STUDENT GRADE PREDICTION

ANALYSIS REPORT

Guided by:

Dr. Amir H. Gandomi

Gaurav Sawant

Vikram Singh

Vipul Gajbhiye

Fall 2017

Table of Contents

Contents

1. Abstract 3
2. Introduction 4
3. Problem Description 5
4. Evaluation of Database 6
5. Data Processing and Preparation 7
6. Multiple Regression 9
6.1 Results and Analysis: 10
6.2 Stepwise Linear Regression 11
6.3 Regression Results and Analysis: 11
7. Classification Methods 13
7.1 Logistic Regression algorithm and Results 13
7.2 Naïve Bayes Algorithm and Results 15
7.3 k-Nearest Neighbors and Results 16
8. Dimensionality Reduction Methods 18
8.1 Principal Component Analysis: 18
8.2 Normalizing before performing PCA: 19
8.3 PCA Results, Comparison and Analysis: 19
9. Conclusion 22
10. Future Scope 23
11. References 25

1. Abstract

The goal of this paper is to put forth the analyses and results obtained by us while trying answer the various questions that we encountered while analyzing the chosen dataset which in our case is the ‘Student Alcohol Consumption’ dataset (Source: Student Alcohol Consumption, https://www.kaggle.com/uciml/student-alcohol-consumption). The data is obtained in a survey of students in a math course in secondary school and it contains interesting information about demographic and social factors (Student Alcohol Consumption, kaggle.com). We have tried to predict the final student grade using social and demographic information and have used multivariate analytical methods like Multiple Regression, Step-wise regression, classification techniques like Logistic Regression, Naïve Bayes algorithm, K Nearest Neighbors algorithm and dimensionality reduction using Principal Component Analysis. A detailed report of all the analysis and results that we got by using the aforementioned techniques along with their interpretations and discussions can be found in the following sections.

1. Introduction

With several social and demographic factors available including workday alcohol consumption and weekend alcohol consumption, our aim is to assess what factors really play a significant role in determining the final grade that a student gets. The dataset conations information of 395 students and in all, there are 33 variables or attribute attached to each student. In order to perform meaningful analysis on the data we firstly perform data some preparation steps. This includes taking care of categorical variables and missing values. After data preparation we found significant variables by performing multiple regression and step-wise regression. We then tried to predict grades that students would obtain by performing different classification algorithms. Lastly, we performed dimensionality reduction and tried to see if that makes any difference to the classification results.

In section 3 we will learn more about the problem that we are trying to address. Section 4 will give us a brief overview about the database and the different variables found in the database. In section 5 we will go through all the data preprocessing and data preparation performed on the dataset. In section 6 we will understand how we found the significant variables for predicting student grades by performing Multiple regression and step-wise regression. Then in section 7 we will have a look at the results obtained in various classification techniques which includes results obtained from Logistic Regression, Naïve Bayes and K-Nearest Neighbors algorithm. Section 8 will give us a brief overview about dimensionality reduction and Principal component analysis along with the analysis and discussion about the results. Section 9 will list all the learnings and conclusions from this project followed by section 10 where we will have a look at the future scope of this project. Lastly, we have a list of all the references.

It is to be noted that all the results are explained in the same section along with the analysis techniques. There is no separate section for results.

3. Problem Description

Student Alcohol Consumption obtained in a survey of students for math and Portuguese language courses in secondary school. It contains social, gender and study information about students. We are considering the database student alcohol consumption to determine how alcohol consumption on workday and weekend alcohol consumption to determine how a student's grade will be affected and also considering the demographic and social factors related to the student. The grade is the final grade being considered. The Demographic Information consists of sex, age, address etc. while social factors include how often the student goes out, what he/she does in the free time etc.

Student Alcohol Consumption, kaggle.com, https://www.kaggle.com/uciml/student-alcohol-consumption

4. Evaluation of Database

The Description regarding the attributes or columns of the database is as follows:

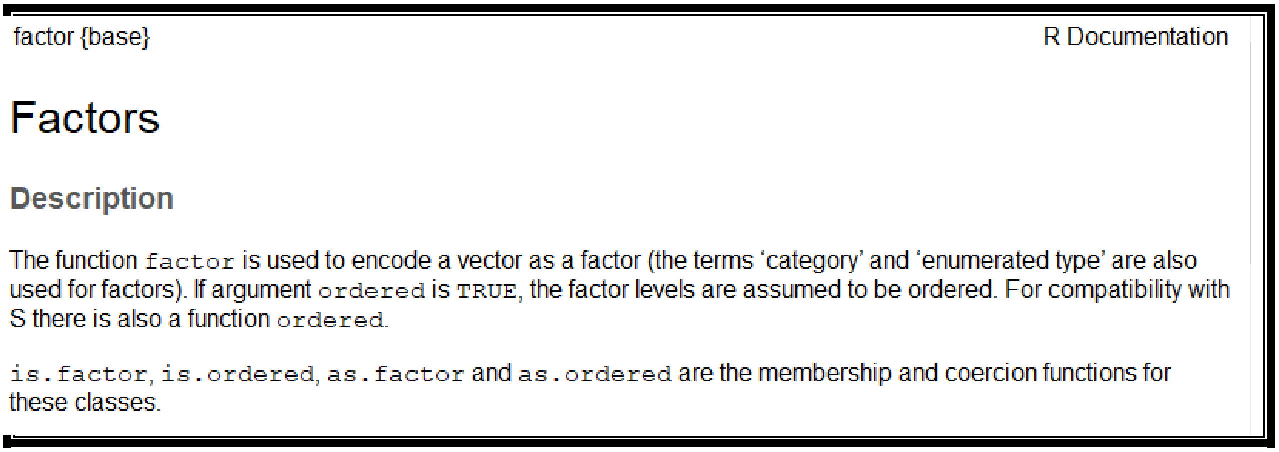
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - 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)
- TARGET VARIABLE G3 - final grade (numeric: from 0 to 20, output target)

Student Alcohol Consumption, kaggle.com, https://www.kaggle.com/uciml/student-alcohol-consumption

5. Data Processing and Preparation

“Data scientists spend 60% of their time on cleaning and organizing data. Collecting data sets comes second at 19% of their time, meaning data scientists spend around 80% of their time on preparing and managing data for analysis” (Press, Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, 2016, www.forbes.com). Preparing data is clearly one of the most time consuming but important part of analytics. All the steps that we performed while preparing the data for analysis are listed below-

1. **Missing Values:** There were no missing values in the dataset. Hence, no efforts we needed for managing missing values.
2. **Categorical Variables:** Categorical variables were converted into factor variables using the function ‘Factor’ in R-studio (Eremenko, de Ponteves, Machine Learning A-Z: Hands-On Python & R in Data Science, www.udemy.com). Variables encoded using the function could be used for all the analysis techniques expect Principal Component Analysis(PCA). For PCA, we used R function ‘dummy.data.frame’ for encoding the categorical variables into dummy variables found in the package ‘dummies’. A brief overview of both the functions ‘Factor’ and ‘dummy.data.frame’ can be seen in Fig. 5.1 and Fig. 5.2 respectively.



factor {base} R Documentation

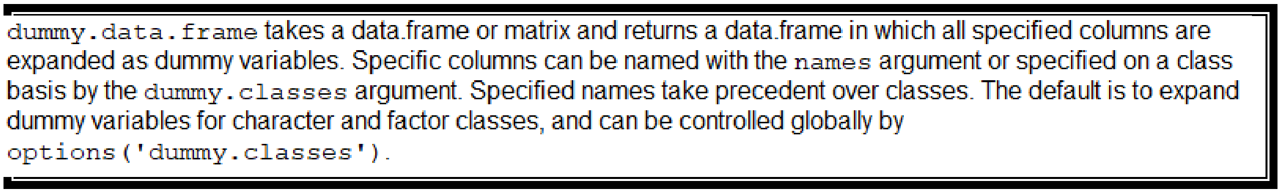
Factors

Description

The function `factor` is used to encode a vector as a factor (the terms ‘category’ and ‘enumerated type’ are also used for factors). If argument `ordered` is `TRUE`, the factor levels are assumed to be ordered. For compatibility with S there is also a function `ordered`.

`is.factor`, `is.ordered`, `as.factor` and `as.ordered` are the membership and coercion functions for these classes.

Fig. 5.1 Description of function ‘factor’

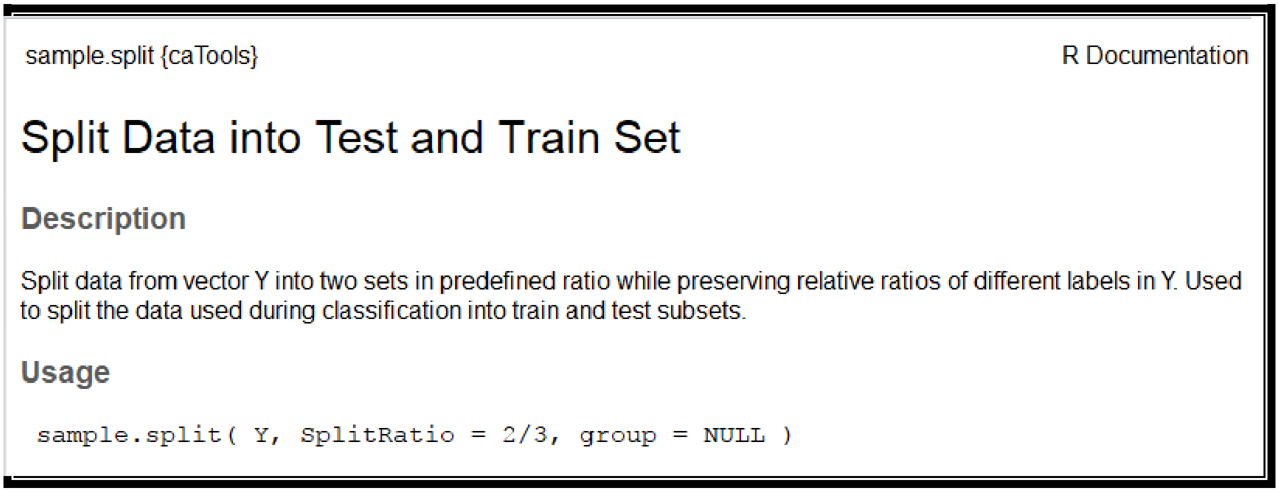


`dummy.data.frame` takes a data.frame or matrix and returns a data.frame in which all specified columns are expanded as dummy variables. Specific columns can be named with the `names` argument or specified on a class basis by the `dummy.classes` argument. Specified names take precedent over classes. The default is to expand dummy variables for character and factor classes, and can be controlled globally by `options('dummy.classes')`.

Fig. 5.2 Description of function ‘dummy.data.frame’

3. **Splitting Data into Training and Test sets:** The dataset is split into training and test sets using R function ‘sample.split’ in a 80:20 ratio respectively (Eremenko, de Ponteves, Machine Learning A-Z: Hands-On Python & R in Data Science, www.udemy.com).

Accordingly, we get a training set consisting of 316 observations and a test set consisting of 79 observations. Fig. 5.3 gives a brief description about the function 'sample.split'



sample.split {caTools} R Documentation

Split Data into Test and Train Set

Description

Split data from vector Y into two sets in predefined ratio while preserving relative ratios of different labels in Y. Used to split the data used during classification into train and test subsets.

Usage

```
sample.split( Y, SplitRatio = 2/3, group = NULL )
```

Fig. 5.3 Description of function 'sample.split'

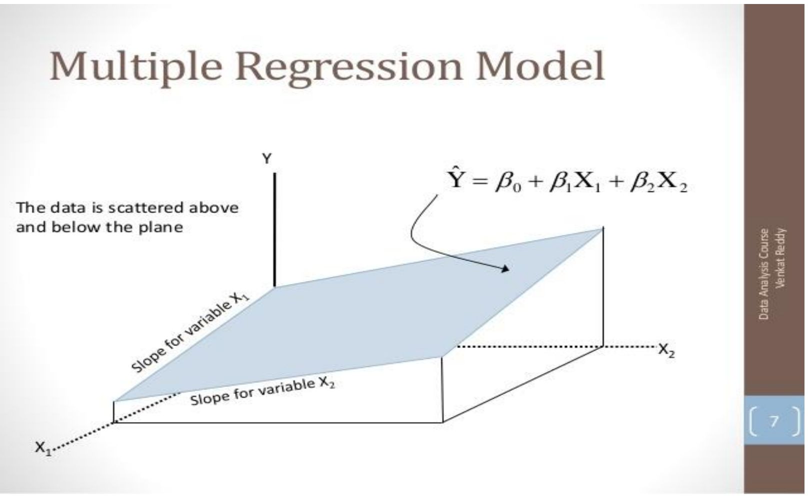
4. **Treatment of Target Variable:** Our target variable G3 is a continuous variable with a range from 0 to 20. This variable has been changed form a continuous variable to as variables which has Pass or Fail grade. Grades in the range of 0-9 have been converted to 'Fail' and grades in the range from 10-20 have been converted to 'Pass' grade. This is helpful in performing classification techniques.

6. Multiple Regression

Multiple regression is performed when an investigator wishes to examine the relationship between a single dependent (outcome) variable Y and a set of independent (predictor or explanatory) variables X1 to XP. The dependent variable Y is of the continuous type. The X variables are also usually continuous, although they can be discrete. (Afifi, May, Clark, Practical Multivariate Analysis, Fifth Edition). Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.

The population regression line for p explanatory variables x1, x2, ... , xp is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values b0, b1, ..., bp estimate the parameters β_0, β_1, ..., β_p of the population regression line. Since the observed values for y vary about their means μ_y, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \beta_p x_p$. The "RESIDUAL" term represents the deviations of the observed values y from their means μ_y, which are normally distributed with mean 0 and variance σ. The notation for the model deviations is ε.

Formally, the model for multiple linear regression, given n observations, is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... \beta_p x_{ip} + \varepsilon_i$ for i = 1,2, ... n.(http://www.stat.yale.edu)


Multiple Regression Model

Y
$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
The data is scattered above and below the plane
Slope for variable X1
Slope for variable X2
X1
X2
Data Analysis Course
Venkat Reddy
7

Fig. 6.1 Multiple Regression Model

(https://image.slidesharecdn.com/multipleregression-130320062840-phpapp02/95/multiple-regression-7-638.jpg?cb=1363760985)

6.1 Results and Analysis:

After performing multiple regression on the data, our goal was to achieve a trend line and also the significant variables that will help us to determine the accuracy of our model to a higher extent. The Plot and table for table for significant variables can be seen in Fig. 6.2:

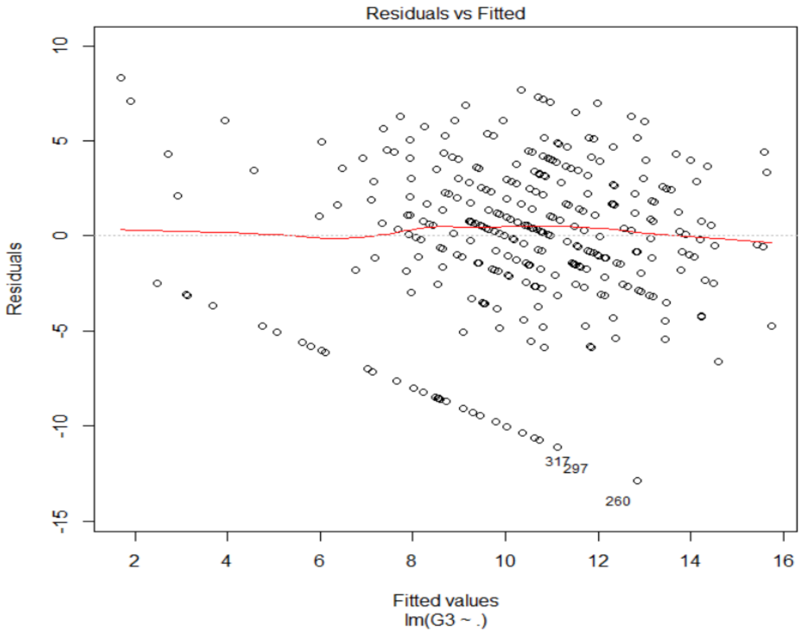


Fig. 6.2 Residuals v/s fitted values for final grade

As we can see the trend line is more towards the clustered data, we can say that our model is correct.

	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

Fig. 6.3 Selected Significant Variables after Multiple Regression

We can conclude most significant variables have $P_r(>|t|)$ less than 0.05.

6.2 Stepwise Linear Regression

Stepwise regression is a way to build a model by adding or removing predictor variables, usually via a series of F-tests or T-tests. The variables to be added or removed are chosen based on the test statistics of the estimated coefficients. (Statistic Regression, statisticshowto, http://www.statisticshowto.com/stepwise-regression/)

How Stepwise Regression Works:

The two ways that software will perform stepwise regression are:

1. **Backward Method**: Start the test with all available predictor variables, deleting one variable at a time as the regression model progresses. Use this method if you have a modest number of predictor variables and you want to eliminate a few. At each step, the variable with the lowest "F-to-remove" statistic is deleted from the model. The "F-to-remove" statistic is calculated as follows:

- A t-statistic is calculated for the estimated coefficient of each variable in the model

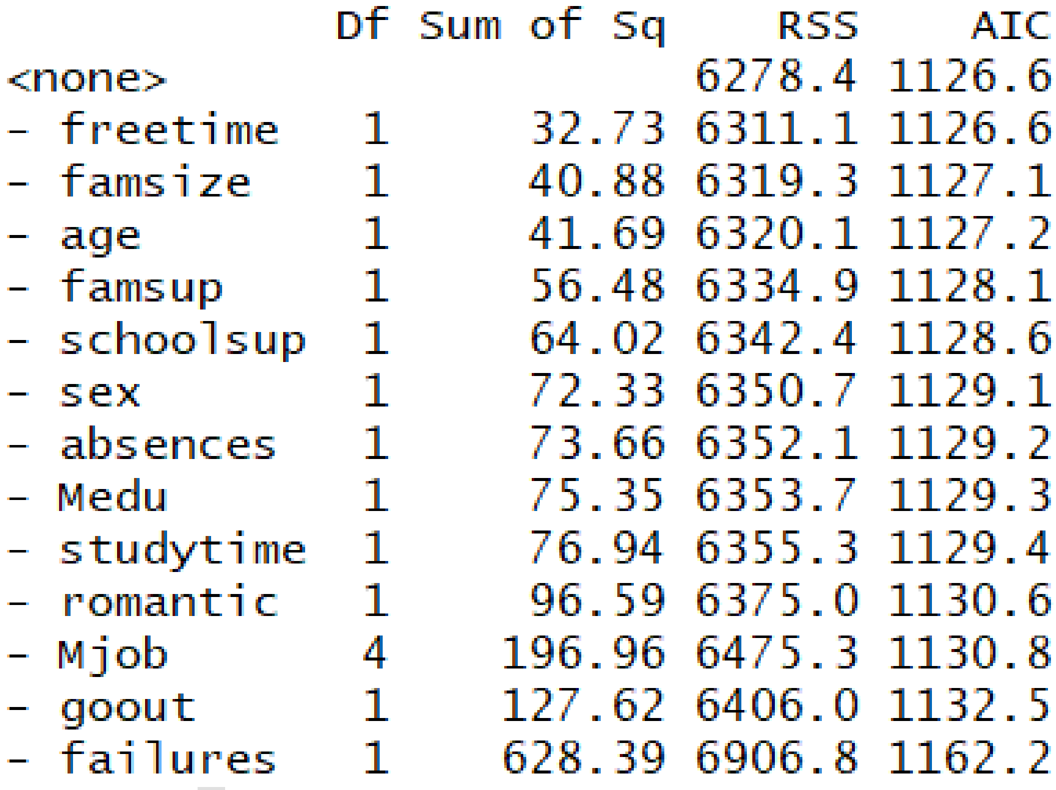
- The t-statistic is squared, creating the "F-to-remove" statistic

2. **Forward Method**: Start the test with no predictor variables, adding one at a time as the regression model progresses. If you have a large set of predictor variables, use this method. The "F-to-add" statistic is created using the same steps above, except the system will calculate the statistic for each variable not in the model. The variable with the highest "F-to-add" statistic is added to the model (Statistic Regression, statisticshowto, http://www.statisticshowto.com/stepwise-regression/).

In comparing the forward and the backward methods, we note that one advantage of the former is that it involves a smaller amount of computation than the latter. However, it may happen that two or more variables can together be a good predictive set while each variable taken alone is not very effective. In this case backward elimination would produce a better equation than forward selection. Neither method is expected to produce the best possible equation for a given number of variables to be included other than one or the total set (Afifi et al. Practical Multivariate Analysis, Fifth Edition).

6.3 Regression Results and Analysis:

After performing both forward and backward step regression, we achieved the same results i.e. the same number of significant variables. The results are as follows:



```
             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
```

Fig. 6.4 Significant variables obtained after stepwise regression

After analyzing the results, we obtain a total of 13 significant variables. However, we move forward with 8 significant variables which were obtained in multiple regression as we achieved higher accuracy while performing classification using these 8 variables.

7. Classification Methods

In Machine learning and statistics, classification methods are used for identifying which set of categories or populations does a new observation belong to and are considered as Supervised Learning techniques as a training set of correctly identified observations is available (Statistical Classification, Wikipedia, https://en.wikipedia.org/wiki/Statistical_classification). Classification algorithms are used when some decision or forecast is made on the basis of presently available information and these methods can be used repeatedly to make decisions in new situations (Sagar S. Nikam, A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental Journal of Classification Techniques in Data Mining Algorithms, www.computerscijournal.org). We have basically used three classification algorithms namely, Logistic Regression, Naïve Bayes algorithm and K-Nearest Neighbors algorithm.

7.1 Logistic Regression algorithm and Results

Logistic regression can be used whenever an individual is to be classified into one of two populations. When there are more than two groups, what is called polychotomous or generalized logistic regression analysis can be used. In the past, most of the applications of logistic regression were in the medical field, but it is also frequently used in epidemiologic research. It has been used, for example, to calculate the risk of developing heart disease as a function of certain personal and behavioral characteristics such as age, weight, blood pressure, cholesterol level, and smoking history. (Afifi et al., Practical Multivariate Analysis, Fifth Edition)

Logistic regression analysis requires knowledge of both the dependent (or outcome) variable and the independent (or predictor) variables in the sample being analyzed. The results can be used in future classification when only the predictor variables are known, like the results in discriminant function analysis. (Afifi et al. Practical Multivariate Analysis, Fifth Edition)
The logistic function has the form as seen in Fig. 7.1:

$$P_z = \frac{e^{\wedge}(\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta pXp)}{1+e^{\wedge}(\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta pXp)}$$

Fig 7.1 Equation for Logistic Regression

(Afifi et al. Practical Multivariate Analysis, Fifth Edition, Pg. 272)

This equation is called the **Logistic Regression Equation**, where Z is the linear function $\alpha + \beta 1X1 + \cdots + \beta PXP$. It may be transformed to produce a new interpretation. Specifically, we define the **Odds** as shown in Fig. 7.2:

$$\text{Odds} = \frac{P_z}{1- P_z}$$

Fig 7.2 Odds in Equation for Logistic Regression

(Afifi et al. Practical Multivariate Analysis, Fifth Edition, Pg. 272)

or in terms of **Pz**,

$$P_z = \frac{\text{Odds}}{1+\text{Odds}}$$

Fig 7.3 Odds in terms of Pz Equation for Logistic Regression

(Afifi et al. Practical Multivariate Analysis, Fifth Edition, Pg. 272)

Logistic regression always produces probabilities that are more than 0 and less than 1. A typical model plot of logistic regression is shown below in Fig. 6.4:

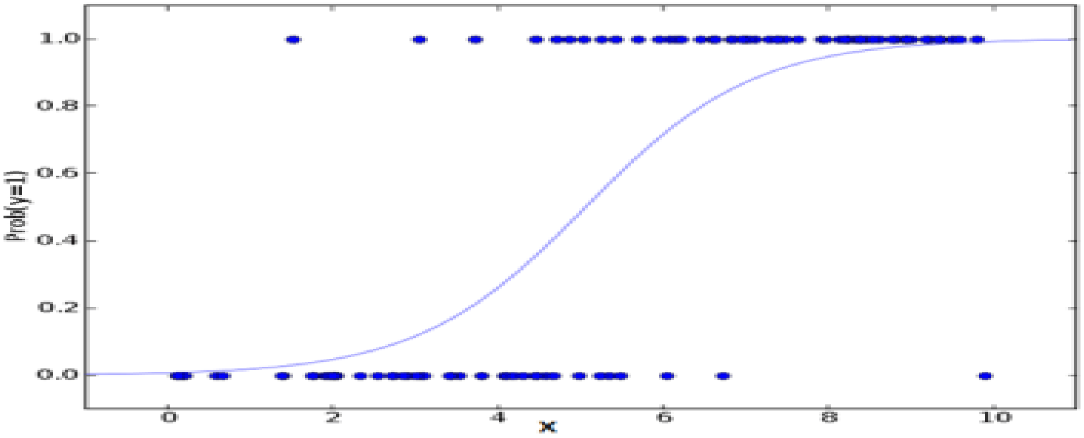
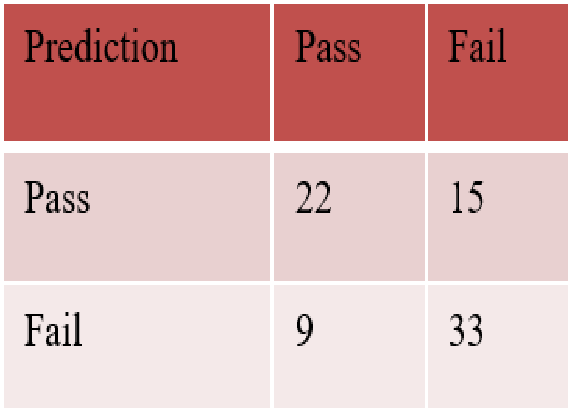


Fig. 7.4 A typical logistic model plot

Source:https://www.analyticsvidhya.com/wp-content/uploads/2015/11/plot.png

Results and Analysis:

The 8 significant variables obtained after performing multiple regression on the data set were further used to perform Logistic regression. Upon performing Logistic regression on the 8 significant variables, the accuracy achieved was 69.62%. The confusion matrix is as seen in Fig. 7.5:



Prediction	Pass	Fail
Pass	22	15
Fail	9	33

Fig. 7.5 Confusion Matrix of Logistic Regression

The curve we obtained after performing logistic regression bears much similarity to the ideal curve and is more than 0 and less than 1. The curve we obtained for the student alcohol consumption dataset upon performing logistic regression is as follows:

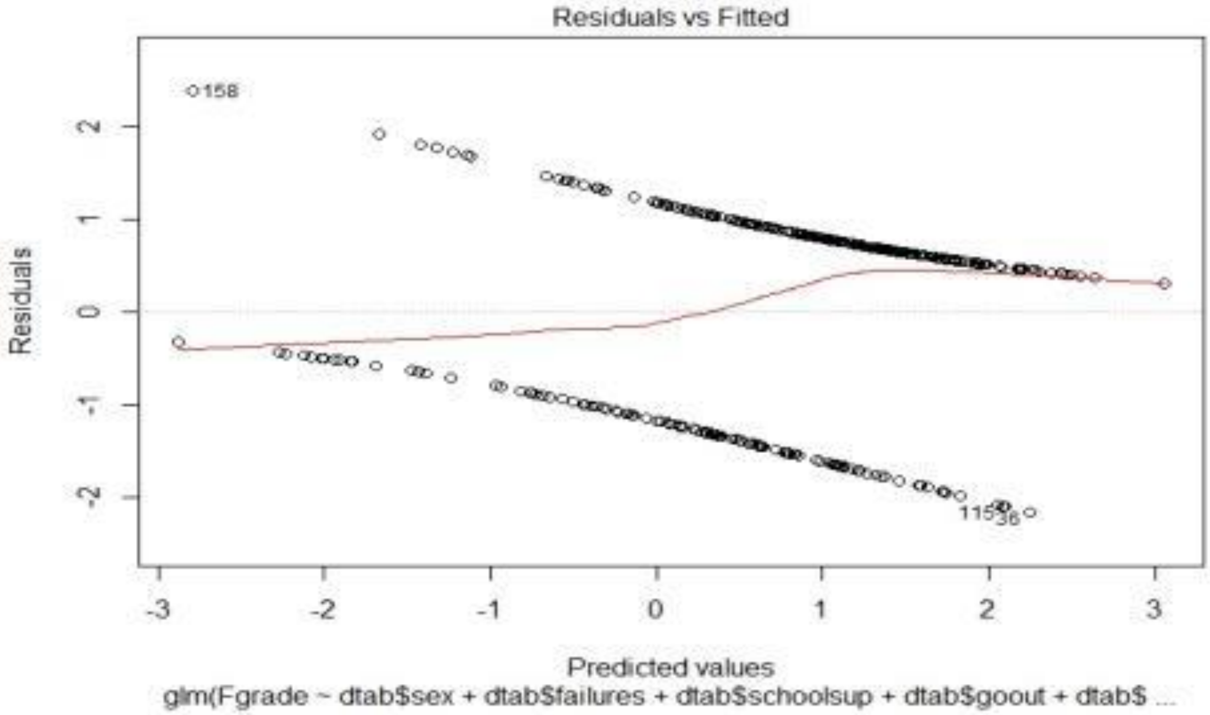


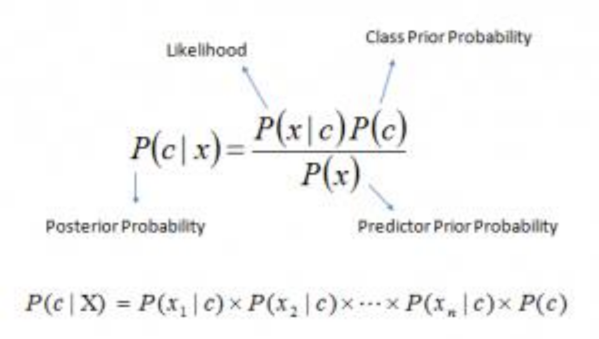
Fig. 7.6 Curve plot of Logistic regression

7.2 Naïve Bayes Algorithm and Results

Naïve Bayes Algorithm is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature. For example, a fruit may be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naïve'.

Source: Sunil Ray, September 11, 2017, 6 Easy Steps to Learn Naïve Bayes Algorithms (with codes in Python and R) https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Naïve Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naïve Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below: in Fig 7.7



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

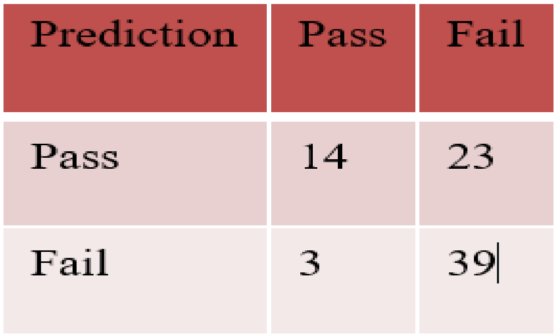
Fig. 7.7 Naïve Bayes equation

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class and P(x) is the prior probability of predictor.
- P(x|c) is the likelihood which is the probability of predictor given class.

Source: Sunil Ray, September 11, 2017, 6 Easy Steps to Learn Naïve Bayes Algorithms (with codes in Python and R) https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Results and Analysis:

Upon performing Naïve Bayes Algorithm for our dataset, we achived an accuracy of 67.08%. Notably, this accuracy percentage is quite close to what we achived using Logistic Regression Algorithm. The confusion matrix is as follows:



Prediction	Pass	Fail
Pass	14	23
Fail	3	39

Fig. 7.7 Confusion Matrix of Naïve Bayes Algorithm

7.3 k-Nearest Neighbors and Results

The KNN or *k*-nearest neighbors algorithm is one of the simplest machine learning algorithms and is an example of instance-based learning, where new data are classified based on stored, labeled instances. More specifically, the distance between the stored data and the new instance is calculated by means of a similarity measure. This similarity measure is typically expressed by a distance measure such as the Euclidean distance, cosine similarity or the Manhattan distance.

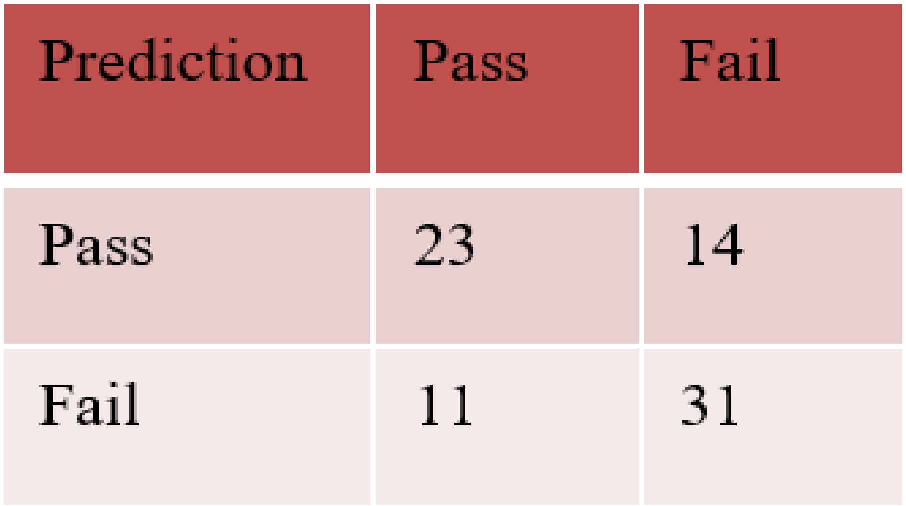
Source: https://www.datacamp.com/community/tutorials/machine-learning-in-r

The *k*-nearest neighbor algorithm does that after the distance of the new point to all stored data points has been calculated, the distance values are sorted, and the *k*-nearest neighbors are determined. The labels of these neighbors are gathered, and a majority vote or weighted vote is used for classification or regression purposes.

Source: https://www.datacamp.com/community/tutorials/machine-learning-in-r

Results and Analysis:

We performed k-nearest neighbors algorithm on our dataset and we calculated the accuracy of the algorithm using the confusion matrix seen in the figure below. The value of k was determined to be as 5. The accuracy achieved was 68.35%. Notably, once again the accuracy achieved here was quite close to the accuracies achieved using both Logistic regression and Naïve Bayes. The confusion matrix is as follows:



Prediction	Pass	Fail
Pass	23	14
Fail	11	31

Fig. 7.8 Confusion Matrix of k Nearest Neighbors for k = 5

8. Dimensionality Reduction Methods

In machine learning classification problems, there are at times a large number of variables or factors based on which the final classification is to be done. Most of these features are often corelated and hence, redundant and hence dimensionality reduction algorithms are used to reduce the number of features considered, by obtaining a set of principal variables (Uberoi, Introduction to Dimensionality Reduction, Geeks for Geeks A computer science portal for geeks, www.geeksforgeeks.org). We have only performed one Dimensionality Reduction algorithm and that is Principal Component Analysis. We'll go into more detail about it in the next sub-section.

8.1 Principal Component Analysis:

Principal Component Analysis (PCA) is used for getting simpler representation of a set of intercorrelated variables (Afifi et al. Practical Multivariate Analysis, Fifth Edition). Variables are not treated as dependent and independent variables, all variables are treated equally, and the original variables are transformed into new, uncorrelated variables called the Principal Components (Afifi et al. Practical Multivariate Analysis, Fifth Edition). These Principal Components are linear combinations of the original variables and each linear combination corresponds to a principal component.

Understanding PCA becomes very easy if we try to understand it in 2 dimensions or with 2 variables. The idea of PCA is to create 2 new components C_1 & C_2 as a linear combination of two variables x_1 and x_2 (Afifi et al. Practical Multivariate Analysis, Fifth Edition). The equation of principal components of only two variables can be written as seen in Fig. 8.1. This concept can be better understood from Fig. 8.2 which is a visual representation of the same idea.

$$C_1 = a_{11}x_1 + a_{12}x_2$$

$$C_2 = a_{21}x_1 + a_{22}x_2$$

Fig 8.1 Equations for Principal Components of 2 variables

(Afifi et al. Practical Multivariate Analysis, Fifth Edition, Pg. 360)

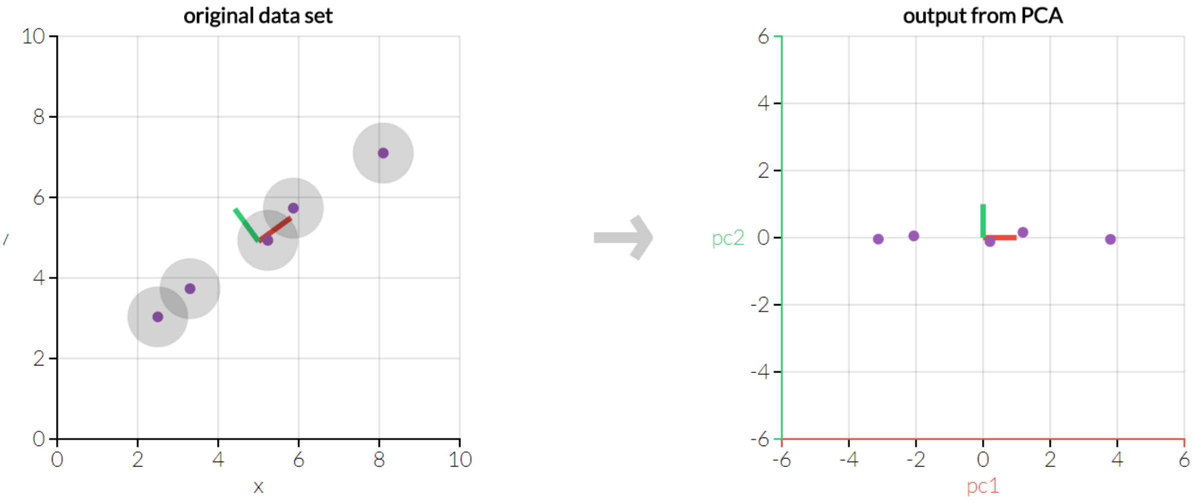


Fig. 8.2 Visual 2D example for first Principal Component

(Powell, Lehe, Principal Component Analysis Explained Visually, setosa.io)

We have seen an example of PCA in two-dimensions above but, it is necessary to understand that PCA remains equally effective even if the dimensionality is increased greatly. The first linear component is the linear combination of variables that has the maximum variance among other principal components and it accounts for the most variance in the data (11.1 Principal Component Analysis (PCA) Procedure, STAT 505 Applied Multivariate Statistical Analysis, onlinecourses.science.psu.edu). The second principal component explains a slightly lesser variance as compared to the first principal component and subsequently the amount of variance explained by principal components goes on decreasing.

8.2 Normalizing before performing PCA:

It is important to normalize the predictors before performing Principal Component Analysis to make sure that all the predictors are on the same scale as performing PCA on un-normalized variables will lead to large loadings for variables with large variance, thus creating dependence of Principal Components on the high variance variables (Analytics Vidhya Content Team, Practical Guide to Principal Component Analysis (PCA) in R & Python, www.analyticsvidhya.com). We have made sure that we have normalized all the predictors before performing Principal Component Analysis.

8.3 PCA Results, Comparison and Analysis:

We were curious to see if performing dimensionality reduction helps us in any way to increase the accuracy of the previously developed models. Hence, we decided to perform PCA. But, the ‘Factor’ function from R that we used for creating dummy variables in Regression and Classification techniques could not be used to perform PCA. We had to add new dummy variables and we did the same using the dummies package in R. After adding the dummy variables, the variable count went up to 57. Refer Fig. 8.3 and Fig. 8.4 to understand the results that we obtained after performing PCA.

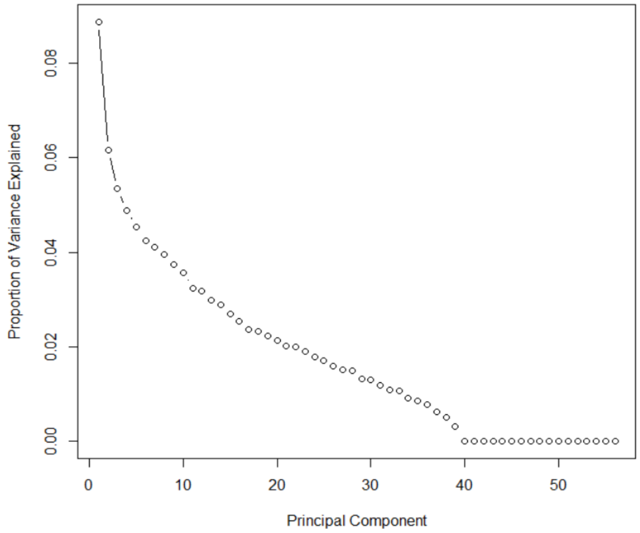


Fig. 8.3 Scree plot of Proportion of variance vs Principal Components

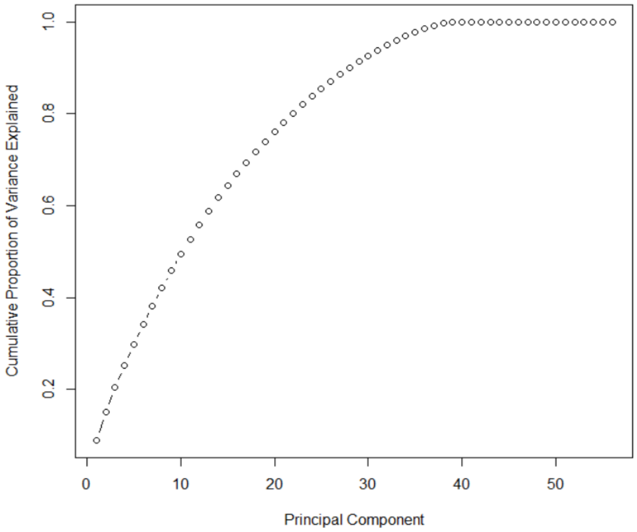
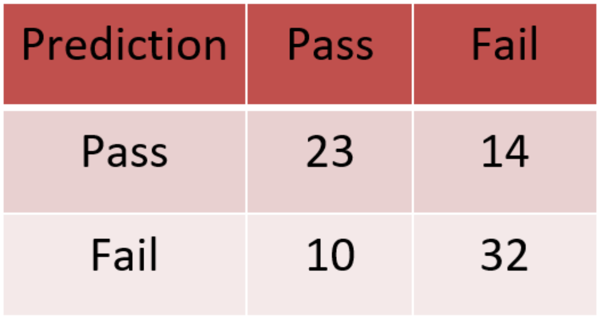


Fig. 8.4 Scree plot of Cumulative variances vs Principal Components

From Fig. 8.3 and Fig. 8.4 which are the scree plots we obtained after performing PCA we interpreted that the traditional methods for selecting the Principal components with most variance like the ‘Elbow Method’ were not suitable for our results. Instead, we chose the first **15** variables that cumulatively explained about **64.44%** of the total variance. We further tried to check if we get better classification results than previously obtained results using the first 15 Principal Components.

We performed Logistic Regression using the chosen principal components. We calculated the accuracy of logistic regression using the confusion matrix results seen in fig 8.5, and we got the exact same accuracy of 69.62% which we had already got using the significant variables form multiple regression. Thus, we concluded that Principal Components did not help us in getting more accurate results and hence, we did not perform any other classification technique using the chosen principal components.



Prediction	Pass	Fail
Pass	23	14
Fail	10	32

Fig. 8.5 Confusion Matrix of Logistic Regression using 15 Principal Component

9. Conclusion

In this section we briefly present all the conclusions, learnings and insights that we have gained after performing all the analyses.

Following are the major conclusions and insights-

1. Owing to the name of the dataset, we thought that the two variables related to alcohol consumption namely, ‘Dalc’ (workday alcohol consumption) and ‘Walc’ (weekend alcohol consumption) would play a significant role in determining the final grade. But our analysis showed that these variables do not play any significant role in determining the final grade that a student will get. Instead, variables such as ‘failures’ (number of past class failures), ‘gout’ (going out with friends), ‘schoolsup’ (extra educational support), ‘sex’ (student’s sex), ‘health’ (current health status), ‘absences’ (number of school absences), ‘age’ (student’s age) and ‘freetime’ (free time after school) are statistically significant and play an important role in determining the final grade.

2. All the three classification algorithms that we performed namely logistic regression, K-Nearest Neighbors and Naïve Bayes return a similar accuracy in the range of 65-70%. Hence, we can conclude that the accuracy predictions from classification algorithm lies in the aforementioned range.

3. We performed Principal Component Analysis to check if we can improve the accuracy of classification algorithms using principal components. We performed logistic regression using the selected principal components and we obtained accuracy figures very similar to what we had already got. Hence, we can conclude that performing Principal Component Analysis was not helpful in increasing the accuracy of our classification models.

10. Future Scope

In this project we have already performed standard and relevant Multivariate Analysis techniques like Multiple Regression, Stepwise regression, classification algorithms like Logistic regression, Naïve Bayes, K-Nearest Neighbors and dimensionality reduction using Principal Component Analysis (PCA). Apart from these techniques, we think performing various clustering techniques on the dataset can help in getting some new insights that we haven't found yet and there is definitely some value in performing clustering. After performing clustering we will be able to check if students having similar backgrounds actually got clustered in a same clusters or not.

Clustering is grouping a set of objects such that objects in the same group or cluster are more similar in some way or other than the objects put in other groups or clusters (Cluster Analysis, Wikipedia, https://en.wikipedia.org/wiki/Cluster_analysis). Clustering is one of the most important unsupervised learning problem and clustering basically deals with finding a structure in a collection of unlabeled data (A Tutorial on Clustering Algorithms, https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/). Clustering can be loosely defined as organizing objects into groups whose members are similar in some way or the other (A Tutorial on Clustering Algorithms, https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/). Fig. 10.1 gives us a good idea of how clustering actually works.

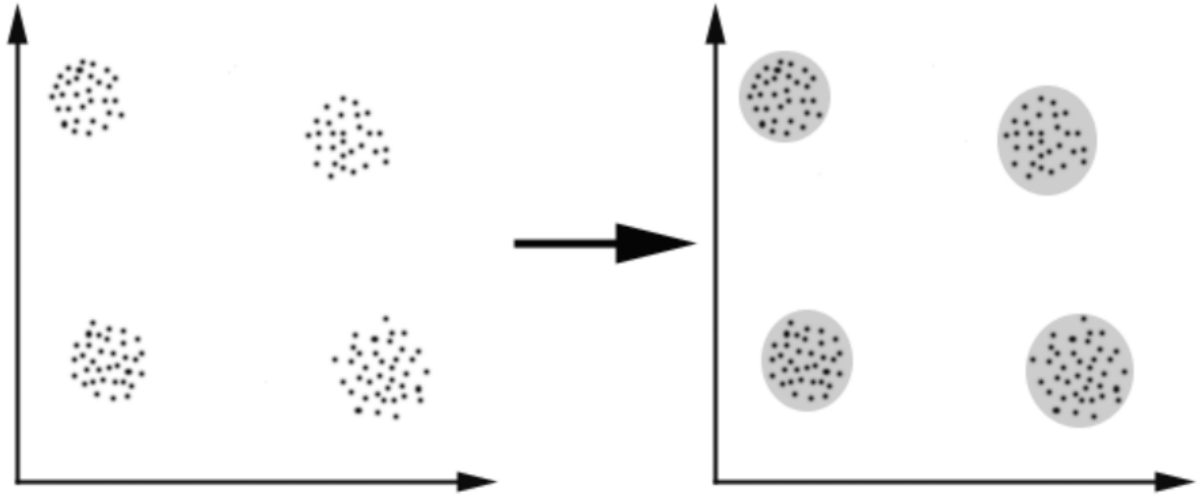


Fig. 10.1 Graphical Representation of Clustering

(A Tutorial on Clustering Algorithms, https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/)

We suggest that simple clustering algorithms like K Means Clustering and Hierarchical Clustering can be performed. K Means is an iterative clustering algorithm that aims to find the local maxima in each iteration and the number of iterations is pre-decided and equal to K (Saurav Kaushik, November 3, 2016, An Introduction to Clustering and different methods of clustering, Analytics Vidhya, www.analyticsvidhya.com). Hierarchical Clustering as the name suggests builds hierarchy of clusters; hierarchical clustering algorithm starts with data points assigned to cluster of their own and the algorithm terminates when there is only a single cluster left (Saurav Kaushik, November 3, 2016, An Introduction to Clustering and different methods of clustering, Analytics Vidhya, www.analyticsvidhya.com).

11. References

Abdelmonem Afififi, Susanne May, Virginia A. Clark, Practical Multivariate Analysis, Fifth Edition

Analytics Vidhya Content Team, Practical Guide to Principal Component Analysis (PCA) in R & Python, www.analyticsvidhya.com

Victor Powell, Lewis Lehe, Principal Component Analysis Explained Visually, setosa.io, http://setosa.io/ev/principal-component-analysis/

11.1 Principal Component Analysis (PCA) Procedure, STAT 505 Applied Multivariate Statistical Analysis, onlinecourses.science.psu.edu, https://onlinecourses.science.psu.edu/stat505/node/51

Anannya Uberoi, Introduction to Dimensionality Reduction, Geeks for Geeks A computer science portal for geeks, http://www.geeksforgeeks.org/dimensionality-reduction/

Student Alcohol Consumption, kaggle.com, https://www.kaggle.com/uciml/student-alcohol-consumption

Gill Press, Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, 2016, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#227d8be26f63

Kirill Eremenko, Hadelin de Ponteves, Machine Learning A-Z: Hands-On Python & R in Data Science, www.udemy.com

Multiple Regression, http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

Multiple Regression, https://image.slidesharecdn.com/multipleregression-130320062840-phpapp02/95/multiple-regression-7-638.jpg?cb=1363760985

Statistic Regression, statisticshowto , http://www.statisticshowto.com/stepwise-regression/)

Slide Share, https://image.slidesharecdn.com/multipleregression-130320062840-phpapp02/95/multiple-regression-7-638.jpg?cb=1363760985

Statistical Classification, Wikipedia, https://en.wikipedia.org/wiki/Statistical_classification

Sagar S. Nikam, A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental Journal of Classification Techniques in Data Mining Algorithms, http://www.computerscijournal.org/vol8no1/a-comparative-study-of-classification-techniques-in-data-mining-algorithms/

Sunil Ray, September 11, 2017, 6 Easy Steps to Learn Naïve Bayes Algorithms (with codes in Python and R https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

https://www.datacamp.com/community/tutorials/machine-learning-in-r

Cluster Analysis, Wikipedia, https://en.wikipedia.org/wiki/Cluster_analysis

A Tutorial on Clustering Algorithms, https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

Saurav Kaushik, November 3, 2016, An Introduction to Clustering and different methods of clustering, Analytics Vidhya, https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/