**Student Performance Prediction Using ML**

Ahmad jabali, Hussam Ziad, Adam Almeedani, Mohammed Aljbali, Abdulrahman aljallal.

[2190009080@iau.edu.sa](mailto:2190009080@iau.edu.sa), [2190005458@iau.edu.sa](mailto:2190009080@iau.edu.sa), [2190003266@iau.edu.sa](mailto:2190009080@iau.edu.sa), [2190005704@iau.edu.sa](mailto:2190009080@iau.edu.sa) , [2180005883@iau.edu.sa](mailto:2190009080@iau.edu.sa)

**Abstract**

One of the goals of education is the success of student learning, it’s a very important factor of nations' evolution, our model will predict student performance using ML.

The problem is that a lot of students get a low mark, and they are in danger of getting fail.

This topic is benefit to all students, for example: if one of students predict to get Fail on our model so the student can take an action and get the assistance he needs before failing the class.

We will use our model on Python Programming language to analyze some data sets from Harvard University, one of the data sets contains 398 rows and 33 columns, and second one contains 649 rows and 33 columns. we classified students into three categories: 1- poor, 2- Fair, 3- Good. We get a most Students that is on category Fair, and the smallest number of students on category Good.

Table of Contents

|  |  |
| --- | --- |
| **Content** | **Page** |
| Abstract | 1 |
| 1. Introduction | 3 |
| 2. Review of Related Literatures | 3 |
| 3.0 Description of the Proposed Techniques | 4-5 |
| 3.1 Technique: Decision Tree Algorithm | 5-6 |
| 3.2 Technique: Logistic Regression Algorithm | 6-7 |
| 3.3 Technique: Random Forest Algorithm | 7 |
| 4.0 Empirical Studies  4.1 Description of dataset | 8-9 |
| 4.1.1 Statistical Analysis of the Dataset | 10-11 |
| 4.2 Experimental Setup | 12 |
| 4.3 Performance Measures | 13 |
| 4.4 Optimization Strategy | 13 |
| 4.5 Further Discussions | 14 |
| 5.0 Result and Discussion | 15 |
| 6.0 Conclusion and Recommendations  7.0 Acknowledgements | 17 |
| 8.0 References | 18 |

**1. Introduction**

In 2020, the pandemic Covid-19 affect negatively on Students performance, the classes shift from in campus classes to studying through online meetings. Students’ performance is determined by several factors. For instance, the amount of study time and the current health of the students.

A lot of factors can negatively affect how students perform in all levels of education, and with pandemic there is big changes with learning, study goes from in campus to online meeting, that’s change a student behavior and might be get a lower result before a pandemic, so we will do in our project a new model to help predicting after the pandemic so these students can get the help they need before a student fails a class.

Our project will be a Machine Learning System that is able to predict if a student is on the trajectory to failure by using certain attributes provided by a dataset. We will use it on Python Programming language to analyze data set and classified students into three categories: 1- poor, 2- Fair, 3- Good.

**2. Review of Related Literatures**

The Table below describe some Related Literatures and the difference between them.

Table 2.1: Literature Review for previous works

|  |  |  |  |
| --- | --- | --- | --- |
| num | Name | Algorithms | Accuracy |
| 1 | Student Performance Prediction Model based on Supervised Machine Learning Algorithms | DT, NB ,SVM ,KNN SMO | 68.7%. |
| 2 | Classification and prediction of student performance data using various machine learning algorithms | decision tree and fuzzy genetic | 71% |
| 3 | Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19 | random forest, decision tree and linear regression | 46.35 - 70.04%. |

After we search about the subject and read multiple of article and papers we notice there multiple point common in a lot of paper, first: divide the dataset – here it's student- into a group depend on their performance and grades to make the prediction easier, second: the common algorithm that used in many papers is decision tree.

**3.0 Description of the Proposed Techniques**

We used many Techniques in out project, we will discuss about the three that we ended up choosing, Decision Tree, Logistic Regression and Random Forest.

**3.1 Technique: Decision Tree Algorithm**

Decision Tree algorithm is one of the easiest to use and most widely used supervised learning algorithms. Like the KNN algorithm, it can solve both classification and regression problems. The Decision Tree model is structured in a hierarchal form, like a tree, that branches from the root-node where each internal node represents a feature, branches represent the decisions and lastly each leaf-node is the outcome. The figure below illustrates the structure of a Decision Tree.

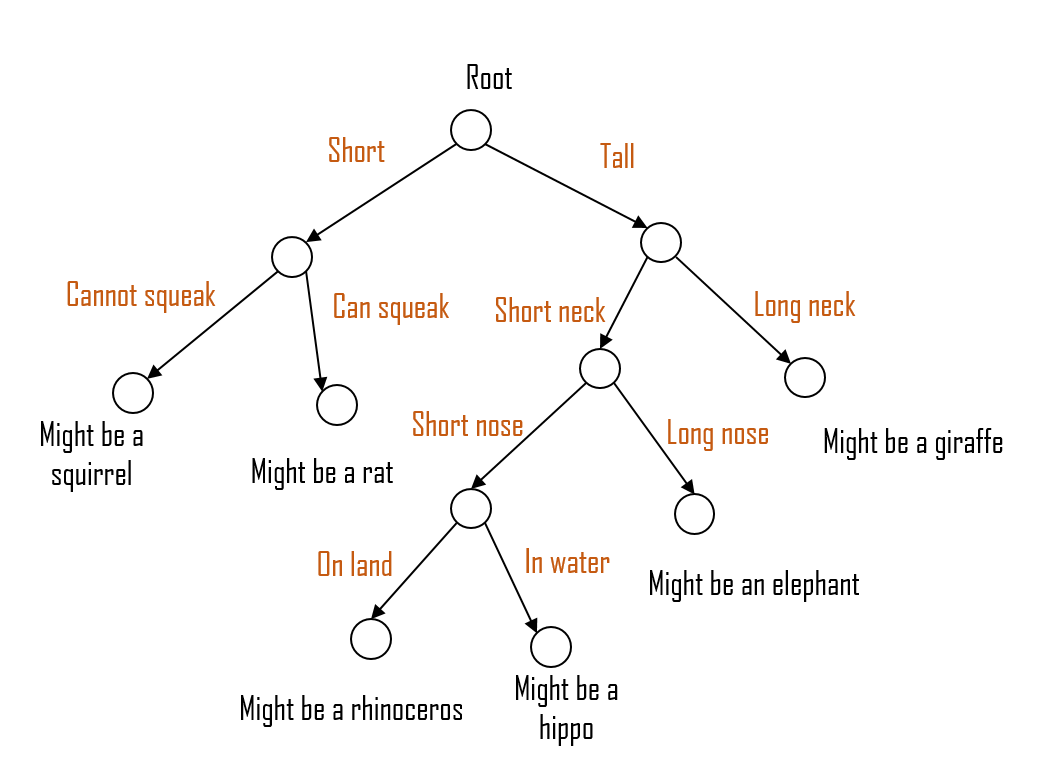


Figure 3.1: Example of a decision tree structure [1]

How does it work? To put it simply, the algorithm starts from the root-node and then compares the value of the attribute given by the root node (from the dataset), then makes a decision and follows the branch to the next node until it reaches a leaf-node which is the final outcome.

For instance, the figure above is a Decision Tree model that predicts which type of animal the input is, based on a set of features. The root-node starts with the feature “Height”, the algorithm then decides whether the height is short or tall so on and so forth until then it reaches the leaf-node.

The Decision Tree construction process includes dividing data of all attributes, compare the results in terms of ‘purity’, then choose the attribute with the highest ‘purity’. Measuring purity is based on the Information Entropy and GINI index [1]. For reference, below is the formula for both.

Formula for Information Entropy:

Formula for GINI Index :

The formulas indicate the probability that the sample belongs to class k (where K is the total number of classes) [1]. The greater the difference between purity before and after segmentation shows that the decision tree is better.

Most common decision tree algorithms: ID3, CART and C4.5.

Decision Tree algorithms use a measure of disorder that goes by “Entropy”. “Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty.” [2]

Mathematical Formula for Entropy:

Where ‘pi’ is the most frequent probability of a class ‘i’ in the dataset. [2]

Entropy is measured between 0 and the number classes in the dataset. The higher the Entropy, the greater the level of disorder. Now that we can measure the amount of disorder / uncertainty in the dataset. We need some way to reduce it. Entropy can be reduced by using a measurement called Information Gain.

Formula for Information Gain:

Information gain is calculated by subtracting the entropies from the original one’s. [3]

**3.2 Technique: Logistic** **Regression Algorithm**

We can say the main point in logistic regression is to describe the relationship between one or more independent variables. This technique can be used in statistics well as ML and it's useful in many fields such as business, marketing and medicine. [4]

Diagram

Description automatically generated

Figure 3.2 : Shows the difference between logistic regression and linear regression

We should use Logistic Regression Algorithm in three cases, if the data is true or false, you want to understand the effect of the features or if you want to get probabilistic results.

To take an example, if we want to make a model that distinguishes from male or female based on their height, then the first part can be considered male and writes a logistic regression model if it is male or not, given a height of is the person male or female?

The coefficients of and have been learned. We can determine the probability of a male given a height of 160cm using the equation above, or more formally: We'll use EXP() for e because that's what you'll see in your screen if you write this example in:

A probability close to zero means that he is male, we can check the probabilities directly because it is a classification and because we want a definitive answer, we can convert it to a binary class. [5]

**3.3 Technique: Random Forest Algorithm**

One of the most popular supervised machine learning algorithms, it's helping to predict behavior and outcomes by combine many classifiers to reach to the goal.

Diagram

Description automatically generatedAfter we describe decision tree in 3.2 there are relation between forest and tree, the model builds many decision trees based on data and when new data come each decision tree give a prediction.

Figure 3.3: Shows the relation between random forest and decision trees.

**4.0 Empirical Studies**

**4.1 Description of dataset**

The data set that was used in this project is the Student Performance Data Set and it contains two subjects: Mathematics and Portuguese. We combined them together and obtained 33 Attributes and 1044 cases. They are described in the following table:

**Table 4.1.1: Dataset Attribute Description**

|  |  |  |
| --- | --- | --- |
| 1 | school | student's school (binary:  "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira) |
| 2 | sex | student's sex (binary:  "F" - female or "M" - male) |
| 3 | age | student's age (numeric  : from 15 to 22) |
| 4 | address | student's home address type (binary:  "U" - urban or "R" - rural) |
| 5 | famsize | family size (binary:  "LE3" - less or equal to 3 or "GT3" - greater than 3) |
| 6 | Pstatus | parent's cohabitation status (binary:  "T" - living together or "A" - apart) |
| 7 | Medu | mother's education (numeric:  0 - none,  1 - primary education (4th grade),  2 – 5th to 9th grade,  3 – secondary education or  4 – higher education) |
| 8 | Fedu | father's education (numeric:  0 - none,  1 - primary education (4th grade),  2 – 5th to 9th grade,  3 – secondary education or  4 – higher education) |
| 9 | Mjob | mother's job (nominal:  "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other") |
| 10 | Fjob | father's job (nominal:  "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other") |
| 11 | reason | reason to choose this school (nominal:  close to "home", school "reputation", "course" preference or "other") |
| 12 | guardian | student's guardian (nominal:  "mother", "father" or "other") |
| 13 | commute\_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) |
| 14 | studytime | weekly study time (numeric:  1 - <2 hours,  2 - 2 to 5 hours,  3 - 5 to 10 hours,  or 4 - >10 hours) |
| 15 | failures | number of past class failures (numeric:  n if 1<=n<3, else 4) |
| 16 | schoolsup | extra educational support (binary:  yes or no) |
| 17 | famsup | family educational support (binary:  yes or no) |
| 18 | paid | extra paid classes within the course subject (Math or Portuguese) (binary:  yes or no) |
| 19 | activities | extra-curricular activities (binary:  yes or no) |
| 20 | nursery | attended nursery school (binary:  yes or no) |
| 21 | higher | wants to take higher education (binary:  yes or no) |
| 22 | internet | Internet access at home (binary:  yes or no) |
| 23 | romantic | with a romantic relationship (binary:  yes or no) |
| 24 | famrel | quality of family relationships (numeric:  from 1 - very bad to 5 - excellent) |
| 25 | freetime | free time after school (numeric:  from 1 - very low to 5 - very high) |
| 26 | goout | going out with friends (numeric:  from 1 - very low to 5 - very high) |
| 27 | Dalc | workday alcohol consumption (numeric:  from 1 - very low to 5 - very high) |
| 28 | Walc | weekend alcohol consumption (numeric:  from 1 - very low to 5 - very high) |
| 29 | health | current health status (numeric:  from 1 - very bad to 5 - very good) |
| 30 | absences | number of school absences (numeric:  from 0 to 93) |
| 31 | G1 | first period grade (numeric:  from 0 to 20) |
| 32 | G2 | second period grade (numeric:  from 0 to 20) |
| 33 | G3 | final grade (numeric:  from 0 to 20, output target) |

**4.1.1 Statistical Analysis of the Dataset**

The statistical analysis of the dataset is presented in table x below. The mean, median, standard deviation, maximum and minimum values of the dataset are presented.

**Table 4.1.1: Statistical Analysis of the dataset**

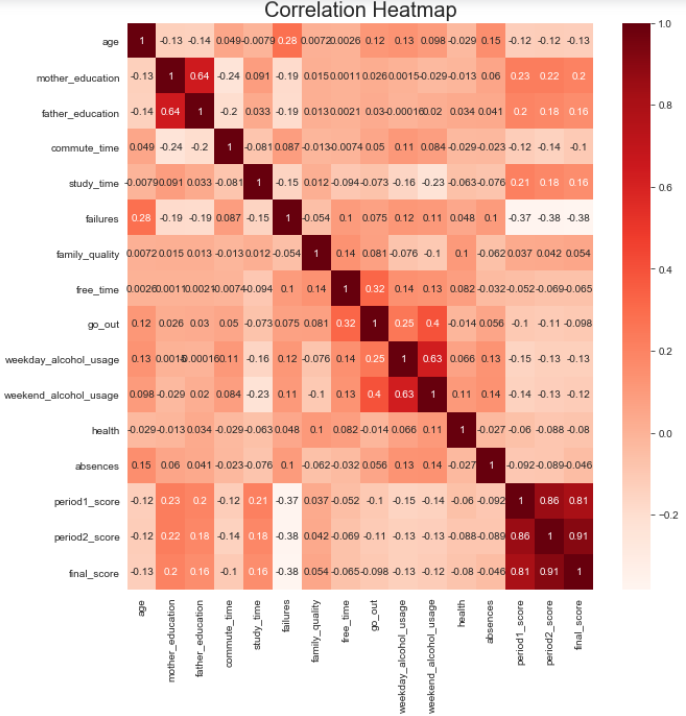


|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| index | count | mean | std | min | 0.25 | 0.5 | 0.75 | max |
| age | 1044 | 16.72605 | 1.239975 | 15 | 16 | 17 | 18 | 22 |
| mother\_education | 1044 | 2.603448 | 1.124907 | 0 | 2 | 3 | 4 | 4 |
| father\_education | 1044 | 2.387931 | 1.099938 | 0 | 1 | 2 | 3 | 4 |
| commute\_time | 1044 | 1.522989 | 0.731727 | 1 | 1 | 1 | 2 | 4 |
| study\_time | 1044 | 1.970307 | 0.834353 | 1 | 1 | 2 | 2 | 4 |
| failures | 1044 | 0.264368 | 0.656142 | 0 | 0 | 0 | 0 | 3 |
| family\_quality | 1044 | 3.935824 | 0.933401 | 1 | 4 | 4 | 5 | 5 |
| free\_time | 1044 | 3.201149 | 1.031507 | 1 | 3 | 3 | 4 | 5 |
| go\_out | 1044 | 3.15613 | 1.152575 | 1 | 2 | 3 | 4 | 5 |
| weekday\_alcohol\_usage | 1044 | 1.494253 | 0.911714 | 1 | 1 | 1 | 2 | 5 |
| weekend\_alcohol\_usage | 1044 | 2.284483 | 1.285105 | 1 | 1 | 2 | 3 | 5 |
| health | 1044 | 3.543103 | 1.424703 | 1 | 3 | 4 | 5 | 5 |
| absences | 1044 | 4.434866 | 6.210017 | 0 | 0 | 2 | 6 | 75 |
| period1\_score | 1044 | 11.2136 | 2.983394 | 0 | 9 | 11 | 13 | 19 |
| period2\_score | 1044 | 11.24617 | 3.285071 | 0 | 9 | 11 | 13 | 19 |
| final\_grade | 1044 | 0.636015 | 0.820218 | 0 | 0 | 0 | 1 | 2 |

**Table 4.1.2: Correlation between each Attribute and the final\_score attribute**



|  |  |
| --- | --- |
| index | final\_score |
| age | -0.125282433 |
| mother\_education | 0.20147169 |
| Sfather\_education | 0.159796049 |
| commute\_time | -0.102627118 |
| study\_time | 0.161628935 |
| failures | -0.38314528 |
| family\_quality | 0.054461059 |
| free\_time | -0.064889679 |
| go\_out | -0.09787726 |
| weekday\_alcohol\_usage | -0.129642125 |
| weekend\_alcohol\_usage | -0.115740004 |
| health | -0.08007864 |
| absences | -0.045670577 |
| period1\_score | 0.809141721 |
| period2\_score | 0.910743163 |
| final\_score | 1 |



**Figure 4.1.1: Correlation Heatmap**

**4.2. Experimental Setup**

Out experiment was carried out by using Python Jupyter Notebook. First, we merged the two datasets, the Math and Portuguese, into one dataset. And then we categorized the final grade into 3, Good:15-20, Fair:10-14, Poor: 0-9.

Then we prepared the dataset for training by using direct partitioning. The total amount of records for the dataset is 1044 and split the dataset into 70% training (730 records) and 30% for testing (313 records). There are 3 main algorithms we focused on, the Decision Tree, Logistic Regression and Random Forest. After finishing the model building for all three algorithms, we found that the Logistic Regression has the highest accuracy. Here’s a flowchart to depict how we worked on the experimentation. Lastly, we did the Exploratory Data Analysis to see how the features correlate to each other.

Diagram

Description automatically generated

Figure 4.2.1: Experimental Setup Flowchart

**4.3 Performance Measures**

In machine learning, we often use 4 things to evaluate models: accuracy, precision, recall, f1 score. In our project we used accuracy equal to

TP: If it is expected that the student will succeed and actually succeed, TN : If the expectation is that the student will fail and he has already failed, FP: If a student is expected to fail, but succeeds , FN: If the student is expected to succeed but fail.

We aim to increase fp and decrease FN

**4.4. Optimization strategy**

Through experimenting on the three algorithms, we applied feature selection method for each of them to further optimize the training. Furthermore, the final best parameters for each of the three algorithms are provided in the tables below.

**Table 4.4.1: Optimum parameters for the Decision Tree Model**

|  |  |
| --- | --- |
| Parameters | Optimal Value chosen |
| criterion | gini |
| splitter | best |
| max\_depth | none |
| min\_samples\_split | 2 |
| min\_samples\_leaf | 17 |

**Table 4.4.2: Optimum parameters for the Random Forest Classifier**

|  |  |
| --- | --- |
| Parameters | Optimal Value chosen |
| n\_estimators | 36 |
| criterion  min\_samples\_leaf | gini  2 |

**Table 4.4.3: Optimum parameters for the Logistic Regression**

|  |  |
| --- | --- |
| Parameters | Optimal Value chosen |
| multi\_class | multinomial |
| solver  fit\_intercept | newton-cg  True |

**4.5 Further Discussions**

Now we will compare our work with previous work on the same data set:

**Table 4.5.1: Further Discussion of previous studies**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Name | Algorithms | Highest accuracy achieved |
| 1 | analyzing-students-performance  P - Yara-Aldajjani[8] | DT, NB ,SVM ,KNN SMO, RandomForest classifier | 78.7%. |
| 2 | Student Data Set Álvaro Bermejo García February 2016[9] | LR,NN,SVR | 85% |
| 3 | Using data mining to predict secondary school student performance[10] | Naive Bayes | 78%. |

We note that we have done a good job and have raised the percentage of accuracy and improved the performance of the model.

**5.0 Result and discussion**

Anticipating students’ performance has a major role in improving their academic results, as it is possible to know the weaknesses of each student and work to improve them. So, our model is to improve students’ performance through the application of a machine learning approach that will help us in predicting students’ performance. So, After Importing the Packages, Load the Dataset, Data Preparation and EDA, we tested the performance of different models we applied on the dataset, the winner is Logistic Regression model, With accuracy of 89%. Below table shows all the accuracies for each of our algorithms tested.

**Table 5.1: Accuracies of all the techniques used.**

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Logistic Regression | 89% |
| Random Forest Classifier  Decision Tree Classifier | 87%  85% |

**Calendar

Description automatically generated**

**Figure 5.1: Confusion Matrix for Logistic Regression Model**

Based on the model findings and correlations, The excellent student profile is likely to have these following qualities:

- Mother is a healthcare professional

- Father is a teacher

- Chose to attend the school based its reputation

- Study more than 10 hours weekly

-Have access to Internet at home

- Is healthy

- No absences to classes

**6.0 Conclusion and Recommendations**

Our model’s purpose is to improve students’ performance through the application of a machine learning approach that will help us in predicting students’ performance. So, After Importing the Packages, Load the Dataset, Data Preparation and EDA, we tested the performance of different models we applied on the dataset, the winner is logistic regression model, With accuracy of 88%. Based on the model findings, The excellent student profile is likely to have these qualities:

- Mother is a healthcare professional

- Father is a teacher

- Chose to attend the school based its reputation

- Study more than 10 hours weekly

-Have access to Internet at home

- Is healthy

- No absences to classes

We recommend to experiment further with the parameters of the proposed models and gather a larger amount of samples.

**7.0 Acknowledgements**

In the end, we, the members of the group, would like to thank Dr. Adam for his support and assistance in developing our skills and working on this project.

# 8.0 **References**

|  |  |
| --- | --- |
| [1] | olive.zhao, "Machine Learning Algorithms: Decision Trees," Huawei, 2021 March 24 . [Online]. Available: https://forum.huawei.com/enterprise/en/machine-learning-algorithms-decision-trees/thread/710283-895. |
| [2] | S. T, "Entropy: How Decision Trees Make Decisions | By Sam T | Towards Data Science," 2019 Jan 11. [Online]. Available: https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8.. |
| [3] | Machinelearninglastery.com, "Information Gain and Mutual Information," [Online]. Available: https://machinelearningmastery.com/information-gain-and-mutual-information/.. |
| [4] | N. S. Chauhan, "Decision Tree Algorithm, Explained - KDnuggets," [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. |
| [5] | A. Saini, "Decision Tree Algorithm - A Complete Guide - Analyitics Vidhya," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/#h2\_9. |
| [6] | J. Brownlee, "Machine Learning Algorithms," 4 1 2016. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. |
| [7] | "Logistic Regression Model for the Academic Performance of First-Year Medical Students in the Biomedical Area," 12 1 2016. [Online]. Available: https://www.researchgate.net/publication/308383245\_Logistic\_Regression\_Model\_for\_the\_Academic\_Performance\_of\_First-Year\_Medical\_Students\_in\_the\_Biomedical\_Area. |
| [8] | rroaa, "analyzing-students-performance," Github, 2020 11 12. [Online]. Available: https://github.com/rroaa/analyzing-students-performance. |
| [9] | camilalobianco, "Students Performance," Github, 2020 8 9. [Online]. Available: https://github.com/camilalobianco/StudentsPerformance.. |
| [10] | H. Teguim, "Prediction of Student Performance Using Machine Learning Algorithms," 9 2017. [Online]. Available: [Online]. Available: https://www.researchgate.net/publication/351707716\_Prediction\_of\_Student\_Performance\_Using\_Machine\_Learning\_Algorithms.. |