Mini Project Report on

“Students Academic Performance Prediction System”

by

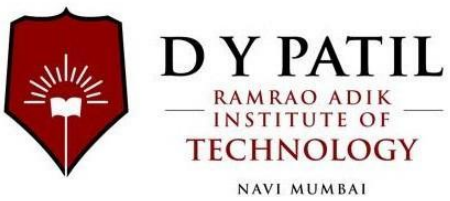
Shubham Patil (19CE5016)

Chetan Khairnar (18CE1034)

Shivam Kendre (18CE1009)

Under the guidance of

Mrs. Rashmi Dhumal



Department of Computer Engineering

Ramrao Adik Institute of Technology

Dr. D. Y. Patil Vidyanagar, Nerul, Navi Mumbai

University of Mumbai

May 2021

Ramrao Adik Institute of Technology

Dr. D. Y. Patil Vidyanagar, Nerul, Navi Mumbai

CERTIFICATE

This is to certify that Mini Project report entitled

"Students Academic Performance Prediction System"

by

Shubham Patil (19CE5016)

Chetan Khairnar (18CE1034)

Shivam Kendre (18CE1009)

is successfully completed for Third Year Computer Engineering as prescribed by University of Mumbai.

Supervisor	Project Coordinator
(Mrs. Rashmi Dhumal)	**(Dr. Bharti Joshi)**
Head of Department	Principal
(Dr. Leena Ragha)	**(Dr. Mukesh D. Patil)**

Mini Project Report Approval

This is to certify that the Mini Project entitled ***" Students Academic Performance Prediction System"*** is a bonafide work done by ***Chetan Khairnar***, ***Shivam Kendre*** and ***Shubham Patil*** under the supervision of ***Mrs. Rashmi Dhumal***. This Mini Project has been approved for Third Year Computer Engineering.

Internal Examiner :

1. . . . . .

2. . . . . .

External Examiners :

Date : . . . /. . . /. . . . . .

Place : . . . . . .

DECLARATION

We declare that this written submission represents our ideas and does not involve plagiarism. We have adequately cited and referenced the original sources wherever others' ideas or words have been included. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will cause disciplinary action against me by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: _____

Shubham Patil (19CE5016)

Chetan Khairnar (18CE1034)

Shivam Kendre (18CE1009)

Contents

Abstract

Self-regulated Student performance prediction is an important job because of the large volume of data generated in educational databases. These methods must be used for knowing students and their learning environment well. The educational institutions are most peculiar about the number of students who will pass/fail for considerable arrangements.

So to solve this problem, we have designed the “Students Academic Performance Prediction System”. Many renowned classification algorithms have been applied in this domain but we have proposed a student performance prediction model based on supervised learning Multiple Linear Regression. In support of it, an ensemble method is applied to ameliorate the performance of the classifier, Also the ensemble methods approach is designed to solve regression, prediction and classification problems

List of Figures

Chapter 1

Introduction

1.1 Overview

The Financial achievement of any country is largely proportional to making secondary and further education more affordable and that concerns most to any government. One of the factors that contribute to the educational expenses is the studying time spent by students to graduate. For example, the number of American students having loan debt has risen because of the failure of many students to graduate at the proper time. Higher education is provided for free to the students in Iraq by the government. Yet, it costs the government additional expenses due to failing graduates.

To elude these expenses, the student graduates must be ensured by the government on time. Hence machine learning techniques should be incorporated to forecast the performance of the students and identify the slow or aberrant students as early as possible so that prior actions can be taken to improve their performances. The most crucial steps to take when using these techniques in choosing the correct attributes or features which are descriptive must used as input to the machine learning algorithm

In today's rapidly changing world it very important to understand how to make students more productive and active during their Education years hence to make it possible for Institutions very important that they understand the student well thereby making it possible to design a curriculum which makes helps every student overcome their fears of a specific subject, therefore a system is required which will be able to predict the student's marks and thereby making it easy for institutions to design syllabus which fitts each students requirements.

And the Linear Regression predicted the most accurate values, hence further going to include more regression and classification algorithms that can give more accurate values. It will not able students in improving the overall performance, but will also help the teacher in getting an idea about the student's needs and help them accordingly

1.2 Objectives

- Data mining techniques based on Machine learning are used to self-regulate the process of using linear regression techniques for student performance prediction.
- The idea behind this analysis is to predict the marks of students by their studying hours.
- How many hours need to do the study to get 99% marks
- If I study x() hours, free time, travel time per day, how much marks I will get in comparison to it.

1.3 Motivation

Numerous Educational Data Mining (EDM) studies have processed data to forge ahead through the extraction of traditional data, consisting of the data preparation process. Diverse stages are used to process the data in order to explicit an absolute factor through which students' module marks are refined during the data preparation stage. **Therefore Observations show that the final marks of students must not be secluded from the environment of the matriculated module's assessment method.**

Hence this makes it necessary to generate a Students marks prediction system, so as to help the institutions to predict the student's marks prior and which will help them in designing a Syllabus that not only focus on exams but takes in many aspects of students and making the learning suitable according to the student's current requirements.

1.4 Organization of report

The details for the organization of the report are as follows:

Chapter 1 includes the introduction about the problem and the method we are trying to employ. The Literature survey which includes the existing systems and the notable works related to the project are included in Chapter 2. The Project system and the methodology used are described in Chapter 3.Then Chapter 4 includes the analysis and the result of the experiment we tried. The conclusion of our project and Outcomes are described in Chapter 5.

Chapter 2

Literature Survey

2.1 Existing Systems

2.1.1 Student Academic Performance Prediction using Supervised Learning Techniques

In this paper it is stated that the class quizzes, assignments, lab exams, mid, and final exams are the factors which affect the student's academic performance. In this paper, Decision tree, Support Vector Machine (SVM), and Naive Bayes machine learning classification algorithms are implemented to predict the student's academic performance.

Also,other factors that affect the student's academic performance are course structure, assignment marks, final exam scores, and extracurricular activities. This dataset for the current study was collected from the website https://www.kaggle.com.

It consists of 480 instances, 16 independent variables, and one dependent variable. The student's academic performance dataset is shown in Table 1. It consists of 305 male and 175 female students.

Parameters used are the following shortcuts are used such as G–(gender), N–(NationalITy), PB-(PlaceofBirth), SID- (StageID), GID- (GradeID), SID-(SectionID), T-(Topic), SE-(Semester), R-(Relation), RH-(Raised Hands), VR- (VisITedResources), AV-(AnnouncementsView), D-(Discussion), MS-(Middle School), Lwl(Lower level), PAS- (ParentAnsweringSurvey), PS-(ParentschoolSatisfaction), SAD-(StudentAbsenceDays).

Eg. *G N PB SI D GI D SI D T S E R R H V R A V D PA S PS SAD Class M KW Kuwa IT Lw l G04 A IT F Fath er 15 16 2 2 0 Yes Good Under7*

Table 4. Confusion Matrix Values for the algorithms

Decision Tree			
Actual Value	Predicted Value		
	Low Level	Middle Level	High Level
Low Level	31	0	21
Middle Level	2	30	4
High Level	10	4	42
SVM			
Actual Value	Predicted Value		
	Low Level	Middle Level	High Level
Low Level	0	0	52
Middle Level	0	0	36
High Level	0	0	56
Navie Bayes			
Actual Value	Predicted Value		
	Low Level	Middle Level	High Level
Low Level	45	1	6
Middle Level	0	28	8
High Level	13	5	38

Table 5. Performance Metric Values for the algorithms

Algorithm	Accuracy	Precision	Recall	F1 score
Decision Tree	0.71	0.74	0.72	0.73
SVM	0.38	0.39	1.0	0.56
Navie Bayes	**0.77**	**0.73**	**0.68**	**0.70**

Fig 2.1 Confusion Matrix Value of Algorithms

Based on confusion matrix, accuracy, precision, recall, and F1 score the performance of an algorithm is evaluated. The classification algorithm is analyzed based on a confusion matrix which has been used to describe the performance of a classification model on a set of test data for which the true values are known .

Accuracy:

Decision tree provides 71% accuracy, Support Vector Machine provides 38% accuracy, and Naïve Bayes gives 77% accuracy. **The Naïve Bayes algorithm gives better results.**

2.1.2 Student Academic Performance Prediction using Machine Learning

This paper has four machine learning techniques that are been used to construct a classifier that predicts the performances of Students enrolled in Computer Science that is offered by Al-Muthanna University (MU), College Of Humanities. The machine learning techniques consisted of Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression. Extra efforts were made to analyze the effects of using the internet and the time spent by students on social networks on the students' performance. The modules constructed are compared with ROC index performances and classification accuracy. The dataset was collected from 2015-2016 using a survey.

The dataset consists of information from 161 students. A 3 layer fully connected feedforward ANN has been used in this research. Network was made using 1input,
1 output and 2 hidden layers.
The input layer has twenty input units, neurons, while the first hidden layer has six hidden units. The Rectifier Linear Unit has been used as the hidden units' activation function.

Model	TP	FP	TN	FN	Precision	Recall	F- Measure	Accuracy	Classification Error	ROC index
ANN	67	18	57	19	79.17	77.92	78.47	77.04	22.96	0.807
DT	67	19	56	19	77.96	77.83	77.88	76.93	23.61	0.762
Logistic Regression	62	17	58	24	79.23	71.91	74.87	74.53	25.47	0.767
Naïve Bayes	55	23	52	31	70.51	64.27	67.21	66.52	33.48	0.697

Table 3- The accuracy and the performance measures for the Models

Fig 2.2 Accuracy and Performance measure for Models

$$Vmax = P(vj)\ vj \in V\ Max \prod P(ai\ i\ |vj)\ \text{s.}$$

The dataset contains 161 student records, 76 male and 85 female. The dataset consisted of 20 attributes. The attributes were classified into five different categories.

Accuracy:

T**he model with the highest accuracy is using the Artificial Neural Network classification technique with an accuracy of 77.04.** In addition, it has the best performance based on the ROC index which equals to 0.807, and the lowest classification error that is equal to 22.96. Naive Bayes model has the least accuracy which equals 66.52 and the highest error of 33.48.

2.2.3 Student grade prediction using neural network

In this paper data of 25 students was taken to build the model
Final year grade was predicted using second year grade which students got
The target variable was final year score and independent variables were s1 – average score in first semester of previous year , s2 – average score in second semester of previous year, Nof- number of failures, Noa – number of absence and gender
They tried to find out the relationship between independent variables and dependant variable
For the type of neural network, a multilayer perceptron was selected.
Due to this, the task is decoded.
They created an artificial neural network to improve the accuracy of the model as compared to the linear regression model as in neural network they backpropagate and update weights.

2.2.4 Student grade prediction using PCN method

In this paper PCN method is used to predict grade of students

PCN are comments, where P is comment before classroom, C is comment during classroom activities, N is comment before next classroom

LDA and PLSA models are employed to grasp students learning attitude and learning situations in this study.

Data were collected from 2 classes class A of 60 students and class B of 63 students

P-comment: I can do that easier I tried to follow the procedure to run program.

N-comment: I will study hard at home to improve my level ▪ Based on comment they categorized into grades

C-comment: I am able to do the exercise without delay

▪ TP – true positive
▪ TN- true negative
▪ FP- false positive
▪ FN- false negative

2.2.5 Predicting Students' Academic Performance Through Supervised Machine Learning

In this research paper published in 2020 Engr. Sana Bhutto, Dr. Qasim Ali Arain Dr. Isma Farah Siddiqui have stated their research on predicting the Performance of Students based on online learning attributes like engrossing in sample tests, referring to the provided materials, student behavior, or interaction with a system like involvement in discussion groups, and also in quizzes plays answering questions.

They collected/pre-processed and transformed data to apply Data Mining Technique.

They used 2 algorithms
a)Logistic Regression
b) Support Vector Machine and they found that
Accuracy:
Logistic regression has **71% accuracy** whereas **SMA had 78% accuracy**

2.2.6 Unsupervised learning based mining of academic data sets for students' performance analysis

In this paper, two Unsupervised machine learning algorithms have been used to analyze the student academic dataset and predict their performance incepted on their grades found during the academic semester by Liana Maria Crivelli, Gabriela Czibula, George Ciubotariu, Mariana Dindelegan of Babes¸-Bolyai University, Cluj-Napoca, Romania.

The machine learning techniques used here include Principal component analysis(PCA) and Relational association rule mining(RAR).
It was found that RARs has an accuracy of 0.714 and an overall error of 0.422.
It was also found that PCA visualization is only for the subset of students with grades higher than or equal to 5. **The precision P computed for this PCA mapping using $\tau = 1$ is 0.786, higher than the precision value 0.722.** It was also observable that when ignored the students with the final grade 4, supporting their previous assumption the PCA shape showed significant changes.

Summary Table

Sr no	Research Paper	Algorithm	Accuracy(%)
1	Student Academic Performance Prediction using Supervised Learning Techniques	Decision Tree	71
2	Student Academic Performance Prediction using Supervised Learning Techniques	SVM	38
3	Student Academic Performance Prediction using Supervised Learning Techniques	Naive Bayes	77
4	Student Academic Performance Prediction using MachineLearning	ANN	67
5	Student Academic Performance Prediction using MachineLearning	Decision Tree	67
6	Student Academic Performance Prediction using MachineLearning	Logistic Regression	62
7	Student Academic Performance Prediction using MachineLearning	Naive Bayes	55
8	Predicting Students' Academic Performance Through Supervised Machine Learning	Logistic Regression	71
9	Predicting Students' Academic Performance Through Supervised Machine Learning	SMA	78
10	Unsupervised learning-based mining of academic data sets for students' performance analysis	RAR	71.4
11	Unsupervised learning-based	PCA	78.6

	mining of academic data sets for students' performance analysis		

2.2 Limitations of Existing System

- Existing system are predicting grades on the basis of the grades which student got in previous year exam
- But apart from previous year marks many other factors affect this like style of studying and attendance.
- Because of this they were not getting much accurate model
- The Dataset Parameters are of varied type, to find the most suitable dataset matching current scenarios is difficult.

Chapter 3

Proposed System

3.1 Problem Statement

The Task of estimating the value of an unknown output variables based on the values of input variable is Prediction. Output variables that refer to students' performance here i.e in Education can be in the form of marks . In this research, we applied linear regression based model to predict the final grade for student . It is difficult for teachers to identify students who are unlikely to pass the final exams during the ongoing academic year. If the faculty gets to know the students who are likely to fail in the final exam it will help them to give more attention towards such students. If these students are not given required attention they will not be able to cope with other students and may lag behind in their academics also creating an impact on their self-esteem.

.

3.2 Proposed Methodology

In this Project We are using Linear Regression Algorithm to predict Marks of next sem, on basics of parameters like student's earlier performance in exam, travel time, study time, health, absency.

Students' performance is evaluated by taking students' marks of previous exams as input. Features and labels are extracted from the dataset after preprocessing it and then we have split the dataset into 2 parts i.e train and test sets then linear regression is applied to the dataset for prediction Linear regression algorithm is one of the most widely used machine learning algorithms. linear regression is based on supervised learning is one of the algorithms that is widely known and it is easily understood even by the person who is not so familiar with machine learning algorithms.Regression is performed in linear regression algorithm as the name suggest.The relationship between two variables is defined in Linear Regression algorithm by fitting regression line to the data.one variables is dependent and other is independent variable dependent variable is dependent on independent variable. One should make sure that there exists a relationship between the dependent and independent variables before modelling. Scatterplot is known the strength of the relationship between the variables . Linear regression line can be represented in the following form of:

y=a*x+b

y-Dependent Variable

a-slope

x-Independent Variable

b-Intercept

With the best fit regression line to the data the error rate between the predicted and true values can be minimized. Linear Regression can be classified into two types i.e Simple linear regression and Multiple linear regression. Simple Linear Regression is regression, in which only one independent variable is used.In Multiple Linear Regression multiple independent variables are used which we are presently using for the thesis.

3.3 Design of the System

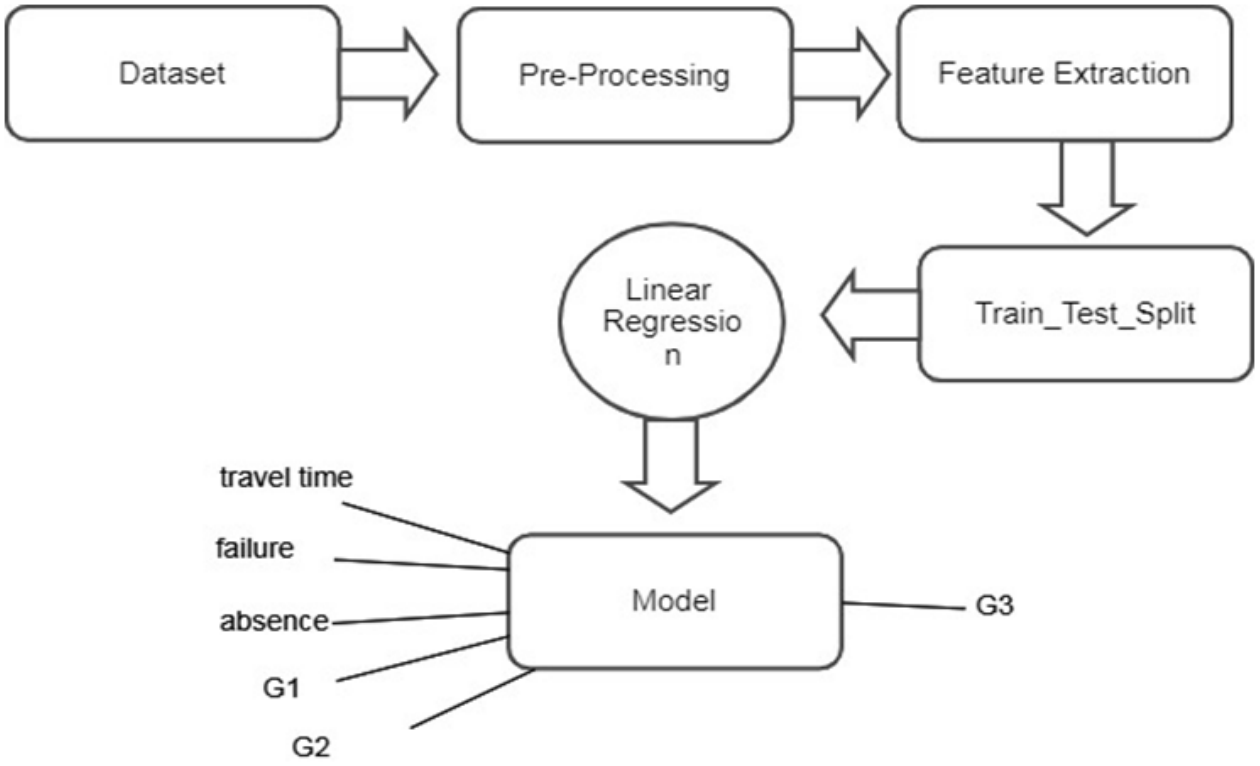


Fig no. 3.1 System Architecture

Here the above figure is of the Architecture of our System which involves basically 4 stages which are

1. Finding Dataset and Processing it:
   Our first aim was to find a Dataset matching our requirements. The dataset can be found based on various platforms like kaggle or can be found in reference paper. After finding our dataset the next task was to process the dataset i.e to transform raw dataset to pure dataset by removing the attributes which consist of null values. After this step we have our dataset ready to be processed.

   2.Feature Extraction:
   In this step we select only the required features from the attributes which are helpful for our system and remove all other features this help us to get a more accurate result.Also outliers are removed at this stage.

   3.Split Dataset:

In this step we split the dataset into 2 parts one to train the dataset and other to test the dataset. We split the dataset in 75%-25% i.e 75% to train the model and 25% to test our model.

4. Apply Linear Regression:
This is the final stage where we provide required features to our model i.e G1, G2, Travel time, failure, absence and then our model gives us required output i.e G3 for each data passed to it.

3.4 Hardware/Software Requirement

Hardware Requirement:

- Minimum i3 processor
- 4GB RAM

Software Requirement:

- Python
- Necessary libraries(Pandas, Sklearn,etc)
- IDE (Jupyter Notebook)

3.5 Implementation Details

Data on student academic performance was collected from a dataset of 20 undergraduate students from kaggle the fig. below shows student demographics

]:

	traveltime	studytime	failures	health	absences	G1	G2	G3
0	2	2	0	3	6	5	6	6
1	1	2	0	3	4	5	5	6
2	1	2	3	3	10	7	8	10
3	1	3	0	5	2	15	14	15
4	1	2	0	5	4	6	10	10
5	1	2	0	5	10	15	15	15
6	1	2	0	3	0	12	12	11
7	2	2	0	1	6	6	5	6
8	1	2	0	1	0	16	18	19
9	1	2	0	5	0	14	15	15

Fig no 3.2 Dataset filtered after preprocessing

We have used Linear regression algorithm to implement marks prediction system which basically finds a linear relationship between independent variable & dependant variable.

```
data.describe()
```

	traveltime	studytime	failures	health	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	1.448101	2.035443	0.334177	3.554430	5.708861	10.908861	10.713924	10.415190
std	0.697505	0.839240	0.743651	1.390303	8.003096	3.319195	3.761505	4.581443
min	1.000000	1.000000	0.000000	1.000000	0.000000	3.000000	0.000000	0.000000
25%	1.000000	1.000000	0.000000	3.000000	0.000000	8.000000	9.000000	8.000000
50%	1.000000	2.000000	0.000000	4.000000	4.000000	11.000000	11.000000	11.000000
75%	2.000000	2.000000	0.000000	5.000000	8.000000	13.000000	13.000000	14.000000
max	4.000000	4.000000	3.000000	5.000000	75.000000	19.000000	19.000000	20.000000

Fig no. 3.3 Descriptive analysis of all attributes of dataset

y= m1x1 + m2x2 + m3x3 + ... + c

where x1,x2,...xn are independent variable

m1,m2,m3, … mn are slopes amd

c is intercept variable

```
[14]: x.head()
```

[14]:

	traveltime	studytime	failures	health	absences	G1	G2
0	2	2	0	3	6	5	6
1	1	2	0	3	4	5	5
2	1	2	3	3	10	7	8
3	1	3	0	5	2	15	14
4	1	2	0	5	4	6	10

```
[15]: y.head()
```

```
[15]: 0     6
      1     6
      2    10
      3    15
      4    10
      Name: G3, dtype: int64
```

Fig no 3.4 Set of independent and dependant variable

If value of mi is greater than xi than it has more effect on y lesser the mi is less affect xi on y. Here job of linear regression algorithm is to find best fit lines for this it has calculate such values of m1,m2,....mn.

Mathematical Calculations

Linear Regression Model

G3 = 0.15(Travel Time) - 0.24(Failure) + 0.10(Health) + 0.031(absence) + 0.14(G1) + 0.98(G2) – 2.37

ROOT MEAN SQUARE ERROR

- Root mean square i.e RMSE can be calculated by finding mean of the squared residuals and identifying residuals and squaring residuals and then calculating the root of the mean squared gives RMSE. By using RMSE we can identify which algorithm gives high accuracy results for the given datasets. RMSE for Linear Regression is 3.6

Fig no. 3.5 RMS

Chapter 4

Results and Discussion

4.1 Result and Analysis

1)Plotted graphs between various factors

we plotted graphs between various features and we found that relation between G1 and G3 is linear also relation between G2 and G3 is linear. G1 and G2 are highly correlated variables whereas studytime, traveltime, absences are also correlated with G3 but not as muc as G1 and G2.

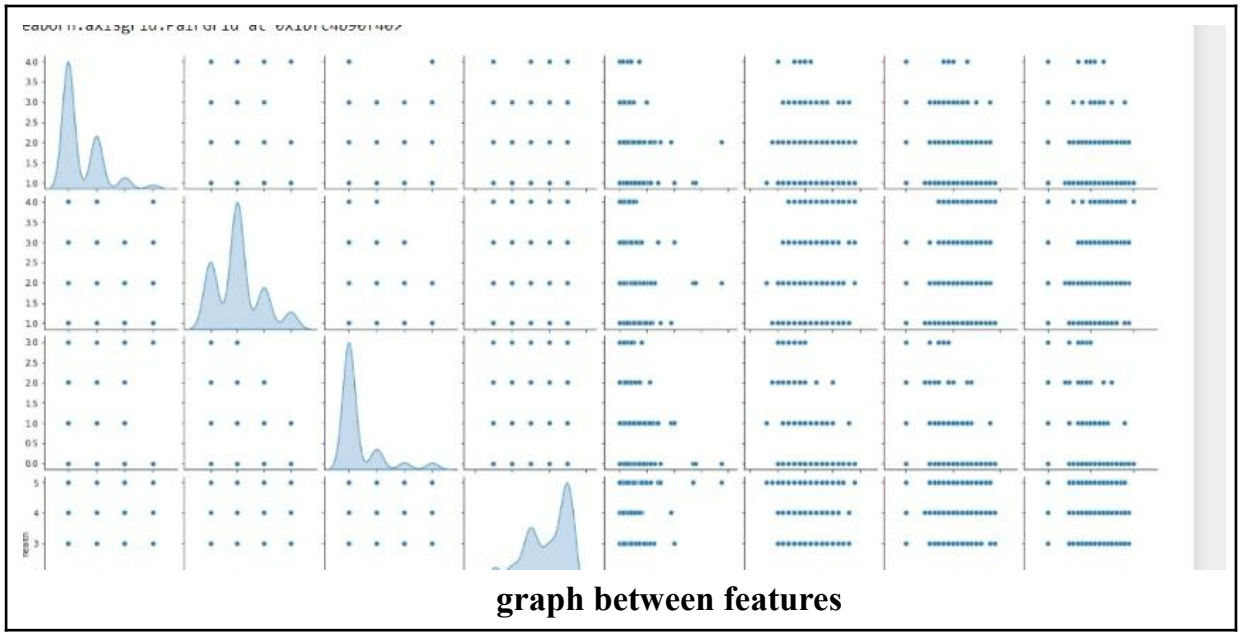
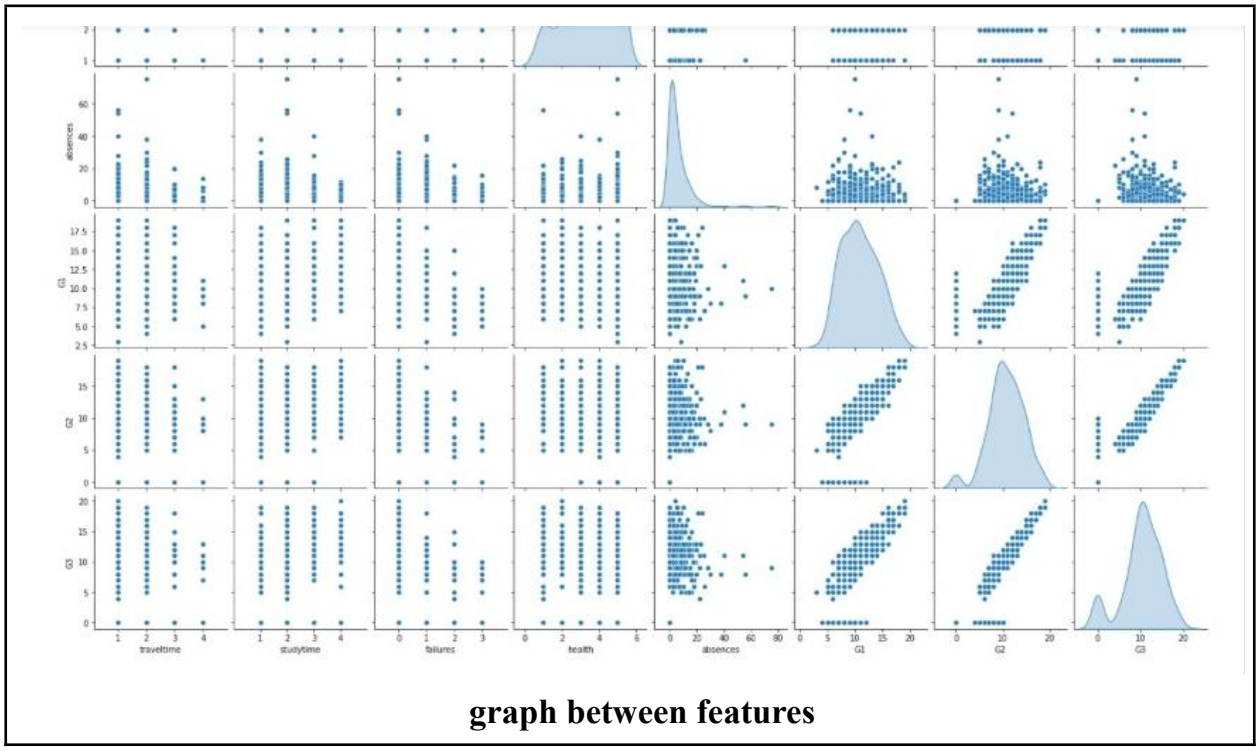


Fig 4.1 Graph between features

2)From these graphs it is clear that g1, g2 affects much on g3. there is linear relation between g1 and g3, g2 and g3



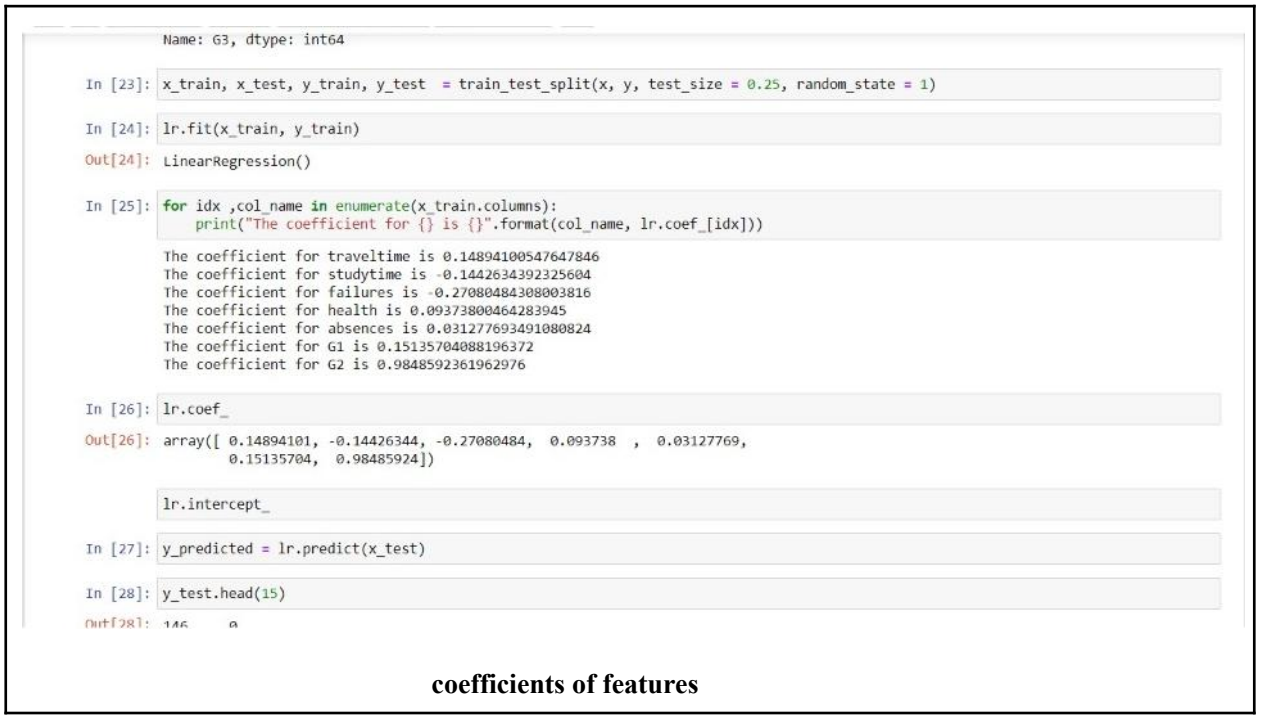
graph between features

Fig 4.2 Graph between features

After applying the algorithm we got the values of m1, m2, m3, m4, m5, m6 and m7 as

0.14894101,
-0.14426344,
-0.27080484,
0.093738 ,
0.03127769,
0.15135704,
0.98485924

3) After training the model



```
Name: G3, dtype: int64

In [23]: x_train, x_test, y_train, y_test  = train_test_split(x, y, test_size = 0.25, random_state = 1)

In [24]: lr.fit(x_train, y_train)

Out[24]: LinearRegression()

In [25]: for idx ,col_name in enumerate(x_train.columns):
             print("The coefficient for {} is {}".format(col_name, lr.coef_[idx]))

         The coefficient for traveltime is 0.14894100547647846
         The coefficient for studytime is -0.1442634392325604
         The coefficient for failures is -0.27080484308003816
         The coefficient for health is 0.09373800464283945
         The coefficient for absences is 0.031277693491080824
         The coefficient for G1 is 0.15135704088196372
         The coefficient for G2 is 0.9848592361962976

In [26]: lr.coef_

Out[26]: array([ 0.14894101, -0.14426344, -0.27080484,  0.093738  ,  0.03127769,
                 0.15135704,  0.98485924])

         lr.intercept_

In [27]: y_predicted = lr.predict(x_test)

In [28]: y_test.head(15)

Out[28]: 146     0
```

coefficients of features

Fig 4.3 coefficients of features

we compared the value of actual y and predicted y
error of this was in range -1.5 to 1.5

4) Mean valued error was calculated as 3.66

mean squared error is mean of square of predicted values and actual values. which is found to be 3.66

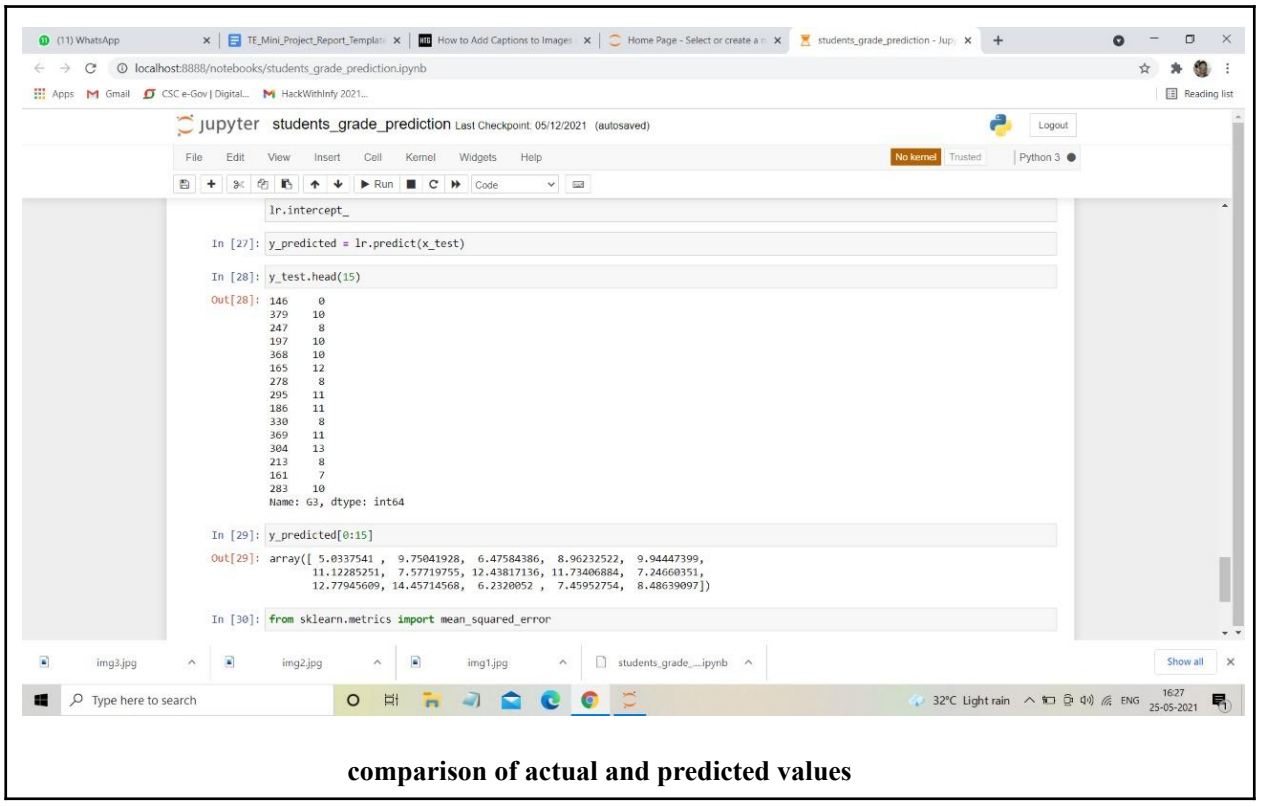


Fig 4.4 comparison of actual and predicted values

Chapter 5

Conclusion and Further Work

5.1 Conclusion

We used a machine learning algorithm to construct a predictive model to predict grade of student in next semester based on grades in previous two semesters. other factors like traveltime, studytime. absence were included in dataset to help predict grades with much accuracy. these features are not mostly used in datasets of other prediction system which makes our system different. mean squared error after applying linear regression algorithm was found to be 3.68 and percentage error was found to be 7.72% which means accuracy of model is 92.28%

5.2 Further Work

Currently we have applied linear regression algorithm. to improve the performance of system

concept of neural network should be applied in which we back propagate and updates weight

References

[1] Virachilai S, Vamshidharreddy. “Student’s Academic Performance Prediction Using Machine Learning Approach.” 2020 International Journal of Advanced Science and Technology,2020.

[2]Liana Maria Crivei; Gabriela Czibula; George Ciubotariu; Mariana Dindelegan “Unsupervised learning based mining of academic datasets for students’ performance analysis” 2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI), 21-23 May 2020.

[3]Engr. Sana Bhutto; Isma Farah Siddiqui; Qasim Ali Arain; Maleeha Anwar “Predicting Students Academic Performance Through Supervised Machine Learning”2020 International Conference on Information Science and Communication Technology (ICISCT), 8-9 Feb. 2020.

[4]Ekaterina E. Vasileva; Daniil S. Kurushin; Sergey S. Vlasov “Early Prediction of the Grade Point Average of University Students Diploma: Neural Network Approach”, 2019 XXII International Conference on Soft Computing and Measurements (SCM)), 23-25 May 2019

[5] K. P. Shaleena and S. Paul, “Data mining techniques for predicting student performance,” in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March.

Acknowledge

We take this opportunity to express our profound gratitude and deep regards to our guide Mrs. Rashmi Dhumal for his/her exemplary guidance, monitoring and constant encouragement throughout the completion of this report. We are truly grateful to his/her efforts to improve our understanding towards various concepts and technical skills required in our project. The blessing, help and guidance given by him/her from time to time shall carry us a long way in the journey of life on which we are about to embark.

Date: _____