iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: XI Month of publication: November 2021

DOI: https://doi.org/10.22214/ijraset.2021.38786

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

Student Academic Performance Prediction under Various Machine Learning Classification Algorithms

M. Nirmala[1], T. Seeni Selvi[2], V. Saravanan[3]
[1]Department of Computer Applications,
[2]Department of Computer Science,
[3]Department of Information Technology
[1]Hindusthan College of Engineering and Technology,
[2, 3] Hindusthan College of Arts and Science

Abstract: Data Mining in Educational System has increased tremendously in the past and still increasing in present era. This study focusses on the academic stand point and the performance of the student is evaluated by various parameters such as Scholastic Features, Demographic Features and Emotional Features are carried out. Various Machine learning methodologies are adopted to extract the masked knowledge from the educational data set provided, which helps in identifying the features giving more impact to the student academic performance and there by knowing the impacting features, helps us to predict deeper insights about student performance in academics. Various Machine learning workflow starting from problem definition to Model Prediction has been carried out in this study. The supervised learning methodology has been adopted and various Feature engineering methods has been adopted to make the ML model appropriate for training and evaluation. It is a prediction problem and various Classification algorithms such as Logistic Regression, Random Forest, SVM, KNN, XGBOOST, Decision Tree modelling has been done to fit the student data appropriately.

Keywords: Scholastic, Demographic, Emotional, Logistic Regression, Random Forest, SVM, KNN, XGBOOST, Decision Tree.

I. INTRODUCTION

Machine Learning [1] commonly deals with big data where the size of the data is massive and the data can be both in structured and unstructured format. It endows the computers with the ability to learn from 'DATA' and make sensible decisions. The main focus of this research it to perform a step by step process of the Machine Learning approach from Problem definition to Prediction. Educational sector is a domain where outsized amount of data is being bred every day. The generated existing data and the about to receive data if analysed in the right format can bring tremendous changes in the Scholastic field. The Machine Learning technique is able to perfectly analyze the data and can bring lot of changes in improving the scholastic performance of the students. The other features which included demographic, behavioural can also create an impact in the academic performance of the students.

II. LITERATURE SURVEY / RELATED WORK

Numerous data mining tasks [2] were used to create qualitative predictive models to predict the students' grades from a collected training dataset. During the survey, university students were aimed and collected multiple personal, social, and academic data of them. Pre-processing of the collected were done to make it suitable for data mining tasks. Third, the classification models were tested on the pre-processed data. On the whole this study motivated the universities to do data mining tasks on their students' data regularly to get interesting results and patterns which in turn can be more effective and helpful for university as well as the students in many ways. A similar research on Educational Data Mining; Student's performance was predicted based on academic records and their forum participation in [3] . Two undergraduate course data were collected. To predict student's performance three classification models like Naive Bayes, Neural Networks and Decision Trees were used. The results show that Naive Bayes model gave better result comparing to other two models.

Another comparative study was done by [4]. They compared six algorithms like J48, Random Forest, Naive Bayes, Naive Bayes Multinomial, K-Star and IBK. The data set contains 480 records and Weka Tool were used for implementation. The Survey conducted based on seven attributes and found Random Forest algorithm provides more accuracy compared to other algorithms.

A survey was conducted over 200 college students. In this research [5] classification algorithms were adopted on student dataset to foretell the learning behavior of student's. Slow learners were identified, and actions were taken to reduce the failure count and correct actions could be adopted to make the weaker students suitable for learning. In this study the J48, Naive Bayes and Random forest algorithms were compared. Finally the researcher got accuracy using Random forest algorithm when the data set is in massive size.

The study about students' educational behavior done by [6] proposed framework having a category of a feature called "Behavioral feature" is introduced where they focus on student's behavioral features and their relationship with student's academic success. They used the same framework to examine student's progress by using ensemble techniques which enhance the overall accuracy of results. Classification task on student database to predict the academic performance of student was carried by [7]. Bayesian Network Classifiers is used in this study. Information like Previous semester marks, Internal Assessment Marks, Performance during Seminars, Assignment, Attendance, Co-Curricular Activities were collected to predict the performance of the end semester marks. This study will help the students improve their performance. The students who require special responsiveness will be effectively identified and the failure rate of students would be decreased considerably.

A Student performance through a study was done by [8]. The sample contains 300 students out of which 225 are males and 75 are females. The performance of the students in the class are affected by various parameters such as student attendance, hours spent in class, family income, students mother's age and her education.

Educational Data Mining to be a upcoming research area which deals with computational methods to explore educational data was explained by [9]. It also explains the types of Educational Environments, Educational data and different group of people in education field. It helps us to explore educational phenomenon better and to get enhanced insights into it. This also says about the current affairs in the EDM field.

III. RESEARCH METHODOLOGY

The various methods adopted during the research process have been portrayed. This is a Descriptive Research problem where the study of student data set is explored. It performs the prediction of Academic performance of students of an educational body by applying various methodologies with respect to Machine Learning.

A. Research Data

The data collected from secondary data sources are tabulated in the **Table 1**.

Table 1 : Data Source Details

Data sources	xAPI-Edu-Data.csv
Dataset characteristics	Multivariate
Number of Instances	480
Number of Attributes	17
Attribute Type	Categorical and Numerical
Dataset Owner	Ibrahim Alijarah Professor (Assistant) at The University of Jordan Fargo, North Dakota, United States
Link	https://www.kaggle.com/aljarah/xAPI-Edu-Data/metadata

B. Proposed System Method Of Analysis

The proposed system states the prediction of the Academic performance of the student using various Features depicted in **Table 2** are classified as Demographic, Scholastic and Emotional.

Table 2 : Students Features

Demographic Features (Related to Population)	Scholastic Features	Emotional Features
gender	Educational Stages	Raised Hands
Nationality	Grade Levels	Visited Resources
Place of Birth	Section ID	Viewing Announcements
Parents responsible for student	Semester	Discussion Groups
	Topic	Parents Answering Survey
	Student Absence Days	Parents School Satisfaction
	Class (L,M,H) based on the total grade marks classified into 3 classes	

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

Machine Learning workflow has various steps to be followed starting from Problem definition to Model Prediction. Various steps required to be followed before fitting the model are shown in the Figure 1.

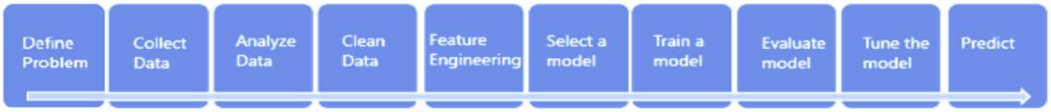


Figure 1 : Machine Learning Process Pipeline

C. Machine Learning Pipeline

Machine learning methodology is adopted for problems when traditional programming cannot be done, and when the system itself needs to solve the problem rather than a programmer, and if the size of the data is very large.

Steps to be followed for Machine Learning Process

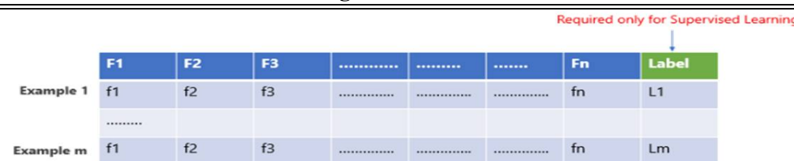
Define Problem	Be clear with what the model is expected to do. Ensure that all the inputs are available during prediction. In this system the academic performance of students need to be predicted based upon various features.
Collect Data	The data is collected from xAPI-Edu-Data.csv data repository. It contains 480 rows and 17 Columns. It contains both categorical and Numerical data. The data collected is in the format shown in **Figure 2**.
	 Figure 2 : Data Format for Supervised Learning

Table 3 : Students Features and its Descriptions

Feature	Datatype	Description
gender	Categorical	Male or Female
NationalITy	Categorical	Student Nationality
PlaceofBirth	Categorical	Place of Birth of the Student
StageID	Categorical	Stage refers to Primary, Middle or High School
GradeID	Categorical	Grade Category varies from G-01 to G-12
SectionID	Categorical	Classroom Section, either A or B or C
Topic	Categorical	Refers to Course Topic such as Math, Quran etc.
Semester	Categorical	Either First semester or Second Semester
Relation	Categorical	Either Father or Mum, who is responsible for Student
raisedhands	Numerical	Count of students Interacted during the class room by raising hands.
VisiTedResources	Numerical	Count of the students who visited the course content.
AnnouncementsView	Numerical	Count of the students who checks the new Announcements
Discussion	Numerical	Count of the students who participated on discussion groups.
ParentsAnsweringSurvey	Categorical	Whether Parent Answered Survey provided from school or not.
ParentsschoolSatisfaction	Categorical	Degree of Parent satisfaction from School
StudentAbsenceDays	Categorical	Either Nominal above 7 or under 7
Class	Categorical	Based on the total grade / marks it is classified as Low-level, Middle Level, High Level.

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (both graphical and quantitative) to better understand data. This system contains 4 Numerical Columns and 13 Categorical Columns and the description about each and every feature, its datatype, its category and its description are explained in the table 3.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

D. Exploratory Data Analysis

1) Univariate Analysis – Individual Features / Variables

Analyze Data	Identify the Null Values present in each column and after analysing it shows that the given data set contains No Null values. Data visualization is the graphical representation of data in the form of charts, diagrams etc. Visualization helps to understand the data much quicker than quantitative methods and as a part of visualization various methods are performed to Analyze the data in a better format. UNIVARIATE ANALYSIS – Individual Features / Variables BIVARIATE ANALYSIS – Relationship of a feature with Target Variable

The Univariate analysis does a single variable analysis. It does not infers its relationship with any other variables. In general count plot could be used for this analysis. It helps to portray the data and it's respective patterns for the user to get a better insight about the single variable and the graphical representation helps us to view maximum, minimum, mean values etc. The Univariate Analysis and its visualization inferences are described using below mentioned charts.

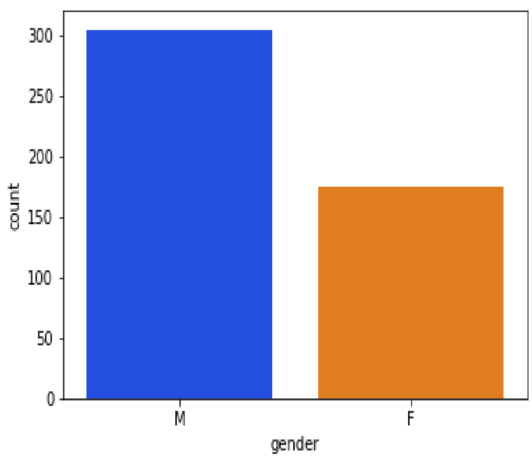


Figure 3 : Univariate Analysis - gender

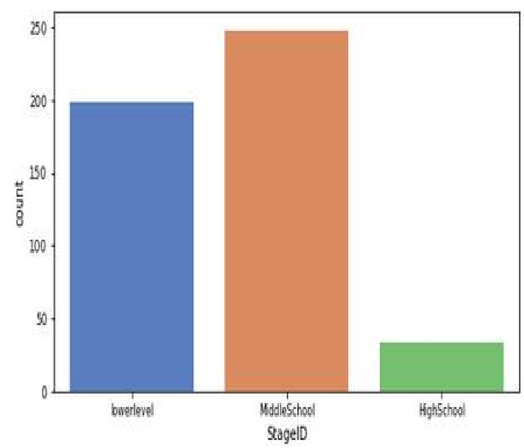


Figure 4 : Univariate Analysis – Stage ID

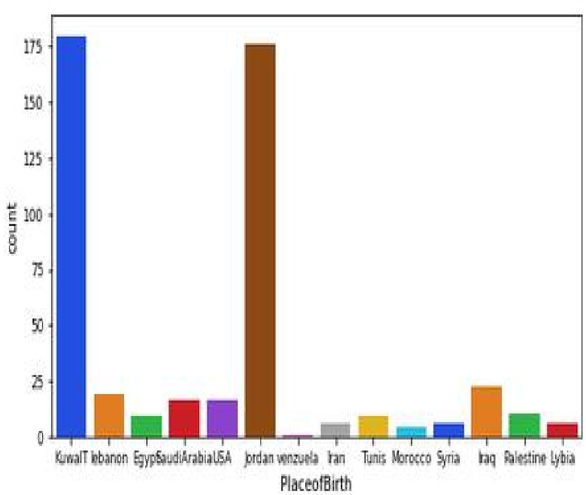


Figure 5 : Univariate Analysis – PlaceofBirth

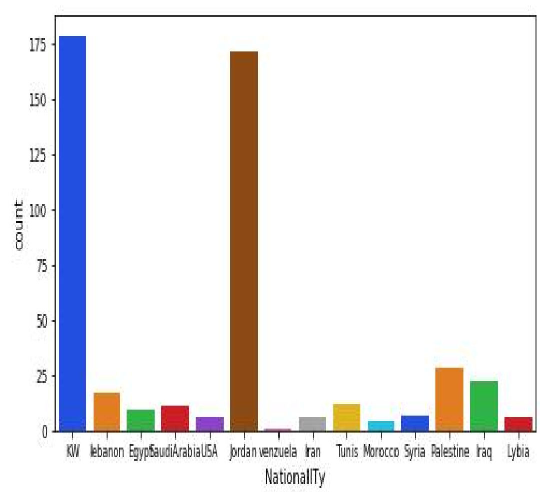


Figure 6 : Univariate Analysis – Nationality

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

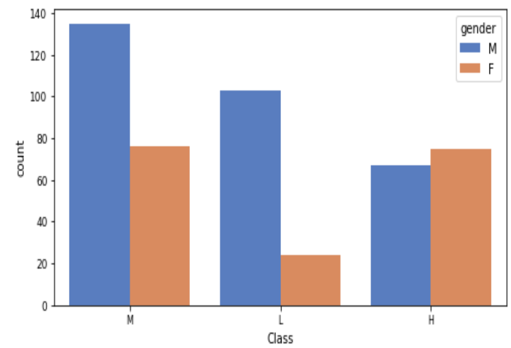


Figure 7 : Univariate Analysis – Class

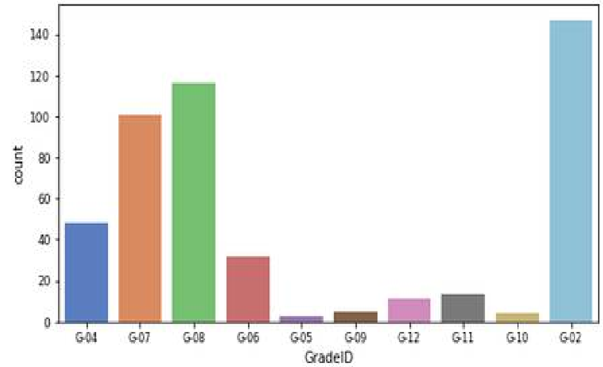


Figure 8 : Univariate Analysis – Grade ID

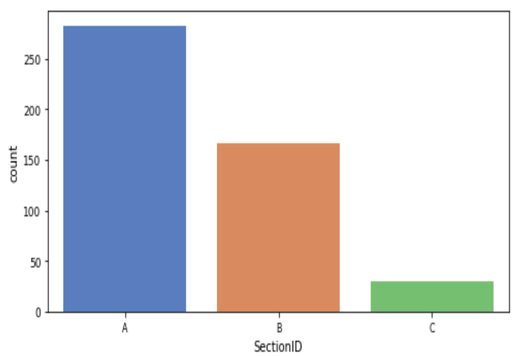


Figure 9 : Univariate Analysis – Section ID

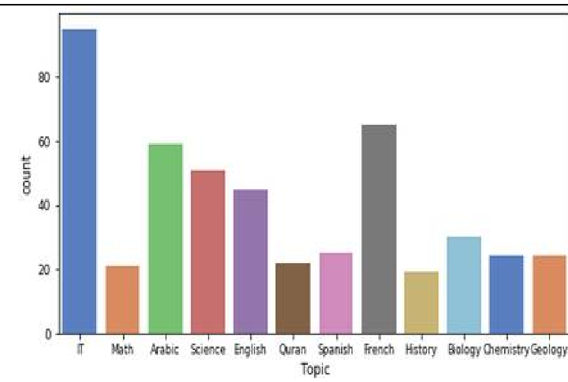


Figure 10 : Univariate Analysis – Topic

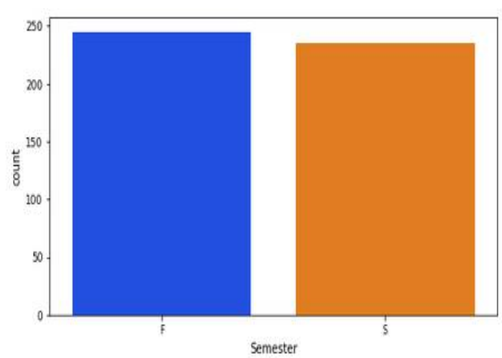


Figure 11 : Univariate Analysis – Semester

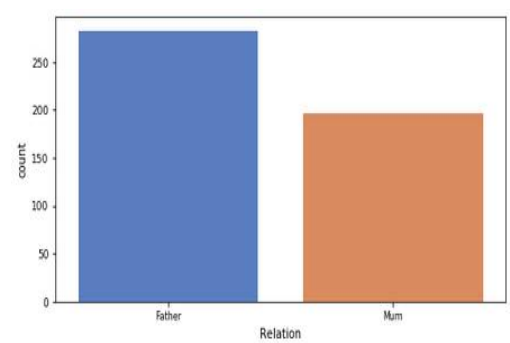


Figure 12 : Univariate Analysis – Relation

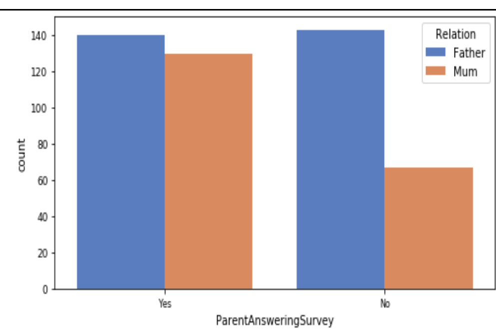


Figure 13 : Univariate Analysis – ParentAnsweringSurvey

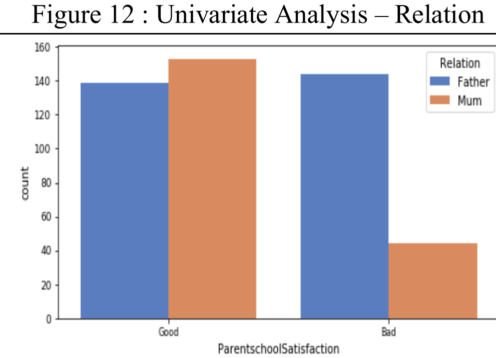


Figure 14 : Univariate Analysis – ParentschoolSatisfaction

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

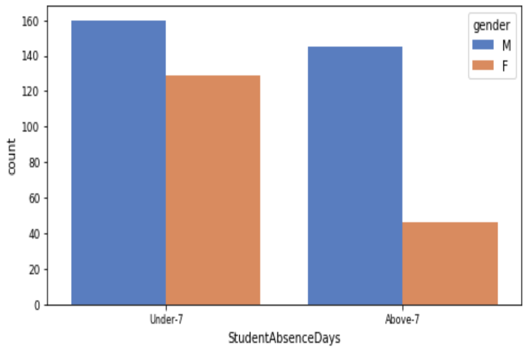


Figure 15 : Univariate Analysis – StudentAbsenceDays

2) Univariate Analysis –Report

<table>
<tr><td>Gender</td><td colspan="3">Male is 63.5% and Female is 36.4% . The gender feature infers that the maximum count of students from the data set is Male.</td></tr>
<tr><td>Nationality</td><td colspan="3">Under Nationality feature KW has 37.3% and Jordan has 35.8% and Venezuela has the least % of 0.2%</td></tr>
<tr><td>PlaceofBirth</td><td colspan="3">The % ratio of Nationality and Place of Birth is almost same and as per the analysis any one column could be dropped.</td></tr>
<tr><td>StageID</td><td colspan="3">Out of the total 51.7 % students are studying in MiddleSchool, 41.5% are in Lowerlevel and only 6.9% are in High School.</td></tr>
<tr><td>GradeID</td><td colspan="3">Out of the total G-02 is 30.6%,G-08 is 24.2% ,G-07 is 21%, G-04 is 10%, G-06 is 6.7%, G-11 is 2.7%, G-12 is 2.3%, G-09 is 1.04%, G-10 is 0.83% and G-05 is 0.63%.</td></tr>
<tr><td>SectionID</td><td colspan="3">Out of the total 59% are studying in A section. 34.8% are studying in B section and 6.25% are studying in C Section.</td></tr>
<tr><td>Topic</td><td colspan="3">Out of the total students 19.8% area of interest topic is IT, 13.5% is French, 12.3 % is Arabic, 10.6% is Science, 9.8% is English, 6.25% is Biology, 5.2% is Spanish, 5% for both Geology and Chemistry , 4.58% for Quran, 4.37% is Mathematics and 3.95% for History.</td></tr>
<tr><td>Semester</td><td colspan="3">51% of students are in First Semester and 48.95% are in Second Semester.</td></tr>
<tr><td>Relation</td><td colspan="3">Parent Responsible for student can be either Father or Mum. Out of the total % 58.9% is for Father and 41.04% is for Mother.</td></tr>
<tr><td>ParentAnsweringSurvey</td><td colspan="3">ParentAnsweringSurvey towards the school improvement is an important factor and 56.25% gave an Answer of ‘YES’ and 43.75% gave an answer of ‘NO’</td></tr>
<tr><td>ParentschoolSatisfaction</td><td colspan="3">ParentschoolSatisfaction is also an important factor and this helps to identify whether the student will continue in the same school or not. Out of the Total percentage 61% opinion towards the School was Good and remaining of 39% opinion towards school was Bad.</td></tr>
<tr><td>StudentAbsenceDays</td><td colspan="3">Out of the total 60% students are regular and 40% has taken more than 7 days leave. Female has more attendance than Male.</td></tr>
<tr><td rowspan="3">StudentAbsenceDays with respect to gender</td><td>StudentAbsenceDays/ Gender</td><td>Male</td><td>Female</td></tr>
<tr><td>Under 7</td><td>160</td><td>129</td></tr>
<tr><td>Above 7</td><td>145</td><td>46</td></tr>
<tr><td>Class</td><td colspan="3">Out of the Total Low Level score is acquired by 26.5%, Medium Level Score is acquired by 44% and High Level score is acquired by 30%of students.</td></tr>
</table>

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

3) Bivariate Analysis – Relationship Of a Feature With Target Variable

Bivariate Analysis is performed to find the associativity between every variable in the data set with the Target Variable (Class in this system). It also checks for association and the strength of this association or whether there are differences between two variables and the significance of these differences.

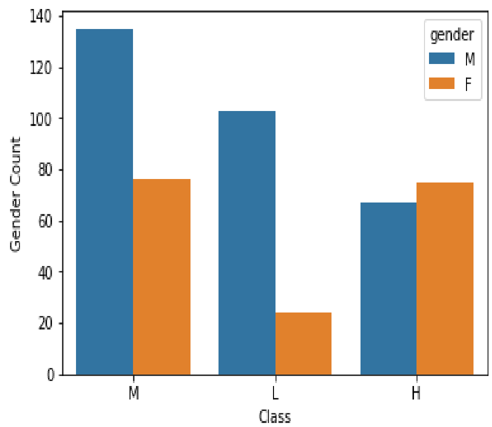


Figure 16 : Bivariate Analysis –Gender & Class

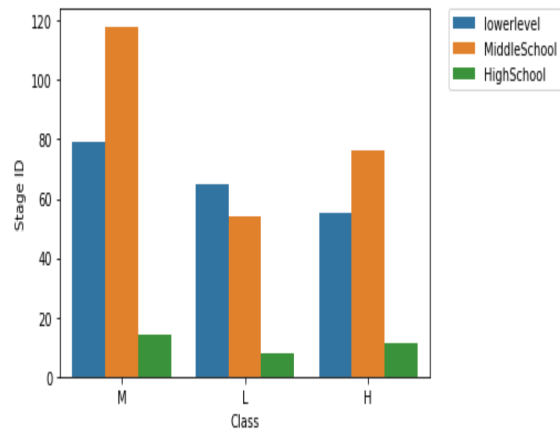


Figure 17 : Bivariate Analysis – Stage ID & Class

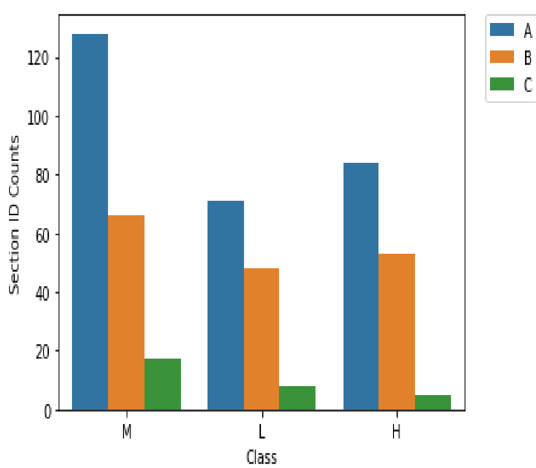


Figure 18 : Bivariate Analysis – Section ID & Class

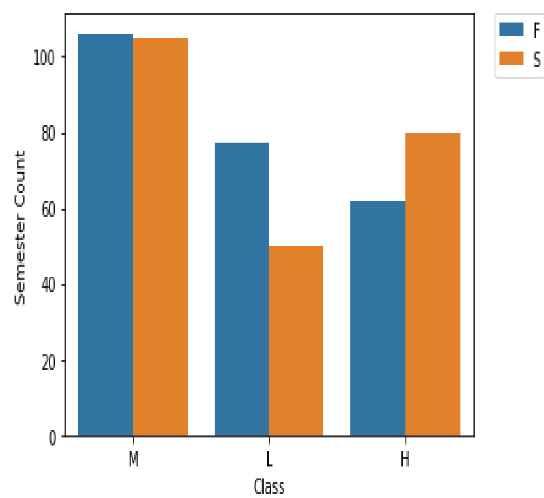


Figure 19 : Bivariate Analysis – Semester & Class

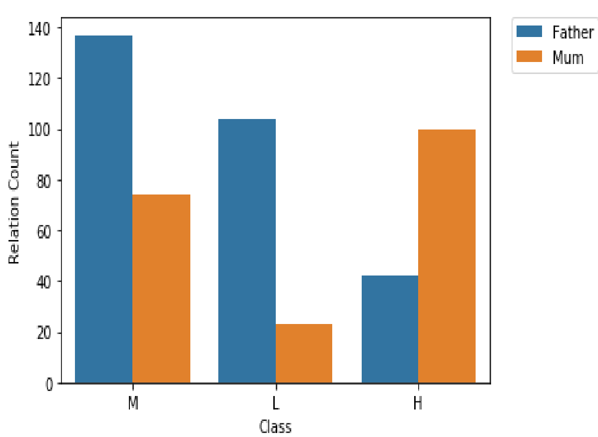


Figure 20 : Bivariate Analysis – Relation & Class

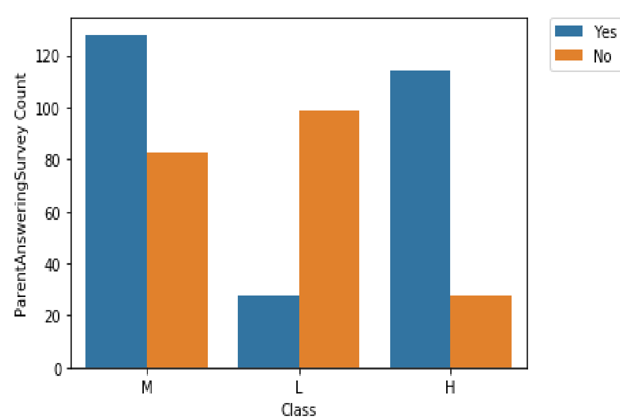


Figure 21 : Bivariate Analysis – ParentAnsweringSurvey & Class

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

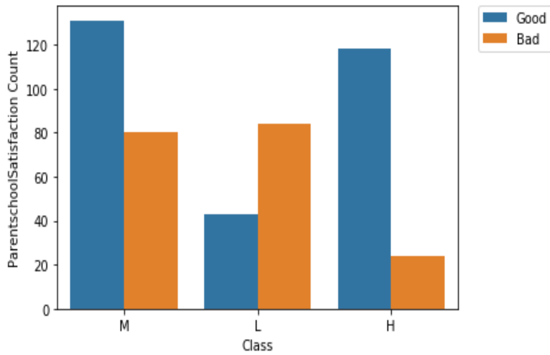


Figure 22 : Bivariate Analysis – ParentSchoolSatisfaction & Class

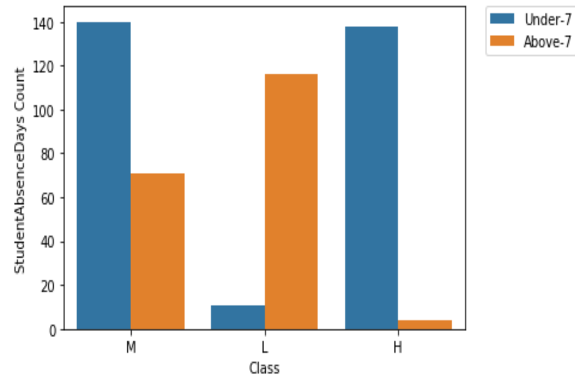


Figure 23 : Bivariate Analysis – StudentAbsenceDays & Class

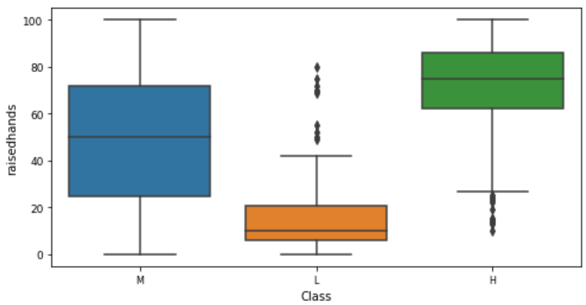


Figure 24 : Bivariate Analysis – raisedhands & Class

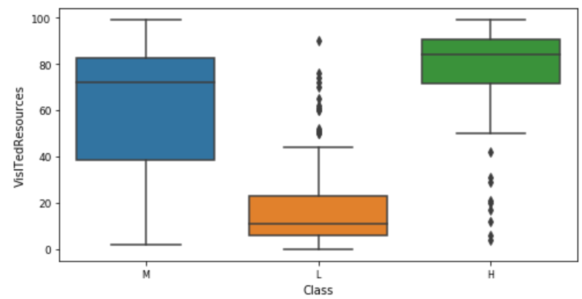


Figure 25 : Bivariate Analysis – Visited Resources & Class

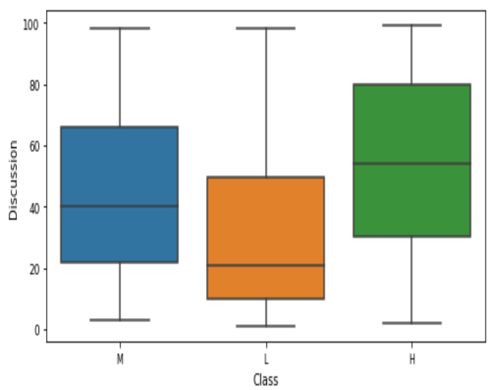


Figure 26 : Bivariate Analysis – Discussion & Class

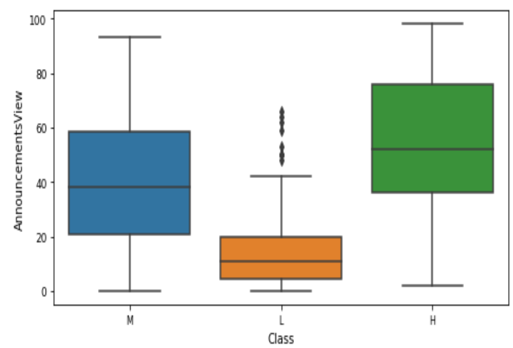


Figure 27 : Bivariate Analysis – Announcements View & Class

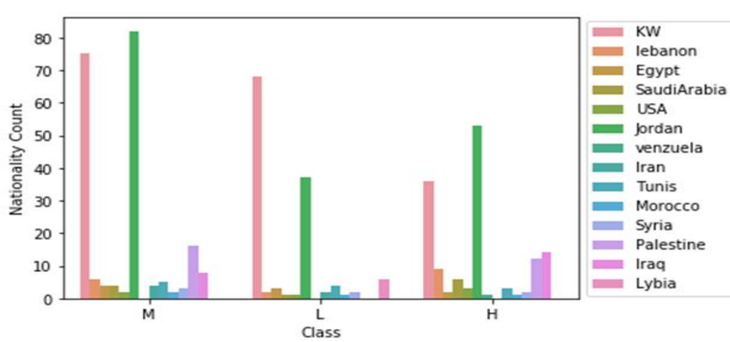


Figure 28 : Bivariate Analysis – Nationality & Class

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

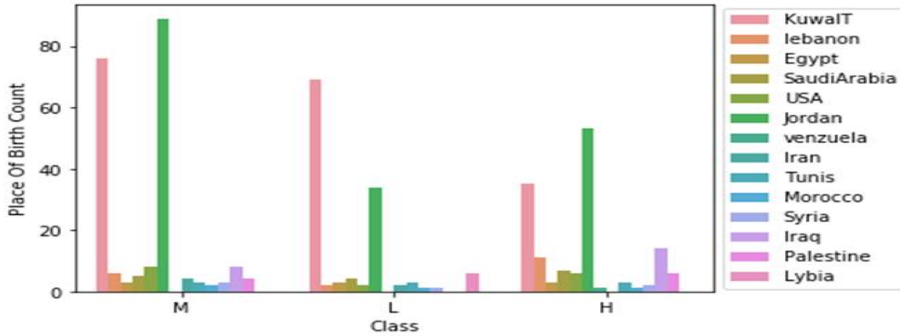


Figure 29 : Bivariate Analysis – Place of Birth & Class

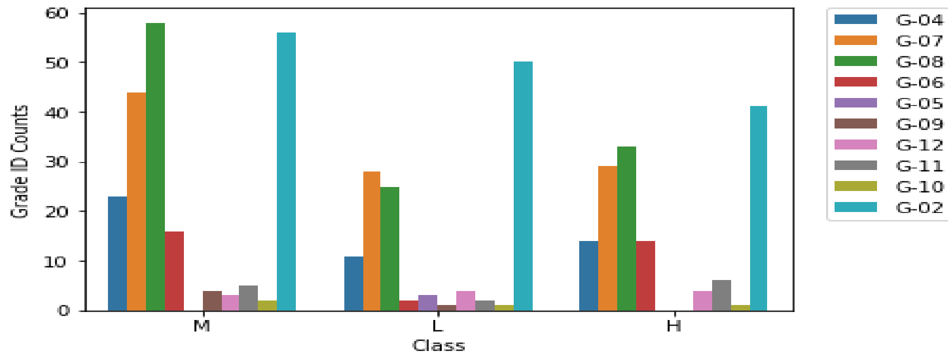


Figure 30 : Bivariate Analysis – Grade ID & Class

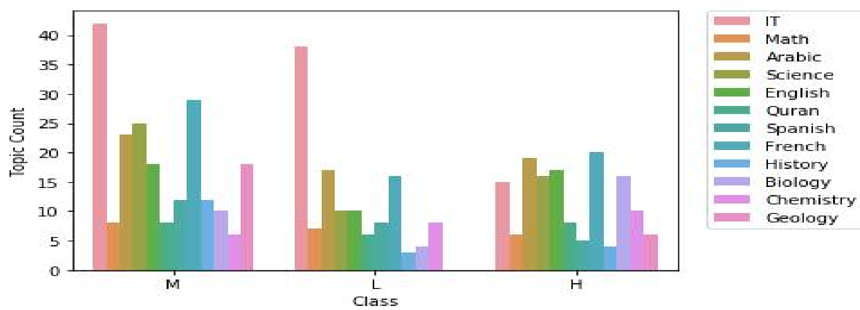


Figure 31 : Bivariate Analysis – Topic & Class

4) Bivariate Analysis –Report – Target Variable = Class

Gender

gender	F	M
Class		
H	52.82%	47.18%
L	18.90%	81.10%
M	36.02%	63.98%

Table 4 : Gender & Class Score

With respect to gender compared with class, female has the highest score with respect to High level and Male has Highest score with respect to Low Level. Female Academic performance is more compared to Male.

Nationality

NationalITy	Egypt	Iran	Iraq	Jordan	KW	Lybia	Morocco	Palestine	Saudi Arabia	Syria	Tunis	USA	lebanon	venzuela
Class	%	%	%	%	%	%	%	%	%	%	%	%	%	%
H	1	0	10	37	25	0	1	8	4	1	2	2	6	1
L	2	2	0	29	54	5	1	0	1	2	3	1	2	0
M	2	2	4	39	36	0	1	8	2	1	2	1	3	0

Table 5 : Nationality & Class Score

With respect to Nationality compared with class, Jordan and Egypt has got highest percentage compared to other countries

PlaceofBirth

Placeof Birth	Egypt	Iran	Iraq	Jordan	KuwaIT	Lybia	Morocco	Palestine	Saudi Arabia	Syria	Tunis	USA	lebanon	venzuela
Class	%	%	%	%	%	%	%	%	%	%	%	%	%	%
H	2	0	10	37	25	0	1	4	5	1	2	4	8	1
L	2	2	0	27	54	5	1	0	3	1	2	2	2	0
M	1	2	4	42	36	0	1	2	2	1	1	4	3	0

Table 6 : PlaceofBirth & Class Score

With respect to PlaceofBirth compared with class, Jordan and Egypt has got highest count value compared to other countries.

StageID

Stage ID	High School	Middle School	lower level
Class	%	%	%
H	8	54	39
L	6	43	51
M	7	56	37

Table 7 : Stage ID & Class Score

With respect to StageID Middle School and Lower Level has got high level of scores with respect to Class.

GradeID

Grade ID	G-02	G-04	G-05	G-06	G-07	G-08	G-09	G-10	G-11	G-12
Class	%	%	%	%	%	%	%	%	%	%
H	29	10	0	10	20	23	0	1	4	3
L	39	9	2	2	22	20	1	1	2	3
M	27	11	0	8	21	27	2	1	2	1

Table 8 : Grade ID & Class Score

G-02, G-08, G-09 has the highest scores compared to other grades

SectionID

Section ID	A	B	C
Class	%	%	%
H	59	37	4
L	56	38	6
M	61	31	8

Table 9 : Section ID & Class Score

With respect to SectionID compared with class, Section A is ranking high in all 3 class categories.

Topic

Topic	Arabic	Biology	Chemistry	English	French	Geology	History	IT	Math	Quran	Science	Spanish
Class	%	%	%	%	%	%	%	%	%	%	%	%
H	13	11	7	12	14	4	3	11	4	6	11	4
L	13	3	6	8	13	0	2	30	6	5	8	6
M	11	5	3	9	14	9	6	20	4	4	12	6

Table 10 : Topic & Class Score

Semester

Semester	F	S
Class	%	%
H	44	56
L	61	39
M	50	50

Table 11 : Semester & Class Score

In case of second semester, it is less in the Low Level and in other cases it is more.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

<table>
<tr><td>Relation</td><td><table><tr><th>Relation</th><th>Father</th><th>Mum</th></tr><tr><th>Class</th><th>%</th><th>%</th></tr><tr><td>H</td><td>30</td><td>70</td></tr><tr><td>L</td><td>82</td><td>18</td></tr><tr><td>M</td><td>65</td><td>35</td></tr></table>Table 12 : Relation & Class Score</td><td>With respect to Relation compared with class, the highlevel learning students are greatly supported and motivated by mothers.</td></tr>
<tr><td>ParentAnsweringSurvey</td><td><table><tr><th>Parent Answering Survey</th><th>No</th><th>Yes</th></tr><tr><th>Class</th><th>%</th><th>%</th></tr><tr><td>H</td><td>20</td><td>80</td></tr><tr><td>L</td><td>78</td><td>22</td></tr><tr><td>M</td><td>39</td><td>61</td></tr></table>Table 13 : ParentAnsweringSurvey & Class Score</td><td>With respect to ParentAnsweringSurvey compared with class, there was more yes for H and M and less for L.</td></tr>
<tr><td>ParentschoolSatisfaction</td><td><table><tr><th>Parent school Satisfaction</th><th>Bad</th><th>Good</th></tr><tr><th>Class</th><th>%</th><th>%</th></tr><tr><td>H</td><td>17</td><td>83</td></tr><tr><td>L</td><td>66</td><td>34</td></tr><tr><td>M</td><td>38</td><td>62</td></tr></table>Table 14 : ParentSchoolSatisfaction & Class Score</td><td>With respect to ParentSchoolsatisfaction compared with class, large majority of parents are satisfied with the education they received. In case of least satisfied parent the count is comparatively less.</td></tr>
<tr><td>StudentAbsenceDays</td><td><table><tr><th>Student Absence Days</th><th>Above-7</th><th>Under-7</th></tr><tr><th>Class</th><th>%</th><th>%</th></tr><tr><td>H</td><td>3</td><td>97</td></tr><tr><td>L</td><td>91</td><td>9</td></tr><tr><td>M</td><td>34</td><td>66</td></tr></table>Table 15 : StudentAbsenceDays & Class Score</td><td>The biggest visual trend can be seen is how frequently the student was absent. Over 90% of the students who did poorly were absent more than seven times, while almost none of the students who did well were absent more than seven times.</td></tr>
<tr><td>Raisedhands</td><td colspan="2">Total Students Raised Hands count : 22452
Total student count is : 480
Average Male Student Raised : 74.0
Average Female Student Raised : 128.0</td></tr>
<tr><td>AnnouncementsView</td><td>Total Students Viewed Announcements : 18201
Total student count is : 480
Average Male Student Viewed Announcement : 60.0
Average FeMale Student Viewed Announcement : 104.0</td><td>Female student have participated more in viewing announcements.</td></tr>
<tr><td>visitedResources</td><td>Total Students visited Resources : 26303
Total student count is : 480
Average Male Student visited Resources : 86.0
Average FeMale Student Visited Resources: 150.0</td><td>Female student have visited the resources more in number.</td></tr>
<tr><td>Discussion</td><td>Total Students Participated in Discussion : 20776
Total student count is : 480
Average Male Student Participated in Discussion : 68.0
Average FeMale Student participated in Discussion : 119.0</td><td>Female Students have more participated in Discussion.</td></tr>
</table>

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

5) *Correlation:* Coorelation [10] is a bivariate analysis that measures the strength of association between 2 variables and the direction of the relationship. The correlation value will be between +1 and -1.

Types of Coorelation are :

Numeric Vs Numeric	Categorical (Binary Feature) Vs Numerical	Ordinal With Ordinal	Categorical vs categorical
Pearson	Pointbiserialr	Spearman Rho	Cross Tab

Different types of correlation has been implemented depending upon the type of variable. For the given data set, the following coorelation methods have been adopted which is depicted in the

Table 16

Table 16 : Correlation Methods Applied for the Dataset

Features/ Data Type			Nominal	Nominal	Nominal	Ordinal	Ordinal	Ordinal	Nominal	Nominal	Nominal					Nominal	Nominal	Nominal	ordinal
			gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raised hands	VisITedResources	AnnouncementsView	Discussion	ParentsAnsweringSurvey	ParentschoolSatisfacon	StudentAbsenceDays	Class
gender	Categorical	Nominal		Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	point biserial	point biserial	point biserial	point biserial	Cross Tab	Cross Tab	Cross Tab	Cross Tab
NationalITy	Categorical	Nominal	Cross Tab		Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	Spearman Rho
PlaceofBirth	Categorical	Nominal	Cross Tab	Cross Tab		Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	Spearman Rho
StageID	Categorical	Ordinal	Cross Tab	Cross Tab	Cross Tab		Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	Spearman Rho
GradeID	Categorical	Ordinal	Cross Tab	Cross Tab	Cross Tab	Cross Tab		Cross Tab	Cross Tab	Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	Spearman Rho
SectionID	Categorical	Ordinal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab		Cross Tab	Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	Spearman Rho
Topic	Categorical	Nominal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab		Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	Spearman Rho
Semester	Categorical	Nominal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab		Cross Tab	point biserial	point biserial	point biserial	point biserial	Cross Tab	Cross Tab	Cross Tab	Cross Tab
Relation	Categorical	Nominal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab			point biserial	point biserial	point biserial	point biserial	Cross Tab	Cross Tab	Cross Tab	Cross Tab
raised hands	Numerical		point biserial							point biserial	point biserial					point biserial	point biserial	point biserial	
VisITedResources	Numerical		point biserial							point biserial	point biserial					point biserial	point biserial	point biserial	
AnnouncementsView	Numerical		point biserial							point biserial	point biserial					point biserial	point biserial	point biserial	
Discussion	Numerical		point biserial							point biserial	point biserial					point biserial	point biserial	point biserial	
ParentsAnsweringSurvey	Categorical	Nominal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	point biserial	point biserial	point biserial	point biserial		Spearman Rho	Cross Tab	Cross Tab
ParentschoolSatisfacon	Categorical	Nominal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	point biserial	point biserial	point biserial	point biserial	Spearman Rho		Cross Tab	Cross Tab
StudentAbsenceDays	Categorical	Nominal	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	Cross Tab	point biserial	point biserial	point biserial	point biserial	Cross Tab	Cross Tab		Cross Tab
Class	Categorical	ordinal	Cross Tab	Spearman Rho	Spearman Rho	Spearman Rho	Spearman Rho	Spearman Rho	Spearman Rho	Cross Tab	Cross Tab					Cross Tab	Cross Tab	Cross Tab	

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

The following inferences has been drawn from the

Table 17. It shows that correlation between various features among other feature using crosstab function, Spearman RHO, Pearson, point biserialr shows that the following features are coo related and could be included for modelling. Nationality, Place of Birth, Stage ID, Grade ID, Section ID, Topic, Semester, Relation, Class, parent Answering Survey, Parent School Satisfaction, Student Absence Days to be included for model along with numerical features. Other features if required using the Feature importance could be later included for modelling.

Table 17 : Correlation Methods Tabulated Values

Features / Data Type			Nominal	Nominal	Nominal	Ordinal	Ordinal	Ordinal	Nominal	Nominal	Nominal					Nominal	Nominal	Nominal	ordinal
			gender	NationalITy	PlaceofBirth	StageID	Grade ID	SectionID	Topic	Semester	Relation	raised hands	VisiTedResources	AnnouncementsVie	Discussion	ParentsAnsweringS	ParentsschoolSatisf	StudentAbsenceDa	Class
gender	Categorical	Nominal	1	0.235	0.258	0.079	0.167	0.059	0.219	0.045	0.191	0.15	0.211	0.052	0.125	0.018	0.089	0.205	0.264
NationalITy	Categorical	Nominal	0.235	1	0.874	0.314	0.294	0.247	0.28	0.274	0.366					0.226	0.341	0.275	0.277
PlaceofBirth	Categorical	Nominal	0.258	0.874	1	0.353	0.296	0.255	0.287	0.228	0.376					0.241	0.316	0.255	0.281
StageID	Categorical	Ordinal	0.079	0.314	0.353	1	0.998	0.086	0.536	0.153	0.044					0.127	0.019	0.12	0.086
GradeID	Categorical	Ordinal	0.167	0.294	0.296	0.998	1	0.4	0.523	0.329	0.144					0.184	0.072	0.167	0.174
SectionID	Categorical	Ordinal	0.059	0.247	0.255	0.086	0.4	1	0.56	0.048	0.035					0.033	0.071	0.052	0.068
Topic	Categorical	Nominal	0.219	0.28	0.287	0.536	0.523	0.56	1	0.528	0.36					0.195	0.224	0.159	0.217
Semester	Categorical	Nominal	0.045	0.274	0.228	0.153	0.329	0.048	0.528	1	0.145	0.178	0.173	0.287	0.019	0.019	0.021	0.068	0.128
Relation	Categorical	Nominal	0.191	0.366	0.376	0.044	0.144	0.035	0.36	0.145	1	0.364	0.36	0.34	0.027	0.16	0.283	0.215	0.412
raisedhands	Numerical		0.15							0.364	0.317					0.317	0.297	-0.464	
VisiTedResources	Numerical		0.211							0.36	0.382					0.382	0.364	-0.499	
AnnouncementsView	Numerical		0.052							0.34	0.396					0.396	0.299	-0.312	
Discussion	Numerical		0.125							0.027	0.232					0.232	0.061	-0.219	
ParentsAnsweringSurvey	Categorical	Nominal	0.018	0.226	0.241	0.127	0.184	0.033	0.195	0.019	0.16	0.317	0.382	0.396	0.232	1	0.536	0.257	0.446
ParentsschoolSatisfaction	Categorical	Nominal	0.089	0.341	0.316	0.019	0.072	0.071	0.224	0.021	0.283	0.297	0.364	0.299	0.061	0.536	1	0.224	0.378
StudentAbsenceDays	Categorical	Nominal	0.205	0.275	0.255	0.12	0.167	0.052	0.159	0.068	0.215	-0.464	-0.499	-0.312	-0.219	0.257	0.224	1	0.685
Class	Categorical	ordinal	0.264	0.277	0.281	0.086	0.174	0.068	0.217	0.128	0.412					0.446	0.378	0.685	1

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

E. Feature Engineering Concepts [11]

It is the process of converting data into features to act as inputs to machine learning models. Variable transformation type is applied in this study, where in the given data set most of the columns are categorical and need to be converted to numerical. The conversion process is done through Label encoding method [12] and the output of the Label Encoding is shown in the **Figure 34** and the formula applied for the label encoding is shown in the **Figure 32**

```
gender_map = {'M':1,'F':2}
NationalITy_map =
{'Iran':1,'SaudiArabia':2,'USA':3,'Egypt':4,'Lybia':5,'lebanon':6,'Morocco':7,'Jordan':8,'Palestine':
9,'Syria':10,'Tunis':11,'KW':12,'Iraq':13,'venzuela':14}
PlaceofBirth_map =
{'Iran':1,'SaudiArabia':2,'USA':3,'Egypt':4,'Lybia':5,'lebanon':6,'Morocco':7,'Jordan':8,'Palestine':
9,'Syria':10,'Tunis':11,'KuwaIT':12,'Iraq':13,'venzuela':14}
StageID_map = {'HighSchool':1,'MiddleSchool':2,'lowerlevel':3}
GradeID_map = {'G-02':2,'G-04':4,'G-05':5,'G-06':6,'G-07':7,'G-08':8,'G-09':9,'G-10':10,'G-
11':11,'G-12':12}
SectionID_map = {'A':1,'B':2,'C':3}
Topic_map =
{'Arabic':1,'Biology':2,'Chemistry':3,'English':4,'French':5,'Geology':6,'History':7,'IT':8,'Math':9,'
Quran':10,'Science':11,'Spanish':12}
Semester_map = {'F':1,'S':2}
Relation_map = {'Mum':1,'Father':2}
ParentAnsweringSurvey_map = {'Yes':1,'No':0}
ParentschoolSatisfaction_map = {'Bad':0,'Good':1}
StudentAbsenceDays_map = {'Under-7':0,'Above-7':1}
Class_map = {'H':3,'M':2,'L':1}
```

Figure 33 : Label Encoding Code

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisITedRes	Announceme	Discussion	ParentAnswer	Parentschool	StudentAbsen	Class
0	1	12	12	3	4	1	8	1	2	15	16	2	20	1	1	0	2
1	1	12	12	3	4	1	8	1	2	20	20	3	25	1	1	0	2
2	1	12	12	3	4	1	8	1	2	10	7	0	30	0	0	1	1
3	1	12	12	3	4	1	8	1	2	30	25	5	35	0	0	1	1
4	1	12	12	3	4	1	8	1	2	40	50	12	50	0	0	1	2

Figure 34 : Label Encoder: Categorical to Numeric Converted Values

Various proposed Classification Algorithms [13] used in this paper are :

1) Logistic Regression — Decision Tree
2) Random Forest — XG Boost
3) K Nearest Neighbors Algorithm — Support Vector Machine

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

IV. EXPERIMENTAL RESULTS

The transformed data set is partitioned into training data set and the test data set where the training data is 70% of the whole data set and the remaining unused 30% is used as Test data set. The random state is set as 0. The parameters applied for various algorithms are depicted in **Table 18**. The experimented results before feature engineering is depicted in

Table 19. Sample code for Logistic Regression and its classification Report has been shown in Table 20 & Figure 35.

Table 18 : Parameters For Model Fitting

Model Type	Parameters for Fitting the Model
Logistic Regression	solver='lbfgs',multi_class='auto', max_iter=2000
Random Forest	RandomForestClassifier(n_jobs=-1, random_state=123, criterion='gini', max_depth=3,)
KNN	KNeighborsClassifier(n_neighbors=7
SVM	svm.SVC(kernel='rbf',gamma='auto') # Linear Kernel
XGBOOST	xgb.XGBClassifier(max_depth=10, learning_rate=0.1, n_estimators=100, seed=10)
DECISION TREE – Gini	DecisionTreeClassifier(criterion = "gini", random_state = 100, max_depth=7, min_samples_leaf=5)
DECISION TREE - Entropy	DecisionTreeClassifier(criterion = "entropy", random_state = 100, max_depth=7, min_samples_leaf=5)

Table 19 : Experimented Results – Before Feature Engineering

Model Type	Training Score	Testing Score
Logistic Regression	79.16	75.0
Random Forest	82.44	75.69
KNN	75.0	61.1
SVM	99.70	50.0
XGBOOST	100.0	74.30
DECISION TREE – Gini	86.90	70.83
DECISION TREE - Entropy	85.11	67.36

Table 20 : Training & Testing Code – Logistic Regression Algorithm

Training Score Code	Testing Score Code
from sklearn.linear_model import LogisticRegression Logit_Model=LogisticRegression(solver='lbfgs', multi_class='auto', max_iter=2000) Logit_Model.fit(X_train,Y_train) Logit_Model.score(X_train,Y_train)	from sklearn.metrics import accuracy_score from sklearn.metrics import classification_report prediction=Logit_Model.predict(X_test) score = accuracy_score(Y_test,prediction) report=classification_report(Y_test,prediction)

```
              precision    recall  f1-score   support

           2       0.85      1.00      0.92        23
           5       0.76      0.69      0.72        45
          10       0.64      0.64      0.64        28

    accuracy                           0.75        96
   macro avg       0.75      0.78      0.76        96
weighted avg       0.75      0.75      0.75        96
```

Figure 35 : Logistic Regression – Classification Report

A. Feature Importance

1) *Random Forest Feature Importance [14]:* Random forests are among the most popular machine learning methods thanks to their relatively good accuracy, robustness and ease of use. They also provide two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy.
2) *Experimented Results after Feature Engineering:* The Feature Engineering process applied data set is divided into training data set and the test data set where the training data is 70% of the whole data set and the remaining unused 30% is used as Test data set. The random state is set as 50 here, whereas in the previous phase it was set as 0.

Table 21 : Experimented Results –After Feature Engineering

Model Type	Training Score	Testing Score	Remarks
Logistic Regression	87.20	86.81	Good
Random Forest	94.05	90.97	Fair
KNN	81.54	82.63	Good
SVM	97.91	83.33	Needs more Testing Effort
XGBOOST	97.02	90.27	Needs more Testing Effort
DECISION TREE – Gini	81.25	76.38	Needs more Testing Effort
DECISION TREE - Entropy	80.65	81.25	Good

V. CONCLUSION

The Machine learning methodology is rapidly increasing and the impact of the machine able to predict the result of a system by itself and also it is able to train a data over a period of time and also test the trained model with a different set of data to prove that the model is working efficiently and effectively. In this research study it has been apparently proved that Logistic Regression has got a training score of 87.20 and a testing score of 86.81 has proved that the model is working effectively without any bias or variance concept. KNN and Decision Tree Entropy also works good and other implemented algorithms in this research study needs some more feature engineering concepts and data analysis in a stronger term. The model deployment has been done for all algorithms and the sample input has been given for evaluation, which classified perfectly in all algorithms.

VI. FUTURE SCOPE

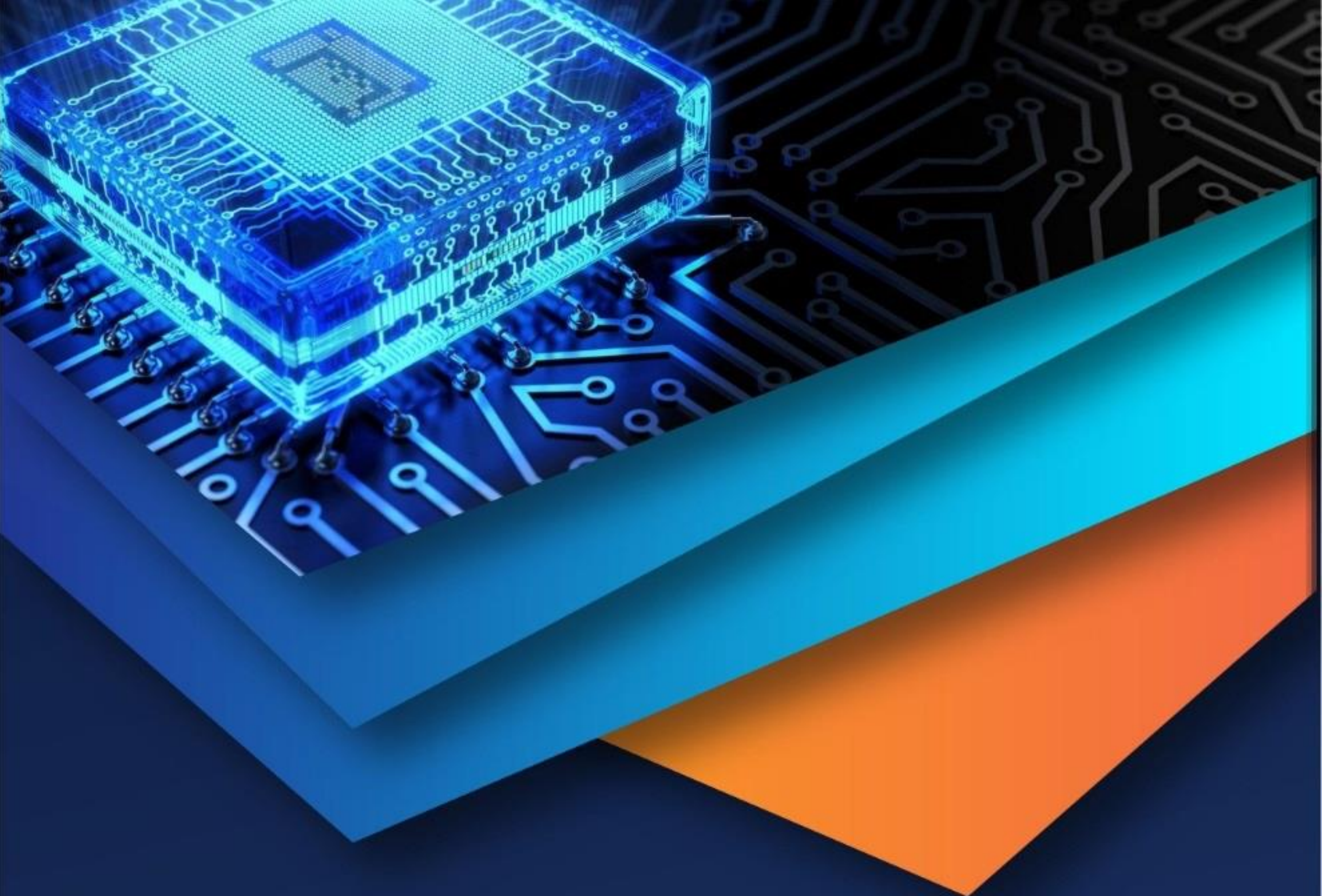
The present study predicting the Academic performance of students with respect various features have considerably proved positive results. This research work increases the performance prediction process of student in an effective way. When considering the future this work can be further extended by using other feature(s) as Target Variable.

A. Other Features such as Financial Impacting feature, Physical Health Impacting feature and practicing food habits feature can also be included in the upcoming research study.
B. As the above factors also can create an impact on the academic performance of the student directly or indirectly.
C. Since the present study focused on predicting the academic performance [5] of the student other factors included can also be experimented to predict the performance of the student not only in academic point of view but also in a behavior perspective.

REFERENCES

[1] Smola, Alex, and S.V.N. Vishwanathan. Introduction to Machine Learning. Cambridge University Press, 2008. N.p., 2008. Web.
[2] Amjad Abu Saa. (2016) "Educational Data Mining & Students' Performance Prediction" International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.
[3] Ahmed Mueen, Bassam Zafar and Umar Manzoor. (2016) "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques" I.J. Modern Education and Computer Science, 2016, 11, 36-42.
[4] Bhrigu Kapur, Nakin Ahluwalia and Sathyaraj R, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms", International Journal of Advanced Research in Computer Science, 8 (3), March-April 2017,632-636
[5] Prasada Rao, K. , M. V.P. Chandra Sekhara, and B. Ramesh. "Predicting Learning Behavior of Students using Classification Techniques." International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, April 2016.

[6] Amrieh, E. A., Hamtini, T. & Aljarah, I. (2016). Mining educational data to predict Student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), pp. 119–136. doi: 2016.9.8.13.
[7] Sundar PVP. A Comparative Study For Predicting Students Academic Performance using Bayesian Network Classifiers. IOSR Journal of Engineering. 2013 Feb; 3(2):37–42.
[8] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006
[9] C. Romero, "Educational Data Mining: A Review of the State of the Art", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 40, 2010.
[10] https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/
[11] https://www.kdnuggets.com/2018/12/feature-engineering-explained.html
[12] https://towardsdatascience.com/encoding-categorical-features-21a2651a065c
[13] https://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf
[14] https://blog.datadive.net/selecting-good-features-part-iii-random-forests/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY