ANALYZING THE RISK FACTORS AND PREDICTING THE LEARNING ABILITY OF STUDENTS DURING PANDEMIC AND COMPARING MACHINE LEARNING ALGORITHMS USING ORANGE TOOL.

V. Belsini Gladshiya¹, Dr. K. Sharmila²,
¹Research Scholar, Department of Computer Science, VISTAS, Assistant Professor, Department of Computer Science, Agurchand Mannull Jain College, Chennai, India.
²Research Supervisor & Associate Professor, Department of Computer Science, VISTAS, Chennai, India.

ABSTRACT

A sudden revamp in the field of education due to Covid-19 made educators, learners and educational institutions set foot into a back breaking environment in education. For the most part the students are facing barriers in learning. Since e-learning plays a radical way for the students to learn, they face towards some difficulties in e-learning, missing the opportunities most probably the students from rural place, middle way families etc. Today Machine learning Algorithms effectuate in data analysis for classification and predicting the facts according to the data. It acts as an eminent role in the field of education to analyze the activities of students and making fine predictions to support them to take better decisions. This research study intends the analysis of student data base to determine the ability and disability factors to pursue the students in learning. This study also focuses better factors in learning process in future. It also gives the risk factors of the students during their online learning. The data is collected by means of a questionnaire in the form of Google forms reporting the self situation of the students. Orange tool is used for analyzing the data. Different Machine Learning algorithms are used to manipulate the data and the comparisons of algorithms have done according to its inferences.

Key words: Machine Learning Algorithms, Orange tool, learning abilities

I. INTRODUCTION

The pandemic situation of Covid-19 made challenges and strain in the field of education. Hence there is a downturn in the process of learning. Students are the building blocks of the learning task. Nowadays educational institutions introduce e-learning systems for learning. Despite the fact, technology has improved and students are learning through digital devices there exists some dilemma the students face during learning. “Early detection of student risk improves their success.”[1].

Machine Learning Algorithms are the boon in the field of Data Analytics for prediction by analyzing the data. “Machine Learning enhanced research works in education for decision making and Policy formulation”[2].

Predictive analytics is the process of attaining the future probabilities and expectations of an event from the data gathered. This study focuses the risk factors of the student’s e-learning course of action. The classification of various possibilities of difficulties and predicting the student’s ability in learning strategy are under taken using machine learning tools and algorithms.

II. MACHINE LEARNING ALGORITHMS

The subfield of Artificial Intelligence and Computer Science which deals with set of algorithms and statistical models to study the outcomes and patterns without direct programming is termed as Machine Learning. Machine Learning algorithms marked a predominant role in Predictive analytics.
Predictive Analytics is the procedure of taking the information from the existing data sets to establish new patterns and future probabilities to take better decisions according to it.

**Types of Machine Learning:**

The major types of machine Learning are

![Figure 1: Types of Machine Learning](image)

**A Supervised Learning:**

The process of inferring a function or a model from the training data set is called Supervised Learning. If \((x)\) is an input variable the supervised learning will predict the output variable\((Y)\) according to \((X)\). Hence the mapping function is given as \(Y=f(X)\). The supervised algorithm detects the relationships between the variables in the data set.

**B Unsupervised Learning:**

Unsupervised Learning algorithms extract the output from the unorganized input data set. It consists of Clustering and Association. The aim of unsupervised learning is to create a model or distribution to know more about the data.

**C Reinforcement Learning:**

Reinforcement is a machine learning paradigm in which the agent learns to attain a goal in an undetermined composite environment. Its aim is to maximize the total reward with the combination of neural networks.

### III. ORANGE TOOL

The open source tool for data analysis and data visualization is Orange. It has multiple classification and regression algorithms used for machine learning and containing the th features for data analytics. Orange is composed of Python libraries and run in a window terminal and integrated environments like Pycharm, Pythonwin or iPython. It contains canvas interface in which user places widgets and creates a data analysis workflow.

**Features of Orange:**

- Orange runs on windows, MacOSX and Linux Operating Systems.
- The files can be read in native and other data formats.
- The data analysis process can be done using visual programming.
- Different visualization objects such as data reading, shows data table, predicting the data, comparing the machine learning algorithms can be done easily.

**Applications:**

- Orange tool was used in the field of medicine to study the data analysis in Clinical Decision making [3].
The technical analysis and prediction in share prices can be done using orange tool for companies.[4]

Orange tool aid in predictive modeling, data analytics, subset selection and empirical analysis. Data manipulation and data transformation can also be performed.[5]

IV. LITERATURE REVIEW

There are many works done in analyzing the students risk in the field of education. Some of the related works would be given as literature review so that the application of the machine learning algorithms and predictive analytical tools used to analyze the risk of the students in the field of education with various applications.

Analyzing and predicting the performance of students by using machine learning algorithms. Many research papers were analyzed and using the machine learning algorithms students drop out and performance were determined.[6] (Juan L. Rastrollo-Guerrero, Juan A. Gómez-Pulido * and Arturo Durán-Domínguez, 2020)

Measuring and enhancing the performance of college students and the performance were analyzed predicted the final marks to provide information to the administration, decision makers to direct the students to improve their performance.[7] (Safaa Alhusban, Mohammed Shatnawi, Muneer Bani Yasin, Ismail, Hmeidi , 2020)

Analyzing and predicting the students risk in K12 multimodal online environment.[8] (Hang Li, Wenbiao Ding, Zitao Liu, 2020).

Predicting the graduating students in tertiary institution and comparing the performance of machine learning algorithms. In this study Artificial Neural Networks performed best.[9] (Ajinaja Micheal Olalekan, Ojonukpe Sylvester Egwuche, Sylvester Oluyemi Olutunji, 2020)

The Academic performance of the students had predicted and identifies the factors contributed to the students in academic through Supervised Machine Learning algorithms using Weka tool.[10] (Engr. Sana Bhutto, Jamshoro, 2020)

Predicting the graduating students in tertiary institution and comparing the performance of machine learning algorithms. In this study Artificial Neural Networks performed best.[9] (Ajinaja Micheal Olalekan, Ojonukpe Sylvester Egwuche, Sylvester Oluyemi Olutunji, 2020)

Analyzing and predicting the students risk in K12 multimodal online environment.[8] (Hang Li, Wenbiao Ding, Zitao Liu, 2020).


Predicting the risk of the students in Virtual learning environment in University and proposed a reduced training vector-based support vector machine (RTV-SVM) for their accurate prediction.[14] (SKwok Tai Chuia,* Dennis Chun Lok Fungb, Miltiadis D. Lytrasc, & Tin Miu Lamb, 2018)


V. DATASET DESCRIPTION

The set of data used in this study is collected from the II and III year under graduate students through the questionnaire created in Google form and circulated though the internet. Around 115 responses were taken and it is formatted as csv file format. There are nine attributes having categorical values. The risk factors are selected according to the parameters such as device, network, residence, interest, awareness, time management, social media usage, e-mail checking.

Using Orange tool the data is processed and the probability classification of students using the attributes has found using Sieve diagram. The table below shows the students details according to the categorical data.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Classification and number of probabilities and percentage for each category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device used</td>
<td>Smart Phone: 5.4%, Laptop: 2.2%, Desktop: - Tablet: -</td>
</tr>
<tr>
<td>108.94%</td>
<td></td>
</tr>
</tbody>
</table>

www.turkjphysiotherrehabil.org
<table>
<thead>
<tr>
<th>Residence</th>
<th>Village</th>
<th>Town</th>
<th>City</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15,13%</td>
<td>22,19%</td>
<td>78,68%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network status</th>
<th>Very fast</th>
<th>Fast</th>
<th>Slow</th>
<th>Very slow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8,7%</td>
<td>63,55%</td>
<td>33,29%</td>
<td>11,10%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interest in e-learning</th>
<th>Very interested</th>
<th>Interested</th>
<th>Some</th>
<th>No interest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13,11%</td>
<td>65,57%</td>
<td>26,23%</td>
<td>11,10%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Awareness in e-learning</th>
<th>Well known</th>
<th>Known</th>
<th>Some what</th>
<th>Not known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23,20%</td>
<td>59,51%</td>
<td>21,18%</td>
<td>12,10%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time for e-learning</th>
<th>More than two hours</th>
<th>Two hours</th>
<th>One hour</th>
<th>Not decided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14,12%</td>
<td>27,23%</td>
<td>36,31%</td>
<td>38,33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time for social media</th>
<th>More than two hours</th>
<th>Two hours</th>
<th>One hour</th>
<th>Not decided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28,24%</td>
<td>37,32%</td>
<td>37,32%</td>
<td>26,23%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Social Media</th>
<th>What’s app</th>
<th>Instagram</th>
<th>Facebook</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>45,39%</td>
<td>27,23%</td>
<td>14,12%</td>
<td>36,31%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Checking e-mail</th>
<th>Daily</th>
<th>Two days</th>
<th>Weekly</th>
<th>Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63,55%</td>
<td>21,18%</td>
<td>17,15%</td>
<td>14,12%</td>
</tr>
</tbody>
</table>

Table: 1 Attributes and Classification

From this the risk factors are identified as slow network, not interested, Not having the awareness, Not decided to spend time to study, Spending more time in social media, type of social media they use and period of checking e-mail. According to the risk factors the data is processed to predict the ability of e-learning during this period.

VI. METHODOLOGY

Figure 2: Methodology
The CSV file formatted is loaded to the Orange tool and the target variable is given. The data is a classified using classification tree and it is visualized using tree viewer. The visualization is given as.

![Figure 3: Visualization of Classification Using Tree Viewer](image)

The data is analyzed with different algorithms and its results are noted.

i) Logistic Regression:

Logistic regression (LR) is one of the most widely used methods in data analysis and particularly used for classifying binary data. Logistic regression is well suited algorithm which describes the relationships between a outcome variable which is categorical and more predictor variables which is continuous. When logistic regression is applied to a new data it produces a probability ranging from zero to one that belongs to the target variable and the known predictor variables. Using the probability, the data analyst can make decisions making according to the data.

ii) Naive Bayes:

Naive Bayes is the simplest form of Bayesian network, in which all attributes are independently given the value of the class variable. This feature is called conditional independence. The algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset.

iii) Random Forest:

Random forest is an ensemble model used in predictive analytics. It takes a group of decision trees to create the model. It is a model averaging procedure in which each tree is established according to the bootstrap sample of the data set. It is flexible and easy to use in machine learning and it is used both for classification and regression methods.

iv) Support Vector Machine (SVM):

The data classifier algorithm assigning new data elements to one of labeled categories is called support vector machine (SVM) which is commonly used in predictive analysis. It is a binary classification algorithm which has two possible target values. It is meant for its popularity of machine learning method for classification, regression, & other learning tasks and mostly used in classification problems.

VII. RESULT AND DISCUSSION

From the algorithms used and compared with the accuracy using test score under cross validation. Cross validation is used to find the predictive performance of the model and from the result we can understand how the model can perform to a new data set.

![Figure 4: Results of Classification Methods](image)
From this the classification accuracy of Support Vector Machine is highest accuracy and it has the highest value of F1 score also high in SVM. It can be shown through confusion matrix. A confusion matrix is a table used to find the accuracy or the performance of the machine learning algorithms in a classification model a data set for classification. The confusion matrix for the algorithms is shown below.

Logistic Regression:

![Confusion Matrix](image)

Naive Bayes:

![Confusion Matrix](image)

SVM:

![Confusion Matrix](image)

Random Forest:

![Confusion Matrix](image)

Figure 5: Confusion matrix of Algorithms

In the method random sampling the Logistic regression algorithm performs best. The technique in which every sample has an equal probability that chose randomly for process is called random Sampling. A randomly chosen sample is the impartial description of the total population. Also if the percentage of sample input is increased Logistic regression performs well.
From the results evaluated through cross validation and random sampling the machine learning algorithms such as Logistic regression and Support vector machine is chosen for Predictive analytics of predicting the students who are able to study and disable to study during e-learning during the covid 19 situation.

Due to this covid -19 situation it is essential for e-learning. During the Severe Acute Respiratory Syndrome or SARS in2003 the schools and colleges were closed and the teachers and institution provide learning through online. Colleges had given the education through their websites in Hong hong. The same situation arises in Tamilnadu now but the e-learning is the new phase for our education. So it will be difficult to the students to practice this learning. The institutional persons who have already started the e-learning process noted that the students facing difficulties in network problem were losing their opportunities of learning. From the data collected and the risk factors it is found and according to this around 50% of the students are not able to learn.

It is also noted around 94% of the students use their smart phones for learning. Since the device is very small it will affect the health particularly eyes and brains. This suggestion has been given by many medical advisors through Medias.

Another risk the students face through this learning is slow networks in rural areas and the amount spending for the network by the middle class people. Due to lack of network the well performing students are missing their opportunities.
From the analysis it should be noted to improve the network facilities, giving correct awareness about the learning, motivating the students to get interest in learning and also to provide learning devices.

The future research can be followed through collecting a large data set from various students and identifying other risk factors of students in learning. The engagement level of the students can be found. Machine learning algorithms can also be used effectively in further research and efficient models can be created using orange tool.

REFERENCES


[4] Muhammad Farooq Ishaq, “DATA MINING FORCASTING Oil and Gas Development Company Ltd. Share Prices Using Orange”. Easy Chair preprints are intended for rapid dissemination of research results and are integrated with the rest of Easy Chair. June 10.2020


