***A***

***PROJECT WORK REPORT ON***

**“Assessing the Behavioural Features and Predict the Academic Performance of Student using Data Mining and Machine Learning Algorithms”**

***This project report submitted towards partial fulfilment for the requirement for the degree of***

**MASTER OF BUSINESS ADMINISTRATION 2022-2024**

****

**BIJU PATNAIK UNIVERSITY OF TECHONOLOGY, ODISHA SUBMITTED BY**

***AMAN MOHANTY***

**2206260002**

**MBA 4th Semester (2022 - 2024 Batch)**

**UNDER THE GUIDANCE OF**

**Dr. SREEKUMAR**

**Director I/c (R&D)**

**ROURKELA INSTITUTE OF MANAGEMENT STUDIES**

**INSTITUTIONAL, AREA, GOPABANDHU NAGAR, CHHEND COLONY ROURKELA: 769015, ODISHA**

**DECLARATION**

I **Aman Mohanty** hereby declare that this project titled **“Assessing the Behavioural Features and Predict the Academic Performance of Student using Data mining and Machine learning Algorithms”** is submitted to the Biju Patnaik University of Technology as a partial requirement for the award of Degree of Master of Business Administration, of Rourkela Institute of Management Studies during the year 2022-2024. It is a research paper project carried out by me, under the supervision of my internal guide **Dr. Sreekumar**, faculty member of Rourkela Institute of Management Studies. This project report has not been submitted earlier by me or by anybody else for the award of any other degree in any University in India or abroad.

**Aman Mohanty**

Registration Number- 2206260002

**ACKNOWLEDGMENT**

First and foremost, I thank the God for his substantial blessing and mercy at all stages in the completion of my project.

I express my sincere gratitude and indebtedness to **Dr.Subrat Prasad Pattnayak, HOD (MBA), RIMS** for helping me find out this opportunity.

I extended my immense gratitude to the faculty guide **Dr.Sreekumar, RIMS** for his initiative in this field of research, for his valuable guidance, sympathies and kind attitude always encourage me to carry out the present work firmly. I thank him for overall supports.

Aman Mohanty

Date

**INTERNAL GUIDE CERTIFICATE**

**This is to certify that this project report entitled “Assessing the Behavioural Features and Predict the Academic Performance of Student using Data mining and Machine learning Algorithms” submitted to Rourkela Institute of Management Studies is a bonafide record of documents done by Aman mohanty, MBA 4th Semester (2022- 24), having Regd. No. – 2206260002**

**Starting Date- 20 February 2024**

**Ending Date- 16 March 2024**

**Signature**

**Dr. Sreekumar RIMS, Rourkela**

**CONTENT**

**Declaration I**

**Acknowledgement II**

**Internal guide certificate III**

|  |  |  |
| --- | --- | --- |
| **Chapter** | **Particulars** | **Page No.** |
| **1** | **Abstract** | **1** |
| **1.0** | **Introduction** | **1-2** |
| **2.0** | **Literature Review**  **2.0.** E-Learning in Education  **2.1.** Educational Data Mining | **2-4** |
| **3** | **Dataset description** | **4-6** |
| **4** | **Research Methodology**  **4.1.** Research design  **4.2.** Classification field  **4.3.** Methods | **6-8** |
| **5** | **Results and Discussion**  **5.1.** Pre-processing  **5.2.** Data visualisation  **5.3.** Summary of results | **9-19** |
| **6** | **Conclusion and Future work** | **19** |
| **7** | **References** | **20-21** |
| **8** | **Power BI Dashboard** | **22** |

**Assessing the Behavioural Features and Predict the Academic Performance of Student using Data mining and Machine learning Algorithms**

**Abstract:**

Creating learning environments, where students, parents, and teachers are linked to a learning process, helps study their overall impact on the students’ performance. Data mining and machine can analyse these inter-relationships and thus enable the prediction of academic performance to improve the student’s academic level. The main factors that affect the student’s performance were selected using feature selection methods. An analysis of the crucial features was investigated to better understand the data. One of the main outcomes found is the impact of the behavioural features on the students’ academic performance. Moreover, gender and relation demographical features are another important features found. It was evident that there is an academic disparity between genders, as females constitute the most outstanding students. Furthermore, mothers have a clear role in student academic excellence. The data mining and machine learning methods were used and tested to predict the student’s performance, namely random forest, logistic regression, Decision tree,Support vector machine, MLP, and ensemble learning using bagging and voting. From the results, it was proven that Support vector machine algorithm is most appropriate for predicting student academic performance. Support vector machine gives 70.8% prediction which is relatively higher than other algorithms. The results show that data mining can accurately predict the students’ performance level, as well as highlight the most influential features.

**Keywords:** Educational Data Mining, Machine Learning, Deep Learning, Learning Analytics.

1. **Introduction:**

Building learning environments, where students have an active interaction in their learning, has become a priority for educational institutions. Learning Analytics (LA) is an emerging area for the collection, analysis, and presentation of learners' data for the purpose of studying the influencing factors on the learning process with the aim of understanding and developing the learning environment. Sin and Muthu (2015) LA provides all parties (parents, teachers, and students) with the appropriate and quick feedback about the educational process. On the other hand, behaviour analytics helps us understand the behaviour of students and how they interact during the learning period with contributing influences.

Predicting the performance of students has attracted the attention of several authors due to its importance in helping teachers identify and support their students according to the level of difficulty. Hussain, Zhu, Zhang, Muhammad, Abidi, and Ali (2018). Considerable works have been done in recent years to analyse the behaviour of students and extract the significant patterns that can be used to predict the students’ performance.

This study intends to use data mining methods to: (i) find the strongest features that can help in the prediction of students’ performance, (ii) analyse the most important behaviour and demographical features to have a better understanding of the features that affect the students’ performance level, (iii) predict the students’ performance by using data mining techniques and show how feature selection, oversampling, ensemble learning, and parameter tuning can enhance the predictive power of the models and resolve over fitting. A summary of the previous work has been presented in the literature review section. Explanation of the data used and the data mining methods applied in this research have been discussed in the data and methodology sections. In the results section, we show the best selected features, provide a detailed analysis of the main selected features, and evaluate and discuss the results of our prediction models. We finally conclude our research in the conclusion section.

**2. Literature Review**

**2.1 E-Learning in Education**

Online-based learning environments, such as learning management systems (LMSs), allow teachers to study and track students' performance by recording and keeping student information online. Hussain, Zhu, Zhang, Muhammad, Abidi, and Ali (2018). E-Learning environments have been used to monitor and record all educational processes and actions done by students, thus it could provide useful information on the progress that each student achieves as mentioned in .Romero and Ventura (2007). In order to achieve the best level of electronic learning, it is necessary to evaluate the processes of learning and teaching continuously by observing all aspects, from the level of interaction between the parties involved in the quality of teaching through the reactions of students and their initiative. In addition to the use of multiple sources, the effect of management, and other aspects, on the development of cognitive skills, should be considered .Rodrigues, Isotani and Zárate (2018).

In recent years, there has been a significant growth of using methods to facilitate the analysis of existing processes related to learners and e-learning systems. In order to provide a more efficient learning environment, data mining techniques have been used in this field, for processing the data and extracting information patterns that can be useful indicators. Romero and Ventura (2007).

**2.2 Educational Data Mining**

Data mining in the educational analysis is described as the process of automatic extraction of a meaningful chain from a large dataset. It is used not only to train the model on the learning process but also to evaluate and develop e-learning systems. Romero and Ventura (2007).

The efficiency of machine learning methods has been analysed in predicting the difficulty the students will face in the next session to support the students and help them according to the level of difficulty. Five of the well-known machine learning algorithms have been used for the prediction process, namely artificial neural network (ANN), logistic regression (LR), naïve Bayes (NB), support vector machine (SVM), and decision tree (DT). The methods have been selected based on their suitability for the dataset and insensitivity towards over fitting. For evaluation, authors in e.g. it used multiple techniques such as root-mean-square error (RMSE) and Cohen’s kappa coefficient. Feature selection techniques focus on reducing the dimensionality and avoid unrelated data to the research interest .Alpha-investing feature selection for ranking was used to minimize the input features to be used in the student prediction model. Their results showed that SVMs and ANNs are the most suitable models to predict student’s performance .Hussain, Zhu, Zhang, Muhammad, Abidi, and Ali (2018).

Two datasets were analysed in .Costa, Fonseca, Santana, Araújo and Rego (2017) to predict students failing in the early stages, exactly after the first exam in introductory programming courses available from a Brazilian Public University. A noticeable result was shown by using the SVM algorithm and F-measure for evaluation. They claimed that female students had less in-class participation than male students. In addition to that, after the first exam application, the F-measure value reached approximately 0.92 and 0.83 with distance and on-campus datasets, respectively.

Other research Amrieh , Hamtini and Aljarah (2016) focused on a new side of educational analytics called behavioural features. The authors chose ANN, DT, NB and ensemble models (bagging, boosting, and random forest) to build a performance predictive model. The information gain algorithm has been used to build the students’ performance model. The results were presented with and without behavioural characteristics using 10-fold cross-validation. The result showed how behavioural characteristics had a strong effect on the students’ academic achievement. In addition, the ANN technique overcomes other methods, as its accuracy was around 79.1%. The result and quality of the classifiers were measured by four common measures; Accuracy, Precision, Recall, and F-2-1Measure. Amrieh, Hamtini and Aljarah (2016) .The lack in this paper is that it did not predict early enough student performance due to the lack of information about midterm exams and assignments, the prediction was after the finals.

Using data mining to predict the student’s dropouts was explored with a dataset of 165, 715 high school students from different schools Chung and Lee (2018) .The authors selected significant features that were presented by using the random forests model with out of bag (OOB) estimate. The unauthorized absence was the most significant variable in predicting students' dropouts, followed by unauthorized lateness. The random forests model predicts students' dropouts with a high accuracy of 95% using 10 folds cross-validation. AUC score got 97% which represents an outstanding performance. The work in this article was excellent, but they noted some shortcomings with the calculation of the model features, that used inaccuracy weights. Chung and Lee (2018).

The researches which were discussed in this section varied by using data mining techniques to predict at-risk students. Findings cannot be generalized due to their limited domain, but their work is worthy of praise. The following section concentrates on one main stage in data analysis which is feature selection and its effect on the predicted results and data visualization and its importance in better understanding the data.

**3. Dataset description**

A LMS called Kalboard 360 has been used to collect educational dataset. This system gives users (students, teachers and parents) synchronous access to reach the educational resources from all devices by using an internet connection. The original source of the dataset is found in Amrieh, Hamtini and Aljarah (2016). The dataset is available on Kaggle.com under the name of BStudents’ Academic Performance Dataset. In total 480 students with 16 features are analysed in this project which can be divided into four basic categories. Details of these features and their number of instances have been presented in Table 1.

**Table 1**. Features’ description of the dataset used in this study adapted from Chung and Lee (2018).

|  |  |  |  |
| --- | --- | --- | --- |
| Features  Category | Feature | Description | Number of  Instances |
| Demographical  Features | Gender | Male/Female | 2 |
| Nationality | Kuwait/Lebanon/Saudi Arabia etc. | 14 |
| Place of birth | Kuwait/Lebanon/Saudi Arabia etc. | 14 |
| Relation | Parent responsible for student(Mother/father) | 2 |
| Academic  Background  Features | Educational Stages | Lower level /Middle School /High School | 3 |
| Grade Levels | G-01/G-02/……. /G-12 | 10 |
| Section ID | Classroom student belongs to (A/B/C) | 3 |
| Topic | Course (English/Spanish/French/IT  etc. | 13 |
| Student Absence  Days | above-7/under-7 | 2 |
| Semester | First/Second | 2 |
| Behavioural  Features | Raised hand | How many times the student raises his/her hand on classroom (numeric:0-100) | 101 |
| Visited resources | How many times the student visits a course content (numeric:0-100) | 101 |
| Viewing  announcements | How many times the student checks the new announcements (numeric:0-100) | 101 |
| Discussion groups | How many times the student participate on discussion groups (numeric:0-100) | 101 |
| Parents  Participation  on learning  process | Parent Answering  Survey | Yes/No | 2 |
| Parent School  Satisfaction | Yes/No | 2 |

The data set consists a total of 305 males and 175 females where 179 are from Kuwait, 28 are from Palestine, 172 are from Jordan, 22 are from Iraq, 17 are from Lebanon, 12 are from Tunis, 11 are from Saudi Arabia, 9 are from Egypt, 6 are from USA, 7 are from Syria, 6 are from Iran, 6 are from Libya, 4 are from Morocco and 1is from Venezuela. The data set is collected in two academic semesters, on first semester they were 245 student records and 235 student records on semester two. The data set also covers the features of the school attendance days. There are two categories based on their number of absence days. We found that there were 289 students their absence days were under 7 days and 191 students were absent for more than 7 days. This data set also contains parent’s participation on their children academic process. There are two categories: first being the parent answering survey and secondly being parent school satis- faction. We found that 270 parents managed to answer the survey and 210 parents did not answer the survey. We also found that 292 parents are happy and satisfied with the school and 188 parents are not satisfied.

The class feature contains the performance level which is the total mark of the student in a subject decided in each record. This performance level is categorized into three levels (High, Medium, and Low). Marks below 70 are belonging to the low level, marks between 70 and 89 are belonging to the medium, and marks higher than 89 considered as a high level. Amrieh, Hamtini and Aljarah (2016)

**4. Research Methodology**

This chapter presents the methods that is proposed in this study to predict performance of students using student’s demographical Information and student’s online logging’s data.

**4.1 Research design**

Based on the literature the study proposed using data mining and machine learning techniques to find whether the total spent time on online learning affect the student’s performance and how have learning management system features affect the student performance.

**4.2 Classification field**

There are three numerical intervals of student grades. The first interval is for students who obtained a failing percentage (L), the interval includes values from 0% to 69%. The second interval is for students who obtained low passing percentage (M), the interval includes values from 70% to 89%. Lastly is for students who got high passing percentage (H), the interval includes values from 90% to 100%.

**4.3 Methods**

Predictive Models: The popular classification methods (Decision trees and Perceptron classification and Support vector machines, Logistic regression and Random forest) are built and compared to other making use of their predictive accuracy on the given data samples. Brief description of the predictive models that will be used in this study.

Support Vector Machines: Support vector machines (SVMs). hua and sun(2001) helps in detecting the outliers on the data set and it also perform classification. SVMs are set of supervised learning methods. SVMs used kernel trick to modify data and use the modified data to find the difference between the possible end results. SVM finds the optimal solution by computing on each feature by using partial differentiation after employing the Lagrange multiplier Burman and Som(2019).The model decreases convolution of the training data consequential subset of support vectors. Given a data set containing a training set of N data points, {xk, yk}N k=1 and input data, which is an n-dimensional data vector (x\_k∈R∧ N) and output, which is the one-dimensional vector space (y\_k∈ r); SVM create the classifier as shown below in this equation.

Where αk are positive real constants and b is a real constant.

Decision trees (DTs): In this study Decision trees is used to find the predictor variables to the predicted variable and shows the targeted discrete value. "Decision trees uses variable values to create a structure that has nodes and edges”. Esposito, Malerba, Semeraro, and Kay(1997). A DT has internal nodes and leaves, rectangles represents nodes and ovals represents leaves. Data set features are represented by the internal node and it contain two or more child. The value of these features is found at the branches. Each leaf contains a classification label. Jauhari and Supianto(2012). Decision trees are established from a training set. A tree is called the hierarchy and a node is called segment. The entire data set is contained at the original segment called the node of the tree. The branches are formed by the node with its successors that created it and the leaves are final nodes. The decision is made on each leaf and it is applied to all the observations in the leaf. The decision is the predicted value.

Perceptron classification: The Perceptron classifier. Chaudhuri and Bhattacharya (2000) is a set of supervised learning, the classification field of a sample can be predicted using Perceptron classifier. Perceptron classifier accept numerous input and if the number of inputs is more than the specified condition, it does not return the output, it output the massage for corrections. In the Perceptron algorithm features on the data set are taken as inputs and it is represented by x1, x2, x3, x4,...,xn where features value is indicated by x and the total occurrences is represented by n. The required features to be trained is stored as input in the first layer. Now the total inputs and weights will be multiplied and add their outcome. The weights are the values obtained through the training of the model and are denoted by w1, w2, w3... wn. The output function will be shifted by the bias value and this value will later be presented to the activation function then the output value is obtained after receiving the value on the last step.

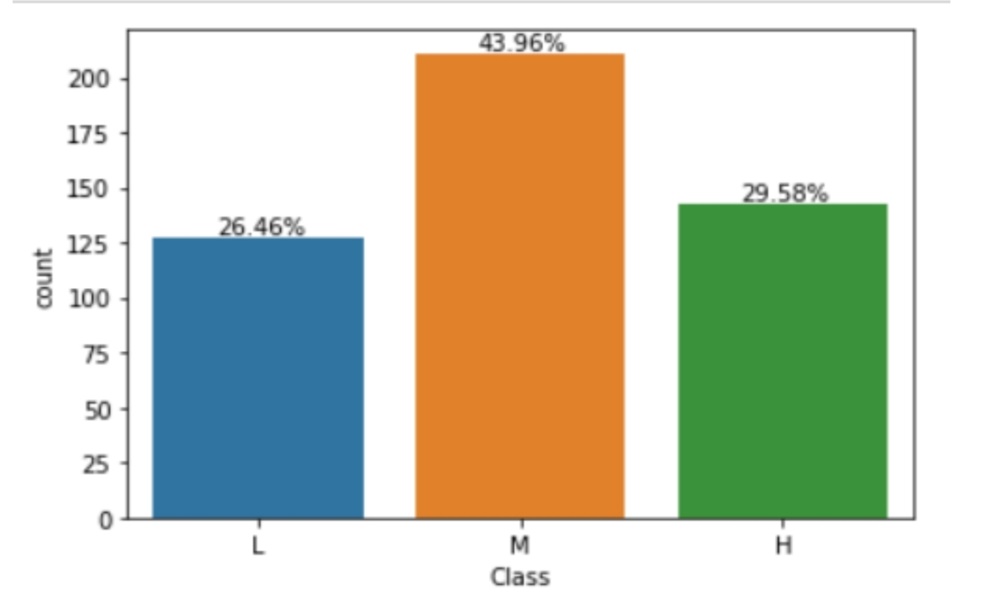
Logistic regression (LR): Logistic regression describe the association among variables and it was used to predict student academic performance by estimating the probability of an event occurring. Hosmer, Lemeshow, and Sturdivant (2013). It also shows the probability of two categories by fitting the explanatory variables and log odds to model using this equation.

Where Y= (0, 1) is the binary variable; 1 if it is higher than the reference level and 0 if not, X = (X1... Xn) are n explanatory variables and β = (β0.....βn) are the estimated regression coefficient.

**Random forest (RF):** Random forest uses begging method to generate trees in which its prediction is more accurate than that of any individual tree. Liaw and Wiener (2002). Random forest was also used to avoid over fitting on the training set and limiting errors due to bias hence yield accurate and useful results. RF can handle outliers and noise in the data and gains high classification accuracy. RF generate numerous decision trees in the training phase and output class labels. Deepika and Sathyanarayana (

2019). RF is used in this study since it is permissible to less over-fitting and it has proofed to be good classification results previously.

RF is a theoretical framework grounded on mixture of decision trees; {T1 X... TBX}. The ensemble produces B outputs {ˇY1 = T1(X)... ˇYB = TB(X), where ˇYb, b = 1... B is the predicted grades by the bth tree. Output of all trees are aggregated to produce one final prediction ˇY, which is the class predicted by majority of trees



**Figure 1**: This figure depicts the proportion of student academic performance of the classification field.

**5. Results and Discussion**

The result and performance of the model depend on the techniques used in the pre-processing. The original data is converted into a form that is suitable for use with data mining such as data cleaning and data conversion. Hussain, Zhu, Zhang, Muhammad, Abidi, and Ali (2018) .The whole 480 records in the dataset are clean from any missing values, and outliers. The class feature has three values, High, Medium, and low levels which contain 141, 211, and 127 cases, respectively. This slight difference in the number of cases is not considered as imbalance dataset Chawla, Bowyer, Hall and Kegelmeyer (2002). Normalization is applied to prevent misleading variance between features’ values. After these steps, the dataset is ready to be used with classification methods.

**5.1 Pre-processing**

Typically in machine learning and data mining before processing and running a test on a data set, it is necessary to prepare the data and select the targeted attribute. Selecting attributes requires putting all the matching combination of attributes in the data set in order to find which combination is suitable in predicting student academic performance.

Our goal with pre-processing was to change our numerical fields that have a value like Grade ID to a numerical only value in a way that we preserve that distance in a meaningful way. We also assign our three classes to numerical outcomes with a preserved distance and set setting L = -1, M = 0, and H = 1. We chose to preserve the distance between the categorical values and scale our numerical fields so that they would be more meaningful when compared together.

The five machine learning and data mining models are used to evaluate the student’s academic performance and to check which model best predict students’ performance.

**5.2 Data visualisation**

The data set consist a total of 480 records. In this study, the purpose of selecting an attributes was to find the attributes that contain numerical data, attributes that contain categorical data and classification label. Our goal with data visualisation is to get an idea of the shape of the data set and to see if we can easily identify any possible outliers and also look to see if any of data is unclear or redundant.

**5.3 Summary of results**

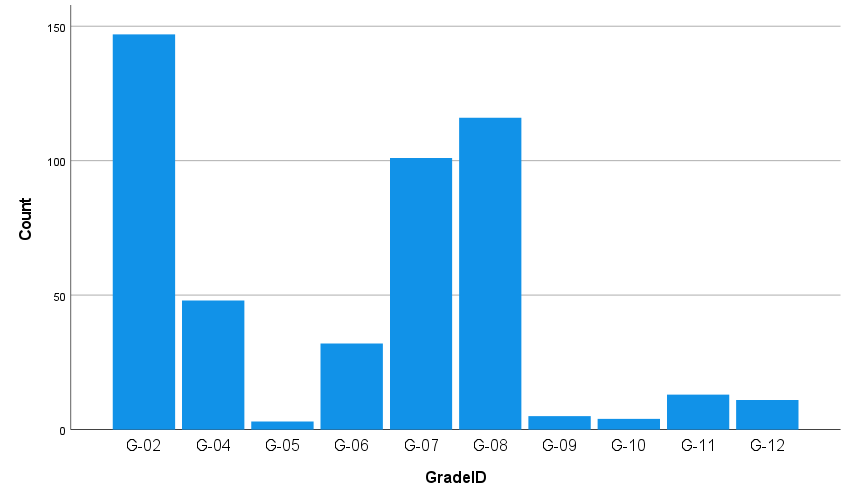
In the previous section, student academic performance was discussed and how it will be evaluated. In this section we discuss the performance of five machine learning models that was specified in section 4. First of all, we performed data visualisation after performing data pre-processing by generating simple plots of data distributions to get an idea of the shape of the data set and then 5 main machine learning and data mining techniques was evaluated and also describe the variables contained on the data set. The evaluation was done on the full data set that consist of 480 features. Sievert (2020)

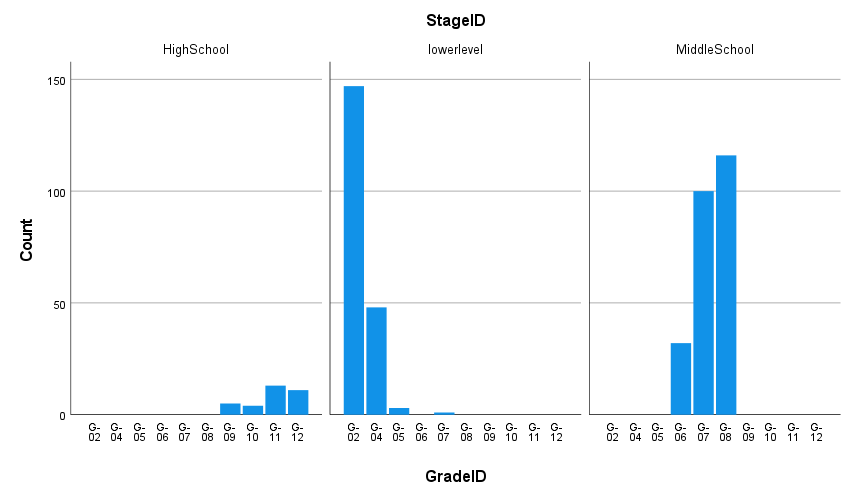
From the results, it was proven that Support vector machine algorithm is most suitable in predicting the performance of students. SVM is relative higher than other algorithms and it has 70.8% prediction accuracy, followed by Random forest with 69.7% accuracy, Logistic regression with 67.7% accuracy, Perceptron with 64.5% and lastly Decision tree with 46.8%. We found the overall percentage of passing rate using class variable. There were 26.46% students who got a failing percent (less than 69%), 43.96% students who got a low passing grades (between 70% and 89%) and 29.58% who achieved high marks in their course (90% to 100%).

Student absence days seems to have a strong correlation with class variable. Very few students who missed more than 7 days managed to achieve high marks and very few students who missed less than 7 days failed their course. From grade 2 to grade 12 we found that grade 5, 9, and grade 10 have very few counts. No 5th grade students pass and no 9th grade students achieve high marks.

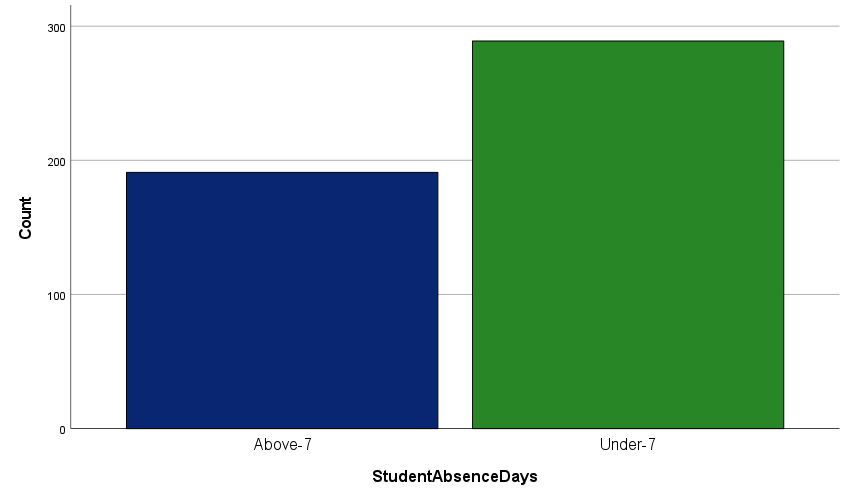
In figure 4, the bar plot shows the accuracy of five popular machine learning models that is used to evaluate the student performance. The legend of the plot indicate that blue colour is Support vector machine, Yellow colour indicate Logistic regression, orange colour indicate Decision tree, light blue indicate Random forest and grey colour indicate Perceptron classifier.

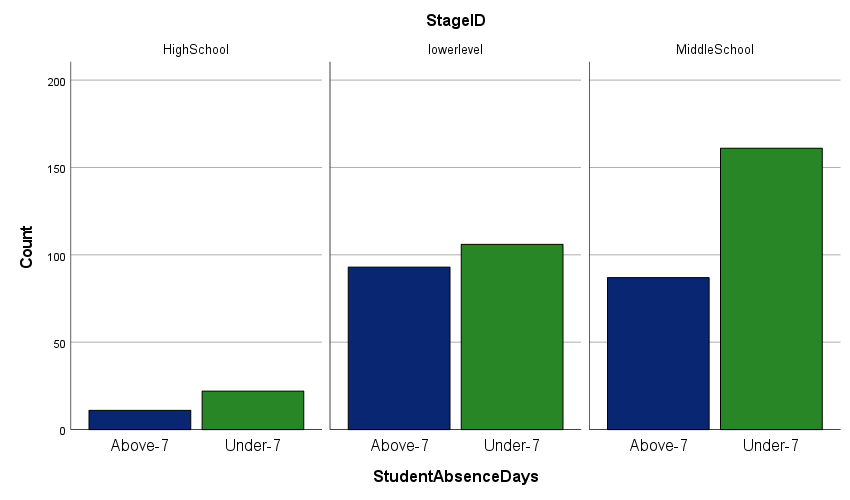
Support vector machine has performed well when compared to other machine models. 78.75% instances was correctly classified and 21.25% instances was incorrectly classified. Another way of representing accuracy of the machine learning models that are in figure 4 is through confusion matrices. Table 2 describe how each algorithm has performed.





**Figure 2**: This figure depicts the grade the student is in and it also shows the average student performance for each grade.

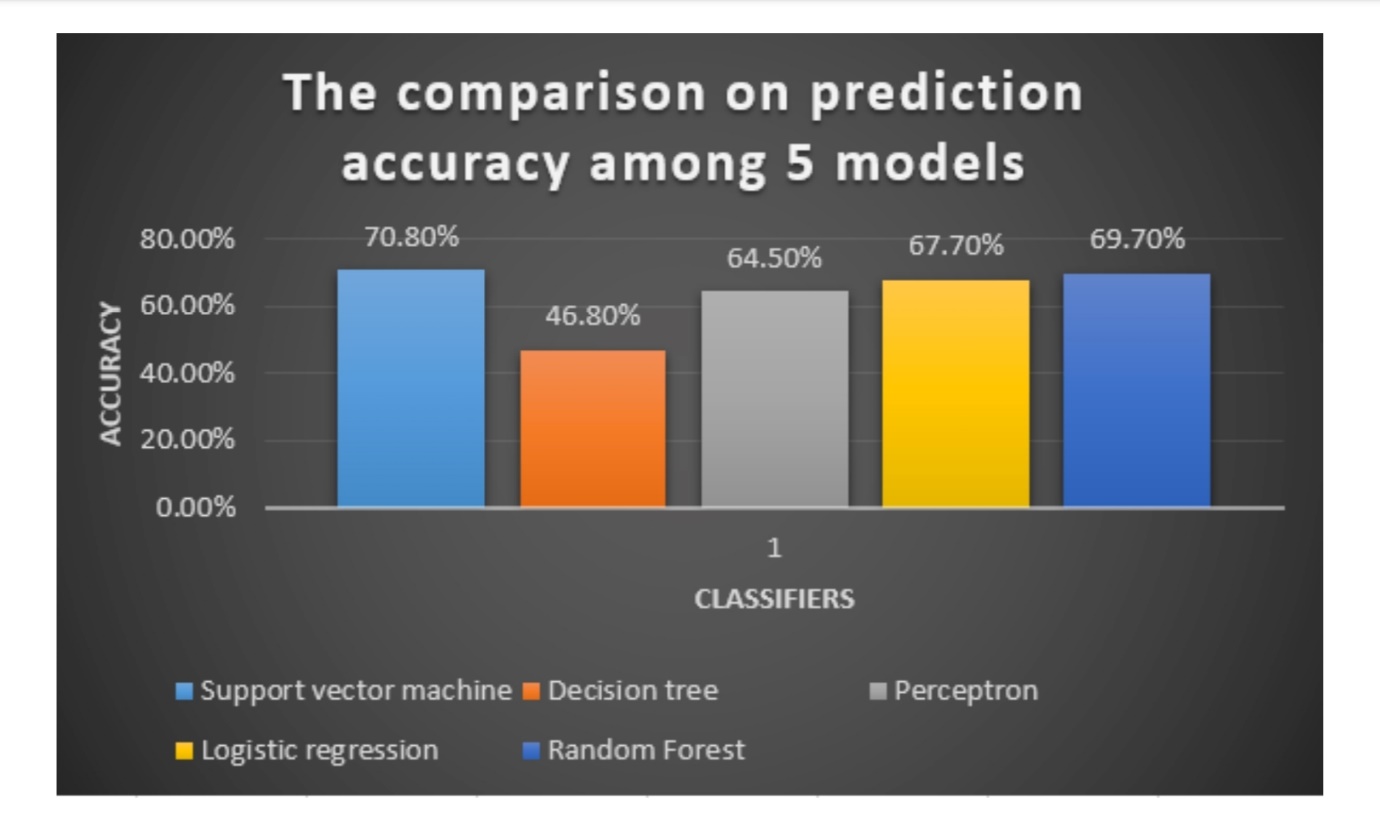




**Figure 3**: This figure depicts the proportion of student absence days and it also shows whether or not a student was absent for more than 7 days.

**Table 2**: This table shows the confusion matrices that describe the performance of the classification models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Support vector machine** | | | |
| **Actual Values** | **Predicted Values** | | |
| **High** | **Low** | **Medium** |
| **High** | **23** | **1** | **11** |
| **Low** | **0** | **19** | **3** |
| **Medium** | **5** | **8** | **26** |
| **Decision tree** | | | |
| **Actual Values** | **Predicted Values** | | |
| **High** | **Low** | **Medium** |
| **High** | **20** | **1** | **14** |
| **Low** | **1** | **15** | **6** |
| **Medium** | **7** | **8** | **24** |
| **Perceptron** | | | |
| **Actual Values** | **Predicted Values** | | |
| **High** | **Low** | **Medium** |
| **High** | **8** | **4** | **25** |
| **Low** | **0** | **21** | **1** |
| **Medium** | **2** | **19** | **18** |
| **Logistic Regression** | | | |
| **Actual Values** | **Predicted Values** | | |
| **High** | **Low** | **Medium** |
| **High** | **8** | **4** | **25** |
| **Low** | **0** | **21** | **1** |
| **Medium** | **2** | **19** | **18** |
| **Random Forest** | | | |
| **Actual Values** | **Predicted Values** | | |
| **High** | **Low** | **Medium** |
| **High** | **21** | **1** | **12** |
| **Low** | **0** | **16** | **6** |
| **medium** | **4** | **6** | **29** |



**Figure 4**: The comparison on prediction accuracy among 5 models.

**Table 3**: This table depicts the machine learning algorithms used in this study and their performance accuracy.

|  |  |
| --- | --- |
| Classifier | Accuracy |
| Support vector machine | 70.8% |
| Decision tree | 46.8% |
| Perceptron | 64.5% |
| Logistic regression | 67.7% |
| Random forest | 69.7% |

**Table 4**: This table depicts detailed accuracy of Logistic regression.

|  |  |
| --- | --- |
| Features | Weighted Average |
| Correctly classified Instances | 73.5% |
| Incorrectly classified Instances | 26.25% |
| Mean absolute error | 21.33% |
| Root mean squared error | 37.53% |
| Relative absolute error | 49.27% |
| Root relative squared error | 80.6% |
| Precision | 73.8% |
| Recall | 73.8% |
| F-Measure | 73.8% |
| Roc Area | 83.8% |
| Total number of Instances | 480 |

Table 5: This table depicts detailed accuracy of Perceptron.

|  |  |
| --- | --- |
| Features | Weighted Average |
| Correctly classified Instances | 79.37% |
| Incorrectly classified Instances | 20.62% |
| Mean absolute error | 14.88% |
| Root mean squared error | 34.59% |
| Relative absolute error | 34.36 |
| Root relative squared error | 74.35% |
| Precision | 79.3% |
| Recall | 79.4% |
| F-Measure | 79.3% |
| Roc Area | 89.3% |
| Total number of Instances | 480 |

Table 6: This table depicts detailed accuracy of Random Forest.

|  |  |
| --- | --- |
| Features | Weighted Average |
| Correctly classified Instances | 76.66% |
| Incorrectly classified Instances | 23.33% |
| Mean absolute error | 24.28% |
| Root mean squared error | 33.37% |
| Relative absolute error | 56.09% |
| Root relative squared error | 71.73% |
| Precision | 76.6% |
| Recall | 76.7% |
| F-Measure | 76.6% |
| Roc Area | 89.7% |
| Total number of Instances | 480 |

Table 7: This table depicts detailed accuracy of Decision tree

|  |  |
| --- | --- |
| Features | Weighted Average |
| Correctly classified Instances | 72.70% |
| Incorrectly classified Instances | 27.29% |
| Mean absolute error | 29.53% |
| Root mean squared error | 37.18% |
| Relative absolute error | 68.19% |
| Root relative squared error | 79.92% |
| Precision | 72.8% |
| Recall | 72.7% |
| F-Measure | 72.7% |
| Roc Area | 84.2% |
| Total number of Instances | 480 |

Table 8: This table depicts detailed accuracy of Support Vector Machine.

|  |  |
| --- | --- |
| Features | Weighted Average |
| Correctly classified Instances | 78.75% |
| Incorrectly classified Instances | 21.25% |
| Mean absolute error | 27.22% |
| Root mean squared error | 35.22% |
| Relative absolute error | 62.78% |
| Root relative squared error | 75.71% |
| Precision | 78.8% |
| Recall | 78.8% |
| F-Measure | 78.7% |
| Roc Area | 86% |
| Total number of Instances | 480 |

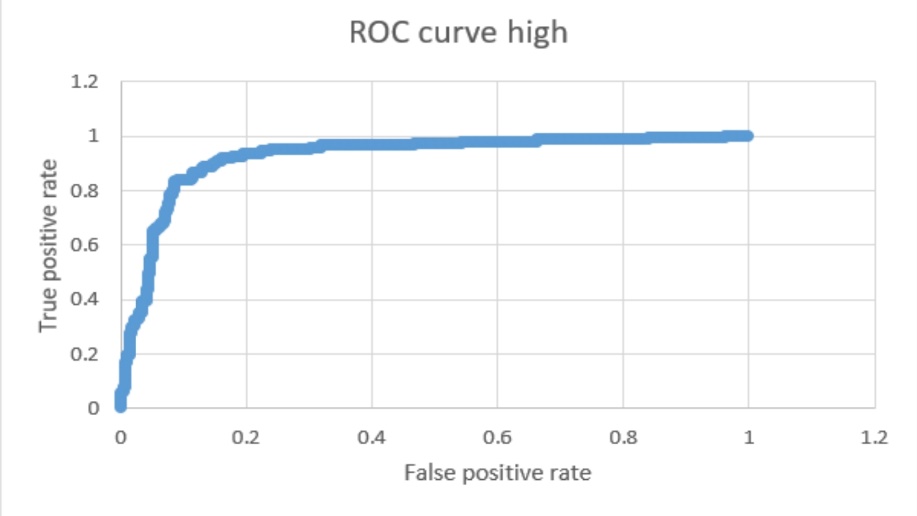


Figure 5: Receiver operating characteristic curve (class high).

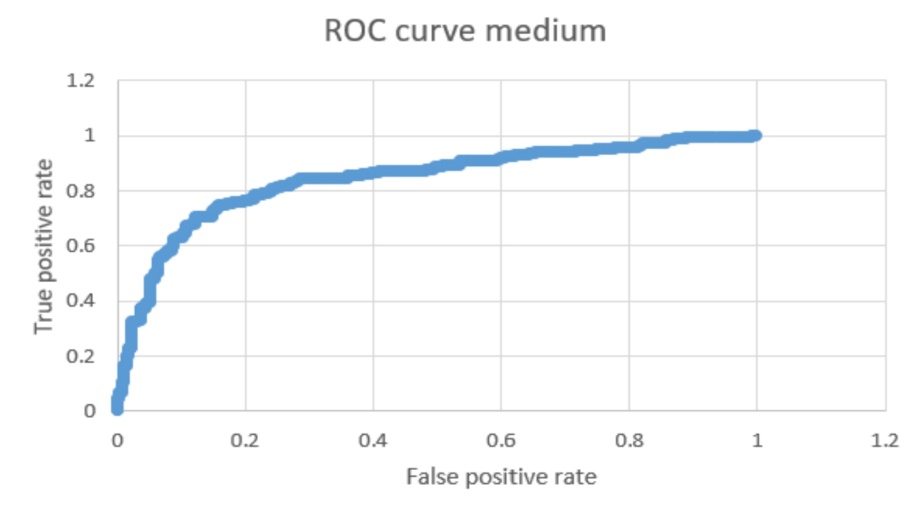


Figure 5: Receiver operating characteristic curve (class medium).

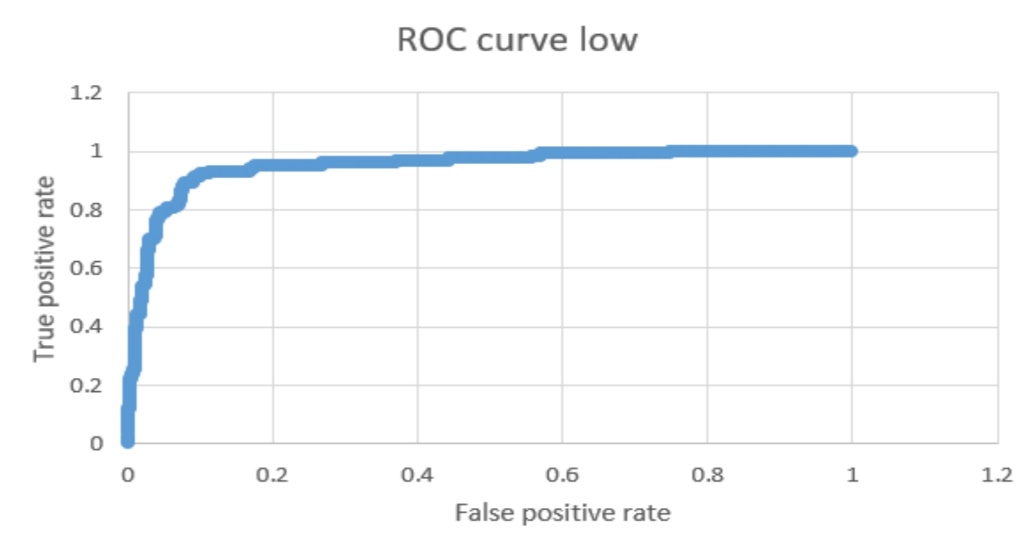


Figure 5: Receiver operating characteristic curve (class low).

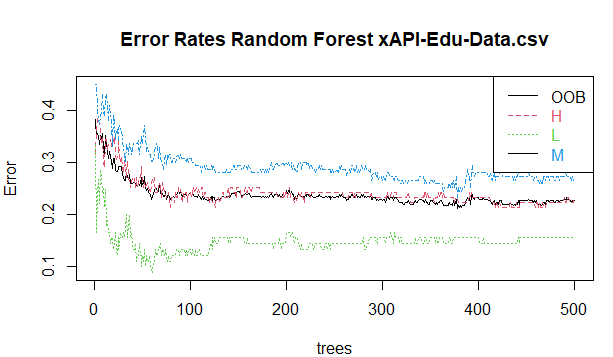
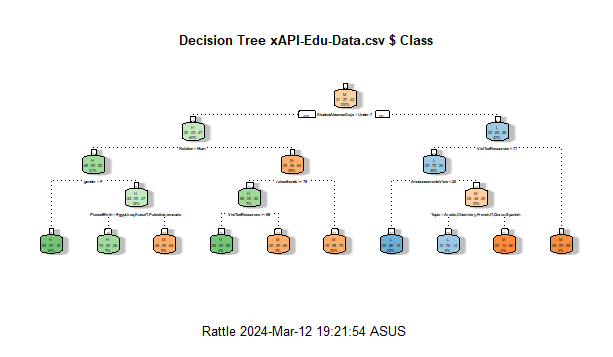


Figure 6: Error rates random forest



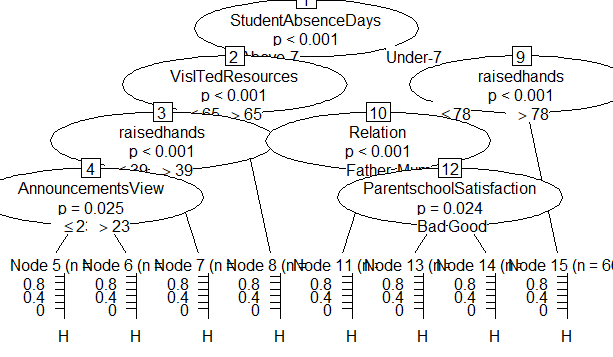


Figure 7: Show the decision tree

The graphical representation of receiver operating characteristic curve shows the performance of our best classification model,Support vector machine at all classification thresholds . Kannan and Vasanthi(2019). The ROC curve figure 5, figure 6 and figure 7 shows the classification performance of our best classifiers which is Support vector machine. There are three numerical intervals of student grades (low, medium and high). As can be seen from figure 5, figure 6 and figure 7.The ROC curve occupies the upper left corner, which means the classifier (SVM) used in this paper indicate a better performance and the prediction of positive value is specific in some degree, with AUC of 86%.

**6. Conclusion and Future work**

Companies and educational institutions uses learning management systems to create and manage lessons, courses, quizzes and other training materials. Hodges, Moore, Lockee, Trust and Bond(2020). Student’s success needs to be predicted to help an instructor identify academic performance and helps with identifying struggling students more easily and giving teachers a proactive chance to come up with supplementary resources to learners to improve their chances of increasing their grades. It may be difficult for students to learn virtually than in a traditional class hence the student’s performance vary due to difference methods of delivering the course materials . Niemi and Kousa(2020).

Various machine learning and data mining models were used to predict student success using the learning management system. Each model indicate different percentage of accuracy when is tested with different features that are associated in an online learning platforms. Hussain, Zhu, Zhang, Muhammad, Abidi, and Ali (2018).

Student’s performance was evaluated by five Data mining and machine learning techniques which is Perceptron classifier, Support vector machine and Decision trees, Logistic regression and Random forest. Support vector machine ends up handling the data the best with 70.8% accuracy. The obtained results shows that the student absence days influence student academic performance on the other hand student class grades is not influencing student academic performance.

I aim to extend the study by collecting more additional features such as encouraging and motivational strategies taken by facilitators and teachers and considering more materials available for students in an E-learning platforms. We will also consider features such as psychological factors available which affect student’s performance. We also intent to use more interesting and detailed data set to predict student academic performance in our future studies

**References**

1. Sin, K., Muthu, L. (2015), Application of Big Data in Education Data Mining and Learning Analytics – a Literature Review. SOFT COMPUTING MODELS FOR BIG DATA.6956, 1035–1050.
2. Hussain, M., Zhu, W., Zhang, W., Muhammad, S., Abidi, R., Ali, S(2018), Using Machine Learning to Predict Student Difficulties from Learning Session Data. Artif. Intell. Rev.
3. Romero, C., Ventura, S(2007), Educational Data Mining: a Survey from 1995 to 2005.Expert Syst. Appl. 33, 135–146 .
4. Rodrigues, M.W., Isotani, S., Zárate, L.E(2018), Educational Data Mining: a Review of Evaluation Process in the E-Learning. Telemat. Informatics. 35, 1701–1717.
5. Costa, E.B., Fonseca, B., Santana, M.A., De Araújo, F.F., Rego, J(2017), Evaluating the Effectiveness of Educational Data Mining Techniques for Early Prediction of Students’ Academic Failure in Introductory Programming Courses. Comput .Human Behav. 73, 247–256 .
6. Amrieh, E.A., Hamtini, T., Aljarah(2016), Mining Educational Data to Predict Student’s Academic Performance Using Ensemble Methods. Int. J. Database Theory Appl. 9, 119–136 .
7. Chung, J.Y., Lee, S(2018), Dropout Early Warning Systems for High School Students Using Machine Learning. Child. Youth Serv. Rev. 96, 346–353.
8. Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., Montenegro, M(2018), Centralized Student Performance Prediction in Large Courses Based on Low- Cost Variables in an Institutional Context. Internet High. Educ. 37, 76–89 .
9. Latham, A., Crockett, K., Mclean, D(2013), Profiling Student Learning Styles withMultilayer Perceptron Neural Networks. 2013 IEEE Int. Conf. Syst. Man, Cybern.2510–2515
10. Chawla, N. V, Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P(2002) SMOTE: Synthetic Minority Over-Sampling Technique SMOTE: Synthetic Minority Over-Sampling Technique.
11. S. Hua, Z. Sun (2001), “Support vector machine approach for protein subcellular localization prediction,” Bioinformatics, 17(8), 721–728.
12. I. Burman, S. Som(2019), “Predicting students’ academic performance using support vector machine,” in 2019 Amity International Conference on Artificial Intelligence (AICAI), 756–759, IEEE.
13. F. Esposito, D. Malerba, G. Semeraro, J. Kay (1997), “A comparative analysis of methods for pruning decision trees,” IEEE transactions on pattern analysis and machine intelligence, 19(5), 476–491.
14. F. Jauhari, A. A. Supianto(2012), “Building student’s performance decision tree classifier using boosting algorithm,” Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1298–1304, 2019, doi:10.1016/j.compedu.2012.08.015.
15. B. Chaudhuri, U. Bhattacharya(2000), “Efficient training and improved performance of multilayer perceptron in pattern classification,” Neurocomputing, 34(1-4),11–27, 2000
16. D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant (2013), Applied logistic regression,volume 398, John Wiley & Sons,.
17. A. Liaw, M. Wiener, et al(2002), “Classification and regression by randomForest,” Rnews, 2(3), 18–22, 2002.
18. K. Deepika, N. Sathyanarayana(2019), “Relief-F and Budget Tree Random Forest Based Feature Selection for Student Academic Performance Prediction,” International Journal of Intelligent Engineering and Systems, 12(1), 30–39, 2019,doi:10.1016/j.5896408.015.
19. C. Sievert(2020), Interactive Web-Based Data Visualization with R, plotly, and shiny,CRC Press, 2020.
20. R. Kannan, V. Vasanthi(2019), “Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease,” in Soft Computing and Medical Bioinformatics, 63–72, Springer, 2019, doi:10.1016/j.compedu.2012.08.015.
21. C. Hodges, S. Moore, B. Lockee, T. Trust, A. Bond(2020), “The difference between emergency remote teaching and online learning,” Educause Review, 27.
22. H. M. Niemi, P. Kousa, et al(2020), “A case study of students’ and teachers’ perceptions in a Finnish high school during the COVID pandemic,” International journal of technology in education and science.

