

**MINI PROJECT REPORT**

**ON**

**STUDENT GRADE PREDICATION MODEL**

Submitted for the partial fulfillment of award of

# BACHELOR OF TECHNOLOGY

**In**

**Computer Science & Engineering**

**By**

Shubham Vishwakarma (2202221649006)

Aditya Yadav (2202221649002)

Afzal Ansari (210222164008)

**under the Guidance of**

Dr. Vrinda Sachdeva



**I.T.S ENGINEERING COLLEGE**

**GREATER NOIDA**

# Declaration by the Students

This is to certify that the Mini Project Report on “STUDENT GRADE PREDICATION MODEL” by Shubham Vishwakarma , Aditya Yadav and Afzal Ansari have been submitted for the partial fulfilment of the requirements of B.Tech. in Artificial Intelligence and Machine Learning (AIML). The report is our own work. Also, we certify that it is our original work and free from any plagiarism.

Shubham Vishwakarma (2202221649006)

Aditya Yadav (2202221649002)

Afzal Ansari (2102221640008)

# Certificate

This is to certify that the Mini Project Report on “STUDENT GRADE PREDICATION MODEL” by Shubham Vishwakarma , Aditya Yadav and Afzal Ansari have been submitted for the partial fulfilment of the requirements of B.Tech. in Artificial Intelligence and Machine Learning (AIML). The work is carried out under my supervision and free from plagiarism.

Dr. Vrinda Sachdeva

Associate Professor-CSE

(Supervisor)

**2023-2024**

# CERTIFICATE

This is to certify that the project titled “Student Grade Prediction Model” is the bonafide work carried out by Afzal Ansari, Shubham Vishwakarma, Aditya Yadav of B.Tech (CSE) of ITS Engineering College (Greater Noida) affiliated to AKTU University, Lucknow, U.P.(India) in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology (Computer Science and Engineering ) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

|  |  |
| --- | --- |
| **Project Guide** | **Project Faculty** |
| Dr. Vrinda Sachdeva | Bhupendra Bhadana |
| Associate Professor | Assistant Professor |
| Department of CSE | Department of CSE |
| ITSEC, Greater Noida | ITSEC, Greater Noida |

# ABSTRACT

**Prediction of student performance has become an essential issue for improving the educational system. However, this has turned to be a challenging task due to the huge quantity of data in the educational environment. Educational data mining is an emerging field that aims to develop techniques to manipulate and explore the sizable educational data. Classification is one of the primary approaches of the educational data mining methods that is the most widely used for predicting student performance and characteristics. In this work, three linear classification techniques; logistic regression, support vector machines (SVM), and stochastic gradient descent (SGD), and three nonlinear classification methods; decision tree, random forest and adaptive boosting (AdaBoost) are explored and evaluated on a dataset of Assistment system. A k-fold cross validation method is used to evaluate the implemented techniques. The results demonstrate that decision tree algorithm outperforms the other techniques, with an average accuracy of 0.7254, an average sensitivity of 0.8036 and an average specificity of 0.901. Furthermore, the importance of the utilized features is obtained and the system performance is computed using the most significant features. The results reveal that the best performance is reached using the first 80 important features with accuracy, sensitivity and specificity of 0.7252, 0.8042 and 0.9016, respectively.**

# TABLE OF CONTENT

**S. No. Chapter**

1. Introduction
   1. Problem Statement
2. Dataset Description
   1. Details about the dataset obtained from Kaggle.com
   2. Data Points and Parameters
3. Methodology
   1. Workflow
   2. Libraries used
   3. Data Collection and Processing
   4. Machine learning algorithms and techniques
   5. Model Training
4. Source Code
   1. Importing the Libraries
   2. Data Collection and Processing
   3. Splitting the Features and Target
   4. Splitting the Data into Training And Testing Data
   5. Model Training
   6. Model Evaluation / Testing:
   7. Building A Predictive System:
5. observations
   1. Do urban student perform better than rural students?
   2. previous failure vs final grades.
   3. go out vs final grade.
   4. Results
6. Conclusion

a. Summary of findings

* + 1. Reflection on Model Performance
    2. Applications to predict student grades
    3. Limitations and Future Directions

1. future use

## 1. INTRODUCTION

A student grade prediction model is a data-driven approach that utilizes various machine learning algorithms and statistical techniques to forecast or estimate the academic performance of students. By analyzing historical academic data, such as previous grades, attendance records, test scores, demographics, and other relevant factors, these models aim to predict future academic outcomes, such as final exam scores or overall grade point average (GPA). The development of such models involves collecting and preprocessing relevant data, selecting appropriate features, and training predictive algorithms. Machine learning techniques like regression, classification, and ensemble methods are commonly employed to build these models. These models can offer insights into factors that significantly influence student performance and help educators identify students who might be at risk of academic underachievement. Additionally, student grade prediction models have practical applications in educational institutions. They can assist teachers and administrators in identifying struggling students early on, enabling proactive intervention strategies to support their academic progress. Moreover, these models can aid in resource allocation, curriculum planning, and designing personalized learning experiences tailored to individual student needs. Ethical considerations surrounding student data privacy and fairness in model predictions are crucial aspects that need careful attention when developing and deploying such predictive systems in educational settings. Transparency, fairness, and accountability are essential elements to ensure that these models benefit students while avoiding biases or discriminatory outcomes. Overall, student grade prediction models serve as valuable tools in education by leveraging data-driven insights to help improve educational outcomes, support student success, and enable educators to make informed decisions for the benefit of their students.

## 1.1 Problem Statement

The problem statement can be defined as follows Given a dataset containing attribute of 396 Portuguese students where using the features available from dataset and define classification algorithms to identify whether the student performs good in final grade exam, also to evaluate different machine learning models on the dataset.”

### 2. Description of the Dataset

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two data sets were modelled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

**2.1 Attribute Information:**

* School - Student's School (Binary: 'GP' - Gabriel Pereira or 'MS' – Mousinho Da Silveira)
* Sex - Student's Sex (Binary: 'F' - Female Or 'M' - Male)
* Age - Student's Age (Numeric: From 15 To 22)
* Address - Student's Home Address Type (Binary: 'U' - Urban Or 'R' - Rural) Famsize - Family Size
* (Binary: 'LE3' - Less or Equal To 3 Or 'GT3' - Greater Than 3)
* Pstatus - Parent's Cohabitation Status (Binary: 'T' - Living Together Or 'A' - Apart)
* Medu - Mother's Education (Numeric: 0 - None, 1 - Primary Education (4th Grade), 2 - “ 5th To

9th Grade”, 3 - “ Secondary Education Or 4 - Higher Education”)

* Fedu - Father's Education (Numeric: 0 - None, 1 - Primary Education (4th Grade), 2 - “ 5th To
* 9th Grade”, 3 - “ Secondary Education Or 4 - “ Higher Education”)
* Mjob - Mother's Job (Nominal: 'Teacher', 'Health' Care Related, Civil 'Services' (E.G. Or Police),
* 'At Home' Or 'Other')
* Fjob - Father's Job (Nominal: 'Teacher', 'Health' Care Related, Civil 'Services' (E.G. Administrative or Police), 'At\_Home' Or 'Other')
* Reason - Reason to Choose This School (Nominal: Close To 'Home', School 'Reputation', 'Course' Preference Or 'Other') Guardian - Student's Guardian (Nominal: 'Mother', 'Father' Or 'Other')
* Travel time - Home to School Travel Time (Numeric: 1 - <15 Min., 2 - 15 To 30 Min., 3 - 30 Min.

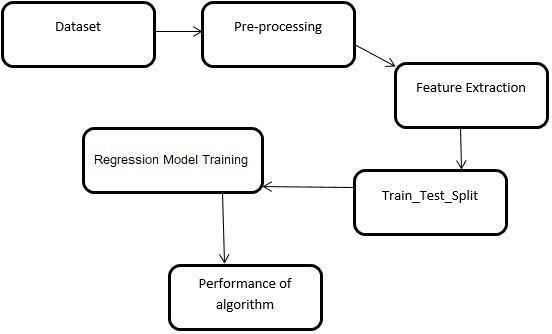
To 1

Hour, Or 4 - >1 Hour)

* Studytime - Weekly Study Time (Numeric: 1 - <2 Hours, 2 - 2 To 5 Hours, 3 - 5 To 10 Hours, Or 4 - >10 Hours)
* Failures - Number of Past Class Failures
* (Numeric: N If 1<=N<3, Else 4)
* Schoolsup - Extra Educational Support (Binary: Yes or No)
* Famsup - Family Educational Support (Binary: Yes or No)
* Paid - Extra Paid Classes Within the Course Subject (Math or Portuguese) (Binary: Yes or No)
* Activities - Extra-Curricular Activities (Binary: Yes or No)
* Nursery - Attended Nursery School (Binary: Yes or No)
* Higher - Wants to Take Higher Education (Binary: Yes or No)
* Internet - Internet Access At Home (Binary: Yes Or No) 
* Romantic - With A Romantic Relationship (Binary: Yes or No)
* Famrel - Quality of Family Relationships (Numeric: From 1 - Very Bad To 5 - Excellent)
* Freetime - Free Time After School (Numeric: From 1 - Very Low To 5 - Very High)
* Goout - Going Out with Friends (Numeric: From 1 - Very Low To 5 - Very High)
* Dalc - Workday Alcohol Consumption (Numeric: From 1 - Very Low To 5 - Very High)
* Walc - Weekend Alcohol Consumption
* (Numeric: From 1 - Very Low To 5 - Very High)
* Health - Current Health Status (Numeric: From 1 - very bad to 5 - Very Good)
* Absences - Number of School Absences (Numeric: From 0 To 93)

#### 3. Methodology

Since universities are prestigious places of higher education, students’ retention in these universities is a matter of high concern. It has been found that most of the students’ drop-out from the universities during their first year is due to lack of proper support in undergraduate courses. Due to this reason, the first year of the undergraduate student is referred as a “make or break” year. Without getting any support on the course domain and its complexity, it may demotivate a student and can be the cause to withdraw the course. There is a great need to develop an appropriate solution to assist student retention at higher education institutions. Early grade prediction is one of the solutions that have a tendency to monitor students’ progress in the degree courses at the University and will lead to improving the students’ learning process based on predicted grades. Using machine learning with Educational Data Mining can improve the learning process of students. Different models can be developed to predict students’ grades in the enrolled courses, which provide valuable information to facilitate students’ retention in those courses. This information can be used to early identify students at-risk based on which a system can 1 suggest the instructors to provide special attention to those students. This information can also help in predicting the students’ grades in different courses to monitor their performance in a better way that can enhance the students’ retention rate of the universities. Using various packages such as cufflinks, seaborn & mat plotlib to represent the data along with different attributes graphically or pictorially to analyse the dataset for predicting the Final Grade(G3).

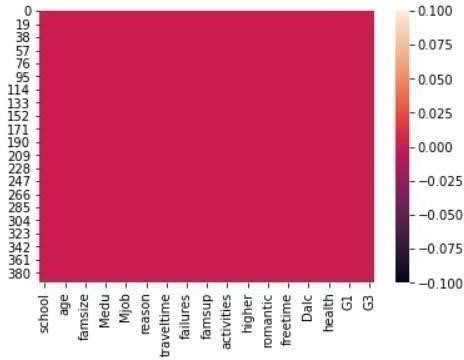


### A. Dataset

The dataset is taken from the UCI Machine Learning Repository [5] of two different schools in Portugal comprising secondary education. The dataset contains information about student performance with several parameters: previously obtained grades, study time, past failures, parent's education, presence in class, etc., and it is the result of data collected using school reports and questionnaires. The datasets are of two subjects: Mathematics and Portuguese language, which were designed according to binary three-level classification and regression.

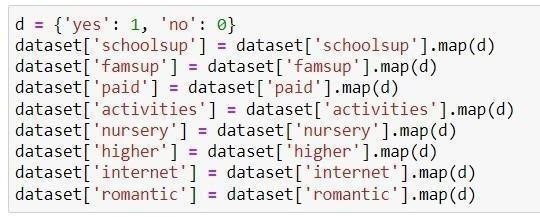
### B. Data Preprocessing

First, we'll check for potential null or nan values in our dataset. Since our dataset is already clean, no null values or missing values present in the dataset. so, we leave out the data cleaning phase and directly proceed with the data preprocessing phase.



#### Figure 2: Heatmap showing no null values in dataset

Our machine learning model needs data to be present in numerical form. So we need to transform categorical values into numerical values, for this purpose we have used ordinal Encoding.

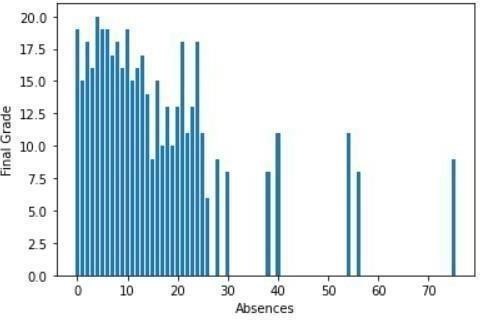




**Figure 3:** Ordinal encoding

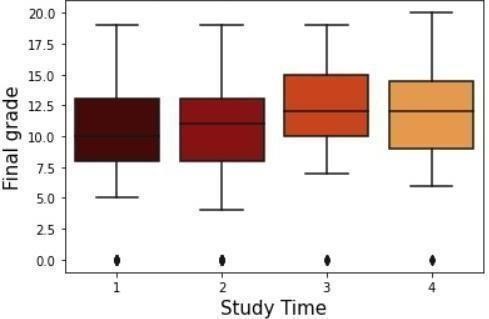
#### C. Exploratory Data Analysis

Before applying the machine learning model, some performance analysis needs to be done to analyze which factor affects the student performance most and how dependent and independent features are related to each other. In EDA, various graphs are plotted based on the dataset. The barplot below indicates that the majority number of students have fewer absences and tend to perform better. At the same time, the students who are absent for most of the classes will get their grades low.



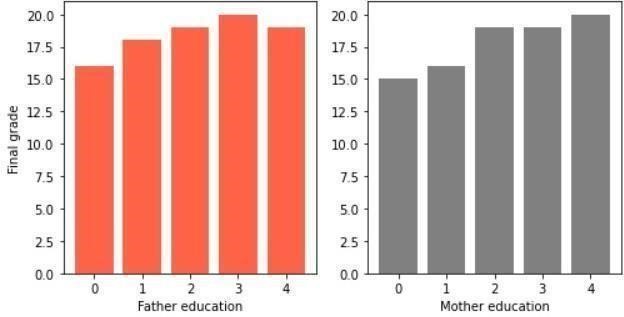
##### **Figure 4:** Bar graph showing student absence in class vs final grade

A box plot is a statistical plot to visualize descriptive statistics, including mean, median, quartile 1, quartile 2, minimum, maximum values. From Fig. 5, it can be inferred that students with study time belonging to group 3(5-10 hours) or group 4(>10 hours) will perform better in the final exams as they have the maximum median final score. Students who belong to group 1(1-2 hours) and group 2(2-5 hours) have the lower final grade. This led to the inference that more study time will always result in better final grades.



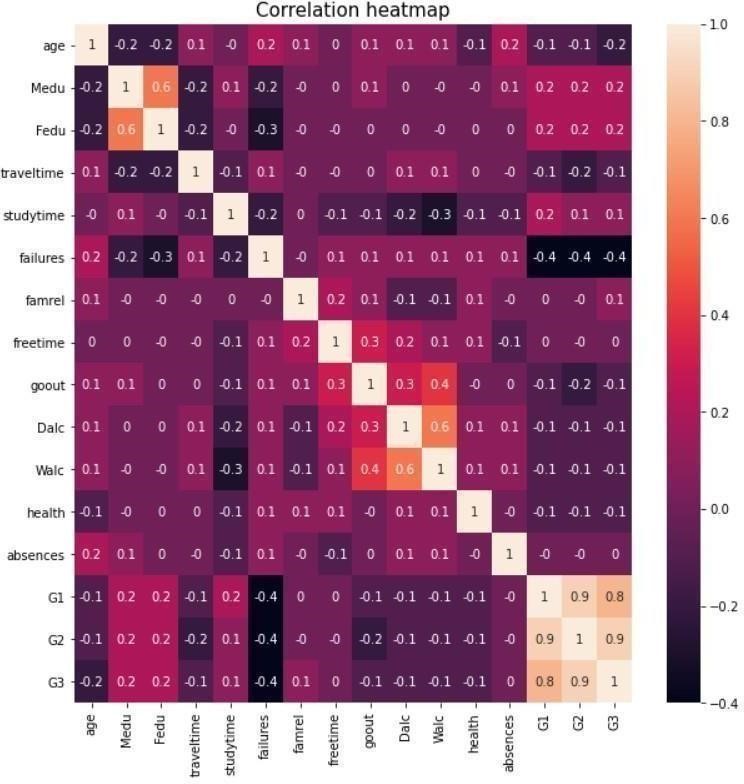
**Figure 5:** Box plot showing Study time vs Final grade.

From the below bar graph, we can see that the parent's education level plays a significant role in student grades. The below graph shows that the connection between the father's education and student's grades is weak, but for the mothers who are educated, their student's final grade increased. It may be because mothers teach their children and spend most of their time with them.



**Figure 6:** Bar graph showing Parent’s education vs Final grade

The below heatmap shows that parents' education, number of past class failures, study time, alcohol consumption, travel time, the student who wants to pursue higher education, class attendance, and previously obtained grades are the eight most correlated factors with G3. Here, G1 and G2 (independent feature) are highly correlated with G3 (dependent feature), but G1 and G2 are highly correlated with each other, so we can remove one of them to avoid overfitting, in this case, we'll only consider G1 (previously obtained grade) to predict G3 (final grade).



**Figure 7:** Correlation Heatmap

By seeing the correlation heatmap

we can observe that:

1. The final grade G3 is highly correlated with the parent's education, and since both mother education and father education are highly correlated, we will only consider mother education.
2. Final Grade G3 negatively correlates with past failures, frequency of going out, and daily alcohol consumption.
3. Previously obtained grades G1 and G2 are highly correlated with final grade G3.

#### D. Various Machine Learning Algorithms

For predicting the student's final grade, we implemented four different regression machine learning Algorithms on our dataset, i.e., multivariate linear regression, random forest, gradient boosting, and bayesian ridge regression.

### E. Testing and Training the data

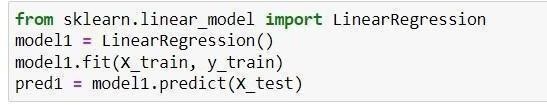
For training and testing, we use the 8:2 ratio, i.e., 80% data for training and the remaining 20% data for testing. Using the scikit-learn library, we split the data into X\_train, y\_train, X\_test, and y\_test. Training data is used for training the model, and testing data is used for predicting and checking the accuracy against training data. At each run, the final result and accuracy score may vary. To avoid this, we set the random state attribute to 0.



**Figure 8:** train\_test\_split

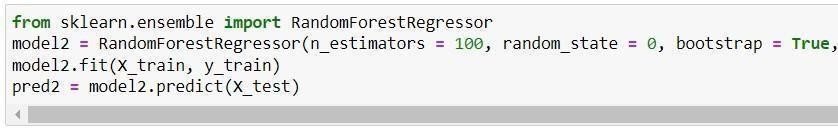
#### 1. Linear regression

Linear regression is a regression technique where a straight line is used to model the relationship between dependent variable and independent variable. More than one feature is used in multivariate linear regression to predict the target value. As to predict the value of the dependent variable, it requires the set of data, so it belongs to the supervised learning category.



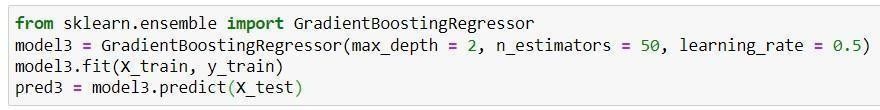
#### 2. Random Forest

Random Forest is a basic yet adaptable machine learning method that delivers excellent results in the vast majority of cases. Due to its simplicity, it is widely used in a variety of problem statements, and it is best suited for both regression and classification issues. Forest is a collection of Decision Trees that have been trained mostly using the "bagging" approach, in which the combination of various learning algorithms helps in improving accuracy.



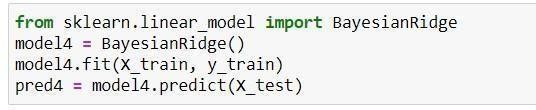
#### 2. Gradient Boosting Regression

Gradient boosting is another machine learning algorithm in which many models are trained consecutively. Using the Gradient Descent approach, each new model gradually reduces the loss function (y = axe + b + e, where 'e' is the error component) of the entire system. The learning process fits new models in a sequence to provide a more precise estimate of the response variable. Gradient boosting regression is also used for both regression and classification problems.



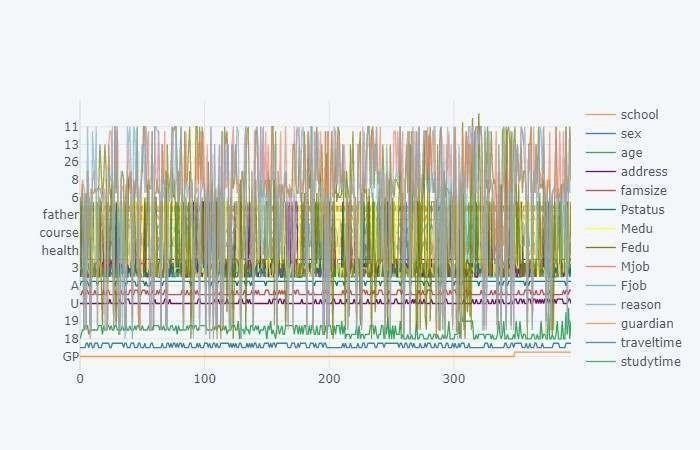
#### 3. Bayesian Ridge Regression

Bayesian Regression is a regression algorithm that comes in handy when there isn't enough data in a dataset or when the data isn't evenly distributed. In contrast to traditional regression techniques, where the output is derived from a single value of each attribute, the output of a Bayesian Regression model is derived from a probability distribution. A normal distribution (where mean and variance are normalized) is used to generate the output.



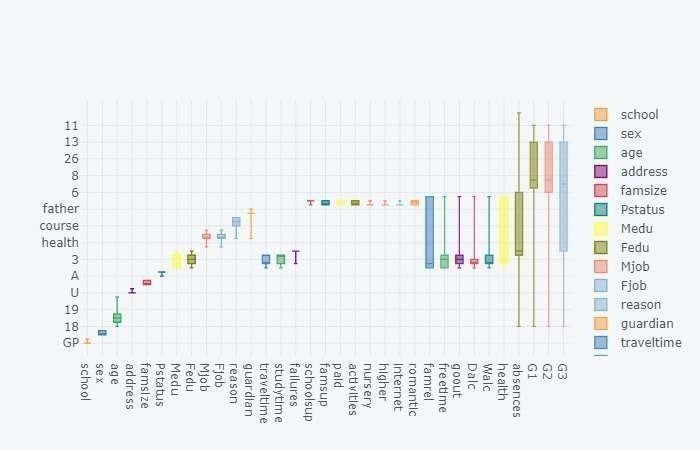


1. **Experimental Results** 
   1. - KDE Plot to view all attributes using cufflinks.

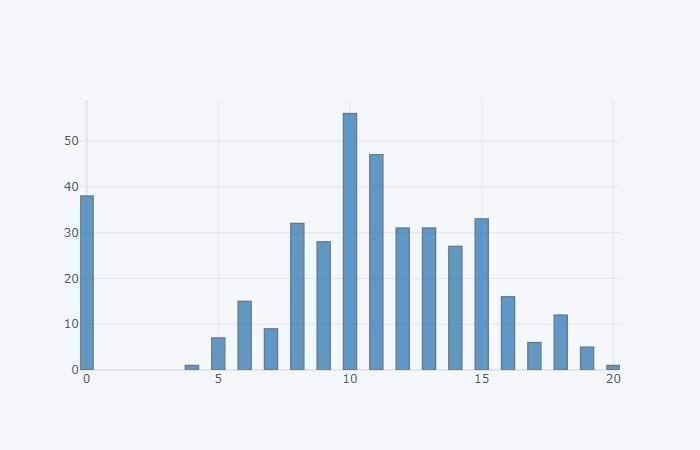


Observation: cufflink connects plotly with pandas to create graphs and charts of data frames directly.

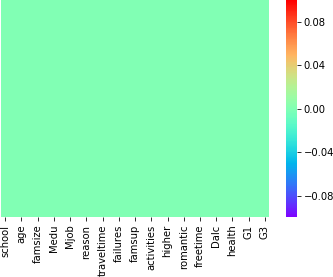
* 1. Box Plot to view all attributes using cufflinks



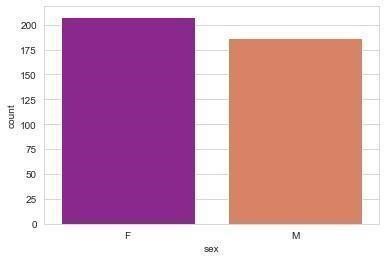
* 1. Histogram Plot for G3 (Final Grade) using cufflinks.



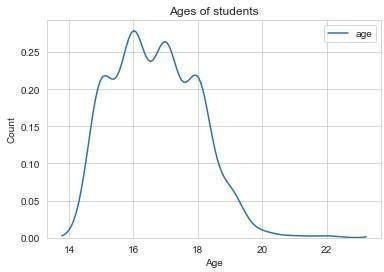
* 1. Pictorial representation of any null data present in the dataset.



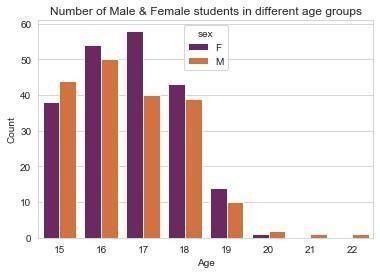
* 1. Count Plot for Student Sex Attribute



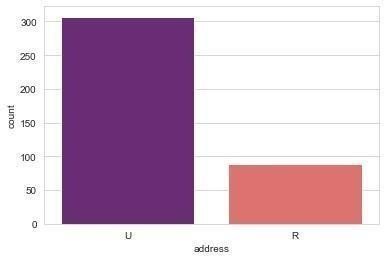
* 1. Kernel Density Estimation for Age of Students.



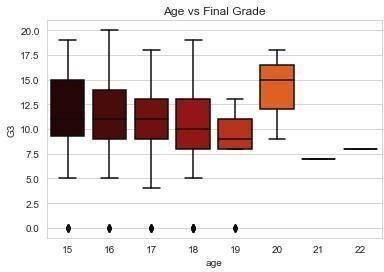
* 1. Count Plot for Male & Female students in different age groups.



* 1. Count Plot for students from Urban & Rural Region.



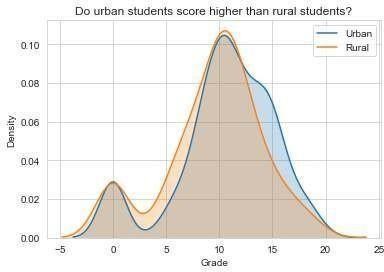
* 1. Does age affect final grade?



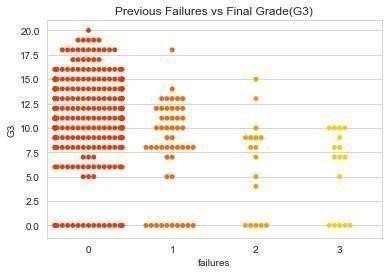
**Observation:**

* + - Plotting the distribution rather than statistics would help us better understand the data.
    - The above plot shows that the median grades of the three age groups (15,16,17) are similar. Note the skewness of age group 19. (may be due to sample size). Age group 20 seems to score highest grades among all.

* 1. Do urban students perform better than rural students?

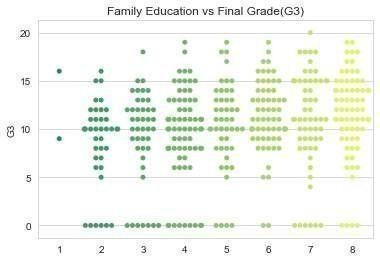


* 1. Previous Failures vs Final Grade (3)

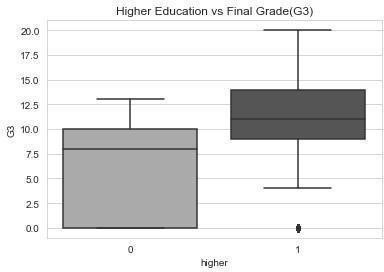


Observation: Student with less previous failures usually score higher.

* 1. Family Education vs Final Grade(G3)

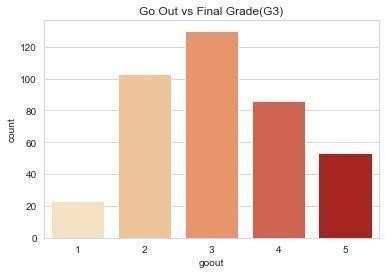


* 1. Higher Education vs Final Grade(G3)



Observation: Students who wish to go for higher studies score more.

* 1. Go Out vs Final Grade(G3)



Observation:

The

students

have

an

average

score

when it

comes

to

going

out

with

friends

&

Students who go out a lot score less.

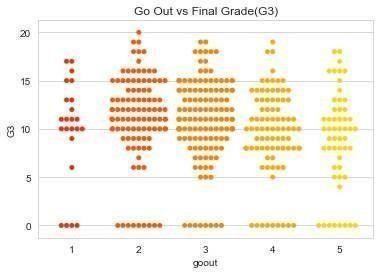
4.15

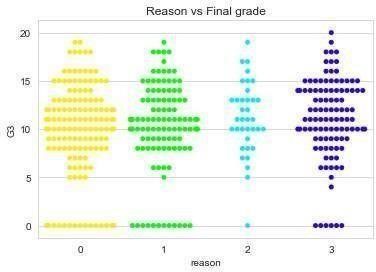
-

Reason vs

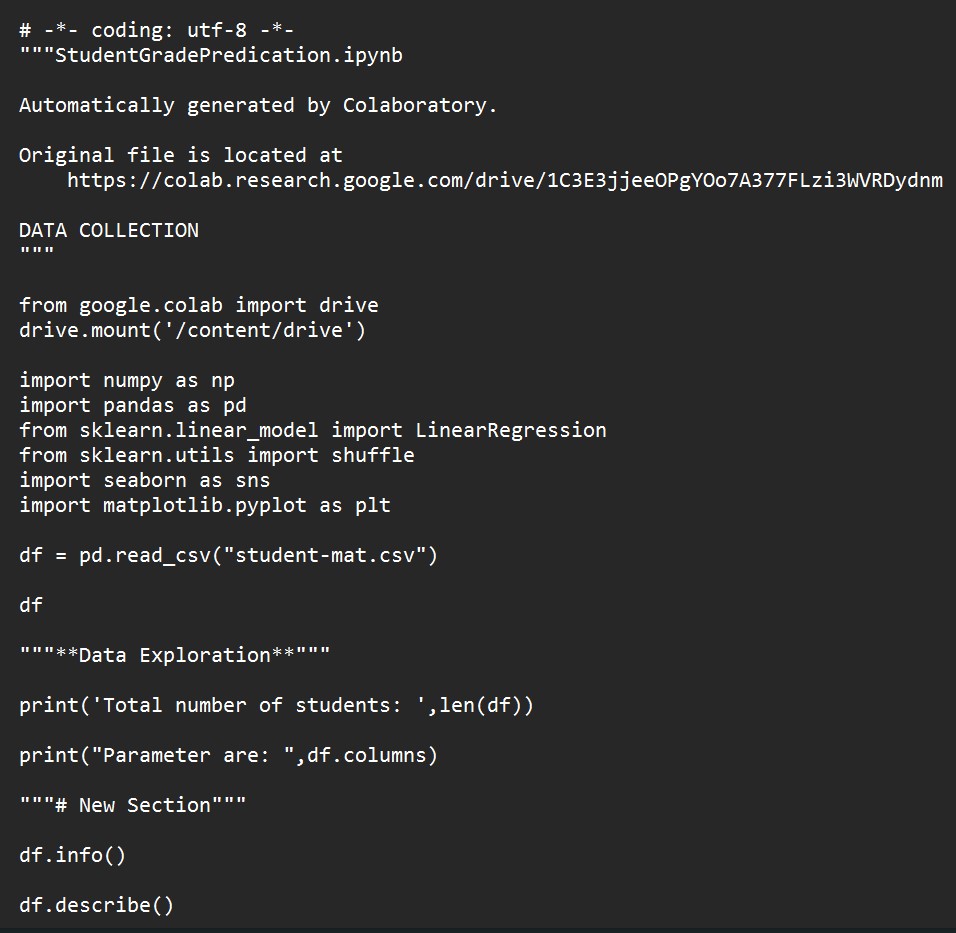
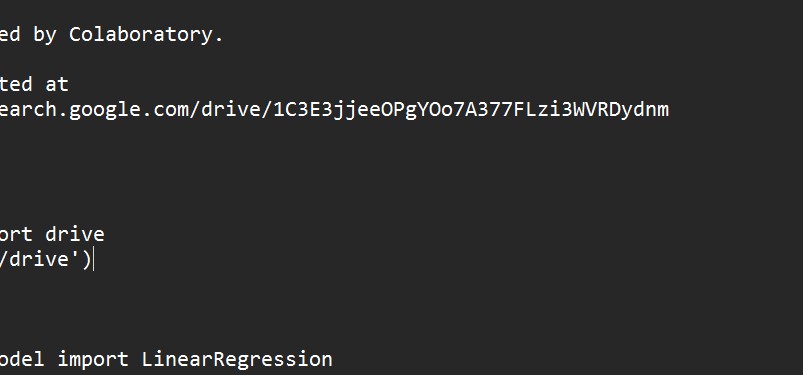
Students

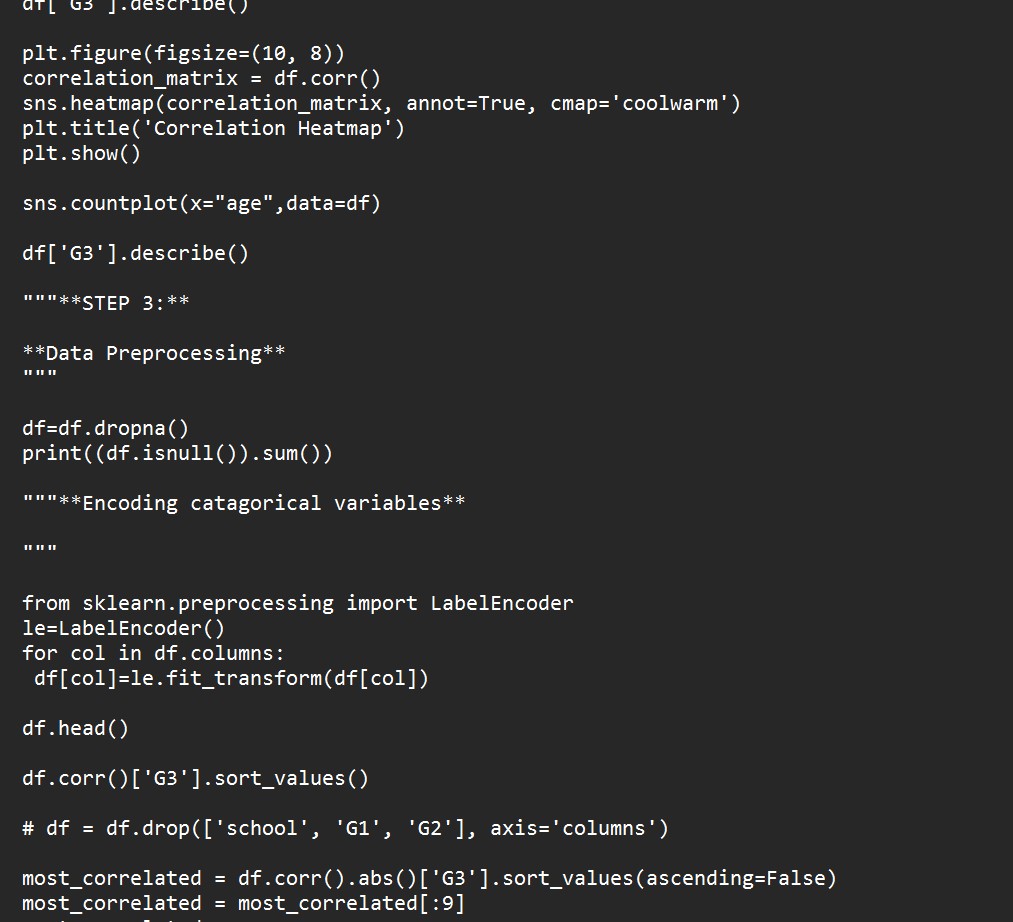
Count

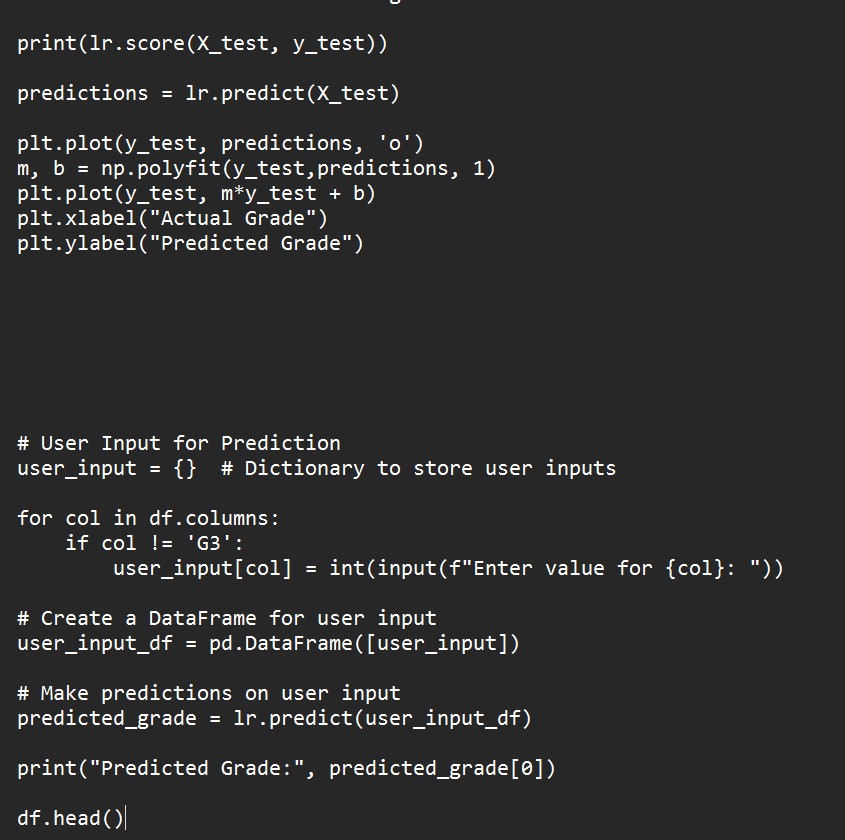
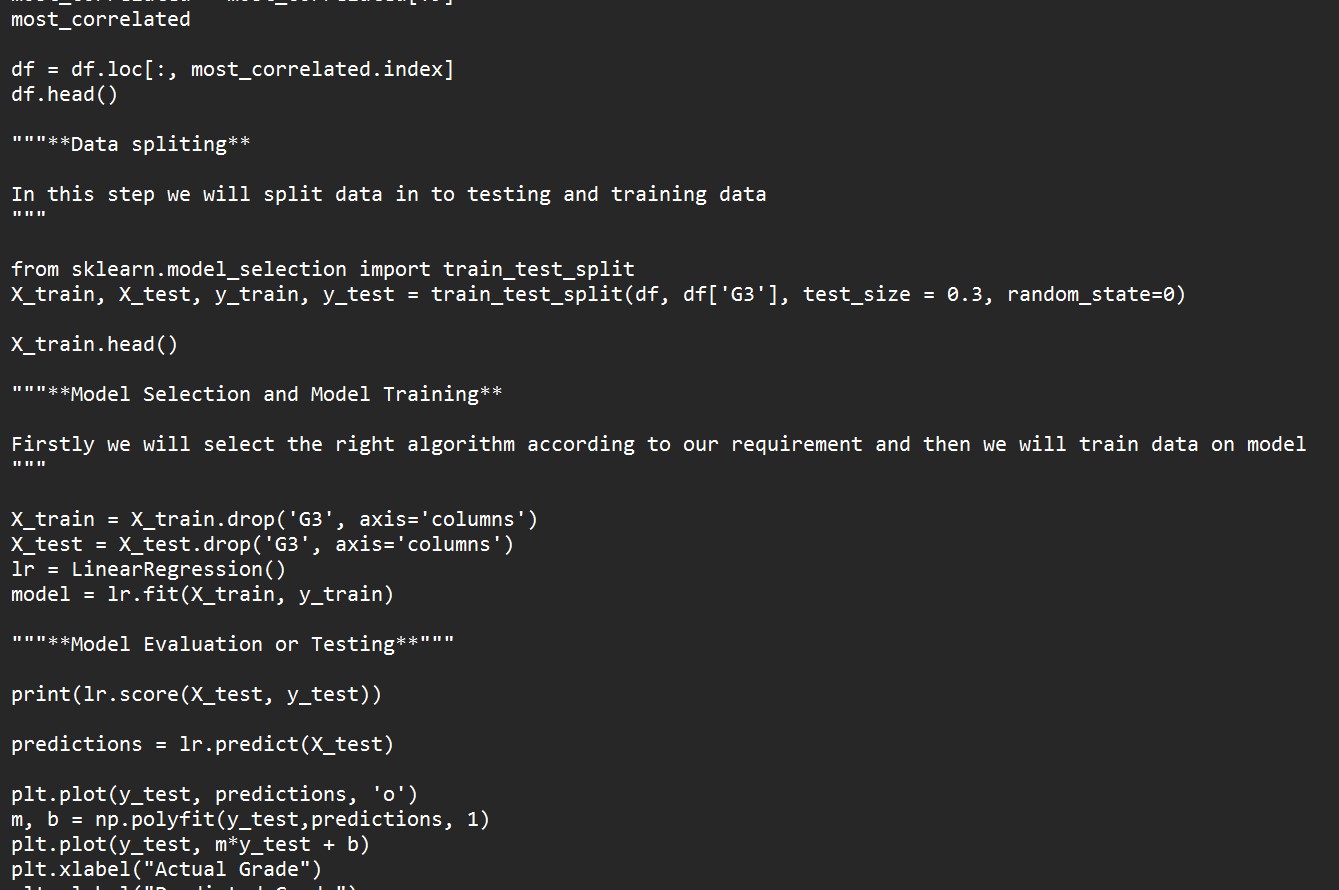




Observation: The students have an equally distributed average score when it comes to reason attribute.





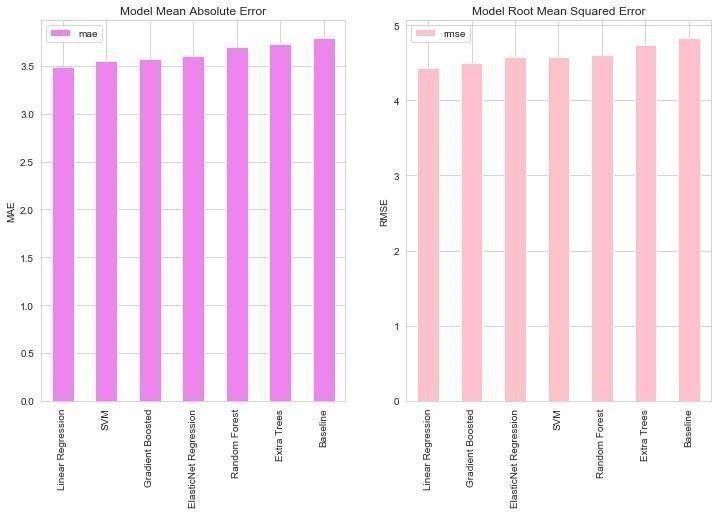


## Results

While considering the student grade dataset, four algorithms, linear regression, random forest, gradient boosting, and bayesian ridge, are tested and under different permutations and combinations. The below bar graph shows the accuracy of each algorithm under these parameters, and gradient boosting comes out to be the best algorithm to predict the student grade based on the given parameters with an accuracy of 79%. The random forest has the second-best result with an accuracy of 74%. Linear regression and bayesian ridge regression show the least accuracy, i.e., 69%.

### 2.Conclusion

As we see both MAE & Model RMSE that the Linear Regression is performing the best in both cases. .



**3. FURTURE USES** :

The future use of student grade prediction models holds significant promise in various aspects of education, leveraging technology and data-driven insights to enhance student outcomes and educational systems. Some potential future uses include: Personalized Learning: Student grade prediction models can facilitate personalized learning experiences by identifying individual student strengths, weaknesses, and learning styles. Educators can tailor instruction and educational resources to better meet each student's needs, fostering improved engagement and academic success. Early Intervention and Support: These models can aid in early identification of students at risk of academic underachievement or dropout. Early intervention strategies, such as targeted tutoring, counseling, or additional academic support, can be implemented to help struggling students before issues escalate. Curriculum Enhancement: Analyzing historical academic data can offer insights into the effectiveness of different teaching methods, curriculum components, or educational programs. Institutions can use this information to refine curricula, update teaching strategies, and optimize educational resources for better student performance. Resources Allocation: Schools and educational institutions can allocate resources more efficiently by identifying trends or patterns that correlate with student success. This may involve adjusting staffing, budget allocation, or implementing specific programs or initiatives to better support students Predictive Analytics for Educational Policy: Governments and educational policymakers can use aggregated data from these models to inform decision-making processes. This might involve understanding broader trends in student achievement, identifying disparities in educational outcomes, and designing policies aimed at improving overall educational equity. Integration with Learning Management Systems (LMS): Integration of predictive models into learning management systems can provide real-time insights to educators. This can help teachers make timely adjustments to their teaching methods or interventions for individual students. Ethical Considerations and Fairness: Future development of these models will emphasize ensuring fairness, transparency, and ethical use of student data. Continuous efforts will be made to mitigate biases, uphold privacy standards, and ensure that these models benefit all students equitably. Adaptive Assessments: These models can potentially assist in the development of adaptive assessment tools that adjust difficulty levels based on predicted student performance. This can offer more accurate evaluations and a better understanding of student mastery levels.

REFERENCES

1. Baradwaj, Brijesh & Pal, Saurabh. (2011). Mining Educational Data to Analyze Students' Performance.

International Journal of Advanced Computer Science and Applications.

2. 63-69. 10.14569/IJACSA.2011.020609.

1. Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of Students Performance using Machine learning. IOP Conference Series: Materials Science and Engineering, 1055(1), 012122.doi:10.1088/1757-899x/1055/1/0121

1. S. Huang and N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques," 2012 Frontiers in Education Conference Proceedings, 2012, pp. 1-2, doi:

10.1109/FIE.2012.6462242.

1. J. Gamulin, O. Gamulin and D. Kermek, "Comparing classification models in the final exam performance prediction," 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 663-668, doi:

10.1109/MIPRO.2014.6859650.

1. https://archive.ics.uci.edu/ml/datasets/student+performance
2. Ajay Ohri (2017, Feb 16). Popular regression algorithms [Online]. Available:

https://[www.jigsawacademy.com/popular-regression-algorithms-ml/ a](http://www.jigsawacademy.com/popular-regression-algorithms-ml/)ccessed on 25.10.2021.

1. A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” in Procedia Computer Science, 2015.
2. P. Guleria, N. Thakur, and M. Sood, “Predicting student performance using decision tree classifiers and information gain,” Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126– 129, 2015