Prediction of Student Performance on Virtual Platform Using Machine Learning Algorithm

D.Vamsika[1],Y.Anjani Priya[2] ,Sk.Sameena[3],R.Chaitra[4],S V N Sreenivasu[5]

[1,2,3,4]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

[5]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

vamsika.danda@gmail.com[1] ,anjanipriya93@gmail.com[2] , sameenashaik2801@gmail.com[3] ,regalagaddachaitra@gmail.com[4], drsvnsrinivasu@gmail.com[5]

1.ABSTRACT- Predictive analysis is a machine learning analytical technique that is the focus of this research. The problem of reliable performance prediction is addressed by a number of online learning systems, including number of Courses and number of learning platforms. We are recommending this work by the contrasting techniques like the regression and the classification, which are useful for the prediction modelling to obtain the most accurate outcomes. The prediction model is trained the data and tested the data by using random forest and different models to explain the learning behaviour of the students in connection with their study factors. These predictive model was trained with random forest and it has the higher accuracy.

2.KEYWORDS: Supervised learning, Predictive analysis, Performance prediction, Machine learning, and featureselection.

3.INTRODUCTION

Machine learning ML is the superset of the deep learning, while Artificial intelligence is the superset of ML. ML is helpful in model creation as data is fed to the machine, employing algorithms for additional training and testing on those enormous data so that the machine can conduct operations on its own when given fresh data. There are 3 different categories:

Supervised learning: Throughout the machine's learning phase, supervision is needed. It contains both the input and the desired output, and the model is set up to forecast the desired outcome.

Example: Classification and Regression**.**

Unsupervised learning: The model learns by itself by identifying the pattern inside the dataset without any supervision. Only input is provided, the model self-trains, and output results.

Ex. Clustering and Association.

Because it best fits the needs of predictive analysis, we adopted the supervised learning approach in this study. Prediction is carried out as historical data is gathered, and the model is trained to handle fresh input and forecast the desired result.

This paper's main goal is to use machine learning approaches to identify characteristics related with students' learning development and how they engaged with the virtual learning environment in order to identify students who are at danger of dropping out as early as feasible.

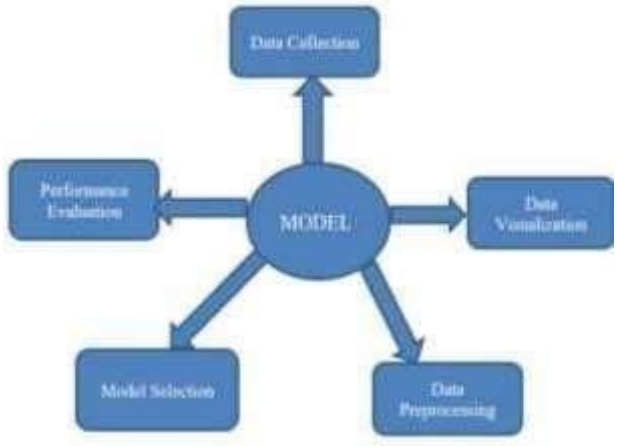

Fig.3(a) 5 Steps involved in the Model

Machine learning can assist students in making judgements about whether or not to continue in the course when it comes to early risk prediction of people.These algorithms had capability of effectively self-preparation and can also deal with the early risk of prediction for students to pass there courses.

4. LITERATURE SURVEY

We must take into account demographic information, a crucial component for early intervention, in order to anticipate children at an earlier stage. Assessment results and Clickstream data are significant time-dependent variables in addition to Demographic data.With an accuracy rate of 85%, it uses a random forestmodel to identify at-risk students in online courses and provides them with intense early assistance.

5. PROPOSED SYSTEM

5.1 Pre-processing

Pre-processing refers to the modifications done to our data before we give it to the algorithm. Data Pre- processing is a technique used to create clean data collections from raw data. In other words, if data are gathered from various sources, they are gathered in an unprocessed way that prevents analysis A data collection should be structured to allow for the usage of multiple Machine learning algorithms in order to achieve the best results.

5.1.1 Dataset

We used the datasets from the Opened university of learning analytics Dataset, which the University makes available. Seven tables include structured data on students. It includes information on demographics (student information and registration), student involvement learning multiple number of courses in the platforms. The student VLE table stores submitted work of the student. The dataset triplet known as student-module presentation contains the assessment

	code_mod	code_pres	id_student	gender	region	highest_ed	imd_band	age_band	num_of_p	studied_cr	disability	final_result
2	AAA	2013J	11391	M	East Anglia	HE Qualific	90-100%	55<=	0	240	N	Pass
3	AAA	2013J	28400	F	Scotland	HE Qualific	20-30%	35-55	0	60	N	Pass
4	AAA	2013J	30268	F	North Wes	A Level or	30-40%	35-55	0	60	Y	Withdrawn
5	AAA	2013J	31604	F	South East	A Level or	50-60%	35-55	0	60	N	Pass
6	AAA	2013J	32885	F	West Midl	Lower Tha	50-60%	0-35	0	60	N	Pass
7	AAA	2013J	38053	M	Wales	A Level or	80-90%	35-55	0	60	N	Pass
8	AAA	2013J	45462	M	Scotland	HE Qualific	30-40%	0-35	0	60	N	Pass
9	AAA	2013J	45642	F	North Wes	A Level or	90-100%	0-35	0	120	N	Pass
10	AAA	2013J	52130	F	East Anglia	A Level or	70-80%	0-35	0	90	N	Pass

Fig 5.1.1 Dataset

5.1.2 Dataset Description

1.Code Presentation term represents the semester of the course for which student joined.

1. The student id identifies the each student by using distinct numbers.
2. Gender specifies the student gender
3. Region specifies location in the region student living.
4. Highest education specifies that student completed before taking the course.
5. The Index of Multiple Deprivation band (IMD band) indicates that the percentile students with an IMD band of below of20 are from the most poor areas.
6. Age specifies age of the Student.
7. The number of priorities refers to the number of students who have already attempted the chosen course and their credit totals The total number of credits a student is enrolled in during the course at the Open University.
8. Disability Student alleges a logical disability of any kind.
9. Final result specifies the final result of the student in the course

5.1.3 Dataset Preprocessing and Merging

Effective preprocessing and the choice of the best classifier have a significant impact on the outcome and precision of the machine learning-based approach. A data mining approach called data preprocessing purifies raw data and produces more pertinent data. Raw data is full of outliers, unneeded features, and many missing values. The required data for the prediction are merged and performed preprocessing on the required dataset.

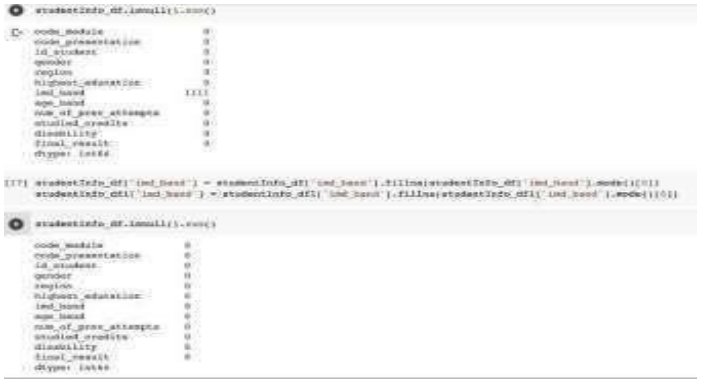

Fig 5.1.3 Replacing the Missing Values in the dataset

5.1.4 Outlier analysis

Since there are many noisy data, cleaning the data is crucial for improved model performance. In this instance, the Turkey's method is employed as a

traditional method of eliminating outliers. We use box plots to show the outliers. Many insightful conclusions may be drawn from these plots, including the fact that each of the data sets contains a significant number of outliers and extreme values.

	id_student	num_of_prev_attempts	studied_credits
count	3.259300e+04	32583.000000	32583.000000
mean	7.066877e+05	0.163225	79.758691
std	5.491673e+05	0.479758	41.071900
min	3.733000e+03	0.000000	30.000000
25%	5.085730e+05	0.000000	60.000000
50%	5.903100e+05	0.000000	60.000000
75%	6.444530e+05	0.000000	120.000000
max	2.716795e+06	6.000000	655.000000

Fig 5.1.4 Outlier Analysis

5.1.5 Splitting the Set of data

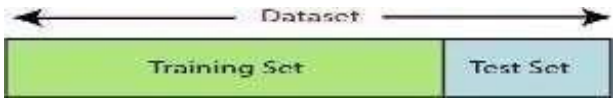

Fig 5.1.5 Spitting of data

The data set is split at a ratio of 20:80. In this scenario, 20% of the data is used for testing, and 80% is used for training.

5.2 Correlation

The correlation matrix is helpful in showing the variables' strong correlation. The matrix shows that because they all have an impact on the prediction variable, each independent variable is significant for it.

Fig 5.2 Correlation Heat Map

5.3 Feature Engineering

It is crucial to draw some conclusions from the available data and create data sets that contain more insightful information. As the dataset already contains the course id and student information, additional calculations are made to the registration**,** click sums, before click, after click, and score that may affect the student's ultimate result.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32593 entries, 0 to 32592
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   code_module           32593 non-null  object
 1   code_presentation     32593 non-null  object
 2   id_student            32593 non-null  int64
 3   gender                32593 non-null  object
 4   region                32593 non-null  object
 5   highest_education     32593 non-null  object
 6   imd_band              31482 non-null  object
 7   age_band              32593 non-null  object
 8   num_of_prev_attempts  32593 non-null  int64
 9   studied_credits       32593 non-null  int64
 10  disability            32593 non-null  object
 11  final_result          32593 non-null  object
dtypes: int64(3), object(9)
memory usage: 3.0+ MB
```

Fig 5.3 Feature engineering

5.4 Model Selection

The ultimate accuracy of the model is estimated once it has been tested using test data. Machine learning models were trained and evaluated using various dataset splits to produce accurate findings. There are four steps that training and testing might take

PHASE-1:DEMOGRAPHIC DATA

PHASE-2:DEMOGRAPHIC and CLICKSTREAM DATA

PHASE-3:DEMOGRAPHIC, CLICKS TRACK,and EVALUATION

FEATURE ENGINEERING:

Merging the various classes, such as Distinction- Pass being combined into the Pass class and Withdrawn-Fail being combined into the Fail class, can help the performance outcomes even more. beat all other models in performance.

5.4.1 Random Forest

A random forest model is created using a large number of decision trees. The programmes areessentially creates a forest out of the forecasted results of trees. The algorithm also employs three random ideas: randomly choosing training data when creating trees, randomly selecting specific subsets of variables when dividing nodes, and using only a small fraction of all variables to split each node in each basic decision tree.

5.5 Data Visualisation

Using various graphs, charts, plots, and other visual aids, data visualization makes it easier to understand the data.

The below graph specifies about the comparison of code modules. Based on this graph we can predict the highly popular module. Based on the graph we predict the students are very much interested towards the BBB ,FFFmodules**.**

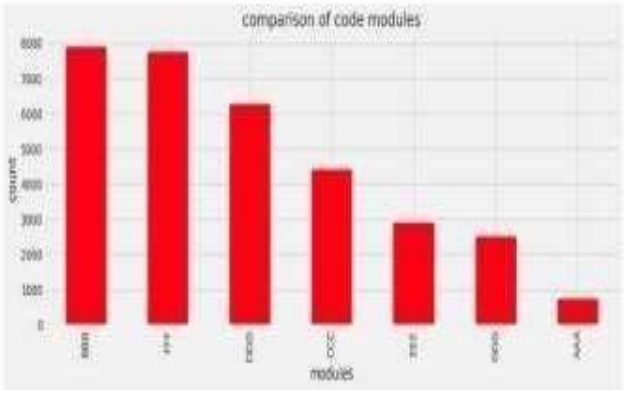

Fig 5.5.1(a) Comparison of Code modules

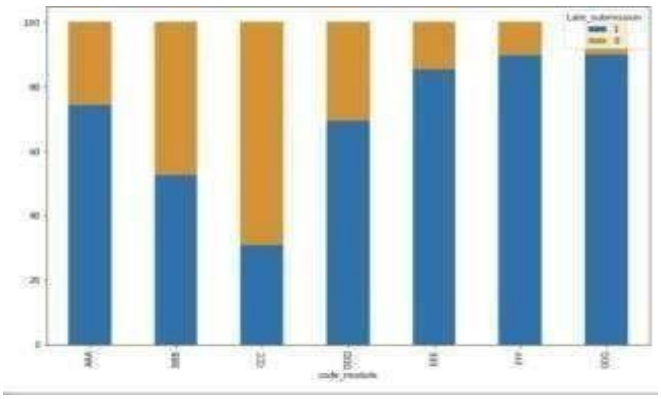

Fig 5.5.1(b) Late submission of assessments

The above graph specifies about the late submissions in the courses. In this graph "0" representing the late submission of assessment and "1" representing the submission of courses by students. Most of the students in FFF and GGG are submitted their assessment

5.6 Proposed Model

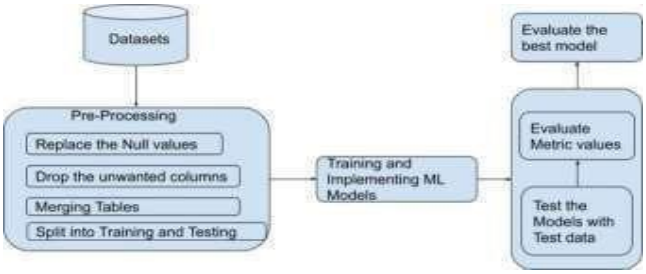

Fig 5.6 Proposed Model Architecture

For the earliest possible prediction of student's performance, divide the course length into 20%,40%,60%,80% and 100% of course completed.

6. RESULT ANALYSIS

The primary objective of this project is the earliest possible identifying students who may be at risk dropouts by leveraging Machine Learning techniques to understand variables associated with the learning behavior of students and how they interact with the

Models	Accuracy
Random Forest	85.15%
Support Vector Machine	79%
Extra Tree Classifier	78%
Adaboost Classifier	75%
Gradient Boost	85%

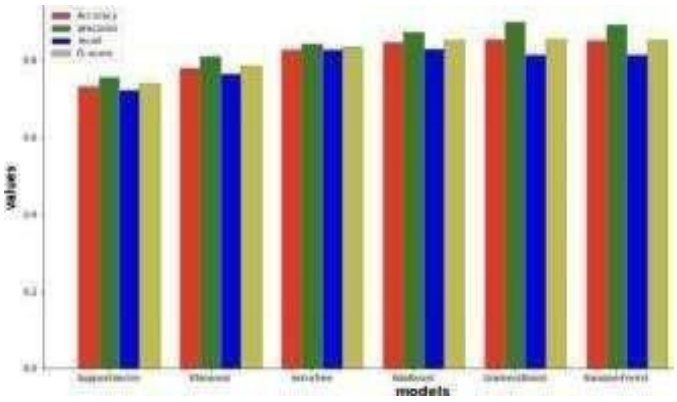

Fig 6(a) Comparison of Models

Virtual Learning Environment. so ,we consider Random Forest as our model with highest accuracy 85.15%.

The above graph specifies the comparision between modules and their accuracy.After the performance of 6 models. Accuracy, Precision,Recall, F1score are predicted.Based on the prediction graphs are plotted**.**

6. OUTPUT SCREENS

Fig 7(a) Data Uploading Screen

Fig 7(b) Output Screen

7. CONCLUSION

The main goal of the project to use the Machine learning algorithm to estimate student risk over a range of course lengths. Four classification metrics were used in the study and for evaluations. The research revealed that using clickstream data and assignment ratings significantly improved the models performance. The best performing Random Forest predictive model (85%)was chosen to forecast student performance. Clickstream data and assessment scores have the biggest influence on the outcome of all the variables.

8. REFERENCES

[1] A. Oritgosa , R. M. Carro, J. Bravo-Agapito, D. Lizcano, J. J. Alcolea, and O. Blanco, ‘‘From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system,’’ IEEE Trans. Learn. Technol., vol. 12, no. 2, pp. 264–277,Apr. 2019.

[2]A. S. Imran, F. Dalipi, and Z. Kastrati, ‘‘Predicting student dropout in a MOOC: An evaluation of a deep neural network model,’’ in Proc. 5th Int. Conf. Comput. Artif. Intell. (ICCAI), 2019, pp. 190–195.

[3] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, ‘‘Predictive learning analytics using deep learning model in MOOCs’ courses videos, ’’ Edu. Inf.Technol.,vol. 6, pp. 1–22, Jul. 2020.

[4] B. Sekeroglu, K. Dimililer, and K. Tuncal, ‘‘Student performance prediction and classification using machinelearning algorithms,’’ in Proc. 8th Int. Conf. Educ. Inf. Technol., Mar. 2019, pp. 7–11.

[5] A.Behr,M.Giese,and K.Theune,”Early Prediction of University droputs-A random forest approach,”J.Nat. stat., vol. 1,pp. 743-789,Feb.2020.

[6] J. Figueroa-Ca nas and T. Sancho-Vinuesa, ‘‘Predicting early dropout student is a matter of checking completed quizzes: The case of an online statistics module,’’ in Proc. LASI- SPAIN, 2019, pp. 100–111.

[7] J. Xu, K. H. Moon, and M. van der Schaar, ‘‘A machine learning approach for tracking and predictingstudent performance in degree programs,’’ IEEE J. Sel. Topics Signal Process.

[8] S. Lee and J. Y. Chung, ‘‘The machine learning-based dropout early warning system for improving the performance of dropout prediction,’’ Appl. Sci., vol. 9, no. 15, p. 3093, Jul. 2019.