Machine learning algorithm, scaling technique and the accuracy: an application to educational data.

WICKRAMASINGHE, I. and KALUTARAGE, H.

2024

Machine Learning Algorithm, Scaling Technique and the Accuracy: An Application to Educational Data

Indika Wickramasinghe
Department of Mathematics
Prairie View A&M University
Prairie View, TX, USA
ipwickramasinghe@pvamu.edu

Harsha Kalutarage
School of Computing
Robert Gordon University
Aberdeen, UK
h.kalutarage@rgu.ac.uk

***Abstract*—Machine learning (ML) applications in educational data mining have become an increasingly popular research area. Literature indicates a lack of research investigating the impact of data scaling techniques, ML algorithms, and the nature of data on the classification's accuracy. This study aims to fulfill the above. In that direction, we use three linear and three non-linear ML classifiers and six scaling techniques to evaluate the impact of the data scaling technique and the ML algorithm on four selected educational datasets. According to the experimental outcomes for data set #1, classification accuracy was significantly influenced (p-value < 0.01) by the nature of the data. All the performance indicators except detection rate and prevalence were highly influenced by the type of ML technique used for the classification. Furthermore, there was a significant (p-value < 0.05) interaction impact of two-way interactions of the nature of the data and the type of ML technique for F1 value and balanced accuracy. Further analysis indicates that the classification accuracy varies with the level of the class variable.**

Keywords—machine learning, educational data mining, educational data analysis, data normalizing, scaling techniques

I. INTRODUCTION

The development of technology, the popularity of learning management systems, and fast-growing online education generate a plethora of educational data. This makes the educational researcher's life easy due to the access to various forms of educational data. Educational data analysis has benefited from data science, enabling us to explore the relationships among the data that were not possible with the existing conventional statistical procedures.

The role of educational data mining is to make educational decisions using extensive data repositories and technology [1]. Traditional educational data analysis primarily relied on conventional statistical procedures. In tandem with the availability of large volumes of data, a more comprehensive selection of methods emerged by challenging the traditional statistical procedures. A section of artificial intelligence called Machine Learning (ML) comprises a collection of computer programs providing systems that can automate learning and progress with the experience. The application of ML techniques to analyze educational data has become a trendy research area with the development of technology.

One of the aims of any data classification application is to improve classification accuracy. With the availability of a plethora of information, the use of ML has increased dramatically in education. At times, the use of ML in the educational domain for the data classification task has become cumbersome as the prediction accuracy depends on various factors [2], including the type of classification algorithm used, the nature of the data [3, 4, 5], preprocessing techniques and more [6, 7].

Due to the large volume of data, there are irrelevant inconsistencies, an obstacle to accurately analyzing educational data [8]. Therefore, the data should be validated and checked for consistency before the analysis. This step is considered data preprocessing when using ML to analyze data. Data scaling is considered one of the main tasks at the data preprocessing stage. In the ML and data mining community, the two terminologies, data scaling, and normalization, are interchangeably used and refer to the same data preprocessing procedure [9]. Literature reveals that no attention has been given to studying the impact of data preprocessing on the accuracy of data educational data classification.

This study investigates the impact of ML and data scaling techniques on the accuracy of educational data classification. The importance of the balanced nature of educational data toward classification accuracy is also studied. The findings of this study will help to fill the vacuum of investigations of this type in data mining for educational data.

II. Literature Findings

A comparative study was conducted by Ramaswami [10] to investigate the impact of selecting the most relevant features in educational data mining. Six filtered feature selection techniques, together with the Naive Bayes algorithm, have been used in this study. The study's findings show the increment of higher prediction accuracy with the availability of a minimum number of features and reduced computational time.

The study conducted by Lile [11] aims to identify and locate areas that need improvement in the learning management system to perform the system smoothly. The researcher proposes a model for this task using several ML techniques. According to the findings, the proposed model has shown promising outcomes for obtaining comprehensible, actionable feedback about students' learning patterns. Another study [12] utilizes several ML algorithms, including regression, to predict students' GPA accurately. According to the experimental outcomes, the

regression approach has the highest prediction accuracy compared to the other counterparts.

A study by Anjewierden [13] shows the application of chat data classification to improve learning environments using the Naive Bayes technique. The results indicate that analyzing chat messages to understand learner behavior is a reality, although several practical issues remain. Another study conducted by Pal [14] focuses on predicting dropout students. Using the Naive Bayes classifier and student-related features, the proposed model predicts the student dropout and identifies the students who need special attention to mitigate the dropout rate.

A completely different aspect of educational data analysis using Ml techniques is discussed by Hämäläinen and Vinni [15]. The study discusses one issue related to educational data analysis using ML techniques. This is the lack of the data set for training ML models. Furthermore, the paper provides general outlines about classifying numerical and categorical educational data of small size.

As educators, classifying the types of exam questions that students are given is essential when preparing for exams. Fei [16] describes a question classification algorithm that can be used in an e-learning system. The importance of the artificial neural network approach is that the algorithm can handle tests in the form of multiple-choice exams, fill-in-the-blank, or short-answer tests.

There is a vacuum of research investigations related to the impact of the nature of the data and scaling techniques on classification accuracy. As discussed below, only a few such studies are available in the literature.

Using more robust ML algorithms, namely, XGB, LR, DT, RF, and scaling techniques such as Standard Scaler, MinMax Scaler, Max Abs Scaler, Robust Scaler, and Quantile Transformer, Balabaeva [17] researches to study the impact of scaling on classification. According to their findings, RF achieves the highest performance with the Standard Scaler, and DT unchanged the performance with the scaling. An extension of similar work is carried out by [18], in which the authors consider eleven ML algorithms and six scaling techniques to identify the effect of the scaling technique on the accuracy of the ML algorithm.

Another study was conducted by Pan et al. [19] to investigate the impact of the scaling technique when using SVM to forecast the price movement of the stock index. According to the experimental outcomes, the authors suggest paying attention to selecting suitable data scaling techniques to refrain from having a negative influence on prediction accuracy. Furthermore, they argue that the failure to do so can impact the processing time of training.

Cao et al. [9] present the Generalized Logistic (GL) algorithm to scale data into intervals of (0,1) using the cumulative density function of each variable in the data set. This simple yet effective algorithm is used in diagnostic and classification modeling in medical applications. Some of the critical features of this GL algorithm are the simplicity, efficiency, and robustness of outliers. According to the experimental findings, the GL algorithm outperforms two of the most frequently used scale algorithms: the Min-max and Z-score algorithms.

According to [13], ML models trained using scaled data perform significantly better than those trained using unscaled data. Furthermore, scaling is critical for models based on distance measures such as clustering and nearest-neighbor classification. According to [18], it is required to use scaled data to have a stable and faster learning process when using artificial Neural Networks. Though directly related studies to data scaling in the educational domain are hard to find, some of the investigations of educational data analysis and data mining can be seen in [20], [21], [22], [10], and [23].

The type of ML classifier plays a significant role when classifying data [24]. One cannot know which algorithm will perform best on their data set, and finding the best algorithm for a given problem requires a systematic spot-checking of different algorithms. Literature indicates the use and comparison of various ML classifiers in different domains. A study compares the performances of supervised ML algorithms and determines the effective classifier in [25]. In another work, ML algorithms are compared over data from communications network traffic domain [26].

III. Methodology

This section describes the six scaling techniques and the ML techniques used in this study. Let's represent N1 as the non-scaled data. Furthermore, let's use N2, N3, N4, N5, N6, and N7 to represent used scaling techniques Minmax, Mean Centered, MAD, Median Normalize, Z-score, and Sigmoid.

A. Classifiers

Data classifiers can be broadly divided into linear and non-linear classifiers. Linear classifiers are the classification mechanisms to categorize data points into respective classes using a linear combination of the explanatory variables. In other words, these techniques classify data that can be done linearly. On the other hand, non-linear classification techniques attempt to classify data that cannot be separated linearly. In this study, we select three linear classifiers, namely, Naive Bayes [27], Logistic Regression, and SVM (linear kernel), and three non-linear classifiers, namely Kth -Nearest Neighbour, Random Forest, and SVM (non-linear kernel).

- Support Vector Machines (SVM)

SVM is regarded as one of the most widely used classification techniques [28] introduced by Cortes and Vapnik [29]. This technique is based on statistical learning theory and it classifies data into two types of classes (Class A and Class B) by creating a hyper-plane in terms of maximizing the boundary of the separation. Consider the n-tuple training data set,

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \quad (1)$$

Suppose the set of feature vector is M-dimensional and the class vector is 2-dimensional. i.e., $x_i \epsilon R^M$ and $y_i \epsilon \{-1, +1\}$. If the feature space can be separated linearly, then the optimal hyperplane can be constructed by solving the following

$$min \ \| w \|^2 \ s.t. \ y_i \ (x_i . w + b) \geq 1, i = 1, 2, \cdots, n \quad (2)$$

- Naive Bayes (NB)

Consider a n-dimensional vector of random variables $X = (X_1, X_2, \ldots, X_n)$ representing the features from a domain $D(x)$. Let's take a set of instances and $x_1, x_2, , .., x_n$ from X. Furthermore, assume that Y be an unobserved random variable from binary domain $D(y) = 0, 1$. Here the aim is to select the appropriate class $Y = y$ for a given $X = x$ maximizing the posterior probability, $P(Y = y|X = x)$.

- Multinomial Logistic Regression

Let X be the vector of features and $Y = \{0, 1\}$ be the class variable. Also, let $p(x) = P(Y = 1|X = x)$ and $1 - p(x) = P(Y = 0|X = x)$. The logistic regression model is defined as follows.

$$log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_{p-1} x_{p-1} \quad (3)$$

Parameters $\beta = (\beta_0, \beta_1, \ldots \beta_{p-1})$ are estimated by maximizing the following Likelihood function.

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i|X = x_i) \quad (4)$$

- K-nearest Neighbors Algorithm (k-NN)

Though k-NN is considered as one of the simplest ML techniques, it is still used widely. As the name suggests, the algorithm uses the k-nearest neighbors to classify unknown object. In this process, distances from the new objects to all the existing objects are calculated. Then the nearest k objects are selected and the assignment of the new object to the appropriate class is done accordingly. Let $X \in \Re^n$ and $Y \in \Re^n$ be coordinates of two objects. The distance between the objects is calculated using the following formula.

$$d(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (5)$$

- Random Forest (RF)

RF algorithm broaden the idea of decision trees with the aggregating higher number of decision trees to minimize the variance of the novel decision tree [30]. Each of the tree is constructed upon a group of random variables (features) and a group of such random trees is named a Random Forest. Due to the increased classification accuracy shown by this algorithm, RF is considered as one of the most successful classification algorithms in modern-times [31, 32].

B. Performance Indicators

In this study, we use the following performance indicators to classify the accuracy given by each of the used ML classifiers. These performance indicators are based on the four values (TP-True Positive, FP-False Positive, TN-True Negative, and FN-False Negative) in the confusion matrix. Finding a uniformly best performance indicator across any data set is impossible, but depending on the application, one can decide the required performance indicator from the list of Accuracy, Balanced Accuracy, Detection Rate (D. Rate), Detection Prevalence (DP), F1, Negative Prediction Value (NPV), Positive Prediction Value (PPV), Precision, Prevalence, Recall, and Specificity.

IV. DATASETS

Four educational data sets for this study were selected, representing the areas of student performance, the impact of COVID-19 on students' education, the rating of a course, and student's adaptability level for online courses. Four data sets are briefly described below. In some instances, certain features were not considered from the original data due to in-applicability.

- Data-set #1: Student Performance

This data set shows the impact of Gender, Race, and Education Level of Parents on Student Performance in Exams. A complete data set is available [33]. The data set contains features of students such as their gender, ethnicity, parental level of education, and some information about their lunch and test scores. Here, the class variable represents the grades (A, B, C, D, and F).

- Data-set #2: Impact of COVID on Students

This data set [34] represents the outcome of a survey about students' experience with an online course during the COVID-19 pandemic. The class variable is an online experience of the class (Excellent, Good, Average, Poor, Very Poor). This data set comprises students' information, including region, age, time spent online classes, and time spent on various activities (self-study, fitness, sleep, and social media). Furthermore, it includes the number of meals per day, weight change, and information such as their stress and what they missed during the pandemic.

- Data set #3-Rating of Finance Courses

This data set contains students' ratings of several development courses in Finance and related fields. Complete information about this data can be found in [35]. The features represent information about the course, payment for the course, the number of students registered, and the course ratings. In this data set, the class variable is the ratings were categorized into three (Poor, if the rating is less than 2.00; okay, if the rating is between 2.00 and 3.50; and Good: if the rating is over 3.50).

- Data set #4-Students Adaptability Level

This data represents the levels of students for online education. The original data for this study. Attributes include students' gender, educational background, available internet quality, class time, and the family's economic condition. The class variable in this data set is adaptability (high, moderate, low). More information about this data can be found in [36].

Table I summarizes the characteristics of each data set using the following indicators. The number of instances, number of features, each class type (number of levels in each class), whether each class has an equal number of instances or not (balanced or unbalanced), number of continuous and categorical features, maximum and minimum correlation, sparsity, and the Pseudo. Here, sparsity measures the proportion of zeros of a matrix. For instance, .95 sparsity shows that 95% of the values are zero. The coefficient of determination value is used in linear regression to measure the goodness of the fit. When the response variable is a nominal or ordinal data type, the regular is not applicable, and the Pseudo-R squared instead. A summary of the above data sets can be seen in Table I.

V. RESULTS

Six ML techniques were applied to each data set to discover the impact of the ML technique, scaling technique, and classification accuracy. Measurements of eleven performance indicators were recorded. Figures 1, 2 and 3 show the recorded sensitivity, specificity, and F1 value for each of the six ML techniques with all four data sets. Figures 4 shows the relationship among balanced accuracy, ML technique, and scaling technique for data set 1.

TABLE I. SUMMARY OF THE DATA SETS

Description	Dataset #1	Dataset #2	Dataset #3	Dataset #4
Name	Students Pfm	Class Eval	Course Rating	Course Adaptability
# instances	1000	1182	13608	1205
# features	5	15	5	11
Class Type	5 classes	5 classes	3 classes	3 classes
Balance/ Unbalance	unbalance	unbalance	unbalance	unbalance
# cont features	0	8	4	0
# cat feature	4	7	1	11
Max Corr	0.047	0.247	0.784	0.505
Min Corr	-0.064	-0.219	-0.073	-0.209
Sparsity	0.249	0.208	0.189	0.219
Psedo R^2 (McFadden, Nagelkerke)	(0.083, 0.232)	(0.1514, 0.3728)	(0.122, 0.185)	(0.166, 0.312)

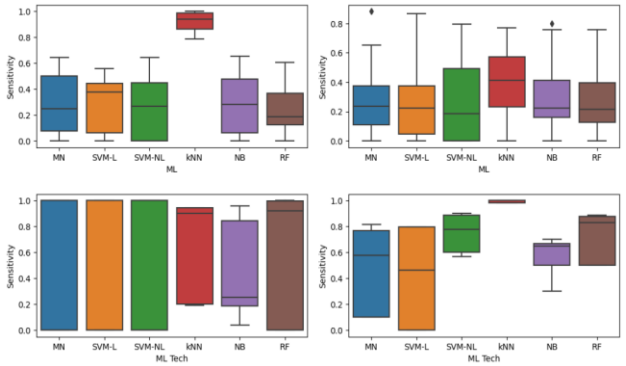

Fig. 1. Sensitivity, ML technique, and Data Sets (1-4).

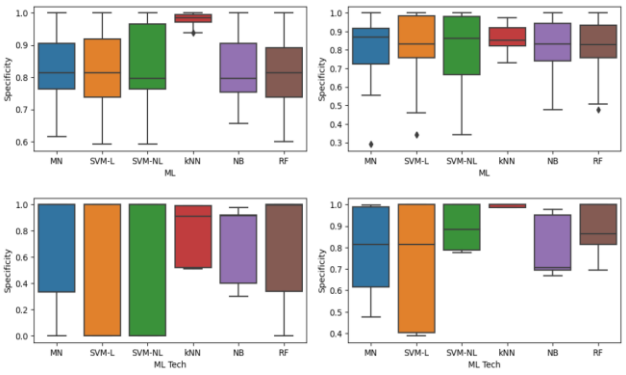

Fig. 2. Specificity, ML technique, and Data Sets (1-4).

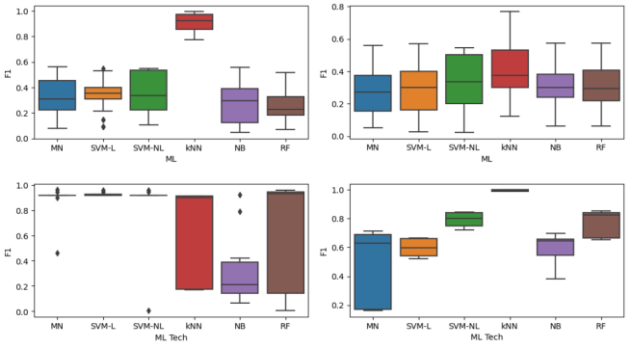

Fig. 3. F1 value, ML technique, and Data Sets (1-4).

Furthermore, Figure 4 shows how the balanced accuracy values change for data set #1 when the scaling and ML techniques change.

The used datasets were not balanced. Therefore, to investigate the impact of the balanced nature of the data on the classification accuracy, we randomly resampled the data sets to convert them to balanced data.

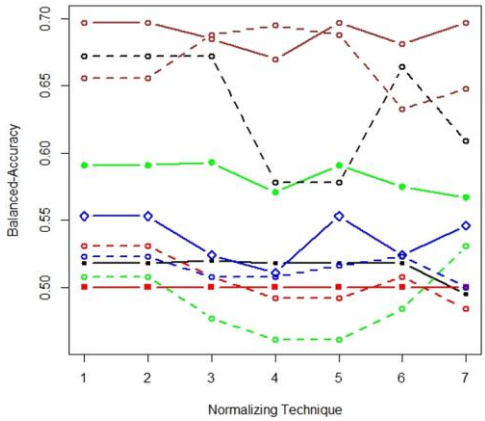

Fig. 4. Change of Balanced-Accuracy, ML technique, and Scaling Technique for Data set #1 (each color line represents each class value, solid line-original data, dash line - balanced data).

A three-way analysis of variance (ANOVA) test was utilized to investigate the influence of various factors on classification accuracy. Three main factors, namely the ML technique, the scaling technique, and the balanced nature of the data, were selected as the three factors.

Classification accuracy was selected as the dependent variable in the three-way ANOVA analysis. The following notations were used in the ANOVA analysis. F1 represents the nature of the data set (balanced or unbalanced), F2 represents the ML technique, and F3 represents the scaling technique. Interaction effects of two levels and three levels are represented by $F1 * F2$, $F1 * F3$, $F2 * F3$, and $F1 * F2 * F3$. Furthermore, the following symbols were used to indicate the level of significance of each factor. Here p is the level of significance. If $p<0.001$, we use '***, if $0.001<p<0.01$, we use '**', if $0.01<p<0.05$, we use '*', otherwise we use '>'.

The experimental outcomes given by ANOVA to investigate the impact of each of the above factors for each data set is represented by the Tables II–V, respectively.

VI. Discussion

According to the experimental outcome, it is clear that balanced accuracy (BA) depends on the type of ML technique and the scaling technique used. According to the experimental outcomes, even the other ten performance indicators vary based on the ML and scaling techniques. Furthermore, the indicators of the classification accuracy change for each level of the class variable differently. Therefore, the classification accuracy depends on the levels of the class variable.

TABLE II. Significance of the Balanced-nature, ML, Scaling Techniqueand the Classifying Accuracy for Data Set #1

Performance Indicator	F1	F2	F3	F1*F2	F1*F3	F2*F2	F1*F2*F3
Sensitivity	>	***	>	>	>	>	>
Specificity	>	***	>	>	>	>	>
PPV	***	***	>	>	>	>	>
NPV	**	***	>	>	>	>	>
Precision	***	***	>	>	>	>	>
Recall	>	***	>	>	>	>	>
F1	***	***	>	*	>	>	>
Prevalence	*	***	>	>	>	>	>
D.Rate	>	>	>	>	>	>	>
DP	>	>	>	>	>	>	>
B.Accuracy	*	**	>	*	>	>	>

TABLE III. Significance of the Balanced-Nature, ML, Scaling Technique and the Classifying Accuracy for Data Set #2

Performance Indicator	F1	F2	F3	F1*F2	F1*F3	F2*F2	F1*F2*F3
Sensitivity	>	**	>	>	>	>	>
Specificity	>	>	>	>	>	>	>
PPV	**	>	***	>	>	>	>
NPV	>	>	>	>	>	>	>
Precision	**	>	***	>	>	>	>
Recall	>	**	>	>	>	>	>
F1	***	***	>	>	>	>	>
Prevalence	>	>	>	>	>	>	>
D.Rate	**	>	>	>	>	>	>
DP	>	>	>	>	>	>	>
B.Accuracy	*	***	>	*	>	>	>

TABLE IV. Significance of the Balanced-nature, ML, Scaling Technique and the Classifying Accuracy for Data Set #3

Performance Indicator	F1	F2	F3	F1*F2	F1*F3	F2*F2	F1*F2*F3
Sensitivity	>	>	>	>	>	>	>
Specificity	>	>	>	>	>	>	>
PPV	>	*	>	>	>	>	>
NPV	*	>	>	>	>	>	>
Precision	>	*	>	>	>	>	>
Recall	>	>	>	>	>	>	>
F1	**	***	>	***	>	>	>
Prevalence	>	>	>	>	>	>	>
D.Rate	>	>	>	>	>	>	>
DP	>	>	>	>	>	>	>
B.Accuracy	*	***	>	*	>	>	>

TABLE V. Significance of the Balanced-Nature, ML, Scaling Technique and the Classifying Accuracy for Data Set #4

Performance Indicator	F1	F2	F3	F1*F2	F1*F3	F2*F2	F1*F2*F3
Sensitivity	***	***	>	***	>	>	>
Specificity	***	***	>	***	>	>	>
PPV	***	***	***	***	***	***	***
NPV	***	***	***	***	>	***	***
Precision	***	***	***	***	***	***	***
Recall	***	***	>	>	>	>	>
F1	***	***	>	***	>	>	***
Prevalence	***	***	*	***	>	***	***
D.Rate	***	***	>	***	>	>	***
DP	***	***	>	***	>	>	*
B.Accuracy	***	>	>	>	>	>	>

Based on the recorded values of Sensitivity, Specificity, Precision, Recall, F1, Prevalence, and BA, it is clear that the classification accuracy depends on the ML technique, the scaling technique, and the balanced nature of the data set.

According to the figures and the graphs representing performance indicators, it is evident that none of the ML techniques and scaling techniques dominates the others uniformly. The graphical illustration does not quantify the statistical evidence of the significance of the existing influential factors on classification accuracy. Therefore, ANOVA analysis is considered to address this.

Three factors, namely the nature of the data (balanced or unbalanced)-(F1), the used ML technique (F2), and the used scaling technique (F3) were considered to run the three-way ANOVA procedure. According to the experimental outcomes for data set #1, PPV, NPV, precision, and F1 are significantly influenced (p-value < 0.01) by the nature of the data. This significance is very high. Precision and balanced accuracy also change significantly (p-value < 0.05) on the nature of the data. All performance indicators except detection rate and prevalence are highly influenced by the type of ML technique used for the classification in data set #1. Furthermore, table 2 does not show any significant impact of the scaling technique on the accuracy of data set #1. The significant impact (p-value < 0.05) of two-way interactions of the nature of the data and the type of ML technique is shown for F1 and balanced accuracy for this dataset.

For data set #2, table 3 illustrates the impact of the above-selected factors on classification accuracy. According to table 3, the balanced nature of the data set significantly depends on the PPV, precision, F1, detection rate, and balanced accuracy. Similarly, the ML technique significantly impacts recall and F1. Furthermore, there is strong evidence of the two-factor interaction between the balanced nature of the data and the ML technique, which significantly impacts the PPV and precision. Also, there is statistical evidence for the significant impact of the above two-way interaction on balanced accuracy.

According to the output given in Table 4, the balanced nature of the data significantly impacts NPV, F1, and Balanced accuracy, while the ML technique succinctly impacts PPV, precision, F1, and balanced accuracy. For this data set #3, the scaling technique significantly impacts the balanced accuracy. Furthermore, the balanced nature of the data and ML technique impact F1 and balanced accuracy significantly. In order to identify which scaling techniques are different, Dunnett's test was utilized. According to these outcomes, there is no significant difference among N1, N2, N3, N4, N5, and N6, but it shows that the performance of N7 is significantly ($p=0.006$) higher than N1.

Finally, data set # 4 shows that the balanced nature of the data set significantly impacts all the performance indicators. ML techniques also significantly impact all the performance indicators except balance accuracy. The scaling technique significantly impacts several performance indicators, namely, PPV, NPV, and prevalence. This is illustrated in table 5.

VII. Conclusion

The accuracy of educational data classification is the prime aim of any educational data mining project. It enables learners, educators, and administrators to interpret the findings meaningfully. Furthermore, educational data analysis enables

educational institutions to complete resource allocation effectively. As a result, the learning process can be properly organized to improve students' learning experience.

In this study, we implemented six ML techniques, together with six scaling techniques and four data sets, to investigate the impact of the nature of the data and the scaling technique on the accuracy of educational data classification. Based on the outcomes of the conducted study, the impact of the scaling technique, the balanced nature of the data, and the ML technique impact the accuracy of the data classification. Though performance indicators are not uniformly influenced by the above factors across any given data set, these experimental findings showcase the importance of testing for various ML and scaling techniques when applying ML-based classification techniques in any data mining project. Further analysis indicates that the selected factors impact differently on the levels of the class variable. In addition, this study reveals the significance of the balanced nature of the data on classification accuracy.

As the performance accuracy depends on the given data set, additional studies are required to make any generalized statement about choosing the best scaling technique for any given data set. According to the findings, some scaling techniques performed poorer than the non-scaled data. Therefore, it is essential to test for several such techniques before adhering to a particular technique due to the sole popularity of the technique. The authors have identified some of the limitations of this study. Therefore, future research is expected to include expanding analysis for additional datasets, algorithms, and performance metrics to produce more robust and generalizable conclusions.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their insightful comments and the funding support given by the Burroughs Wellcome Fund to conduct this study.

REFERENCES

[1] H. Mannila, Data mining: machine learning, statistics, and databases, *IEEE*, 1996.

[2] I. Wickramasinghe and H. Kalutarage, Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation, *Soft Computing*, 2020.

[3] G. W. Foody and M. K. Arora, An evaluation of some factors affecting the accuracy of classification by an artificial neural network, *International Journal of Remote Sensing*, 18(4), pp. 799–810, 1997.

[4] J. W. Lee, J. B. Lee, M. Park M, and S. H. Song, An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis*, 48(4), pp 869–885, 2005.

[5] P. W. Novianti, V. L. Jong and K. C. B. Roes, Factors affecting the accuracy of a class prediction model in gene expression data, *BMC Bioinformatics*, 16, 199, 2015.

[6] G. Chandrashekar and F. Sahin, A survey on feature selection methods, Computers *& Electrical Engineering*, 40 (1), pp. 16–28, 2014.

[7] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Machine Learning Research*, 3, pp 1157–1182, 2003.

[8] A. J. Almalki, Accuracy analysis of Educational Data Mining using Feature Selection Algorithm. arXiv preprint arXiv:2107.10669. 2021.

[9] X. H. Cao, I. Stojkovic and Z. Obradovic, A robust data scaling algorithm to improve classification accuracies in biomedical data, *BMC Bioinformatic*, 17, pp. 359–369, 2016.

[10] M. Ramaswami and R. Bhaskaran, A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924, 2009.

[11] A. Lile, Analyzing e-learning systems using educational data mining techniques. *Mediterranean Journal of Social Sciences*, 2(3), pp. 403-403, 2011.

[12] B. Al Breiki, N. Zaki, and E.A. Mohamed, Using educational data mining techniques to predict student performance. In 2019 *International Conference on Electrical and Computing Technologies and Applications (ICECTA)* (pp. 1-5). IEEE, 2019.

[13] A. Anjewierden, B. Kolloffel and C. Hulshof, Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. *In International Workshop on Applying Data Mining in e*-Learning *(ADML 2007)*, pp 27-36, 2007.

[14] S. Pal, Mining educational data using classification to decrease dropout rate of students. arXiv preprint arXiv:1206.3078, 2012.

[15] W. Hämäläinen and M. Vinni, Comparison of Machine Learning Methods for Intelligent Tutoring Systems. In: Ikeda, M., Ashley, K.D., Chan, TW. (eds) Intelligent Tutoring Systems. ITS 2006. *Lecture Notes in Computer Science*, vol 4053. Springer, Berlin, Heidelberg.

[16] T. Fei, W. J. Heng, K. C. Toh and T. Qi, "Question classification for e-learning by artificial neural network," *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, Singapore, pp. 1757-1761 vol.3, 2003.

[17] K. Balabaeva and S. Kovalchuk, Comparison of Temporal and Non-Temporal Features Effect on Machine Learning Models Quality and Interpretability for Chronic Heart Failure Patients. *Procedia Computer Science*, 156, pp. 87–96, 2019.

[18] M. M. Ahsan, M. A. P. Mahmud, P.K. Saha, K. D. Gupta and Z. Siddique, Methods on Machine Learning Algorithms and Model Performance.*Technologies*, 52(9), 2019.

[19] J. Pan, Y. Zhuang and S. Fong, The Impact of DataNormalization on Stock Market Prediction: *Using SVM and Technical Indicators, Communications in Computer and Information Science*, 652, pp 16–28, Springer, Singapore. 45, pp. 199–209, 2016.

[20] B. K. Baradwaj and S. Pal, Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417, 2012.

[21] P. Cortez and A. Silva, Using Data Mining to Predict Secondary School Student Performance, In A Brito and J Teixeira Eds, *Proceedings of 5th Future BUsiness TEChnology Conference (FUBUTEC 2008)*, pp 5–12, 2008.

[22] G. Mahajan, and B. Saini, Educational Data Mining: A state-of-theart survey on tools and techniques used in EDM. *International Journal of* Computer *Applications & Information Technology*, 12(1), pp. 310-316, 2020.

[23] J. C. F. Wong and T. C. Y. Yip, "Measuring Students' Academic Performance through Educational Data Mining," International Journal of Information and Education Technology vol. 10, no. 11, pp. 797 804, 2020.

[24] I. Wickramasinghe, Applications of Machine Learning in cricket: A systematic review, *Machine Learning with Applications*, 10. 2022.

[25] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi, Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT),* 48(3), pp. 128-138, 2017.

[26] P. Perera, Y. C. Tian, C. Fidge and W. Kelly, A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, ES. (eds) *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science*, vol 10634. Springer, Cham, 2017.

[27] C. D. Manning, Introduction to information retrieval. *Syngress Publishing*. 227, 2008

[28] E. Carrizosa, B. Martin-Barragan, and D. R. Morales, Binarized support vector machines. *INFORMS Journal on Computing*, 22(1), 154-167, 2010.

[29] C. Cortes and V. Vapnik, Support-vector networks. *Machine learning*, 20, 273-297, 1995.

[30] R. Couronn´e, P. Probst and A. L. Boulesteix, Random forest versus logistic regression: a large-scale benchmark experiment, *BMC Bioinformatic*, 19(1), 270, 2018.

[31] L. Breiman, Random forests, *Machine Learning*, 45(1), pp. 5–32, 2001.

[32] G. Biau and E. Scornet, A random forest guided tour, *Test*, 25(2), pp 197–227, 2016.

[33] J. Seshapanpu, Students Performance in Exams, https://www.kaggle.com/spscientist/students-performance-in-exams,

[34] K. Chaturvedi, D. K. Vishwakarma and N. Singh, COVID-19 and its impact on education, social life and mental health of students: A survey. *Children and youth services review*, 121, 105866, 2021.

[35] J. Kothari, Finance & Accounting Courses -Udemy (13K+ course), Kaggle, https://doi.org/10.34740/KAGGLE/DSV/1494962, 2020.

[36] M. H. Suzan, N. A. Samrin, A. A. Biswas and A. Pramanik, Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-7, 2021.