SJIF Impact Factor 2022: 8.197| ISI I.F. Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)

EPRA International Journal of Research and Development (IJRD)

Volume: 7 | Issue: 4 | April 2022 - Peer Reviewed Journal

ANALYZING & PREDICTING STUDENTS' PERFORMANCES USING MACHINE LEARNING

Atharva M. Wankhede[1], Suraj M. Jha[2], Varad R. Patil[3], Mr. Pravin V. Shinde[4]

[1]*Student, Dept. of I.T. VPPCOE & VA, Mumbai University, Mumbai, India*
[2]*Student, Dept. of I.T. VPPCOE & VA, Mumbai University, Mumbai, India*
[3]*Student, Dept. of I.T. VPPCOE & VA, Mumbai University, Mumbai, India*
[4]*Professor, Dept. of I.T, VPPCOE & VA, Mumbai University, Mumbai, India*

ABSTRACT

Literacy is very critical to economic development as well as individual and community well-being. Effective literacy skills open the doors to more educational & employment opportunities. Moreover, the economic success of any country highly depends on making higher education more affordable and that considers one of the main concerns for any government. Thus, this Machine Learning project is to classify and predict the future academic grades and leadership scores of the students. The ability to predict student performance in education is very significant in educational environments. This will help teachers and professors in school and universities consolidate the student on improving and developing each student's curriculum record. The main intention is to identify and support the students to score better marks for their betterment.

KEYWORDS—*Prediction, Performance, Education, Marks, Machine Learning, Linear Regression Algorithm, Dataset, Evaluation, Data Preparation, feature selection, Libraries, Results.*

I.INTRODUCTION

Machine Learning can be used to predict the performance of the students and identifying the risk as early as possible, so appropriate actions can be taken to enhance their performance. ML techniques would help students to improve their performance based on predicted grades and would enable instructors to identify such individuals who might need assistance in the courses.

Our primary rationale behind making this project is to ease the preparations of students by predicting their academic future outcome based on their previous results. ML techniques would help students to improve their performance based on predicted grades and other crucial factors and it would enable instructors to identify such individuals who might need assistance in the courses.

Machine learning techniques can be used to forecast the performance of the students and identifying the at risk as early as possible so appropriate actions can be taken to enhance their performance. The aim is to help the students to avoid his/her predicted poor result using ML, were we are trying to find out student's current status and further predict his/her future results.

- System can help a teacher about the students like which student needs what kind of help.
- Train the student by finding out the student's weakness & strengths for final examinations.
- Predict the performance of students to derive various correlations of student's performance

It is difficult to analyse the exam manually, most of the time results are not precise as the calculated ones and evaluations are also done manually i.e., time consuming. Result processing after summation of exam takes time as it's done manually. So, we introduce an examination portal system, which is computerized.

System is an emerging field and is very crucial to schools and universities in helping their students and professor to develop performance and grades while keeping in mind other personality factors like interests, attributes and opinions (IAO) which affect their daily lifestyle. Using that information, we can analyze the performance, which will help for both students and mentors.

II.METHODOLOGY

Predictive model for predicting students' performance using Machine Learning (Asaad Masood - 2019) Algorithms used: Decision Tree, Logistic Model Tree (LMT), Association Rules Mining. Data Set used 1021 Records from examination Database. Advantages: Analysis of Factor influencing student performance, Real time intelligent & accurate decision-making Disadvantages: Dataset used is relationally old.

Machine Learning algorithm for student performance prediction (H.M. Rafi Hasan-2019) Algorithms used: Support Vector Clustering (SVC), Decision Tree, K-Nearest Neighbor (KNN). Data Set used 1170 Students Record from University Database. Advantages: Learning about each student, the method can identify weaknesses & suggests ways to improve. Disadvantages: Less accuracy, Non-academic attributes are not considered.

Student Performance Assessment and Prediction System using Machine Learning (Mehil Shah-2019) Algorithms used: Support Vector Clustering (SVC), Decision Tree, K-Nearest Neighbor (KNN) Random Forest Model. Data Set used Portuguese data set using 33 different attributes modeled under binary/five-level classification and regression tasks. Advantages: Considers wide range of non-academic attributes like travel time, father's or mother's job, etc. Disadvantages: Less accuracy Neural network, Accuracy rate of all algorithm used is considerably low

Student marks prediction using Matrix factorization (Thai Nghe et al - 2011)
Algorithms used: Factorization Matrix (FM), Stochastic Gradient Descent (SGD), Logistic Regression, Personalized Multi-Linear Regression (PMLR). Dataset of 171 students uses 2 different attributes performance of Algebra & Bridge to Algebra courses Advantages: Explores a completely different way of approaching the problem by checking the ability to solve the tasks when interacting with the tutoring system. Disadvantages: Comparatively smaller data set making predictions a bit unreliable.

Student marks prediction using Factorization Machine (Mack Sweeney-2016)
Algorithms used: Collaborative Filtering (CF), Sequential Coordinate-wise Descent (SCD). Data Set used 1021 Records from examination Database. Advantages: Minimal prediction error, accurately predict grades for both new and returning students taking both new and existing courses. Disadvantages: Requires additional info on the courses from historical transcript & instructions, Algorithms used are hard to understand.

Predicting student performance: an application of data mining methods with an educational web-based system (Behrouz Minaei-Bidgoli) Algorithm used Genetic Algorithms have been shown to be an effective tool to use in data mining and pattern recognition. Dataset used students' final grades based on their web-use features, which are extracted from the homework data. We design, Implement & evaluate a series of pattern classifiers with many parameters in order to compare the performance on a dataset from LON-CAPA. Advantages: This paper presents on This paper presents on approach to classifying students in order to predict their final grade based on features extracted from logged on to in an education web-based system. Disadvantage: The time overhead for evaluation is therefore a critical issue.

Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations (HAYAT SAHLAOUI). Algorithm used is a simple grid search algorithm provided by scikit to estimate the optimal parameters of our model. Data Set used three nominal intervals depending on the student's average grade (high performer, average performer, and low performer). Advantages: Deals with students' observed records that represent the training set and matching student historical data that represent features with their label that represents the actual performance. Disadvantage: Student performance value could be numeric in the case of a regression problem or categorical in the case of a classification problem.

Early Prediction for At-Risk Students in an Introductory Programming Course Based on Student Self-Efficacy (Mona M. Jamjoom, Eatedal A. Alabdulkareem). Algorithm used Decision Tree, Bayesian classifier, artificial neural network (ANN), SVM, and kNN algorithm. Data set used dataset was derived from students of the CS department of the College of Computer and Information Sciences (CCIS) at Princess Nourah bint Abdulrahman University (PNU). Advantages: The experiments of this study yielded good results in term of accurate prediction for student performance in the final test. Disadvantage: predicting who would quit the course was difficult, as there was an increase in the dropout rate amongst students who attained high score.

In the above section we discussed how Machine Learning can be used to predict the performance of the students and identifying the risk as early as possible so appropriate actions can be taken to enhance their performance. This paper is aimed to groom the student by finding out the dependencies for examinations. The details are given below.

Software used Tech: Python, Machine Learning. Libraries: NumPy, pandas, matplotlib, seaborn, scikit-learn. Implementation: Google Colab, Jupyter Notebook. Website: HTML, CSS, JavaScript, Flask.

Hardware used Processor: i3 4th Generation or higher, 4GB RAM, 1GB Storage.

Dataset used

This data approaches student's achievement in secondary education of two Portuguese schools. The data includes student grades, social and school-related features), demographic and was collected by using school reports and questionnaires. Pair of datasets are provided including the performance in only two distinct subjects: Math's (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two data sets were modeled under a binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1.

This occurs cause the G3 attribute is the final year grade (issued at 3rd period), while G1 & G2 corresponds to the 1st and 2nd period grades. It's more difficult to predict G3 without G2 and G1, therefore such prediction is much more useful.

The entire dataset was extensively analyzed and features (attributes) were tested against the label (final attribute) using Pearson's Correlation Coefficient, Chi-Squared Test and ANOVA Test, the dataset was splinted in 2 parts to avoid overfitting for training, testing and validating sets, we split the dataset into 75:25 for training to testing sets ratio. From this, 75% of the dataset is used for training sets & 25% of the dataset is used as testing sets.

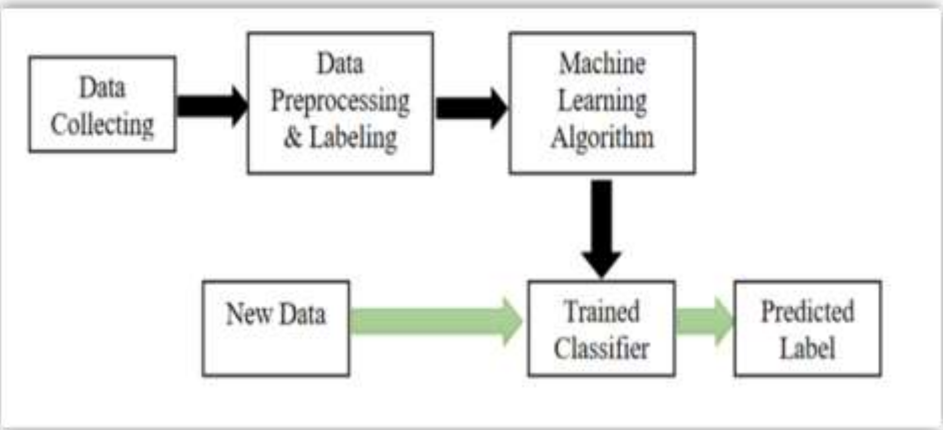


Fig.1: (DFD) Dataflow Diagram

To get the right predictions, we must construct the data set and transform the data correctly.

- The first step is collecting the data from the data sources. In our case, the data has been collected using a survey given to the students and the students' grade book.
- The second step is pre-processing the data in order to get a normalized dataset and then labelling the data rows.
- In the third step, the result of the second step, the training and testing dataset, is fed to the Machine Learning algorithm.
- The Machine Learning Algorithm builds a model using the training data and tests the model using the test data.
- Finally, the Machine Learning Algorithm produces a trained model or a trained classifier that can take as an input a new data row and predicts its label.

Input Data

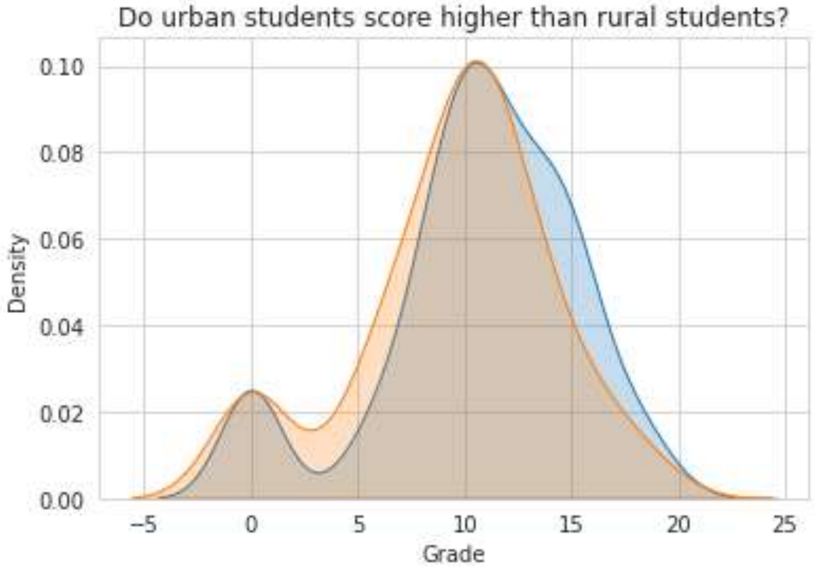
We use dataset containing attribute of 396 Portuguese students were using the features available from dataset and define classification algorithms to identify whether the student performs good in final exam to evaluate different machine learning models on to the dataset.

Attribute Information

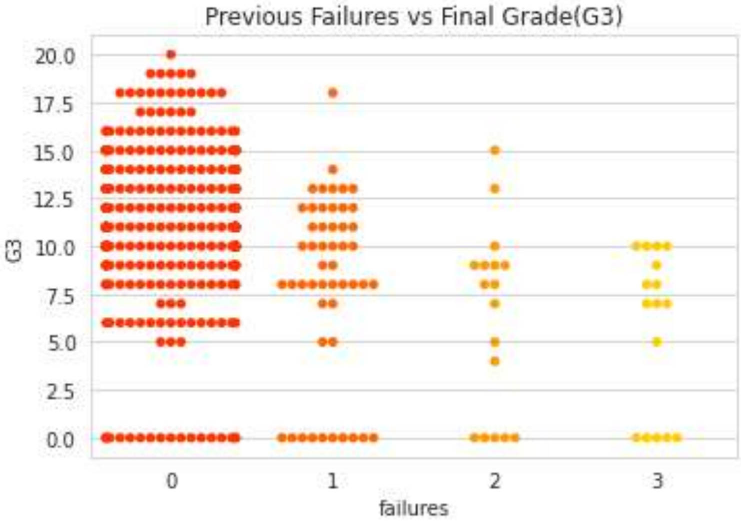
school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho Silveira), **sex** - student's sex (binary: 'F' – female/'M' - male), **age** - student's age (numeric: from 15- 22), **address** - student's home address type (binary: 'U' - urban or 'R' - rural), **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3),**Pstatus** - parent's 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. administrative 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), **study time** - 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), **schools**- extra educational support (binary: yes or no) (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),

IV.MODELING AND ANALYSIS

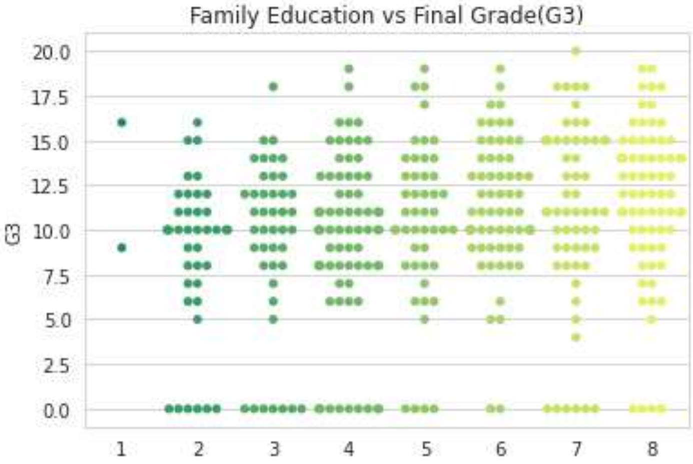
- It was found that the students' location does not matter as students from both urban and rural areas almost scored identically. [attribute: address] (FIG.2)



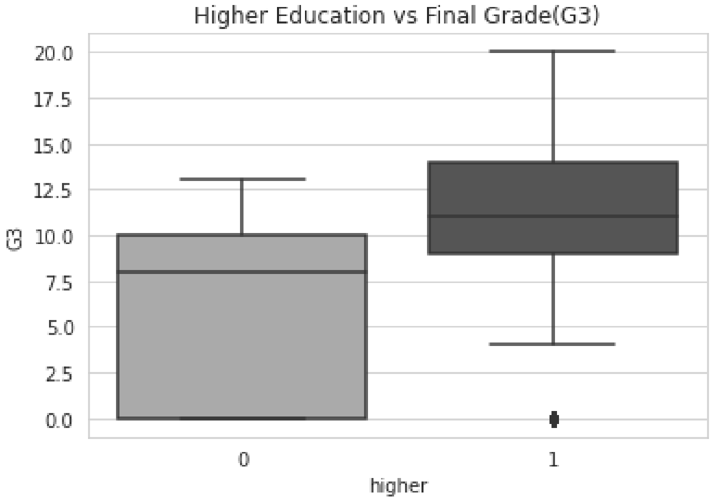
- Students with less previous failures usually scored higher. [attribute: failures] (FIG.3)



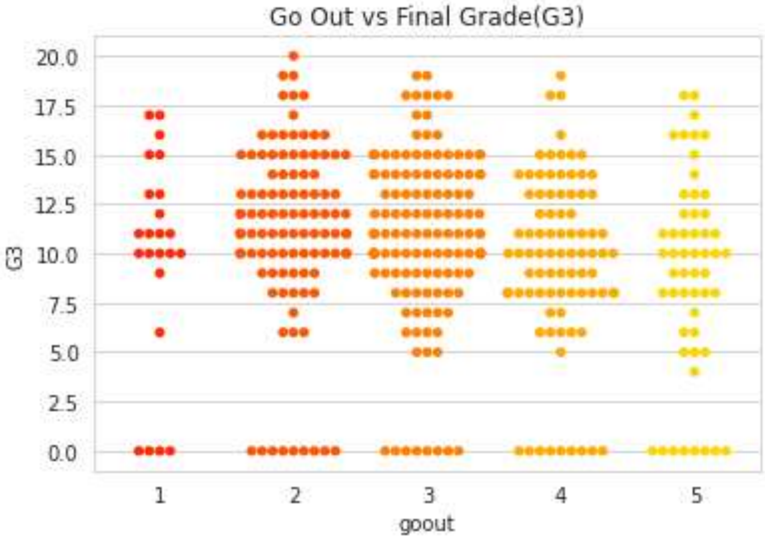
- Student coming from Educated Families scored higher. [attribute: Fedu + Medu] (FIG.4)



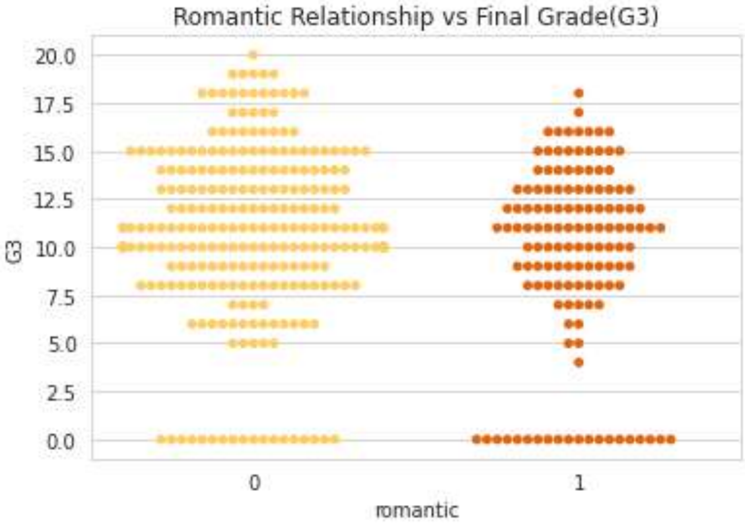
- Students wishing to pursue Higher Education had better grades than those who didn't wish to go for Higher Education. [attribute: higher] (FIG.5)



- Students who hangout or engage in social outings a lot usually scored lesser. [attribute: gout] (FIG.6)



- Students who were not involved in any romantic relationship scored higher grades. [attribute: romantic] (FIG.7)



Dataset Information:

1. The gender distribution is very even as number of female students were 208 while the rest 187 were male.
2. Approximately, 77.72% students come from urban region and 22.28% from rural region.

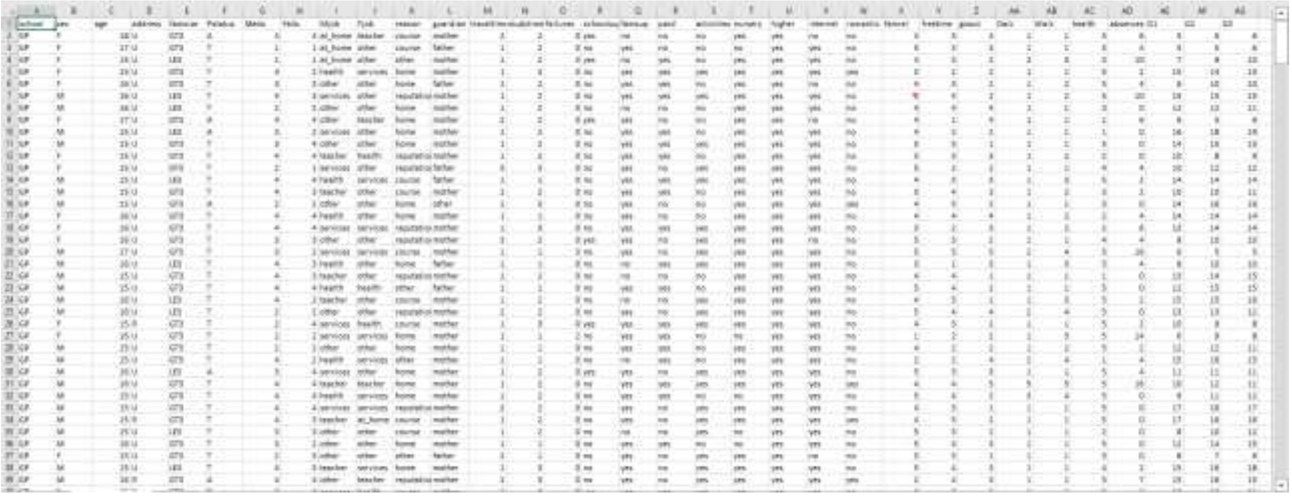


Fig.8: (Dataset) Set of Attributes

Machine Learning Algorithms used:

Linear Regression: Linear Regression is a machine learning algorithm i.e. based on supervised learning. It performs regression task. A regression models a target prediction value based on independent variables provided. It is mostly used to find out the relationship between variables and forecasting.

Random Forest: Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is simple i.e. to find a hyperplane in an N-dimensional space that distinctly classify the data points.

SVM: Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

Gradient Boosting: It is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Extra Trees: It is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as-good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble.

Bayesian Regression: allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response 'y' is assumed to drawn from a probability distribution rather than estimated as a single value.

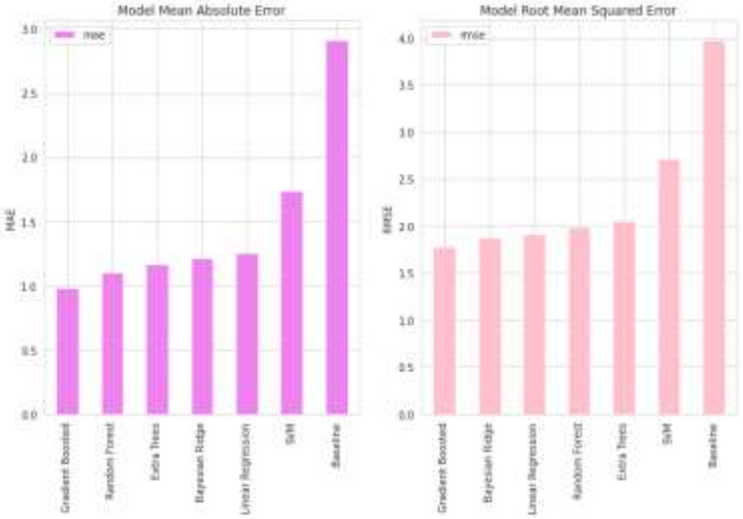


Fig.9: Popular Supervised Learning Algorithms Compared

ALGORITHMS	MAE	RMSE
Gradient Boosted	0.983401	1.779369
Random Forest	1.109453	2.019044
Extra Trees	1.112205	2.031504
Bayesian Ridge	1.216506	1.874521
Linear Regression	1.252096	1.912315
SVM	1.740928	2.711474

Table.1: Mean Absolute Error & Root Mean Squared Error Statistics for all the mentioned algorithms

V. RESULTS AND DISCUSSION

After all the analysis and testing, 8 features used for the purpose of predictions were:

- Travel time: The time required to reach the school/college to home.
- Study time: Number of Hours spent on studying by a student in a week.
- Past failures: Number of backlogs/Arrears in past.
- Higher education: If the student wishes to pursue higher education.
- Health: State of physical, mental & social wellbeing of the student.
- Absences: Number of classes/Lectures missed by a student
- G1 and G2 (first and second period grades).

The following findings are considered to judge our proposed system some major attributes are considered out of the total 33 attributes.

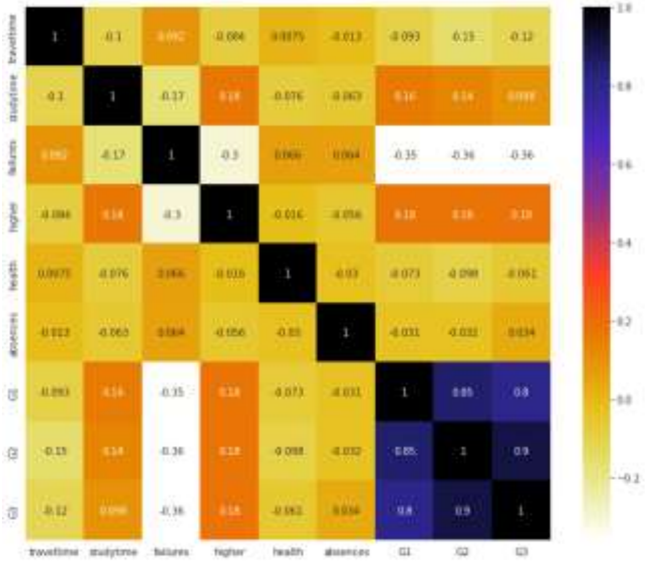
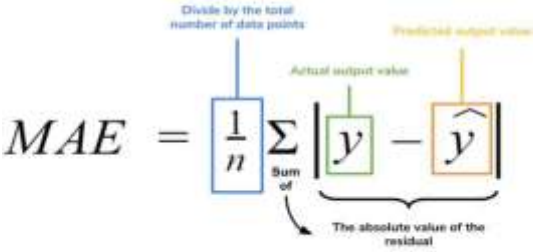


Fig.10: Heatmap showing correlation between all the mentioned features (attributes) and the label (target attribute)

Mean Absolute Error: MAE is the absolute difference between the target value and the value predicted by the model. MAE is a linear score which means all the individual differences are weighted equally.


Divide by the total number of data points
Predicted output value
Actual output value
MAE = 1/n Σ |y − ŷ|
Sum of
The absolute value of the residual

Root Mean Squared Error: RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

VI. CONCLUSION

We proposed a novel method for predicting students' future performance given their past performances & other real-life attributes. We believe our project will be able to help students in preparing for their next exams by letting them help in estimating their score. In general, it will aware the student about their performance.

The analysis of student's dropout during the very early stages of their levels is very interesting, as there are still many opportunities to research about helpful & predictive tools to enable the prevention mechanisms. In this sense, a good approach to research would really be to apply the same predictive techniques used for the academic performances to this case.

Mean Absolute Error & Root Mean Squared Error was minimum in the Gradient Boosted Algorithm which means that Gradient Boosted Algorithm is likely to give best accuracy with the selected attributes.

VI. REFERENCES

1. *Asaad Masood-” Predictive model for predicting students' performance using Machine Learning”, 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA).*
2. *Rafi Hasan H.M.- “Machine Learning algorithm for student performance prediction”,2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT).*
3. *Mehil B Shah- “Student Performance Assessment and Prediction System using Machine Learning”, 2019 4th International Conference on Information Systems and Computer Networks (ISCON).*
4. *Thai-Nghe Nguyen- “Matrix and Tensor Factorization for Predicting Student Performance”,2011 CSEDU-Proceedings of the 3rd International Conference on Computer Supported Education, Volume 1, Noordwijkerhout, Netherlands.*
5. *Sweeney Mack- “Student marks prediction using Factorization Machine”,2016 Journal of Educational Data Mining, Volume 8, No 1.*
6. *Behrouz Minaei-Bidgoli- “Predicting student performance: an application of data mining methods with an educational web-based system”,2003 Frontiers in Education, 2003. FIE 2003. 33rd Annual Volume: 1.*
7. *HAYAT SAHLAOUI- “Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations”, 2016 IEEE Access. VOLUME 4.*
8. *Jamjoom M. Mona “Early Prediction for At-Risk Students in an Introductory Programming Course Based on Student Self-Efficacy”, 2021 Informatica 45.*