# A MINI PROJECT REPORT

# ON

**STUDENT GRADE ANALYSIS AND PREDICTION-USING MACHINE LEARNING**

**Submitted to**

**Sri Indu College of Engineering and Technology, Hyderabad**

**in partial fulfilment of the requirement for the award of degree of**

**BACHELOR OF TECHNOLOGY**

**In**

**INFORMATION TECHNOLOGY**

**Submitted**

**By**

**M.NIKHITHA (20D41A1233)**

**S. RAJASREE (20D41A1249)**

**S. HARSHITHA (20D41A1247)**

**Under the esteemed guidance**

**Of**

**PROF. J.S RADHIKA**



DEPARTMENT OF INFORMATION TECHNOLOGY

**SRI INDU COLLEGE OF ENGINEERING AND TECHNOLOGY**

## (An Autonomous Institution under UGC, Accredited by NBA, Affiliated to JNTUH)

Sheriguda, Ibrahimpatnam (2023-2024)

# SRI INDU COLLEGE OF ENGINEERING AND TECHNOLOGY

## (An Autonomous Institution under UGC, Accredited by NBA, Affiliated to JNTUH)

**DEPARTMENT OF INFORMATION TECHNOLOGY**



**CERTIFICATE**

Certified that the Technical Seminar Work entitled “**STUDENT GRADE ANALYSIS AND PREDICTION – USING MACHINE LEARNING”** is a Bonafide work carried out by **M.NIKHITHA (20D41A1233), S.RAJASREE (20D41A1249), S.HARSHITHA (20D41A1247)** in partial fulfillment for the award of **BACHELOR OF TECHNOLOGY** in **INFORMATION TECHNOLOGY** of SICET, Hyderabad for the academic year 2023- 2024.The project has been approved as it satisfies academic requirements in respect of the work prescribed for the **IV YEAR, I-SEMESTER** of **B. TECH** course.

**INTERNAL GUIDE HOD/IT**

**PROF.J.S. RADHIKA Dr. P. EPSIBA**

**DEAN External Examiner**

**Dr. K. S. SADASIVA RAO**

**ACKNOWLEDGEMENT**

With great pleasure we want to take this opportunity to express our heartfelt gratitude to all people who helped in making this project work a success. We thank the almighty for giving us the courage & perseverance in completing the project.

We are thankful to Principal **Prof. Dr. G. Suresh**, for giving us the permission to carry out this project and for providing necessary infrastructure and labs.

We are highly indebted to, **Dr. P. Epsiba**, Head of the department of Information Technology, project for providing valuable guidance at every stage of this project.

We are grateful to our internal project guide, **Prof. J. S. Radhika**, for his constant motivation and guidance given by her during the execution of this project work.

We would like to thank the teaching & non- teaching staff of department of Information Technology for sharing their knowledge with us.

Last but not the least we express our sincere thanks to everyone who helped directly or indirectly for the completion of this project.

**M.NIKHITHA (20D41A1233)**

**S. RAJASREE (20D41A1249)**

**S. HARSHITHA (20D41A1247)**

**ABSTRACT**

The goal of this paper is to put forth the analysis and results obtained by me while trying to answer the question of predicting the students grade using the chosen dataset which is ’Student Alcohol Consumption’. The data for this dataset was obtained in a survey of students of a secondary school. This dataset includes information about demographic and social factors. In this project we have tried to predict the final grades of the students. For prediction we have used 5 machine learning algorithms which are Multiple Regression, Stepwise Regression, Logistic Regression, Naive Bayes and K nearest Neighbour. A detailed report of all the analysis and results that I got by using the Machine Learning techniques along with their interpretations and discussions can be found in the following sections. **Keywords**: - Student Alcohol Consumption, Multiple Regression, Stepwise Re- gression, Logistic Regression, Naive Bayes and K Nearest Neighbour.

# CHAPTER-1

# INTRODUCTION

The education is very important in everyone’s life and for maintaining the society and for living a happy life it is very important to study. Alcohol is one of the major drink that is known to be an enjoyment drink and it is also known to be a status drink in the world. This kind of thinking have had bad impact on the today’s teenagers as drinking these days is highly regarded as cool and no drinking not so cool. For being cool, students start drinking alcohol at very early age as everyone wants to have fun and no one wants to be left alone.

## MOTIVATION

Alcohol consumption also affect the students academically. Due to the consumption of high level of alcohol many students can’t get enough marks to pass in their examinations. Alcohol consumption, bad behaviour, relationship with their parents impacts the bodies of students and their mental health too. There are many countries in the world which don’t let teenagers drink who are less than 18 years old but somehow these young people always found a way to drink. If young people have their hands on alcohol at an early age then the drinking can be a addiction to them for their whole life. So we can say that while alcohol costs lots of money it also costs life. There are many cases which led to suicide due to alcohol or self harm or loss of memory. It also has been found out that alcohol also affects people’s mood and over consumption of alcohol makes the people less conscious as well. Alcohol is also the reason which helps people to make unwanted mistakes. Alcohol wasted that time of students which should be invested in education but for looking cool, students drink alcohol and wasted their whole life.

## RESEARCH QUESTION

What are the main factors that led the students to fail in their examinations?

## PLAN OF PAPER

For this project, I am going to use Student Alcohol Consumption dataset which is a res- ult of survey of students for maths and Portuguese language courses in secondary school. It contains social, gender and study information about students. The dataset contains information of 1045 students and in dataset there are 382 students whose are students of both Maths and Portuguese Class, there are 33 attributes attached to each student. There are many several social and demographic factors available.

In order to perform meaningful analysis on the data we firstly perform some data prepar- ation steps. This include taking care of categorical variables and missing values. After preparation of data we found significant variables by performing multiple regression and stepwise regression. We then tried to predict grades that students would obtain by per- forming different classification algorithms. Lastly, we performed dimensionality reduction technique and tried to see if that makes any difference to the classification results.

In upcoming section [2](#_bookmark0) we’ll come to know about the researches that has already been done in this area on this topic. After that in further section [3](#_bookmark1) we’ll discuss about the methodology that we are going to use in my project. In the next section [4](#_bookmark3) we’ll discuss about the algorithms, dataset and the evaluation method that we are going to use in our project. After the Specification, in next section [5](#_bookmark5) we’ll go through data preparation and then in the next section [6](#_bookmark6) we will talk about the results that we got after applying all the models and then in the next and last section [7](#_bookmark13) we’ll conclude our project and in that section we will also talk about the future work that can be done in this area.

# CHAPTER-2

# RELATED WORK

In this section, we will go through some of the researches that has already been done on this topic.

In this research, researcher [Pagnotta and Hossain (2016)](#_bookmark22) used same dataset ”Student Alcohol Consumption” and by using this dataset, the researchers tried to approach the addiction of alcohol in secondary school students. These researchers used Business Intel- ligence and different Data Mining techniques to predict the level of addiction. By this research, the researchers also came to know that the alcohol drinking was also impact- ing the final results of the students. The researchers have used Decision tree, Random Forest for classification and they have also used KNIME Analytics Platform. In the data pre-processing, they have used KNIME rule engine for data reduction and KNIME concatenate component for merging two datasets. After that they performed linear cor- relation for column filtration. After that the researchers also removed backward feature. And for elimination the researchers have used loop with cross validation and they have used cross validation with random forest for prediction and testing the result and in the last, they filter backward elimination feature data so that they can rescue the original data. By this research, researchers came to know that the males were more involved in drinking. They also came to know that the person who goes out too much with friends drinks more alcohol. According to the authors, More free time and less study time are two of the main characteristics. The researchers also came to know that the students who weren’t frequent to university they will drink more. In their research paper, the re- searchers got 8.0.18% of average error rate and an accuracy level of 92% with no missing values.

In the work of Researchers [Pal and Chaurasia](#_bookmark23) [(2017),](#_bookmark23) they tried to identify those students who need counselling to understand the bad effects of alcohol on life and for identifying the researchers have used four different data mining techniques and those techniques are Sequential Minimal Optimization, Bagging, REP Tree and Decision Table. For implementing the machine learning models the researchers have used WEKA Toolkit. First the researchers collect the data and then they converted it into arff file format so that the data can be mined by the WEKA toolkit. While the Sequential Minimal Optimization model took 0.97 Seconds for building, the Bagging, REPTree and Decision Tree took 0.14 seconds, 0.02 seconds and 0.23 seconds respectively. While Sequential Minimal Optimization model classified 290 instances correctly, other models like Bagging, REPTree and Decision Tree were able to classify 317, 316 and 313 instances correctly. The accuracy level of Sequential Minimal Optimization model, Bagging model, REP Tree and Decision Model were 73%, 80%, 80% and 79.24%. The researchers concludes that legitimate admittance to liquor influences understudy execution. The outcomes of this research recommend that among the AI calculation tried, Bagging classifier can possibly fundamentally further develop the ordinary order strategies utilized in the examination. In another research, [Pisutaporn et al.](#_bookmark24) [(2018)](#_bookmark24) tried to analyse the importance of dif- ferent variables by using different data mining algorithms. Their main goal was to study the student alcohol consumption. They used the same dataset ”student alcohol consump- tion”. First they applied two of the well known models of data mining world to analyze the importance of the variables and those models were Decision Tree and Random Forest then they also applied a regression model to understand and present the relationship between alcohol consumption level and student’s final grades. The researchers got the the accuracy level of 56.36% when they tried the decision tree with Walc. Walc is week- end alcohol consumption level and when they tried decision tree with Dalc, they were able to achieve the accuracy level of 72.84%. Dalc is Weekdays Alcohol Consumption level. The researchers got the the accuracy level of 88.07% when they tried the Random Forest with Walc. The researchers got the the accuracy level of 79.43% when they tried the Random Forest with Dalc. And when the researcher tried to predict the final grades of the student using regression, they got the root mean square error of 3,827 and r-square value of 0.019. They conclude that there is a negative relationship between Dalc and G3 and Walc and G3. G3 is the final grades of students. It has been clearer in this research that the random forest works very good in this area and surprisingly it has been found out that there is no connection between alcohol consumption level and the student grade. In the research of [Trivedi and Kotak](#_bookmark28) [(2019),](#_bookmark28) they tried to predict if a student is addicted to alcohol or not and for the prediction of alcoholic students, the researchers have used data mining methods like clustering, classification and some other filtering methods. First thing the researchers have done is they applied K-means clustering, after applying k means clustering they tried to find the best accuracy by comparing two classifiers. The researchers have used ID3 algorithm of decision tree. When researchers used Z-transformation for classification, for Decision tree they got the accuracy level of 98.77%, the recall of 9.57% and the Precision level of 58.99% and for Naive Bayes they got the accuracy level of 98.00%, the recall of 57.39% and the Precision level of 56.46%. When researchers used Rang transformation for classification, for Decision Tree they got the accuracy level of 98.77%, the recall of 59.57% and the Precision level of 58.99% and for Naive Bayes they got the accuracy level of 98.00%, the recall of 57.39% and the Precision level of 56.24%. When researchers used Preposition transformation for classification for Decision tree they got the accuracy level of 98.77%, the recall of 59.57% and the Precision level of 58.99% and for Naive Bayes they got the accuracy level of 39.30%, the recall of 11.31% and the Precision level of 11.43%. When researchers used Interquatile transformation for classification they got the accuracy level of 98.77%, the recall of 9.57% and the Precision level of 58.99% and for Naive Bayes they got the accuracy level of 98.15%, the recall of 57.83% and the Precision level of 56.87% and when they use two of their own algorithms they got different results. When they used their own Clustering method they got the Precision of 64%, Recall level of 64%, Fscore level of 64% and the Accuracy of 64%. When they used their own Decision Tree, they got the Precision of 62%, Recall level of 62%, Fscore level of 62% and the Accuracy of 62%. The researchers conclude that the DT algorithm was the best algorithm used in this research and they also came to know that there is a positive correlation between student’s behaviour and their academic performance.

In their research, researchers [ElTayeby et al. (2018)](#_bookmark19) tried to identify the alcoholic posts

on Facebook by mine the text, images and videos using machine learning algorithms. Te researchers select 4266 posts from a group called ”I’m Shmacked”. The researchers build different types of models for different type of posts like different model for text data, a different model for photos and another different model for videos. After building the models, the researchers sorted the models according to their performance and in last they combined top performing models so that the performance of the models can be improved. First the researchers cleaned text data by removing special characters like links, emojis and hash tags. All these types of data comes under noise. Then the researchers build models. They build SVM classifier for recognizing the data. The researchers also used LLDA for classifying the posts. They used two LLDA algorithms which are Gibbs sampling and Bayes. When the researchers build image classification model, they reset all images into a specific format which is of 256 X 256. Researchers did that so that they can be fitted into Neural Network. For all the videos they extracted one image from every 100 frames for total of 462 videos and then these images were added to training dataset. When researchers used SVM with Linear Kernel they were able to get the precision of 84%, Recall of 63%, F-Score of 72% and Support of60% and when researchers used SVM with customized poly degree and gamma they were able to get the precision of 88%, Recall of 60%, F-Score of 71% and Support of 60%. When researchers used SVM with RBF they were able to get the precision of 88%, Recall of 60%, F-Score of 71% and Support of 60%. When researchers used SVM with Sigmoid they were able to get the precision of 77%, Recall of 55%, F-Score of 64% and Support of 60%. When researchers used LLDA with Gibbs Sampling they were able to get the precision of 68%, Recall of 71%, F-Score of 69% and Support of 58%. When researchers used LLDA with CVB0 they were able to get the precision of 64%, Recall of 60%, F-Score of 62% and Support of 60%. When Researchers used SVM on text, hints they got the precision of 86%, Recall of 60%, F-Score of 71% and Support of 60%. When Researchers used SVM on text, hints and links they got the precision of 84%, Recall of 63%, F-Score of 72% and Support of 60%. When Researchers used Alexnet on Images they got the precision of 29%, Recall of 55%, F-Score of 38% and Support of 20%. When Researchers used Alexnet on Videos, they got the precision of 83%, Recall of 93%, F-Score of 88% and Support of 27%. When Researchers used Combined model, they got the precision of 78%, Recall of 65%, F-Score of 71% and Support of 60%. The researchers conclude that online media clients regularly erase unseemly substance, especially in the arrangement of picture and video after unrehearsed posting.

With the help of this study which is conducted by [Htet et al. (2020)](#_bookmark20) where they tried to find the impact of alcohol on the university students in Myanmar. In their research, they have used a sample of conducted study of 15 years old students to 24 years old students. In this study, the students were selected from six universities from Mandalay, Myanmar. This study was conducted in 2018 and there were total of 3456 students who participated in this study. For this research, the researchers have used Multiple Logistic Regression. The researchers have used multiple logistic regression for estimating the adjusted odds ratio and the researchers have also used 95% confidence interval. In their research, they found out that in the previous 30 days 36% of males and 10.8% of females were involved in alcoholic activities. It has been found out that males were more interested in alcoholic activities. It also has been found out the main factors for consuming alcohol were age, sex, monthly expenses, parental alcohol consumption, peer alcohol consumption, truancy, and feelings of sadness or hopelessness. On that account, It can be said that things like counselling is very much needed and it is also important that government created some strict rules and regulations so students can’t drink alcohol that will also help the society. In the work of [Butler et al.](#_bookmark15) [(2010),](#_bookmark15) they examined 106 employed students for their relationship between work stressors and alcohol consumption for 14 days. The researchers have used a framework which is known as tension reduction, by using that framework the researchers have came to know that the work stress increase the chance of consumption of the alcohol. It also has been found out that more males were involved in alcohol activities. The researchers found out that hour works were also positively related to the number of drinks consumed. They found out that either student get employment when he is student affected the college student drinking problems as well. The researchers found out that male may drink more alcohol than females but there are different reasons why males and females drink alcohol. Women drinks alcohol when they are under any event stress and men drinks alcohol when they are under work stress. There are many limitations in this study like the sample was of limited size, dataset was of a survey result which was self

reported so the people can lie and dataset can be biased.

In this paper, researcher [Kitsantas et al.](#_bookmark21) (2[008)](#_bookmark21) tried to find the subgroups for drinking alcohol amid college students. The researchers used the survey of American and Greek understudies addressed inquiries regarding liquor utilization, strict convictions, mental- ities toward drinking, promotion impacts, parental checking, and drinking results. It has been found that in America the higher amount of alcohol is consumed by younger people and the consumers are also less religious and Greek students didn’t believe in the after impacts of alcohol and those people who drink less in Greek they were monitored by their parents. By this research, the researchers found out that parental checking and an accentuation on advising understudies about the adverse consequences regarding liquor on their well being and social and scholarly lives might be viable techniques for lessening liquor utilization. The lenient perspectives toward drinking in undergrads who live in Greece may go about as a cradle against hazardous understudy drinking. When to a lesser extent an untouchable encompasses drinking, it might turn out to be less appealing to understudies, or maybe understudies are associated into savoring a less risky way by Greek guardians and the parental monitoring in Greece is also very high. The main for higher parental monitoring in Greece is that more of the students in Greece lives with their parents than the America. We can say that the size was small of this research and also the number of attributes are also less and it was also a self reported survey so as we know that the students can lie.

In an another research, [Singleton](#_bookmark27) [(2007),](#_bookmark27) tried to analyse the relationship between alcohol and grades of students. The researcher of this research conducted personal in- terviews of students in a college. The researcher interviews total of 754 people. In the interview, the researcher asked questions about alcohol consumption, gender, athletic status, parent’s education, income and frequency of attending off-campus parties. 94% of students gave their permission to access their grades. In this research the researcher have used Least square regression method to analyse. By using regression method the researcher came to know that the gender and partying were responsible for 43% of the alcohol consumption. It also has been came to know that the parents income is also responsible for how much student is drinking. Researcher created four modes in total, in first model, all independent variables were included but they didn’t include amount of alcohol and rank in high school. It has been found out in first model that the SAT score were positively related to GPA. In model 2 the researchers add the amount consumed too and they came to know that the man’s consumption level were also responsible for getting low grades. Models 3 and 4 present the relapses of aggregate GPA on the subset of respondents for whom secondary school class rank was accessible.

In work of [Davoren et al.](#_bookmark18) [(2017),](#_bookmark18) they distributed questionnaires to students of ran-

domly selected students. The researchers got replied on total of 2332 questionnaires in which 84% students responded. There were a total 49 questions.It has been came to know that the 65% of the male students were consuming too much of the alcohol and 68% of the women were also involved in high level of drinking.

In the work of [Chudasama and Joglekar](#_bookmark17) [(2016),](#_bookmark17) the researchers tried to predict the performance of student using Artificial Neural Network. They tried to find what are the factors that affect student performance. The included only 13 variables in their research and they used Feed-form=ward topology of Artificial Neural Network. In their research, the researchers were able to get the accuracy level of 78.94%. More accuracy can be achieved if there were more variables.

In another research [Aissaoui et al. (2019)](#_bookmark14) tried to build a prediction model. For cre- ating the prediction model, the researchers have used multiple linear regression. First they used linear regression model for finding the relationship between one dependent and one independent variable after that, they applied multiple linear regression for finding relationship between one dependent variable and many independent variables. The re- searchers build 7 different regression models in which the best model was the model with the heist R-squared and lowest RMSE and MAE. It also have negative correlation coeffi- cients. The researchers came to know that the increase in number of age and go out will impact and decrease final grades.

In another research of [Shukla et al. (2018),](#_bookmark26) they tried to identify those attributes which affect the student performance. They used discretization and feature selection in their pre-processing and called it Multistage pre-processing. Via their system the researchers tried to predict if a student is addicted to alcohol or not. The researchers performed many different types of tests like correlation based feature selection (CFS), Information Gain (IG), Chi-Square and Relief-F. In their research, the researchers have used 6 classifiers namely Naive Bayes, Support Vector Machine, J48 Decision Tree, k-Nearest Neighbor, Random Forest and Multi Layer Perceptron and with all classifiers the researchers have used 10-fold cross validation. When the researchers have used Random Forest with CFS they achieved an accuracy level of 67.59%. When the researchers have used Random Forest with IG they achieved an accuracy level of 69.87%. When the researchers have used Random Forest with Chi-Square they achieved an accuracy level of 70.37%. When the researchers have used Random Forest with Relief-F they achieved an accuracy level of 71.39%. When the researchers have used Random Forest with All features they achieved an accuracy level of 71.24%. When the researchers have used SVM with CFS they achieved an accuracy level of 67.87%. When the researchers have used SVM with IG they achieved an accuracy level of 69.11%. When the researchers have used SVM with Chi-Square they achieved an accuracy level of 67.8%. When the researchers have used SVM with Relief-F they achieved an accuracy level of 69.01%. When the researchers have used SVM with All features they achieved an accuracy level of 69.01%. When the researchers have used NB with CFS they achieved an accuracy level of 70.12%. When the researchers have used NB with IG they achieved an accuracy level of 68.6%. When the researchers have used NB with Chi-Square they achieved an accuracy level of 68.86%. When the researchers have used NB with Relief-F they achieved an accuracy level of 68.86%. When the researchers have used NB with All features they achieved an accuracy level of 68.60%.When the researchers have used k-NN with CFS they achieved an accuracy level of 66.83%. When the researchers have used k-NN with IG they achieved an accuracy level of 64.05%. When the researchers have used k-NN with Chi-Square they achieved an accuracy level of 64.05% and When the researchers have used k-NN with Relief-F they achieved an accuracy level of 66.91%. The researchers conclude that the dataset was a small dataset and they need more data to obtain more higher level of accuracy.

In another work of [Sakaray et al.](#_bookmark25) [(2017](#_bookmark25)), the researchers tried to elevate the per-

formance of students using machine learning algorithms so that those students who need guidance can be recognized earlier. The researchers used same dataset as us and they applied decision Tree algorithm and Random Forest algorithm. In this research, the re- searchers have used KNIME Analytics. The main goal of this research is to find the alcoholic consumption by secondary school students. The researchers also used WEKA tool. In this research, for selecting the best attributes the researchers have used Neural Networks. The researchers have also used Apriori algorithm. When researchers applied Neural Network on 40% of dataset, they were able to get 50% Accuracy and when they applied it on 50% of the dataset, they were able to get 50% accuracy as well but when they applied it on the 60% and 70% of the dataset, they were able to get the accuracy level of 60% and 70%. Tee researchers of this research concludes that the results of the students is totally based on the performance of the students in the previous exams. It has been came to know that the performance of the neural network is a lot better than others like decision tree, multi linear regression and Apriori algorithms.

## 

# CHAPTER-3

# METHADOLOGY

The methodology that I am going to use in my project is CRISP-DM. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. CRISP-DM is a extensive data mining system and measure model that gives anybody from amateurs to data mining specialists with a total plan for directing a data mining project.

As we can see in the Figure [1](#_bookmark2) there are total of 6 stages in CRISP-DM methodology which are following:

* Business Understanding:
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

# 

Figure 1: CRISP-DM

**Business Understanding** is the phase in which the analyst tries to understand the objectives and requirement of project. This knowledge is very useful for the analyst to solve the data mining problem and it is also very useful to develop plan for the project.

**Data Understanding** is the phase in which a analyst do collection of data and then they tried to understand about the data. In this stage, the expert may likewise distinguish intriguing subsets to shape theories for covered up data

**Data Preparation** is the phase in which the analyst take initial raw data and convert it into final dataset from the information that he achieved from previous section.

**Modelling** is the phase in which the analyst select and implement the best data mining techniques onto the dataset. There may be some things that are required more attention that’s why there is a loop in this section to preparation of the data.

**Evaluation** is the phase in which analyst checks the outcomes. Analyst checks if the outcomes of their program meets the objectives of the business.

**Deployment** is the phase in which the analyst prepare the plan for the deployment. The analyst also plan for the monitoring and the maintenance and they also prepare final report and then deploy it forward.

# CHAPTER-4

# DESIGN SPECIFICATION

## 4.1 Algorithms

In this section I am going to discuss about the algorithms that I am going to use in my project:

### **4.1.1 MULTIPLE REGRESSION**

Multiple Regression is used when an analyst wants to check the relationship between a dependent and a set of independent variables. The dependent variable in the multiple regression is always of continuous type and the independent variables can be of discrete or continuous type. Although, they are usually of continuous type.

The model for multiple regression is: y = B 1 \* x 1 + B 2 \* x 2 + . . . + B n \* x n + A. Here the subscripts means independent variables. A is the constant stating the value of dependent variable, y, when all of the xs (independent variables) are zero. B are the coefficients linked to each independent variable.

### 

### **4.1.2 STEPWISE REGRESSION**

Stepwise regression is used when the analyst wants to construct their regression model step by step. It also includes the adding of independent variables and it also includes removing the variables which are not useful for the model. Stepwise relapse is a technique that iteratively inspects the factual meaning of every independent variable in a linear regression model. There are two types of Stepwise regression approach: Forward Selection and Backward elimination method. Forward selection starts with 0 variable and adds new variable as it goes. Backward Elimination method starts with all the variables loaded in a model and then it removes variables one by one as it goes.

### 

### **4.1.3 LOGISTIC REGRESSION**

Logistic regression is the type of regression which is mostly used for binary type of classification problem. Logistic Regression is used for prediction of categorical variable by taking all of the given independent variables. Output of logistic regression always returns in 0 or 1 in which 0 means false and 1 means true.

The equation of Logistic Regression is:

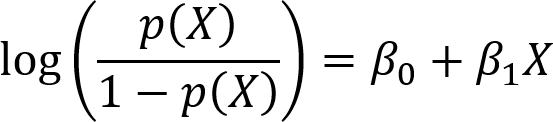


Figure 2: Logistic Regression

**4.1.4 NAÏVE BAYES**

Naive Bayes is one of the classification algorithm. In straightforward terms, a Naive Bayes classifier accepts that the presence of a specific feature in a class is independent to the presence of some other feature. It is very easy to apply and very much useful as well. The main advantage of using Naive Bayes is that it is very easy and very fast for prediction. For using Naive Bayes Classifier there is less training data required. There are some cons of using it as well. If categorical data is available in the test data which was not present in the training data, then there will be less chance for a correct prediction. The equation for Naive Bayes is:



Figure 3: Naive Bayes

### **4.1.5 K-NEAREST NEIGHBOUR**

K-Nearest Neighbors algorithm is one of the algorithm which can be used for the regres- sion as well as the classification problems. k-Nearest Neighbors algorithm is one of the supervised machine learning algorithm. k-Nearest Neighbors algorithm assumes that the similar type of things are close to each other. First it loads the data and then initialize K to the user’s chosen number of neighbors. The predictions become less stable if the value of k goes below 1. The predictions become more stable if the value of k is above

1. The higher the value of k the better will be the results. The main advantages of using k-Nearest Neighbors is it is very simple and very easy to use and it can be used for classification and regression problems but as user increases the number of variables the slower it gets.

Example of k-Nearest Neighbors is given below:

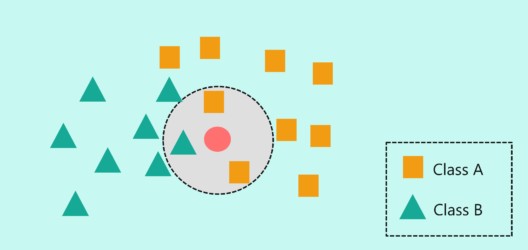


Figure 4: k-Nearest Neighbors

### **4.1.6 PRINCIPAL COMPONENT ANALYSIS**

Principal Component Analysis is a dimensionality reduction strategy that is used to di- minish the dimensionality of data set, by changing a huge number of variables into a modest one that actually contains the majority of the data in the huge set. Diminishing the quantity of factors of a dataset normally comes to the detriment of precision, however the stunt in dimensionality decrease is to exchange a little exactness for straightforward- ness. Since more modest datasets are simpler to investigate and envision and make breaking down information a lot simpler and quicker for calculations. So to summarize, the possibility of Principal Component Analysis is basic, lessen the quantity of factors of a dataset, while safeguarding however much data as could reasonably be expected.

## 4.2 EVALUAITION METHOD

**Confusion Matrix** is a table which is used to check how the classification model has performed. There are four cubes in a confusion matrix. One contains True Positives, another one contains True Negatives, another one contains False Positives and the last one contains False Negatives. True Positives are those which we predicted true and they are true. True Negatives are those which we predicted False and they are False. False Positives are those which we predicted True but they are false. False Negatives are those which we predicted False but they are true.We can clearly see it in the Figure [5.](#_bookmark4)

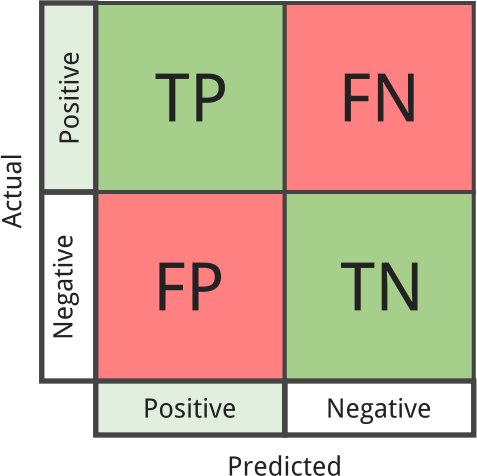


Figure 5: Confusion Matrix

We will also use other types of methods for evaluation of our different models.

## 4.3 DATASET DESCRIPTION

The Description of attributes pf the dataset is as follows:

### **Attributes Description**

School School’s name

Age Age of students

Sex Gender of students Address Home address of students Famsize Family size of students

Pstatus Cohabitation status of parent of students Medu Education level of mother of students Fjob Education level of father of students Reason Reason for choosing this school

Guardian Guardian of student

Traveltime How far is school from home in hours Studytime Weekly study time

Failures Number of past class failures Schoolsup Extra educational support Famsup Educational support by family Paid Extra paid classes

Activities Extra-curricular activities Nursery Attended nursery school Higher Wants to take higher education Internet Internet access at home Romantic In a romantic relationship Famrel Quality of family relationships Freetime Free time after school

Goout Going out with friends

Dalc Workday alcohol consumption

Walc Weekend alcohol consumption

Health Current health status Absenses Number of school absences

There are three more attributes which are grades (G1, G2 and G3) and there are 382 students which appeared in both maths and portuguese classes.

# CHAPTER-5

# IMPLEMENTATION

## 5.1 DATA PREPARATION

Data preparation and preprocessing is one of the most important process in which the analyst tried to clean and transform the raw data before the processing and analysis. It is a very important part before the processing and this step contains things like formatting of the data, join multiple datasets to make it more useful, delete or add some columns to make it more useful. Data preparation is known as the one of the most time consuming process for data analysts but it is very important to remove the poor quality of the data. In our project, for data preparation and preprocessing we did the following:

* First we checked if there is any missing values in our dataset.
* Then in the next step we encoded categorical variables into factors.
* After transforming the variables, we also created some dummy variables specially for Principal Component Analysis for handling the nominal variables.
* Then in the next step we converted the G3 variable which was a continuous variable and we transformed it into categorical variable.
* Then in the last step we splitted our dataset into test and training subsets in which 80% of the dataset was used for training and 20% of the dataset was used for testing.

**5.2 DESCRIPTION OF THE DATASET**

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

This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

## 5.3 ATTRIBUTE INFORMATION:

* school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
* 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 - “ 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')
* 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 - 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)

**5.4** **Methodology**

Since universities are prestigious places of higher education, students’ retention in these universities is a matter of high concern. It has been found that most of the students’ drop-out from the universities during their first year is due to lack of proper support in undergraduate courses. Due to this reason, the first year of the undergraduate student is referred as a “make or break” year. Without getting any support on the course domain and its complexity, it may demotivate a student and can be the cause to withdraw the course.

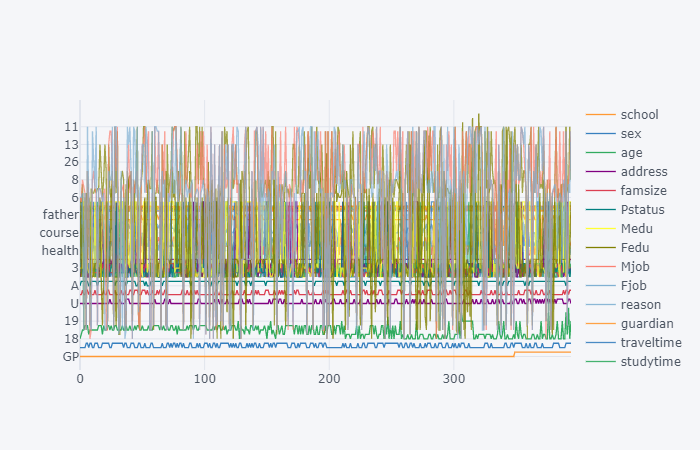
There is a great need to develop an appropriate solution to assist students retention at higher education institutions. Early grade prediction is one of the solutions that have a tendency to monitor students’ progress in the degree courses at the University and will lead to improving the students’ learning process based on predicted grades.

Using machine learning with Educational Data Mining can improve the learning process of students. Different models can be developed to predict students’ grades in the enrolled courses, which provide valuable information to facilitate students’ retention in those courses. This information can be used to early identify students at-risk based on which a system can 1 suggest the instructors to provide special attention to those students. This information can also help in predicting the students’ grades in different courses to monitor their performance in a better way that can enhance the students’ retention rate of the universities.

Using various packages such as cufflinks, seaborn & matplotlib to represent the data along with different attributes graphically or pictorially to analyse the dataset for predicting the Final Grade(G3).

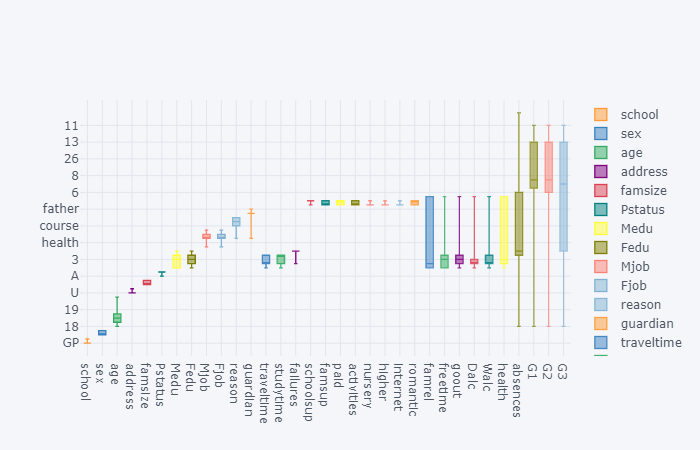
**Experimental Results**

5.4.1 - KDE Plot to view all attributes using cufflinks

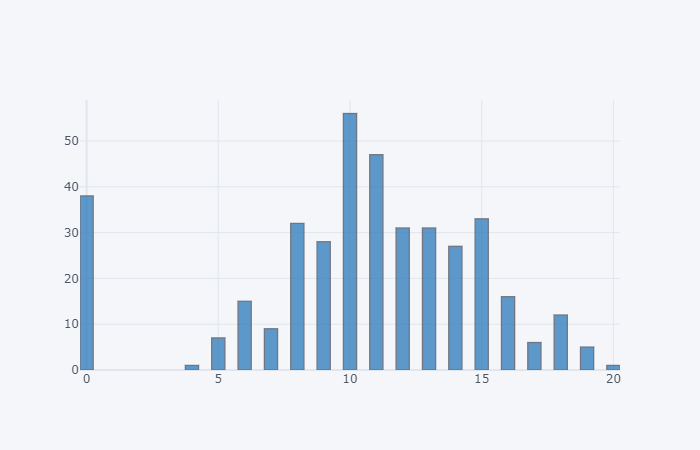


Observation: **cufflink** connects plotly with pandas to create graphs and charts of dataframes directly.

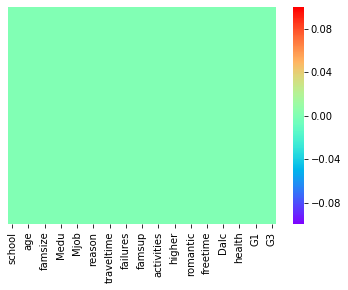
5.4.2 - Box Plot to view all attributes using cufflinks



5.4.3 - Histogram Plot for G3 (Final Grade) using cufflinks



4.4 - Pictorial representation of any null data present in the dataset



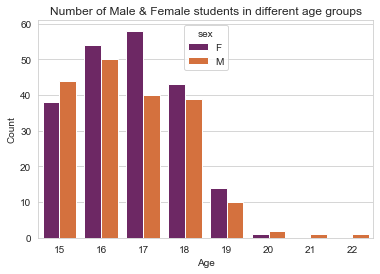
5.4.5 - Count Plot for Student Sex Attribute

# download (1)

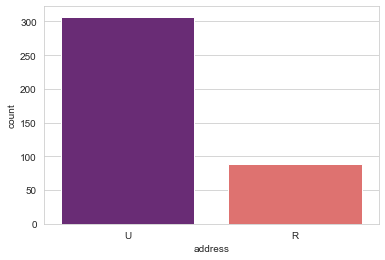
5.4.6 - Kernel Density Estimation for Age of Students.

# download (2)

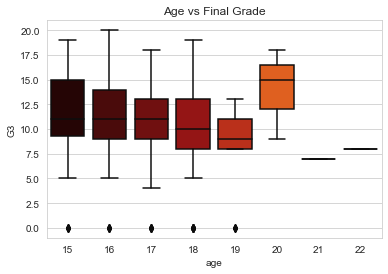
5.4.7 - Count PLot for Male & Female students in different age groups.



5.4.8 - Count Plot for students from Urban & Rural Region.



### 5.4.9 - Does age affect final grade?



Observation:

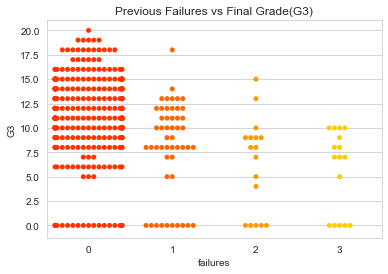
* Plotting the distribution rather than statistics would help us better understand the data.
* The above plot shows that the median grades of the three age groups(15,16,17) are similar. Note the skewness of age group 19. (may be due to sample size). Age group 20 seems to score highest grades among all.

## 5.4.10 - Do urban students perform better than rural students?

## download (7)

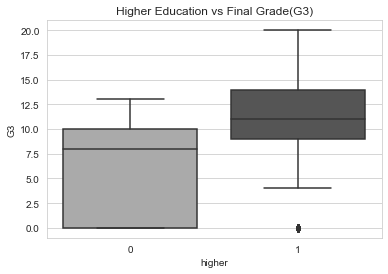
Observation: The above graph clearly shows there is not much difference between the grades based on location.

### 5.4.11 - Previous Failures vs Final Grade(G3)



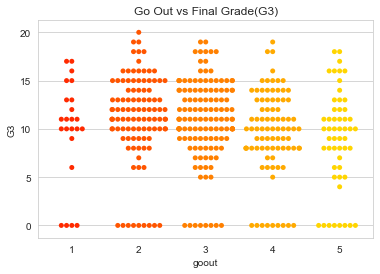
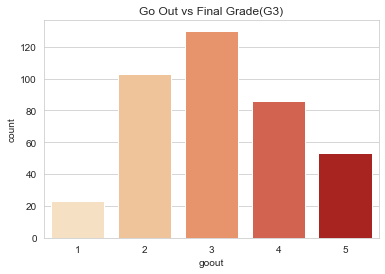
Observation: Student with less previous failures usually score higher.

5.4.12 - Family Education vs Final Grade(G3)



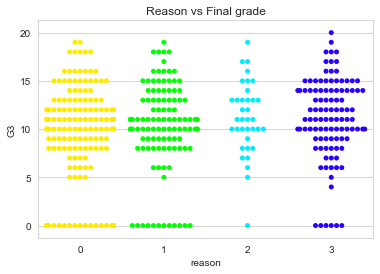
Observation: Students who wish to go for higher studies score more.

5.4.14 - Go Out vs Final Grade(G3)



Observation: The students have an average score when it comes to going out with friends & Students who go out a lot score less.

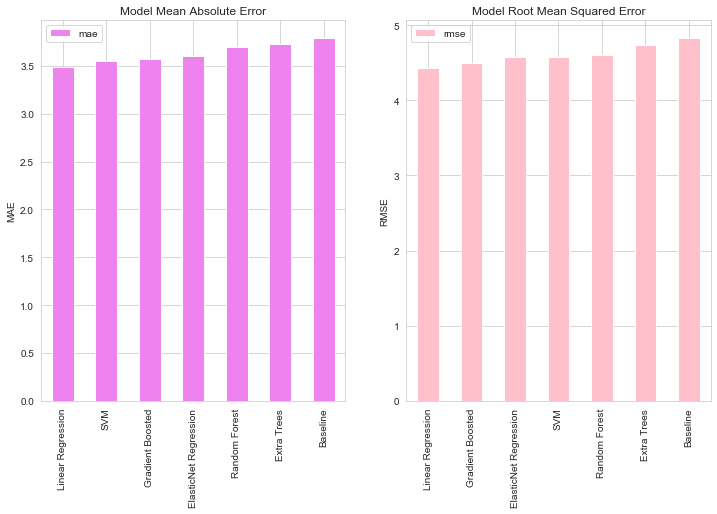
5.4.15 - Reason vs Students Count



**Observation :** The students have an equally distributed average score when it comes to reason attribute

**5.5 Conclusion**

As we see both MAE & Model RMSE that the Linear Regression is performing the best in both cases.

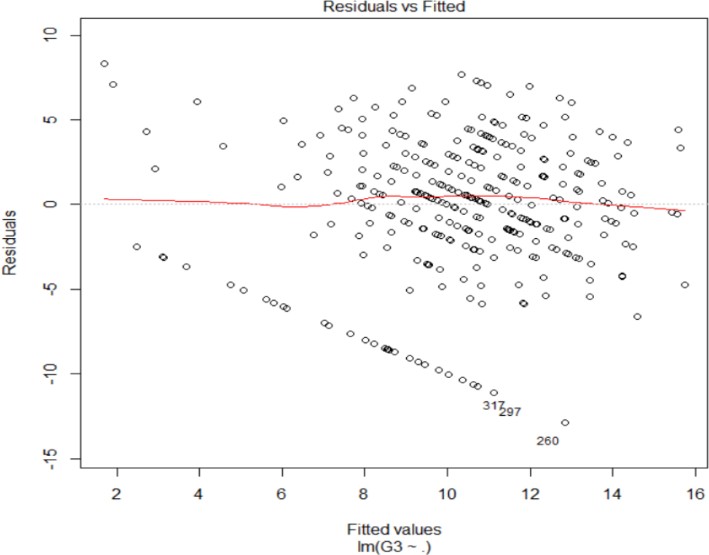


# CHAPTER-6

# EVALUATION

## Multiple Regression

When we performed multiple regression onto our data, our aim was to find a trend line and to find those variables who are significant. So that our model’s accuracy can be higher. Here is the result of plot and table. As we can see in the Figure [6a,](#_bookmark7) pattern line is more onto the side of clustered data. So we can say that our model is working good enough.



1. Residuals v/s fitted values for final grades

A table with numbers and a few digits

Description automatically generated with medium confidence

1. Selected Variables after checking the significance

Figure 6: Multiple Regression Results

## Stepwise Linear Regression

As we can see in Figure [7,](#_bookmark8) after performing both types of stepwise regression, we got these variables as significant variables.

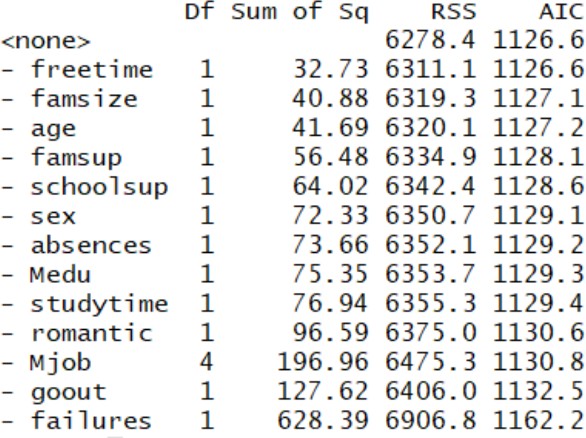


Figure 7: Stepwise Regression Results

## k-Nearest Neighbors

After selecting the 9 most significant variables I have applied different classification mod- els. When I performed k-Nearest Neighbors with k = 5 onto dataset’s significant variables, I was able to get the accuracy level of 82.53%. We can see the results in Figure [8.](#_bookmark9)

A number of numbers and letters

Description automatically generated with medium confidence

Figure 8: Result of k-Nearest Neighbors

## Logistic Regression

I have applied Logistic Regression onto all of the available variables with G3 as target variable and as we know that the target variable should be categorical in Logistic Re- gression so I assign the the values of G3 above 12 as High. When I checked the results of logistic regression, I was able to get the accuracy level of 96.19%. The result can be seen in Figure [9.](#_bookmark10)

A white background with green text

Description automatically generated

Figure 9: Result of Logistic Regression

## Principal Component Analysis

By performing dimensionality reduction I wanted to checked if it was able to increase the performance accuracy of my models. I have used Dummies package of R to create dummies variables. After using dummy package we had 57 variables. We can see the results of PCA in Figure [10.](#_bookmark11)

A graph of a principal component

Description automatically generatedA graph of a principal component

Description automatically generated

Figure 10: PCA Scree Plots

From the scree plots Figure [10](#_bookmark11) I came to know that the first 15 variance contains the 64.44% of the total variance. I have used these 15 Principal components to check if they were able to give the better results. I have applied Logistic Regression using these 15 Principal components and got the accuracy level of 92.82%. The results of which can be seen in the fig. [11.](#_bookmark12) After watching the results of Principal Component Analysis we can conclude that the Principal Component Analysis was not able to make my results more accurate.

A close-up of a computer code

Description automatically generated

Figure 11: Result of Logistic Regression using Principal Components

In the following table we can see all of the models and the result of all of the models in a tabular structure.

Table 1: Result of Models

|  |  |
| --- | --- |
| Model | Accuracy |
| K-nearest Neighbor | 82.53% |
| Logistic Regression | 96.19% |
| Logistic Regression with Principal Components | 92.82% |

## Discussion

In our research, we have used ideas of many researches in this area. As in their research, the researchers [Pagnotta and Hossain (2016)](#_bookmark22) used KNIME rule engine for data reduction, we took idea from their research and used Multiple Regression and Stepwise Regression for choosing the most significant variables. By their research, researchers came to know that the addiction of alcohol in students was one of the main reason for failures in ex- aminations but In our research, we came to know that the alcohol variables were not statistically significant. In their research, researchers [Pisutaporn et al.](#_bookmark24) [(2018)](#_bookmark24) applied different Classification and regression algorithms to check the relationship of different variables with student’s final grades. The researchers also came to know that there is no connection between alcohol and student final grades and in our research we also found out that it was true. In another research of [Shukla et al. (2018),](#_bookmark26) they tried to identify the variables which affect student performance. They used feature selection for pre-processing and applied many data mining models onto the dataset. The higher accuracy they got is 71.24 while they used Random Forest into their research but we were able to get the ac- curacy level of 96.19 with Logistic Regression while applying onto the all of the available variables so we can conclude that our method looks more good than author’s.

# CHAPTER-7

# GOOGLE COLAB

Google is quite aggressive in AI research. Over many years, Google developed AI framework called **TensorFlow** and a development tool called **Colaboratory**. Today TensorFlow is open-sourced and since 2017, Google made Colaboratory free for public use. Colaboratory is now known as Google Colab or simply **Colab**.

Another attractive feature that Google offers to the developers is the use of GPU. Colab supports GPU and it is totally free. The reasons for making it free for public could be to make its software a standard in the academics for teaching machine learning and data science. It may also have a long term perspective of building a customer base for Google Cloud APIs which are sold per-use basis.

Irrespective of the reasons, the introduction of Colab has eased the learning and development of machine learning applications.

So, let us get started with Colab.

**Google Colab - What is Google Colab?**

If you have used **Jupyter** notebook previously, you would quickly learn to use Google Colab. To be precise, Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook

**What Colab Offers You?**

As a programmer, you can perform the following using Google Colab.

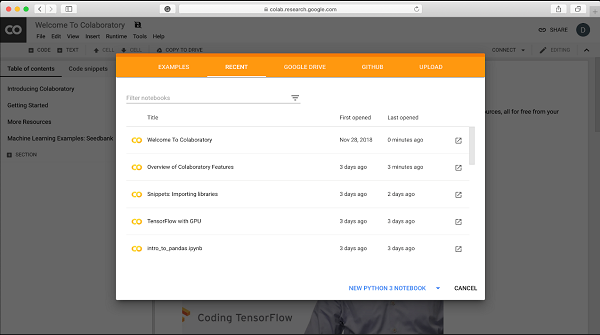
* Write and execute code in Python
* Document your code that supports mathematical equations
* Create/Upload/Share notebooks
* Import/Save notebooks from/to Google Drive
* Import/Publish notebooks from GitHub
* Import external datasets e.g. from Kaggle
* Integrate PyTorch, TensorFlow, Keras, OpenCV
* Free Cloud service with free GPU

**Google Colab - Your First Colab Notebook**

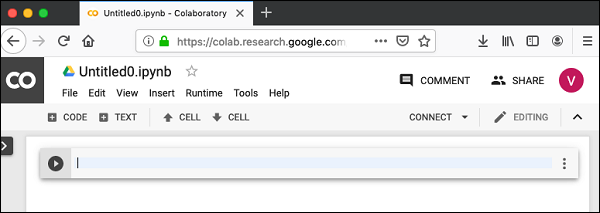
In this chapter, you will create and execute your first trivial notebook. Follow the steps that have been given wherever needed.

**Note** − As Colab implicitly uses Google Drive for storing your notebooks, ensure that you are logged in to your Google Drive account before proceeding further.

**Step 1** − Open the following URL in your browser − [https://colab.research.google.com](https://colab.research.google.com/notebooks/welcome.ipynb) Your browser would display the following screen (assuming that you are logged into your Google Drive) −

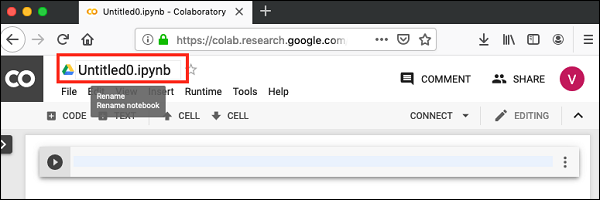


**Step 2** − Click on the **NEW PYTHON 3 NOTEBOOK** link at the bottom of the screen. A new notebook would open up as shown in the screen below.



**Setting Notebook Name**

By default, the notebook uses the naming convention UntitledXX.ipynb. To rename the notebook, click on this name and type in the desired name in the edit box as shown here −



We will call this notebook as **MyFirstColabNotebook**. So type in this name in the edit box and hit ENTER. The notebook will acquire the name that you have given now.

**Entering Code**

You will now enter a trivial Python code in the code window and execute it.

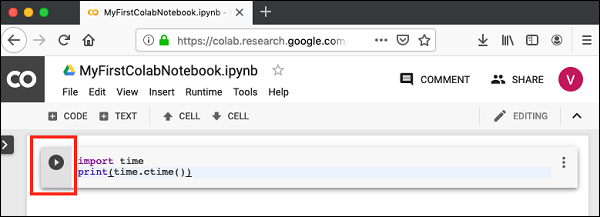
Enter the following two Python statements in the code window −

import time

print(time.ctime())

**Executing Code**

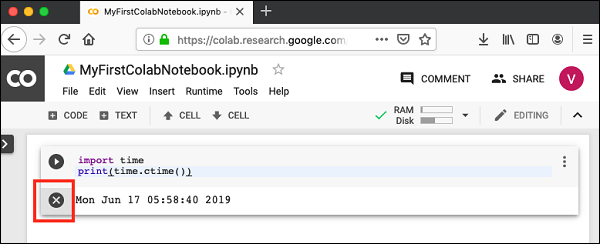
To execute the code, click on the arrow on the left side of the code window.



After a while, you will see the output underneath the code window, as shown here −

Mon Jun 17 05:58:40 2019

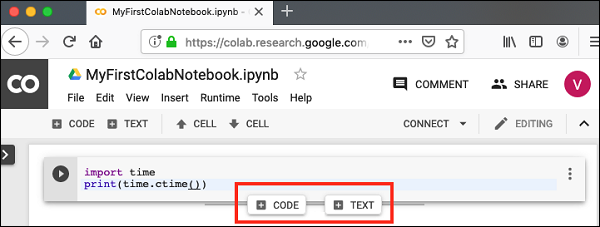
You can clear the output anytime by clicking the icon on the left side of the output display.



To add more code to your notebook, select the following **menu** options −

Insert / Code Cell

Alternatively, just hover the mouse at the bottom center of the Code cell. When the **CODE** and **TEXT** buttons appear, click on the CODE to add a new cell. This is shown in the screenshot below –



A new code cell will be added underneath the current cell. Add the following two statements in the newly created code window −

time.sleep(5)

print (time.ctime())

Now, if you run this cell, you will see the following output −

Mon Jun 17 04:50:27 2019

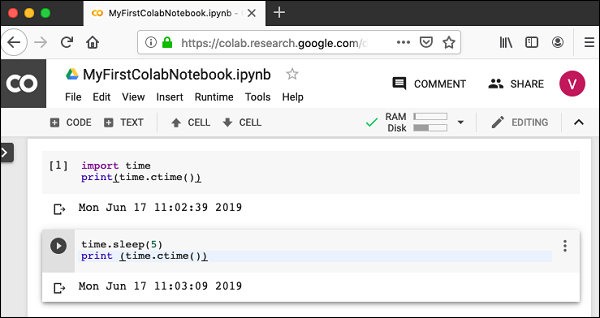
Certainly, the time difference between the two time strings is not 5 seconds. This is obvious as you did take some time to insert the new code. Colab allows you to run all code inside your notebook without an interruption.

**Run All**

To run the entire code in your notebook without an interruption, execute the following menu options −

Runtime / Reset and run all…

It will give you the output as shown below –



Note that the time difference between the two outputs is now exactly 5 seconds.

The above action can also be initiated by executing the following two menu options −

Runtime / Restart runtime…

or

Runtime / Restart all runtimes…

Followed by

Runtime / Run all

Study the different menu options under the **Runtime** menu to get yourself acquainted with the various options available to you for executing the notebook.

**Google Colab - Saving Your Work**

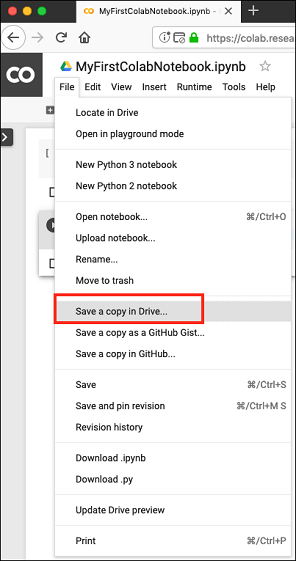
Colab allows you to save your work to Google Drive or even directly to your GitHub repository.

**Saving to Google Drive**

Colab allows you to save your work to your Google Drive. To save your notebook, select the following menu options −

File / Save a copy in Drive…

You will see the following screen –



The action will create a copy of your notebook and save it to your drive. Later on you may rename the copy to your choice of name.

**Google Colab - Invoking System Commands**

Jupyter includes shortcuts for many common system operations. Colab Code cell supports this feature

**Simple Commands**

Enter the following code in the Code cell that uses the system command echo.

message = 'A Great Tutorial on Colab by Tutorialspoint!'

greeting = !echo -e '$message\n$message'

greeting

Now, if you run the cell, you will see the following output −

['A Great Tutorial on Colab by Tutorialspoint!', 'A Great Tutorial on Colab by Tutorialspoint!']

**Getting Remote Data**

Let us look into another example that loads the dataset from a remote server. Type in the following command in your Code cell −

[!wget http://mlr.cs.umass.edu/ml/machine-learning-databases/adult/adult.data](http://mlr.cs.umass.edu/ml/machine-learning-databases/adult/adult.data) -P "/content/drive/My Drive/app"

If you run the code, you would see the following output −

--2019-06-20 10:09:53-- <http://mlr.cs.umass.edu/ml/machine-learning-databases/adult/adult.data>

Resolving mlr.cs.umass.edu (mlr.cs.umass.edu)... 128.119.246.96

Connecting to mlr.cs.umass.edu (mlr.cs.umass.edu)|128.119.246.96|:80... connected.

HTTP request sent, awaiting response... 200 OK

Length: 3974305 (3.8M) [text/plain]

Saving to: ‘/content/drive/My Drive/app/adult.data.1’

adult.data.1 100%[===================>] 3.79M 1.74MB/s in 2.2s

2019-06-20 10:09:56 (1.74 MB/s) - ‘/content/drive/My Drive/app/adult.data.1’ saved [3974305/3974305]

As the message says, the **adult.data.1** file is now added to your drive. You can verify this by examining the folder contents of your drive. Alternatively, type in the following code in a new Code cell −

import pandas as pd

data = pd.read\_csv("/content/drive/My Drive/app/adult.data.1")

data.head(5)

Run the code now and you will see the following output



Likewise, most of the system commands can be invoked in your code cell by prepending the command with an Exclamation Mark (!). Let us look into another example before giving out the complete list of commands that you can invoke.

**Listing Drive Contents**

You can list the contents of the drive using the ls command as follows −

!ls "/content/drive/My Drive/Colab Notebooks"

This command will list the contents of your Colab Notebooks folder. The sample output of my drive contents are shown here −

Greeting.ipynb hello.py LogisticRegressionCensusData.ipynb LogisticRegressionDigitalOcean.ipynb MyFirstColabNotebook.ipynb SamplePlot.ipynb

**Google Colab - Graphical Outputs**

Colab also supports rich outputs such as charts. Type in the following code in the Code cell.

import numpy as np

from matplotlib import pyplot as plt

y = np.random.randn(100)

x = [x for x in range(len(y))]

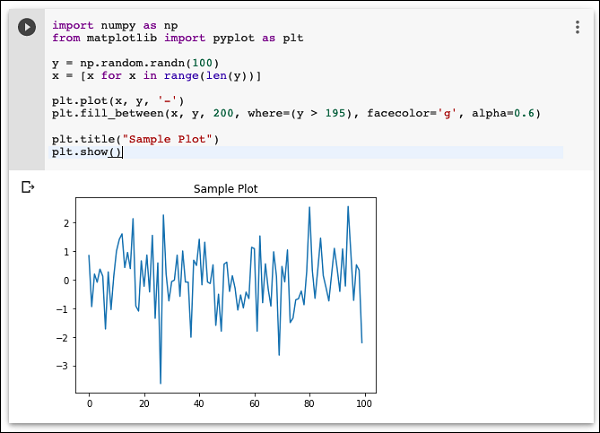
plt.plot(x, y, '-')

plt.fill\_between(x, y, 200, where = (y > 195), facecolor='g', alpha=0.6)

plt.title("Sample Plot")

plt.show()

Now, if you run the code, you will see the following output −



Note that the graphical output is shown in the output section of the Code cell. Likewise, you will be able to create and display several types of charts throughout your program code.

Now, as you have got familiar with the basics of Colab, let us move on to the features in Colab that makes your Python code development easier.

**Testing for GPU**

You can easily check if the GPU is enabled by executing the following code −

import tensorflow as tf

tf.test.gpu\_device\_name()

If the GPU is enabled, it will give the following output −

'/device:GPU:0'

**Listing Devices**

If you are curious to know the devices used during the execution of your notebook in the cloud, try the following code −

from tensorflow.python.client import device\_lib

device\_lib.list\_local\_devices()

You will see the output as follows −

[name: "/device:CPU:0"

device\_type: "CPU"

memory\_limit: 268435456

locality { }

incarnation: 1734904979049303143, name: "/device:XLA\_CPU:0"

device\_type: "XLA\_CPU" memory\_limit: 17179869184

locality { }

incarnation: 16069148927281628039

physical\_device\_desc: "device: XLA\_CPU device", name: "/device:XLA\_GPU:0"

device\_type: "XLA\_GPU"

memory\_limit: 17179869184

locality { }

incarnation: 16623465188569787091

physical\_device\_desc: "device: XLA\_GPU device", name: "/device:GPU:0"

device\_type: "GPU"

memory\_limit: 14062547764

locality {

bus\_id: 1

links { }

}

incarnation: 6674128802944374158

physical\_device\_desc: "device: 0, name: Tesla T4, pci bus id: 0000:00:04.0, compute capability: 7.5"]

**Google Colab - Conclusion**

Google Colab is a powerful platform for learning and quickly developing machine learning models in Python. It is based on Jupyter notebook and supports collaborative development. The team members can share and concurrently edit the notebooks, even remotely. The notebooks can also be published on GitHub and shared with the general public. Colab supports many popular ML libraries such as PyTorch, TensorFlow, Keras and OpenCV. The restriction as of today is that it does not support R or Scala yet. There is also a limitation to sessions and size. Considering the benefits, these are small sacrifices one needs to make