ANALYSIS OF COVID-19 EFFECTS IN EDUCATION USING MACHINE LEARNING TECHNIQUES

Project work submitted to Avinashilingam Institute for Home Science and Higher Education for Women

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

SUBMITTED BY

SABARNA .M (20PIT007)

Under the Guidance of

Dr.N.KRISHNAVENI M.SC., M.Phil., Ph.D, SET.,

Assistant Professor

Department of Information Technology

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND HIGHER EDUCATION FOR WOMEN

SCHOOL OF PHYSICAL SCIENCES AND COMPUTATIONAL SCIENCES

DEPARTMENT OF INFORMATION TECHNOLOGY

COIMBATORE-641043

MAY-2022

ANALYSIS OF COVID-19 EFFECTS IN EDUCATION USING MACHINE LEARNING TECHNIQUES

Project work submitted to Avinashilingam Institute for Home Science and Higher Education for Women

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

SUBMITTED BY

SABARNA .M (20PIT007)

Under the Guidance of

Dr.N.KRISHNAVENI M.SC., M.Phil., Ph.D, SET.,

Assistant Professor

Department of Information Technology

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND HIGHER EDUCATION FOR WOMEN

SCHOOL OF PHYSICAL SCIENCES AND COMPUTATIONAL SCIENCES

DEPARTMENT OF INFORMATION TECHNOLOGY

COIMBATORE-641043

MAY-2022

DECLARATION

DECLARATION

I hereby declare that the project entitled "**ANALYSIS OF COVID-19 EFFECTS IN EDUCATION USING MACHINE LEARNING TECHNIQUES"** is a record of the original work done by **Sabarna.M(20PIT007)** under the guidance of **Dr.N.Krishnaveni M.SC., M.Phil., Ph.D** Assistant Professor, Department of Information Technology, School of Physical Sciences and Computational Sciences**,** Avinashilingam Institute for Home Science and Higher Education for Women in the partial fulfilment for the award of the degree of Master of Science in Information Technology, and this project work has not formed the basis for any Degree/Diploma/Associates.

PLACE:

DATE:

Signature of the Candidate

Countersigned By

Dr.N.KRISHNAVENI M.SC., M.Phil., Ph.D
Assistant Professor,

Department of Information Technology,

School of Physical Sciences and Computational Sciences**.**

CERTIFICATE

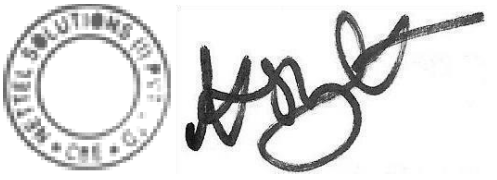
Net Tel Solutions India Pvt Ltd
Excellence in Service
NTS

25.05.2022
Coimbatore

TO WHOMSOEVER IT MAY CONCERN

This is to confirm that **Ms. M. Sabarna (Reg No:20PIT007)** final year student of **M.Sc (Information Technology)** from **"Avinashilingam Institute for Home Science and higher Education for Women, Coimbatore-641043**", has successfully completed her Project work on **"Analysis of Covid-19 Effects in Education using Machine Learning Techniques"** in our esteemed organization from January 2022 to May 2022.

For Net Tel Solution India Pvt Ltd



Authorized Signatory

#203, Ram Nagar, Nehru Streeet, Coimbatore - 641 009
www.nettelsolutions.in
info@nettelsolutions.in

CERTIFICATE

This is to certify that this project work entitled "**ANALYSIS OF COVID-19 EFFECTS IN EDUCATION USING MACHINE LEARNING TECHNIQUES"** done by **Sabarna(20PIT007)** has been submitted to Avinashilingam Institute for Home science and Higher education for women, Coimbatore-43 in partial fulfillment of the requirement for the award of the degree of **MASTER OF SCIENCE IN INFORMATION TECHNOLOGY.** This Project has not found the basis for the award of any Degree/Associate/fellowship or similar title to any Candidate of any University. Certified as a bonafied record of the work submitted for the Viva voce held on_____.

Signature of the HOD

Signature of the Guide

Signature of External Examiner

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I owe my sincere thanks to **Lord Almighty** and **My lovable parents** for showering their generous blessings upon me in all endeavours.

I would like to express my deep sense of reverential gratitude and sincere thanks to **Dr.S.P.Thyagarajan,** Chancellor, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, for the opportunity given to me for undertaking this study and for providing all the needed facilities during my study.

I owe my great deal of gratitude to **Dr.V.Bharathi Harishankar, Ph.D., FRSA Vice-Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for extending all resources that facilitated the conduct of the present study.

I express my gratitude to **Dr.S.Kowsalya, Registrar, M.Sc., M.Phil., Ph.D.** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities necessary for the study.

I would express my boundless thanks to **Dr. G. Padmavathi, M.Sc., M.Phil., Ph.D., and Dean,** School of Physical Sciences & Computational Sciences and Principal Investigator for Center for Cyber Intelligence, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for granting the facility required.

I express my sincere thanks to **Dr. D. ShanmugaPriya, M.Sc., M.Phil., Ph.D., SET,** Head, Department of Information Technology, for her encouragement and inspiration to complete theproject.

Ihe artily thank my esteemed project guide **Dr. N. Krishnaveni M.Sc., M.Phil., Ph.D, SET, Assistant Professor, Department of Information Technology**, for imparting tremendous assistance and well-timed support for triumph of our project.

I express my honorable thanks to our project coordinator **Dr.F.Paulin MCA., M.Phil., Ph.D., Assistant Professor, Department of Information Technology,** for her kind advice and knowledgeable suggestions which helped us to complete our project successfully.

I sincerely thank all **the Staff members** of Department of Information Technology, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for their help and support.

I would like to express my special thanks to **my parents, my friends** and all **my well-wishers** for their constant encouragement, support and help in carrying out this work successfully.

ABSTRACT

ABSTRACT

The urban and rural education systems in India are both still in their infancy. The midday food programme is designed to encourage pupils to attend school. In these conditions, the government enacted a nationwide lockdown on March 25, 2020 to fight COVID-19, which has had a significant influence on schooling. After China, India has the world's second biggest educational system. According to UNESCO, 63 million instructors in 165 countries were affected. A total of 1.3 billion students were unable to attend schools or institutions throughout the world, with nearly 320 million students in India alone. It has shifted the old educational system to a paradigm based on educational technology, in which teaching and evaluations are done online. COVID-19 has both beneficial and bad effects on the Indian educational system. The dataset for the covid19 pandemic on education comes from the kaggle repository. To get the most optimum feature, an exhaustive feature selection strategy employing a random forest classifier is utilised. Using support vector regression and logistic regression, this research examines the impact of COVID-19 on the Indian education system. It focuses on education during online teaching and assessment of students taking online classes from home during this epidemic.

CONTENTS

CONTENTS

INTRODUCTION

INTRODUCTION

Background

As a natural way to enforce social separation among communities, the Indian government has declared the lockout and shutdown of educational institutions. The statewide lockdown has had a significant influence on the country's educational system, particularly for children from remote areas. The current situation has made the running of educational institutions extremely challenging, as the Indian education system is dominated by classroom learning.

During this time, all educational activities, including ajvbbs examinations, school admissions, university entrance exams, fnfnnfsdkfnand competitive examinations, are held. As time passes and no obvious solution to the epidemic emerges, the closure of schools and colleges has a significant impact on learning across the country. The framework of the Indian educational system, including learning methods, teaching strategies, and evaluation methodologies, has been significantly influenced, resulting in a transition to online education with a strong emphasis on virtual education to meet the specified goals and objectives. However, only a few schools and colleges could implement such approaches, and low-income private and public institutions are inefficient in doing so, resulting in a closure.

Online Education System

Teachers can understand the learning effect of students in a typical learning environment by examining each course. Obtaining each student's particular learning process throughout the teaching process is challenging, as is predicting the students' learning success based on their characteristics and learning behaviour data. As a result, it is difficult for teachers to master their pupils' situations and give prompt intervention and assistance. With the fast growth of online learning in recent years, particularly the introduction of MOOCs, a vast quantity of data has been gathered in various online learning platforms.

The data created by learners' interactions with the operating platform throughout the learning process is known as online behaviour data, and it is primarily captured and stored in real time by the learning platform database and other tools. The categorization of a significant amount of data on learning behaviour has certain implications for data gathering.

The researcher chooses the relevant indicators based on the analysis' objectives, in order to define the type of learning behaviour from which data is gathered based on the behaviour classification findings, making the data collecting process quicker and more convenient, and the behaviour description results more accurate. Learners undertake learning activities using online learning platforms depending on their individual requirements, according to behavioural science theory, and create various learning behaviours. The login behaviour is determined by the course plan's needs, the contents of the guide, the contents of the assessment, and academic accomplishments, among other factors.

The behaviour of resource access is determined by the demand for resources. The retrieve behaviour is determined by the query learning material, the topic tool, and the course content, among other factors. The urge to communicate and collaborate with others drives the forum interaction habit. Feedback and reflection behaviour are developed depending on the requirements of reflecting learners, curriculum evaluation, and peer needs. And the learning task-related behaviour is developed depending on the learning task's needs.

Researchers are seeking for strategies to make data intelligible and useful, according to Burgos et al. (2018), as the volume of data on the online learning platform grows. Researchers look into the theory, structure, techniques, and practises of learning analysis in order to examine and uncover new possible educational laws. In recent years, more and more research on learning behaviour analysis, particularly on forecasting students' learning achievement, has been conducted.

According to Valiente et al. (2015), as learning analysis research progresses, researchers have begun to investigate learning predictions using machine learning methodologies and tools. Teachers and students are more likely to use mobile smart terminals and social applications such as WeChat, WhatsApp, Twitter, and Facebook to communicate with each other, discuss learning problems, and even submit homework on tablets or mobile phones in recent years, thanks to widespread use of the internet, particularly the rise of the mobile internet. Researchers have substantially less effective data available from online learning platforms than before since the following data are difficult to gather. As a result, under the assumption of limited data, how to assess online learning behaviour and construct an effective learning performance prediction platform is an important topic.

The Overview

With the advent of educational data and big data-related technological research, online learning behaviour analysis has drawn the attention of an increasing number of scholars. Because online learning behaviour analysis may acquire numerous recorded data of learners' online learning, rather than acquiring subjectively strong data through questionnaires, it is more objective than traditional offline learning analysis. Learners engage in a number of activities to meet their individual requirements.

During the learning process, the learner's expectations of the learning objectives are realised by employing learning tools over time while interacting with learning content, the learning environment, and learning partners. After that, the learning accomplishment is created. A thorough learning process should comprise learners, learning partners, learning content, learning tools, learning objectives, learning accomplishment, learning environment, and learning time. The above variables should be considered during the learning process, whether it is conducted offline or online.

Reading books, answering questions, watching videos, seeing courseware, exploring forums, posting resources, accessing learning platforms, debating and connecting with others, and so on are all examples of learning behaviour. Learning behaviour, when combined with behavioural science theory, may be thought of as a process in which learners engage based on a two-way interaction between a learning objective and a learning environment.

Consider each learning behaviour as a system, which should comprise subjects, objects, operations, surroundings, and consequences of behaviour. Each behaviour activity is a systematic process in which a behaviour actor interacts with a behaviour object in a certain learning environment and time, conducts operations, and creates specific outcomes. Subjective thought guides the occurrence of each learning activity, which is also constrained by the environment. There may be actions such as fast forward and pause when a student obtains knowledge from a video resource; when the learner is engaged in collaborative learning, it may be essential to voice his own viewpoint on the topic. In online learning, operational behaviours are varied and vary with different operating objects. The following four types of online learning habits can be classified.

Trajectory behaviors

Login, exit, and jumping between online pages in the system are all examples of trajectory behaviours. The most general and fundamental sort of learning behaviour is trajectory behaviour. This behaviour belongs to both the resource learning behaviour and the trajectory behaviour when a learner obtains a certain type of resource. The distinction between the trajectory behaviour and the resource learning behaviour is that the resource learning behaviour includes the body of the search, whereas the track behaviour just retrieves the action and does not include the retrieval's major substance.

Social behaviors

Learners, teachers, and other learners can connect and cooperate face to face in a standard offline learning environment, but persons in an online learning environment (OLE) rely on the network, including real-time communication and discussion-based communication. When learners cooperate to accomplish a work, they may express their own thoughts individually; when learning is unclear, they can consult with the teacher or other learners; when they have insights into the study, they can publish in the forum or in their unique learning area, and so on. Learners often provide text-based content that represents their internal knowledge structure and personal traits in online learning social behaviours.

Resource learning behaviors

Learning behaviours that include resources are called resource learning behaviours. The online learning platform offers a wide range of learning tools, including multimedia, text, and other formats. When a learner browses multimedia learning resources, it will display fast-forward, pause, and loop-playing behaviours; when a learner reads text-based learning resources, it will display jump positioning, pause preview, and other similar behaviours.

Evaluation and reflection behaviors

The majority of learner evaluations rely on examinations, tests, assignments, and other similar activities in order to assess learners' learning achievement at a certain level. Learner test scores and submission of responses are frequently recorded in online learning processes, however if just the final answer or grade is recorded, the learner's process information in the answer process

would be lost. As a result, the learner's test answers are amended numerous times, and this process should be documented to better reflect the learner's actual learning environment.

Problem statement:

The major problem that deals in this research is to find at what degree did the fatal corona virus have a cascading impact on Indian education?

Objectives

Some of the major objectives of the impact of corona virus in online education are listed below:

- The depiction of how the virus has impacted underprivileged kids and private teachers.
- To determine if virtual learning can replace instructors.
- To determine the limits of internet access for education in Delhi and beyond.

Motivation and Justification

The motivation and justification of the covid19 in online education are listed below:

- The impact of Covid 19 in Indian online education must be investigated.
- Our primary goal is to assess the influence of Covid 19 on the education platform before it does harm to India's educational system.

Machine learning techniques

Software developers define traditional applications as functioning within confined and limited rules. Because constructing information-driven applications like computer visions or email sorting is almost impossible with typical programming strategies, algorithms' advancement and self-learning qualities encourage them to overcome traditional applications. These algorithms assist us in making better selections and provide consistency. Machine learning is divided into three categories, as follows:

- Supervised
- Unsupervised
- Reinforcement

Supervised Learning

For the most part, supervised learning is used to solve even-minded AI challenges. In supervised learning, there are two factors: one for inputs and one for outputs.

The algorithms are prepared using data sources and the results are thought about in terms of accessible outputs. As a coach or educator, we screen the learning during the training stage.

The last part examines supervised learning techniques in the categories of regression and classification. There are also doubts that these algorithms will be operational after the data has been categorised. Data is no longer free, and social event information for learning may be costly as a result of its use as another oil on the planet. A large number of supervised algorithms have been tested for the common sense element. Each of them has their own set of strengths and weaknesses. In machine learning, there is no one algorithm that performs best for all jobs. As a result, selecting an algorithm is an important issue that everyone working in Machine Learning should consider.

The following are some of the most commonly used managed algorithms:

- Support Vector Regression
- Logistic regression
- Linear regression
- Decision trees
- k-nearest neighbor

Unsupervised Learning

Unsupervised algorithms, unlike supervised algorithms, do not have correct solutions. There is no output variable and no guide or teacher to deal with mistakes. The algorithms

are attempting to grasp the aspects of the data. They look for hidden and hidden instances in the dataset to predict the outcome based on the input parameters alone. Clustering and association problems collect unsupervised learnings.

Clustering: In this type of work, information is divided into groups, such as categorising clients based on their purchase habits.

Association: Algorithms are striving to understand the criteria that can explicate the vast amount of data, such as whether a client who buys a shirt will also buy pants. To offer you some instances of unsupervised learning computations, consider the following:

- Apriori algorithm connection
- K-means clustering

Reinforcement Machine Learning

Reinforcement learning functions similarly to how a kid learns in the early stages of life. When a child does admirably, the individual will be encouraged, and when a child performs poorly, the conclusion will be discipline or advice to avoid repeating the mistake. These algorithms are performing the same duty as an agent who acts as a youngster. The agent cooperates with the environment, and it is rewarded for properly doing tasks and punished for incorrectly executing them. When an agent is attempting to increase the awards while limiting the penalties.

These algorithms are also known as dynamic programming, and a great lot of research and development is being done to enhance them so that they may be used for a wide range of jobs in the near future.

Used Machine learning Techniques

On the same dataset presented before, the author constructed and developed two supervised category algorithms in this research. The following are the algorithms:

- Logistic Regression
- Support Vector Regression (SVM-R)

a) Support Vector Regression (SVM-R):

The Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for classification and regression. It is most commonly used for classification, although it may also be beneficial for regression. SVM basically finds a hyper-plane that separates the different sorts of data. This hyper-plane is nothing more than a line in two-dimensional space.

Each data item in a dataset is represented by an N-dimensional space in SVM, where N is the number of features/attributes in the dataset. Next, determine the best hyperplane for separating the data. Only binary classification is possible with SVM (i.e., choose between two classes). For multi-class situations, however, there are several approaches to employ.

We may use a binary classifier for each class of data to conduct SVM on multi-class situations. Each classifier's two outcomes will be:

- The data point is a member of that class OR
- The data point does not fit into that category.

To do multi-class classification on a class of fruits, for example, we may develop a binary classifier for each fruit. There will be a binary classifier for the 'mango' class to predict whether it is a mango or not. The SVM output is picked as the classifier with the greatest score. SVM works very well without any modifications for linearly separable data. Linearly Separable Data is any data that can be plotted in a graph and can be separated into classes using a straight line.

The supervised learning technique Support Vector Regression is used to predict discrete values. SVMs and Support Vector Regression are both based on the same premise. SVR's primary concept is to identify the optimum fit line. The best fit line in SVR is the hyperplane with the greatest number of points.

The SVR, unlike other regression models, aims to fit the best line within a threshold value, rather than minimising the error between the real and projected value. The distance between the hyperplane and the boundary line is the threshold value. SVR's fit time complexity grows more than quadratically with the amount of samples, making it difficult to scale to datasets with more than a few tens of thousands of samples.

Linear SVR or SGD Regressor are utilised for big datasets. Linear SVR is quicker than SVR, but solely takes into account the linear kernel. Because the cost function ignores samples whose prediction is near to their objective, the model built by Support Vector Regression only uses a part of the training data.

b) Logistic Regression

Logistic regression is an example of supervised learning. It's used to calculate or predict the likelihood of a binary (yes/no) event occurring. Logistic regression is used to determine whether or not a person is likely to be impacted the online education with COVID-19 using machine learning. Because there are only two possible responses to this question: yes, online education is impacted or no, they are not impacted, this is known as binary categorization.

In this hypothetical scenario, the likelihood of a state getting impacted with COVID-19 might be determined by id, region of residence, age of subject, time spend of online class, rating of online class experience, medium for online class, time spend on self study and among others. Our factors (independent variables) would include id, region of residence, age of subject, time spend of online class, rating of online class experience, medium for online class, time spend on self study, time spend on fitness, time spend on sleep time spend on social media, preferred social media platform, time spend on TV, number of meals per day, change in your weight, health issue during lockdown, stress busters, time utilized, do you find yourself more connected with your family, close friends and relatives, what you miss the most, all of which would affect our result by state (Dependent Variable).

A mathematical formula that turns predicted values into probabilities is the sigmoid function. Any real number between 0 and 1 is converted into another number. The value of the logistic regression must be between 0 and 1, and it cannot be more than this, resulting in a "S"-shaped curve. The S-form curve is also known as the Sigmoid function or logistic function. In logistic regression, the threshold value, which represents the likelihood of 0 or 1, is employed. Numbers over the threshold value, for example, are more likely to be 1, whereas values below the threshold value are more likely to be 0.

For training data that fits the following conditions, logistic regression is a good match.

- Either a binary or dichotomous outcome is expected. (In the case of binary logistic regression, this is true.)
- The independent variables, or factors that influence the outcome, are unconnected to one another. In other words, there is little or no multi co-linearity among the independent variables.
- The independent variables can be linearly related to the log probability.
- The sample sizes are colossal.

Logistic regression may not be appropriate for that application if the training data does not match the aforementioned requirements.

Tools Used

This proposed system, like all other research and studies, used some software and tools to create models and tests. Various tools are required for such studies, such as the author's choice of Python as the model development programming language, the dataset containing a large number of vehicle as well as non-vehicle images used for training, or the many Python packages that are required or beneficial in order to generate ML models. This chapter introduces all of the tools that will be utilized in the proposed methodology.

Python

Python is a Guido van Rossum-created general-purpose programming language that is used for a variety of platforms including mathematics, computer GUIs, the web, and a variety of significant scientific applications. Python's fundamental goal was to make programming simple so that anybody in the world could create code. As a result, it is well-known for its simplicity. Python is a space-sensitive language. Ignoring the fact that Python was designed for children, it currently supports all programming languages in a wide range of industries, which is both astounding and incredible. Python can accomplish whatever those other programming languages can. Everything from the web to algorithms to desktop apps is covered.

Python has gained popularity in artificial intelligence and machine learning applications in recent years owing to its abundance of efficient and useful libraries that make the process much easier and faster. Experts in this discipline come from a variety of backgrounds. Python is the easiest and most convenient language to start with if they have no programming experience.

Despite the foregoing characteristics, the researcher's knowledge and self-interest prompted him to chose Python as the model development programming language. The following are the libraries that were used in this project.

a) Numpy

Numpy is a free and open source framework that uses multi-dimensional matrices and arrays to do computations. It has a variety of functions that make working with this sort of data simple. Arrays are used in data analysis to make it faster and more efficient. As a result, this library aids data scientists in working with vast volumes of data. The models function often requires arrays as a parameter for detection and forecasting to operate quickly and reduce training prediction time.

b) Matplotlib

Plotting is becoming more popular in all industries as a way to see and analyse data. As a result, Matplotlib is used as a plotting library to construct various graphs and figures for a range of purposes. Matplotlib has the advantage of producing nice plots and graphs with only a few lines of code. To extract colour characteristics and construct histograms, matplotlib is utilised.

JupyterNotebook

Jupyter Notebook is an online tool that allows us to write and edit live scripts, equations, plaintexts, and visualisations. This is a free notepad that supports a variety of programming languages. This is used for a variety of things, including machine learning, numerical simulation, and data visualisation. This notebook was chosen by the researcher for creating legible code and applying machine learning algorithms.

System Specification

The PC that is being used to train and test the models has the following properties:

Model: Dell

RAM: 16 GB

Processor: Core i7

Quad Core Graphic:

Intel HD8 GB

LITERATURE REVIEW

LITERATURE REVIEW

Prediction of learning behaviour is an important research topic and application objective in learning behaviour analysis. With the advancement of learning analysis technology in recent years, more and more learning prediction research work has appeared.

The study's data and information come from a variety of publications and articles on the impact of the COVID-19 pandemic that have been published by national and international organisations. Data is also gathered from a variety of reliable websites. Some journals on the influence of COVID-19 on the educational system are also recommended.

Table 2.1: Literature survey

Sno	Author	Observation
1.	Dietzuhler and Hurn, 2013	Predicting learners' performance by analysing learning behaviours assists developers in more effectively evaluating online learning systems, continuously improving system availability, and expanding system functions to visualise learners' behaviours and future trends of learning behaviour; assist teachers in understanding the trends of learning behaviour; assist teachers in engaging in appropriate human interventions for learners at the appropriate time; assist teachers in continuously improving their teaching skills. At the same time, the system can assist teachers in providing timely assistance to learners who are performing poorly in order to enhance their learning performance.
2.	Balakrishnan and Coetzee (2013)	On the MOOC platform, it employs four behavioural eigenvalues and hidden Markov models to forecast the likelihood of dropping out.
3.	Kloft et al., 2014	On the MOOC platform, click data and SVM models were employed to forecast student dropout rates.

4.	Jiang et al., 2014	Chosen two logistic regression models to predict learners' success on the MOOC platform over the course of a week
5.	Qiu et al., 2016	developed a unified predictive model for predicting learners' academic achievement and certificate attainment
6.	Sorour and Mine (2016)	Machine learning is used to predict learner performance using predictive trees and random forests. Second, many academics do a range of theoretical study on eigenvalue selection, which includes a variety of eigenvalues that may be connected to learning effects from various viewpoints.
7.	Brown, 2012	Learner traits, learning behaviour features, and student work were defined as three categories for predicting eigenvalues. He explores relevant prediction models and examples for various forms of eigenvalues.
8.	Berry, 2017	Academic variables, demographic factors, and cultural and social factors are proposed as three markers that influence academic attainment. Because not all feature values relevant to learning effects can be gathered in real prediction, which is limited to many objective criteria, there are typically numerous manual trade-offs in the procedure.
9.	Balakrishnan and Coetzee (2013)	The researcher concludes that as predictors, it uses the cumulative percentage of video lectures that can be seen, the number of forum posts, the number of forum responses, and the number of course progress views.
10.	Romero et al., 2013	The participation in the forum directly predicts learner performance. The number of messages written by the learner, the number of new topics made by the learner, the number of posts read by the learner, the length of

		time the learner spent on the forum, the learner's focus, and the learner's adherence were all employed as predictive indicators.
11.	Villagrá-Arnedo et al. (2017)	Researcher points out that white-box model predictions are often regarded to be "interpretable," whereas black-box learning predictions are "unexplained." The decision tree model is a common white box model in machine learning, while the artificial neural networks (ANNs) model is a typical black box model.
12.	Fojtik (2018)	In the previous two decades, conducted a comparative study of undergraduate students who attended courses and studied in the computer science department using the online education and full-time education models.
13.	(Xiao & He, 2020)	It is observed that with the influence of the Covid-19 epidemic, online education research has expanded in the previous two years. Based on qualitative data from a survey study conducted at the Chinese University of Hong Kong at the time the pandemic measures were suddenly launched in China where the epidemic began, the validity and reliability of emergency online classes were investigated from the perspective of students and teachers.
14.	Pete and Soko (2020)	During the Covid-19 epidemic, a systematic study of students and instructors in selected Sub-Saharan African nations, Kenya, Ghana, and South Africa, examined digital competency, cost, and Internet connectivity of online education.
15.	Sözen (2020)	In European nations that are particularly hit by the epidemic, a lot of research is being done on distant education using textual and visual media.
16.	(Yavuz et al., 2020)	The Covid-19 period activity reports issued by the

		Turkish Higher Education Council (YK) were used to analyse the online education applications of higher education institutions throughout the pandemic era.
17.	Şen & Kızılcalıolu, 2020	A group of researchers from Antalya AKEV University used the structured interview approach to create online survey questions to determine the attitudes of university students and academics on online education.
18.	Yılmaz and Güner (2020)	Researchers examined into how online education is delivered to pupils at various levels, taking parental feedback into account.
19.	Keskin and Kaya (2020)	During the epidemic, people of different ages and socioeconomic groups have shared their views on social media on online education.
20.	Özoran (2020)	researched the usage of the 10 most valuable education platform of Turkey in the Covid-19 pandemic period on Twitter

From the above literature study, it is observed that the machine learning algorithm can be applied to analyze the impact of covid-19 in online education using regression techniques. Hence, the regression techniques namely support vector regression and linear regressions are applied to analyze the impact of covid-19 in online education.

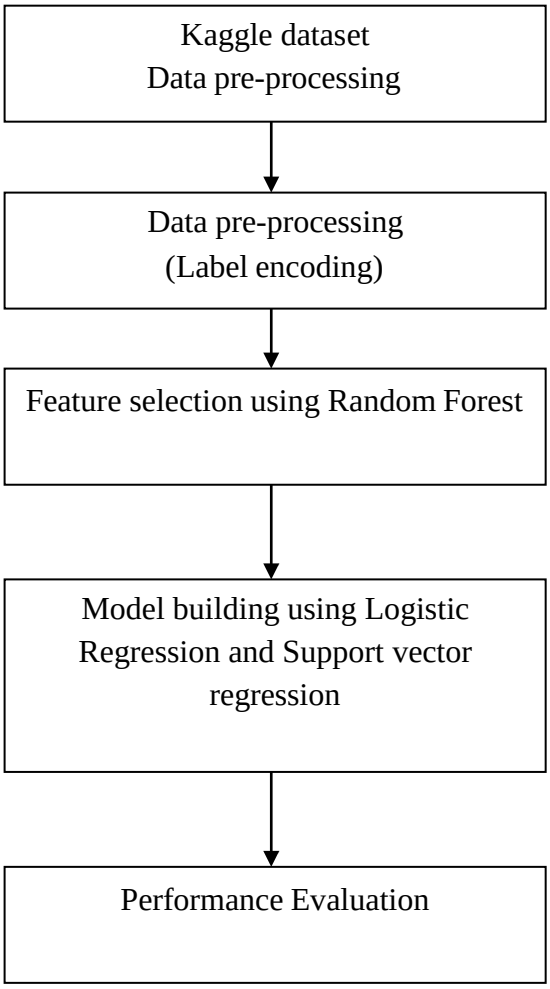
METHODOLOGY

METHODOLOGY

Overall Methodology

The proposed system is model creation for impact of online education by covid19. Initially, pre-processing is accomplished using exhaustive feature selection technique namely random forest. Followed by model building using machine learning techniques namely support vector regression and logistic regression to analyse and detect the impact of covid19 in online education. It is done using covid19 survey response from kaggle repository. Finally, the model is evaluated using performance metrics.

The overview of methodology is shown in figure 3.1



The entire methodology is divided into four modules namely, data acquisition, data preprocessing namely label encoding and feature selection, model building using machine learning techniques and performance evaluation to evaluate the performance.

MODULE 1: Data acquisition

The proposed system uses Benchmark dataset (covid19 student response dataset) from Kaggle. The dataset contains 16697 instances and 19 attributes. The attribute consist of

- id,
- region of residence,
- age of subject,
- time spend of online class,
- rating of online class experience,
- medium for online class,
- time spend on self study,
- time spend on fitness,
- time spend on sleep,
- time spend on social media,
- preferred social media platform,
- time spend on TV,
- number of meals per day,
- change in your weight,
- health issue during lockdown,
- stress busters,
- time utilized,
- do you find yourself more connected with your family,
- close friends and relatives,
- What you miss the most.

Based on the analysis it will predict whether the online education is happy or bad during covid-19. The main purpose of this project is to analyse whether the online education during covid-19 is impactful or not using the above mentioned dataset.

MODULE 2: Data pre-processing

In Machine Learning, data preparation is a critical step that improves data quality and facilitates the extraction of relevant insights from the data. In Machine Learning, data preprocessing refers to the process of cleaning and organising raw data in order to make it appropriate for creating and training Machine Learning models. In basic terms, data preprocessing is a data mining approach in Machine Learning that turns raw data into a legible and intelligible format.

When it comes to building a Machine Learning model, data preparation is the first step. Real-world data is frequently partial, inconsistent, erroneous (including mistakes or outliers), and lacking in exact attribute values and trends. This is where data preparation comes into play: it cleans, formats, and organises the raw data, making it ready for Machine Learning models. Let's have a look at the different stages of data preparation in machine learning.

Steps in Data Preprocessing in Machine Learning

In Machine Learning, there are seven essential processes in data preprocessing:

1. **Acquire the dataset**
2. **Import all the crucial libraries**
3. **Import the dataset**
4. **Identifying and handling the missing values**
5. **Encoding the categorical data**
6. **Splitting the dataset**
7. **Feature scaling**

In this research, the data encoding technique is used to convert the categorical data into numerical one.

Categorical variables are finite in number and are commonly expressed as 'strings' or 'categories'. The variables can only take on certain values. We can also observe that there are two types of categorical data-

- Ordinal Data: There is an intrinsic order to the categories.
- Nominal Data: There is no intrinsic order to the categories.

When encoding ordinal data, it's important to keep track of the category's order. The highest degree a person holds, like in the example above, provides important information about his qualifications. A person's degree is a key factor in determining whether or not they are qualified for a position. We must evaluate the existence or lack of a feature while encoding Nominal data. There is no sense of order in such a situation. For instance, the state of residence. It is critical to keep track of where a person resides either within delhi or outside delhi. We don't have any sort of order or sequence here. It makes no difference if someone lives in Delhi or Bangalore.

When the categorical characteristic is ordinal, we apply this categorical data encoding strategy. It's critical to keep the sequence in this scenario. As a result, encoding should match the sequence. Each label is turned into an integer value during label encoding. We'll make a variable that holds the categories that indicate a person's educational qualifications.

This categorical data encoding approach is used when the categorical characteristic is ordinal. In this circumstance, it's vital to maintain the sequence. As a result, encoding must be consistent with the sequence. During label encoding, each label is converted to an integer value. We'll create a variable that stores the categories that describe a person's educational background.

This is accomplished using the pre-processing module from `sklearn` package. It has LabelEncoder() function that is used to encode the categorical value into numerical value.

MODULE 3: Feature selection

Feature selection is a technique for limiting the input variable to your model by only utilising useful data and eliminating noise. It's the process of selecting appropriate characteristics for your machine learning model based on the sort of problem you're attempting to answer automatically. This is accomplished by adding or removing key characteristics without altering them. It aids in the reduction of noise in our data as well as the amount of our input data.

Feature Selection Models

There are two types of feature selection models:

1. Supervised Models: Supervised feature selection is a method for selecting features that leverages the output label class. They utilize the goal variables to find variables that can improve the model's efficiency.
2. Unsupervised Models: Unsupervised feature selection refers to a feature selection procedure that does not need the output label class. For unlabeled data, we utilize them.

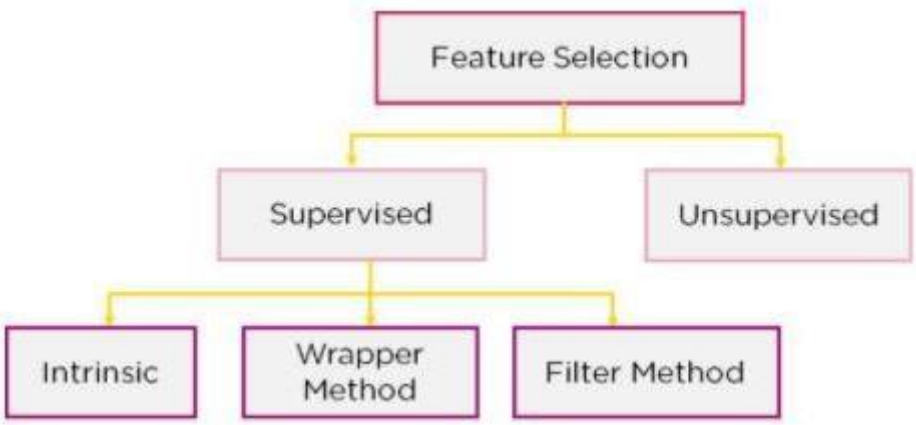


Figure 4: Feature Selection Models

The supervised models may be further divided into three categories:

1. Filter Method:

Features are discarded in this manner depending on their relationship to the output, or how they correlate with the outcome. We utilise correlation to see if the characteristics are favourably or negatively connected with the output labels, and if they are, we drop them. For example, information gain, the Chi-Square Test, Fisher's Score, and so on.

2. Wrapper Method:

We divided our data into subgroups and used this to train a model. We add and eliminate features based on the model's output and retrain the model. It uses a greedy strategy to create subsets and

analyses the accuracy of all potential feature combinations. For example, Forward Selection, Backwards Elimination, and so on.

3. Intrinsic Method:

To construct the optimal subset, this technique incorporates the benefits of both the Filter and Wrapper methods. This approach takes care of the iterative machine training process while keeping the computation cost low. Lasso and Ridge Regression are two examples.

Wrapper approaches use greedy search algorithms to examine all potential feature combinations and pick the one that delivers the best result for a particular machine learning algorithm. One disadvantage of this strategy is that it might be computationally expensive to test all potential combinations of features, especially if the feature collection is big.

Wrapper techniques, as previously stated, can determine the optimum collection of features for a certain algorithm; however, these features may not be optimal for all other machine learning algorithms. Step forward feature selection, Step backwards feature selection, and Exhaustive feature selection are the three types of wrapper approaches for feature selection.

Exhaustive feature selection:

The performance of a machine learning algorithm is tested against all conceivable combinations of the features in the dataframe in exhaustive feature selection. Because it attempts all possible combinations of characteristics and picks the best, the exhaustive search algorithm is the most greedy of all the wrapper techniques presented above. The min features and max features properties in the method can be used to determine the combination's minimum and maximum amount of features. Because it assesses all feature combinations, comprehensive feature selection might be slower than the step forward and step backward methods.

Algorithm of Exhaustive feature selection:

Step 1: Importing the libraries

Step 2: Importing the dataset

Step 3: The model is well defined using random forest classifier

Step 4: Training the model on the Training set

Step 5: Feature that was selected after the step backwards elimination

Step 6: Predicting the Test set using probabilities.

Step 7: Comparing the Test Set with Predicted Values

The above specified algorithm is used to select the best K features from the overall features using random forest classifier in jupyter notebook.

MODEL 4: MODEL BUILDING

In this study, the author created and developed two supervised category algorithms using the same dataset as previously. The algorithms are as follows:

- Logistic Regression
- Support Vector Regression (SVM-R)

Support Vector Regression (SVM-R)

The Support Vector Machine (SVM) is a classification and regression supervised machine learning technique. It's most typically used for classification, but it might also help with regression. SVM locates a hyper-plane that divides the various types of data. A line in two-dimensional space is all that this hyper-plane is.

The technique of guided learning to predict discrete values, Support Vector Regression is employed. Both SVMs and Support Vector Regression work on the same principle. The main idea of SVR is to find the best fit line. In SVR, the hyperplane with the highest number of points is the best fit line.

Unlike other regression models, the SVR seeks to fit the best line within a threshold value rather than minimising the difference between the real and predicted values. The threshold value is the distance between the hyperplane and the boundary line. The complexity of SVR's fit time rises

more than quadratically with the number of samples, making it difficult to scale to datasets with more than a few tens of thousands of samples.

For large datasets, the linear SVR or SGD Regressor is used. Linear SVR is faster than SVR, however it only considers the linear kernel. The model generated by Support Vector Regression only employs a portion of the training data since the cost function rejects samples whose prediction is close to their aim.

Algorithm of SVR:

Step 1: Importing the libraries

Step 2: Importing the dataset

Step 3: Feature Scaling

Step 4: Training the Support Vector Regression model on the Training set

Step 5: Training the Support Vector Regression model on the Training set

Step 6: Predicting the Test set Results

Step 7: Comparing the Test Set with Predicted Values

Logistic Regression

Supervised learning is an example of logistic regression. It's used to figure out or anticipate the chances of a binary (yes/no) event happening. Using machine learning, logistic regression is utilised to evaluate whether or not a person is likely to be influenced by online education using COVID-19. Because there are only two viable answers to this question: yes, online education is influenced or no, online education is unaffected, binary classification is used.

In this hypothetical situation, id, area of residency, age of subject, time spent in online class, rating of online class experience, medium for online class, time spent on self-study, and other factors might all influence whether a state is affected by COVID-19. ID, region of residence, age of subject, time spent in online class, rating of online class experience, medium for online class,

time spent on self-study, time spent on fitness, time spent on sleep, time spent on social media, preferred social media platform, time spent on TV, number of meals per day, change in weight, health issue during lockdown, stress busters, time utilised (Dependent Variable).

Logistic regression is an excellent match for training data that meets the following criteria.

- A binary or dichotomous result is predicted. (This is true in the case of binary logistic regression.)
- The independent variables, or the elements that determine the result, are unrelated. In other words, the independent variables have little or no multi co-linearity.
- The log probability can be linearly connected to the independent variables.
- The sample size is enormous.

If the training data does not meet the aforementioned conditions, logistic regression may not be acceptable for that application.

Algorithm steps of logistic regression

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)

MODEL 5: PERFORMANCE METRICS

We can evaluate the performance of ML algorithms, classification algorithms, and regression algorithms using a variety of measures. Because the measures for measuring ML performance must be properly chosen. The metric you pick will determine how the performance of ML algorithms is assessed and compared. The measure will have a huge impact on how relevant of various attributes in the end outcome.

Metrics used

Because we are predicting a continuous quantity, regression measurements differ from categorization metrics. Furthermore, evaluating regression is often easier than evaluating categorization. The fundamental metrics for evaluating the regression model are listed below.

a) Mean Absolute Error

The mean absolute error (MAE) is one of the most frequent metrics for calculating the model's prediction error. A single row of data has a prediction error of:

PredictionError = ActualValue-PredictedValue

For each row of data, we must compute prediction errors, obtain their absolute value, and then find the mean of all absolute prediction errors. The formula for MAE is as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

where yi is the anticipated value of ŷi

Because MAE employs the absolute value of the residuals, it cannot tell if the model is doing well or poorly. Because we are adding separate residuals, each one adds linearly to the overall error. As a result, a low MAE indicates that the model is good at forecasting. A big MAE, on the other hand, indicates that your model may have difficulty generalising. Our approach produces flawless predictions with an MAE of 0, although this is improbable in real-world settings.

b) Mean Squared Error

The mean squared difference between the target and forecasted values is used to calculate MSE. Many regression issues employ this number, and greater mistakes have correspondingly larger squared contributions to the mean error.

MSE is calculated using the following formula:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where yi is the anticipated value of ŷi.

Because in MAE, residuals contribute linearly to the overall error, but in MSE, the error rises quadratically with each residual, MSE will almost always be larger than MAE. Because MSE heavily penalises the heavy outliers, it is used to determine the extent to which the model fits the data.

c) The coefficient of determination (R2 score)

The R2 value indicates how well the regression predictions match the actual data points.

The following formula is used to compute the value of R2:

$$R^2 = 1 - \frac{\sum_{1=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{1=1}^{N} (y_i - \bar{y})^2}$$

determination coefficient R2 score, where I is the expected value of yi and is the mean of the observed data, computed as formula for calculating the average.

$$\bar{y} = \sum_{i=1}^{N} y_i$$

R2 can have any value between 0 and 1. The regression predictions exactly fit the data if the value is 1.

d) Score:

The scoring is done using a random sample of data and a linear regression technique. The method incorporates scoring systems for variables with linear relationships. If you need to score two different data values, each of which is linked to five different features, you'll need to run 25 linear regression analyses. It returns the coefficient of determination R^2 of the prediction dataset.

Regression Metrics: Important Hints

- We must always ensure that the regression problem assessment measure we chose penalises errors in a way that represents the repercussions of those errors for our application's business, organisational, or user demands.

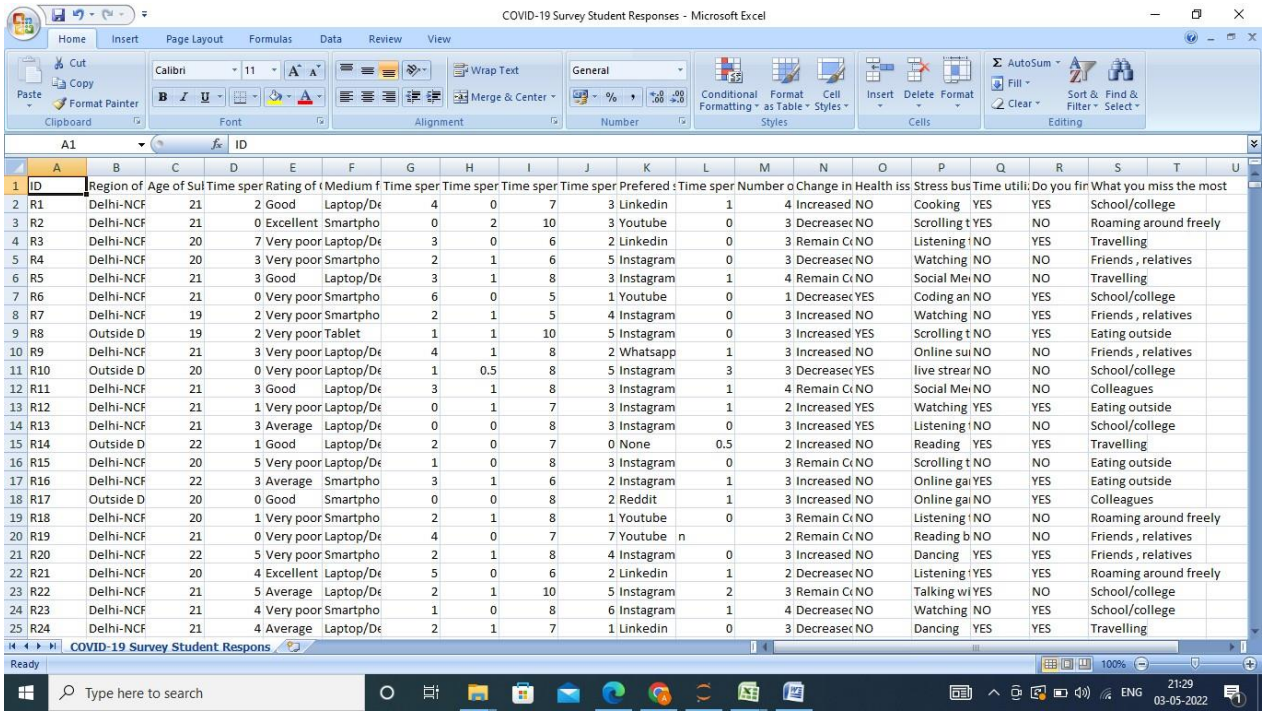
- Outliers in the data might have a negative impact on the overall R2 or MSE values. Because it employs the absolute value, MAE is resistant to the existence of outliers. As a result, if avoiding outliers is crucial to us, we can utilise the MAE score.
- Because it does not reflect huge residuals, MAE is the best statistic for distinguishing between various models.
- The MSE metrics should be used if we want to guarantee that our model takes outliers into account more.

RESULTS AND DISSCUSION

RESULT AND DISCUSSION:

Dataset:

Figure 4.1 shows the dataset that is used for identifying the impact of online education during covid19.



ID	Region of	Age of Sul	Time sper	Rating of	Medium f	Time sper	Time sper	Time sper	Time sper	Prefered	Time sper	Number o	Change in	Health iss	Stress bus	Time utili	Do you fir	What you miss the most
R1	Delhi-NCF	21	2	Good	Laptop/De	4	0	7	3	Linkedin	1	4	Increased	NO	Cooking	YES	YES	School/college
R2	Delhi-NCF	21	0	Excellent	Smartpho	0	2	10	3	Youtube	0	3	Decreasec	NO	Scrolling t	YES	NO	Roaming around freely
R3	Delhi-NCF	20	7	Very poor	Laptop/De	3	0	6	2	Linkedin	0	3	Remain Cc	NO	Listening t	NO	YES	Travelling
R4	Delhi-NCF	20	3	Very poor	Smartpho	2	1	6	5	Instagram	0	3	Decreasec	NO	Watching	NO	NO	Friends , relatives
R5	Delhi-NCF	21	3	Good	Laptop/De	3	1	8	3	Instagram	1	4	Remain Cc	NO	Social Me	NO	NO	Travelling
R6	Delhi-NCF	21	0	Very poor	Smartpho	6	0	5	1	Youtube	0	1	Decreasec	YES	Coding an	NO	YES	School/college
R7	Delhi-NCF	19	2	Very poor	Smartpho	2	1	5	4	Instagram	0	3	Increased	NO	Watching	NO	YES	Friends , relatives
R8	Outside D	19	2	Very poor	Tablet	1	1	10	5	Instagram	0	3	Increased	YES	Scrolling t	NO	YES	Eating outside
R9	Delhi-NCF	21	3	Very poor	Laptop/De	4	1	8	2	Whatsapp	1	3	Increased	NO	Online sul	NO	NO	Friends , relatives
R10	Outside D	20	0	Very poor	Laptop/De	1	0.5	8	5	Instagram	3	3	Decreasec	YES	live strear	NO	NO	School/college
R11	Delhi-NCF	21	3	Good	Laptop/De	3	1	8	3	Instagram	1	4	Remain Cc	NO	Social Me	NO	NO	Colleagues
R12	Delhi-NCF	21	1	Very poor	Laptop/De	0	1	7	3	Instagram	1	2	Increased	YES	Watching	YES	YES	Eating outside
R13	Delhi-NCF	21	3	Average	Laptop/De	0	0	8	3	Instagram	0	3	Increased	YES	Listening t	NO	NO	School/college
R14	Outside D	22	1	Good	Laptop/De	2	0	7	0	None	0.5	2	Increased	NO	Reading	YES	YES	Travelling
R15	Delhi-NCF	20	5	Very poor	Laptop/De	1	0	8	3	Instagram	0	3	Remain Cc	NO	Scrolling t	NO	NO	Eating outside
R16	Delhi-NCF	22	3	Average	Smartpho	3	1	6	2	Instagram	1	3	Increased	NO	Online ga	YES	YES	Eating outside
R17	Outside D	20	0	Good	Smartpho	0	0	8	2	Reddit	1	3	Increased	NO	Online ga	NO	YES	Colleagues
R18	Delhi-NCF	20	1	Very poor	Smartpho	2	1	8	1	Youtube	0	3	Remain Cc	NO	Listening t	NO	NO	Roaming around freely
R19	Delhi-NCF	21	0	Very poor	Laptop/De	4	0	7	7	Youtube	n	2	Remain Cc	NO	Reading b	NO	NO	Friends , relatives
R20	Delhi-NCF	22	5	Very poor	Smartpho	2	1	8	4	Instagram	0	3	Increased	NO	Dancing	YES	YES	Friends , relatives
R21	Delhi-NCF	20	4	Excellent	Laptop/De	5	0	6	2	Linkedin	1	2	Decreasec	NO	Listening t	YES	YES	Roaming around freely
R22	Delhi-NCF	21	5	Average	Laptop/De	2	1	10	5	Instagram	2	3	Remain Cc	NO	Talking wi	YES	NO	School/college
R23	Delhi-NCF	21	4	Very poor	Smartpho	1	0	8	6	Instagram	1	4	Decreasec	NO	Watching	NO	YES	School/college
R24	Delhi-NCF	21	4	Average	Laptop/De	2	1	7	1	Linkedin	0	3	Decreasec	NO	Dancing	YES	YES	Travelling

Figure 4.1: benchmark dataset

Description:

The above figure shows the dataset is gathered from kaggle repository for analyzing the impact of covid-19 in online education. The dataframe shows the benchmark data that is used to analyse and detect the impact of online education during covid19.

Label Encoding

Figure 4.2 shows the label encoding technique applied for categorical data

```
In [118]: from sklearn.preprocessing import LabelEncoder

          obj_df=obj_df.apply(LabelEncoder().fit_transform)
          obj_df
```

Out[118]:

ID	Region of residence	Rating of Online Class experience	Medium for online class	Prefered social media platform	Time spent on TV	Change in your weight	Health issue during lockdown	Stress busters	Time utilized	Do you find yourself more connected with your family, close friends , relatives ?	What you miss the most
R1	0	3	2	3	9	1	0	9	1	1	40
R2	0	2	3	15	1	0	0	48	1	0	37
R3	0	5	2	3	1	2	0	29	0	1	44
R4	0	5	3	2	1	0	0	65	0	0	16
R5	0	3	2	2	9	2	0	54	0	0	44
...	...	...	...	...	...	...	...	...	...	...	...
R16705	1	1	2	2	9	2	0	39	1	1	11
R16706	1	1	2	15	6	2	0	46	0	1	18
R16707	1	2	4	15	18	1	0	29	0	0	37

Figure 4.2: Converting categorical data into numerical data

Description

The dataset is encoded using label encoder to convert the categorical type into numerical type. Label encoding technique is a type of data transformation technique that is used to convert the data type into numerical one.

Feature selection

Figure 4.3 shows the select top features using exhaustive feature selection algorithm

'Age of Subject'
'Time spent on Online Class’
'Rating of Online Class experience'
'Medium for online class'
'Time spent on self study'
'Time spent on fitness'
'Time spent on sleep'
'Time spent on social media'
'Prefered social media platform'
'Time spent on TV'
'Number of meals per day'
'Change in your weight'
'Health issue during lockdown'
'Stress busters'
'Time utilized',
'Do you find yourself more connected with your family, close friends , relatives ?',
'What you miss the most'

Figure 4.3: top features selected using exhaustive feature selection algorithm

Description:

The exhaustive feature selection using random forest is applied to obtain the important feature from the existing 19 features.

Figure 4.4 shows the top features from Exhaustive feature selection using random forest in bar chart

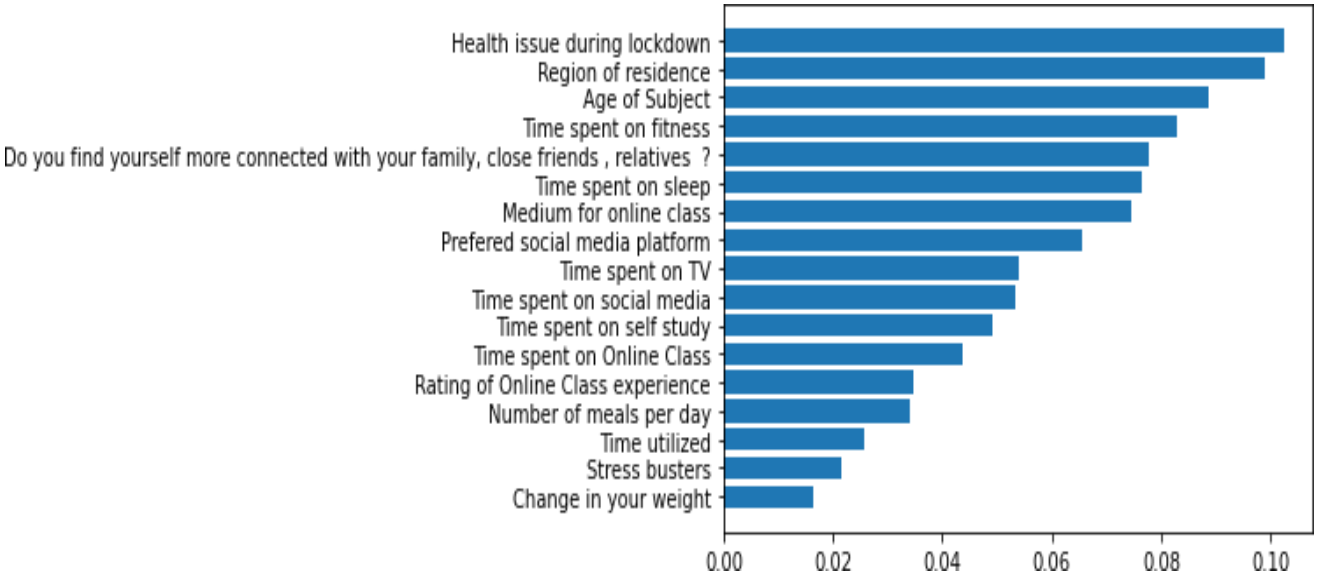


Figure 4.4: Top features from Exhaustive feature selection using random forest in bar chart

Description

Bar chart is a type of data visualization technique that is used to show the best features based on score. The top 17 feature is obtained from the existing 19 features using exhaustive feature selection technique using random forest.

4.4. Train-Test split

The reduced dataset is split into ratio of 80-20 for train data and test data. The two models such as SVR and LR is trained using train data. Both the trained models are tested using test data.

4.5 Model evaluation

Since the model is used for regression problem, the performance metrics such as model score, mean absolute error, mean squared error and r2 score is used to evaluate and detect the impact of covid19 in online education.

Figure 4.5 shows the performance evaluation for analyzing the impact of online education during covid-19 in bar chart

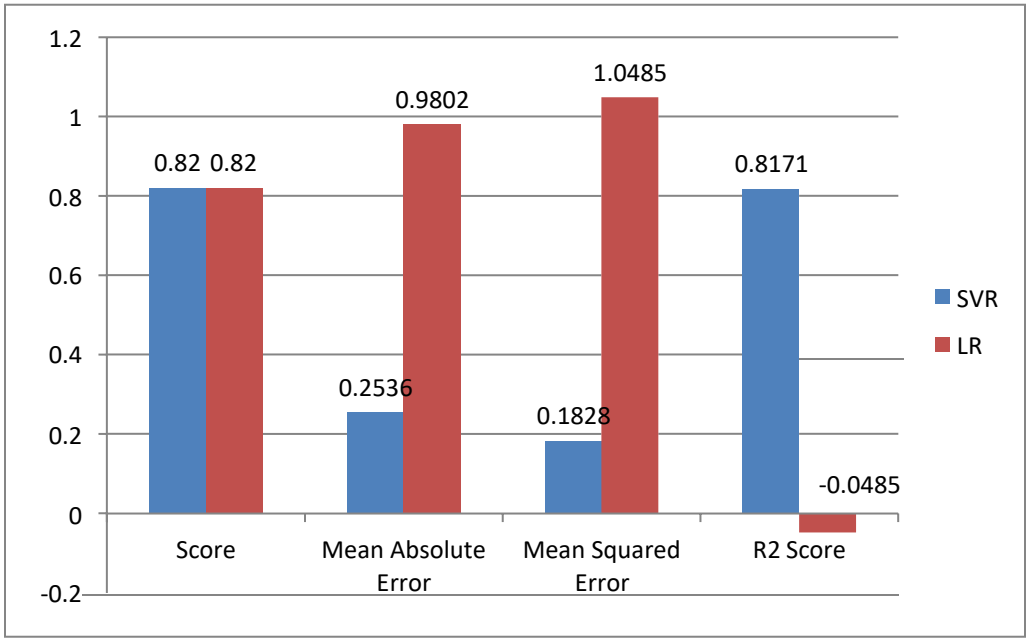


Figure 4.5: performance evaluation for analyzing the impact of online education during covid-19 in bar chart

Description:

The model score for support vector regression and logistic regression is 82%. The model score is equal for Support vector regression (SVR) and Logistic Regression (LR). Support vector regression achieved mean absolute error of 25% and mean squared error of 18%. Logistic Regression obtained mean absolute error of 98% and mean squared error of more than 100%. Support vector regression obtained less mean absolute error and mean squared error than Logistic regression. SVR produce more r2 score with 81% than LR with -0.04%. Hence it shows that support vector regression achieves better result than logistic regression for analyzing the impact of online education during covid-19. It is observed that the people outside the delhi is impacted more than the people within delhi for online education in covid-19 wave.

CONCLUSION

CONCLUSION

The dataset is gathered from kaggle for analyzing the impact of covid 19 in online education. The dataset applies various pre-processing techniques namely label encoding and exhaustive feature selection techniques using random forest. The top 17 feature is obtained from 19 features. The selected features are used train the set of models namely SVR and LR. The models are tested using test data. They are evaluated using performance metrics such as model score, Mean absolute error, mean squared error and r2 score for regression problem. Hence, it is concluded that SVR produce less error and more r2 score than LR. It successfully detects the impact of covid19 in online education. It is analyzed that the impact of covid19 is effective for underprivileged kids and private teachers. Hence, the virtual learning replaced the instructor during covid19 pandemic. But the internet access is much more in delhi than other regions for online education.

SCOPE FOR FUTURE ENHANCEMENT

SCOPE OF FUTURE RESEARCH

In near future, the impact of covid19 for online education can be analyzed using other unsupervised machine learning algorithms such as PCA and K-Means. For the larger dataset to be analysed, the deep learning techniques can be applied for detecting the impact of online education during covid-19 outbreak. The data collection for online education impact can be done in national level in near future to study the impact of online education.

REFERENCES

REFERENCES

Balakrishnan, G.K. and Coetzee, D. (2013), "Predicting student retention in massive open online courses using hidden Markov models", Technical Report No. EECS-2013-109, UC Berkeley, available at: www2.eecs.berkeley.edu/Pubs/TechRpts/2013/ (accessed September 3, 2019).

Berry, L.J. (2017), "Using learning analytics to predict academic success in online and face-to-face learning environments", Dissertation for the Degree of Doctor of Education, Boise State University, Boise, ID, March 6, available at: https://scholarworks.boisestate.edu/cgi/viewcontent.cgi?article=2317&context=td (accessed September 3, 2019).

Brown, M. (2012), "Learning analytics: moving from concept to practice", EDUCAUSE Learning Initiative, July, available at: https://library.educause.edu/-/media/files/library/2012/7/elib1203-pdf (accessed September 3, 2019).

Burgos, C., Campanario, M.L., Peña, D., Lara, J.A., Lizcano, D. and Martínez, M.A. (2018), "Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout", Computers & Electrical Engineering, Vol. 66, pp. 541-556.

Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F. and Alowibdi, J.S. (2017), "Predicting student performance using advanced learning analytics", Proceedings of the 26th International Conference on World Wide Web Companion, Perth, pp. 415-421.

Dietzuhler, B. and Hurn, J.E. (2013), "Using learning analytics to predict (and improve) student success: a faculty perspective", Journal of Interactive Online Learning, Vol. 12 No. 2, pp. 17-26.

Jiang, S., Williams, A.E., Schenke, K., Warschauer, M. and O'Dowd, D.K. (2014), "Predicting MOOC performance with week 1 behavior", Proceedings of the 7th International Conference on Educational Data Mining, International Educational Data Mining Society, London, pp. 273-275

Kloft, M., Stiehler, F., Zheng, Z. and Pinkwart, N. (2014), "Predicting MOOC dropout over weeks using machine learning methods", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) in Doha, Association for Computational Linguistics, pp. 60-65.

Mason, C., Twomey, J., Wright, D. and Whitman, L. (2018), "Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression", Research in Higher Education, Vol. 59 No. 3, pp. 382-400.

Mitchell, M. (2003), Machine Learning (translated by Zeng huayin et al.), China Machine Press, Beijing Osman, A.H. (2016), "An evaluation model of teaching assistant using artificial neural network", VAWKUM Transactions on Computer Sciences, Vol. 11 No. 2, pp. 10-14.

Qiu, J., Tang, J., Liu, T.X., Gong, J., Zhang, C., Zhang, Q. and Xue, Y. (2016), "Modeling and predicting learning behavior in MOOCs", Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, ACM Press, San Francisco, CA, pp. 93-102.

Romero, C., López, M.I., Luna, J.M. and Ventura, S. (2013), "Predicting students' final performance from participation in on-line discussion forums", Computers & Education, Vol. 68 No. C, pp. 458-472.

Sorour, S.E. and Mine, T. (2016), "Building an interpretable model of predicting student performance using comment data mining", 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Kumamoto, IEEE Press, pp. 285-291.

Valiente, J.A.R., Merino, P.J.M., Leony, D. and Kloos, C.D. (2015), "ALAS-KA: a learning analytics extension for better understanding the learning process in the khan academy platform", Computers in Human Behavior, Vol. 47, pp. 139-148.

Villagrá-Arnedo, C.J., Gallego-Durán, F.J., Llorens-Largo, F., Compañ-Rosique, P., Satorre-Cuerda, R. and Molina-Carmona, R. (2017), "Improving the expressiveness of black-box models for predicting student performance", Computers in Human Behavior, Vol. 72, pp. 621-631.

Zacharis, N.Z. (2016), "Predicting student academic performance in blended learning using artificial neural networks", International Journal of Artificial Intelligence and Applications, Vol. 7 No. 5, pp. 17-29.

Fojtík, R. (2018). Problems Of Distance Education. International Journal of Information and Communication Technologies in Education, 7(1), 14–23.

Xiao, J., & He, W. (2020). The Emergency Online Classes During COVID-19 Pandemic: A Chinese University Case Study. Asian Journal of Distance Education, 15(2), 21–36.

Pete, J., & Soko, J. (2020). Preparedness for online learning in the context of Covid-19 in selected SubSaharan African countries. Asian Journal of Distance Education, 15(2), 37–47

Yavuz, M., Kayalı, B., Balat, Ş., & Karaman, S. (2020). Review Of Distance Learning Applications In The Universities In The Covid-19 Period. Milli Eğitim Dergisi [National Education Journal], Special Issue, 49(1), 129–154.

Yavuz, M., Kayalı, B., Balat, Ş., & Karaman, S. (2020). Review Of Distance Learning Applications In The Universities In The Covid-19 Period. Milli Eğitim Dergisi [National Education Journal], Special Issue, 49(1), 129–154.

Şen, Ö., & Kızılcalıoğlu, G. (2020). Determining The Views Of University Students And Academics On Distance Education During The Covid-19 Pandemic. International Journal of 3D Printing Technologies and Digital Industry, 4(3), 239–252.

Yılmaz, E., & Güner, B. (2020). Evaluation Of Distance Education Services Provided To Students At Different Education Levels According To Parents' Opinions. Milli Eğitim Dergisi [National Education Journal], Special Issue, 49(1), 477–503.

Keskin, M., & Özer Kaya, D. (2020). Evaluation of Students' Feedbacks on Web-Based Distance Education in the COVID-19 Process. İzmir Katip Çelebi University Faculty of Health Science Journal, 5(2), 59–67.

Özoran, B. A. (2020). An Analysis on Twitter Usage of Brands in Covid-19 Pandemics. Van Yüzüncü Yıl University The Journal of Social Sciences Institute, Outbreak Diseases Special Issue, 429–458.

ANNEXURE

ANNEXURE

```
import numpy as np

from sklearn.datasets import load_boston

from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import train_test_split

from sklearn.feature_selection import RFECV

import matplotlib.pyplot as plt

import pandas as pd


survey = pd.read_csv('COVID-19 Survey Student Responses.csv', header=0, index_col=0)

survey

obj_df = survey.select_dtypes(include=['object']).copy()

obj_df.head()

obj_df=obj_df.fillna("0")

obj_df.isnull().value_counts()


from sklearn.preprocessing import LabelEncoder


obj_df=obj_df.apply(LabelEncoder().fit_transform)

obj_df

obj_df.iloc[:, 9]
```

```
survey["Rating of Online Class experience"] = obj_df["Rating of Online Class experience"]

survey["Medium for online class"] = obj_df["Medium for online class"]

survey["Prefered social media platform"] = obj_df["Prefered social media platform"]

survey["Change in your weight"] = obj_df["Change in your weight"]

survey["Time spent on TV"] = obj_df["Time spent on TV"]

survey["Health issue during lockdown"] = obj_df["Health issue during lockdown"]

survey["Stress busters"] = obj_df["Stress busters"]

survey["Time utilized"] = obj_df["Time utilized"]

survey.iloc[:, 16] = obj_df.iloc[:, 9]

survey["What you miss the most"] = obj_df["What you miss the most"]

survey.head()


import pandas as pd

from sklearn.feature_selection import SelectFromModel


y=survey['Region of residence']

X=survey.drop('Region of residence', axis=1)

y, X


from sklearn import preprocessing

# label_encoder object knows how to understand word labels.
```

```
label_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'species'.

y=label_encoder.fit_transform(y)

df=pd.DataFrame()


df['y']=y


df


y=df['y'].values

y


X_train,X_test, y_train,y_test  = train_test_split(X,y,test_size=0.3)

X.shape, y.shape, X_train.shape, y_train.shape


rf = RandomForestRegressor(random_state=0)

rf.fit(X_train,y_train)


features=survey.columns


f_i = list(zip(features,rf.feature_importances_))
```

```
f_i.sort(key = lambda x : x[1])

plt.barh([x[0] for x in f_i],[x[1] for x in f_i])


plt.show()


rfe = RFECV(rf,cv=5,scoring="neg_mean_squared_error")

rfe.fit(X_train,y_train)

rfe.get_support(),

features=np.delete(features,0)


selected_features = np.array(features)[rfe.get_support()]

selected_features


from sklearn.preprocessing import StandardScaler

sc_X = StandardScaler()

sc_y = StandardScaler()

X_train = sc_X.fit_transform(X_train)

y_train = sc_y.fit_transform(y_train.reshape(-1, 1))


#SVR (Support Vector Regression)

from sklearn.svm import SVR
```

```
regressor = SVR(kernel='rbf')

regressor.fit(X_train,y_train)


X_test=StandardScaler().fit_transform(X_test)

y_test=StandardScaler().fit_transform(y_test.reshape(-1, 1))


y_pred = regressor.predict(X_test)

print('Accuracy of Support Vector regression classifier on test set: {:.2f}'.format(regressor.score(X_test, y_test)))


y_test.reshape(1, -1), y_pred.reshape(1, -1)


from sklearn.metrics import mean_absolute_error

print(mean_absolute_error(y_test, y_pred))


from sklearn.metrics import mean_squared_error

print(mean_squared_error(y_test, y_pred))


from sklearn.metrics import r2_score

print(r2_score(y_test, y_pred))
```

```
#logistic regression

X_train, y_train

y_train=y_train.astype(int)


from sklearn.linear_model import LogisticRegression

clf = LogisticRegression()

clf.fit(X_train, y_train)


y_pred = clf.predict(X_test)


print('Accuracy of Logistic regression classifier on test set: {:.2f}'.format(regressor.score(X_test, y_test)))


print(mean_absolute_error(y_test, y_pred))

print(mean_squared_error(y_test, y_pred))

#print(root_mean_squared_error(y_test, y_pred))

print(r2_score(y_test, y_pred))
```