VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science Engineering

TO PREDICT STUDENT PERFORMANCE

A PROJECT REPORT

for

FOUNDATIONS OF DATA SCIENCE (BCSE206L)

in

B.Tech (Computer Science)

by

NAMAN GUPTA (21BCE0240)

ARVASU GUPTA (20BCE0841)

4th SEMESTER, 2023

Under the Guidance of

Prof. SHASHANK MOULI SATAPATHY

Associate Professor Sr. SCOPE

March 2023

INTRODUCTION

Higher education's academic community faces a challenge in raising student's academic performance. Engineering and science students academic performance in their first year of college is a turning point in their educational path and typically has a significant impact on their General Point Average (GPA). The evaluation criteria for the students, such as midterm and final exams, assignments, and lab work, are examined. Before the final test is given, it is advised that the class teacher be informed of all this related material. The results of this study will assist teachers in raising student achievement and significantly lowering the dropout rate.

In this research, we provide a hybrid method based on Data Munging , Data Analysis and Data Exploration that enables academics to forecast student (SGPA,CGPA), and based on that, instructors can take the necessary steps to enhance student's academic performance. A frequently used measure of academic performance is the grade point average (GPA). Many universities have a minimum GPA requirement that must be met. As a result, the academic planners continue to use grade point average as their primary indicator of academic achievement. Throughout their time in college, a student's ability to achieve and maintain a high GPA that accurately reflects their overall academic achievement may be hampered by a variety of issues. As data scientists, we need to ask a variety of questions for the issues we focus on.

The faculty members could focus on these elements while creating methods to enhance student learning and enhance their academic success through tracking the development of their performance.

The crucial qualities for future prediction can be found using the data modelling technique's such as SVM, KNN or Random Forest algorithms. The technique of extracting previously undiscovered, reliable, strategically relevant, and concealed patterns from big data sets is known as data clustering which uses decision trees.

Methods like One Hot, 0-1 transformation and suppression methods are used for pre-processing of data and helps to identify the outlier data and also converts the data type of the columns according to the required integer.

We use domain information, statistics, and algorithmic programming together in our project.

BACKGROUND / RELATED WORK

It is important for us to know what are the factors that overall affects the performance of the students. Thus, our motive was to consider all the factors which may or may not affect the grades of a student.

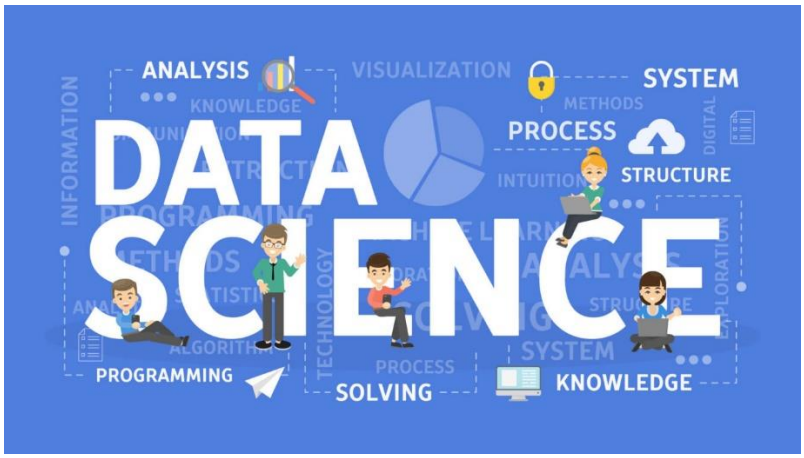
Students Performance Prediction Using Machine Learning Classifiers

By Ahmed Adeel, University Research Journal of Engineering, Science & Technology, Nawab shah 19.1 (2021): 112-121.

This paper implements various machine learning classification techniques on students' academic records for results predication. For this purpose, The Random Forest data modelling has been used for performing all experiments namely, data per-processing, classification, and visualization. For performance measure, classifier models were trained with 5-fold cross validation and 10-fold cross validation methods to evaluate classifiers' accuracy. The results show that bagging classifier combined with support vector machines outperform other classifiers in terms of accuracy, precision, recall, and F-measure score.

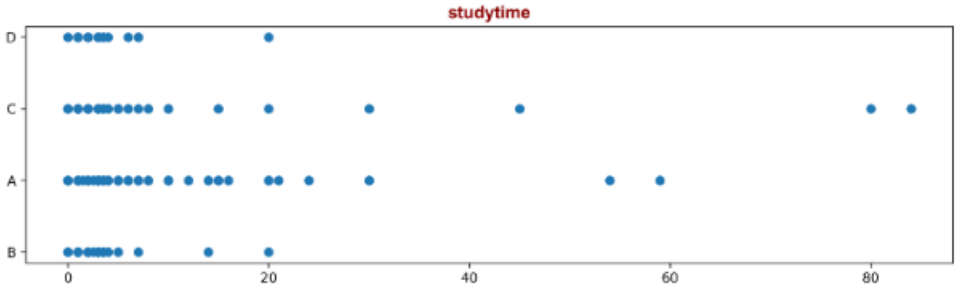
Supervised data Science approach for predicting student performance (2019)

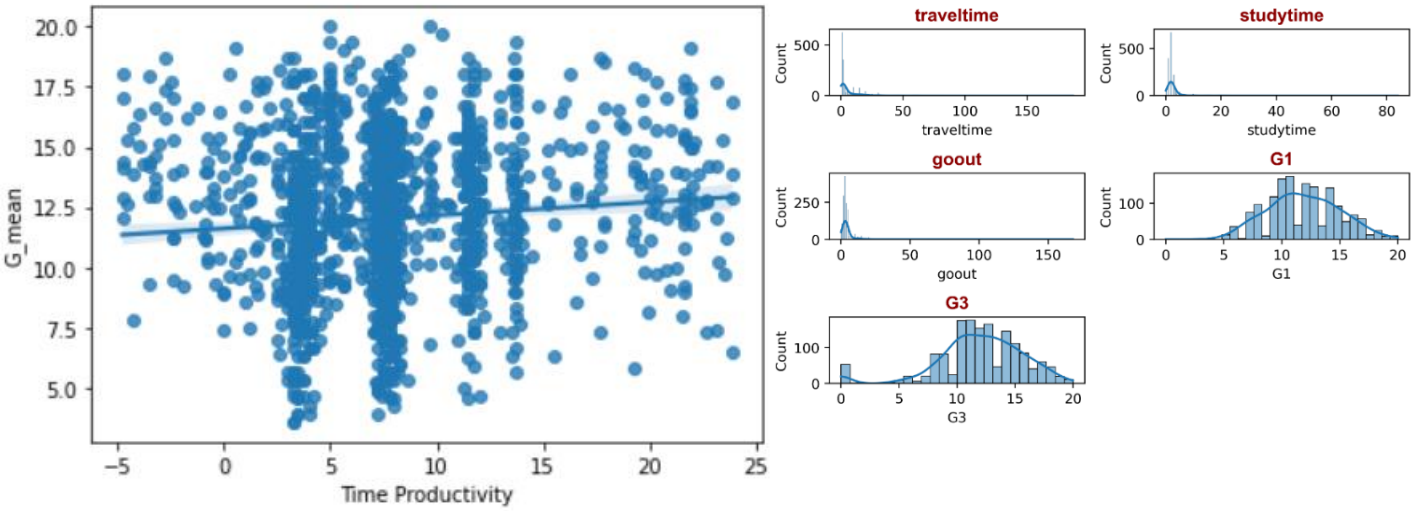
Used Decision Trees, Naïve Bayes, KNearest Neighbour, Logistic Regression for 631 Transcripts from 2013 to 2016



REAL-LIFE APPLICABILITY

- To predict Students GPA (CGPA, SGPA)
- Does higher grades necessarily mean higher study time, The Co-relation between them.
- A Student performance analysis report.
- Traits required for good grades.





INDIVIDUAL CONTRIBUTION

Teamwork simplifies tasks and makes it easier to get them done faster. It teaches how to bond and work with others. Teamwork is the key to success and makes a person humble. Teamwork is crucial to growth and self-development. Communication is an essential skill for good teamwork. In our project we have divided our work in the project equivalently and have made utmost sincerity to coordinate among ourselves towards our final goal.

In this project we have divided our work as:

Naman Gupta-Business Understanding/Data Acquisition/Deployment of Models into web services

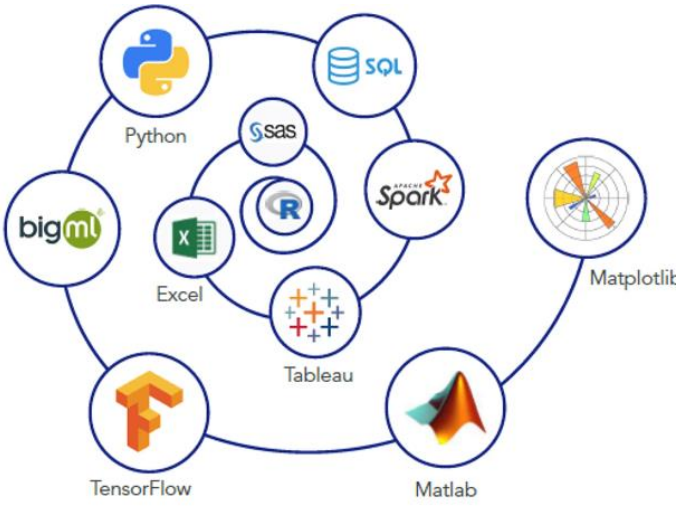
Arvasu Gupta-Feature Engineering/Data Wrangling/Exploratory Data Analysis/Modelling

Through this project we committed ourselves to effectively Develop a curiosity into the field of Data Science, Algorithms and Machine Learning, Modelling and Deployment of models into web services. We have tried our best to explore our data thoroughly as a team.

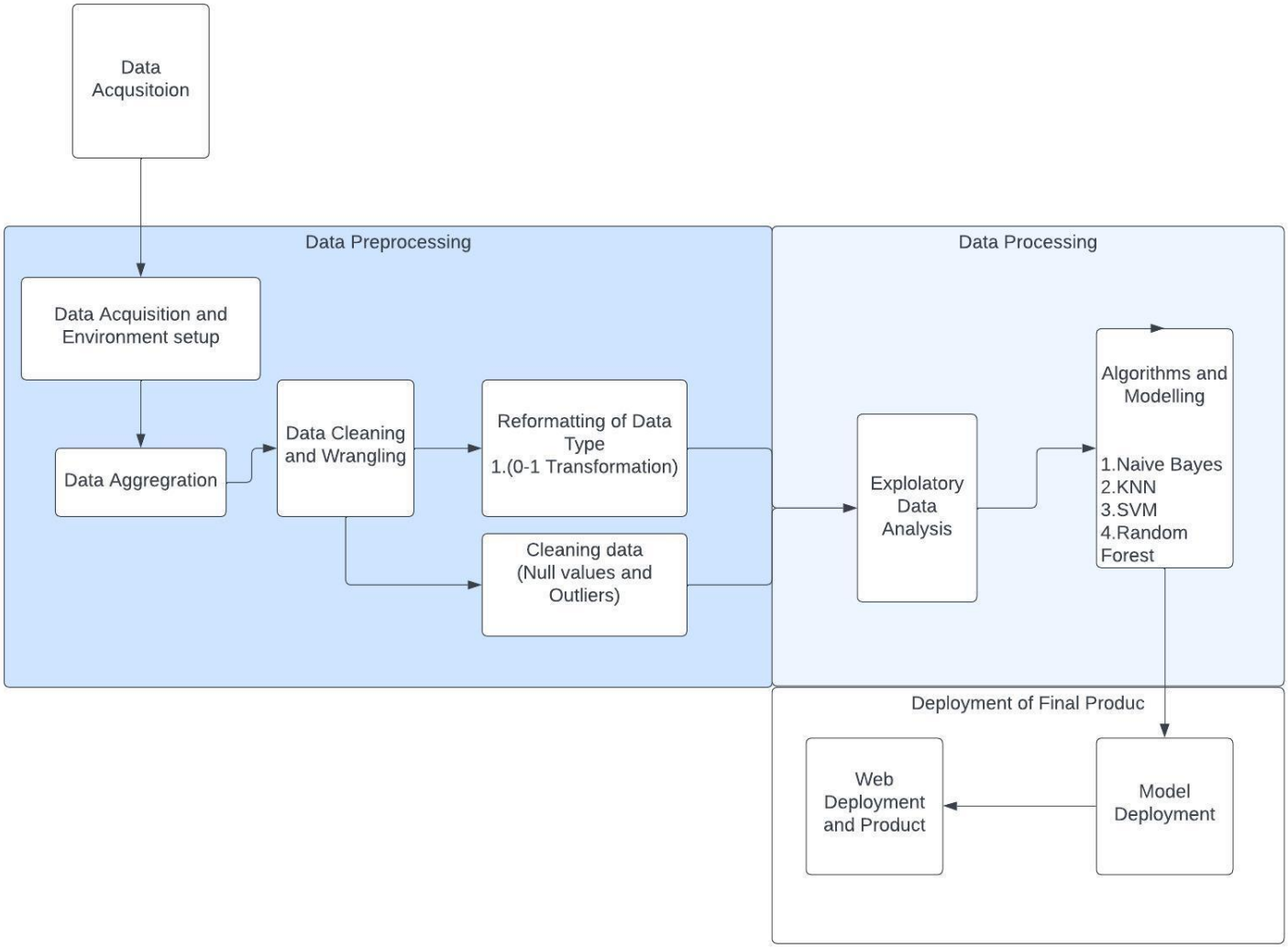
TOOLS AND TECHNOLOGIES USED

Various Tools and technologies are used which are best in their respective field such as-

- PyCharm
- Jupyter Notebook
- VSCode
- Postman
- Flask
- Spark
- Git Bash
- Pandas & Numpy
- Seaborn
- Anaconda cmd prompt



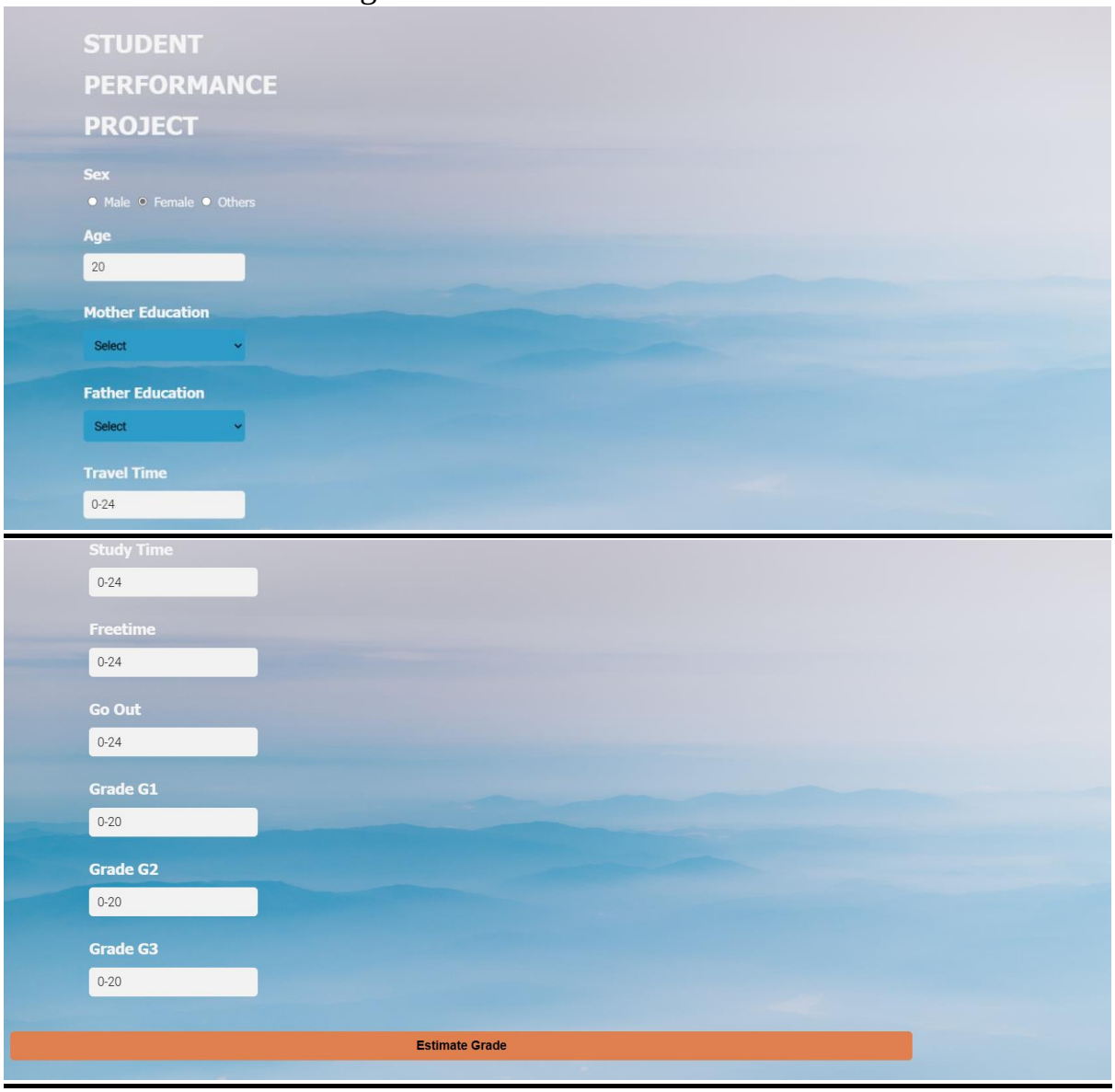
PROPOSED SYSTEM PROCESS FLOW



IMPLEMENTATION RESULTS AND USER INTERFACES

This Project has been deployed as a web server using Flask and has taken necessary Inputs to predict the model efficiently. The webpage has been connected to Flask server and the model.pickle files have been connected to Python server code which GETS and POSTS data from the frontend to backend and back to frontend.
We are taking Gender,Age three previous grades, travel time,study time,freetime, ‘Go-Out’, Mother education and Father education and the basis of their values we will predict our Final Grade- That is letter grade using Random Forest modeling. We chosen the Random Forest

modeling after comparing its results with various other models and their hypertunings.
Thus generating a .PKL file of Random forest model makes us connect the Frontend using Flask server



WORKING METHODOLOGY

As seen from the Process Flow various working methodologies are applied in the project in a proper way to build a model with maximum accuracy.

We can detect/handling outliers through 2 major detections:

- IQR Method
- Z-Score

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

Four Typer of modelling methods are used

- Naive Bayes
- KNN
- SVM
- Random forest

Before applying the necessary models we shall first import models and necessary packages like train_test_split, GridSeachCV, accuracy_score, classification_report etc.

A. Precision(What percent of your predictions were correct)=>

$$Precision = \frac{TP}{(TP + FP)}$$

B. Recall(What percent of the positive cases did you catch=>

$$Recall = \frac{TP}{(TP + FN)}$$

C. F1 Score(What percent of positive predictions were correct)=>

$$F1 = \frac{2 * Recall * Precison}{Recall + Precision}$$

In Naïve Bayes Algirithm

Naives Bayes is Supervised (Probabilistic)Learning Algorithm based on Bayes' Theorem Bayes' Theorem =>

$$P(B \mid A) = \frac{P(A \mid B)\, P(A)}{P(B)}$$

Such taht it is given that A and B are mutually independent events
Hence Basically Naive Bayes gives us the probability of event B when A has already occured.Naive Bayes has high Bias but low variance

Gaussian/Normal Bayes Theorem is a variant of conventional bayes' Theorem, in which the event B(whose probability we need to find) is a continous variable, whereas in the Naive, it was a discrete variable.

Gaussian Theorem=>

$$P(B|A = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp\left(-\frac{(B-\mu_c)^2}{2\sigma_c^2}\right)$$

where c is a class/set of variable A.

In K- Nearest neighbhour is based on Supervised Learning.Basically we can differentiate some clusters of data on basis of their characteriestics and what KNN does is classifies a data on basis of their neighbouring datasets. K in KNN means the number of neighbours. Hence for k=5, 5 neighbour's data would be adjoining.For finding the value of k we need to apply the process of parameter tuning. which is done below.We can use KNN when data is labelled and dataset is small.

In Support Vecotr machine algorithm is a supervised Machine learning algorithm used for both classification and regression.Like KNN,we do better by marking out the middle extreme cases as a plane.Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier. A hyperplane is a decision boundary that differentiates the two classes in SVM
For this context SVM would make classsify the datasets into lettergrades 'A','B','C' and 'D' from the given training data.It would draw hyperplance for each datasets of factors with other factors. Thus on training data set it would classify with the factors given and thus make a prediction model.

In Random forest is a supervised learning algorithm which can be used for both Classification and Regression problems. It is based on on ensemble learning which is a process of combining multiple classifiers to solve a complex problem.

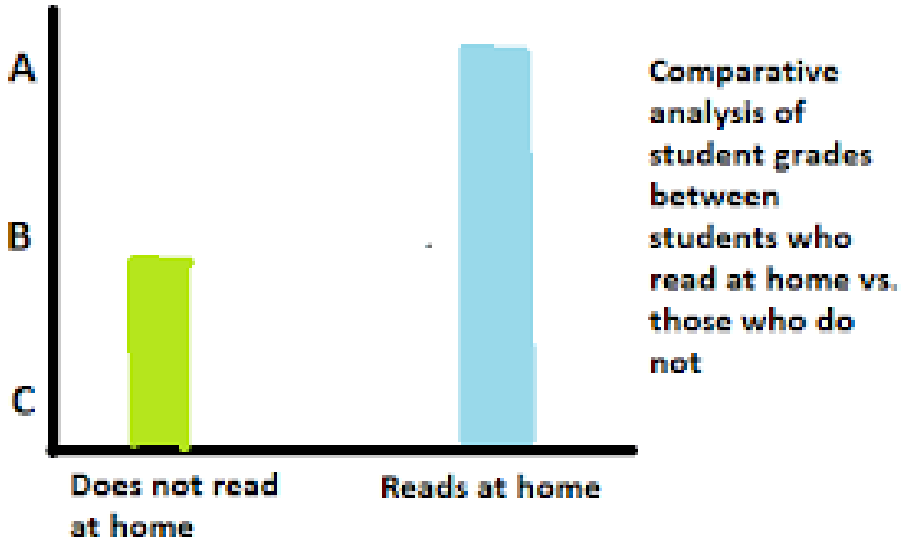
Random forest works on major assumptions

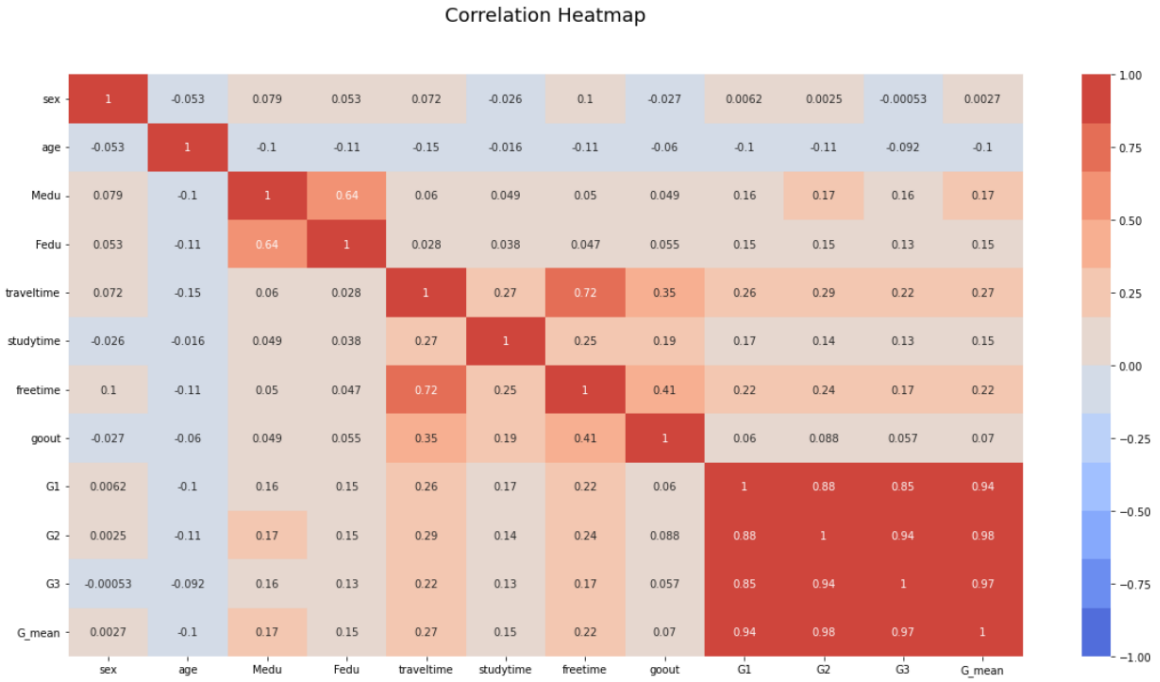
1. There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
2. The predictions from each tree must have very low correlations.

Random forest creates highly accurate models for large database and is thus prefered. Along with that, it also predicts missing data when a large proportion of data is missing.
RF creates multiple decision trees during training phase on the basis of input dataset and the majority of the trees is chosen by the random forest as the final decision.

COMPARATIVE ANALYSIS

Comparison or comparing is the act of evaluating two or more things by determining the relevant, comparable characteristics of each thing, and then determining which characteristics of each are similar to the other, which are different, and to what degree.

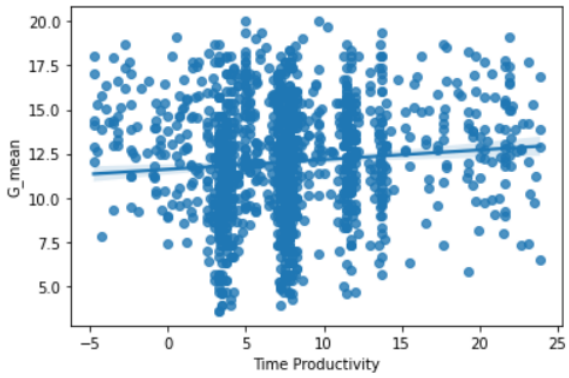




Comparing Time Productivity vs Grade

In [40]:
```
df['Time Productivity'] = 4* df['studytime']-0.25*df['freetime']+0.25*df['goout']-0.25*df['traveltime']
sns.regplot(x='Time Productivity', y='G_mean', data=df)
```

Out[40]: <AxesSubplot:xlabel='Time Productivity', ylabel='G_mean'>



CONCLUSION & FUTURE SCOPE

An essential component of our society is education. The field of education can benefit from business intelligence (BI)/data Science (DS) approaches, which enable the high-level extraction of knowledge from unstructured data. In example, several studies have

employed BI/DS techniques to boost school resource management and educational quality. Using past school grades (first and second periods), demographic, socioeconomic, and other school-related data, we have addressed the prediction of secondary student grades in two key classes (Mathematics and Portuguese). Four alternative DM techniques, including KNN, Support Vector Machine (SVM), Naive Bayes (NB), and Random Forests (RF), were investigated. However, the use of a student prediction engine as a component of a school administration support system has the potential to create an automatic online learning environment. This will enable the gathering of extra information (such as grades from prior academic years) and the acquisition of insightful input from the school personnel. To improve the student databases, we also plan to expand the experiments to new schools and academic years. Because just a subset of the input variables taken into consideration appear to be pertinent, automatic feature selection techniques (such as filtering or wrapping) will also be investigated. This should particularly help nonlinear function approaches (like NN and SVM), which are more susceptible to irrelevant inputs. To comprehend why and how particular factors (such as motivation for choosing school, parent's employment, or alcohol intake) effect student performance, more research is also required (e.g., sociological studies).

REFERENCES

[1] Kumar, Mukesh & Sharma, Chetan & Sharma, Shamneesh & , Nidhi & Islam, Nazrul. (2022). Analysis of Feature Selection and Data Science Techniques to Predict Student Academic Performance. 10.1109/DASA54658.2022.9765236.

https://www.researchgate.net/profile/ShamneeshSharma/publication/359520060_Analysis_of_Feature_Selection_and_Data_Science_Techniques_to_Predict_Student_Academic_Performance/links/6242b4fd57084c718b72cabc/Analysis-ofFeature-Selection-and-Data-Science-Techniques-to-Predict-Student-Academic-Performance.pdf

[2]Ahmed, Adeel, et al. "Students' Class Performance Prediction Using Machine Learning Classifiers." Quaid-E-Awam University

Research Journal of Engineering, Science & Technology, Nawabshah. 19.1 (2021): 112-121. https://pdfs.semanticscholar.org/0e6b/f2516ecb3eebdd1e1f35e16a26ce4d830769.pdf?_ga=2.771 83125.1291724342.1660019231-263005371.1660019231

[4] Ünal, Ferda. "Data Science for student performance prediction in education." Data ScienceMethods, Applications and Systems (2020). https://pdfs.semanticscholar.org/8b76/9fd122c361ee695d71cc1e9aeac36f0cae67.pdf?_ga=2.1427 7687.1291724342.1660019231-263005371.1660019231

[5] Mengash, Hanan Abdullah. "Using data Science techniques to predict student performance to support decision making in university admission systems." IEEE Access 8 (2020): 55462-55470. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9042216 30

[6] Hasan, Raza, et al. "Predicting student performance in higher educational institutions using video learning analytics and data Science techniques." Applied Sciences 10.11 (2020): 3894. https://www.mdpi.com/2076-3417/10/11/3894

[7] Karthikeyan, V. Ganesh, P. Thangaraj, and S. Karthik. "Towards developing hybrid educational data Science model (HEDM) for efficient and accurate student performance evaluation." Soft Computing 24.24 (2020): 18477-18487. https://link.springer.com/article/10.1007/s00500-020-05075-4

[8] Alhakami, Hosam, Tahani Alsubait, and Abdullah Aljarallah. "Data Science for student advising." International Journal of Advanced Computer Science and Applications 11.3 (2020). https://pdfs.semanticscholar.org/6a37/d4fca2abe00eba300d89668965dee95660ab.pdf?_ga=2.725 94451.1291724342.1660019231-263005371.1660019231

[9] Hassan, Hasniza, Nor Bahiah Ahmad, and Syahid Anuar. "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data Science." Journal of Physics: Conference Series. Vol. 1529. No. 5. IOP Publishing, 2020. https://iopscience.iop.org/article/10.1088/1742-6596/1529/5/052041

[10] Yousafzai, Bashir Khan, Maqsood Hayat, and Sher Afzal. "Application of machine learning and data Science in predicting the performance of intermediate and secondary education level student." Education and Information Technologies 25.6 (2020): 4677-4697. https://link.springer.com/article/10.1007/s10639-020-10189-1