Predicting Student Performance Using Regression Analysis

Submitted in partial fulfillment of the requirements of the degree

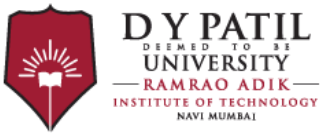
BACHELOR OF ENGINEERING IN **COMPUTER ENGINEERING**

By

Sushree Alina Avipsa Rout	**19CE8013**
Purva Kharsambale	**19CE7003**
Samruddhi Bhosale	**19CE8008**
Chaitali Shetty	**19CE7027**

Name of the Mentor

Dr. Vaibhav Narawade



Department of Computer Engineering

Ramrao Adik Institute of Technology

Sector 7, Phase I, Pad. Dr. D. Y. Patil Vidyapeeth, Nerul, Navi Mumbai, Maharashtra 400706

University of Mumbai

(AY 2021-22)

CERTIFICATE

This is to certify that the Mini Project entitled **"Predicting Student Performance Using Regression Analysis"** is a bonafide work of **Sushree Alina Avipsa Rout (19CE8013), Purva Kharasambale (19CE7003), Samruddhi Bhosale (19CE8008), Chaitali Shetty (19CE7027)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering" .**

(Dr. Vaibhav Narawade)
Mentor

(Dr. Leena Ragha)
Head of Department

(Dr. Mukesh D Patil)
Principal

Mini Project Approval

This Mini Project entitled "**Predicting Student Performance Using Regression Analysis"** by **Sushree Alina Avipsa Rout (19CE8013), Purva Kharsambale (19CE7003), Samruddhi Bhosale (19CE8008), Chaitali Shetty (19CE7027)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1……
(Internal Examiner Name & Sign)

2……
(External Examiner name & Sign)

Contents

ABSTRACT

Prediction of student academic performance helps instructors develop a good understanding of how well or how poorly the students in their classes will perform, so the instructors can take measures to improve the learning of students. Predicting the performance of students becomes more challenging beacuse of the huge amount of databases available. The accomplishment of learning outcomes was measured mainly as ranks and grades. For classifying student performance Regression and supervised machine learning models were frequently employed.

Student online learning activities, term assessment grades, and student academic emotions were the most observable predictors of learning outcomes. We will conclude the survey by highlighting some major research challenges and by suggesting a summary of significant recommendations to motivate future works in this field. Good prediction is defined as the one with the prediction error of ±10%. Results show that developed predictive models have average prediction accuracy of 86.8%-95.7% and they generate good predictions of 44.4%-65.6%.

ACKNOWLEDGEMENT

Success and outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got all of this along with the successful completion of our project. All we have done is only due to such supervision and assistance and we would not forget to thank them. We owe our deep gratitude to our project guide Dr. Vaibhav Narawade, who took keen interest in our project work and guided us all along, until the completion of our project work by providing all the necessary information for developing a good system. We are thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staff of the Computer Department of RAIT who helped us in successfully completing our project work.

Lastly, we would also like to thank our parents and group members who helped us a lot in finalizing this project within the limited time frame.

LIST OF FIGURES

FIGURES **PAGE NO.**

1.Introduction

1.1.Introduction

Prediction of student academic performance is regarded as an important research topic from a long time in many academic disciplines because it has benefits in both teaching and learning. Due to huge amount of data in educational databases, predicting performance of students has become more difficult. Shortage of an established framework for evaluating and tracking success of students is currently not being considered.There are two primary reasons of such kind of occurrence. First is that the research on existing methods of prediction is still insufficient to determine the most appropriate methods for predicting student performance in institutions. Second is that the absence of inquiry of the specific courses.Data mining provides many tasks which are used to predict the student's performance.

Regression task is used to evaluate performance of a student.By using educational machine learning methods, we can potentially improve performance and progress of students in an efficient manner. As a result, it will assist the educators for providing an effective teaching approach. Besides that the educators could also monitor student's achievements. Students could improve their learning activities which will allow the administration to improve the systems performance.

Timely prediction of student performance enables detection of

low performing students which empowers the educators to intervene early during the process of learning and implement the required interventions. Fruitful interventions are included, but they are not limited to, student advising, performance progress monitoring, intelligent tutoring systems development, and policymaking. This endeavor is strongly boosted by computational advances in data mining and learning analytics . A recent comprehensive survey highlights that approximately 70% of the reviewed work investigated student performance prediction using student grades and GPAs, while only 10% of the studies inspected the prediction of student achievement using learning outcomes. This gap incited us to thoroughly investigate the work carried out where the learning outcomes are used as a proxy for student academic performance.

In the model build and training process, a regression algorithm estimates the value of the dependent variable as a function of the predictors in the build data which is based on the independent variables. The relationships between predictors and target are summarized in a model, which are then applied to a different dataset in which target values are unknown. Second approach is finding error rate of regression algorithms by using root mean square error. Obtained results reveal that school background also plays a major role in predicting the grades. Finally, we can identify the students who are at risk and provide better additional training for the students who are weak. Results showed that self-efficacy contributed unique variance in the prediction of student academic performance.

1.2.Motivation

- To develop a validated set of regression models.
- To create a user-friendly web interface.
- To make prediction more efficient.
- To make predictions more accurate.

1.3.Problem Statement & Objectives

In today's environment, there is a lack of quality education, and also competition is increasing day by day. Prediction of student academic performance will help the instructors to develop good understanding of how well or how poor the students in their classes will perform, so instructors can take necessary measures to improve student learning.

Objectives

1. Objective of the present study is to develop validated set of regression models to predict student's academic performance.
2. To develop a regression model to predict a student's performance.
3. To predict the final performance w.r.t. their 3 semester's score.
4. To analyze the factors responsible for scoring high/low marks
5. To create a user-friendly web interface on which the system can be implemented.
6. To make the prediction more efficient and accurate.

2.Literature Survey

2.1.Survey of Existing System/SRS

Student Outcomes

Outcome-based education (OBE) has emerged as a new school of thought in education and has recently enjoyed wide acceptability and adoption . This educational paradigm shifts the focus of the teaching and learning process from the traditional teacher objectives to the so-called student outcomes. The outcomes, representing the targeted competencies, might be defined and measured at the course level, i.e., course outcomes, or program level, i.e., program outcomes. Essentially, course outcomes enable the accomplishment of program outcomes, and their alignment (i.e., courses to program) is performed in a critical activity referred to as curriculum mapping. Computerized tools were developed to assist in realizing the OBE goals and effectively document the educational assessment activities . Their usefulness could be extended by incorporating intelligent models that can prognosticate the attainment of learning outcomes during academic terms.

Prediction in single course :-

Title : Predicting student academic performance in an engineering dynamics course:A comparison of four types of predictive mathematical models

Author : S. Huang, N. Fang
Year : 2013
Description: The work presented in this paper does not predict performance at degree level but at a course level. However, it is interesting as it suggests a kind of upper bound for the accuracy that can be achieved when predicting performance at the end of a degree. They employed four mathematical models namely multiple linear regression, multilayer perception networks, radial basis functions and support vector machines to predict student's academic performance in an engineering dynamics course.It has worked on the data of 323 undergraduate students who took dynamics course at Utah State University in four semesters. Their predictor variables were the students' cumulative GPA; grades earned in four pre-requisite courses i.e. statistics, calculus I, calculus II and physics; and scores on three dynamics midterm examinations. The paper used six combinations of predictor variables to develop a total of 24 predictive mathematical models. For all the four models, they achieved an average prediction accuracy of 81%– 91%. This work shows that previous marks can predict the grade of a course with high accuracy.

<u>Prediction by Using Data Mining Classification :-</u>
Title : Student Performance Prediction by Using Data Mining Classification Algorithms,
International Journal of Computer
Science and Management Research.
Author : DorinaKabakchieva
Year : 4 November 2012

*Description:*The research is focused on the development of data mining models for predicting student
performance, based on their personal, higher secondary and university International Journal of Pure and Applied Mathematics Special Issue performance by using One Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbor classification methods and found that Neural Network model predicts with higher accuracy than other three models. The present study differs from other works in various aspects. First, by using the predicting values that affect the programming skills most are only considered, no socioeconomic data is considered. It also gives some limit to what can be achieved when predicting graduation performance. Indeed the predictors include midterm examinations that can be expected to correlate well with the final exam of the course, more than marks of single courses with the graduation mark.

Factors Affecting Student's Performance

Title: Factors Affecting Student Performance : A Case of Private Colleges Bangladesh e-Journal of Sociology
Author : Hijazi, S. T., and Naqvi, R.S.M.M
Year : 2006.
Description: A study on the student's performance by selecting a sample of 300 students (225males, 75 females) from a group of colleges under Punjab University in Pakistan. The conclusion that was stated as Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' fathers income, mother's age and mother's educational qualification are significantly related to student performance by means of simple linear regression analysis, it was found that the factors like

mother's educational qualification and student's father's income were highly interconnected with the student academic performance. The present study differs from other works in various aspects. First, by using the predicting values that effect the programming skills most are only considered, no socio-economic data is considered.

2.2.Limitation Existing system or Research gap

In existing system i.e. Prediction of student performance using data mining , Indexed Conference(2017); there is lack of unsupervised learning techniques devised to forecast student attainment of the learning outcomes.The existing regression models like data mining,machine learning etc. have less accuracy rate in predicting the student performance. Datasets used in these system were limited to a particular subject or course.Also, the variables used for predicting were limited like only time devoted to study were considered.Therefore existing systems needs an alternative to overcome the above limitations.

.	Focus of Survey and Publication Venue	Models and Approaches Reviewed	Limitations	Strength
1.	Performance prediction using machine learning ; Unindexed Journal (2020)	Machine learning models	→ Did not survey student outcomes	➢ Compared the performance of machine learning models
2.	Features predicting student performance; Indexed Journal (2018)	Data mining techniques, machine learning models	→ Did not discuss student outcomes	➢ Adopted a robust methodology ➢ Highlighted the predictors of student performance ➢ Described the most used prediction methods
3.	Techniques and algorithms used for student performance prediction ; Indexed Conference (2018)	Education analytics	→ Did not assess the quality of the studies → Did not compare the models	➢ Listed the techniques used in the learning analytics for predicting student performance

Figure 1:Limitations & Strength

	Focus of Survey and Publication Venue	Models and Approaches Reviewed	Limitations	Strength
4.	Prediction of student outcome using data mining; Symposium (2018)	Data mining techniques	→ Focused on the Moodle virtual learning system only → Did not address student outcome from different perspectives → Did not discuss the limitations → Did not compare the predictive models	➢ Formulated well defined questions
5.	Performance prediction using data mining techniques; Unindexed Journal (2017)	Data mining techniques	→ Did not survey student outcomes → Reported only five techniques → Did not discuss the limitations – Adopted a weak survey methodology	➢ Discussed the factors predicting student performance

Figure 2:Limitations & Strength

	Focus of Survey and Publication Venue	Models and Approaches Reviewed	Limitations	Strength
6.	Preliminary results of predictive learning analytics ; Indexed Conference (2017)	Machine learning models	→ Presented a preliminary study lacking crucial details → Did not publish the full results of the survey	➢ Reported the overall context of student outcomes
7.	Prediction of student performance using data mining ; Indexed Conference (2015)	Data mining techniques	→ Did not forecast student outcomes → Reviewed a small number of papers → Did not discuss the limitations	➢ Identified attributes and methods that predict student performance ➢ Applied the PICO methodology

	Focus of Survey and Publication Venue	Models and Approaches Reviewed	Limitations	Strength
8.	Data mining techniques to discover knowledge in education; Indexed Journal	Educational data mining	→ Did not focus on student outcomes	➢ Covered different areas such as student performance

Figure 4:Limitations & Strength

2.3.Mini Project Contribution

All of us have equally designed, planned and contributed towards the completion of the successful and developed project. We planned and discussed every detail of the project by holding meetings through the internet and surveying various systems.

Sushree Alina has worked on the research and development of the prediction model along with report making.

Purva Kharsambale has worked on research and designing of the project along with making reports.

Samruddhi Bhosale has worked on the designing and layout of the model with report making.

Chaitali Shetty has worked on the development and research part of the model along with report making.

3.Proposed System (eg New Approach of Data Summarization)

3.1.Introduction

To achieve project's aims, quantitative simulation research methods were conducted as suggested in the framework phases. In these phases dataset will be prepared to be passed through visualization and clustering techniques, i.e. like heat map and hierarchical clustering, to extract the top correlated indicators. Then, the indicators will be used in different classification algorithms and the most accurate model will be chosen for predicting student performance in dissertation projects and all course grades. In between, and before the classification models' evaluation phase, the datasets will pass through a pre-processing (cleansing, missing data imputation) stage to make it ready for the analysis phase. That will be more detailed in the following sections.

Regression is the most often used technique for student performance prediction. Researchers usually examined study-related (SR) data. Our study-related data contained attributes such as higher secondary school background, the medium of study, syllabus, mathematics and English marks. We built a multi-linear regression for a programming course based on the training set and evaluated the results using various models(SVM, RF, Decision tree).The method that achieved best results was subsequently validated on the test set.

The regression is a commonly used technique for student grade prediction. SVM Reg., Random Forest, Decision Tree are used to validate the models. The baseline model predicts the programming marks of the training set of a given instance of

attributes. In addition to accurately predicting student‟s performance,the multi-regression model can be used to analyze how the different features contribute to the predicted grades and thus gain some insights about the student‟s performance. For a proper analysis of the estimated model parameters, it is more convenient that all the attributes have non-negative values which will make all the model‟s components contribute additively to the predicted grades.

3.2.Architecture/ Framework

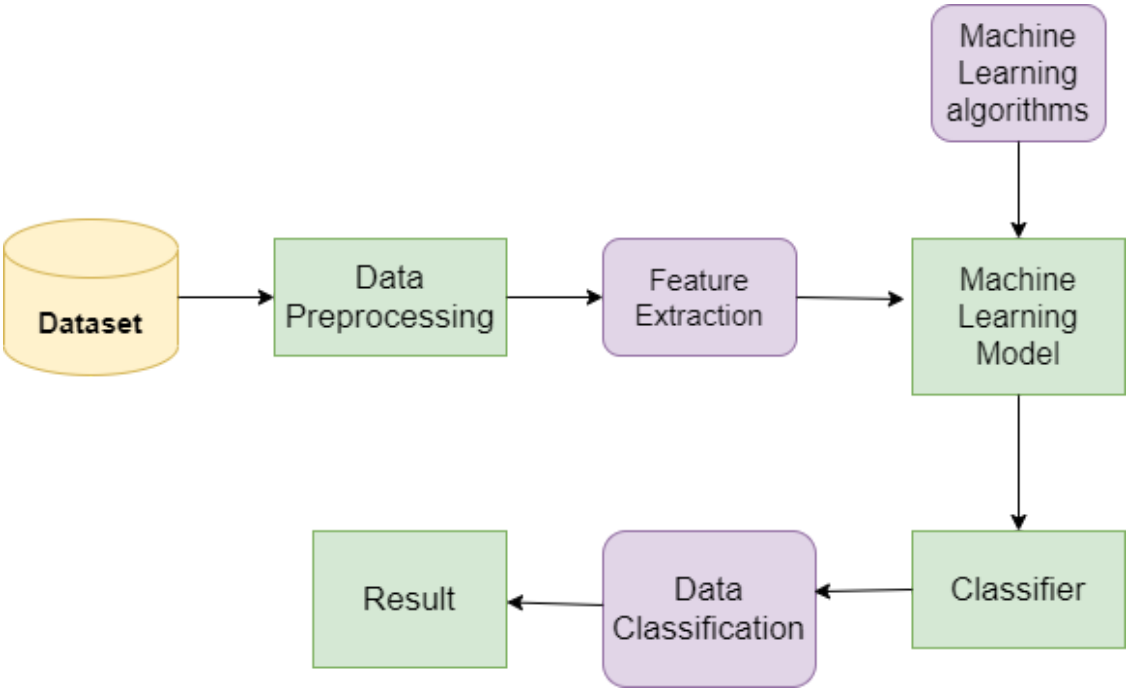


Figure 5:Architecture

1. Dataset: Dataset includes the following attributes like school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime,

failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3.

2. Data Preprocessing: Initially, datasets contained valueless attributes, missing instances, inadequate attributes, data types and other problems that raise the necessity of preparing it first before feeding it to the analysis phase.

3. Feature Extraction: Feature extraction is a part of the dimensionality reduction process, in which an initial set of the raw data is divided and reduced to more manageable groups. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process them. So Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with accuracy and originality.

4. Machine Learning algorithms: Here Logistic regression algorithm is used. Logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college.The resulting analytical model can take

into consideration multiple input criteria. In the case of college acceptance, the model could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

5. Machine Learning Model: Machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

6. Classifier: A classifier is a type of machine learning algorithm used to assign a class label to a data input. ... Classifier algorithms are trained using labeled data; in the image recognition example, for instance, the classifier receives training data that labels images.

7. Data Classification: Data classification is the process of analyzing structured or unstructured data and organizing it into categories based on file type, contents, and other metadata. Data classification helps organizations answer important questions about their data that inform how they mitigate risk and manage data governance policies.

8. Result: The final result will be the predicted model with high accuracy which will predict the performance of the

students.

3.3.Algorithm and Process Design

The proposed model is as follows:

1. Collecting data of student academic performance in 3 semesters.
2. Splitting the data of the first semester into a training and testing dataset.
3. Building regression models to analyse student data in a training set.
4. Testing each regression model using the corresponding testing dataset.
5. Apply the regression model developed in semester 1 to the datasets collected in semester 2 and 3 to determine the prediction of each model.

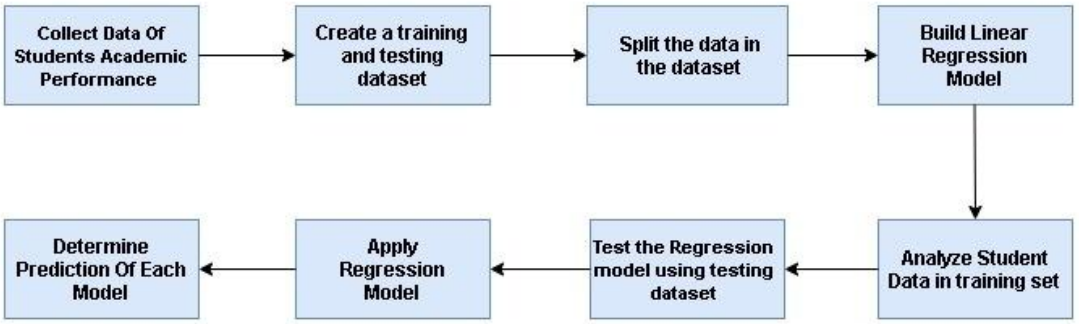


Figure 6:Proposed model Flowchart

3.4.Details of Hardware & Software

Hardware and Software Requirements:

For running this website, the user must have a processor with a

minimum capacity of 1GHz or more.

Hard Drive space should be more than 32 GB.

Recommended space is 64 GB or more.

Memory (RAM): Minimum 1 GB; Recommended 4 GB or above.

Technologies Used**:**

Language: Python

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language.

It supports functional and structured programming methods as well as OOP.

It can be used as a scripting language or can be compiled to byte-code for building large applications.

It provides very high-level dynamic data types and supports dynamic type checking.

IDE-Jupyter Notebook

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R.

3.5.Experiment and Results for Validation and Verification

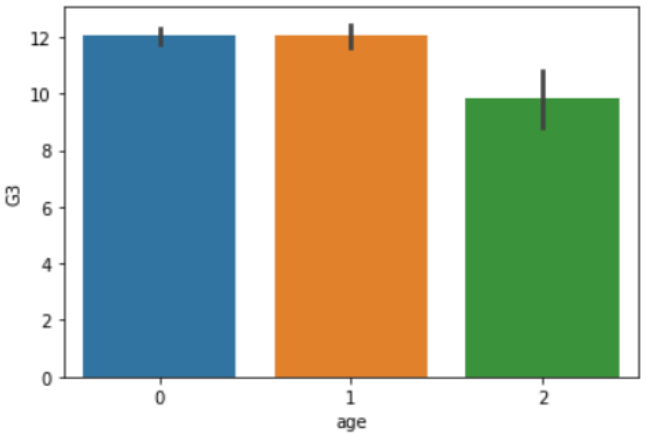
```
df_por.head()
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13

Figure 7 : First 5 rows of dataset

```
sns.barplot(x='age',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='age', ylabel='G3'>
```

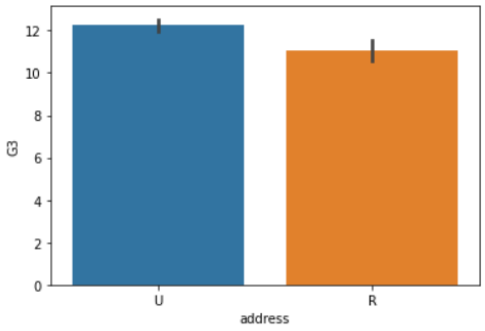


Students of age (19-20) mostly have a score around 10.0

Figure 8 : Barplot for ‘age’

```
sns.barplot(x='address',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='address', ylabel='G3'>
```



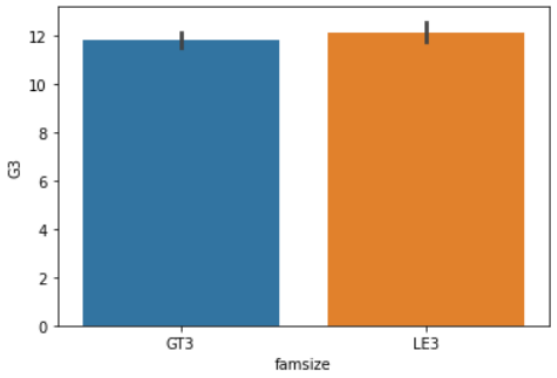
Urban students are distriuted widely in terms of scores, whereas rural students are clustered between 7 and 17

Figure 9: Barplot for ‘address’

```
sns.barplot('famsize','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
s: x, y. From version 0.12, the only valid positional argument will be `data`,
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='famsize', ylabel='G3'>
```

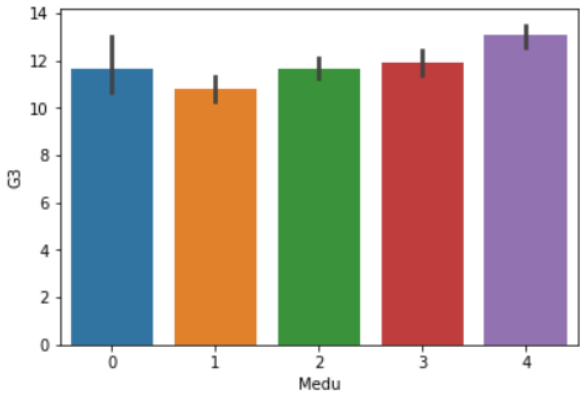


Since the family size of the students doesn't tell much about their scores, we'll drop this attribute.

Figure 10: Barplot for 'famsize'

```
sns.barplot(x='Medu',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='Medu', ylabel='G3'>
```

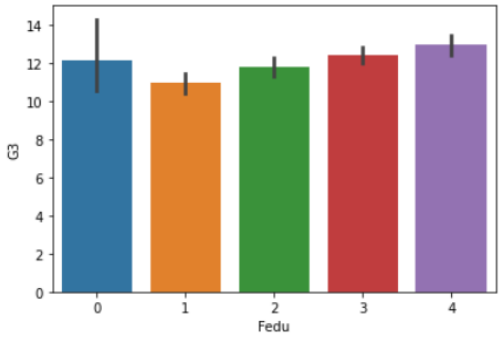


Students whose mother have completed higher education tend to score more than others

Figure 11 : Barplot for 'Medu'

```
sns.barplot(x='Fedu',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='Fedu', ylabel='G3'>
```

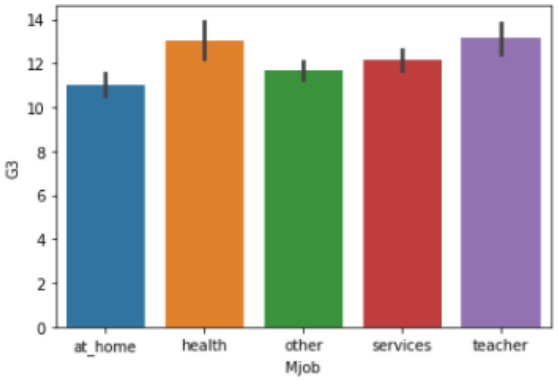


Students whose fathers have completed higher education have slightly better chance of performing well in exams.

Figure 12: Barplot for 'Fedu'

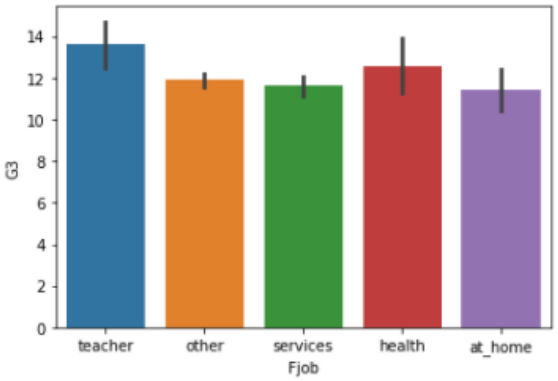
```
sns.barplot(x='Mjob',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='Mjob', ylabel='G3'>
```




```
sns.barplot(x='Fjob',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='Fjob', ylabel='G3'>
```



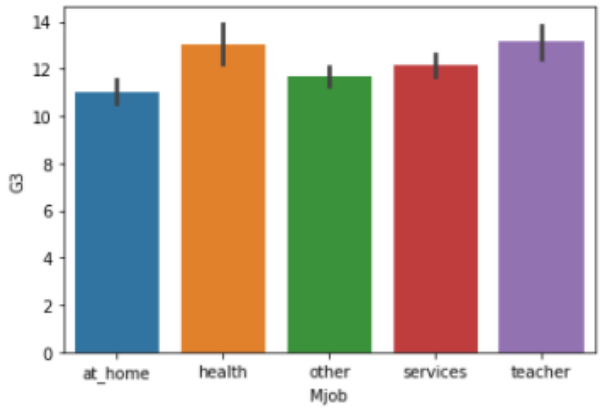
Students whose parents are teachers ,score better in exams.

Students who choose their school by looking at it's reputation ,tend to score better

Figure 13:Barplots for 'Mjob' and 'Fjob'

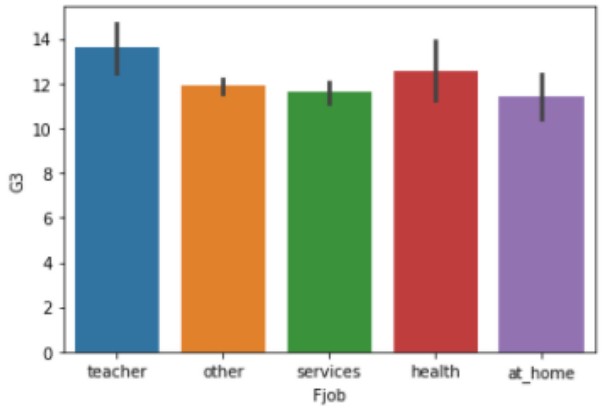
```
sns.barplot(x='Mjob',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='Mjob', ylabel='G3'>
```




```
sns.barplot(x='Fjob',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='Fjob', ylabel='G3'>
```



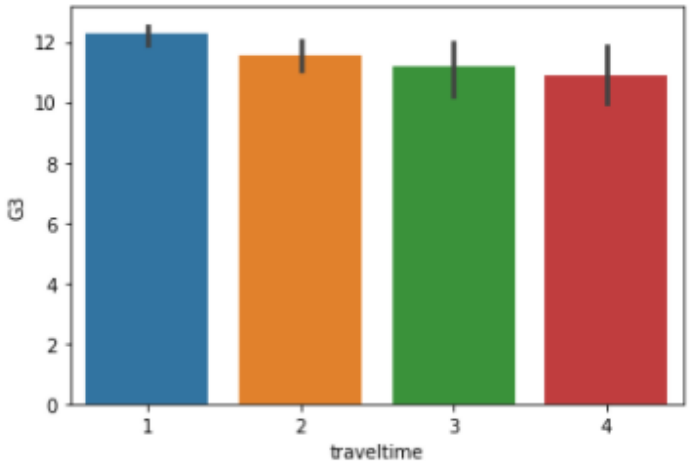
Students whose parents are teachers ,score better in exams.

Students who choose their school by looking at it's reputation ,tend to score better

Figure 14: Barplot for 'Mjob' and 'Fjob'

```
sns.barplot(x='traveltime',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='traveltime', ylabel='G3'>
```

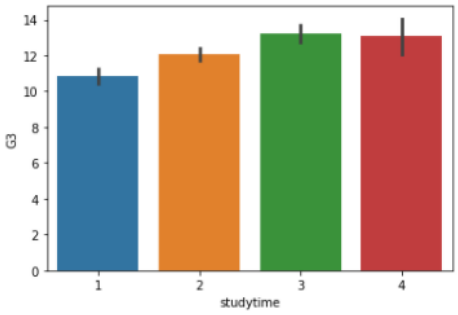


Travel time to school doesn't affect much

Figure 15: Barplot for 'traveltime'

```
sns.barplot(x='studytime',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='studytime', ylabel='G3'>
```

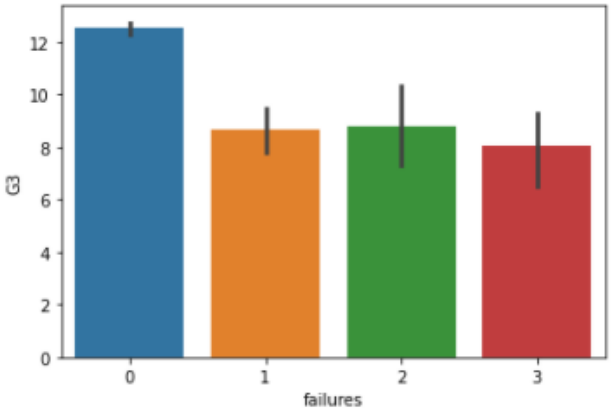


It's pretty clear that if you study for more time,your result will improve. But we see a little downfall in students who study more than 10 hours , maybe because they are not resting.

Figure 16: Barplot for 'studytime'

```
sns.barplot(x='failures',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='failures', ylabel='G3'>
```



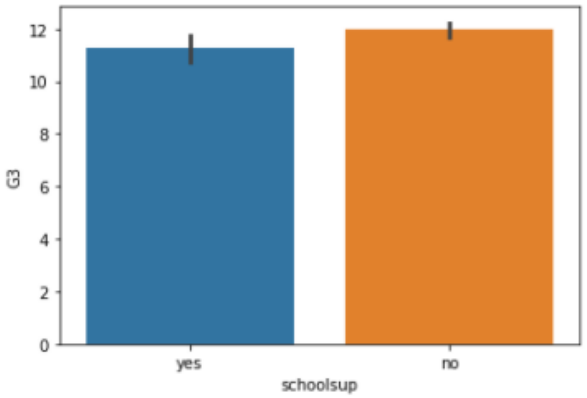
Student's who do not have any past failures record score more.

Figure 17: Barplot for 'Failures'

```
sns.barplot('schoolsup','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decorators.p
s: x, y. From version 0.12, the only valid positional argument
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='schoolsup', ylabel='G3'>
```



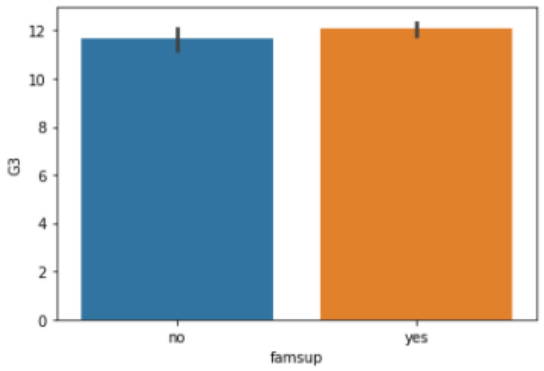
Student's with school support have much higher chance of passing the exam.

Figure 18: Barplot for 'schoolsup'

```
sns.barplot('famsup','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the
s: x, y. From version 0.12, the only valid positional argument will be `data`, and passing
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='famsup', ylabel='G3'>
```



Suprisingly there's almost no difference in student's grades whether their family supports them or not in education.

Figure 19: Barplot for 'famsup'

```
sns.barplot(x='paid',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='paid', ylabel='G3'>
```

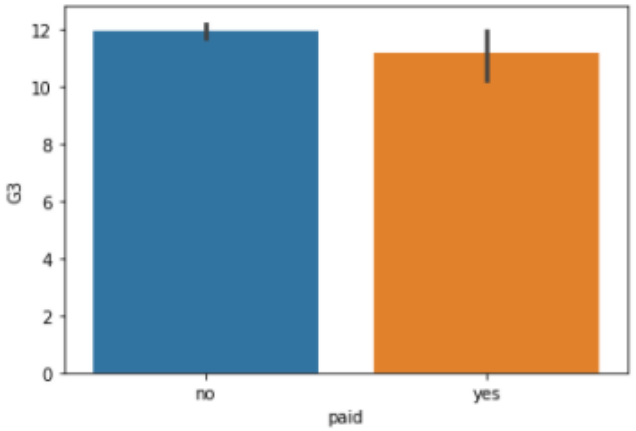


Figure 20: Barplot for 'paid'

```
sns.barplot('activities','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_d
s: x, y. From version 0.12, the only valid positiona
keyword will result in an error or misinterpretation
  warnings.warn(
```

```
<AxesSubplot:xlabel='activities', ylabel='G3'>
```

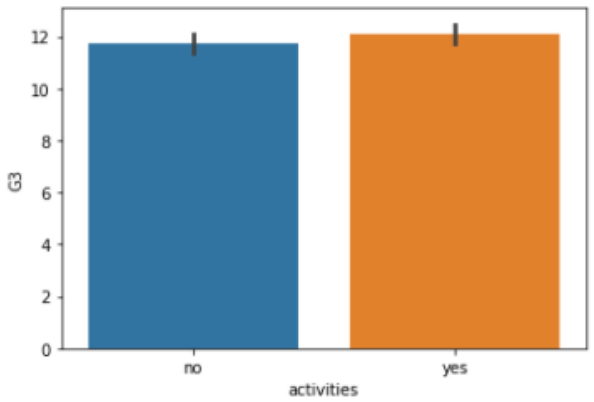
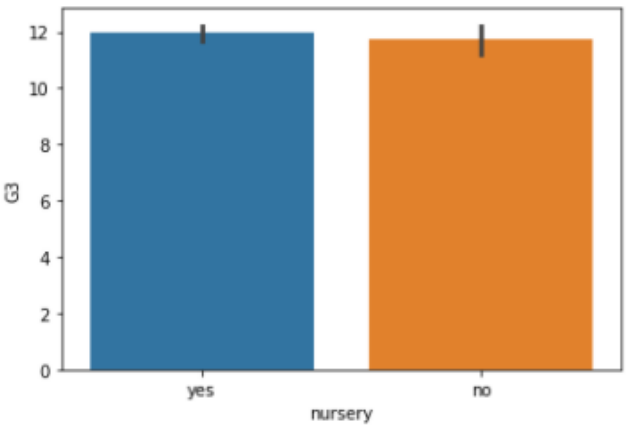


Figure 21: Barplot for 'Activities'

```
sns.barplot('nursery','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_d
s: x, y. From version 0.12, the only valid positiona
keyword will result in an error or misinterpretation
  warnings.warn(
```

```
<AxesSubplot:xlabel='nursery', ylabel='G3'>
```



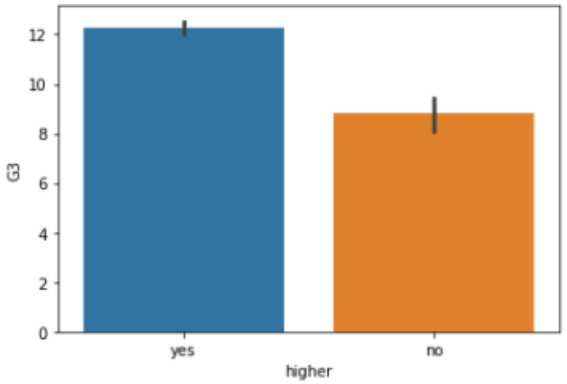
Since there's not much difference in scores. We'll drop this also.

Figure 22: Barplot for 'nursery'

```
sns.barplot('higher','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decorators.py
s: x, y. From version 0.12, the only valid positional argument v
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='higher', ylabel='G3'>
```

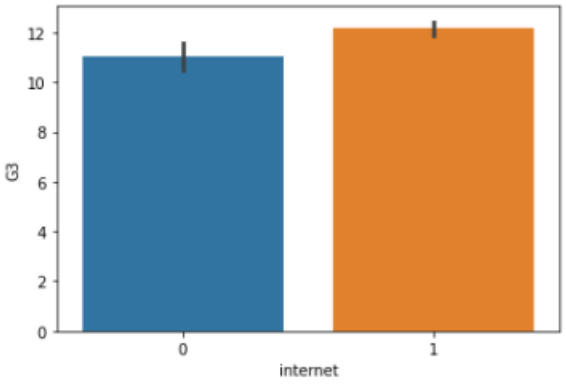


The students who don't want to take higher studies ,have a poor performance .

Figure 23: Barplot for 'higher'

```
sns.barplot(x='internet',y='G3',data=df_por)
```

```
<AxesSubplot:xlabel='internet', ylabel='G3'>
```



Students who do not have internet access at home tend to score average marks.

Figure 24: Barplot for 'Internet'

```
sns.barplot('freetime','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_d
s: x, y. From version 0.12, the only valid positiona
keyword will result in an error or misinterpretation
  warnings.warn(
```

```
<AxesSubplot:xlabel='freetime', ylabel='G3'>
```

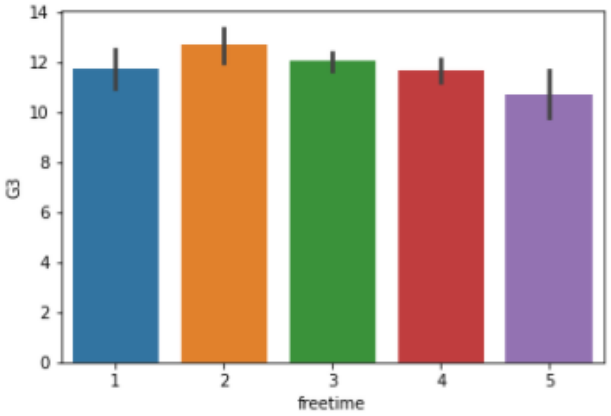
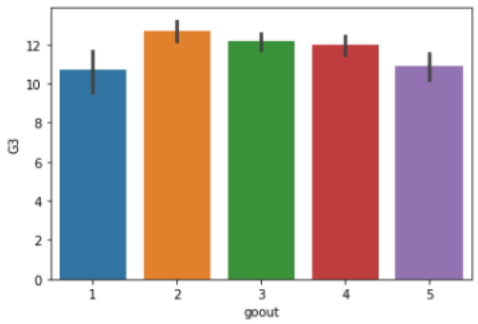


Figure 25: Barplot for 'freetime'

```
sns.barplot('goout','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following
s: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other argum
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='goout', ylabel='G3'>
```



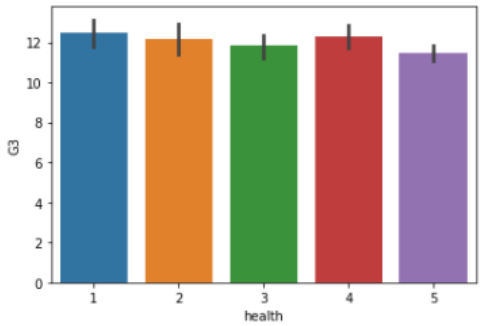
Going out with friends doesn't affect the performance much , but the student's go out very frequently ,tend to score less marks.

Figure 26: Barplot for 'goout'

```
sns.barplot('health','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decorator
s: x, y. From version 0.12, the only valid positional argume
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='health', ylabel='G3'>
```



There's not much difference in the scores with respect to student's health

Figure 27: Barplot for 'health'

```
sns.jointplot(x='absences',y='G3',data=df_por,kind="reg")
```

```
<seaborn.axisgrid.JointGrid at 0x29e717d08e0>
```

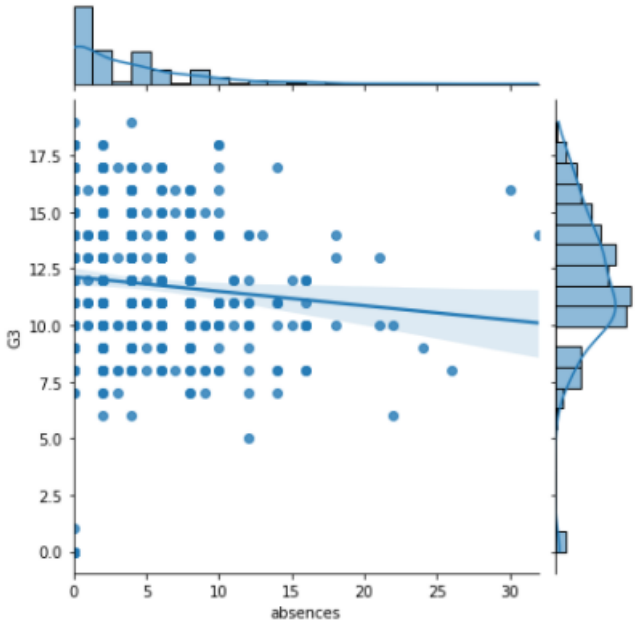
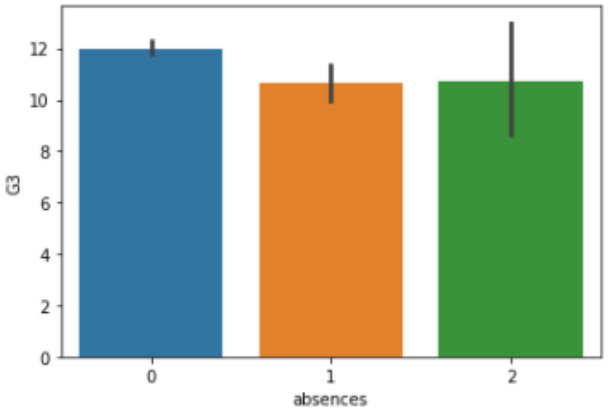


Figure 28: Jointplot for 'absences'

```
sns.barplot('absences','G3',data=df_por)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\_decora
s: x, y. From version 0.12, the only valid positional arg
keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='absences', ylabel='G3'>
```



With increase in absences , the performance of the students decrease

Figure 29: Jointplot for ‘Activities’

```
df_por.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 35 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   age                649 non-null    int64
 1   address            649 non-null    uint8
 2   Medu               649 non-null    int64
 3   Fedu               649 non-null    int64
 4   studytime          649 non-null    int64
 5   failures           649 non-null    int64
 6   schoolsup          649 non-null    uint8
 7   higher             649 non-null    uint8
 8   internet           649 non-null    uint8
 9   famrel             649 non-null    int64
 10  goout              649 non-null    int64
 11  Dalc               649 non-null    int64
 12  Walc               649 non-null    int64
 13  health             649 non-null    int64
 14  absences           649 non-null    int64
 15  G1                 649 non-null    int64
 16  G2                 649 non-null    int64
 17  G3                 649 non-null    int64
 18  school_MS          649 non-null    uint8
 19  sex_M              649 non-null    uint8
 20  Pstatus_T          649 non-null    uint8
 21  Mjob_health        649 non-null    uint8
 22  Mjob_other         649 non-null    uint8
 23  Mjob_services      649 non-null    uint8
 24  Mjob_teacher       649 non-null    uint8
 25  Fjob_health        649 non-null    uint8
 26  Fjob_other         649 non-null    uint8
 27  Fjob_services      649 non-null    uint8
 28  Fjob_teacher       649 non-null    uint8
 29  reason_home        649 non-null    uint8
 30  reason_other       649 non-null    uint8
 31  reason_reputation  649 non-null    uint8
 32  guardian_mother    649 non-null    uint8
 33  guardian_other     649 non-null    uint8
 34  romantic_yes       649 non-null    uint8
dtypes: int64(14), uint8(21)
memory usage: 84.4 KB
```

Figure 30: Info on attributes after dropping columns which are not required for prediction

```
predictions = model.predict(X_test)
predictions
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1], dtype=int64)
```

Figure 31: Prediction of the model

```
print(abs(predictions-y_test).mean())
```

```
0.05116279069767442
```

```
print('Accuracy : {:.2f}'.format(model.score(X_test,y_test)))
```

```
Accuracy : 0.95
```

Figure 32: Accuracy of the model

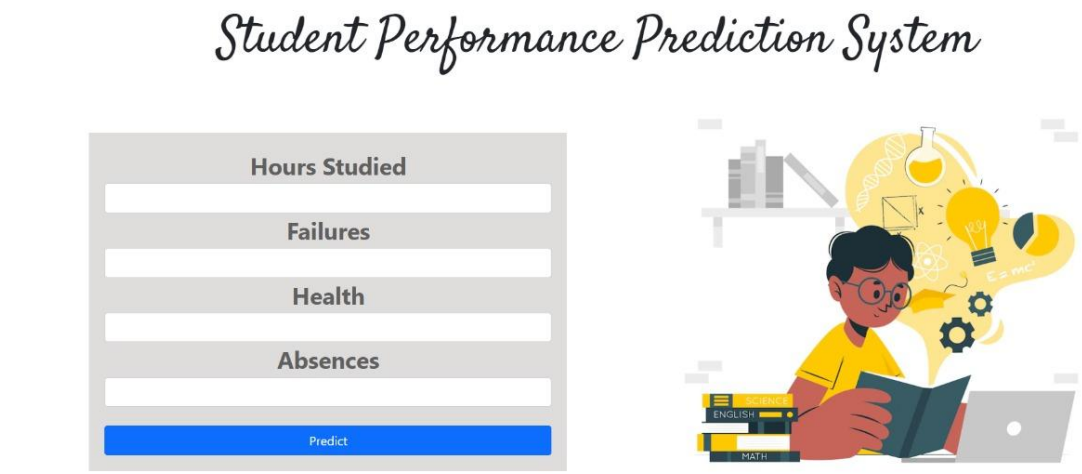


Figure 33: Web-Based Student Prediction

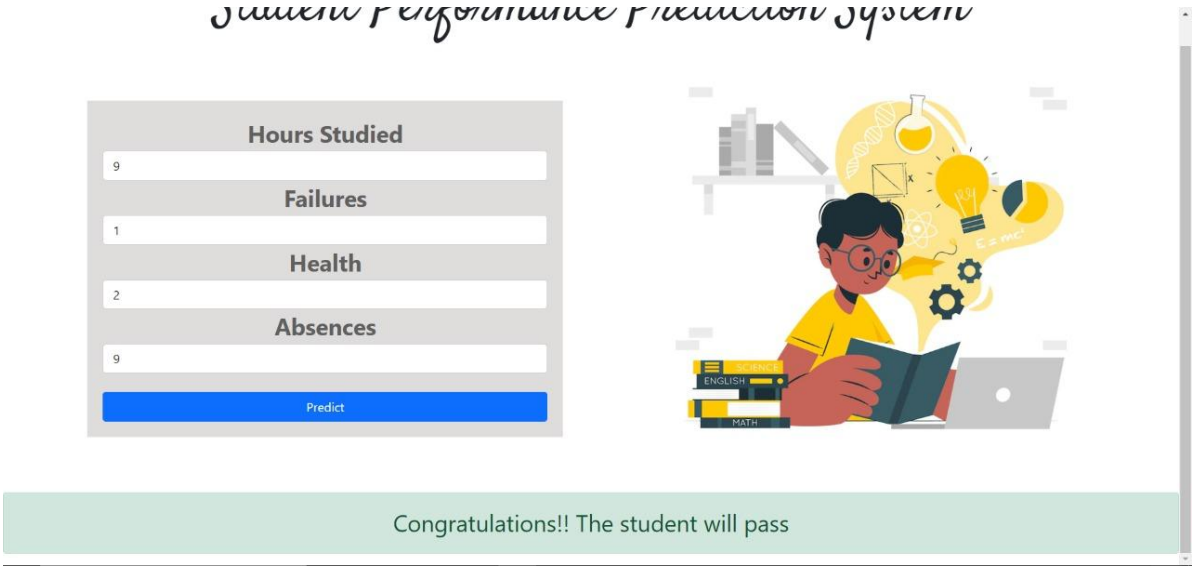


Figure 34: Input the values to predict the final outcome of the student

3.6.Analysis

There exists a variety in semester to semester student body and the classroom composition. The reliability of the predictive models should be tested in different semesters.

We have created a web-app system

3.7.Conclusion

Prediction of student academic performance helps instructors develop a good understanding of how well or how poorly the students in their classes will perform, so instructors can take proactive measures to improve student learning. Predicting student‟s performance would boost the results of student‟s grades and give teachers a better approach for teaching the students who are at risk of failure. Regression models,tree based models and created to make the best predictions with high accuracy. The basic idea is to increase the efficiency of the prediction results using various algorithms. Thus by finding the RMSE, we observed that SVM,linear model gives optimum results.

We have deployed the predicted model in the form of a web-app system so that the prediction of whether a student will be able to clear his /her exams based on the performance can be calculated.

References

[1] G.Sujatha,Sindhu.S and Savaridassan.P,PREDICTING STUDENTS PERFORMANCE USING PERSONALIZED ANALYTICS,Volume 119 No. 12 2018

URL:https://www.acadpubl.eu/hub/2018-119-12/articles/7/1592.pdf

[2] J. Dhilipan, N.Vijayalakshmi , S.Suriya , Arockiya Christopher,Prediction of Students Performance using Machine learning,IOP Conf. Series: Materials Science and Engineering,IVC RAISE 2020

URL:https://iopscience.iop.org/article/10.1088/1757-899X/1055/1/012122/pdf

[3] Arteep Kumar,K.K. Eldhose,R. Sridharan and Vinay V. Panicker,Students' Academic Performance Prediction using Regression, 2020 International Conference on System, Computation, Automation and Networking (ICSCAN)

URL:https://ieeexplore.ieee.org/abstract/document/9262346

[4] Shaobo Huang,Ning Fang,Regression Models of Predicting Student Academic Performance in an Engineering Dynamics Course,AC 2010-190,American Society for Engineering Education, 2010

URL:https://peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course.pdf

[5] Abdallah Namoun and Abdullah Alshanqiti,Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review,Appl. Sci. 2021, 11, 237

URL:https://www.mdpi.com/2076-3417/11/1/237/pdf

[6] Amirah Mohamed Shahiria, Wahidah Husaina , Nur'aini Abdul Rashid,A Review on Predicting Student's Performance using Data Mining Techniques,2015 The Authors,The Third Information Systems International Conference,Procedia Computer Science 72 (2015) 414 – 422

URL:https://www.sciencedirect.com/science/article/pii/S1877050915036182

[7] R.R.Rajalaxmi , P.Natesan , N.Krishnamoorthy ,S.Ponni International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-7 Issue-6S3 April, 2019

URL:https://www.researchgate.net/publication/333102911_Regression_Model_for_Predicting_Engineering_Students_Academic_Performance

[8] Lubna Mahmoud Abu Zohair International Journal of Educational Technology in Higher Education volume 16, Article number: 27 (2019)
URL:https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-019-0160-3

[9] Shakir Khan ,IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.2, February 2021
URL:http://paper.ijcsns.org/07_book/202102/20210221.pdf