Exploring Student Performance Prediction: A Comprehensive Review of Data Mining Techniques

Pooja [1], Dr Rajni Bhalla [2]

[1] Research Scholar (Computer Application) LPU Phagwara, Punjab, India

[2] Associate Professor, School of Computer Application, LPU Phagwara, Punjab, India

Email: [1] Poojagori51@gmail.com, [2] rajni.b27@gmail.com

Abstract

Many studies have been done on predicting students' academic performance over several years. A key indicator of education quality is student performance. In educational institutions, huge amount of data is stored about the students that can be explored to learn more about how they are learning and using data mining methods, to enhance their performance in advance. Machine learning algorithms, including KNN, Naïve Bayes, Decision Tree, ANN, Logistic Regression and Support vector machine, have been widely utilized in these studies, encompassing supervised, unsupervised, and semisupervised learning approaches. The main focus of the study is to identify the most suitable methods for predicting the performance of students from the institutes of Jalandhar. The sample size is 280 respondents of either male or female across Jalandhar. They belong to different classes of schools in Jalandhar city. In this paper, we will look for elements that could improve or help pupils perform better. In this we present the systematic literature review of 30 papers from Google scholar which helps to predict the factors that affecting on students' academic performance Mean is used for descriptive data analysis, Pearson's Correlation is used get the relation between attributes. T-Test and ANOVA are used for Hypothesis testing. All the experiments were conducted on GoogleColab and RapidMiner Studio tool.

Index Terms— Data mining, Machine Learning, Students performance predicttion.

1. Introduction

Education is an essential factor of development and a key contributor to wellbeing. The educational success of the students is an essential part of the education. As there is a lot of data stored in educational databases, students' performance prediction has becoming difficult task. To find enduring patterns that enhance students' knowledge and academic institutions generally, machine learning and statistical techniques for educational data are analyzed. Low-performing students will experience a variety of difficulties, including late graduation and perhaps quitting out [1-2]. Therefore, educational resources need to maintain an open mind on the educational progress of their students and act quickly to help those who perform poorly and wish to drop out. Predicting students' academic success is one way to do that [1]. This technique will assist educational institutions in early detection and support those students having low performance.

Many researches has been done on Student academic performance prediction and the effectiveness Even though the results have been utilized to suggest ways to improve the educational sector, the performance prediction accuracy of the established models is not sufficient to provide early student identification and intervention. Low graduation, student retention, and advancement as a result of this have had a significant negative effect on the economy and the society due to the loss of potential skill and knowledge [3]. Currently many [4] researchers are working on educational datasets to study the aspects that are intended to cause problems with students' performance [5]. Without students, schools, colleges, and universities are worthless. The most valuable resource for any educational institution is its students. Performance of students is closely related to the economic and social development of the nation. The quality of the education provided to students determines whether they will become good leaders and members of the nation, who will then be in charge of the nation's economic and social growth [6(a)].

Using educational data mining (EDM), a model of user thinking, work and experimentation was made. A variety of machine learning methods were used in EDM such as K-NN, Naive Bayes (NB), Neural Networks (NN), Decision Trees (DT), and Support Vector Machines (SVM), and many more. WEKA, KNIME, RAPID MINER, and SSDt (SQL Server data Tools) are just a few examples of the many open source tools that can be used to analyze datasets using machine learning algorithms. These tools were created for pattern discovery and logical arrangement for later use. Educational Data Mining is a new field which focuses on to analyze the information generated by educational environments using DM techniques. It is used to examine the distinctive and increasingly large volumes of information derived from educational settings and use such techniques to effectively comprehend the students. The significance of applying EDM models lies in making use of the previous data of the students for predicting the unseen or upcoming their performances. This idea has attracted several researchers to develop classification prediction models the hidden factors of future instances. In data mining methodology, various steps are used such as:

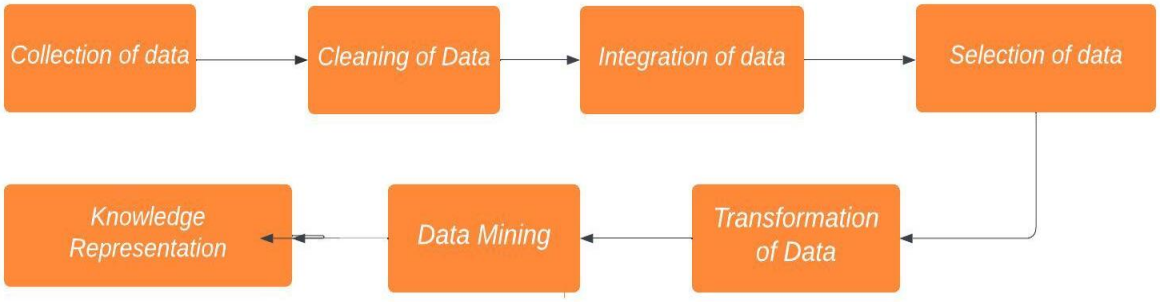


Fig: Data mining methodology

2. Literature Review

A new model was proposed by Mustafa Yagci to predict undergraduate students' final exam grades, based on machine learning algorithms using their midterm exam scores as the main source.The performances of Random Forests, KNN, SVM and Logistic Regression were calculated and compared. KNN provided 78% accuracy [7]. For prediction and classification the students' performance, Sekeroglu et al. [8] used many machine learning algorithms such

as Support Vector Regression,Gradient Boosting Classifier , Back Propagation, and Long short term memory algorithms and resulted SVR with 87.88% accuracy. For students' performance prediction, Yossy et al. [9] used 7 methods, namely Classification and regression trees, Random Forest(RF), Naïve Bayes(NB), Extratree, Bernaoulli naïve bayes, AdaBoost and KNN. They concluded Random Forest G-89.78% accuracy. Kaswan et al. [10] used 5 classification tasks, including ANN, SVM, Decision Tree, Naive Bayes, and KNN to predict student outcomes. The attributes used for this analysis are demographics, internal/ external assessment. The experiment's results demonstrate that a neural network's prediction error is less significant and more precise than other tasks. The WEKA Naive Bayes technique to build prediction models was proposed by Kaswan et al. [11] as a method for predicting student success. Six factors were tested such as gender, university, income households, ethnicity, home town and CGPA. Among them three factors income household, gender and home town contributed in university success.

Mohammadi et al [12] used algorithms such as NB, KNN and Decision Tree (DT) and to discover information in student's raw data, and resulted that KNN 54.64% accuracy which is better than other. Bhalla et at [31] used decision tree, KNN, Naïve Bayes on nursery dataset. All these three techniques achieved precision 72% to 90%.

Ref	Source of Dataset	Features Considered for evaluation	ML Algorithms used	Accuracy
[13]	3166 students from Open University Learning Analytics (OULAD) dataset	Demographics, historic performance, Student engagement	Naive Bayes (NB), K-Nearest Neighbor (K-NN), Decision Tree (DT), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Logistic Regression (LR)	ANN with F1 score 96%
[14]	335 students from Ecuador University	Historic performance	Decision Tree	96.5% accuracy
[15]	323 students from Engineering Dynamics course	Historic performance	Radial Basis Function (RBF) network, Multiple Linear Regression (MLR), Multilayer Perception (MLP) network, and Support Vector Machine (SVM)	MLR gave 89.7% accuracy
[16]	384 students from OU University, United Kingdom	student-system engagement, assessment scores and Demographic	Gradient-Boosted Tree (GBT), CART, Decision Tree (DT), JRIP, J48, and Naive Bayes	J48 gave 88.5% accuracy
[17]	11,556 students University of Minnesota's Moodle	Historic performance, Student engagement	Multi-Regression	RMSE= 0.145
[18]	115 students from DEEDS,	Historic performance Student engagement,	SVMs, ANNs, Naive Bayes classifications, Logistic Regression, and Decision Trees	SVM with accuracy 80%
[19]	115 students from DEEDS	Student engagement, historic performance	Artificial Neural Networks(ANN), Linear Regression(LR) and Support Vector Machine (SVM)	SVM gave 95% accuracy

[20]	115 students from DEEDS	Student engagement, historic performance	Artificial Neural Network(ANN), Decision Tree(DT), Support Vector Machine, and Naïve Bayes (NB) and Logistic Regression(LR)	SVM gave 94.8% accuracy
[21]	696 records from Mathematics dataset	Historic performance, demographic data, personality type of features	Random Forest(RF) and SVM	SVM gave 92% accuracy
[22]	649 records from Portuguese dataset	Historic performance, demographic data, personality	Support Vector Machine(SVM), Random Forest(RF)	RF gave 94% accuracy
[23]	722 students	Historic performance, personality engagement, institutional	Neural Network, LR, Naïve Bayes, SVM, KNN, CN2 Rule Inducer	RF gave 88% accuracy
[24]	3518 students	student engagement and historic performance	Artificial Neural Networks (ANN)	80%. accuracy

3. Proposed Methodology

Our approach for predicting student performance will be based on a hybrid method, which combines the multiple techniques or methodologies to enhance prediction, accuracy and effectiveness. Model Performance Testing: To validate the proposed model, it will be tested using existing performance metrics or benchmarks. This phase will provide insights into model's accuracy in predicting student performance based on collected data. By following this methodology, we aim to develop a robust and accurate model for predicting student performance, leveraging both online/offline data sources and employing a hybrid approach to enhance prediction capabilities.

Methodology used

- **Collection of data:** Primary data is collected for the above mentioned study.
- **Sample Size:** The sample size is 280 respondents of either male or female across Jalandhar. They belong to different classes of schools in Jalandhar city. .
- **Sampling Technique:** convenience sampling has been used for this study.
- **Research Technique:** Mean is used for descriptive data analysis; Pearson's Correlation is used get the relation between attributes. T-Test and ANOVA are used for Hypothesis testing.

Data Profile

The whole sample is divided on the basis of gender

Sr No	Gender	No. of respondents
1	Male	70
2	Female	210
	Total	280

The sample includes 70 male and 210 female respondents on the basis of gender.

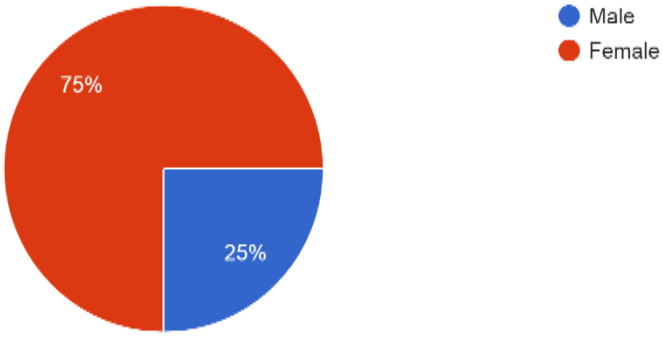


Figure 1a: Pie chart of gender classification

Descriptive statistics (Analysis of Questionnaire)

Descriptive statistical analysis helps to describe the data. It obtains a data summary so that information that is meaningful can be interpreted from it. We don't come to a conclusion using descriptive analysis, but we do understand the contents of data. This paper presents the bar charts and pie-charts of responses received for various questions.

Domestic factors affecting the performance of students

Question 1) Do my parents encourage me to work hard in my studies?

Table 2(a) Summary of response to question1

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	0	0	0	0
Agree	70	25.0	25.0	25.0
Neutral	130	46.4	46.4	71.4
Disagree	50	17.9	17.9	89.3
Stongly Disagree	30	10.7	10.7	100.0
	280	100.0	100.0	

Although 25% respondents agreed that parents encourage them to work hard in their studies. 46.4% member's responded neutral and 10.7% were negative in response.

Question 2) Do my parents live harmoniously?

Table 2 (b) Summary of response to question 2

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	50	17.9	17.9	17.9
Agree	120	42.8	42.8	60.7
Neutral	100	35.7	35.7	96.4
Disagree	10	3.6	3.6	100.0
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 17.9% respondents strongly agree that their parents live harmmoniously, 35.7 % respondents were neutral . 42.8% respondents agreed.

Question 3) My parents are able to pay for my further education.

Table 2(c) Summary of response to question 3

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	40	14.2	14.2	14.2
Agree	140	50	50	64.2
Neutral	50	17.9	17.9	82.1
Disagree	50	17.9	17.9	100
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 14.2% respondents strongly agree that their parents are able to pay for further education,50% respondents agreed and 17.9 % respondents disagreed.

Question 4) My parents attend all meeting whenever they are called by school.

Table 2(d) Summary of response to question 4

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	40	14.3	14.3	14.3
Agree	150	53.6	53.6	67.9
Neutral	90	32.1	32.1	100.0
Disagree	--	--	--	
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 14.3% respondents strongly agree ,53.6% respondents agree that their parents attend all meetiins whenever they are called by school.

Question 5) I am never absent without any good reason.

Table 2(e) Summary of response to question 5

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	10	3.6	3.6	3.6
Agree	130	46.4	46.4	50.0
Neutral	90	32.1	32.1	82.1
Disagree	30	10.7	10.7	92.8
Stongly Disagree	20	7.2	7.2	100.0
	280	100.0	100.0	

Although 46.4% respondents agreed ,32.1% respondents neutral whereas 10.7% respondents disagreed that I am never absent without any good reason .

Question 6) Education helps me to get a good job in future.

Table 2(f) Summary of response to question 6

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	40	14.3	14.3	14.3
Agree	100	35.7	35.7	50.0
Neutral	60	21.4	21.4	71.4
Disagree	80	28.6	28.6	100.0
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 35.7% respondents agreed ,21.4% respondents neutral whereas 28.6 respondents disagreed that education help them to get a good job .

Question 7) Overall, how satisfied are you with your teachers?

Table 2(g) Summary of response to question 7

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	50	17.9	17.9	17.9
Agree	110	39.3	39.3	57.2
Neutral	90	32.1	32.1	89.3
Disagree	30	10.7	10.7	100.0
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 32.1% respondents were neutral, 39.3% respondents agreed that they are satisfied with their teachers.

Question 8)What type of the family do you live in?

Table 2(h) Summary of response to question 8

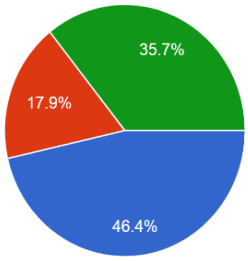
	Frequency	Percent	Valid Percent	Cumulative Percent
Joint	160	57.1	57.1	57.1
Nuclear	120	42.9	42.9	100.0
	280	100.0	100.0	

Although 57.1% respondents agreed that they are live in joint family.

Question 9) Affiliation of the school ?

Table 2(i) Summary of response to question 9

Affiliation of the School	Frequency	Percent	Valid Percent	Cumulative Percent
PSEB	130	46.4	46.4	46.4
ICSE	50	17.9	17.9	64.3
CBSE	100	35.7	35.7	100.0
	280	100.0	100.0	



Although 46.4% respondents belongs to school having PSEB affiliation, 17.9% belongs to ICSE and rest of belongs to school having CBSE affiliation.

School related factors affecting the performance of students

Question 10) Do you/parents look for popularity of school when taking admission ?

Table 2(j) Summary of response to question 10

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	70	25	25	25
Agree	90	32.1	32.1	57.1
Neutral	80	28.6	28.6	85.7
Disagree	40	14.3	14.3	100.0
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 32.11% respondents agreed, 25% respondents strongly agreed that they look for popularity of school when taking admission.

Question 11) Do you/parents look for previous result of the school when taking admission ?

Table 2(k) Summary of response to question 11

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	20	7.1	7.1	7.1
Agree	130	46.4	46.4	53.5
Neutral	80	28.6	28.6	82.1
Disagree	30	10.8	10.8	92.9
Stongly Disagree	20	7.1	7.1	100.0
	280	100.0	100.0	

Although 46.4% respondents agreed, 7.1% respondents strongly agreed that they look for previous result of school when taking admission.

Question 12) Total number of family members?

Table 2(l) Summary of response to question 12

Number of family members	Frequency	Percent	Valid Percent	Cumulative Percent
Between 3-5	90	32.1	32.1	32.1
Between 6-10	150	53.6	53.6	85.7
More than 10	40	14.3	14.3	100.0
	280	100.0	100.0	

Although 32.1% respondents belongs to family having members between 3-5, whereas 536% repondents belongs to family having members between 6-10.

Question 13) Do you /your parents look for Co-curriculum activities in the school when taking admission ?

Table 2(m) Summary of response to question 13

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	10	3.6	3.6	3.6
Agree	130	46.4	46.4	50.0
Neutral	90	32.1	32.1	82.1
Disagree	30	10.7	10.7	92.8
Stongly Disagree	20	7.2	7.2	100.0
	280	100.0	100.0	

Although 46.4% respondents agreed that they look for co-curriculum activities of school while taking admission.

Question 14) Do you /your parents look for extra-curricular activities in the school when taking admission ?

Table 2(n) Summary of response to question 14

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	50	17.9	17.9	17.9
Agree	110	39.3	39.3	57.2
Neutral	90	32.1	32.1	89.3
Disagree	30	10.7	10.7	100.0
Stongly Disagree	--	--	--	-
	280	100.0	100.0	

Although 32.1% respondents were neutral that they look for extra curricular activities of school while taking admission.

Question 15) Do you /your parents look for transoprt facility available in the school when taking admission ?

Table 2(o) Summary of response to question 15

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	40	14.3	14.3	14.3
Agree	160	57.1	57.1	71.4
Neutral	70	25.0	25.0	96.4
Disagree	--	--	--	96.4
Stongly Disagree	10	3.6	3.6	100.0
	280	100.0	100.0	

Although14.3% respondents were strongly agreed that they look for transport facility of school while taking admission.

Skill and responsiveness of teacher that affecting the performance of students

Question 16) Should there be an effective lecturer/demostrator?

Table 2(p) Summary of response to question 16

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	150	53.6	53.6	53.6
Agree	90	32.0	32	85.6
Neutral	20	7.2	7.2	92.8
Disagree	10	3.6	3.6	96.4
Stongly Disagree	10	3.6	3.6	100.0
	280	100.0	100.0	

Although 53.6% respondents were strongly agreed that instructor will be an effective lecturer/demonstrator.

Question 17) Should the presentations be clear and organized?

Table 2(q) Summary of response to question 17

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	70	25	25	25
Agree	130	46.4	46.4	71.4
Neutral	40	14.3	14.3	85.7
Disagree	40	14.3	14.3	100.0
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 46.4% respondents were agree, whereas 25% were strongly agreed that presentation must be clear and organized.

Question 18) Should the instuctor be available and helpful?

Table 2(r) Summary of response to question 18

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Agree	30	10.7	10.7	10.7
Agree	120	42.9	42.9	53.6
Neutral	50	17.9	17.9	71.5
Disagree	80	28.5	28.5	100.0
Stongly Disagree	--	--	--	--
	280	100.0	100.0	

Although 17.9% respondents were neutral, whereas 42.9% were agreed that instructor should be helpful and available.

Question 19) Do you find time to do your studies at home?

Table 2(s) Summary of response to question 19

Time for study	Frequency	Percent	Valid Percent	Cumulative Percent
More than 3 hours	130	46.4	46.4	46.4
only about 1-2 hours	100	35.7	35.7	82.1
I don't have any time	50	17.9	17.9	100.0
	280	100.0	100.0	

46.4% respondents agreed that they have more than 3 hours for study at home whereas 17.9% said they don't have any time.

4. Result of Pearson Correlation

In this study, source of dataset is online from kaggle
http://www.kaggle.com/datasets/larsen0966/student-performance-data-set.
Student-math and Student-por are 2 files names which show the performance of 649 students in mathematic course and Portuguese course respectively. Each file contains thirty three (33) attributes such as age, gender, PStatus, Medu, Fedu, Mjob, Travel Time, G1, G2, G3 and many more.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc
school	1.000000	0.083050	0.087170	0.354520	-0.022252	0.028120	-0.254787	-0.209806	-0.087114	-0.113320	...	-0.031597	0.034666	0.044632	0.047169
sex	0.083050	1.000000	0.043662	0.025503	0.098205	-0.064700	-0.119127	-0.083913	-0.068955	-0.003157	...	-0.083473	-0.146305	-0.058178	-0.282696
age	0.087170	0.043662	1.000000	0.025848	0.002470	-0.005631	-0.107832	-0.121050	-0.014556	0.029903	...	-0.020559	-0.004910	0.112805	0.134768
address	0.354520	0.025503	0.025848	1.000000	0.046113	0.094635	-0.190320	-0.141493	-0.062200	0.066236	...	0.033897	0.036647	-0.015475	0.047304
famsize	-0.022252	0.098205	0.002470	0.046113	1.000000	0.239608	0.014325	0.039538	0.034355	0.013862	...	-0.004641	0.021257	0.004312	-0.060482
Pstatus	0.028120	-0.064700	-0.005631	0.094635	0.239608	1.000000	-0.057174	-0.031856	-0.038208	-0.053428	...	0.051303	0.037585	0.031086	0.041513
Medu	-0.254787	-0.119127	-0.107832	-0.190320	0.014325	-0.057174	1.000000	0.647477	0.101969	-0.030613	...	0.024421	-0.019686	0.009536	-0.007018
Fedu	-0.209806	-0.083913	-0.121050	-0.141493	0.039538	-0.031856	0.647477	1.000000	0.018162	-0.098847	...	0.020256	0.006841	0.027690	0.000061
Mjob	-0.087114	-0.068955	-0.014556	-0.062200	0.034355	-0.038208	0.101969	0.018162	1.000000	0.161032	...	0.005917	0.015351	0.007630	0.007883
Fjob	-0.113320	-0.003157	0.029903	0.066236	0.013862	-0.053428	-0.030613	-0.098847	0.161032	1.000000	...	0.034751	0.021646	0.026614	-0.030667
reason	0.000935	-0.033349	-0.024693	0.023587	-0.004526	-0.007825	0.087954	0.049901	0.012382	-0.007079	...	0.032472	-0.059167	-0.006705	0.060315
guardian	0.033364	-0.019301	0.171760	0.026165	0.019892	0.007587	-0.111274	0.006902	0.036735	-0.045428	...	-0.036994	-0.003350	-0.025124	0.117029
traveltime	0.252936	-0.040880	0.034490	0.344902	-0.012794	0.040633	-0.265079	-0.208288	-0.070649	0.102045	...	-0.009521	0.000937	0.057454	0.092824
studytime	-0.137857	0.206214	-0.008415	-0.062023	0.010945	-0.008748	0.097006	0.050400	0.028265	-0.043094	...	-0.004127	-0.068829	-0.075442	-0.137585
failures	0.113788	-0.073888	0.319968	0.063824	0.066068	-0.009881	-0.172210	-0.165915	-0.073155	0.014310	...	-0.062645	0.108995	0.045078	0.105949
schoolsup	0.123340	-0.111202	0.167841	0.017956	-0.056405	0.009456	0.022168	-0.023572	-0.031313	0.013929	...	0.012038	0.015611	0.058124	0.028076
famsup	0.063720	-0.129467	0.101894	0.005577	-0.039819	-0.010203	-0.120491	-0.135191	0.051871	0.079804	...	-0.015228	-0.003764	-0.017262	0.016844

We import the dataset in GoogleColab and after applying correlation among attributes, we get the result in the form of below table and we obtained that attributes like Medu, Fedu, Mjob, Fjob attributes have positive impact whereas Past_failure, family_size and Travel_time have negative impact on student grades i.e. G3.

Table3: Analysis and interpretation of data

No	**Perception of Respondents**	**Mean**
1	Do my parents encourage me to work hard in my studies?	2.857
2	Do my parents live harmoniously?	3.75
3	My parents are able to pay for my further education.	3.607
4	My parents attend all meeting whenever they are called by school.	3.821

No	Perception of Respondents	Mean
5	I am never absent without any good reason.	3.285
6	Education helps me to get a good job in future.	3.357
7	Overall, how satisfied are you with your teachers?	3.642
8	What type of the family do you live in?	3.357
9	Affiliation of the school?	3.571
10	Do you/parents look for popularity of school when taking admission?	3.687
11	Do you/parents look for previous result of the school when taking admission?	2.571
12	Total number of family members?	3.821
13	Do you /your parents look for Co-curriculum activities in the school when taking admission?	3.285
14	Do you /your parents look for extra-curricular activities in the school when taking admission?	3.642
15	Do you /your parents look for transport facility available in the school when taking admission ?	4.107
16	Should there be an effective lecturer/demonstrator?	3.357
17	Should the presentations be clear and organized?	3.785
18	Should the instructor be available and helpful?	2.925
19	Do you find time to do your studies at home?	3.571

The above table shows the results of perceptions of the respondents. The average score is above 3. So the respondents' perception does not carry a negative opinion on any question. The respondents view it as neutral to agree.

Inferential Analysis of Data:

To make the generalization of population by using samples, inferential statistics are used for it. Where the sample is taken from the whole population. The term "inferential statistics" refers to the fact that sampling is bound to have errors and cannot be assumed to represent the population perfectly. Following steps are taken for drawing inferences:

1) State research hypothesis as a null (H o) and alternate (H a) hypothesis.
2) Collect data in a way designed to test the hypothesis.
3) T-test is performed on Gender category to study the effect of gender on performance of students.
4) ANOVA test is performed on **find time to do studies at home** to study the effect of student performance (Don't have time, only about 1-2 hours, more than 3 hours) for better performance.
5) the null hypothesis is supported if p-value is greater than 0.05 and refuted if p-value is less than 0.05

The paper tested 8 hypotheses using ANOVA test conducted on responses gathered from questionnaire. Moreover 14 hypothesis is tested using T-test conducted on responses collected from various respondents.

Table4: Based on find time to do studies at home

ANOVA TEST

Sr.No.	Question	F-value	P-value	Null Hypothesis(H0)	Accepted/ Rejected
1	Do my parents encourage me to work hard in my studies?	0.716	0.499	There is no difference in online platform for sharing material(p>0.05)	p>0.05 (Accepted)
2	Do my parents live harmoniously?	0.84	0.444	Almost all the parents live harmoniously	p>0.05 (Accepted)
3	My parents are able to pay for my further education	1.262	0.301	Mostly parents are able to pay for my further education	p>0.05 (Accepted)
4	My parents attend all meeting whenever they are called by school.	0.218	0.806	All the meeting attended by parents	p>0.05 (Accepted)
5	I am never absent without any good reason.	0.193	0.826	Never absent without any reason	p>0.05 (Accepted)
6	Education helps me to get a good job in future.	0.106	0.9	Education will help to get better job	p>0.05 (Accepted)
7	Overall, how satisfied are you with your teachers?	0.913	0.403	All the teachers are equivalent providing information among students	p>0.05 (Accepted)
8	Do you/parents look for popularity of school when taking admission?	0.7	0.333	Popularity of school also matters	p>0.05 (Accepted)

Table 5: Two sample T-test Based on Gender for factors affecting students performance

Sr.No.	Question	P-value	Null Hypothesis(H0)	p-value>0.05
1	Do my parents encourage me to work hard in my studies?	0.4382	H0: There is no difference in encouragement of parents amongst Males and females	Accept
2	Do my parents live harmoniously?	0.3685	H0: There is no difference in parents live harmoniously as seen by Males and females	Accept

Sr.No.	Question	P-value	Null Hypothesis(H0)	p-value>0.05
3	My parents are able to pay for my further education.	0.2508	H0: There is no difference in ability to pay fee for education as seen by Males and females	Accept
4	My parents attend all meeting whenever they are called by school.	0.5655	H0: There is no difference in terms of gender in attending meeting called by school.	Accept
5	I am never absent without any good reason.	0.2325	H0: There is no difference in terms of gender in absence from school.	Accept
6	Education helps me to get a good job in future.	0.8062	H0: There is no difference in perspectives of gender that education help them to get good job	Accept
7	Overall, how satisfied are you with your teachers?	0.7932	H0: There is no difference in perspectives of males/females in satisfaction of teachers.	Accept
8	Do you/parents look for popularity of school when taking admission?	0.5412	H0: There is no difference in perspectives of males/females in looking the popularity of school	Accept
9	What type of the family do you live in?	0.2554	H0: There is no difference in terms of males/females type of family they live.	Accept
10	Do you/parents look for previous result of the school when taking admission?	0.722	H0: Males and Females have equal belief in their parents look for previous result of the school	Accept
11	Do you /your parents look for transoprt facility available in the school when taking admission ?	0.8223	H0: Males and Females have equal belief in availability of transport facility	Accept
12	Affiliation of the school ?	0.4948	H0: Males and Females have equal belief in affiliation of school	Accept
13	Should there be an effective lecturer/demostrator?	0.5999	H0: Males and Females have equal belief in effective session of lecturer	Accept
14	Should the presentations be clear and organized?	0.8072	H0: Males and Females have equal belief in organized and clear presentation	Accept

5. Conclusion

In conclusion, the prediction of student academic performance has remained a crucial area of investigation across various academic fields. An important research area in many academic areas for a long time has been predicting students' academic success. In this paper, dataset of 649 students were used to predict the factors affecting the grade G3. This procedure enables instructors to quickly assess student achievement and plan more effective strategies for enhancing their students' educational experience. In addition to an analysis of student achievement, a proposed automated evaluation strategy for evaluating student performance has been made. Naïve bayes is used frequently to find parameter that affects the performance of student. But Naive Bayes faced issue of zero probability [26-27]. Author proposed new algorithm RB- Bayes to solve this issue.

After reviewing the 30 research papers, we concluded that domestics, historical, behavioral and socio-economic cause factors and psychological factors play a vital role in the prediction of students' performance. Future research priorities for higher education should include creating an intelligent learning model that forecasts programme results given the variety of programmes and student circumstances [28] .Students' performance also affected by sleep, mood and eating habits [29]. Parents, administrators, educators, psychologists, and counselors are just a few of the stakeholders who have shown concern over student's poor academic achievement. [30]

The survey was conducted to identifying the factors that affecting the performance of students. The result indicates that most of the respondents accept the domestic factors, factors related to teachers as well as institute/school also affect on the students' performance. T-Test was applied to test the difference in perceptions. It was concluded that male and female respondents carry same view point[32] . ANOVA test is conducted to find time for studies at home Almost in all the cases, null hypothesis is accepted and opinion of respondents is same .

6. Future Work

In future, we will ensemble hybrid model based on ANN and Decision Tree as they gave highest accuracy 96% and 96.5% among various ML algorithms, predict the factors which are mostly affected on students' performance and will perform comparative analysis to find best model for prediction.

References

[1] Costa,E.B.;Fonseca, B.;,Santana, M.A.; de Araujo, F.F.;Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students'academic failure in introductory programming courses- Comput. Hum.Behav. 2017,73,247-256. [CrossRef]

[2] Liao, S.N.; Zingaro, D.;Thai, K.;Alvarado ,C.;Griswold, W.G.;Porter, L. A robust machine learning technique to predict low- performing students. ACM Trans. Comput. Educ. (TOCE) 2019, 19,1-19 [CrossRef]

[3] Crosling ,G. ;Heagney, M. ;Thomas, L. improving student retention in higher education: improving teaching and learning, Australian Universities' Review, 2009, 51,9-18.

[4] Lu, O.H. ;Huang, J.C; Lin, A.J; Ogata, H.;Yang ,S.J.; Applying learning analytics for early prediction of students' academic performance in blended learning, J.Educ. Technol. Soc. 21(2), 2018, 220-232.

[5] Suhem,P.;Zain,Z.;Fatima, M.;Application of data mining in educational databases for predicting academic trends and patterns, in : 2012 IEEE International Conference on Technology Enhanced Education (ICTEE) ,2012, 1-4.

[6(a)] Ali I,S.Haider, Z., Munir1, F. Khan, H. and Ahmed, A. ;Factors Contributing to the students' Academic Performance. A case study of Islamia University, sub-campus, Science and Education Journal, 2013, Vol 1 issue 8, 1-10.

[6(b)] Alapont, J.; Bella-Sanjuán, A.; Ferri, C.; Hernández-Orallo, J.; Llopis-Llopis, J.; Ramírez-Quintana, M. Specialised tools for automating data mining for hospital management. In Proceedings of the First East European Conference on Health Care CModelling and Computation, Craiova, Romania, 31 August–2 September 2005; pp. 7–19.

[7] Mustafa Yagci.; Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 2022, pp1-19. https://doi.org/10.1186/s40561-022-00192-z

[8] Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. Proceedings of the 2019 8th International Conference on Educational and Information Technology - ICEIT 2019. https://doi.org/10.1145/3318396.3318419

[9] Yossy, E. H., Heryadi, Y., & Lukas. (2019). Comparison of Data Mining Classification Algorithms for Student Performance. 2019 IEEE International Conference on Engineering, Technology and Education (TALE). https:// doi.org/10.1109/tale48000.2019.9225887

[10] K. S. Kaswan, ―A new nature inspired optimization technique,‖ International Journal of Advanced Research Engineering and Technology (IJARET), vol. 12, issue 2, 2021.

[11] S. Sharma, N. Kumar, and K. S. Kaswan, ―Big data reliability: A critical review,‖ Journal of Intelligent & Fuzzy Systems, vol. 41, pp. 1-16, 2021.

[12] Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019). Comparative study of supervised learning algorithms for student performance prediction. 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). doi:10.1109/icaiic.2019.8669085

[13] Tomasevic, N.; Gvozdenovic, N.; Vranes, S.: An overview and comparison of supervised data mining techniques for student exam performance prediction. Comput. Educ. 143, 103676 (2020)

[14] Buenaño-Fernández, D.; Gil, D.; Luján-Mora, S.: Application of machine learning in predicting performance for computer engineering students: a case study. Sustainability 11(10), 2833 (2019)

[15] Huang, S.; Fang, N.: Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Comput. Educ. 61, 133–145 (2013)

[16] Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R.: Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput. Intell. Neurosci. (2018)

[17] Elbadrawy, A.; Studham, R.S.; Karypis, G.: Collaborative multi-regression models for predicting students' performance in course activities. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, pp. 103–107 (2015)

[18] Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R.; Ali, S.: Using machine learning to predict student difficulties from learning session data. Artif. Intell. Rev. 52(1), 381–407 (2019)

[19] Sriram, K.; Chakravarthy, T.; Anastraj, K.: A comparative analysis of student performance prediction using machine learning techniques with DEEDS lab. J. Compos. Theory XII(VIII) (2019)

[20] Maksud, M.; Nesar, A.: Machine learning approaches to digital learning performance analysis. Int. J. Comput. Digit. Syst. 10, 2–9 (2020)

[21] Leena, H. A; Ranim, S. A; Mona, S. A; Dana, K. A; Irfan, U. K; Nida, A.: Predicting Student Academic Performance using Support Vector Machine and Random Forest. 3rd International Conference on Education Technology Management. pp. 100–107 (2020)

[22] Leena, H. A; Ranim, S. A; Mona, S. A; Dana, K. A; Irfan, U. K; Nida, A.: Predicting Student Academic Performance using Support Vector Machine and Random Forest. 3rd International Conference on Education Technology Management. pp. 100–107 (2020)

[23] Hasan, R.; Sellappan, P.; Salman, M.; Ali, A.; Kamal, U.S.; Mian, U.S.: Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Appl. Sci. 10(11), 3894 (2020)

[24] Aydoğdu, Ş: Predicting student final performance using artificial neural networks in online learning environments. Educ. Inf. Technol. 25(3), 1913–1927 (2020)

[25] Godfrey, P.; Gryz, J.; Lasek, P. Interactive Visualization of Large Data Sets. IEEE Trans. Knowl. Data Eng. 2016, 28, 2142–2157. [CrossRef]

[26] Pooja and Bhalla R : A Review Paper on the Role of Sentiment Analysis in Quality Education, SN Computer Science (2022) 3:469. https://doi.org/10.1007/s42979-022-01366-9

[27] Bhalla R (2019) A comparative analysis of application of proposed and the existing methodologies on a mobile phone survey.In: International conference on futuristic trends in networks and computing technologies. Springer, Singapore

[28] Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. *IEEE Access, 8*, 203827–203844. https://doi.org/10.1109/access.2020.3036572

[29] Mehta, K.J. Effect of sleep and mood on academic performance—at interface of physiology, psychology, and education. *Humanit Soc Sci Commun* **9**, 16 (2022). https://doi.org/10.1057/s41599-021-01031-1. https://www.nature.com/articles/s41599-021-01031-1

[30] AH, K., Oldayo, A. A., & Fakai, A. A. (2020). Factors and Effects of Poor Background on the Students Academic Performance in Physics at Senior Secondary School in Birnin Kebbi Metropolis. https://www.saudijournals.com/media/articles/SJEAT_53_114-119.pd

[31] Bhalla R., Sharma A., Gupta J.(2022) An Ensemble based Machine Learning model for Accurate Prediction using Multiple Categorical Dataset, IEEE Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), May 2022 https://ieeexplore.ieee.org/document/994124610.1109/ICPS55917.2022.00008

[32] Alani F.S., Hawas A.T.(2021) Factors Affecting Students Academic Performance : A case Study of Sohar Univesity, Psychollogy and Education . Sept 2021. https://www.researchgate.net/publication/354723769