STUDENT PERFORMANCE ANALYSIS AND PREDICTION

A PROJECT REPORT

Submitted by

AADITYA PRABU K

MOHAMMED THABREZ G

submitted to the faculty of

INFORMATION AND COMMUNICATION ENGINEERING

In partial fulfilment for the award of

the degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND TECHNsOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY

CHENNAI 600 025

AUGUST 2022

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report **"STUDENT PERFORMANCE ANALYSIS AND PREDICTION"** is the bonafide work of "**AADITYA PRABU K (2020115001) and MOHAMMED THABREZ G (2020115051)"** who carried out the project work under my supervision

SIGNATURE

DR.T.MALA

SUPERVISOR

ASSOCIATE PROFESSOR

Department of Information Science and Technology,

College of Engineering,

Chennai - 600 025

ABSTRACT

Learning Analytics (LA) focuses on the collection and analysis of learners' data to improve their learning experience by providing informed guidance and to optimize learning materials. Online learning has attracted a large number of learners and is increasingly becoming very popular. Categorizing these learners based on their interaction with the course can help address this need and suggest possible improvements in course design and delivery. The main objective is to find meaningful indicators or metrics in a learning context and to study the inter-relationships between these metrics using the concepts of Learning Analytics and Educational Data Mining (EDM) thereby, analyzing the factors affecting student performance and student dropout. In this project, K-means, a clustering data mining technique, using Davies' bouldin method is used to obtain clusters which are further mapped to find the important features affecting students' performance. Another main objective is to predict a student's final result based on the interactional, course, and grade factors using various prediction models and choosing the best model for prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

S. NO	ABBREVIATION	EXPANSION
1	VLE	Virtual Learning Environment
2	OULAD	Open University Learning Analytics Dataset
3	LA	Learning Analytics
4	LR	Linear Regression
5	LOR	LOgistic Regression
6	RFC	Random Forest Classifier
7	ANN	Artificial Neural Network
8	DBI	Davies Bouldin Index
9	EDM	Educational Data Mining

CHAPTER 1
INTRODUCTION

OBJECTIVES OF THE PROJECT

To understand the factors affecting student performance and dropout using data mining techniques:

- To understand student performance
- To understand student dropout
- To predict student performance

DATA COLLECTION

The data set used for this study is the Open University Learning Analytics Dataset (OULAD). It contains data about courses, students, and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). The dataset consists of tables connected using unique identifiers. All tables are stored in CSV format.

Website: https://analyse.kmi.open.ac.uk/open_dataset#description

The data set has 7 CSV files: ‘assessments.csv’, ‘ courses.csv’, ‘studentAssessment.csv’, ‘ studentInfo.csv‘, ‘ studentRegistration.csv’, ‘studentVle.csv’, ‘ vle.csv’.

CHAPTER 2

STUDENT PERFORMANCE ANALYSIS

DATA PREPARATION

Table Selection

Our object is to analyze student performance. The main attributes that are required to understand the objective are:

- Physical factors
- Demographic factors
- Educational factors
- Interactional factors

These factors in detail are only present in the following tables:

1. ***studentInfo*** : code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result.
2. ***studentVle***:code_module,code_presentation,id_student,id_site, student_interaction_date,sum_click.

Data exploration/relevant attributes selection

The above two tables have fifteen fields, which can be divided into three categories:

1) ***Qualitative*** : code_module, code_presentation, id_student, gender, region, highest_education, disability, id_site, final_result.

2) ***Quantitative*** : imd_band, age_band, num_of_prev_attempts, studied_credits, sum_click.
3) ***Date*** : student_interaction_date.

The date variables were not considered for the purpose of clustering. The studentVle table is grouped by code_module, code_presentation, id_student. As a result, the total sum clicks (summation of sum clicks of a student for different id_sites, visited on different dates) is derived.

Table Merging

It is difficult to analyze tables separately hence the studentInfo and studentVle tables are merged into one using inner join on 'code_module', 'code_presentation', 'id_student'.

Data cleaning

Data cleaning is the process of removing irrelevant items and missing values. The 'imd_band' attribute of this dataset had 55 missing values out of 4137 records, which were then converted to the mean of the relevant values.

Feature selection

Feature selection is an important step in the data preprocessing field. The objective of this process is to select a suitable subclass of features that can competently define the input data, decrease the dimensionality of feature space, and delete redundant and inappropriate data. This process can help in improving the data quality, therefore the performance of the learning algorithm. We have used two methods for feature selection: mutual information classifier and random forest classifier. The top seven features from the mutual information classifier are 'age_band', 'gender', 'highest_education', 'imd_band', 'num_of_prev_attempts', 'region', 'sum_click', and the top seven features from the random forest classifier are 'age_band', 'gender', 'highest_education', 'imd_band',

‘num_of_prev_attempts’, ‘region’, ‘sum_click’. From both of these methods, the attributes selected for clustering are shown in Table 1.

Table 1. Attributes considered for clustering

Attribute name	**Description**
code_module	Identification code of the module, to which the assessment belongs.
gender	The student’s gender.
region	Identifies the geographic region where the student lived while taking the module.
highest_education	Highest student education level on entry to the module presentation.
imd_band	Specifies the Index of Multiple Deprivation band of the place where the student lived during the module.
age_band	Band of the student’s age.
num_of_prev_attempts	The number of times the student has attempted this module.
disability	Indicates whether the student has declared a disability.
final_result	Student’s final result in module
sum_click	The number of times a student interacts with the material.

DATA PREPROCESSING

Categorical Data handling

Categorical data are usually represented as ‘strings’ or ‘categories’ and are finite in number and are of two categories.

1. ***Ordinal data*** - The categories have an inherent order.
2. ***Nominal data*** - The categories do not have an inherent order.

Ordinal data handling

In ordinal data, while encoding, the information regarding the order in which the category is provided, is retained. From the table, ‘imd_band’, ‘age_band’, and ‘highest_education’ have an ordinal relationship within themselves. Hence it is mapped with their ordinal equivalent numbers.

Nominal data handling

While encoding nominal data, the presence or absence of a feature is considered. In such a case, no notion of order is present and hence One Hot encoding method is used to encode nominal data. In this method for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category. From the table, ‘region’, ‘code_module’, ‘gender’, ‘disability’, and ‘final_result’ have a nominal relationship within themselves. Hence it is mapped with their One Hot encoded values.

Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data. Since the range of values of raw data varies widely,

in some machine learning algorithms, objective functions will not work properly without normalization. The most common techniques of feature scaling are Normalization and Standardization.

Normalization (Min-Max)

Min-max normalization performs a linear transformation on the original data. This technique gets all the scaled data in the range (0, 1). Formula shown in Fig 2.1

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Fig 2.1 Min-Max Normalization formula

CLUSTERIZATION

Clustering is an unsupervised learning technique to identify hidden patterns or structures in the data. Clustering helps to partition the data into homogeneous groups such that the observations in one group are more similar to each other than to the observations in other groups. The various partition-based clustering techniques used are K-means, K-medoids, CLARA, and so on.

K-Means Clustering

The K-means clustering algorithm is the most common clustering algorithm used to generate insights in terms of a grouping of data. In a data set having “N”

observations, all in real “d”-dimensional space, the problem is to find a set of “K” centers. The main objective of this clustering technique is to minimize the mean squared distance from each element to its nearest center. Each group or cluster is represented by its mean. This algorithm has an issue in that the number of clusters must be determined before the iterative procedure.

Selection of K in K-Means clustering

Elbow Method

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula for this method is given in Fig 2.2

$$WCSS = \sum P_{\text{i in Cluster 1}}\, distance(P_i C_1) + \sum P_{\text{i in Cluster 2}}\, distance(P_i C_2) + \ldots\ldots + \sum P_{\text{i in Cluster k}}\, distance(P_i C_k)$$

Fig 2.2 Elbow method formula

$\sum P_{\text{i in Cluster1}}$ distance$(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms. To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance. The sharp point of the bend of the plot looks between calculated WCSS values and the number of clusters K, like an arm, is considered the best value of K. The sharp bend is shown in Fig 2.3.

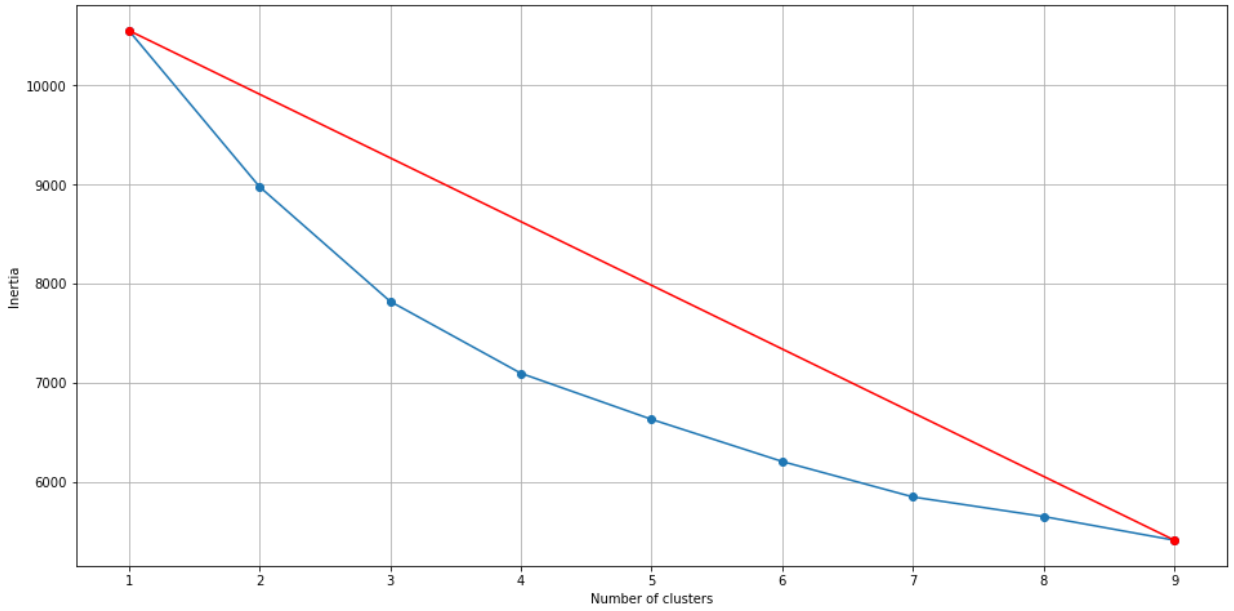

Fig 2.3 Elbow method graph

Davies Bouldin Method

The Davies-Bouldin index (DBI) is one of the clustering algorithms evaluation measures. It is most commonly used to evaluate the goodness of split by a K-Means clustering algorithm for a given number of clusters. In a few words, the score (DBI) is calculated as the average similarity of each cluster with a cluster most similar to it. The lower the average similarity is, the better the clusters are separated and the better the result of the clustering performed. The formula for this method is given in Fig 2.4. The graph for Davies Bouldin implementation is shown in Fig 2.5.

$$\bar{R} = \frac{1}{N} \sum_{i=1}^{N} R_i$$

A smaller $\bar{R}$ represents better defined clusters.

Fig 2.4 Davies Bouldin formula

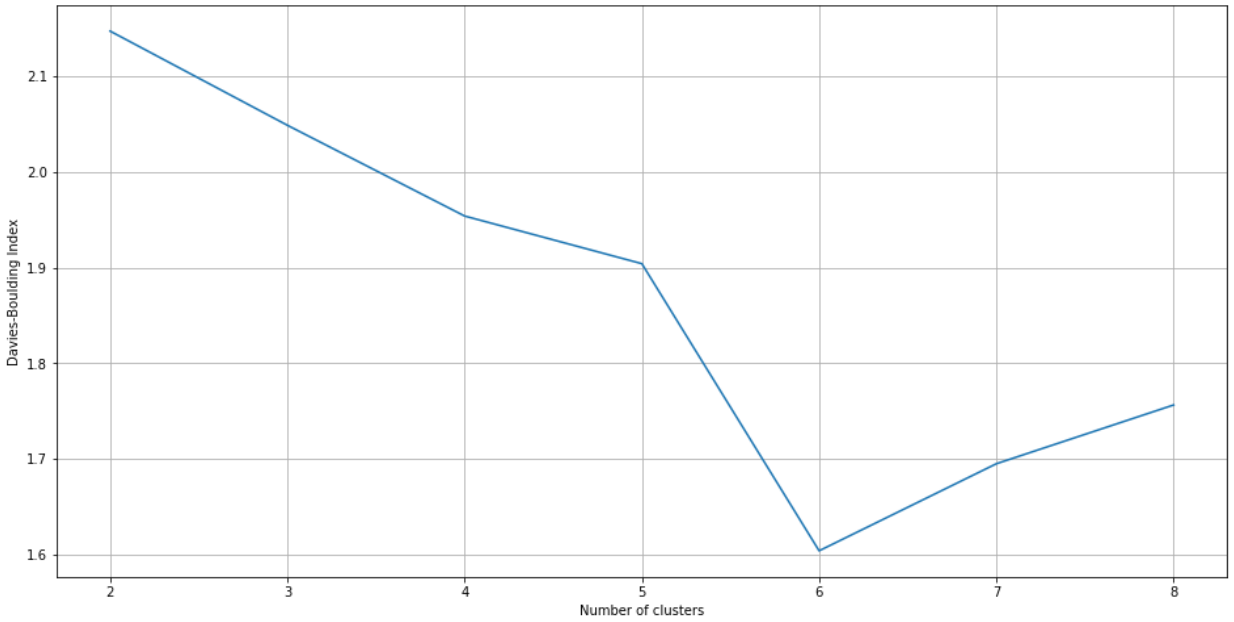

Fig 2.5 Davies Bouldin graph

The obtained k value for the Elbow method is four and for Davies Bouldin method is six. When the number of clusters increases the splitting of data also increases, hence it is easier to identify trends when k is six than when k is four. Hence six is chosen as the K value for K means Clusterization.

RESULT AND ANALYSIS

Table 2 Final result

Feature Final Result	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
Pass	0%	0%	77.34	100%	0%	41.44%
Fail	0%	100%	8.54	0%	0%	27.54%
Distinction	100%	0%	5.09	0%	0%	5.35%
Withdrawn	0%	0%	9.03	0%	100%	25.67%

Based on final results we can classify these clusters as Achievers, Failed, Performer, Dropout and mixture of these.

Table 3 Region

Feature Region	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
East Anglian Region	9.88%	10.01%	13.14%	10.33%	11.86%	15.24%
East Midlands Region	7.10%	6.84%	6.90%	5.34%	7.74%	9.36%
Ireland	3.40%	4.52%	2.96%	6.55%	4.78%	1.34%
London Region	7.41%	12.45%	8.87%	6.91%	10.38%	7.75%
North Region	3.40%	3.54%	3.94%	3.85%	3.95%	1.07%
North Western Region	4.94%	10.13%	6.90%	4.91%	9.06%	7.49%
Scotland	9.26%	8.79%	9.03%	10.04%	7.58%	10.43%
South East Region	11.73%	3.66%	7.72%	7.55%	5.60%	9.36%
South Region	11.73%	5.74%	11.17%	8.90%	8.40%	11.50%
South West Region	10.19%	6.23%	10.34%	7.98%	6.59%	6.68%
Wales	8.95%	12.70%	3.45%	12.18%	8.73%	8.02%
West Midlands Region	5.86%	7.69%	8.21%	8.69%	8.90%	8.29%
Yorkshire Region	6.17%	7.69%	7.39%	6.77%	6.43%	3.48%

Majority of the achievers are coming from South East Region, South Region and South West Region (i.e) mostly from southern regions. Failed students are mostly found in the East Anglian Region, London Region, North Western Region and Wales. A similarity can be found between Failed students and Dropouts as they are coming from the same region i.e East Anglian Region and London Region. Performers are from the East Anglian Region, Scotland, Wales. It is noticed that the performers and failed students share a common region which is Wales and East Anglian Region. The connection between the regions of Performers and Failed students can be understood by analyzing Mixed cluster C2. In Mixed cluster C2 majority of the people have passed. Hence these two regions comprise both performers and failed students. Mixed Cluster C5 is a combination of the remaining clusters.

Table 4 Disability

Feature Disability	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
Yes	0%	0%	2.46%	0%	0%	0%
No	100%	100%	97.54%	100%	100%	100%

All the clusters are not affected by disability, hence disability is not considered as a factor in this scenario.

Table 5 Code Module

Feature Code Module	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
AAA	4.01%	4.03%	100%	0%	8.90%	6.95%
BBB	95.99%	95.97%	0%	100%	91.10%	93.05

In Mixed Cluster C2, the major people who have passed found Course 'AAA' to be less difficult to understand, but in Achievers, Failed students, Performers, Dropout, Mixed Cluster C5, the majority of courses found to be 'BBB'. Hence course BBB can be inferred to be popular and difficult among the students.

Table 6 Gender

Feature Gender	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
Male	10.80%	12.09%	70%	10.80%	12.09%	8.57%
Female	89.20%	87.91%	29.39%	89.20%	87.91%	91.43%

In Mixed Cluster C2, the majority of the males have passed but when other clusters are taken into account it is inferred that females are actively participating in online education when compared to male.

Table 7 Age Band

Feature Age Band	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
0-35	52.47%	72.04%	46.96%	64.39%	65.73%	67.38%
35-55	47.53%	27.84%	46.47%	35.54%	33.44%	32.62%
55<=	0%	0.12%	6.57%	0.07%	0.82%	0%

Achievers comprising fifty percent of the 35-55 age category, mostly females (inference from the previous table), are interested to learn skills by taking up online education. Major students who fail come under the age band of 0-35, and also dropouts. The minimum age to study a degree programme at a UK university is 17 years old. Hence 0 - 16 age groups are neglected. The average age for men marrying is 36.7 years, while for women it is 34.3 years in UK. Now we take in the age range from 17 to 35 who are assumed to be unmarried. Hence we can conclude that the dropouts/ failed students are mainly not interested in the course. And those above the 35 - 55 age group are mainly affected by less time spent for education.

Table 8 Imd Band

Feature Imd Band	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
0-10%	6.48%	16.48%	4.11%	9.54%	12.19%	16.31%
10-20%	7.41%	13.31%	6.08%	10.97%	12.19%	13.10%
20-30%	7.41%	12.94%	8.05%	10.61%	15.32%	13.90%
30-40%	12.04%	11.60%	9.52%	11.32%	10.71%	13.64%
40-50%	12.65%	11.48%	10.34%	10.11%	12.19%	10.70%
50-60%	12.96%	9.65%	12.48%	12.46%	8.90%	6.95%
60-70%	10.19%	8.42%	7.06%	10.04%	8.24%	8.02%
70-80%	8.02%	7.08%	14.29%	10.40%	7.08%	8.29%
80-90%	12.04%	6.35%	11.82%	8.05%	7.58%	4.01%
90-100%	12.04%	2.69%	16.26%	6.48%	5.60%	5.08%

Dropouts and failed students are mostly found in the imd_band of 0-50 %. It is inferred that they mostly have the facilities or money to acquire online education but are mainly not interested in the course. Even despite a high imd_band, achievers tend to score distinction shows their interest to learn.

Table 9 Highest Education

Feature Highest Education	**C0** (Achiever)	**C1** (Failed)	**C2** (Mixed)	**C3** (Performer)	**C4** (Dropout)	**C5** (Mixed)
No Formal quals	0.62%	1.71%	0%	0.64%	2.31%	0.27%
Lower Than A Level	19.75%	51.28%	20.36%	40.31%	51.07%	54.01%
A Level or Equivalent	57.72%	37.36%	52.22%	45.94%	37.07%	41.18%
HE Qualification	20.68%	9.65%	26.11%	12.89%	9.56%	4.55%
Post Graduate Qualification	1.23%	0%	1.31%	0.32%	0%	0%

Achievers and Performers mostly have their highest education as “A Level or Equivalent”. Slightly more than half of the dropouts, have their highest education as “Lower than A level”. This infers that dropouts from this educational level find it difficult to understand their courses and hence dropped the course. About thirty-seven percent of the dropouts, whose highest education is “A Level or Equivalent”, find the course to be less difficult than that of “Lower than A level” dropouts. It is inferred that as the educational level increases people’s understanding of the course also increases, and the dropout rate decreases. This is the same for failed students.

CONCLUSION AND FUTURE SCOPE

The above study shows that the Majority of the students are from the East Anglian region and there is a larger number of females who are actively participating even in the age group of 35-55. The student's performance mainly depends on their interest in learning despite a poor Imd Band index. The level of education is directly dependent on the student’s understanding of the course. This cluster-based approach for Learning analytics can be used in practice and is applicable at classroom-level as well as distance learning levels. In the future, further analysis can be carried on to use mixed-method evaluation approaches to study the inter-relationships between the different features. Also, various other clustering techniques like DB-SCAN, Agglomerative, etc. can be further used to improve the clusters and help achieve different outcomes for numerous other Learning management systems. Further analysis can also be performed to study the trends in various educational systems which in turn can help in improving the learning systems and quality of education.

CHAPTER 3

STUDENT PERFORMANCE PREDICTION

FEATURE ENGINEERING

Assessments

The performance in each assessment is a good indicator of the student's knowledge of the course and, as it has the grade for the final evaluation, it's interesting to make it a feature in the final model. But, as there are many different courses, each with a different structure, it's unfeasible to create a feature for each assessment. In order to include the assessments, a new feature is derived, called ‘Weighted Grade’ which is the sum of all weighted assessment scores of a student in a module-presentation. Final exams from the other assessments are split, given that their status and participation in the final evaluation is different from the other assessments.Along with ‘weighted grade’ another two attributes ‘code_presentation’ and ‘code_module’ are considered for predicting, because it has the important factor of course difficulty.

Virtual learning environment

The datasets referring to the VLE (Virtual Learning Environment) contain the interaction feed of the students with the content available for reference throughout the duration of the period. From this data, it can be inferred how in touch a student was with their subjects, whether they studied it on a solid basis and how they used the content**.** Two attributes are derived related to vle which are ‘site_visited’ and ‘tot_click’. ‘site_visited’ is the total number of vle material visited by a student in the module-presentation and ‘tot_click’ is the total number of clicks on vle

material by a student in the module-presentation. In turn we derive another attribute called ‘avg_click_per_site’ as ‘tot_click’ by ‘site_visited’ which gives the average clicks per site.

Student Information

The ‘studentInfo’ table contains various info about the students, but the relevant one for this analysis is the students' final results which is our interest variable as we build our prediction model.

The final attributes selected for student performance prediction are shown in Table 10

Table 10 Attributes considered for modeling

Attributes	**Description**
code_module	Code name of the module, which serves as the identifier.
code_presentation	Code name of the presentation.
avg_click_per_site	It is Total number of clicks done by the student in a module-presentation by Total number of vle material visited by the student in the same module presentation
weighted_grade	Which is the sum of all weighted assessment scores of a student in a module-presentation
final_result	Student’s final result in the module-presentation.

MODELING

Linear Regression

Linear Regression (LR) is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Logistic Regression

Logistic regression (LOR) is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Random Forest Classifier

Random forest is a supervised (RFC) learning algorithm that is used for both classifications as well as regression. But it is mainly used for classification problems. As we know that a forest is made up of trees and more trees mean a more robust forest. Similarly, the random forest algorithm creates decision trees on data samples and then gets the prediction from each of them, and finally selects

the best solution by means of voting. It is an ensemble method that is better than a single decision tree because it reduces the over-fitting by averaging the result.

Artificial Neural Network

Artificial neural networks (ANNs), usually simply called neural networks (NNs) or neural nets, are computing systems inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives signals then processes them and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold.

RESULT

The accuracy scores obtained from the different methods are given in Table 11 and Table 12

Table 11 Binary classification results

Modelling	**Accuracy**
Linear Regression	0.53
Logistic Regression	0.90
Random Forest Classifier	0.89
Artificial Neural Network	0.91

Table 12 Multi classification results

Modelling	**Accuracy**
Linear Regression	0.027
Logistic Regression	0.79
Random Forest Classifier	0.79
Artificial Neural Network	0.80

CONCLUSION

From Table 11 and 12 we can conclude that linear regression is not suitable for student performance as their data is not linear. Out of all the prediction models, Artificial Neural Networks gives the best accuracy of 80 % for Pass/Fail/Distinction Prediction (Multi-Class Classification) and 91% for Pass/Fail Prediction (Binary Classification).

REFERENCES

1) Francis, B. K., & Babu, S. S. (2019) Predicting academic performance of students using a hybrid data mining approach. Journal of medical systems.

2) Karimi, H., Derr, T., Huang, J., & Tang, J.(2020) Online Academic Course Performance Prediction Using Relational Graph Convolutional Neural Network.

3) SANDRA PRUSAKA (2019) Student Failure | Modelling with a messy dataset.

4) Sanyam Bharara & Sai Sabitha & Abhay Bansal (2017) Application of learning analytics using clustering data Mining for Students' disposition analysis.

5) Skand Arora , Manav Goel , A. Sai Sabitha & Deepti Mehrotra (2017) Learner Groups in Massive Open Online Courses, American Journal of Distance Education.

6) VICTOR RÉGIS (2020) Student Performance Prediction: Complete analysis.