Student Grade Prediction

Vruddhi Mehta, Rajasi Adurkar, Kriti Srivastava

Abstract*: Education is a dominating strand in accomplishing indelible economic progress. It is complex and nuanced. Grades outline the shape of our institutional system. It is the most powerful bargaining chip, at once cherished and dreaded by most students, the unyielding mallet of teachers and parents compressed into a single letter. However, the grading system is not an efficient way to gauge intelligence. The domains of Data Mining (DM) and Business Intelligence (BI) aim at deriving impactful insights from unprocessed data and propose techniques that can encourage a change in the education system. Our work plans to analyze students in secondary year of education using Business Intelligence and Data Mining techniques. These algorithms assist in finding patterns. It covers a broad scope of statistics, machine learning, and database systems. Past evaluations are influential in their performance. Insightful research shows that there are some other pertinent features (for example, department, age, romantic relations, outings, and goals). The methodology uses seven different algorithms and compares them to find the most suitable one. Visualizations help understand each factor thoroughly. As a result of this research, we can also analyze the reason behind a student's achievements. Each student faces several hurdles. The system should not focus only on improving student's grades but should also be concerned with the other aspects affecting their scores. The paper presents the research of the factors affecting the student's grades the most.

***Keywords* : Analysis, Education, Grade Prediction, Machine learning, Regression.**

I. INTRODUCTION

Today, the world has realized the importance of education in a person's life which, in turn, has led to a revolution in the field of education. The current education system typically instructs the academics and teaching personnel to use grades in evaluating students. The grading system is used to assess a student's educational progress in a particular subject and completely depends on the points alone. Some of the crucial purposes of the grading system are to provide feedback on the student's achievement and progress, to be used for various administrative purposes, and to provide guidance to the student concerning improvements in a particular area. For the supervisory objective, grades assist in informing us about a student's retention as well as serve as an indicator when students transfer from one school to another.

The remarkable progress in the field of data analytics has made it possible to reach the crux of any problem. Educators generally tend to only give out the grades and the areas where the student needs to work on. However, the poor performance of the pupil in that subject may be the result of other factors and not merely insufficient efforts. Factors such as economic conditions of the family, family dysfunction, improper time management, substance abuse, limited or no access to technology, etc. play a consequential role in affecting the academic prowess of a student. In this era, considerable attention should be invested by the educational institutions in not just improving a student's grade but also analysing the aspects which affect their grades.

With a vast amount of data available and advancement in the field of analytics, it is now possible for academic establishments to collect and visualize information. Further, meaningful insights and conclusions can be found out. Remarkable success in Smart Education or Learning these days has encouraged educators to find more ways in which the efficiency can be increased, and students can benefit maximum from their knowledge. Earlier classes required a teacher to be present physically, but now remote education is possible through online classes on the internet. Such facilities that technology has provided us with, must be exploited in order to gain maximum from them. Tremendous data that is being generated by the numerous educational institutions must be collected and stored properly so that through knowledge, interesting trends could be understood. This will help the academics and educators correctly identify their students' weaknesses and strengths, hence allowing them to construct efficient strategies to improve their students' performance. In this paper, we aim to analyse various economic and social attributes related to a student affecting his/her grade and further we try to find a machine learning algorithm to anticipate the future grade of a student. This paper is organized as follows: Section II describes the methodology being used, Section III provides a detailed analysis of the dataset as well as depicts the visualizations made in order to get a deeper insight into the data and the results obtained, finally, Section IV specifies the conclusion drawn from this work.

II. METHODOLOGY

A. Learning Methodology

The learning methodology used for this project is supervised learning also called Inductive learning. We are analyzing the ML models by training on the training set and testing on the testing set. Training data includes the desired output and therefore the learning is supervised. Since we are trying to predict the effect of different parameters/attributes on our primary target attribute, our input is the data(x) and the output is (f(x)).

Revised Manuscript Received on September 05, 2020.

***** Correspondence Author

Vruddhi Mehta*, Computer department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India. Email: vruddhimehta@gmail.com

Rajasi Adurkar, Computer department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India. Email: adurkar.rajasi562@gmail.com

Kriti Srivastava, Computer department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India. Email: kriti.srivastava@djsce.ac.in

The goal of the selected learning methodology is to learn the function of x. Based on the modelling approach chosen, the function being learned is continuous rather than discrete or probabilistic.

B. Modelling Technique

We can create a model for the dataset [4] in 3 ways

- Binary classification
  - G3 > 10: pass
  - G3 < 10: fail
- Five level classification
  - Between 16 and 20: very good
  - 14 or 15: good
  - 12 or 13: satisfactory
  - 10 or 11: sufficient
  - Below 9: fail
- Regression (Predicting target attribute)

Reasons for choosing the Regression model :

- Since one of the primary objectives of this project is to predict the behavior of the primarily targeted attribute regression model implies to be the best fit for the task.
- The regression model allows the user to perform predictive analysis with the utmost precision.
- Moreover, all the values in the primary target attribute are numerical values that are highly suitable for the regression model. Had the values for the primary target attribute been categorical, the classification model would have been a better fit.

1) Linear Regression

In order to identify relationships among variables, Linear Regression is one of the widely used predictive models. It tries to model the relationship between the input and target variables by applying a linear equation onto the data. The number of input variables tells us if the Linear regression is Simple or Multiple. Linear regression proves to be most useful if there exists a correspondence between the independent and the output variables. The prediction of the output variable is done using the association between two or more variables [5].

2) Elastic Net Regression

Elastic Net regression is a regularization and variable selection method which through simulation studies has proven to outperform the lasso [6]. Elastic fit is a good fit for this dataset as the number of predictors or input variables is more than the output or target variable. It is a mix between Lasso and Ridge Regression methods and the mix ratio can be controlled. This type of regression is efficient as it can reduce the weight of the useless features.

3) Random Forest Regression

In Random forests, also called “Regression Forests”, a horde of decision trees are built during the training period and the output is obtained by selecting the mean or mode of the individual trees [7]. It performs regression as well as classification tasks with the help of multiple trees and a statistical technique called bagging [7]. Its advantages include high precision, ability to handle a large number of input variables without deletion, and ability to estimate missing data accurately.

4) Extra Trees

Extra Trees, is also known as “An Extremely Randomized Trees”. It is a variant of a random forest. The framework is a decision tree-based ensemble learning method. They perform regression by fitting an assemblage of decision trees and bagging relevant features. We make use of the extra trees in our proposed methodologies as it is less susceptible to overfitting.[1] Extra trees work by forming several trees, where the complete original training set is used to construct each tree. The splits and features set at random. The average prediction of each tree is taken to obtain the aggregate prediction .[2] They exhibit low variance. In reality, the performance is equipotential to random forests.

5) SVM

The support vector machine applies to regression and classification problems. The goal is to discover a hyperplane in an N-dimensional area that enormously classifies the recorded data points. Support vector regression is a part of SVM which applies to regression. The only difference is the direct outcome of SVR is a real number. SVR tries to minimize the error while limiting the margin violations.[3] The hyperparameter Epsilon controls the best fit margin. We can think of SVR as if each data point in the training represents its dimension. When you evaluate your kernel between a test point and a training point, the resulting value gives you the coordinate of your test point in that dimension. The vector we get when we evaluate the test point for all the points in the training set, 'k' is the representation of the test point in the higher dimensional space. Regression uses this vector for implementation.

6) Gradient Boosting

Gradient Boosting is an effective and interpretable machine learning algorithm. Learning takes place by optimizing the loss function. It mainly consists of trees with leaf nodes with a range from 8 to 32. The main idea is to have a model, and we have multiple versions of that model chained together. So each tree is going to boost the attributes that led to misclassifications from the previous trees. It uses regularized boosting to ensure that the model is generalized and prevents overfitting. The technique can handle missing values automatically and can process data parallelly. [8] There are two novel techniques used for achieving our goal : exclusive feature bundling and gradient-based one-side sampling. It helps the algorithm to deal with big data with multiple feature sets[2].

III. EXPERIMENTAL SETUP AND RESULTS

A. Dataset

The dataset contains the performance of students. It is obtained from two public schools, from the Alentejan region of Portugal. The data attributes include social parameters, demographic details, personal choices, and more. It is collected using school reports and a series of questionnaires. The data is for the period 2005-2006. The dataset is regarding the performance of students in two distinct subjects: Mathematics and Portuguese language. The figure given below depicts all the attributes with an explanation

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	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 $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
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)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday 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)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Fig. 1. Pre-processed student related variables [4]

In order to get more clarity correlation analysis was performed. Correlation analysis yields a correlation coefficient between a source attribute and a target attribute. Correlation coefficient ranges from -1 to +1, where -1 represents high negative correlation, 0 represents no correlation and +1 represents high positive correlation. The Correlation analysis for the dataset is represented below.

Now that the correlation analysis is completed. Although G1 and G2 which are period grades of a student and are highly correlated to the final grade G3, we drop them. It is more challenging to predict G3 leaving out G2 and G1, but such predictions are much more favourable because we want to find other factors that affect the grade.

The column "Higher education" was a categorical variable with values yes and no. Since we used one hot encoding it has been converted to 2 variables. So, we can safely eliminate one of them (since the values are compliments of each other).

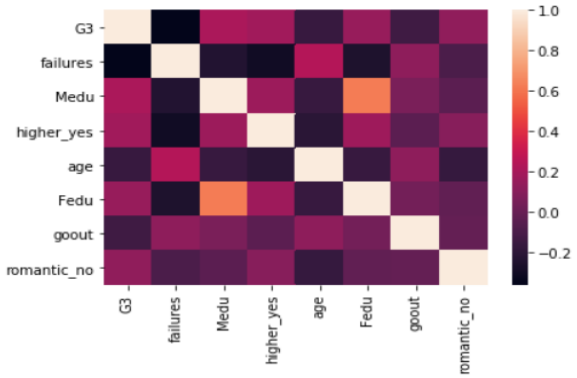

Fig. 2. Correlation heat map

We will eliminate higher_no, since higher_yes is more intuitive.

B. Analysis

We analyze the effect of various attributes on the primary target attribute, i.e. the final grade. The results obtained in the various plots as depicted below give us an insight into the factors(demographic, social, and cultural) which might be responsible for the high, average, or low student grades.

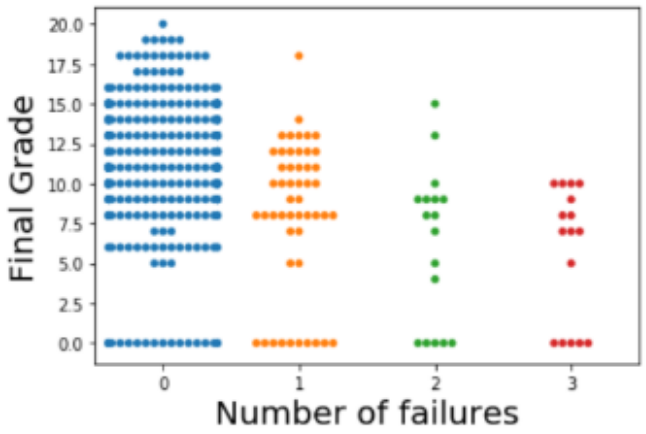

Fig. 3. Swarm plot for Number of Failures v/s Final Grade

Final grade was plotted against the number of failures and a swarm plot was obtained. We can see that students who had failed in one or more subjects previously had low grades than the students who did not have any previous failures. Failure shows the strongest negative correlation with the grade.

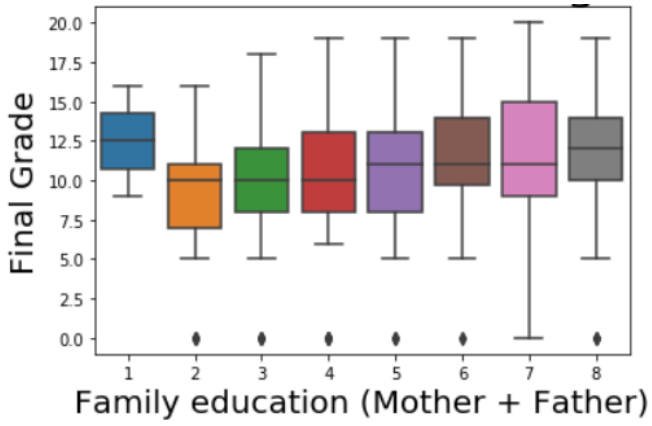

Fig. 4. Box plot for Family Education v/s Final grade

Mother's and father's education are scored on a scale of 0 to 4 each (0-No education, 1-completed primary school, 2-completed middle school, 3-completed high school, 4-completed graduate studies). The scores are added and plotted against the final grade. We conclude from the box plot obtained that students whose parents have completed a higher degree of education are likely to score more in their exams.

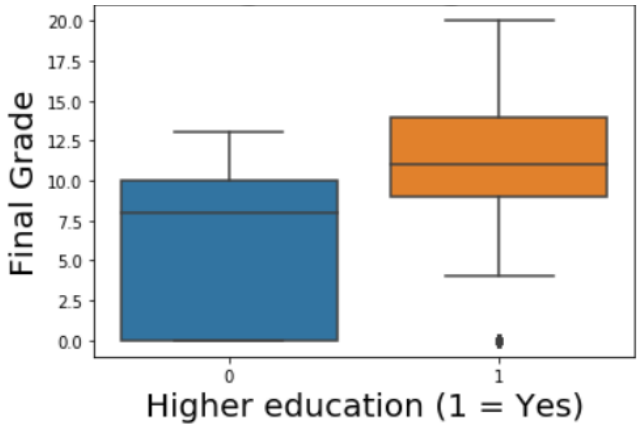

Fig. 5. Box plot for Higher education v/s Final grade

From the box plot shown in figure 5, we see that students who aspire for higher studies or post graduate degree are likely to score more. The students are more focused when they have their goals set Their intent to pursue studies keeps them motivated. As the graph shows, there are some outlier possible. The student may aim for higher education, but due to some other factors might not score well.

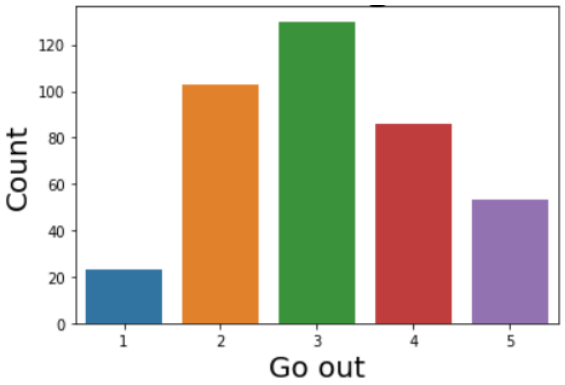

Fig. 6. Bar plot for frequency of students going out

Students' social outings are scored from 1 to 5 (1 being the lowest and 5 the highest), and it is found that most students have an average score when it comes to going out with friends. The graph is normally distributed. The count of students going out is more. It can be stated that going out can freshen up your mood and release the stress. This helps a student perform better in exams. However, extremes are always dangerous.

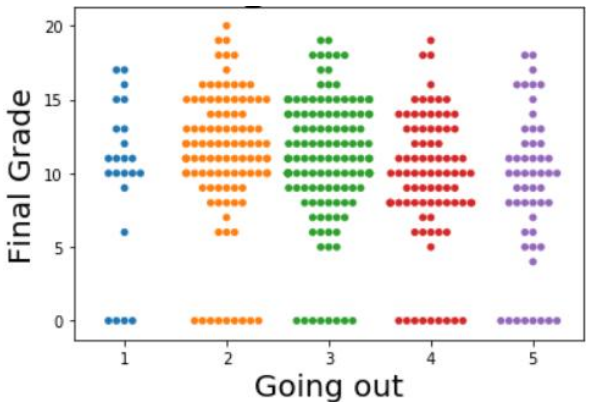

Fig. 7. Swarm plot for Going out v/s Final grade

Further, the frequency of going out is compared with the final grade and we observe that it is inversely proportional to the final grade. Students who spend a lot of time in social outings or gatherings are likely to score less in their final term exam. The lack of focus and time in studying reflects on the grades of the student

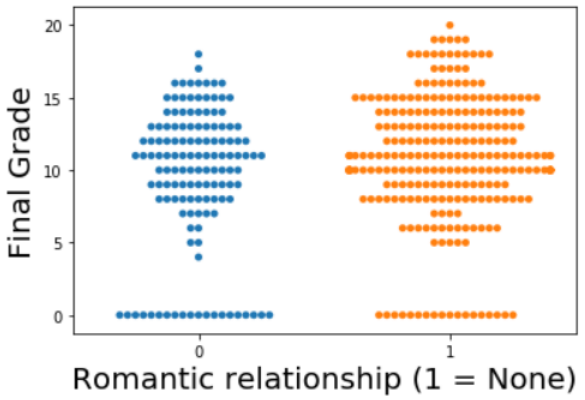

Fig. 8. Swarm plot for Romantic Relationship status v/s Final grade

During school days, being in a relationship is common. It can affect some aspects of life positively and negatively. From our analysis, we can state that the students not engaging in a relationship have better grades. The figure above clearly demonstrates the same. The graph also shows that the number of students not engaging in a relationship is more.

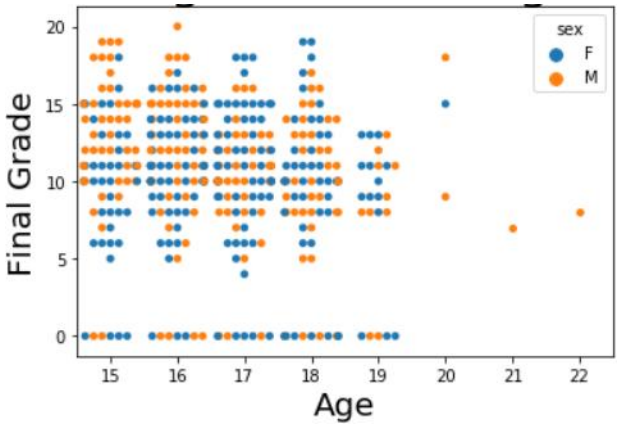

Fig. 9. Swarm plot for Age v/s Final grade

Students have greater grasping potential when they are young. The graph above demonstrates the same. Studying at the right age is very important. Students younger than their batch mates find it more difficult to get good grades. The graph shows strong positive correlation with the grades. The students below the age of 19 have a higher chance of scoring good on final exams. The graph also points out a minor detail the students who are studying at the age of 20 and above. The ratio of the men to women studying is higher. Men are more likely to complete their education even in late stages of their life.

C. Grade Prediction

TABLE-I: Mean Absolute Error and Root Mean Squared Error values for six regression models

MODEL	MAE	RMSE
Linear regression	3.51289	4.45104
ElasticNet regression	3.61061	4.57647
Random Forest	3.7981	4.83544
Extra Trees	3.89519	5.05534
SVM	3.58885	4.60437
Gradient Boosted	3.60464	4.48663
Baseline	3.78788	4.82523

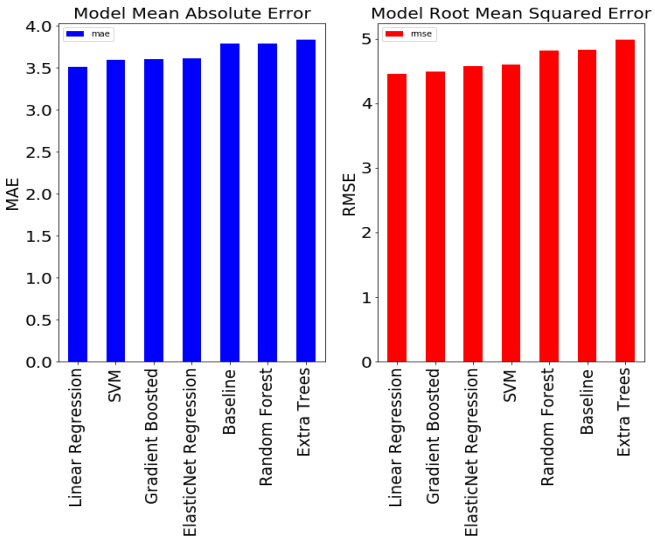

Fig. 10. Comparison of MAE and RMSE values for all models

All the six regression models explained in the methodology are fit into the selected dataset and their Mean Absolute Error (MAE) and Root Mean Squared Error(RMSE) are calculated. From fig. 10, as well as fig. 11, we see that linear regression is the best fit for this dataset as it gives lowest MAE and RMSE values. Further, after finding out the R2 value of these values, we find that linear regression has the highest value, nearly 0.1400, hence is the best suited for this dataset.

IV. CONCLUSION

In this paper we have tried to find out if various demographic, social and economic factors affect a student's academic grade, and if they do, does it have a positive or negative influence on their scores. We see that attributes like romantic relationship status of a student, family's education, frequency of social outings and desire to pursue higher education have a strong influence on the final grade. However, attributes like age, gender, and geographic area of settlement (urban or rural) seem to have no clear relation to the final grade. Further, from the results obtained from the analysis, MAE and RMSE for the linear regression is the least. Which means that the output obtained from analysis on the training dataset matches the predictive analysis performed on the test dataset most accurately on the linear regression model. Moreover, the r-squared value for the linear regression model was highest relative to the other regression model. Therefore, it can be concluded with high confidence that Linear regression model is the best approach to do the following:

- Finding the relation between the varied parameters and the target attribute
- Predict the behavior of the primary target attribute based on the defined parameters.

Thus, we see that data mining and machine learning methodologies are a boon, in today's world, and can be exploited to their maximum extent to get meaningful insights from data. Grading system paired with a system to analyze the students, as depicted in the paper, could be beneficial to the academic professionals as well as the students.

REFERENCES

1. V. John, N. M. Karunakaran, C. Guo, K. Kidono and S. Mita, "Free Space, Visible and Missing Lane Marker Estimation using the PSINet and Extra Trees Regression," 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, 2018, pp. 189-194, DOI: 10.1109/ICPR.2018.8546108.
2. B. V. Mbuwir, F. Spiessens, and G. Deconinck, "Benchmarking regression methods for function approximation in reinforcement learning: heat pump control," 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), Bucharest, Romania, 2019, pp. 1-5, DOI: 10.1109/ISGTEurope.2019.8905533.
3. S. S. Arun and G. Neelakanta Iyer, "On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, In- dia, 2020, pp. 1222-1227, DOI: 10.1109/ICICCS48265.2020.9121027.
4. P. Cortez and A. Silva. "Using Data Mining to Predict Secondary School Student Performance." In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April 2008, EUROSIS, ISBN 978-9077381-39-7.
5. U. Gülden and N. Güler, "A Study on Multiple Linear Regression Analysis." Procedia - Social and Behavioral Sciences. 106. 234–240. 10.1016/j.sbspro.2013.12.027.
6. H. Zou, and T. Hastie. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) 67, no. 2 (2005), pp. 301-320.
7. L. Breiman, "Random Forests.", *Machine Learning* 45, 5–32 (2001). https:///doi.org/10.1023/A:1010933404324.
8. T. Chen and C. Guestrin. " Xgboost: A scalable tree boosting system" . In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

AUTHORS PROFILE

Ms. Vruddhi Mehta has completed her Diploma in Computer Engineering from Shri Bhagubhai Mafatlal Polytechnic, Mumbai, Maharashtra, with distinction. She is pursuing a BE in Computer Engineering from Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra. She has published various papers in the fields of blockchain, databases, and cybersecurity in reputed journals. Her area of interest lies in artificial intelligence and cybersecurity.

Ms. Rajasi Adurkar is currently in her 4th year of computer engineering. She is pursuing Bachelor of Engineering degree from Dwarkadas J. Sanghvi College of Engineering, Mumbai, India. She has done a lot of resprojects in analytics, machine learning and software development arena. Recently she published a research paper in big data and analytics domain that proposed the use of newer technologies like NoSQL and XML for clinical data storage. Data analytics, business analytics and machine learning are her main area of interests.

Mrs. Kriti Srivastava is an Assistant Professor in the Computer Engineering department of Dwarkadas J. Sanghvi College of Engineering, Mumbai, India. She has an M. Tech degree in Computer Engineering and her main are of interest is Machine Learning and Artificial Intelligence.