2018). New smoothed location models integrated with PCA and two types of MCA for handling large number of mixed continuous and binary variables. *Pertanika Journal of Science & Technology, 26*(1), 247-260.

Hamid H., Zainon, F., & Yong T. P. (2016). Performance analysis: An integration of principal component analysis and linear discriminant analysis for a very large number of measured variables. *Research Journal of Applied Sciences, 11*(11), 1422-1426.

Huberty, C. J., & Holmes, S. E. (1983). Two-group comparisons and univariate classification. *Educational and Psychological Measurement, 43*(1). https://doi.org/10.1177/001316448304300103.

Jabatan Siatan dan Penguatkuasaan Trafik, PDRM (2020).

Jimoh, R. G., Abisoye, O. A., & Uthman, M. M. B. (2022). Ensemble feed-forward neural network and support vector machine for prediction of multiclass malaria infection. *Journal of Information and Communication Technology, 21*(1), 117-148. https://doi.org/10.32890/jict2022.21.1.6.

Johnson, R. A., & Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis* (3rd ed). New Jersey: Prentice-Hall, Inc, Englewood Cliffs.

Johnson, W. (1987). The detection of influential observations for allocation, separation, and the determination of probabilities in a bayesian framework. *Journal of Business and Economic Statistics, 5*(3). https://doi.org/10.1080/07350015.1987.10509601.

Kementerian Kesihatan Malaysia. (2020, November 10). *Covid-19 Malaysia*. http://covid-19.moh.gov.my/.

Levy, P. (1967). Substantive significance of significant differences between two groups. *Psychological Bulletin, 67*(1). https://doi.org/10.1037/h0020415.

Lok, P. Y. (2018). Sleep quality among undergraduates during pre-examination period in Universiti Utara Malaysia (Unpublished final year project). Universiti Utara Malaysia, Malaysia.

Ma, D., Wei, W., Hu, H., & Guan, J. (2011). The application of bayesian classification theories in distance education system. *International Journal of Modern Education and Computer Science, 3*(4). https://doi.org/10.5815/ijmecs.2011.04.02.

Mohd Noor, N. K., Mohd Noah, S. A., & Ab Aziz, M. J. (2020). Classification of short possessive clitic pronoun nya in Malay text to support anaphor candidate determination. *Journal of Information and Communication Technology, 19*(4), 513-532. https://doi.org/10.32890/jict2020.19.4.3.

NCDC (@NCDCgov) / Twitter (2020, June 19). https://twitter.com/ncdcgov.

Okwonu, F. Z., Ahad, N. A., Ogini, N. O., Okoloko, I. E., & Wan Husin, W. Z. (2022). Comparative performance evaluation of efficiency for high dimensional classification methods. *Journal of Information and Communication Technology*, *21*(3), 437-464. https://doi.org/10.32890/jict2022.21.3.6

Okwonu, F. Z., Ahad, N. A., Okoloko, I. E., Apanapudor, J. S., Kamaruddin, S. A., & Arunaye F. I. (2022). Robust Hybrid Classification Methods and Applications. *Pertanika Journal of Science and Technology*, *30*(4), https://doi.org/10.47836/pjst.30.4.29.

Okwonu, F. Z., Dieng, H., Othman, A. R., & Ooi, S. H. (2012). Classification of aedes adults mosquitoes in two distinct groups based on fisher linear discriminant analysis and FZOARO techniques. *Mathematical Theory and Modeling, 2*(6), pp. 22-30.

Pang, Y. S., Ahad, N. A, Syed Yahaya, S. S., & Lim, Y. F. (2019). Robust linear discriminant rule using novel distance-based trimming procedure. *Journal of Advance Research in Dynamical & Control Systems, 11*(05-Special Issue), pp. 969-978.

Sainin, M. S., & Alfred, R., & Ahmad, F. (2021). Ensemble meta classifier with sampling and feature selection for data with imbalance multiclass problem. *Journal of Information and Communication Technology, 20*(2), 103-133. https://doi.org/10.32890/jict2021.20.2.1.

Sheth, M., Gerovitch, A., Welsch, R., & Markuzon, N. (2019). The univariate flagging algorithm (UFA): An interpretable approach for predictive modeling. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0223161.

Song, K., Wang, N., & Wang, H. (2020). A metric learning-based univariate time series classification method. *Information (Switzerland), 11*(6), https://doi.org/ 10.3390/INFO11060288.

Sun, Y., Li, J., Liu, J., Sun, B., & Chow, C. (2014). An improvement of symbolic aggregate approximation distance measure for time series. *Neurocomputing, 138*. https://doi.org/10.1016/j.neucom.2014.01.045.

Taheri, S., & Mammadov, M. (2013). Learning the naive bayes classifier with optimisation models. *International Journal of Applied Mathematics and Computer Science, 23*(4), https://doi.org/10.2478/amcs-2013-0059.

Thet Zan, C., & Yamana, H. (2016). An improved symbolic aggregate approximation distance measure based on its statistical features. *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services,* 72-80. https://doi.org/ 10.1145/3011141.3011146.

Theodoridis, S., & Koutroumbas, K. (2009). Classifiers based on bayes decision theory. *Pattern Recognition*, 13-89.

Trovato, G., Chrupala, G., & Takanishi, A. (2016). Application of the naive bayes classifier for representation and use of heterogeneous and incomplete knowledge in social robotics. *Robotics, 5*(1). https://doi.org/ 10.3390/robotics5010006.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *The Annals of Mathematical Statistics, 15*(2), https://doi.org/ 10.1214/aoms/1177731280.

Wei, Q. (2018). Understanding of the naive Bayes classifier in spam filtering. *AIP Conference Proceedings*, *1967*. https://doi.org/10.1063/1.5038979.