Educational Grade Prediction System Using Machine Learning Algorithm

[1]Kalyani Darekar, [2]Shweta Khilari, [3]Hritik Shirsath, [4]Shweta Mate, [5]Prof. J. V. Borase

[1,2,3,4]Student, [5]Professor
Department of Computer Engineering,
PVG's COE & SSDIOM, Nashik, Maharashtra, India

Abstract- **Predictive analytics applications are highly desired in higher education. Predictive Analytics used advanced analytics, including applications of machine learning, to deliver high-quality performance and valuable information at all academic levels. Researchers have developed many variations of machine learning approaches in education over the past decade. This system enables in-depth study of machine learning approaches for predicting students' final grades in the first-semester course by improving prediction accuracy. The proposed system emphasizes two components. First, using a dataset of real student course performance, we tested six well-known machine learning algorithms including Decision Trees (J48), Nave Bayes (NB), Logistic Regression (LR), and Random Forest (RF). Evaluate the accuracy of the approach's performance. Next, to avoid overfitting and misclassification results resulting from imbalanced multiple classifications, we provide a multiclass prediction model that can improve the unbalanced multiple classification prediction performance model for predicting student performance.**

Keywords—**machine learning, algorithms, educational data mining, predictive modeling, unbalanced problems, and predicting student grades.**

I. Introduction

Educational institutions are an important part of our society and play an important role. It plays an important role in the growth and development of any country. Educational data mining is an application of data mining. An emerging interdisciplinary research area deals with the development of methods for investigating data originating from the educational system context. Data mining in education is a new trend aimed at automation. Explore unique data types from a large repository of educational data. This data is often large, fine-grained, and precise. The main goal of the project is to study student performance using a machine learning models course. ML Technologies offers many tasks that you can use to study student performance. The report uses a regression task to assess student performance. Since many approaches are used for data regression techniques such as linear regression, decision trees, etc. are used here. In the project, the accuracy of regression methods for predicting student performance. Faculty discovering and developing student skills and interests is not an easy task in it. This can affect bad grades in college, bad jobs, and bad careers. Individually. As a result, it helps to achieve the company's mission and vision institute. If the project succeeds, it will be of great help to teachers and improve the education system.

II. MOTIVATION

The education system is getting closer to digitization. Therefore, adapting to the new regulations, machine-learning models drive student academic development to drastic change.

III. OBJECTIVES

- Learn machine-learning algorithms.
- Find effective algorithms for data science.
- To improve module efficiency.
- To improve system accuracy.

IV. LITERATURE REVIEW

Multiclass Prediction Model for Student Grade Prediction Using Machine Learning- There is an urgent need for predictive analytics applications in higher education institutions. Predictive analytics uses advanced analytics across multiple machines. Introduce learning that brings quality performance and meaningful information to all levels of education. Most people know that student grades are one of the key performance indicators that help educators monitor academic performance. Over the past decade, researchers have proposed many variations of machine learning methods for education. However, there are significant challenges in improving the performance of predicting student grades by processing imbalanced datasets, so this application focuses on machine learning for predicting student final grades in freshman courses through improvement. Provides a comprehensive analysis of technology. Predictive accuracy performance. [1]

The Prediction of Students' Academic Performance Using Classification Data Mining Technique - Data mining provides a powerful technology for various fields, including education. Education research is growing rapidly thanks to the vast amount of student data that can be used to discover valuable patterns in student learning. This application proposes a framework for predicting student numbers. Academic performance of undergraduate students in computer science courses. Data were collected from her 8-year record from July 2006/2007 to July.2013/2014 including student demographics, previous academic performance, etc. Background information about the family. Decision trees, naive Bayes, and rule-based classification techniques are applied to student data to get the best results. A model for predicting student academic performance. Test results show that Rule-based is the best model among other techniques because it gives the best values. The accuracy score is 71.3. [2]

Improving Academic Performance Prediction by Dealing with Class Imbalance - The application presents and compares a few procedures utilized to foresee understudy execution at the college. As of late, analysts have centered on applying machine learning in higher instruction to back both the understudies and the teachers in getting superior in their exhibitions. A few past papers have presented this issue but the forecast comes about as unsuitable since of the course awkwardness issue, which causes the corruption of the classifiers. The reason for the application is to handle the lesson lopsidedness by progressing the prediction/classification comes about by over-sampling procedures as well as utilizing cost-sensitive learning (CSL). The paper appears that the comes about has been moved forward when comparing as it was utilizing pattern classifiers such as Choice Tree (DT), Bayesian Systems (BN), and Back Vector Machines (SVM) to the initial information sets.[3]

V. PROPOSED SYSTEM

The proposed system is a supervised machine-learning model that is being trained on labeled data and produces results on a training basis. A dataset containing student information is used for training. The dataset contains linked tuples to student privacy and academic data. The machine-learning module is built to predict student performance by tracking their studies and extracurricular activities. The module will use different machine learning algorithms such as random forest, decision tree, linear regression, slope regression, and noose regression to predict students' grades based on their recent grades.

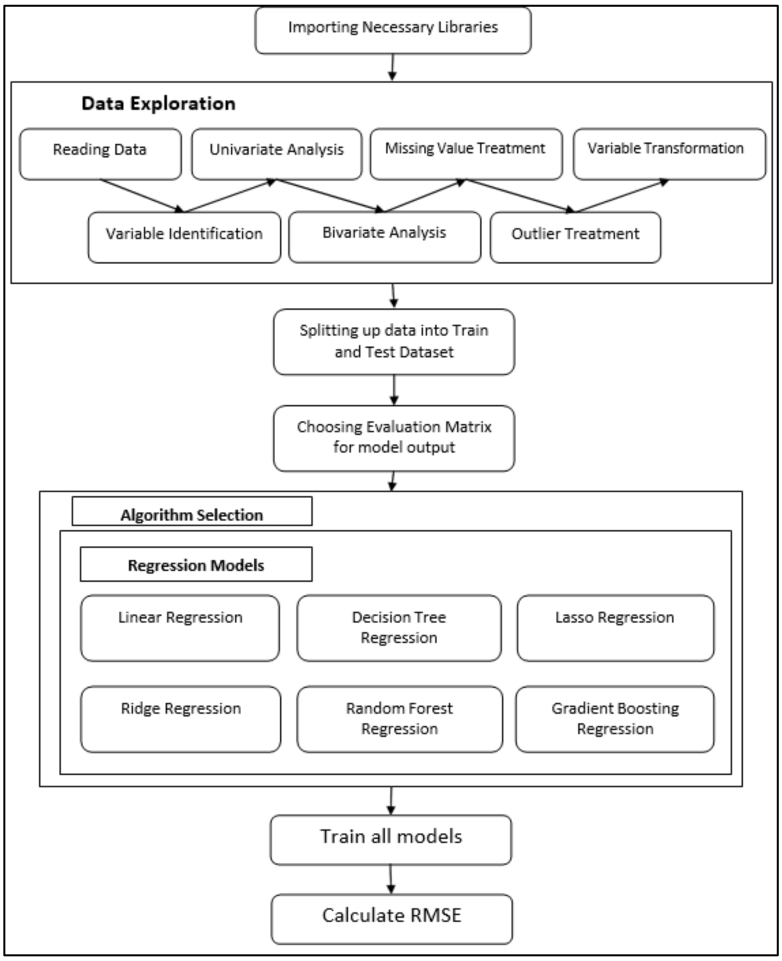


Fig.1. System Architecture

VI. METHODOLOGY

Regression Analysis: Regression models predict continuous values by modeling the relationship between dependent and independent variables. Regression analysis predicts real values like temperature, age, salary, price, etc. by examining how the dependent variable changes with an independent variable while keeping other independent variables constant. Please condense this text.

Bias: Bias simplifies the objective function, improving generalization and reducing sensitivity to individual data. Less training time with simple objective functions. High bias means more assumptions in the objective. This may cause underfitting. Biased algorithms: linear regression, logistic regression. Please condense this text.

Variance: In ML, Variance is when a model is sensitive to small data fluctuations causing errors, such as modeling outliers or noise in the training set. High variance algorithms: Decision Tree, KNN, etc.

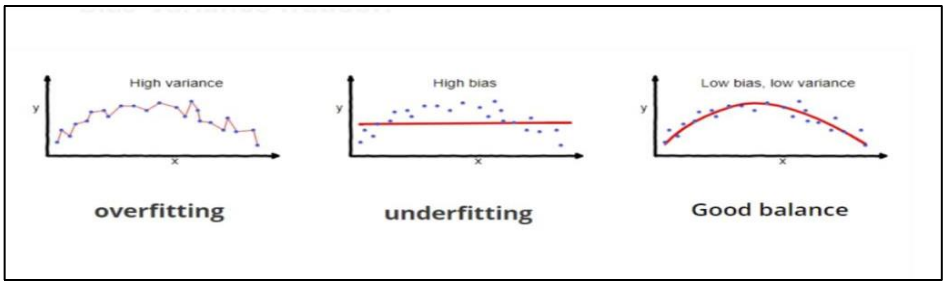


Fig.2. Regression Analysis

Types of Regression: Different regression methods are used in data science and machine learning to analyze the effect of independent variables on dependent variables. Please condense this passage.

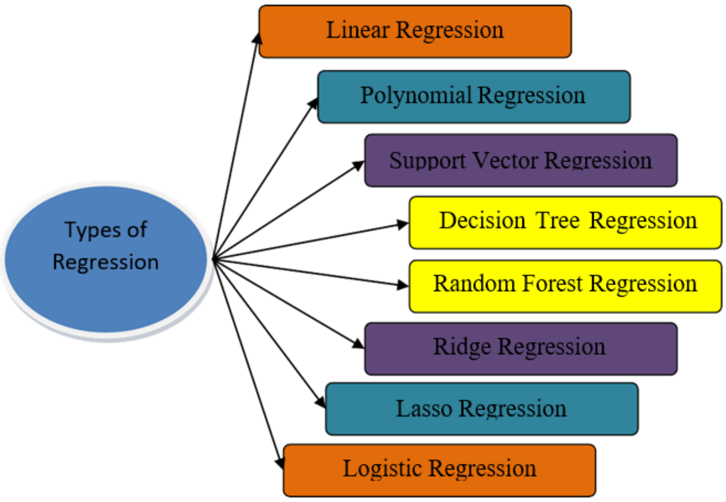


Fig.3. Types of Regression Algorithm

VII. IMPLEMENTATION

Training data has 4000 rows while the test data has 2000 rows with empty dependent variable columns. We split our training data into two sets: training and validation.

Data Representation and Exploration:

school - student's school ()
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 â€“ 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 – a reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian - student's guardian (nominal: 'mother', 'father' 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<=n<3, else 4)
schoolsup - extra educational support (binary: yes or no)
famsup - family educational support (binary: yes or no)

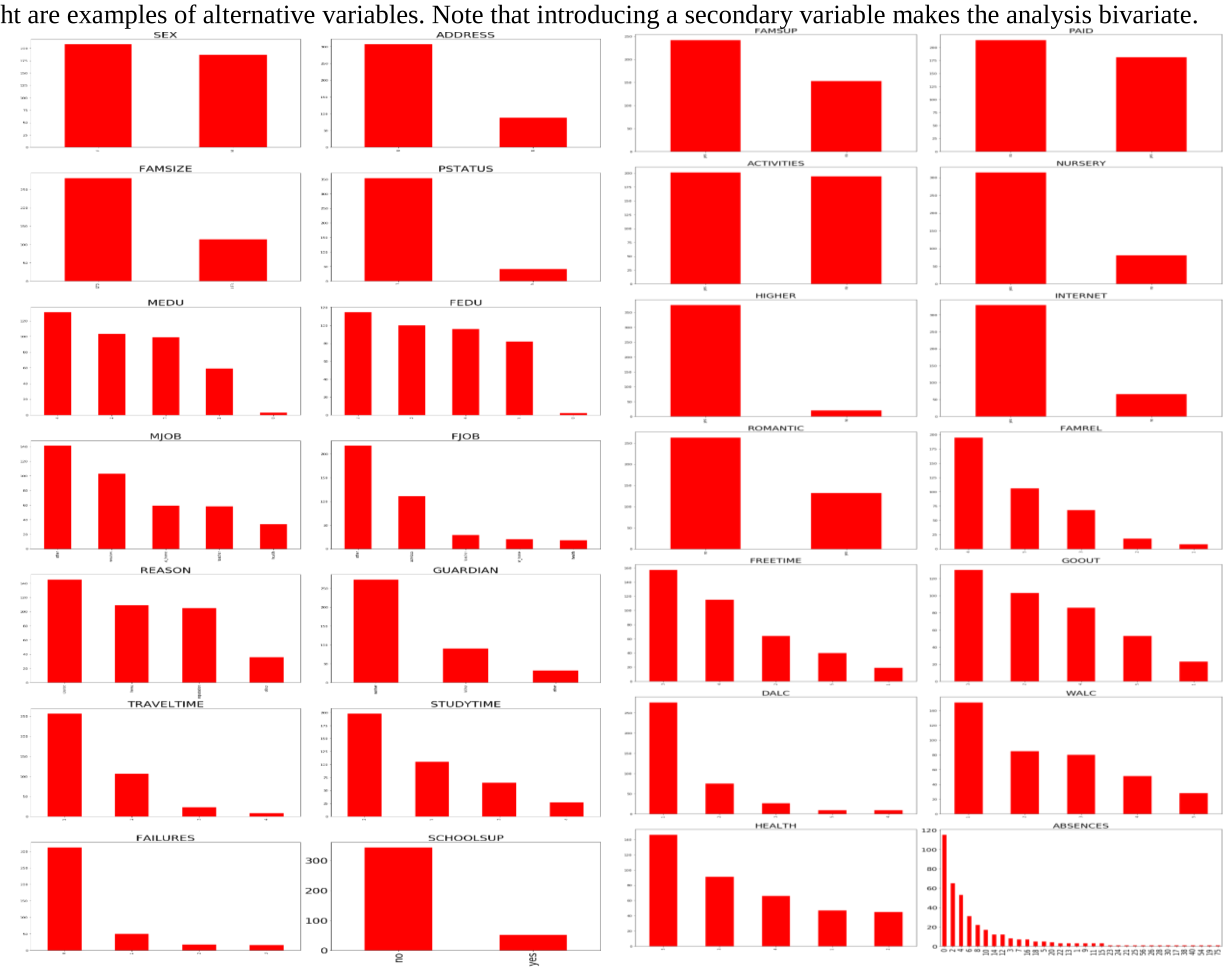
paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities - extra-curricular activities (binary: yes or no) nursery - attended nursery school (binary: yes or no) higher - wants to take higher
education (binary: yes or no) Internet - Internet access at home (binary: yes or no)

romantic - with a romantic relationship (binary: yes or no)
famrel – the 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)

# These grades are related to the course subject, Math or Portuguese:
G1 - first unit marks (numeric: from 0 to 20)
G2 - second unit marks (numeric: from 0 to 20)
G3 - final marks (numeric: from 0 to 20, output target)

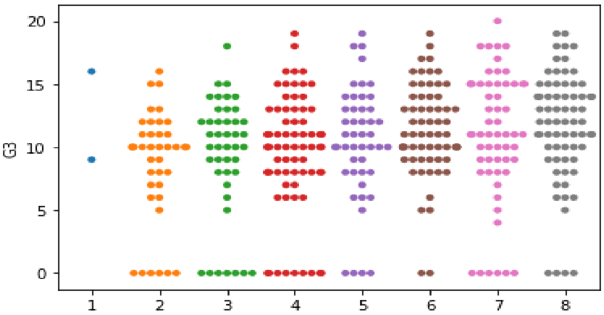
VIII. DATA VISUALIZATION

Univariate Analysis: Univariate analysis analyzes data where there is only one variable. Univariate Analysis describes data and finds patterns, without exploring causes or relationships like regression. It focuses on examining the effects of one variable on a set of data. A frequency distribution table is a univariate analysis that measures only frequency, while other variables like age, height, and weight are examples of alternative variables. Note that introducing a secondary variable makes the analysis bivariate.



Bivariate Analysis:

a. Compare family (father & mother) education with G3



Students having good family education backgrounds are performing better.

b. Compare travel and study time with G3

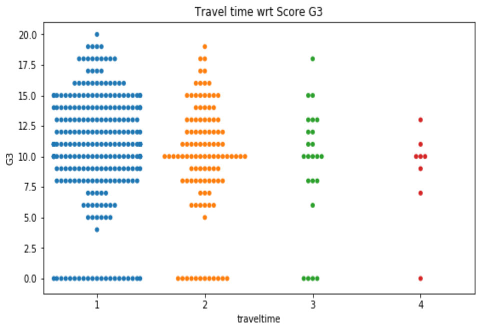


Fig.4(a). Travel-time vs G3

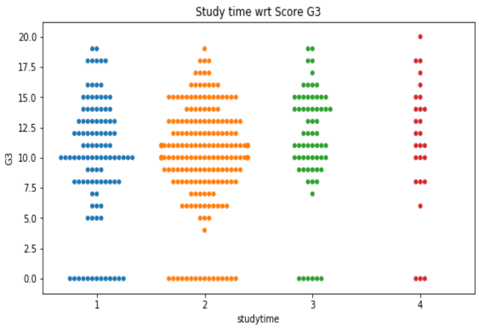


Fig.4(b). Study-Time vs G3

Fig. 4(a). Students living near the school scored better than the far students.

Fig. 4(b). Students having more weekly study time (over 10hrs) are getting a better score

c. Compare Other Features

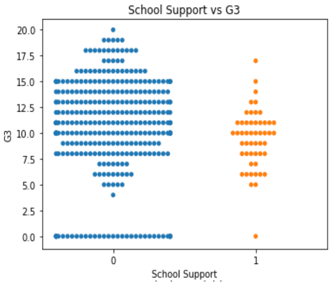


Fig.5(a). School Support vs G3

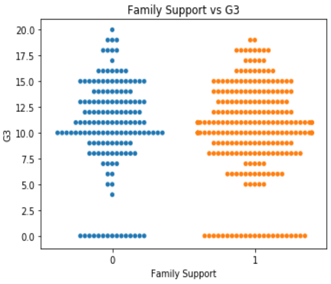


Fig.5(b). Family Support vs G3

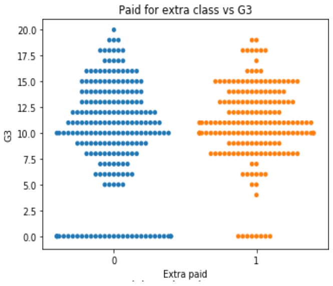


Fig.5(c). Extra Paid vs G3

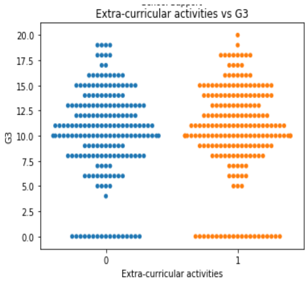


Fig.5(d). Extra Curriculum vs G3

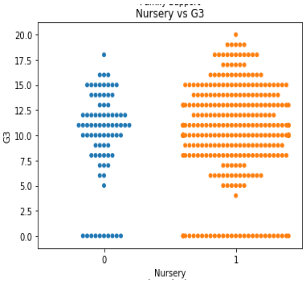


Fig.5(e). Nursery vs G3

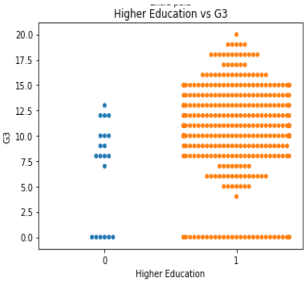


Fig.5(f). Higher Edu. vs G3

Fig. 5(a). Students who don’t have school support are showing a negative trend.Fig. 5(c). Students who paid for extra classes are showing a negative trend.
Fig. 5(e). Students who went to nursery schools are performing better.
Fig. 5(f). Students who proceed with higher education are performing better.Fig. 5(g). Students having internet access are performing better.
Fig. 5(h). Students having no romantic relations are performing better.

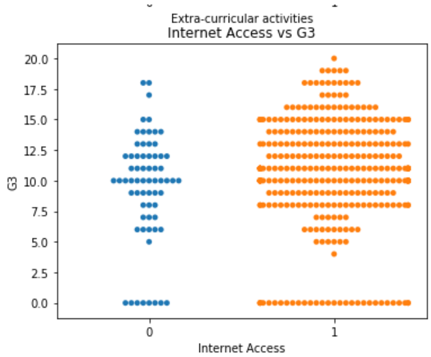


Fig.5(g). Internet Access vs G3

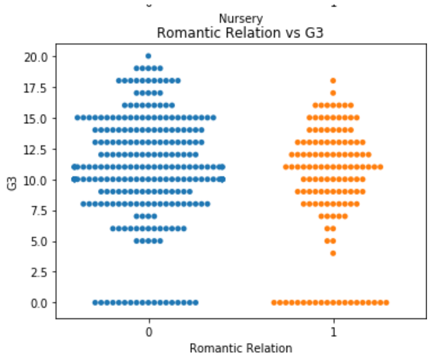


Fig.5(h). Romantic Relation vs G3

d. Checking Alcohol Consumption

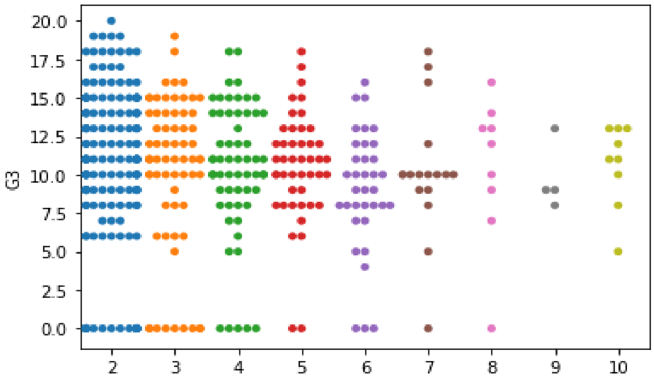
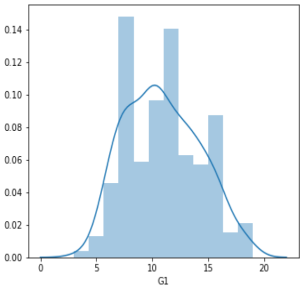
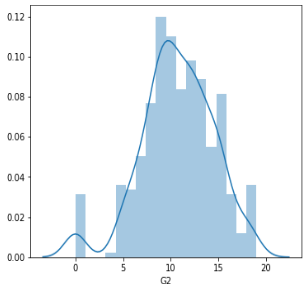


Fig.6. Alcohol vs G3

Students having more alcohol consumption has performed very poorly.

e. Probability Distribution of Marks





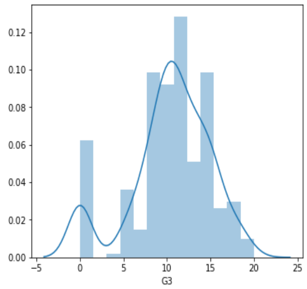


Fig.7(a). G1 Marks Fig.7(b). G2 Marks Fig.7(c). G3 Marks

The above figures show, the probability distribution of Marks feature i.e. G1, G2 and G3

f. Compare G1 and G2 with G3

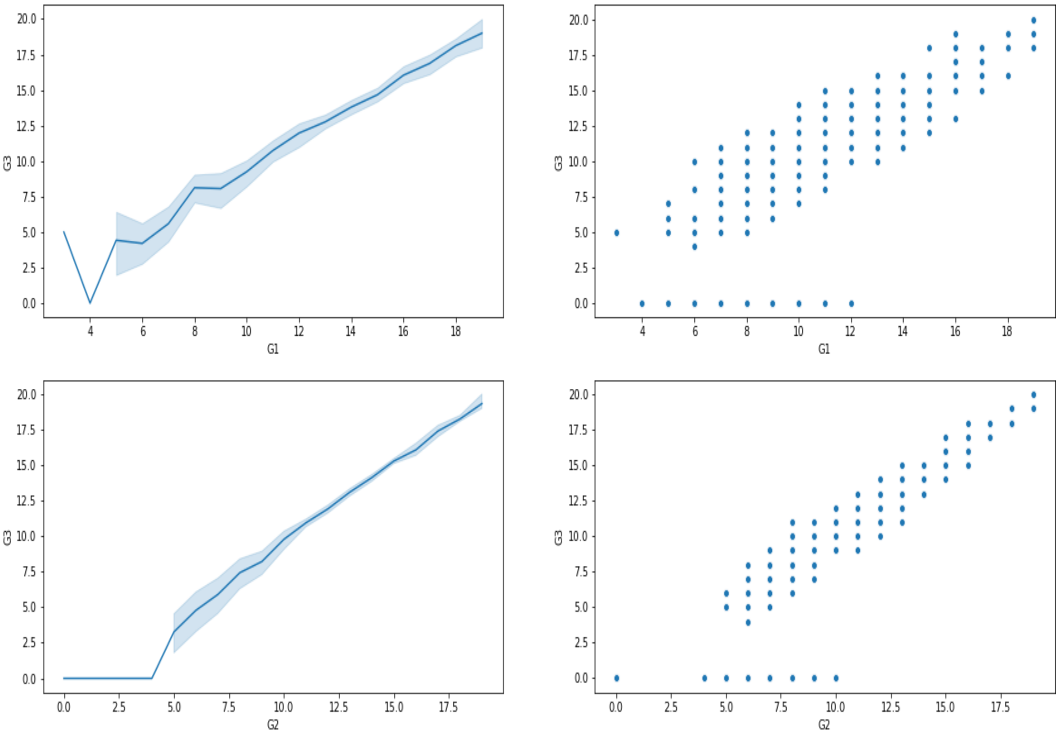


Fig.8. G1 vs G3 and G2 vs G3

In the above fig., we have visualized a comparison between G1 and G2 with G3

g. Student Count

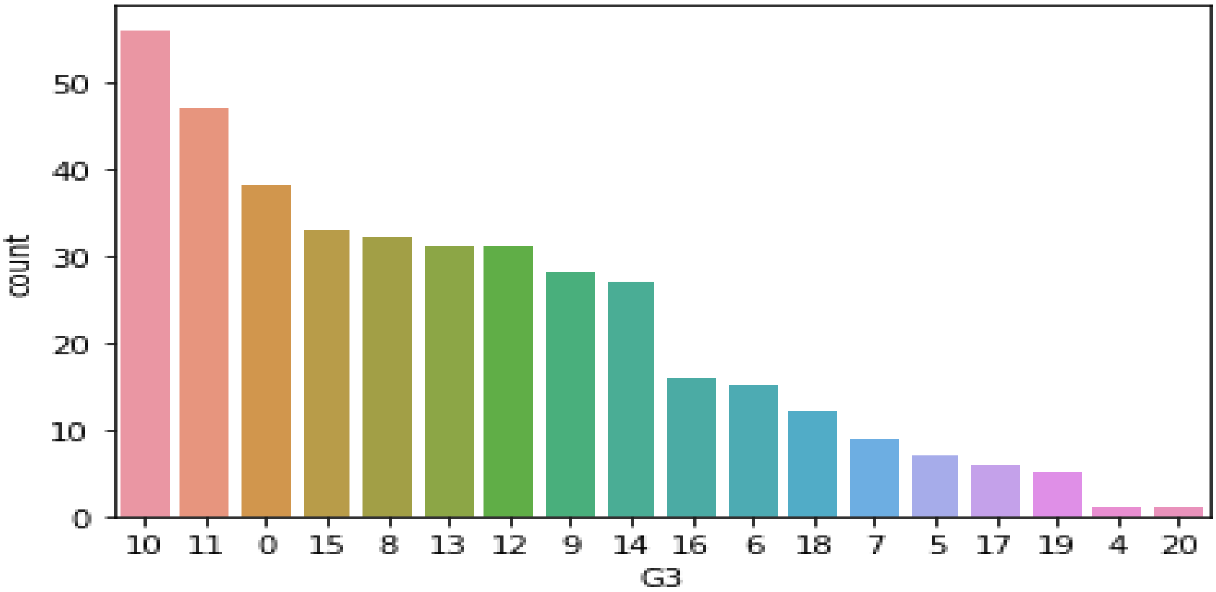


Fig. 9. Student Count

In the above fig., we have displayed student marks distribution as per G3 marks.

Feature of Engineering: In feature engineering, we convert the text values to numerical values to be readable by Machine Learning Algorithms.

Covert to Numerical: ML models need numeric input/output variables. Need to encode categorical data before fitting/evaluating the model. We utilized Ordinal Encoding in this project instead of One-Hot Encoding, which is the two most commonly used techniques.

In this project, we used Ordinal Encoding Method as follows:
{'yes': 1, 'no': 0},
{'F': 1, 'M': 0}
{'U': 1, 'R': 0}
{'F': 1, 'M': 0}
{'U': 1, 'R': 0}

{'LE3': 1, 'GT3': 0}
{'T': 1, 'A': 0}
{'teacher': 0, 'health': 1, 'services': 2, 'at_home': 3, 'other': 4}
{'home': 0, 'reputation': 1, 'course': 2, 'other': 3}
{'mother': 0, 'father': 1, 'other': 2}

Feature Selection
There are 33 columns in the dataset and this is going to be cumbersome for trainingour model. Let's train with the best 15 columns only
From the plot, absences, G2, G1, and failures are having best scores.

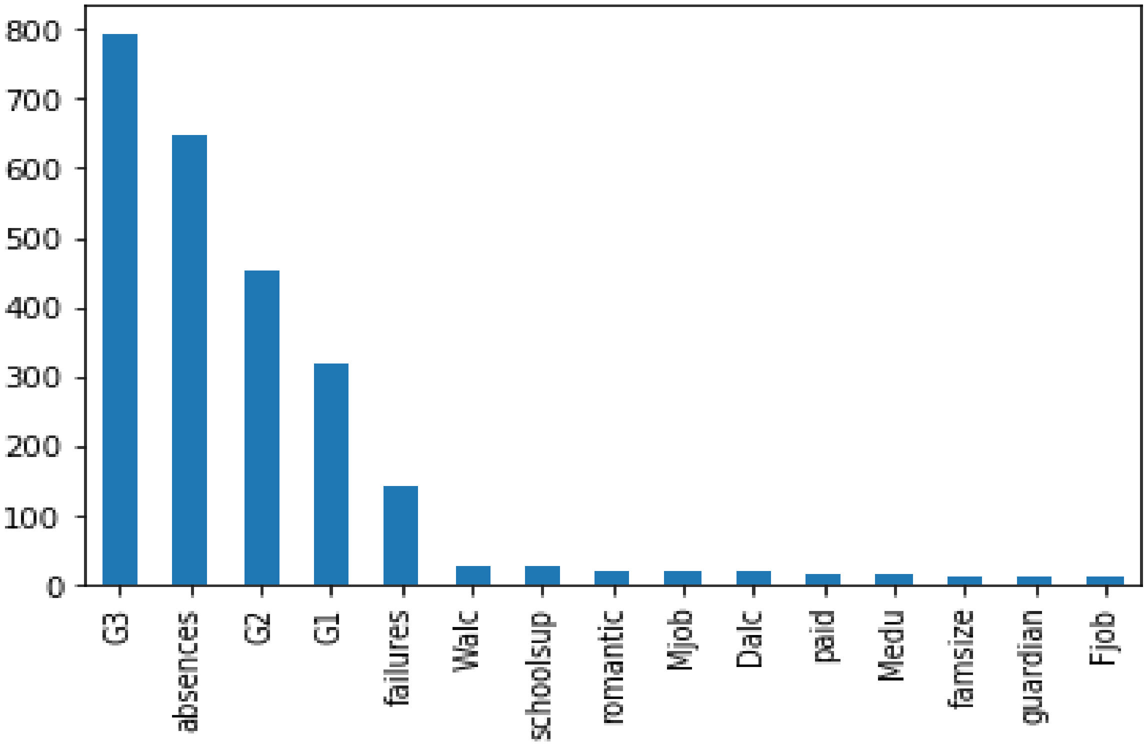


Fig.10. Most affected Feature

From the feature selection score, let us define the independent variable i.e. x.

Heatmap Correlation: Correlation is a measure of how strongly one variable depends on another.Correlation can be an important tool for feature engineering in building machine learning models.
From the heatmap plot, G1 & G2 are highly correlated with G3

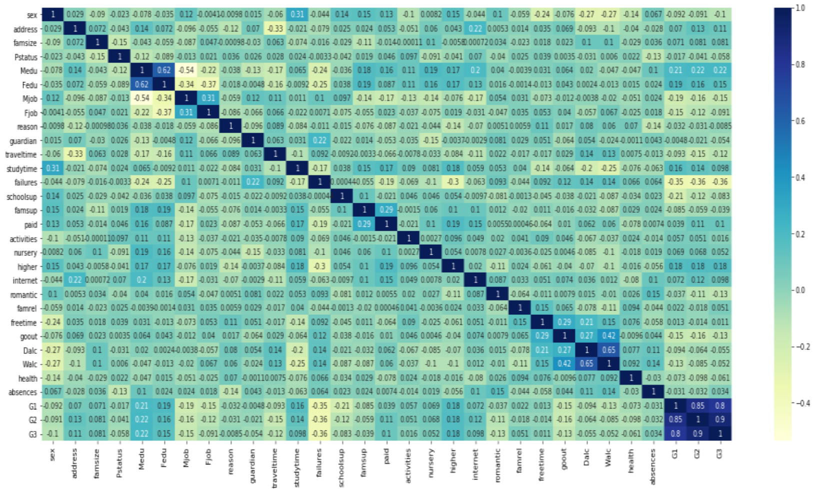


Fig.11. Heatmap Correlation

Splitting data into Train and Test

As we mention in the feature selection, we have used the best 15 columns to reducecumbersome which causes a problem during model training.

We have X_train, X_test, y_train, y_test splits with k_fold 10.

IX. RESULTS

All the models are implemented over a pipeline of cross-validation of 10 folds. The pipeline included steps of data pre-processing and model evaluation to achieve the best hyper-parameters and scores for the data.

Following Regression models have been implemented to estimate the marks of a student.

- Linear Regression
- Ridge Regression
- Lasso Regression
- Random Forest Regression
- Gradient Boost Regression
- Decision Tree Regression

Our experiments with various algorithms, but Random Forest Regression have proven to be successful giving the least error rate (31% test error) when compared to the other Regression models like Lasso and Ridge.

Model	Train RMSE Error	Test RMSE Error
Linear Regression	42%	39%
Lasso Regression	99%	99%
Ridge Regression	42%	39%
Random Forest Regression	16%	31%
Gradient Boost Regression	22%	33%
Decision Tree Regression	28%	34%

Table 1. Overall RMSPE

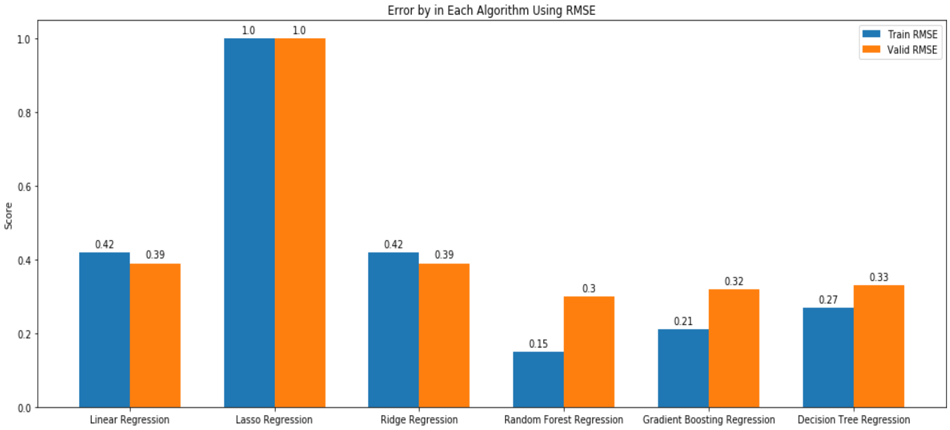


Fig.12. RMSE with Actual Error

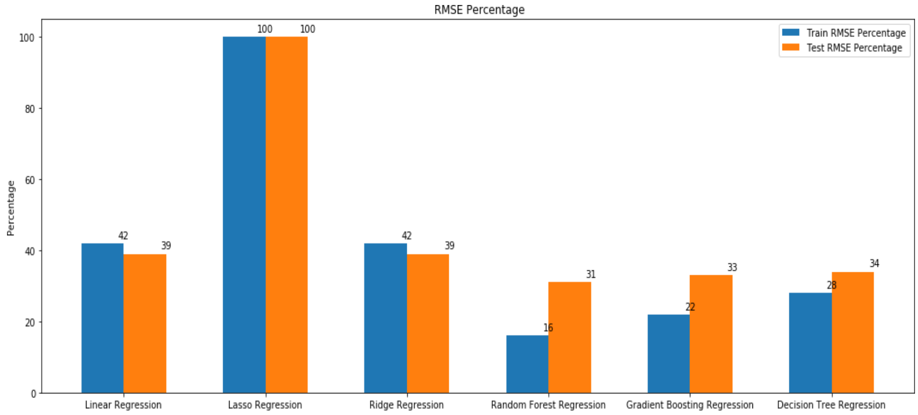


Fig.13. RMSE Percentage

Above figs., shows that Random Forest has the lowest error compared to Lasso and Ridgeregressions

First fig,18 shows error values in the RMSE calculation format. i.e. Root MeanSquared Error.

Fig. 19. briefly shows the calculated error output in percentage i.e. Root MeanSquared Percentage Error

X. CONCLUSION

Despite initial expectations, the variables Dalc and Walc, which represent workday and weekend alcohol consumption, were found to be statistically insignificant in student grades. Interestingly, all of the classification algorithms used in the study performed similarly in terms of accuracy, with Logistic Regression performing the best at 96. However, Principal Component Analysis did not make any significant difference to the Logistic Regression model. Notably, the use of Principal Component Analysis did not have a significant effect on the Logistic Regression model. The study results revealed that the variables Dalc and Walc, which represent workday and weekend alcohol consumption, did not have a significant impact on student grades. Interestingly, the inclusion of Principal Component Analysis did not have a significant impact on the Logistic Regression model.

REFERENCE:

1. “Predicting performance and potential difficulties of university students using classification: Survey Paper” by D Solomon, S Patil, and P Agarwal –2018
2. “Early segmentation of students according to their academic performance: A predictive modeling approach” by V L Migueis, A Freitas, P J V Garcia, A Silva – 2018
3. “Tracking student performance in introductory programming using machine learning” by I Khan, A Al Sadiri, A R Ahmad, N Jabeur – 2019
4. “Prediction of student’s performance by modeling small dataset size” by L M Abu Zohair –2019
5. “Educational data mining and learning analytics for 21st century higher education in A review and synthesis” by H Aldowah, H Al-Samarraie, W M Fauzy – 2019
6. “Educational data mining and learning analytics for 21st century higher education: A review and synthesis” by H Aldowah, H Al-Samarraie, W M Fauzy – 2019
7. “Get more from less: A hybrid machine learning framework for improving early predictions in STEM education” by Mohammad Rashedul Hasan, Mohamed Aly –2019
8. “Student performance predictor using multiclass support vector classification algorithm” by Suhas S Athani, Mayur N Banavasi, Sharath A Kodli, P G Sunitha Hiremath – 2017
9. “An empirical analysis of classification techniques for predicting academic performance” by S Taruna, Mrinal Pandey – 2014 61
10. Chen, Q. Liu, Z. Huang, L. Wu, E. Chen, R. Wu, Y. Su, and G. Hu. Tracking knowledge proficiency of students with educational priors. In Proceedings of the 26th ACM International Conference on Conference on Information and Knowledge Management pages 989–998. ACM, 2017
11. Zhang, G. Su, C. Qing, X. Xu, B. Cai, and X. Xing. Hierarchical life-long learning by sharing representations and integrating hypotheses. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019