International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1220-1233 www.ijaem.net ISSN: 2395-5252

Ensemble Machine Learning-Based Sentiment Analysis Model for Teachers' Performance Evaluation

Vedaste Nyandwi, Olivier Habimana, Nyesheja M.Enan
Faculty of Computing and Information Sciences, University of Lay Adventists of Kigali (UNILAK)
School of Education Lecturer at University of Rwanda, College of Education,Rukara Campus
Dean of Faculty of Computing and Information Sciences, University of Lay Adventists of Kigali (UNILAK)

Date of Submission: 15-04-2023 Date of Acceptance: 25-04-2023

ABSTRACT
Teacher evaluation has emerged as a central theme in school reform efforts. Students' rating decisions are often used by HLI leaders when hiring, promoting, determining tenure, raising salaries, and making other performance measurements. Student evaluation is important because it may be the only opportunity for students to provide constructive feedback that may improve their learning outcomes in the future. Sentiment analysis and data mining techniques (such as Lexicon-Based and machine learning algorithms like Naïve Bayes, support vector machine, LDA, etc…)were introduced as a tool for automatically extracting insights and useful information from user-generated data. This study analyzed students' instructional feedback using machine learning algorithms. The researcher designed a suitable machine learning model for sentiment analysis to predict students' polarity. At the training size of 80%, stacking ensemble machine-based classifier produced the precision (weighted avg) of 86%, recall of 86%, f1-score of 85% and the accuracy of 86%. The qualitative analysis also demonstrated that stacking ensemble machine-based classifier which combined the advantages of three classifiers (LDA, SVM, and NB) dominated individual classifiers and other ensemble machine-based classifiers which combined two different classifiers.

Keywords: Feedback, Sentiment Analysis, Performance Evaluation, Linear Discriminant Analysis, Naïve Bayes, Support Vector Machine, Teacher Evaluation.

I. INTRODUCTION

Education causes a natural and long-lasting shift in an individual's way of thinking and ability to achieve desired outcomes. Education is a process that begins at birth and continues **to** the end of life. It is the path to our destiny because it can only be attained if people have the necessary knowledge, skills, and spirit(Reinhardt & Beu, 2015). There are three main types of education: formal, informal, and informal.

In education, teacher performance evaluation is a continuous, routine, and mandatory exercise. In higher education institutions, student ratings for teaching are a common way to assess the effectiveness of the teaching and learning process. Students' ratings are typically collected at the end of the semester via paper-based surveys or Google form surveys. Though, online reviews have become more popular these days. Students' rating decisions are often used by HLI leaders when hiring, promoting, determining tenure, raising salaries, and making other performance measurements.

Student evaluation is important because it may be the only opportunity for students to provide constructive feedback that may improve their learning outcomes in the future(Baddam et al., 2019). The development of the data mining approach facilitates research into the classification of characteristics of the provided datasets. Applications such as student/teacher performance evaluation play an important role in measuring the effectiveness of educational teachers. The traditional way to evaluate an educator's performance is from a student perspective(Vijayalakshmi et al., 2020).

Students' emotions and opinions provide valuable information not only for analyzing student behavior toward courses, subjects, or teachers, but also for reforming strategies and institutions(Kastrati et al., 2021). The main necessary benefit of this analysis is the form of feedback provided to educators to refine their courses and teaching practices to produce students with a higher learning experience. As a result, sentiment analysis was introduced as a tool for extracting insights and useful information from user-generated data automatically. One of the tasks of natural language processing (NLP) is sentiment analysis. For example (Vijayalakshmi et al., 2020) used different machine learning algorithms such as Naive Bayes, KNearest Neighbor, Random Forest, Support Vector Machine, and Decision Tree. The main objective was to demonstrate the variable that relies on the teachers' performance such as accuracy, precision, recall, specificity, and sensitivity using different machine learning algorithms.

The studies showed that the proposed machine learning algorithms (Vector Support Machine, Linear Discriminant Analysis, and Naïve Bayes) have advantages and disadvantages when extracting text features, tokenizing the text,tripping tags, multiple white spaces, punctuations, numeric characters, and short words from the text and remove stop words from the text**.**

Naive Bayes is used for document-level sentiment classification, and it produces relatively good output and performance. Because it simply updates the counts required to estimate the conditional and algorithmic probabilities, this algorithm is simple to use and apply to a data stream(Malviya et al., 2020). However, its performance is often imperfect because it does not do well in extracting key patterns, and by inappropriate feature selection(Qiang, 2010). Support Vector Machine gives efficient results in traditional text categorization.

Naive Bayes is a fast algorithm that can handle both continuous and discrete data; training and classification can be done in a single pass over the data. It is also impervious to noise features. It works with small amounts of data, handles multiple classes, and is unaffected by irrelevant characteristics(Kalcheva et al., 2020). In addition and Naive Bayes produced lower F1-score metrics when trained with a small amount of data (Ahmad et al., 2018).

SVM essentially finds the best possible boundaries to distinguish between positive and negative training samples(Malviya et al., 2020). To classify data points, the SVM classifier estimates the hyperplane based on the feature set. The dimensions of the hyperplane change depending on the number of features(Lee et al., 2022), though the study showed that Support vector machinesproduced lower F1-score metrics when trained with a small amount of data (Ahmad et al., 2018).

LDA extracts text features by determining the linear combination of independent variables that models and classifies the response variable and calculating discriminant scores(Akbarzadeh et al., 2022). LDA models the distribution of the independent variables (X) separately in each response class. The Bayes theorem is then used to calculate the likelihood of the X values' response levels. LDA computes discriminant scores by calculating the linear combination of independent variables that models and categorizes the response variable(Akbarzadeh et al., 2022), LDA, on the other hand, has poor performance with non-linear problems and a small sample size(Wu & Feng, 2015).

To address the challenges of the above-mentioned machine learning models, we propose the stacking ensemble machine learning-based that has the advantage of combining the capabilities of several high-performing machine learning models on a classification or regression task to make predictions that gives better results compared to the individual machine learning model. In summary, this study focuses on text feedback in form of sentences from students to analyze students' instructional feedback using machine learning algorithms and design a suitable machine learning model for sentiment analysis to predict students' polarity.

II. RELATED WORK

In this section, we briefly discussed machine learning algorithms for sentiment analysis aiming to analyze students' instructional feedback using machine learning algorithms and design a suitable machine learning model for sentiment analysis to predict students' polarity. These approaches fall into four categories: Linear Discriminant Analysis, Support Vector Machines, Naïve Bayes and Ensemble Machine Learning.

2.1 Linear Discriminant Analysis

Linear Discriminant Analysis is a dimension reduction technique commonly used for supervised classification problems. The method's goal is to maximize the ratio of between-group variance to within-group variance. When the value of this ratio is at its maximum, the samples within each group have the least amount of scatter and the groups are the most separated from one another(Vaibhaw et al., 2020).

The LDA has different applications including customer identification. Assume the owner of the business wants to identify the types of customers most likely to purchase a specific product in a shopping mall. He/she can gather all of the characteristics of the customers by conducting a simple question-and-answer survey. In this case, a linear discriminant analysis will assist him/her in identifying and selecting the features that can describe the characteristics of the group of customers who are most likely to purchase that particular product in the shopping mall.

2.2. Support Vector Machines

Support vector machines or SVMs are one of the most common supervised learning algorithms used for both classification and regression problems. SVMs are based on statistical learning theory(Malik et al., 2021). Support Vector machines are systems that use the hypothesis space of linear functions in a high dimensional feature space and are trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory(Nguyen, 2016). The goal of the SVM algorithm is to create optimal lines or decision boundaries that can divide n-dimensional space into classes so that new data points can be easily placed in the correct category in the future. This best decision boundary is called the hyperplane(Noble, 2006). SVMs are used in many applications, such as categorizing reviews based on quality.

2.3 Naive Bayesian Classification

The Naive Bayes algorithm is a simple and effective predictive modeling algorithm. The model contains two types of probabilities that can be calculated directly from the training data. The probabilities of each class

(i) Each class's conditional probabilities at each x value. The probabilistic model, once calculated, can be used to predict new data using Bayes' theorem(Shobha & Rangaswamy, 2018).

The formula for Bayes' theorem is given as:

$$\mathrm{P}(yj|xi) = \frac{\mathrm{P(xi|yj)P(yj)}}{\mathrm{P(xi)}}, \quad i = 1,2,3, \ldots \ldots . \; n \quad (1)$$

Where:

$(yj|xi)$ is a posterior probability, the probability of event yj (hypothesis) given event xi (prior knowledge), $(xi|yj)$ is a likelihood probability, or the probability that a hypothesis is true given the evidence. (yj) is a prior probability, or the probability of a hypothesis prior to seeing the evidence, and (xi) is an evidence probability.

2.3 Ensemble Machine Learning

Ensemble machine learning is a method of building multiple machine learning models and combining them to obtain better results. An ensemble machine learning model is more likely to produce accurate results than a single model(Necati Demir, 2016). The three main types of ensemble learning methods are bagging, stacking, and boosting. Bagging ensemble learning method that searches for diverse sets of ensemble members by changing training data, boosting is an ensemble learning method that combines groups of weak learning methods into one strong learner to minimize training errors, and stacking is an ensemble technique that finds a diverse set of members by varying the type of model that fits the training data and combining predictions using one model(Brown, 2010).

The ensemble model has the following advantages in general:

- The ensemble reduces the spread in a predictive model's average skill.
- The ensemble outperforms any contributing member in terms of average prediction performance.
- The reduction in the variance component of prediction errors made by the contributing models is frequently the mechanism for improved ensemble performance.

2.4 Related Case Studies

According to M.O. et al.(2016), teacher performance evaluation is among the ways to achieve the highest standards in higher education. This study used data mining techniques to present a practical system model for assessing and forecasting teacher performance in higher education institutions. The data set was composed of 216 (61.89%) permanents, 72(20.34%) temporal, and 61 (17.48%) employees’ contracts ranging from professors to assistant professors. MLP proposed by M.O. et al.(2016), produced 82.5% of accuracy, 82.8 % of precision, 82.5% of recall, and 82.4 % of F1-Score. (M.O. et al., 2016) suggested other classification algorithms that improve classification accuracy.

Higher education is undergoing a major shift in the large-scale transfer of faculty knowledge and experience to the student body (T. Manjunath Kumar, 2019). According to T. Manjunath Kumar, 2019) study, The faculty assignment process's main goal is to maximize student learning capacity by assigning the best faculty to the right courses based on the teacher's qualifications, skills, and abilities. In this study, different machine learning algorithms such as SVM, NB, LR, etc… were implemented to evaluate teachers’ performance. Different performance metrics were used to evaluate the algorithms. The size of the dataset used was composed of 892 students' comments. SVM produced 63% of accuracy, 73% of F1-Score, 67.27% of precision, and 80% of recall while Naïve Bayes produced 61.89 % of accuracy, 74.24 % of F1-Score, 64.47% of precision, and 85.50% of recall. According to T. Manjunath Kumar (2019), more experiments were needed to enhance the model's efficiency and use a huge dataset.

Machine learning algorithms are used in different domains in higher learning institutions. According to Xia & Yan (2021), more precise and accurate assessment models of the effectiveness and nature of music teacher assessment are needed for educational leaders to make effective decisions about music education in schools. In this study, researchers used the Naïve Bayes classifier for the valuation of music performance. The dataset used, was composed of 290 samples divided into 220 training samples and 70 testing samples. The classifier produced an accuracy of 76.7%.

III. PROPOSED METHOD

This section presents a detailed description of the framework design, tools, processes, and data collection techniques that were employed in this study to determine the findings required to meet the study objectives.

3.1. Research approach

The goal of student sentiment analysis is to look into the mechanisms that cause language and text to elicit emotions like joy, anger, sadness, and happiness, as well as the emotions that students may feel after reading the text. The evaluation of education by students is an important link in the development and support systems for teachers at various universities (Peng et al., 2022). Several machine learning algorithms, including Support Vector Machines, Nave Bayes, and Linear Discriminant Analysis, were proposed in this study to predict whether comments were positive, negative, or neutral. As evaluation measures, this study proposed the stacking ensemble machine learning algorithm for improved sensibility, specificity, and predictive values.

Figure 3below describes the steps of building a machine learning algorithm from libraries importation to classification metrics. The library is made up of related modules. It includes code bundles that can be reused in multiple programs. Python libraries are extremely useful in fields such as machine learning, data science, and data visualization. Exploratory data analysis focuses on gaining insights into the data. Model training and development are carried out using training data. Training sets are frequently used to estimate various parameters and compare the performance of various models.

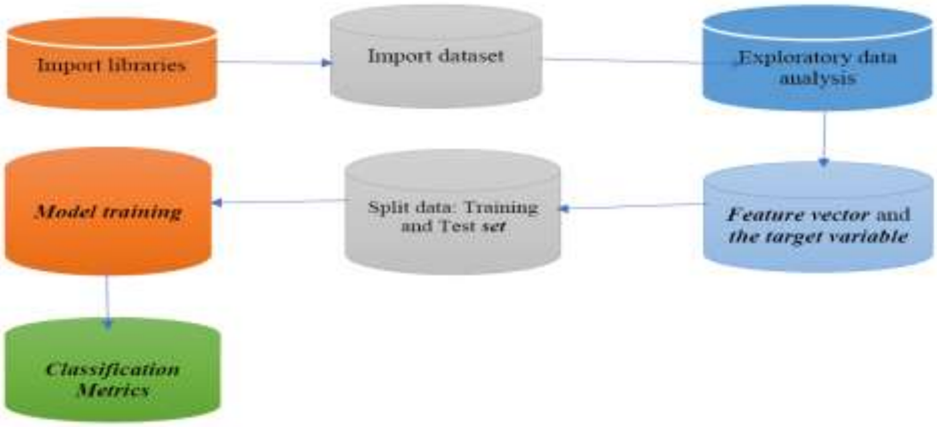

Import libraries
Import dataset
Exploratory data analysis
Model training
Split data: Training and Test set
Feature vector and the target variable
Classification Metrics

Figure 1: Building machine learning algorithm

3.2 Research Design

In this study, secondary data were used that are in text format. Much focus was put on the data downloaded from **kaggle.com (ref https://www.kaggle.com/datasets/brarajit18/student-feedback-dataset)** whose data were taken and fit into the machine learning-based sentiment analysis algorithms to meet our specific objectives.

3.3 Description of the population

The dataset for this study was gathered from students at a well-known university in North India. Based on student feedback data, data were collected and analyzed to create the overall Institutional Report. In this dataset, students provided feedback on the areas that impacted teaching and learning in HLI (teaching, course content, examination, lab work, library facilities, and extra-curricular activities).

3.4 Data collection

Only secondary data were downloaded for this study from the kaggle.com dataset, a platform for big data sets. Kaggle is the world's largest data scientist and machine learning community. Kaggle began by only offering machine learning contests, but has since evolved into a public cloud-based data science platform.

Kaggle is not only helping the researchers to solve difficult problems, employs strong teams, demonstrates the power of data science, and also accesses various libraries and frameworks that were incompatible with the author's local device. The dataset has 925 rows and 2 columns that represent Label and Teaching and learning feedback (in form of texts) from a prominent university in North India.

3.5 Development Technologies

While implementing the ensemble machine learning-based sentiment analysis models on students' survey feedback, different technologies were used. CSV file that contains students' feedback on teaching. A programming language is required to analyze data and build a stacking ensemble machine learning-based sentiment analysis model for teachers' performance evaluation. Python took a significant lead in determining the best programming language for the research's specific goals. Python is currently one of the most popular programming languages. It was created by Guido Van Rossum in 1991, and became one of the most commonly used languages, alongside C++, Java, and others (Insights, 2016).

Python is available on all OS and is also available as open source software under the name CPython. Python also provides the flexibility to use various pre-built libraries such as Numpy and Scipy, and also supports other Machine Learning libraries like Scikit Learn, Keras, and Pytorch, NLP libraries like NLTK, and also frameworks to deploy the algorithms such as Flask and which are very suitable to predict students' polarity in sentiment analysis. Kaggle gives access to various libraries and frameworks that were incompatible with the author's local device.

IV. EXPERIMENTS

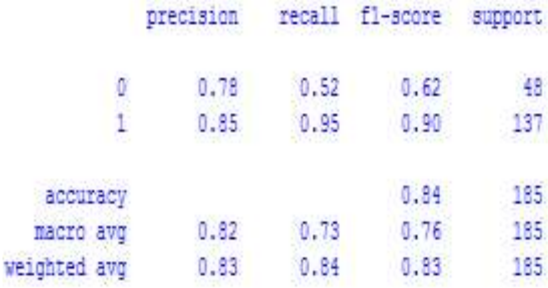
We evaluate the effectiveness of stacking ensemble machine learning algorithm on dataset downloaded from **kaggle.com**. This dataset was gathered from students at a well-known university in North India. Therefore, this section presents the empirical results obtained.

4.1 Dataset description

Data used in this research were downloaded from an online community platform for data scientists and machine learning enthusiasts. The dataset has 2 columns (Label and Teaching) and 925 rows. The teaching column contained students' feedback submitted on different learning areas such as teaching, course content, exams, lab work, library resources, and extracurricular activities. Contents submitted by students were in sentence case, upper case, or lower case text formats. The data were used in the machine learning algorithms proposed to predict students' polarity. Different graphs were plotted to visualize the results of the machine learning algorithms used.

4.2 Performance of Individual Algorithms

Support Vector Machines, Linear Discriminant Analysis and Naïve Bayes classifiers were applied to classify data points on teaching and learning feedback students submitted and predicted students' polarity. The performance of this classifier was determined on the basis of its precision, recall, f1-score, accuracy, the area under curve (AUC), and the Matthews correlation coefficient (MCC). After training the classifier using a training size of 80%, SVM produced 83.78% of accuracy, 73.48% of AUC and 54.43% MCC, LDA produced t 77.84% of accuracy, 60.00% of AUC, and 32.11% of MCC,NB produced 83.24 % of accuracy, 69.13 % of AUC and 51.44 % of MCC, and stacking ensemble machine algorithm produced the accuracy of 85.95%, ACC of 75.16% and MCC of 60.22%

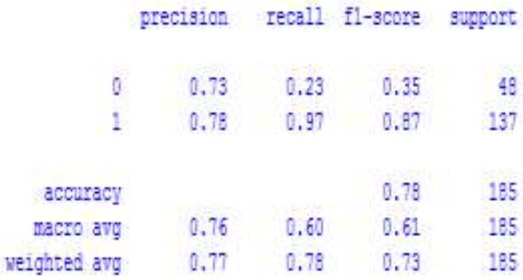


```
              precision    recall  f1-score   support

           0       0.78      0.52      0.62        48
           1       0.85      0.95      0.90       137

    accuracy                           0.84       185
   macro avg       0.82      0.73      0.76       185
weighted avg       0.83      0.84      0.83       185
```

Table 1:Classification report for Support Vector Machines.

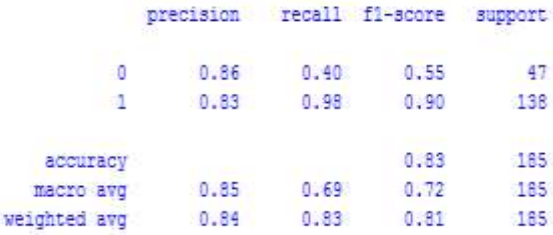


```
              precision    recall  f1-score   support

           0       0.73      0.23      0.35        48
           1       0.78      0.97      0.87       137

    accuracy                           0.78       185
   macro avg       0.76      0.60      0.61       185
weighted avg       0.77      0.78      0.73       185
```

Table 2: Classification report for Linear Discriminant Analysis.

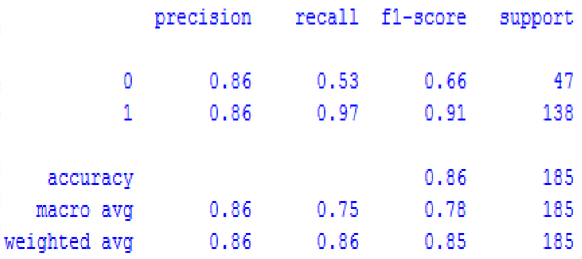


```
              precision    recall  f1-score   support

           0       0.86      0.40      0.55        47
           1       0.83      0.98      0.90       138

    accuracy                           0.83       185
   macro avg       0.85      0.69      0.72       185
weighted avg       0.84      0.83      0.81       185
```

Table 3: Classification report for Naive Bayes Classifier



```
              precision    recall  f1-score   support

           0       0.86      0.53      0.66        47
           1       0.86      0.97      0.91       138

    accuracy                           0.86       185
   macro avg       0.86      0.75      0.78       185
weighted avg       0.86      0.86      0.85       185
```

Table 4: Classification Report for Ensemble SVM, LDA, and NB.

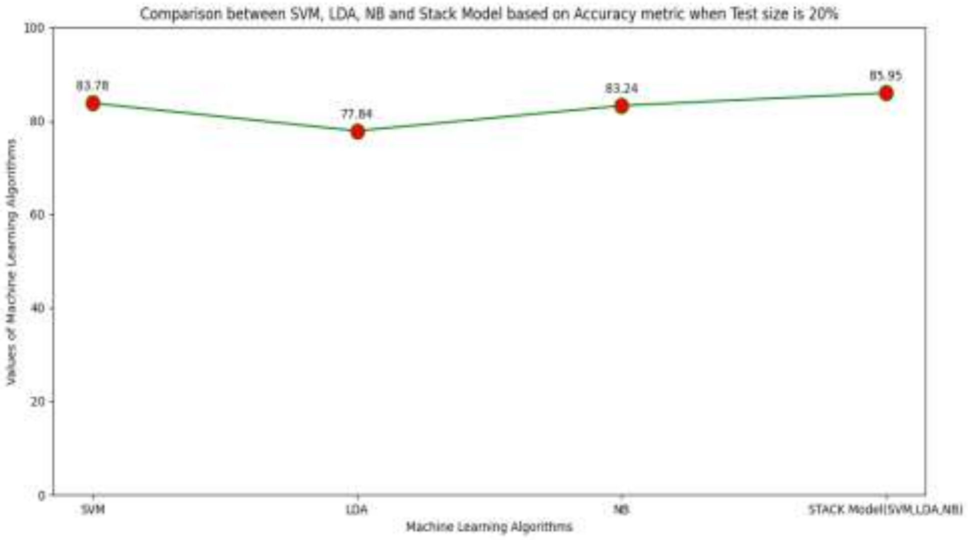


Figure 2: Comparison between SVM, LDA, NB, and Stack Model based on Accuracy metric

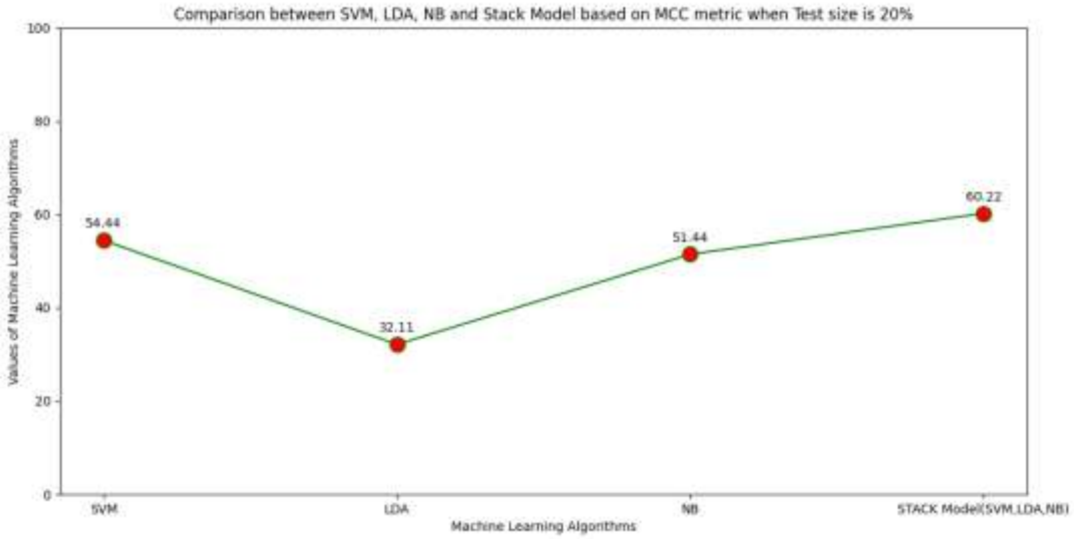


Figure 3: Comparison between SVM, LDA, NB, and Stack Model based on MCC metric.

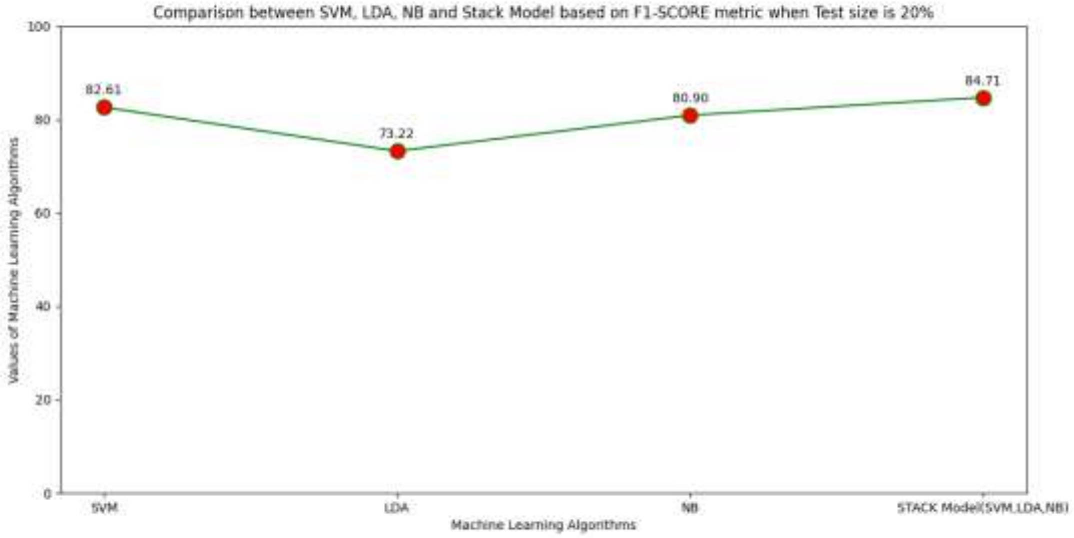


Figure 4: Comparison between SVM, LDA, NB, and Stack Model based on F1-SCORE metric.

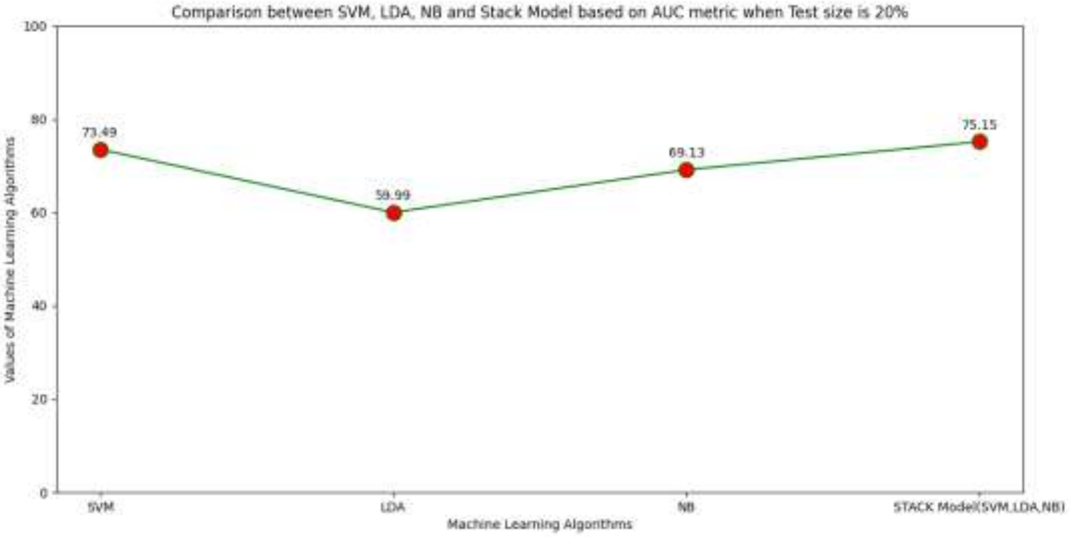


Figure 5: Comparison between SVM, LDA, NB, and Stack Model based on AUC metric.

4.3 Baseline methods

Experimental results were obtained with stacking ensemble machine learning-based sentiment analysis model and the related studies machine learning algorithms. Most importantly, the experimental results show that the stacking ensemble machine learning-based sentiment analysis model performs better than all outcomes of the studies discussed in the previous sections.

Table 5: Discussion of the results (in accuracy, AUC, Precision, Recall, F1-Score, and MCC) on MLP, SVM, NB, and SEMLM.

		PERFORMANCE METRICS					
AUTHORS	**MLAs**	**ACC**	**AUC**	**PREC**	**RECALL**	**F1-SC**	**MCC**
M.O. et al., 2016	**MLP**	82.50%	-	82.80%	82.50%	82.40%	-
T. Manjunath Kumar, 2019	**SVM**	63.00%	-	67.27%	80.00%	73.00%	-
	NB	61.89%	-	64.47%	85.50%	74.24%	-
Xia & Yan, 2021	**NB**	76.70%	-	-	-	-	-
OURS	**SEML**	**85.95%**	**75.16%**	**86.00%**	**86.00%**	**85.00%**	**60.22%**

Naive Bayes and SVM produce relatively good output and performance. However, the performance of NB is often imperfect due to its poor performance in extracting important patterns and poor feature selection (Qiang, 2010). On the other hand, the SVM essentially finds the best possible bounds to distinguish between positive and negative training patterns (Malviya et al., 2020). To classify the data points, the SVM classifier estimates a hyperplane based on the feature set. However, support vector machines produced lower F1 score metrics when trained on small amounts of data. LDA includes poor performance with non-linear problems and small sample size.

Due to the above advantages and disadvantages discussed in the above sections, the Stacking ensemble machine learning-based (SEML) model performs better than SVM, NB, and LDA due to its ability of combining the advantages of SVM, NB, and LDA in extracting important features. The SEML model performs better than the above model in Table 5 with an increase:

(a) In accuracy of 3.45%, 22.95%, 24.06%, and 9.25% on MLP (M.O. et al., 2016) SVM(T. Manjunath Kumar, 2019), NB (T. Manjunath Kumar, 2019), and NB(Xia & Yan, 2021) respectively.
(b) In precision of 3.2%, 18.73%, and 21.53% on MLP (M.O. et al., 2016), SVM (T. Manjunath Kumar, 2019), NB (T. Manjunath Kumar, 2019)
(c) In recall of 3.5%, 6%, and 0.5% on MLP (M.O. et al., 2016), SVM (T. Manjunath Kumar, 2019), NB (T. Manjunath Kumar, 2019)
(d) In F1-SCORE of 2.6%, 12%, 10.76% on MLP (M.O. et al., 2016), SVM (T. Manjunath Kumar, 2019), NB (T. Manjunath Kumar, 2019)

In brief, our experimental results prove that the proposed machine learning model performs well compared to individual machine learning algorithms in metrics such as accuracy, f1-score, the area under curve, MCC, precision, and recall.

4.4 Ablation studies

Ensemble techniques are methods that use multiple learning algorithms or models to generate optimal predictive models.In this research, different stacking ensemble machine learning classifiers were built to predict students' polarity from students' feedback on teaching and learning. The experiments carried out in this research showed that ensembling Support Vector Machines, Linear Discriminant Analysis, and Naïve Bayes classifiers produced better results compared to ensembling two different classifiers as shown in Table 6 below:

Table 6: Comparison between four implemented stacking ensemble classifiers in terms of ACC, MCC, F1-SCORE, and AUC.

	Classifiers' performance for test size of 20%			
	Accuracy	**Matthew Correlation Coefficient (MCC)**	**F1-SCORE**	**Area Under Curve (AUC)**
En SVM and NB	85.40	58.50	84.03	74.08
En LDA and NB	84.86	56.77	83.33	73.02
En SVM and LDA	78.37	34.43	74.10	61.04
En SVM, LDA, NB	85.94	60.21	84.71	75.14

Staking ensemble machine learning model of three classifiers produces an increase in the accuracy of 0.54%, 1.08%, and 7.57 on ensemble SVM and NM, ensemble LDA and NB, and ensemble SVM and LDA. The model produces an increase in the Mathew correlation coefficient (MCC) of 1.71%, 3.44%, and 25.78% on ensemble SVM and NM, ensemble LDA and NB, and ensemble SVM and LDA. The model produces an increase in the F1-SCORE of 0.68%, 1.38%, and 10.61% on ensemble SVM and NM, ensemble LDA and NB, and ensemble SVM and LDA. The model also produces an increase in the area under curve (AUC) of 1.06%, 2.12%, and 14.1% on ensemble SVM and NM, ensemble LDA and NB, and ensemble SVM and LDA.

4.5Discussion and qualitative analysis.

Different qualitative analysis were performed to analyze the performance of the stacking ensemble machine learning-based model (**SEML**) compared to individual machine learning algorithms. The analysis was performed on different training sizes 60, 70, 80, and 90 as shown in Table 7 and Figure 8, Figure 9, Figure 10, and Figure 11 below.

Table 7: Qualitative analysis of machine learning algorithms and SEMLM using different training sizes

		Performance Metrics in percentage			
Training Size (%)	**MLA**	**ACC**	**MCC**	**F1-SCORE**	**AUC**
60	**SVM**	84.05	55.02	82.55	72.65
	LDA	78.38	34.37	73.88	60.70
	NB	82.70	51.25	80.15	68.67
	SEMLM	84.32	56.29	82.93	73.43
70	**SVM**	83.45	53.22	82.12	72.57

	LDA	77.70	31.77	73.58	60.56
	NB	82.01	51.00	79.43	68.75
	SEMLM	84.53	58.47	83.35	74.99
80	**SVM**	83.78	54.44	82.61	73.49
	LDA	77.84	32.11	73.22	60.00
	NB	83.24	51.44	80.90	69.13
	SEMLM	85.95	60.22	84.71	75.15
90	**SVM**	80.65	44.74	79.24	69.29
	LDA	77.42	30.96	70.13	56.25
	NB	88.17	65.80	86.63	75.00
	SEMLM	89.25	68.55	88.34	78.84

The stacking ensemble machine learning-based model that combined three machine learning algorithms (SVM, LDA, and NB) performed well compared to the Stacking ensemble machine learning-based model that combined two machine learning algorithms ((SVM and LDA), (SVM and NB), and (LDA and NB)) and individual algorithms as shown in Table 7 and Figure 8 below. On the training size of 60%, the SEML model performs better than other ensemble machine-based models with an increase in accuracy of 0.27%, 5.94%, and 1.62% on SVM, LDA, and NB. On the training size of 70%, the SEML model performs with an increase in the accuracy of 1.08%, 6.83%, and 2.52% on SVM, LDA, and NB. On the training size of 80%, the SEML model performs with an increase in the accuracy of 2.17%, 8.11%, and 2.71% on SVM, LDA, and NB, and on the training size of 90%, the SEML model performs with an increase in the accuracy of 8.6%, 11.83% and 1.08% on SVM, LDA, and NB.

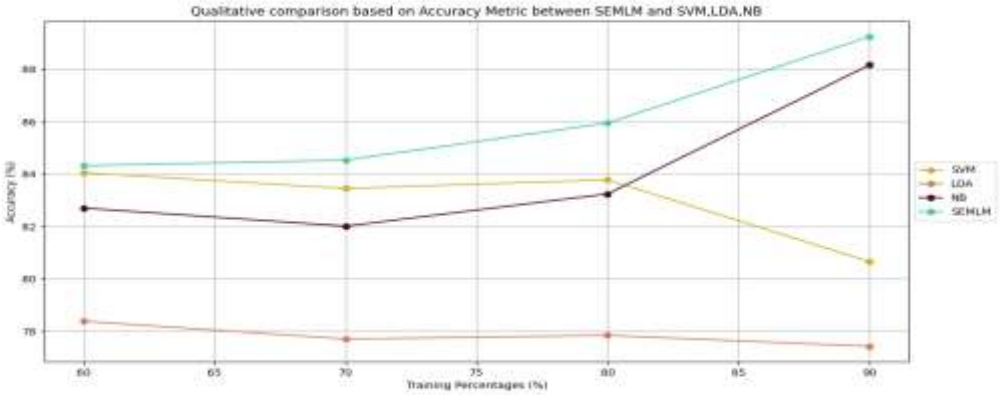


Figure 6: Sentiment accuracy metric versus training size

In terms of MCC, the model produced an increase of 1.27%, 21.92%, and 5.04% on SVM, LDA, and NB at the training size of 60%, an increase of 5.25%, 26.7%, and 7.47% on SVM, LDA, and NB at the training size of 70%, an increase of 5.78%, 28.11% and 8.78% on SVM, LDA and NB at the training size of 80%, and an increase of 23.81%, 37.59% and 2.75% on SVM, LDA and NB at the training size of 90% as shown in Table 7 and Figure 9.

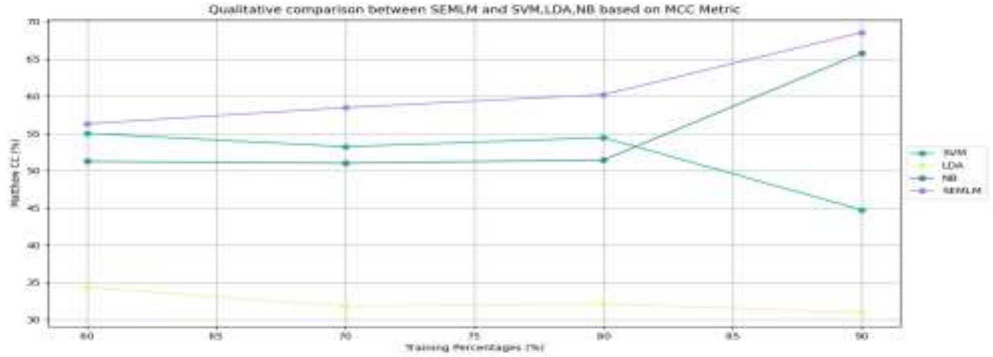


Figure 7: Sentiment Matthews Correlation Coefficient (MCC) metric versus training size.

In terms of F1-SCORE, the model produced an increase of 0.38%, 9.05%, and 2.78% on SVM, LDA, and NB at the training size of 60%, an increase of 1.23%, 9.77%, and 3.92% on SVM, LDA, and NB at the training size of 70%, an increase of 2.1%, 11.49% and 3.81% on SVM, LDA, and NB at the training size of 80%, and an increase of 9.1%, 18.21% and 1.71% on SVM, LDA, and NB at the training size of 90% shown in Table 7 andFigure 10.

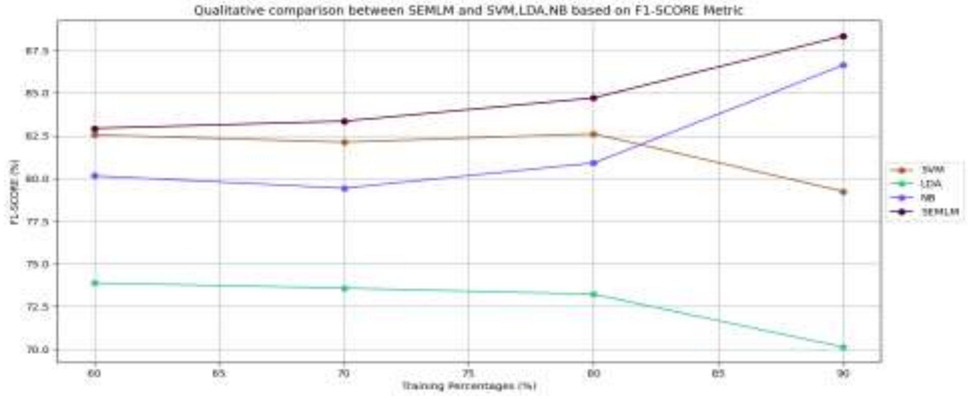


Figure 8: Sentiment F1-SCORE Metric versus Training size

In terms of AUC, the model produced an increase of 0.78%, 12.73%, and 4.76% on SVM, LDA, and NB at the training size of 60%, an increase of 2.42%, 14.43%, and 6.24% on SVM, LDA, and NB at the training size of 70%, an increase of 1.66%, 15.15% and 6.02% on SVM, LDA, and NB at the training size of 80%, and an increase of 9.55%, 22.59% and 3.84% on SVM, LDA, and NB at the training size of 90% as shown in Table 7 and Figure 11.

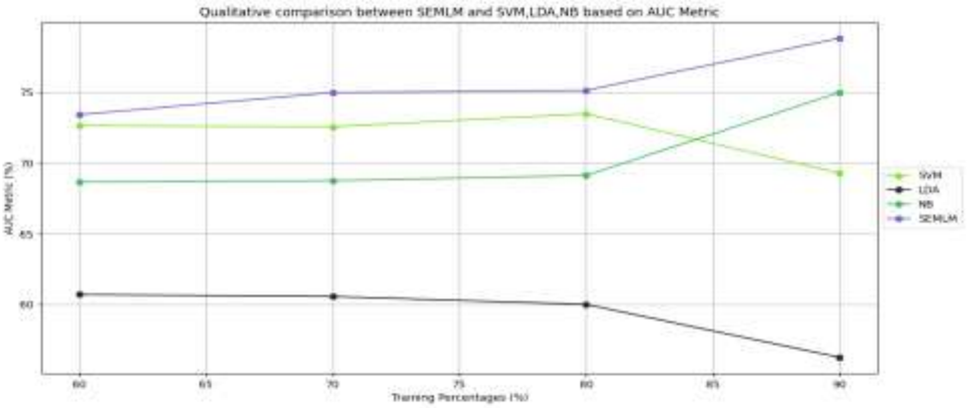


Figure 9: Sentiment Area under curve (AUC) Metric versus Training Size.

The above results showed that ensembling three machine learning algorithms dominated other ensemble machine learning. It is because the machine learning discussed in this study demonstrated advantages and challenges. Stacking ensemble machine learning-based gained their advantages to make predictions that gave better results compared to the individual machine learning model.

V. CONCLUSION

In this study, the researcher used multiple machine learning algorithms to analyze students' instructional feedback, predict the students' polarity based on the importance of features, and design a suitable machine learning model for sentiment analysis that gives better results compared to the individual machine learning model. Different classification models were used: Support Vector Machines (SVM) classifier, Naive Bayes (NB) classifier, Linear Discriminant Analysis (LDA) classifier, and stacking ensemble machine learning-based classifier for analyzing students' instructional feedback and predictingthe students' polarity. The researcher compared the results produced by each machine learning algorithm.

Due to the challenges and advantages of each classifier, the stacking ensemble machine-based classifier was implemented to improve its performance. The researcher used different tests and training sizes to prove the performance of the proposed machine learning algorithm whose goal was to improve the performance. Different stacking ensemble machine learning-based classifiers were implemented such as ensembling the Support Vector Machines and Naïve Bayes, ensembling Support Vector Machines and Linear Discriminant Analysis, ensembling Linear Discriminant Analysis and Naïve Bayes, and ensembling three classifiers LDA, SVM, and NB on different training size 60%, 70%, 80% and 90% respectively.

The results showed that stacking ensemble machine-based classifier which combined the advantages of three classifiers (LDA, SVM, and NB) produced good results compared to individual classifiers and ensembling two classifiers.

This research can be extended in multiple dimensions. Futurists (Future researchers) can expand this work by collecting big dataset in HLI that contains feedback in different languages and ensemble different better-performing machine learning algorithms to analyze the feedback of students in HLI and to test the behaviors of the model in terms of sentence size.

REFERENCES

[1]. Ahmad, M., Aftab, S., Bashir, M. S., Hameed, N., Ali, I., & Nawaz, Z. (2018). SVM optimization for sentiment analysis. International Journal of Advanced Computer Science and Applications, 9(4), 393–398. https://doi.org/10.14569/IJACSA.2018.090455

[2]. Akbarzadeh, M., Alipour, N., Moheimani, H., Zahedi, A. S., Hosseini-Esfahani, F., Lanjanian, H., Azizi, F., & Daneshpour, M. S. (2022). Evaluating machine learning-powered classification algorithms which utilize variants in the GCKR gene to predict metabolic syndrome: Tehran Cardio-metabolic Genetics Study. Journal of Translational Medicine, 20(1), 1–12. https://doi.org/10.1186/s12967-022-03349-z

[3]. Baddam, S., Bingi, P., & Shuva, S. (2019). Student Evaluation of Teaching in

Business Education: Discovering Student Sentiments Using Text Mining Techniques. E-Journal of Business Education & Scholarship of Teaching, 13(3), 1–13. http://www.ejbest.org

[4]. Brown, G. (2010). Ensemble Learning Motivation and Background. 1–24. http://www.cs.man.ac.uk/~gbrown/research/brown10ensemblelearning.pdf

[5]. Kalcheva, N., Todorova, M., & Marinova, G. (2020). NAIVE BAYES CLASSIFIER, DECISION TREE AND ADABOOST ENSEMBLE ALGORITHM – ADVANTAGES AND DISADVANTAGES. 6th ERAZ Conference Proceedings (Part of ERAZ Conference Collection), 153–157. https://doi.org/10.31410/eraz.2020.153

[6]. Kastrati, Z., Dalipi, F., Imran, A. S., Nuci, K. P., & Wani, M. A. (2021). Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. In Applied Sciences (Switzerland) (Vol. 11, Issue 9, p. 3986). Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/app11093986

[7]. Lee, E., Rustam, F., Washington, P. B., Barakaz, F. El, Aljedaani, W., & Ashraf, I. (2022). Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. IEEE Access, 10, 9717–9728. https://doi.org/10.1109/ACCESS.2022.3144266

[8]. M.O., A., A.O., O., & W.F., W. (2016). Teachers' Performance Evaluation in Higher Educational Institution using Data Mining Technique. International Journal of Applied Information Systems, 10(7), 10–15. https://doi.org/10.5120/ijais2016451524

[9]. Malik, H., Fatema, N., & Iqbal, A. (2021). Intelligent Data Analytics for Transmission Line Fault Diagnosis Using EEMD-Based Multiclass SVM and PSVM. In Intelligent Data-Analytics for Condition Monitoring (pp. 115–140). Elsevier. https://doi.org/10.1016/b978-0-323-85510-5.00006-5

[10]. Malviya, S., Tiwari, A. K., Srivastava, R., & Tiwari, V. (2020). Machine Learning Techniques for Sentiment Analysis: A Review. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 12(02), 72–78. www.ijmse.org

[11]. Necati Demir. (2016). Ensemble Methods in Machine Learning | Toptal. Toptal. https://www.toptal.com/machine-learning/ensemble-methods-machine-learning

[12]. Nguyen, L. (2016). Tutorial on Support Vector Machine. Special Issue "Some Novel Algorithms for Global Optimization and Relevant Subjects", Applied and Computational Mathematics (ACM), 6(4–1), 1–15. https://doi.org/10.11648/j.acm.s.2017060401.11

[13]. Noble, W. S. (2006). What is a support vector machine? In Nature Biotechnology (Vol. 24, Issue 12, pp. 1565–1567). https://doi.org/10.1038/nbt1206-1565

[14]. Peng, H., Zhang, Z., & Liu, H. (2022). A Sentiment Analysis Method for Teaching Evaluation Texts Using Attention Mechanism Combined with CNN-BLSTM Model. Scientific Programming, 2022. https://doi.org/10.1155/2022/8496151

[15]. Qiang, G. (2010). An effective algorithm for improving the performance of Naive Bayes for text classification. 2nd International Conference on Computer Research and Development, ICCRD 2010, 699–701. https://doi.org/10.1109/ICCRD.2010.160

[16]. Reinhardt, E., & Beu, F. A. (2015). An introduction to education. In An introduction to education. https://doi.org/10.1037/14681-000

[17]. Shobha, G., & Rangaswamy, S. (2018). Machine Learning. In Handbook of Statistics (Vol. 38, pp. 197–228). Elsevier B.V. https://doi.org/10.1016/bs.host.2018.07.004

[18]. T. Manjunath Kumar, R. M. (2019). Predicting Faculty Performance in Higher Education using Machine Learning. International Journal of Recent Technology and Engineering, 8(4), 9472–9478. https://doi.org/10.35940/ijrte.d9750.118419

[19]. Vaibhaw, Sarraf, J., & Pattnaik, P. K. (2020). Brain-computer interfaces and their applications. In An Industrial IoT Approach for Pharmaceutical Industry Growth: Volume 2 (pp. 31–54). Elsevier. https://doi.org/10.1016/B978-0-12-

821326-1.00002-4

[20]. Vijayalakshmi, V., Panimalar, K., & Janarthanan, S. (2020). Predicting the performance of instructors using Machine learning algorithms. 26(12). https://www.researchgate.net/publication/347935410_Predicting_the_performance_of_instructors_using_Machine_learning_algorithms

[21]. Wu, G., & Feng, T. T. (2015). A theoretical contribution to the fast implementation of null linear discriminant analysis with random matrix multiplication. Numerical Linear Algebra with Applications, 22(6), 1180–1188. https://doi.org/10.1002/nla.1990

[22]. Xia, X., & Yan, J. (2021). Construction of Music Teaching Evaluation Model Based on Weighted Naïve Bayes. Scientific Programming, 2021. https://doi.org/10.1155/2021/7196197