National College of Ireland

Student Grade Prediction

MSc Research Project
Data Analytics

Vivek Kumar
Student ID: x19201885

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Vivek Kumar
Student ID:	x19201885
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	16/08/2021
Project Title:	Student Grade Prediction
Word Count:	6411
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vivek Kumar
Date:	16th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Student Grade Prediction

Vivek Kumar
x19201885

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 Regression, Logistic Regression, Naive Bayes and K Nearest Neighbour.
Dataset Link: - Student Alcohol Consumption

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.

1.1 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.

1.2 Research Question

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

1.3 Plan of Paper

For this project, I am going to use Student Alcohol Consumption dataset which is a result 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 preparation 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 performing 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 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 we'll discuss about the methodology that we are going to use in my project. In the next section 4 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 we'll go through data preparation and then in the next section 6 we will talk about the results that we got after applying all the models and then in the next and last section 7 we'll conclude our project and in that section we will also talk about the future work that can be done in this area.

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) 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 Intelligence 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 impacting 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 correlation 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 researchers 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 (2017), 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. (2018) tried to analyse the importance of different variables by using different data mining algorithms. Their main goal was to study the student alcohol consumption. They used the same dataset "student alcohol consumption". 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 weekend 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 (2019), 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) 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 of 60% 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) 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. (2010), 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. (2008) 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, mentalities 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 (2007), tried to analyse the relationship between alcohol and grades of students. The researcher of this research conducted personal interviews 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. (2017), they distributed questionnaires to students of randomly 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 (2016), 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) tried to build a prediction model. For creating 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 researchers 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 coefficients. 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), 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. (2017), the researchers tried to elevate the performance 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 researchers 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.

3 Methodology

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 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.

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 Linear 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:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Figure 2: Logistic Regression

4.1.4 Naive 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:

$$P(X|y) = P(x_1|y) * P(x_2|y) * ... * P(x_n|y)$$

Figure 3: Naive Bayes

4.1.5 k-Nearest Neighbors

K-Nearest Neighbors algorithm is one of the algorithm which can be used for the regression 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:

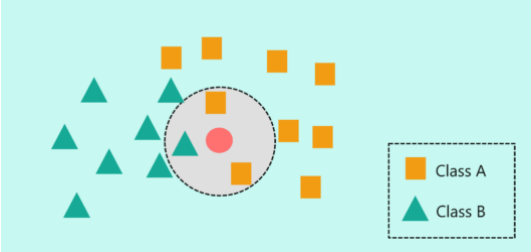


Figure 4: k-Nearest Neighbors

4.1.6 Principal Component Analysis

Principal Component Analysis is a dimensionality reduction strategy that is used to diminish 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 straightforwardness. 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 Evaluation 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.

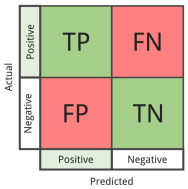


Figure 5: Confusion Matrix

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.

5 Implementation

5.1 Data Preparation

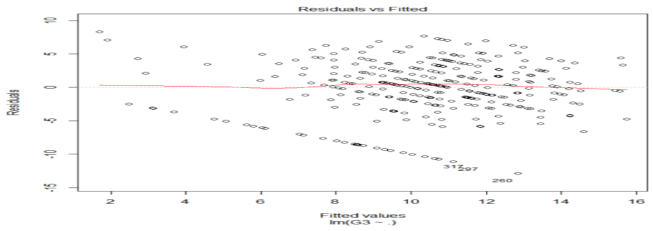
Data preparation and preprocessing is one of the most important process in which the analyst tried to clean and transform the raw data before the processing and analysis. It is a very important part before the processing and this step contains things like formatting of the data, join multiple datasets 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.

6 Evaluation

6.1 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, pattern line is more onto the side of clustered data. So we can say that our model is working good enough.



(a) Residuals v/s fitted values for final grades

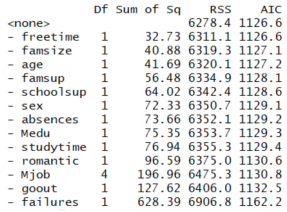
	Estimate	Std. Error	T-value	Pr(>|t|)
intercept	16.59743	4.58358	3.621	0.000349
sex	-1.22655	0.56493	-2.171	0.030765
age	-0.43309	0.24357	-1.778	0.076477
failures	-1.68769	0.38958	-4.332	2.07e-05*
schoolsup	1.64477	0.74195	2.217	0.02745*
freetime	0.47227	0.26573	1.777	0.076619
goout	-0.58106	0.25336	-2.293	0.022570
health	-0.37157	0.18309	-2.029	0.043368
absences	0.06339	0.03317	1.911	0.057010

(b) Selected Variables after checking the significance

Figure 6: Multiple Regression Results

6.2 Stepwise Linear Regression

As we can see in Figure 7, after performing both types of stepwise regression, we got these variables as significant variables.



	Df	Sum of Sq	RSS	AIC
<none>			6278.4	1126.6
- freetime	1	32.73	6311.1	1126.6
- famsize	1	40.88	6319.3	1127.1
- age	1	41.69	6320.1	1127.2
- famsup	1	56.48	6334.9	1128.1
- schoolsup	1	64.02	6342.4	1128.6
- sex	1	72.33	6350.7	1129.1
- absences	1	73.66	6352.1	1129.2
- Medu	1	75.35	6353.7	1129.3
- studytime	1	76.94	6355.3	1129.4
- romantic	1	96.59	6375.0	1130.6
- Mjob	4	196.96	6475.3	1130.8
- goout	1	127.62	6406.0	1132.5
- failures	1	628.39	6906.8	1162.2

Figure 7: Stepwise Regression Results

6.3 k-Nearest Neighbors

After selecting the 9 most significant variables I have applied different classification models. 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.

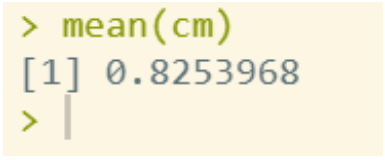


Figure 8: Result of k-Nearest Neighbors

6.4 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 Regression 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.

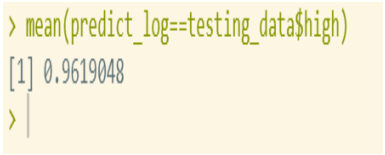


Figure 9: Result of Logistic Regression

6.5 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.

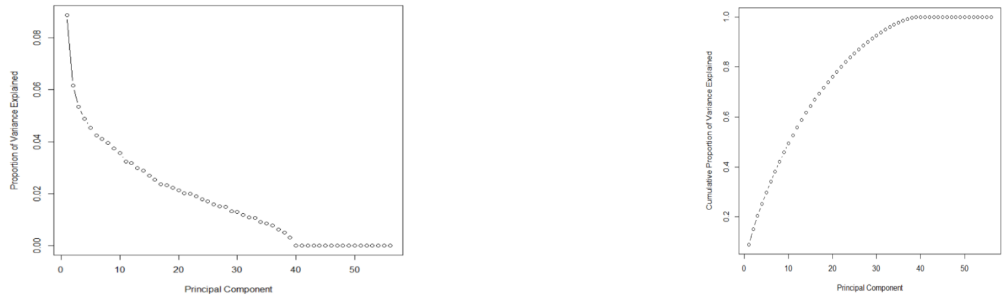


Figure 10: PCA Scree Plots

From the scree plots Figure 10 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. 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.

```
> predict_log <- round(predict_log)
> mean(predict_log==test_set_pca$G3)
[1] 0.9282297
>
>
```

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%

6.6 Discussion

In our research, we have used ideas of many researches in this area. As in their research, the researchers Pagnotta and Hossain (2016) 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 examinations but In our research, we came to know that the alcohol variables were not statistically significant. In their research, researchers Pisutaporn et al. (2018) 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), 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 accuracy 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.

7 Conclusion and Future Work

7.1 Conclusion

To sum up, we can say that surprisingly the variables Dalc (Workday alcohol consumption) and Walc (Weekend alcohol consumption) were not significant, so I can say that the consumption of alcohol don't impact the grades of the students much and the variables like failures, health, free time were much more impactful and they were also much far significant than Walc and Dalc. We also found out that all the classification algorithms that I have applied in my research returned almost similar accuracy level with Logistic Regression performing best with 96.19% Accuracy and the Naive Bayes Model was not able to perform well so I removed it from the research and we also came to know that the Principal Component Analysis also returned the same results and did not make any difference to the Logistic Regression model.

7.2 Future Work

In this project we have applied many techniques and got almost similar results but in our research we found out that the Logistic Regression performed best with 96.19% Accuracy. In the future researchers can apply clustering techniques onto the dataset to gain some new insights. After the use of clustering, we will also come to know if the students with similar interests and similar knowledge got into same cluster or not. In the future, researchers can apply clustering techniques like K means and Hierarchical Clustering.

Acknowledgement

I would like to thank my supervisor Dr. Christian Horn who gave me the golden opportunity to do this project and also helped me a lot in completing this project. Due to him, I came to know about many new things. I would also like to thanks my parents and family for their encouragement and motivations. I would also like to thank my friends who helped me by providing their valuable feedback and in the last I specially want to thank my late father who always inspires me and supported me in every possible way.

References

Aissaoui, O. E., Madani, Y. E. A. E., Oughdir, L., Dakkak, A. and Allioui, Y. E. (2019). A multiple linear regression-based approach to predict student performance, *Advances in Intelligent Systems and Computing* **1102**.

Butler, A. B., Dodge, K. D. and Faurote, E. J. (2010). College student employment and

drinking: A daily study of work stressors, alcohol expectancies, and alcohol consumption, *Journal of Occupational Health Psychology* **15**(3): 291–303.

Chudasama, R. and Joglekar, A. (2016). Innovative technique of improving student performance using data mining algorithm, *International Journal of Computer Science Trends and Technology (IJCST)* **4**(2).

Davoren, M. P., Dahly, D., Shiely, F. and Per, I. J. (2017). Alcohol consumption among university students: A latent class analysis, *Drugs: Education, Prevention and Policy* .

ElTayeby, O., Eaglin, T., Abdullah, M., Burlinson, D. and Dou, W. (2018). A feasibility study on identifying drinking-related contents in facebook through mining heterogeneous data, *Health Informatics Journal* **1**(12).

Htet, H., Saw, Y. M., Saw, T. N., Htun, N. M. M., Mon, K. L., Cho, S. M., Thike, T., Khine, A. T., Kariya, T. and Hamajima, E. Y. N. (2020). Prevalence of alcohol consumption and its risk factors among university students: A cross-sectional study across six universities in myanmar, *PLoS ONE* **15**(2).

Kitsantas, P., Kitsantas, A. and Anagnostopoulou, T. (2008). A cross-cultural investigation of college student alcohol consumption: A classification tree analysis, *The Journal of Psychology: Interdisciplinary and Applied* **142**(1): 5–20.

Pagnotta, F. and Hossain, A. (2016). Using data mining to predict secondarty school student alcohol consumption.

Pal, S. and Chaurasia, V. (2017). Performance analysis of students consuming alcohol using data mining techniques, *International Journal of Advance Research in Science and Engineering* **6**(2).

Pisutaporn, A., Chonvirachkul, B. and Sutivonhg, D. (2018). Relevant factors and classification of student alcohol consumption, *2018 IEEE International Conference on Innovative Research and Development (ICIRD)* **1**(6).

Sakaray, P., Kankariya, S., Lulla, C., Agarwal, Y. and Alappanavar, P. (2017). Review on student academic performance prediction using data mining techniques, *International Journal of Advanced Research in Computer and Communication Engineering* **6**(2).

Shukla, A. K., Singh, P. and Vardhan, M. (2018). Predicting alcohol consumption behaviours of the secondary level students, *International Conference on Internet of Things and Connected Technologies* **17**(4): 1369.

Singleton, R. A. (2007). Collegiate alcohol consumption and academic performance, *J Stud Alcohol Drugs* **68**(4): 548–55.

Trivedi, T. and Kotak, D. (2019). Exploring prediction modeling of students alcohol and drug addiction affecting performance using data mining approach, *International Journal of Engineering Research Technology (IJERT)* **8**(12).