|  |
| --- |
|  |

Text

Description automatically generated with medium confidence

**DS7001 – DATA ECOLOGY (M-LEVEL MODULE)**

**EVALUVATED REVIEW AND PORTFOLIO- (FINAL).**

**NAME:** Kalai Selvi Gandhi **STUDENT ID:** U2216077

**MODULE NAME:** Data Ecology **MODULE CODE:** DS7001

**COURSE NAME:** MSC, Data Science (January 2022)

**APPLICATION AREA:** Education.

**TOPIC:** Data Science in the age of COVID (A Statistical modelling approach of factors affecting Education during Covid-19 using Artificial Intelligence and Machine Learning).

**MODULE LEADER:** Dr. Yang Li.

**DATA SCIENCE IN THE AGE OF COVID.**

**A STATISTICAL MODELLING APPROACH OF FACTORS AFFECTING EDUCATION DURING COVID-19 USING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING.**

**ABSTRACT:**

The COVID-19 pandemic had an impact on many aspects of global economy, includes the sudden shutdown of educational facilities at several countries around the world. Most educational institutions have changed their online teaching and learning activities because of this abrupt closure, enabling students to submit assignments and complete assessments on different learning management systems from the comfort of their homes. During COVID-19, several factors will impact remote learning. The three primary factors are **affordability**, **infrastructural** and **training**.

In this review, features are investigated and academic achievement of students throughout the pandemic is predicted using **classification algorithms** using **Machine Learning**. The key objectives of educational organizations are the achievements in their studies. In order to, effectively direct management, educators, and the government in the development and implementation of new educational policies, it is essential to be able to accurately forecast students' performance. Additionally, the accuracy of **Random Forest** and **Decision Tree** approaches was evaluated. However, an analysis of the correlations between the factors impacting digital education is performed. In addition to **correlation analysis**, **linear regression** has been used to determine the effects of infrastructure and affordability or training factor. Analysis illustrated that affordability had a positive adversely affect the training component whereas the impact of infrastructure is negative. Furthermore, this review shows an intelligent method for predicting student performance considering dataset from Global University used **Artificial Intelligence (AI)** during Covid-19.

***Keywords:*** *COVID-19, Educational Data Mining, Classification, Digital Education, Exploratory analysis, Correlation, Artificial Intelligence, Machine Learning, Prediction.*

**Table of Contents**

[I.INTRODUCTION 3](#_Toc123136248)

[II.LITERATURES 4](#_Toc123136249)

[III.METHODOLOGY 6](#_Toc123136250)

[III). A. Hypothesis and Discussion 7](#_Toc123136251)

[i) Null Hypothesis (H0) 7](#_Toc123136252)

[ii).Alternative Hypothesis (H1) 7](#_Toc123136253)

[III). B. Data preprocessing 7](#_Toc123136254)

[i). Factor Analysis 7](#_Toc123136255)

[ii). Statistical Analysis 8](#_Toc123136256)

[IV.PRESENTATION OF MAIN ELEMENTS 8](#_Toc123136257)

[a). Statistical analysis of using Data Ecology Elements 8](#_Toc123136258)

[i). Data Integration 8](#_Toc123136259)

[ii). Social media Data Analysis 9](#_Toc123136260)

[iii). Genomic Data Analysis 9](#_Toc123136261)

[b). Techniques used in Data Analysis 10](#_Toc123136262)

[i). EFA Assessment 10](#_Toc123136263)

[ii). Model fitting tests 10](#_Toc123136264)

[iii). Factor Correlation 12](#_Toc123136265)

[iv). Algorithms used 12](#_Toc123136266)

[V.CRITICAL DISCUSSION 14](#_Toc123136267)

[VI.ASSESTS AND LIABILITY 15](#_Toc123136268)

[VII.CONCLUSION 15](#_Toc123136269)

[REFERENCE 16](#_Toc123136270)

# **I.INTRODUCTION:**

The appropriate person in the right position can help to rationalize decisions by raising the human element's level of qualification, which is a vital component of good governance. The procedure is continued during the student's time in school. And each academic year's final months are defined by a busy orientation period. This is seen as the result of numerous campaigns to raise awareness, get people together in groups, and have one-on-one dialogues that begin at the beginning of the academic term. Filling out cards, choosing a career path, and other activities are just a few of the steps that learners, families, and institutions must engage in as part of this process, schedule cards for adults with unique needs and schedule cards for apprentices. These procedures include an agreement for the council to use pre-screened people and restricted routes, as well as studies, plans, and classification. It is also said that it is active during the second year of a bachelor's program because this is when students are assigned to enroll higher education institutions and academies, whether or not they are affiliated with institutions and regardless of whether they have open or restricted access. The production of the numerous documents that constitute up the application files, as well as electronic registration, are required for this appointment.

The academic achievement of pupils is the main goal of educational institutions. In order to effectively direct management, educators, and the government in developing new educational policies and implementing changes, it is essential to be able to accurately forecast students' performance. Finding elements that directly and indirectly affect learners' success rates is one of the biggest obstacles in assuring this success. Universities also deal with the ability to discover and apply significant patterns from the vast numbers of student statistics that are regularly created. The novel COVID-19 pandemic outbreak that occurred at educational institutions in 2019 added to these difficulties. Academic levels from the greatest to the lowest, the COVID influence significantly changed the ways that teaching, and learning were done. It is employed in a variety of corporate sectors, and the term "Educational Data Mining" has recently become extremely popular in educational contexts (EDM). One of the main uses of EDM is modelling student behavior, academic performance, career, social network analysis, and groups.

In order to create prediction models for their respective domains, these applications have made use of semi-supervised, supervised, and unsupervised learning. Reinforming Learning (RL) algorithms are being used more frequently in EDM to create instructional games and agents. This class of algorithm is essential, for customized learning and the creation of revolutionary educational gaming content.

Predicting students' academic achievement in the brand-new learning and teaching environment that COVID-19 has built is the main goal of this assignment. The goals is to identify the weak and strong characteristics which will influence students' educational achievement, also to develop the classification model for further forecasts and the factors determining online education. It also discusses remote learning that is AI-based.

# **II.LITERATURES:**

During the COVID-19 epidemic, **Samsudin** created a vector support network (SVM) learning technique to identify trends in students' academic and creative output. The classifier was created using a student database of 225 examples from Institute Pendidikan Sultan Idris (UPSI). The students' Point Average (GPA) before throughout virtual classrooms, their ages, and their overall GPA were included as attributes for the pattern analysis. The linear kernels produced the best accuracy, 73.68%, out of the four components utilized to build the SVM model: linear, radial basis function, polynomial, and sigmoid.

In order to group poorer pupils for academic support, **Prakash** suggested an improved support vector machine (ESVM) using clustering algorithms to identify students' performance during the pandemic. Following pre-processing, the Enhanced Clustering Algorithm and Particles Swarm Optimization (PSO) (IFCM)were used to identify clusters and choose features, respectively. The model was created using 1182 examples from colleges in the Delhi region and 420 student examples of classification results from an open-source SARS-CoV-2. The IFCM-ESVM achieved with an efficiency of 93.187% in comparison to the ESVM's performance of 91.892%.

An ensemble data mining technique was put out by **Krishna** and **Kumar** to pinpoint the main issues that students encountered during the epidemic. The classifier was developed with a total of 734 sample data and 7 features, including the impact of being physically within the classroom during the epidemic, beginning concentration during online classes, and the effect of being physically present during the pandemic, the degree of precision and interaction regarding online lessons, and the difficulty of lessons. As a predictive classifier, Decision Tree technique outperformed SVM, Logistic Regression, Bayesian Network, KNN, and Random Forests with an accuracy of 94.85%.

**Ilieva et al**. study's quantitative and qualitative data analysis and artificial intelligence approaches to develop a recommendation system that can identify, clarify, and foresee the difficulties that pupils may have during the pandemic. The classifier and statistical model were developed using data from 134 students who responded to questions that were posted on Facebook. Based on the dangers posed by COVID-19 and disruptions to students' academic success, a drop-out of schooling classification model was developed utilizing the classification and Regression Trees (CART), Conditional inference Tree (CTREE), Random Forest (RF), and SVM methods. With a training and validation of 88% and a subset of the training accuracy of 100%, SVM fared best.

**Ahmed et al.** and colleagues developed a model to predict the no-detriment proportion for supportive evaluations in relation to academic accomplishment during the epidemic by analyzing the effectiveness of the SVM, RF, Linear Regression, and Naive Bayes classification algorithms. The Global Institute of Technology and Engineering (GCET) provided the classifier with 1020 student data, of which 15% were utilized to anticipate the no detrimental policy and 85% were employed to create the model. Random Forests provided the highest accuracy of classification of 99.29% with an enhanced confusion matrix.

For COVID-19, **Zoric** proposed using a neural network and a rapid propagation algorithm to predict students' academic success. The University of Applied Research Baltazar, Zapresic provided the student data from 76 cases, with the following characteristics: state of studying, sex, parents' and educational levels, high school average grade, housing while studies, financial help, and employment during studies. A median ROC curve area under curve of 0.82 and a median accuracy of 93.42% were achieved by the neural network.

A thorough analysis of intelligent computation methods and their applications is provided by **Jennifer S. Raj** and **Thana Kumar Iwin.** The examination of data mining methods is presented by **Joseph** and methods that the computer system with intelligence could use. Data mining techniques were used by **Fernandes**, Maristela, Marcio, Vinicius, Carvalho, and **Gustavo** to do a descriptive and preventive statistical study to forecast student performance in the Bronze Capital. In a separate study, Tomasevic, **Gvozdenovic**, and **Vranes** looked at machine learning algorithms to assess student exam performance to identify students who were at "high risk" of failing the class. A project that examines eight machine learning (ML) algorithms for forecasting students' academic performance in a course has been proposed by Uskov, Bakken, Byerly, and Shah.

**Sekeroglu**, **Dimililer**, and **Tuncal** proposed a method for forecasting and categorizing student performance using five various machine learning algorithms; **Abu Saa**, **Al-Emran**, and **Shaalan** used random forest algorithm to forecast trainee learning achievement using a brand-new set of data from a private university in the United Arab Emirates (UAE); **Sossi Safae Alaoui**, Brahim Aksasse, and **Yousef Farhaoui** investigated the likelihood of combining.

**Manouselis** and **Drachsler** discuss the significance of technology-enhanced learning in whilst **Thai-Nghe et al**. suggested a method that projected student performance, Thai-Nghe and **Drumond** suggested a recommendation system that evaluated student performance, Romero et al. evaluated by comparing data mining technique to rank students as according to their use of Moodle, and Bekele and Menzel utilized data analysis methodologies for predicting student performance. Data mining techniques were utilized by **Thai-Nghe et al.** to forecast pupils' academic success, using a recommendation system.

**O. Chavarriaga** and **B. Florian-Gaviria** suggest that students can improve their learning abilities; A suggestion mechanism for high school students' orientation is proposed by **Ahajjam Tarik** and **Farhaoui**.

Artificial intelligence (AI) and machine learning models are mostly employed to increase the screening and diagnosis accuracy of non-infectious disorders. Additionally, machine learning techniques are frequently employed in the study and forecasting of COVID-19 survival rates, patient discharge times based on clinical data, and the likelihood of a second COVID-19 pandemic wave. **Mishra et al.** recognized teacher and student perspectives of online learning modes and web-based learning models using both quantitative and qualitative data. They investigated the use of cutting-edge web-based learning tools, such as tablets and smart phones, and discovered that they have an impact on both teachers' and students' emotional health. To further these endeavors, researchers investigate whether continuous usage of online learning tools by students as the outcomes of COVID-19 pandemic affects their psychological, also mental health.

The COVID-19 pandemic and other natural disasters shed some lights on the growth of the educational methodology start-up, and **Shivangi** includes recommendations for academic institutions on how to handle challenges associated with distance learning. The author examined the degree of educational stress that student encounter while pursuing an online degree as well as their coping mechanisms during the COVID-19 epidemic. In their statistical (Regression) model analysis of the data, **Onyema et al.** demonstrated that COVID-19 has negative consequences on education, including disruptions in learning, students losing their employment, and increased student duties. The research revealed that many professors and students utilized technology to carry on their online education. However, several challenges, including a bad network, a lack of electricity, and infrastructural problems, hindered online learning.

Different machine learning regressor models for the COVID-19 pandemic and its second rebound were suggested by **Zohair et al.** Additionally, the models are used to extrapolate the relationship between the quantity of stated cases and atmospheric parameters in specific places in order to calculate the impact of climate on the transmission of COVID-19.

# **III.METHODOLOGY:**

**DATA SOURCE:**

In this coursework, Data for the 2019–2020 academic year was gathered from the Department of ICT Education at the University of Education in order to undertake factor analysis and random forests algorithms to anticipate the effects of distance learning. The information was gathered in accordance with pertinent variables that might affect students' accomplishment and was connected to their overall mean test scores. An online survey was used to collect information on a range of topics, including demographics, use of technology, sleeping habits, social interactions, educational achievement, psychological problem, and a measure for anxiety and melancholy. The principal analysis is split into two sections using descriptive and inferential statistics, chi-square, Variance Analysis (ANOVA), and logistic regression.

A Chi-Square test for independence examines the relationship between the two variables in a contingency table below. It typically examines to see if categorical variable distribution diverges from one another. A low Chi-Square test score denotes a significant correlation or a fit between the observed and expected data. A big Chi-Square test statistic suggests that there is no link between the data. The p-value is used to determine whether results are significant. The number of categories less one determines the degree of freedom. The alpha level is set to 0.05 (5%), while it can also be set to 0.01 or 0.10. If Statistical value >= Critical Value, the dependent variable is rejected as true, and the null hypothesis (H0) is accepted. If Statistic Critical Value: not a substantial effect, don't rule out the independent null hypothesis (H0).

The hypotheses for this work are:

• H0: Academic achievement and the characteristics of digital instruments are independent.

• H1: Academic achievement and digital tools are not mutually exclusive.

## **III). A. Hypothesis and Discussion:**

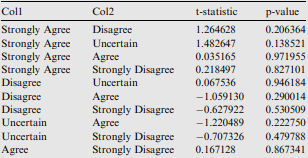
### **i) Null Hypothesis (H0):**

The null hypothesis is being taken into consideration because this study is experimental and there is no prior review of literature to support an alternative hypothesis. H0 here indicates that there is no evidence of a meaningful relationship between virtual learning and factors like infrastructure, accessibility, or training for distance learning.

### **ii). Alternative Hypothesis (H1):**

The hypotheses will be assessed at significance levels of 5% and 1%. The null hypothesis will be taken into consideration for acceptance when the alternative hypothesis has been ruled out. H1 here denotes whether there is a strong association between virtual learning and the mentioned variables.

The obtained result for t-test is illustrates as follows.



Preparing the data for learning, choosing the best model, and selecting the concept are the three main procedures. The dataset underwent the common data pre-processing procedures of missing data, normalization, and feature selection at the first block. For the purpose of learning, the database was randomly split into the Training dataset (70%) and Test dataset (30%). Most of the time, the dataset is inadequate to train the models when a portion of it is used for validating tasks. The dataset loses significant patterns as a result of reducing the training data, which increases the bias error. K-Fold Cross Validation was used to split the data into k subsets and verify the accuracy of the machine learning model. The super parameters optimization technique was used to improve the performance of the selected algorithms.

## **III). B. Data preprocessing:**

The missing data, inaccurate data, and null values from the dataset are removed at this stage by data pre-processing. Depending on repeated entries and lacking variables, a total of 536 samples were used to build the classifier.

### **i). Factor Analysis:**

In order to comprehend the issue statement, enormous amounts of data must be reduced to relevant variables through factor analysis. Principal Component factor Analyses (CFA) and Experimental Factor Analysis (EFA) are two types of it (CFA). When there is no accompanying literature, EFA has been employed, and scholars have investigated it to find novel facets of the topic. A theory put out by CFA researchers is supported by earlier research.

In order to mitigate the effects of an epidemic like COVID-19, there is yet no evidence that demonstrates that online schooling is necessary. Because of this, the need for widespread online education was unmet in the past; it only applicable to distance learning. Therefore, it is believed that exploratory research would be more appropriate for this coursework.

### **ii). Statistical Analysis:**

Factor analysis, and more especially EFA, is the most appropriate method for assessing the data when using Jamovi software because comments are on a five- point Likert scale and are not validated variables. The most recent developments in statistical methods are presented using open-source software named Jamovi. Among the statistical software tools related to it are ANOVA, t-test, dependability and correlation analysis, regression, and factor analysis.

# **IV.PRESENTATION OF MAIN ELEMENTS:**

**Affordability**, **Infrastructure**, and **Training** are the three main aspects that influence online learning during COVID-19. However, it is not an easy process to apply new regulations in schooling given the limited epidemic time. It is evident from this latest epidemic that there are threats outside the academic system. Additionally, it should be mentioned that the transition to digital academic delivery is fraught with logistical difficulties. E-learning is fully dependent, from a technology standpoint, on the availability of PCs, smartphones, and the internet; students or teachers with poor internet connections may find it difficult to access online learning. For students, teachers, and institutions, providing the technical tools necessary to organize virtual sessions is a significant problem. In his research, **Feldman J.** outlined some difficulties in dealing with student evaluation while managing digital education.

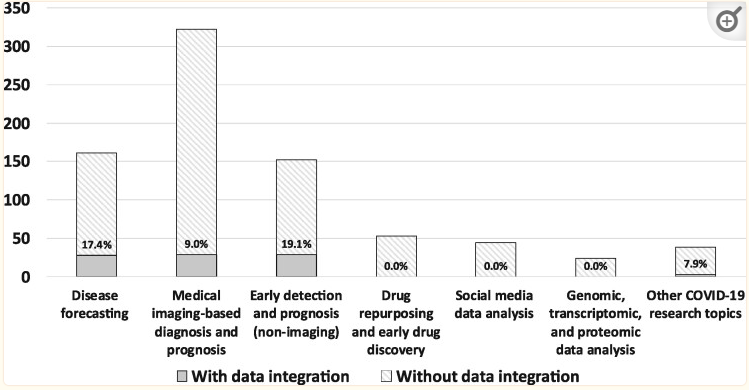
The academic achievement of students can initially be significantly impacted by disparities in economic, ethnic, and resource distribution. Worry and anxiety associated to the pandemic might also negatively affect pupils' ability to learn. Furthermore, not all teachers are equipped to deliver superior instruction remotely. The current study performed an excellent job of addressing the challenges of implementing digital learning from the viewpoint of the student consequently. To the greatest of our knowledge, it is the only research that considers all demographic, scientific, device usage, compatibility, and supervision-related limitations and has a significant participant (student) population.

**I****V. a). DATA EXPLORATION:**

**a). Statistical analysis of using Data Ecology Elements:**

### **i). Data Integration:**

A total of 24 Data integration is the process of combining information from several sources so that users can see things from a single perspective. Data integration is built on making data more easily accessible, consumable, and usable by systems and humans. The process of combining data from various sources into a single dataset is known as data integration. Its main goal is to serve the information needs of all applications and company processes by providing users with reliable access to and data delivery across a broad range of themes and structure types.



### **ii). Social media Data Analysis:**

A total of 44 studies in all discussed the application of AI to the study of social media data. With 32 studies examining tweets from across the world, Twitter was the most widely used data source in these studies. The other 12 studies made use of data from websites like Facebook, Reddit, YouTube, and Weibo. A common analytical strategy was used in most social media studies: text extraction and processing using NLP tools and methods, followed by topic modelling and/or sentiment analysis. Latent Dirichlet allocation was the most popular approach for topic modelling, whilst SVM, Naive Bayes, k-NN, random forest, etc. were among the machine learning models used for sentiment analysis. The social media studies did not incorporate diverse data for modelling.

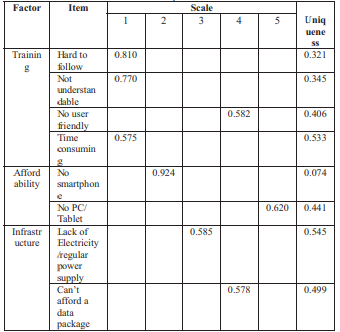
### **iii). Genomic Data Analysis:**

A total of 24 studies in total discussed the use of AI to examine SARS-CoV-2 sequence data. Finding the distinctive SARS-CoV-2 RNA or protein traits that might be targeted for disease detection and therapeutic or vaccine creation was one common analysis focus of several of these investigations. The SARS-CoV-2 genome sequences in GenBank at the National Center for Biotechnology Information were examined in more than half of these research. The Protein Data Bank, the National Genomics Data Center of China, and self-generated sequence data are among more sources of data. These experiments used a wide range of AI models, including deep learning models (CNN, RNN) and conventional machine learning methods (k-NN, SVM, random forest, GBM).

## **b). Techniques used in Data Analysis:**

### **i). EFA Assessment:**

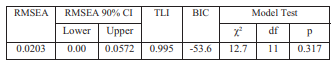
The four main reasons why adoption of e-learning during the current epidemic is challenging are presented in the table below together with 10-item questionnaires. The least residual extraction approach and a variance method were coupled for factor labelling. The training component, for instance, struggled with four categories, including priority for offsite learning (0.810), comprehending difficulty (0.770), not viewer (0.582), and much more time-consuming (0.582). (0.575). Lack of a mobile (0.924), a laptop or iPad (0.620), or both were affordability difficulties compared to infrastructural issues including no power source (0.585), difficulty to recharge data plans (0.578), and no connection to the internet (0.578). (0.514). The inability of certain students to operate e-learning with an outer loading of 0.684 was something they openly stated to me. However, as this element represents an individual or distinct perspective and cannot be categorized as an unique component under the EFA approach, it is ignored.



**Table of Factors affecting distance learning.**

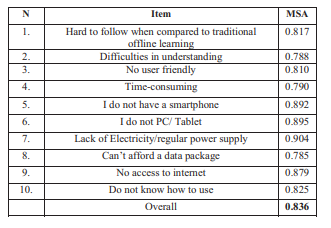
### **ii). Model fitting tests:**

The table below displays the models’ fitting variables for online learning during COVID-19. So, when Root Mean Squared Error of Approximate (RMSEA), which runs from 0 to 1, is small, a model fit better. The result model's RMSEA value of 0.0203 (2%, 90% CI) shows that the model was created and fitted precisely. According to Newsom J. (2015), if parameter estimates such the Tucker-Lewis Index (TLI) are greater than 0.95 and the Bayesian Information Criteria (BIC) has a low or negative value, the model is well-specified. Because of this, the proposed model's TLI value of 0.995 and low BIC value (-53.6) both show how well it suited the requirements of the research.



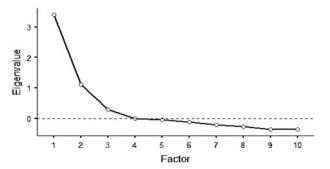
**Model fitting parameters.**

Another test that assesses the model fitting is Bartlett's test of sphericity. The model fitting is thought to be completely valid if the null hypothesis (H0), which works on the assumption that all variables are uncorrelated, is close to zero. It is clear from table 1 that p0.001 exhibits significant model significance. If the sample size is insufficient to accurately represent the factor items, the Kaiser-Meyer-Olkin (KMO) analysis, also known as the Measurement of Sample Adequacy (MSA) test, is used to determine this. The model was successfully fitted for this test if MSA > 0.50. All of the item values in table 2, including the sum of 0.836, satisfy this condition.

**Bartlett’s test of sphericity. (Table1) MSA Values. (Table 2)**

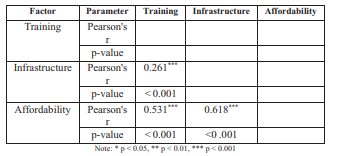
The number of components that can be retained in the EFA is determined by the scree plots used during multivariate statistical analysis. In the scree plot below, eigenvalues are on the Y-axis and factors are on the X-axis. The remaining factors are situated below a line, as opposed to the first three components, which are situated above the zero lines. Thus, it is simple to maintain the elements above the line, and these three elements also passed all the tests for model fitting.



**Scree plot.**

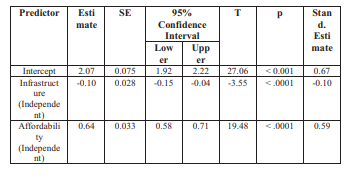
### **iii). Factor Correlation:**

The correlation matrix for the three criteria is shown in the table below. There is a 0.261 positive correlation between infrastructure and training, a 0.531 positive correlation between training and affordability, and a 0.618 positive correlation between infrastructure and affordability. It also demonstrates the initial link between the variables. The theory was evaluated at significance levels of 1% and 5%. For all three variables, the null hypothesis (H0) was rejected, and no evidence of a meaningful relationship between these variables was discovered. As a result, the alternative hypothesis was confirmed by a substantial association.



**Correlation Matrix for factors identified.**

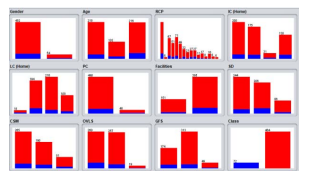
The outcomes of a linear regression study with a variable coefficient of training are shown in the table below (dependent). Infrastructure and affordability are projected to have estimative coefficients of -0.102 and 0.647, respectively (e.g., every unit change in affordability corresponds to a 0.647 change in training).



**Linear Regression.**

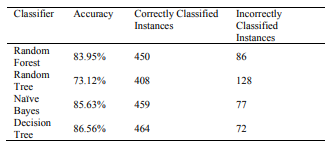
### **iv). Algorithms used:**

This coursework evaluates the precision of the classification methods Random Tree, Random Forest, Decision Tree, and Naive Bayes, as well as the confusion matrix of each classifier. By removing the least useful variables and rebuilding the classification model, feature selection is used more frequently.



**Attributes and Frequencies in Weka.**

* The bagging technique was used to create a **Random Forest (RF) algorithm**, which combines several decision tree algorithms. While growing trees select the best characteristic from a random set of features, random forests add extra unpredictability to the model.
* Ensemble basis models are employed in the **Random Tree (RT)** supervised classification to produce predictions. In order to select the most famous class from a group of trees, the random tree introduces more tree randomization.
* According to the Bayes Theorem, the **Naive Bayes (NB)** scientific method a posterior probability. Nave Bayes performed the best with categorical inputs and assumes that the validity of a certain quality exists independently of other attributes.
* A training model is produced via the **Decision Tree** (DT) method by inferring straightforward decision rules from the information. A node is divided up till the terminal node by the decision tree according to information gained.

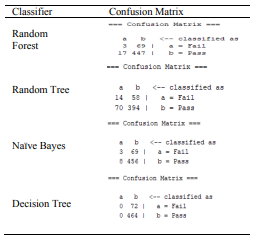
 **Chart, bar chart

Description automatically generated**

**Classification Accuracy of various algorithms.**

The **J48 Decision Tree** approach as from table above has the highest precision without selecting features when compared with Random Tree, Random Forest, and Naive Bayes.

The **J48 algorithm** successfully classified all instances of passed students while misclassifying all instances of unsuccessful students on the basis on the confusion matrix in the table below. Despite having an accuracy rate of 83.95%, Random Forest incorrectly labelled 17 passing students as failing while correctly labelling 3 students as failing. Naive Bayes incorrectly classified 69 cases as passes and 8 instances as failures, with an accuracy rate of 85.63 percent. The least accurate classification, Random Tree, identified 14 students as failing and 394 as passing with an accuracy of 73.12%.



**Confusion Matrix.**

 The feature selection applied to the model analysis to get a higher AUC classifier. The dataset's low-performing features were eliminated, leaving the classifier with only the high-performing features. The RF's accuracy fell from 83.95% to 81.90%. From 73.12% to 75.56%, RT accuracy increased. While J48 DT slightly decreased from 86.56% to 86.55%, NB climbed from 85.63% to 85.82%. The greatest marginal improvement in classification accuracy comes from RT owing to this feature selection. The classification accuracy of NB increased slightly by 0.16%, whereas RF's accuracy fell back by 0.18%. DT grew by 0.02% point while RT climbed by 2.43% gap. RT once more has the largest marginal improvement in classification accuracy following the second feature selection.

# **V.CRITICAL DISCUSSION:**

The current coursework investigates how COVID19 is harming higher education and the various frustrations that keep students from learning properly. As a result, the COVID-19 epidemic has a negative effect on educational activities, limiting access to laboratories, raising student debt, and decreasing learners' enthusiasm to learn. The COVID-19 pandemic caused issues in various spheres of life, but particularly in the educational system because it reduced educational chances for many pupils. The main barriers preventing students from enrolling in online education in this area are a lack of energy, digital literacy, accessibility, network problems, and poor facilities. More than 70% of those who responded to the suggested question concurred that the considerations prevented them from engaging in online education.

According to the study, the COVID-19 lockdown has a variety of repercussions, including putting more strain on parents, students, and educational institutions. The study's findings concur with those of earlier academics who believed that integrating technology into traditional education was the best course of action. Additionally, the COVID-19 epidemic will usher in a new era of online education and enable people to see the positive aspects of these tools. The study's limitations are summarized as follows: The proposed machine learning model has several noteworthy advantages, but it also has a number of disadvantages, the most remarkable of which are that it needs to perform correspondingly to a conventional one, that it requires a lot of processing time, and that the model's dimensions must be precisely calibrated. The sample of students who answered to an internet questionnaire on COVID-19 was also too small to generate viable machine learning models, so for this course, all the data from potential learners was used.

# **VI.ASSESTS AND LIABILITY:**

The EFA approach is used to find significance among three main elements, including affordability, infrastructure, and training, which were all included in 10-item questionnaires. To assess the effectiveness of the model and determine the link between the variables, several analytical tests were performed. Results showed that for every change in infrastructure and affordability, there can be a proportionate shift in training.

The classification accuracy of NB increased slightly by 0.16%, whereas RF's accuracy fell back by 0.18%. DT grew by 0.02% point while RT climbed by 2.43% gap. The greatest marginal improvement in categorization accuracy is shown with RT.

# **VII.CONCLUSION:**

The purpose of this study is to better understand the key variables influencing the acceptability of digital education in the classroom during COVID-19. Moreover, the classification methods to data collected are used to forecast individuals' academic success during the COVID-19 outbreak. Also, the analyzed pertinent features that mostly contribute to the class label. In order to predict future student performance using this dataset, the Naive Bayes classification model is utilized after analyzing the AUC, ROC, and accuracy without feature choice. The two top predictors employing feature selection are the Naive Bayes and Random Forest classifiers.

The results of this experimental study show that student participation is favorable, and that digital training is constructive and valuable, despite any technical issues that might disrupt and limit our capacity to participate in it during the COVID-19 epidemic. However, most participants agreed that due to various issues with the current infrastructure, engagement tactics, web information strategies, and the need for better information management, it could not completely replace the traditional method of learning. To manage the post-COVID-19 period, education sector stakeholders should develop strong mechanisms. Furthermore, there is a connection between the use of digital technologies for distance learning and academic achievement during the COVID-19 epidemic. This pandemic will also usher in a new era for online education and inspire greater faith in the system.

Future studies can focus on expanding the recommended methodology to additional countries in order to evaluate distance classes for undergraduate students with respect to the COVID-19 outbreak. Research must be done into the benefits of online education as well as the tensions and fears that such pandemics cause among students. Researchers may also investigate deeper deep learning models based on feature extraction achieved by sophisticated optimization methods. The suggested strategy can be used by researchers to tackle more difficult optimization problems.

The Study developed with the use of this dataset and its properties may serve as the foundation for future forecasts in educational settings and be pertinent to students' achievements during the COVID-19 epidemic.

# **REFERENCES:**

[1]. Aki, S. (2021). IMPACT OF COVID-19 ON UNIVERSITY EDUCATION. *International Journal of Research in Education and Sustainable Development*, 16(4)(8), pp.31–40.

[2]. Tarik, A., Aissa, H. and Yousef, F. (2021). Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science*, 184, pp.835–840. doi: 10.1016/j.procs.2021.03.104.

[3]. Aristovnik, A., Keržič, D., Ravšelj, D., Tomaževič, N. and Umek, L. (2020). Impacts of the COVID-19 Pandemic on Life of Higher Education Students: A Global Perspective. *Sustainability*, [online] 12(20), p.8438. Available at: <https://www.mdpi.com/2071-1050/12/20/8438>.

[4]. Jandrić, Petar, et al. “Teaching in the Age of Covid-19 - a Longitudinal Study.” *Postdigital Science and Education*, vol. 3, no. 3, 19 Aug. 2021, pp. 743–770, 10.1007/s42438-021-00252-6. Accessed 4 May 2022.

[5]. BURADA, Prof.Ph.D.A. (2021). IMPACT OF THE COVID-19 PANDEMIC ON EDUCATION AND TRAINING. *Pro Edu. International Journal of Educational Sciences*, 3(4), pp.42–48.

[6]. Aldhafeeri, Fayiz M., and Asmaa A. Alotaibi. “Effectiveness of Digital Education Shifting Model on High School Students’ Engagement.” *Education and Information Technologies*, vol. 6, no. 4, 27 Jan. 2022, 10.1007/s10639-021-10879-4.

[7]. Arndt, Sonja, et al. “Between the Blabbering Noise of Individuals or the Silent Dialogue of Many: A Collective Response to ‵Postdigital Science and Education′ (Jandrić et Al. 2018).” *Postdigital Science and Education*, vol. 1, no. 2, 14 Mar. 2019, pp. 446–474, 10.1007/s42438-019-00037-y. Accessed 10 Nov. 2022.

[8]. Australian Curriculum, Assessment and Reporting Authority. “Information and Communication Technology (ICT) Capability.” *Australian Curriculum*, 2017, [www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/information-and-communication-technology-ict-capability/](http://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/information-and-communication-technology-ict-capability/).

[9]. Bogdandy, Bence, et al. “Digital Transformation in Education during COVID-19: A Case Study.” *2020 11th IEEE International Conference on Cognitive Info communications (CogInfoCom)*, vol. 8, no. 6, 23 Sept. 2020, 10.1109/coginfocom50765.2020.9237840.

[10]. Bozkurt, Aras, et al. “A Global Outlook to the Interruption of Education due to COVID-19 Pandemic: Navigating in a Time of Uncertainty and Crisis.” *Asian Journal of Distance Education*, vol. 15, no. 1, 5 June 2020, pp. 1–126, zenodo.org/record/3878572#. YQOa-Y4zbIV, 10.5281/zenodo.3878572. Accessed 30 July 2021.

[11]. Ahshan, R. (2021). A Framework of Implementing Strategies for Active Student Engagement in Remote/Online Teaching and Learning during the COVID-19 Pandemic. *Education Sciences*, 11(9), p.483. doi:[10.3390/educsci11090483](file:///C:\Users\TEMP\Downloads\10.3390\educsci11090483).

[12]. Arora, M. and Mittal, M. (2020). Relation of Live Projects with Employability using Path Analysis Model. *Procedia Computer Science*, 167(6), pp.1675–1683. doi:[10.1016/j.procs.2020.03.378](file:///C:\Users\TEMP\Downloads\10.1016\j.procs.2020.03.378).

[13]. Chawla, S., Mittal, M., Chawla, M. and Goyal, L. (2020). Corona Virus - SARS-CoV-2: An Insight to Another way of Natural Disaster. *EAI Endorsed Transactions on Pervasive Health and Technology*, 6(22), p.164823. doi:[10.4108/eai.28-5-2020.164823](file:///C:\Users\TEMP\Downloads\10.4108\eai.28-5-2020.164823).

[14]. Emery, A. (2017). E-lectures and online learning: Not a replacement for live teaching. *Medical Teacher*, 39(12), pp.1292–1292. doi:[10.1080/0142159x.2017.1367373](file:///C:\Users\TEMP\Downloads\10.1080\0142159x.2017.1367373).

[15]. F.Y, O., J.E.T, A., O, A., J. O, H., O, O. and J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, 48(3), pp.128–138. doi:[10.14445/22312803/ijctt-v48p126](file:///C:\Users\TEMP\Downloads\10.14445\22312803\ijctt-v48p126).

[16]. Jide E. T., A., Oludele, A., Sunday A., I. and Shade O., K. (2020). SQL Injection Attacks Predictive Analytics Using Supervised Machine Learning Techniques. *International Journal of Computer Applications Technology and Research*, 9(4), pp.139–149. doi:[10.7753/ijcatr0904.1004](file:///C:\Users\TEMP\Downloads\10.7753\ijcatr0904.1004).

[17]. Ledesma, R.D., Valero-Mora, P. and Macbeth, G. (2015). The Scree Test and the Number of Factors: a Dynamic Graphics Approach. *The Spanish Journal of Psychology*, 18(4). doi:[10.1017/sjp.2015.13](file:///C:\Users\TEMP\Downloads\10.1017\sjp.2015.13).

[18]. Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R. and Sindhi, S. (2018). Online Education: Worldwide Status, Challenges, Trends, and Implications. *Journal of Global Information Technology Management*, 21(4), pp.233–241. doi:[10.1080/1097198x.2018.1542262](file:///C:\Users\TEMP\Downloads\10.1080\1097198x.2018.1542262).

[19]. Parvin, H., MirnabiBaboli, M. and Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence*, 37(8), pp.34–42. doi:[10.1016/j.engappai.2014.08.005](file:///C:\Users\TEMP\Downloads\10.1016\j.engappai.2014.08.005).

[20]. Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3). doi:[10.1002/widm.1355](file:///C:\Users\TEMP\Downloads\10.1002\widm.1355).