STUDENT BEHAVIOUR PREDICTION USING DATA MINING

Submitted in partial fulfillment

Of

Project in Bachelor of Technology

Submitted by
1. Y.MANASA (411582)
2. M. SAI SHUSHMA (411545)
3. Y. LAKSHMI SRAVANTHI (411583)

Under the Supervision of
Mr. BKSP Kumar Raju Alluri

NATIONAL INSTITUTE OF TECHNOLOGY,
Andhra Pradesh.

CONTENTS

SUMMARY

The project is about predicting student's grade and behavior. There will be influence of many factors on student behavior and grade. We have taken some of those factors into consideration and tried to predict student's grade and behavior. Many parents worry that their children are getting less grade. So, if we divide the students getting different grades into different groups, then we will know what is influencing the student to get a particular grade. So he/she can rectify their problems and can improve their grade. We have also predicted the levels of alcohol consumption for a student

Objectives:

- The main objective of our project is to predict student's performance.
- Training the existing dataset and use that as a model to predict students grade.
- To predict alcohol consumption of a student. That is at what level he/she is consuming alcohol.
- Use template matching to arrive at some association rules we are interested in which satisfy minimum confidence and support.
- Use clustering and divide the dataset into groups, where you can know the possible values of other attributes for a particular grade.
- To predict grade not only based on historical dataset but by considering some external factors such as friends influence, number of educated people around him, number of employed people etc.

Challenges:

- First we found our dataset in excel file. To work in weka we have to convert it into arff file. Initially to convert that we installed an app and tried a lot. But we cannot succeed in that because it need an excel 2007 version which is not available with us. Later with the help of notepad we finally able to complete that task. First excel file has to convert into csv file and later we can directly save it as .arff which give successful result.
- To find relation between attributes we used weka and found association rules but as our data set has more accuracy we found many rules even at high confidence and support values .So we cannot get any knowledge out of it.
- So we used template matching in R and fix left and right side of association rule and got some knowledge out of it.
- We wanted to make al1 the students into clusters. We thought to it in R. But in R we cannot get any sense out from that clusters .So we again go to weka which gave us meaningful clusters. The main difference in both is that in R the centroid of the cluster does not give any sense and taking only numeric values where as we get exact values in weka.

- For attribute subset selection we first thought of constructing correlation matrix and find the best attributes that related to class label. But we directly found an option in weka which directly gives required attributes as output.
- While performing template matching in R we faced a big issue that we cannot directly read csv because if we read csv file there are any numerical attributes and that must be converted to nominal then only we can get association rules. So what we did is we first convert csv into arff and using weka converted numeric into nominal and saved it as a separate file. We installed separate packages in R and read the arff file directly .Now we applied template matching easily.

Methodology:

- **Relationship mining**
  Association rule mining
  Template matching
  Correlation mining
- **Prediction**
  Classification
  Regression
- **Clustering**
- **Analysis of variance**
- **Preprocessing**
  Numeric to nominal

Outcome:

- The outcome of our project is predicting student's grades and finding how their grades are affected by some external factors.

- Also our project is used for predicting Alcohol Consumption of students and finding the effects of Alcohol Consumption.

1. Introduction:

Data mining refers to mining or extracting knowledge from large amount of data .Data mining focuses on extraction of information from a large set of data and transforms it to an easily interpretable structure for further use. As data is growing rapidly, there is a need to analyze large, complex and information datasets to achieve hidden information.
Data mining is divided into predictive and descriptive. Association mining and Clustering comes under descriptive and classification comes under predictive technique.
Understanding and analyzing the factors for poor performance is a complex and incessant process hidden in past and present information congregated from academic performance and students' behavior. Powerful tools are required to analyze and predict the performance of students scientifically.

I. Association mining:

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

II. Classification:

Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis. Also called sometimes called a Decision Tree, classification is one of several methods intended to make the analysis of very large data sets effective. Classification is used to map data in predefined groups.

III. Clustering:

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).

In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

TECHNIQUES USED:

Association Mining:

Aprioiri algorithm:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database

Key Concepts:

• Frequent Itemsets: The sets of item which has minimum support (denoted by Li for ith-Itemset).

• Apriori Property: Any subset of frequent itemset must be frequent.

• Join Operation: To find L k , a set of candidate k-itemsets is generated by joining Lk-1 with itself.

Template matching:

Filter the original data by using the itemsets that the users are interested in. At the time of producing the rule, making use of the template matching method to mine the association rule that the users are interested in and confirm the previous item and the consequent item of the association rule, we can confirm the mining direction of the association rule.

For example: A=>B is a template, all association rules
matching this template will be found.
Here A is set of attributes.
B is set of attributes.

Classification:

J48:

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable.

J48 is an extension of ID3. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Features of J48 Algorithm:

1. Both the discrete and continuous attributes are handled by this algorithm. A threshold value is decided by C4.5 for handling continuous attributes. This value divides the data list into those who have their attribute value below the threshold and those having more than or equal to it.
2. This algorithm also handles the missing values in the training data.
3. After the tree is fully constructed, this algorithm performs the pruning of the tree. C4.5 after its construction drives back through the tree and challenges to remove branches that are not helping in reaching the leaf nodes.

SVM:

In machine learning, **support vector machines** (**SVM)** are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear

classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Random Forest:

Random forests are ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set

Clustering:

K-means:

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering:

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3.

2. DATASET:

Source:

https://data.world/data-society/student-alcohol- consumption

Attributes:

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)

These grades are related with the course subject, Math or Portuguese:

G1 - first period grade (numeric: from 0 to 20)

G2 - second period grade (numeric: from 0 to 20)

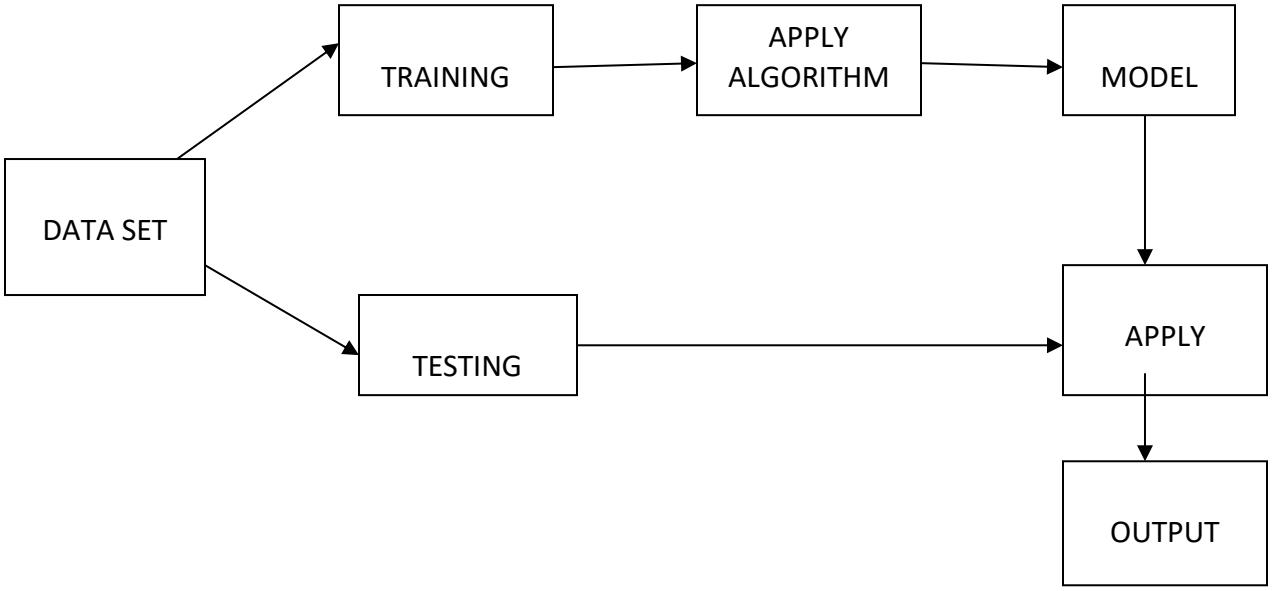
G3 - final grade (numeric: from 0 to 20, output target)

There are 649 tuples in the dataset.

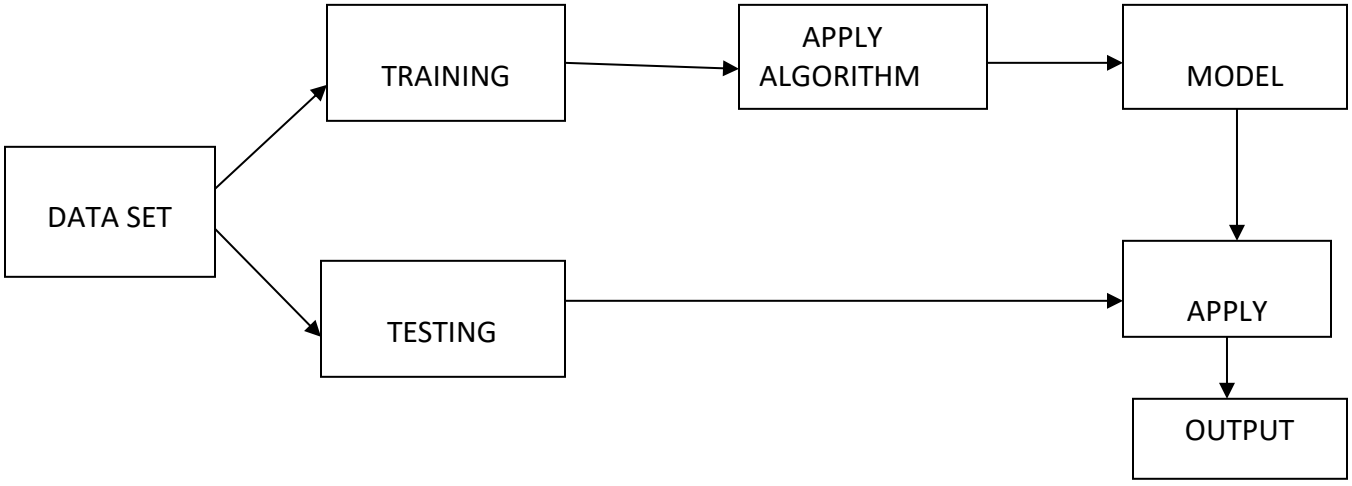
3. Schematic view of Project:

Data Flow Diagrams

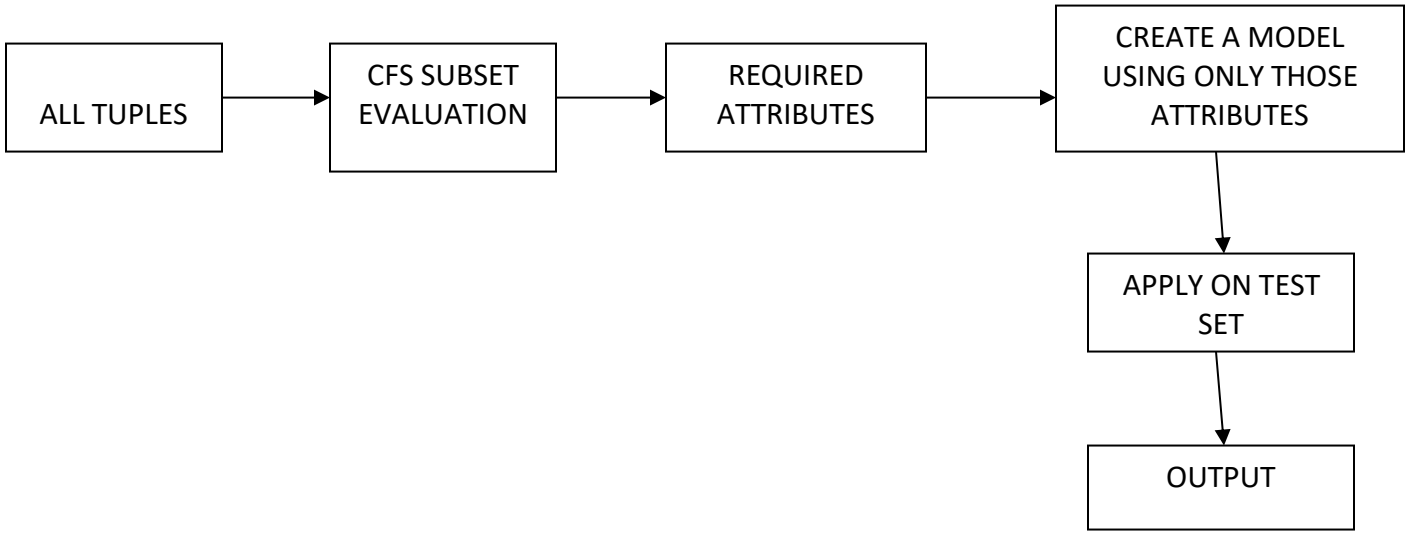
Module 1:



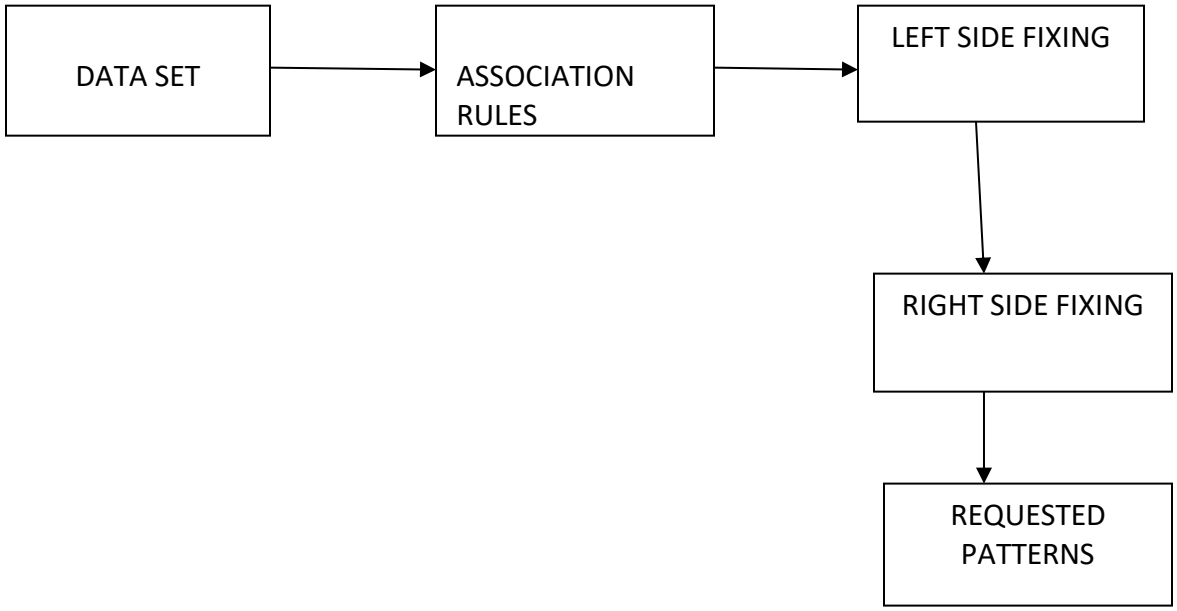
Module 2:



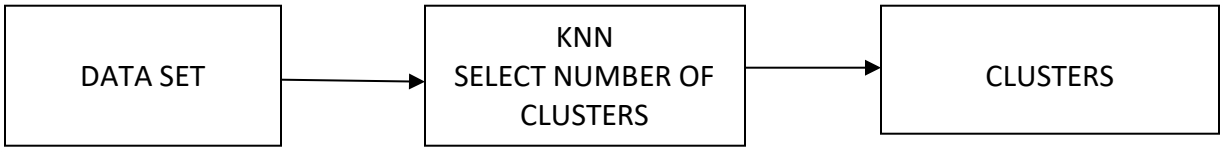
Module 3:



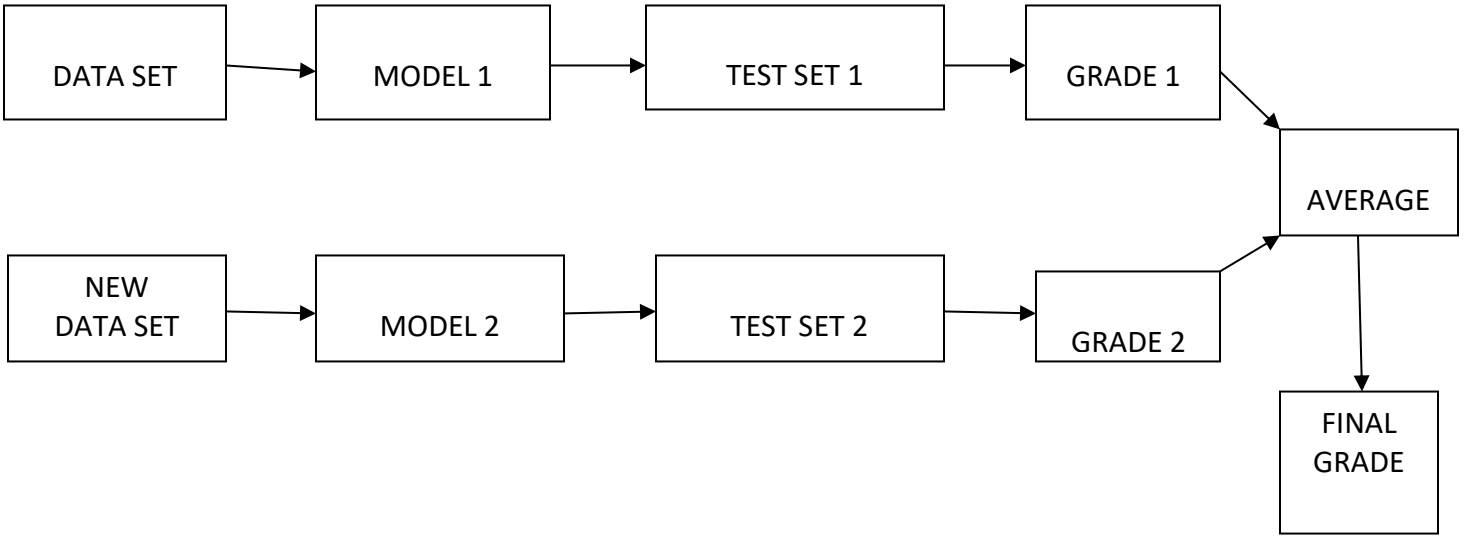
Module 4:



Module 5:



Module 6:



4. MODULES:

MODULE 1

Existing module 1:

Prediction of Grade

The availability of data has been growing rapidly, and there is a need to analyze huge amounts of data generated from this system. There are many techniques in data mining that can be applied to educational data, such as classification, clustering, and association rules to name a few. These techniques will help extract hidden knowledge and useful information.

Classification is one of the supervised learning techniques that build a model to classify a data item into a predefined class label. The aim of classification is to predict the future output based on the available data. Hence, we can predict grade of a student based on the attributes specified in the dataset.

A.Data Collection

A data set is collected and was organized in Microsoft Excel sheet with 649 tuples and 33 attributes. Each student record had the grade, sex, school, Medu, Fedu etc.

B. Tools Used

To apply the classification algorithm, we used WEKA toolkit, widely used software for data mining. This toolkit provides a wide range of different data mining algorithms. It has been widely used in educational data mining researches and for teaching purposes.

C. Data Preparation and Pre-Processing

Initially (in the Preprocess tab) click "open" and navigate to the directory containing the data file (.csv or .arff). In this case we will open the above data file. Once the data is loaded, WEKA will recognize the attributes and during the scan of the data will compute some basic statistics on each attribute. During this phase, we applied some pre-processing for the collected data to prepare it for the mining techniques. We use a "Nominal to Numerical "filter for turning numeric attributes into nominal ones.

D.Classification

In our project, we aim to predict students' final GPA based on their features mentioned in data set. This will give us an insight on how much these features affect the student's grades. We chose to use classification because the objective of classification techniques in data mining is to identify what are the important factors that contribute to categorizing students' final grades. Decision trees are the most popular classification technique in data mining. They represent the group of classification rules in a tree form. The simplicity of its presentation makes them easy to understand. They can work for different types of attributes, nominal or numerical.

- First we loaded the input file and train it.
- Here grade G2 is the class label.
- Applied Random forest and save as a model.
- Load the model and supply test data set and apply.
- Result is obtained as follows:

=== Summary ===

Correctly Classified Instances	149	98.6755 %
Incorrectly Classified Instances	2	1.3245 %
Kappa statistic	0.9854	
Mean absolute error	0.0359	
Root mean squared error	0.0891	
Relative absolute error	29.6446 %	
Root relative squared error	36.2196 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0
0.667	0.000	1.000	0.667	0.800	0.814	0.999	0.917	5
1.000	0.007	0.833	1.000	0.909	0.910	0.999	0.967	6
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	7
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	8
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	9
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	10

Output

MODULE 2:

Existing module 2:

Prediction of Alcohol Consumption

The availability of data has been growing rapidly, and there is a need to analyze huge amounts of data generated. There are many techniques in data mining that can be applied to data, such as classification, clustering, and association rules to name a few. These techniques will help extract hidden knowledge and useful information.

Classification is one of the supervised learning techniques that build a model to classify a data item into a predefined class label. The aim of classification is to predict the future output based on the available data. Hence, we can predict alcohol consumed by the student based on the attributes specified in the dataset.

A.Data Collection

A data set is collected and was organized in Microsoft Excel sheet. with 649 tuples. Each student record had the grade, sex, school, Medu, Fedu etc.

B. Tools Used

To apply the classification algorithm, we used WEKA toolkit, a widely used software for data mining. This toolkit provides a wide range of different data mining algorithms. It has been widely used in educational data mining researches and for teaching purposes.

C. Data Preparation and Pre-Processing

Initially (in the Preprocess tab) click "open" and navigate to the directory containing the data file (.csv or .arff). In this case we will open the above data file. Once the data is loaded, WEKA will recognize the attributes and during the scan of the data will compute some basic statistics on each attribute. During this phase, we applied some pre-processing for the collected data to prepare it for the mining techniques. We use a "Nominal to Numerical " filter for turning numeric attributes into nominal ones.

Classification

In our project, we aim to predict students' alcohol consumption based on their features mentioned in data set. This will give us an insight on how much these features affect the student . We chose to use classification because the objective of classification techniques in data mining is to identify what are the important factors that contribute to categorizing students' alcohol consumption. Decision trees are the most popular classification technique in data mining. They represent the group of classification rules in a tree form. The simplicity of its presentation makes them easy to understand. They can work for different types of attributes, nominal or numerical.

- First we loaded the input file and train it.
- Here dalc is the class label.
- Applied Random forest and save as a model.
- Load the model and supply test data set and apply.
- Result is obtained as follows:

```
Correctly Classified Instances         112               74.1722 %
Incorrectly Classified Instances        39               25.8278 %
Kappa statistic                          0.4372
Mean absolute error                      0.1372
Root mean squared error                  0.2588
Relative absolute error                 65.0642 %
Root relative squared error             80.3635 %
Total Number of Instances              151

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.500    0.780      0.929    0.848      0.494    0.872     0.922     1
                 0.290    0.042    0.643      0.290    0.400      0.346    0.844     0.571     2
                 0.500    0.051    0.500      0.500    0.500      0.449    0.913     0.435     3
                 0.000    0.000    0.000      0.000    0.000      0.000    0.913     0.072     4
                 0.800    0.007    0.800      0.800    0.800      0.793    0.996     0.823     5
Weighted Avg.    0.742    0.341    0.716      0.742    0.711      0.463    0.875     0.790
```

Output

MODULE 3:

Novelty 1:

Machine learning data contains a mixture of attributes, some of which are relevant to making predictions. In our project, there are two existing modules:

- Prediction of Grade
- Prediction of Alcohol Consumption

As our dataset is large, it is difficult to predict. So we use Attribute subset selection. In large data sets, Grade are predicted using all the attributes .Prediction becomes difficult with increasing number of attributes and tuples .How to decide which attributes to use and which attributes to remove? The process of selecting attributes in your data to model your problem is called as **Attribute Subset Selection.** This technique is used in order to decrease the complexity and increase of accuracy.

Top reasons to use Attribute subset selection are:

- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It improves the accuracy of a model if the right subset is chosen.
- It reduces over fitting.

Attribute Subset Selection internally applies CfsSubset Eval algorithm.

CfsSubset Eval:

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

After applying attribute subset selection for
->Prediction of Grade

The output is as follows:

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 207
        Merit of best subset found:     0.552

Attribute Subset Evaluator (supervised, Class (nominal): 32 G2):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 4,20,28,29,31,33 : 6
                     address
                     nursery
                     Walc
                     health
                     G1
                     G3
```

Here G2 is taken as class label

After applying attribute subset selection for

=>Prediction of alcohol consumption

```
              G3
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 193
        Merit of best subset found:     0.244

Attribute Subset Evaluator (supervised, Class (nominal): 27 Dalc):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 28,30 : 2
                     Walc
                     absences
```

Here dalc is taken as class label

In order to predict, we applied an algorithm “RANDOM FOREST” using the obtained attributes by attribute subset selection technique.

Prediction of Grade

Before attribute subset selection

=== Summary ===

Correctly Classified Instances	149	98.6755 %
Incorrectly Classified Instances	2	1.3245 %
Kappa statistic	0.9854	
Mean absolute error	0.0359	
Root mean squared error	0.0891	
Relative absolute error	29.6446 %	
Root relative squared error	36.2196 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0
0.667	0.000	1.000	0.667	0.800	0.814	0.999	0.917	5
1.000	0.007	0.833	1.000	0.909	0.910	0.999	0.967	6
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	7
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	8
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	9
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	10

After attribute subset selection

=== Summary ===

Correctly Classified Instances	151	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0392	
Root mean squared error	0.0853	
Relative absolute error	32.3301 %	
Root relative squared error	34.6795 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	5
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	6
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	7
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	8
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	9
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	10
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	11
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	12
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	13
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	14

Prediction of Alcohol Consumption

Before attribute subset selection

Correctly Classified Instances	112	74.1722 %
Incorrectly Classified Instances	39	25.8278 %
Kappa statistic	0.4372	
Mean absolute error	0.1372	
Root mean squared error	0.2588	
Relative absolute error	65.0642 %	
Root relative squared error	80.3635 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.929	0.500	0.780	0.929	0.848	0.494	0.872	0.922	1
	0.290	0.042	0.643	0.290	0.400	0.346	0.844	0.571	2
	0.500	0.051	0.500	0.500	0.500	0.449	0.913	0.435	3
	0.000	0.000	0.000	0.000	0.000	0.000	0.913	0.072	4
	0.800	0.007	0.800	0.800	0.800	0.793	0.996	0.823	5
Weighted Avg.	0.742	0.341	0.716	0.742	0.711	0.463	0.875	0.790	

After attribute subset selection

=== Summary ===

Correctly Classified Instances	151	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0706	
Root mean squared error	0.115	
Relative absolute error	33.4755 %	
Root relative squared error	35.7253 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	2
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	3
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	4
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	5
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

MODULE 4

Novelty 2:

Association mining with template matching:

In association mining we find the rules based on confidence and support measures. We cannot decide the rule as interesting based on these measures. Using template matching we will find the rules that the user is interested in.
For example:
If A->B is a template then we will fix the LHS of to A and RHS to B.

#Association between internet and paid for extra classes

```
rules <- apriori(x, parameter= list(supp=0.7, conf=0.9), appearance=list(lhs=
c("internet=yes","internet=no"), default="rhs")) // fix LHS to internet
inspect( subset( rules, subset = rhs %pin% "paid=" ) )  //fix RHS to paid
```

Observation:
If there is internet facility then students are not paying for extra classes

#Association between pstaus and failures

```
rules <- apriori(x, parameter= list(supp=0.7, conf=0.8), appearance=list(lhs=
c("Pstatus=T","Pstatus=A"), default="rhs")) // fix LHS to Pstatus
inspect( subset( rules, subset = rhs %pin% "failures=" ) )  // fix RHS to failures
```

Observation:
If parents are living together then there is very less chance of for failures.

#Association between pstaus and higher

```
rules <- apriori(x, parameter= list(supp=0.7, conf=0.8), appearance=list(lhs=
c("Pstatus=T","Pstatus=A"), default="rhs")) // fix LHS to Pstatus
inspect( subset( rules, subset = rhs %pin% "higher=" ) ) // fix RHS to higher
```

Observation:
If parents are living together then there is very high chance for a student to go for higher studies.

#Association between failures and higher

```
rules <- apriori(x, parameter= list(supp=0.7, conf=0.9), appearance=list(lhs=
c("failures=1","failures=2","failures=3","failures=4"), default="rhs")) // fix LHS to failures
inspect( subset( rules, subset = rhs %pin% "higher=" ) ) // fix RHS to higher
```

Observation:

If there are no failures then there is most likely chance for the student to go for higher studies.

#Association between school and higher (wants to take higher education)

```
rules <- apriori(x, parameter= list(supp=0.6, conf=0.8), appearance=list(lhs=
c("school=GP","school=MS"), default="rhs"))              // fix LHS to school
inspect( subset( rules, subset = rhs %pin% "higher=" ) )   // fix RHS to higher
```

Observation:
More number of students from school 'GP' are opting for higher studies when compared to students of school 'MS'.

#Association between school and reason to choose school

```
rules <- apriori(x, parameter= list(supp=0.5, conf=0.8),appearance=list(lhs=
c("school=GP","school=MS"), default="rhs"))              //fix LHS to school
inspect( subset( rules, subset = rhs %pin% "reason=" ) )  //fix RHS to reason
```

Observation:
The reason to join school 'GP' is 'course' for maximum number of students.

#Association between school and freetime

```
rules <- apriori(x, parameter= list(supp=0.5, conf=0.7),appearance=list (lhs=
c("school=GP","school=MS"), default="rhs"))              //fix LHS to school
inspect( subset( rules, subset = rhs %pin% "freetime=" ) )  // fix RHS to freetime
```

Observation:
Free time is more for the students of 'GP' than that of'MS'.

#Association between gender and free time

```
rules <- apriori(x, parameter= list(supp=0.4, conf=0.7),appearance=list (lhs=
c("sex=F","sex=M"), default="rhs"))                      //fix LHS to sex
inspect( subset( rules, subset = rhs %pin% "freetime=" ) )  //fix RHS to freetime
```

Observation:
Females are having more moderate free time than males.

#Association between gender and studytime

```
rules <- apriori(x, parameter= list(supp=0.6, conf=0.6), appearance=list(lhs=
c("sex=F","sex=M"), default="rhs"))                      //fix LHS to sex
inspect( subset( rules, subset = rhs %pin% "studytime=" ) )  //fix RHS to studytime
```

Observation:

Females are having more study time than males.

#Association between gender and paid for extra classes

```
rules <- apriori(x, parameter= list(supp=0.5, conf=0.6),appearance=list(lhs=
c("sex=F","sex=M"), default="rhs"))                    //fix LHS to sex
inspect( subset( rules, subset = rhs %pin% "paid=" ) )  //fix RHS to paid
```

Observation:
Generally, males are paying more for extra classes than females.

#Association between gender and Walc

```
rules <- apriori(x, parameter= list(supp=0.4, conf=0.6),appearance=list(lhs=
c("sex=F","sex=M"), default="rhs"))                     //fix LHS to sex
inspect( subset( rules, subset = rhs %pin% "Walc=" ) )   //fix LHS to Walc
```

Observation:
In the lower level of consumption females are more in number than males,in case of weekly consumption.

#Association between gender and Dalc

```
rules <- apriori(x, parameter= list(supp=0.5, conf=0.7), appearance=list(lhs=
c("sex=F","sex=M"), default="rhs"))                     //fix LHS to sex
inspect( subset( rules, subset = rhs %pin% "Dalc=" ) )  //fix LHS to Dalc
```

Observation:
In the lower level of consumption females are more in number than males, in case of daily consumption.(Same as weekly consumption).

#Association between famsize and pstatus

```
rules <- apriori(x, parameter= list(supp=0.5, conf=0.7),appearance=list(lhs=
c("famsize=LE3","famsize=GT3"), default="rhs"))          //fix LHS to famsize
inspect( subset( rules, subset = rhs %pin% "Pstatus=" ) )  //fix LHS to Pstatus
```

Observation:
If family size is more then, then most probably parents are living together.

#Association between Pstatus and alcohol consumption

```
rules <- apriori(x, parameter= list(supp=0.1, conf=0.1),appearance=list(lhs=
c("Pstatus=T","Pstatus=A"), default="rhs"))  ////fix LHS to Pstatus
inspect( subset( rules, subset = rhs %pin% "Walc=" ) )  // fix RHS to Walc
```

Observation:
If parents are living together there is less chance of alcohol consumption among students
Above all are the examples of how an attribute is related to other. Now we will see the association between multiple attributes.

rules<-apriori(x,parameter= list(supp=0.7, conf=0.8),appearance=NULL,control=NULL)
inspect(rules)

```
[Pstatus=T,failures=0}     => {paid=no}       0.7026194 0.9480249  1.0086364 456
[failures=0,paid=no}       => {Pstatus=T}     0.7026194 0.8769231  1.0002163 456
[Pstatus=T,paid=no}        => {failures=0}    0.7026194 0.8539326  1.0094759 456
[failures=0,higher=yes}    => {schoolsup=no}  0.7010786 0.8869396  0.9907466 455
[failures=0,schoolsup=no}  => {higher=yes}    0.7010786 0.9266802  1.0369232 455
[schoolsup=no,higher=yes}  => {failures=0}    0.7010786 0.8852140  1.0464552 455
[failures=0,higher=yes}    => {paid=no}       0.7488444 0.9473684  1.0079379 486
[failures=0,paid=no}       => {higher=yes}    0.7488444 0.9346154  1.0458024 486
[paid=no,higher=yes}       => {failures=0}    0.7488444 0.8933824  1.0561114 486
[failures=0,schoolsup=no}  => {paid=no}       0.7211094 0.9531568  1.0140964 468
[failures=0,paid=no}       => {schoolsup=no}  0.7211094 0.9000000  1.0053356 468
[schoolsup=no,paid=no}     => {failures=0}    0.7211094 0.8540146  1.0095728 468
[Pstatus=T,higher=yes}     => {paid=no}       0.7349769 0.9352941  0.9950916 477
[Pstatus=T,paid=no}        => {higher=yes}    0.7349769 0.8932584  0.9995254 477
[paid=no,higher=yes}       => {Pstatus=T}     0.7349769 0.8768382  1.0001195 477
[Pstatus=T,schoolsup=no}   => {paid=no}       0.7380586 0.9392157  0.9992639 479
[Pstatus=T,paid=no}        => {schoolsup=no}  0.7380586 0.8970037  1.0019887 479
[schoolsup=no,paid=no}     => {Pstatus=T}     0.7380586 0.8740876  0.9969822 479
[schoolsup=no,higher=yes}  => {paid=no}       0.7457627 0.9416342  1.0018371 484
[paid=no,higher=yes}       => {schoolsup=no}  0.7457627 0.8897059  0.9938367 484
[schoolsup=no,paid=no}     => {higher=yes}    0.7457627 0.8832117  0.9882834 484
```

rules<-apriori(x,parameter= list(supp=0.6, conf=0.9),appearance=NULL,control=NULL)
inspect(rules)

```
{failures=0,nursery=yes,higher=yes}    => {paid=no}       0.6040062 0.9445783  1.0049694 392
{failures=0,paid=no,nursery=yes}       => {higher=yes}    0.6040062 0.9333333  1.0443678 392
{schoolsup=no,nursery=yes,higher=yes}  => {paid=no}       0.6009245 0.9397590  0.9998420 390
{Pstatus=T,failures=0,schoolsup=no}    => {higher=yes}    0.6194145 0.9305556  1.0412596 402
{Pstatus=T,failures=0,higher=yes}      => {paid=no}       0.6579353 0.9467849  1.0073171 427
{Pstatus=T,failures=0,paid=no}         => {higher=yes}    0.6579353 0.9364035  1.0478032 427
{Pstatus=T,failures=0,schoolsup=no}    => {paid=no}       0.6332820 0.9513889  1.0122154 411
{Pstatus=T,failures=0,paid=no}         => {schoolsup=no}  0.6332820 0.9013158  1.0068054 411
{failures=0,schoolsup=no,higher=yes}   => {paid=no}       0.6687211 0.9538462  1.0148298 434
{failures=0,schoolsup=no,paid=no}      => {higher=yes}    0.6687211 0.9273504  1.0376732 434
{Pstatus=T,schoolsup=no,higher=yes}    => {paid=no}       0.6533128 0.9359823  0.9958238 424
```

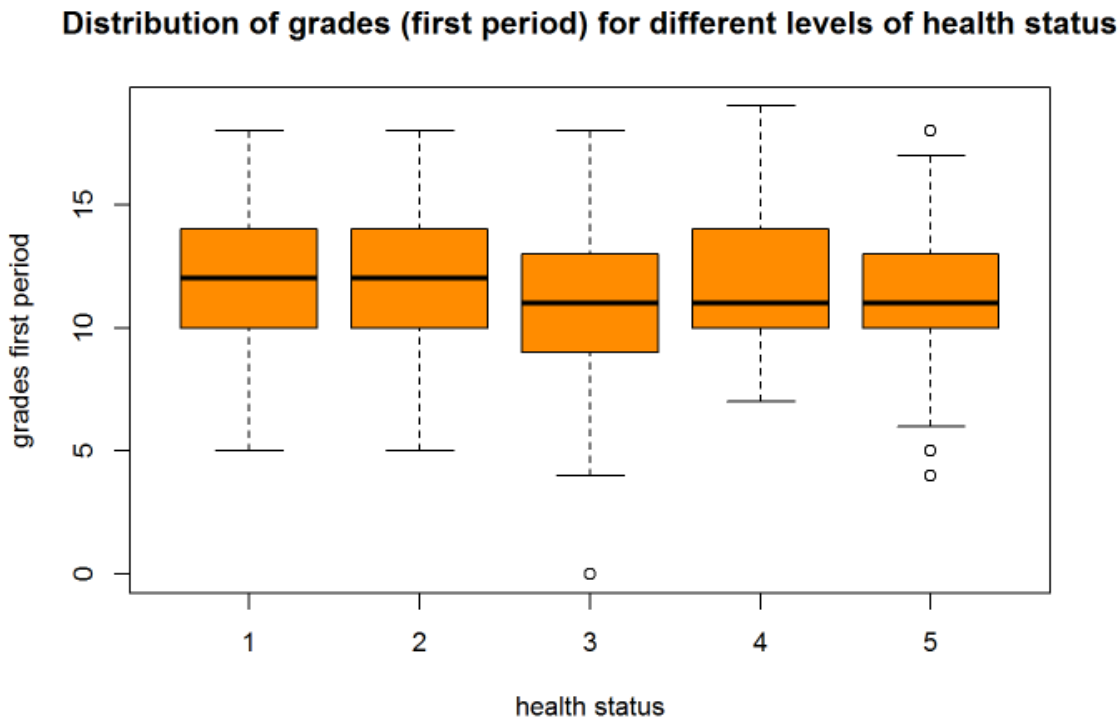
Here the column followed by rules indicates support, second column indicates confidence, third indicates lift and fourth indicates count of tuples which satisfy the rule.

```
x<-read.csv(file.choose())      //Reading a file
head(x)
```

#How health and grades of period 1 are related to each other

```
aggregate( G1 ~ health, data = x, FUN = mean)

boxplot (x $ G1 ~ x $ health, main = "Distribution of grades (first period)
        for different levels of health status", xlab = "health status",
        ylab = "grades first period", col = "darkorange")
```



The boxplot shows that the students' grades for low health status (1 and 2) are slightly better than for high health status (3-5). So, to check if this association is significant, carry out a correlation test between students' grades in first period and their health status.

#Corelation between health and grade

cor.test (x = x $ health, y = x $ G1)

```
##
##  Pearson's product-moment correlation
##
## data:  student.por$health and student.por$G1
## t = -1.3155, df = 647, p-value = 0.1888
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12809950  0.02541487
## sample estimates:
##         cor
## -0.05164742
```

Here, correlation=-0.0516472
So, there is no significant correlation between health and students grade of period1

#---

#Differnce in grades of female and male

t.test (x $ G1 ~ x$ sex, alternative = "two.sided")

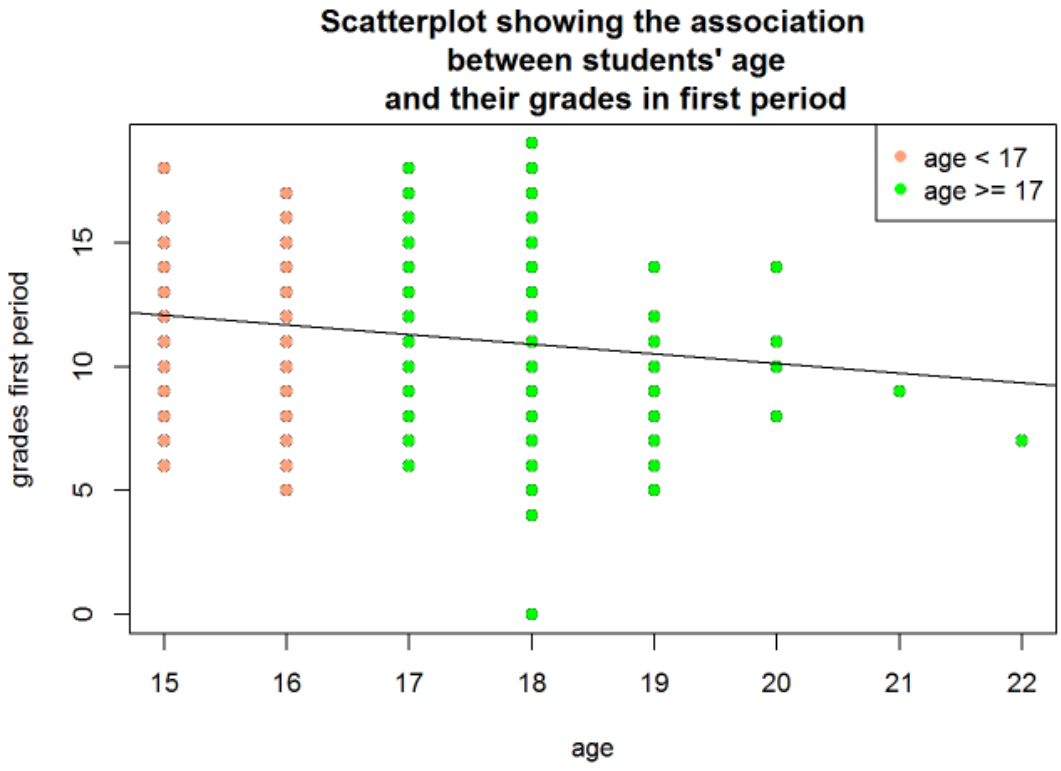
```
##
##  Welch Two Sample t-test
##
## data:  student.por$G3 by student.por$sex
## t = 3.2747, df = 547.44, p-value = 0.001125
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3390334 1.3554639
## sample estimates:
## mean in group F mean in group M
##        12.25326        11.40602
```

There is a significant difference between final grades of female and male students, t (547.44) = 3.2747, p = 0.001125. Female students had better grades (mean = 12.25326) than male students (mean = 11.40602).

#---

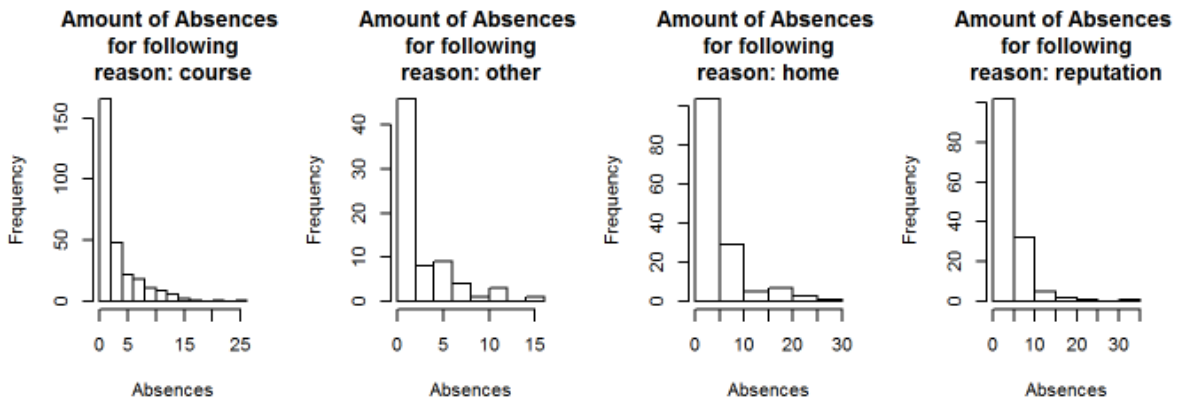
#Scatter plot between age and grades of period 1

```
plot (x = x $ age, y = x $ G1,
      xlab = "age",                    //Plot by taking age on x-axis and grade on y-axis
      ylab = "grades first period",
      main = "Scatterplot showing the association
     between students' age
     and their grades in first period"
)
with (subset(x, age < "17"),
      points(age, G1,
            pch = 16, col = "lightsalmon1") )  //Fixing lightsalmon1 color for points with age<17
with (subset(x, age >= "17"),
      points (age, G1,
             pch = 16, col = "green"))   //Fixing green color for points with age>=17
legend("topright",
       c("age < 17", "age >= 17"),
       pch = 16,
       col = c("lightsalmon1", "green")  //Defining the top right heading
)
abline(a = 17.97725,
       b = -0.39286)
```



```
#-------------------------------------------------------------------------------
#Association between absences and why students choose this school

par(mfrow = c(2, 4))       //Fixing how output looks like
x.hist <- unique (x $ reason)
for (reason.i in x.hist) {      //Histogram for each and every  reason to join school
  data.temp <- subset(x,
                      reason == reason.i)
  hist(data.temp$absences,
       main = paste ("Amount of Absences
                     for following
                     reason:",
                     reason.i),
       xlab = "Absences")
}
```



- The histograms show an association between the amount of absences and the reason why students chose this school. With the help of these histograms I will now find out, if there are significant differences between the different groups of reasons.
- We have conducted an analysis of variance to see if there are significant differences of absences for the different groups of reasons why students attend this school.

```
fit <- aov (formula = x $ absences ~ x $ reason)//analysis of variance between absences
                                                  and reason
summary (fit)
```

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## student.por$reason   3    175   58.22   2.725 0.0434 *
## Residuals          645  13781   21.37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here $F(645) = 2.725$, $p = 0.0434$.

So, there is a significant difference between students' absences for the different reasons why they choose their school.

In histograms we cannot see the significant difference. So there must be some outlayers. Now after removing outlayers conduct analysis of variance.

```
table (x$absences)

mean (x$absences)  //Mean of absences

sd (x$absences)      //Standard deviation of absences


without_outliers <- function(x1) {              //Function to remove outlayers

outlier <- mean(x1)+ 2*sd(x1)

x1[x1>=outlier] <- NA

return(x1)

}
```

 # calculate the mean of absences without the outliers.

```
mean (without_outliers (x$absences), na.rm = T)
```

#AOV for absences without outlayers

```
x$absences_without <- without_outliers (x$absences)

fit <- aov (formula = x$absences_without ~ x$ reason)

summary (fit)
```

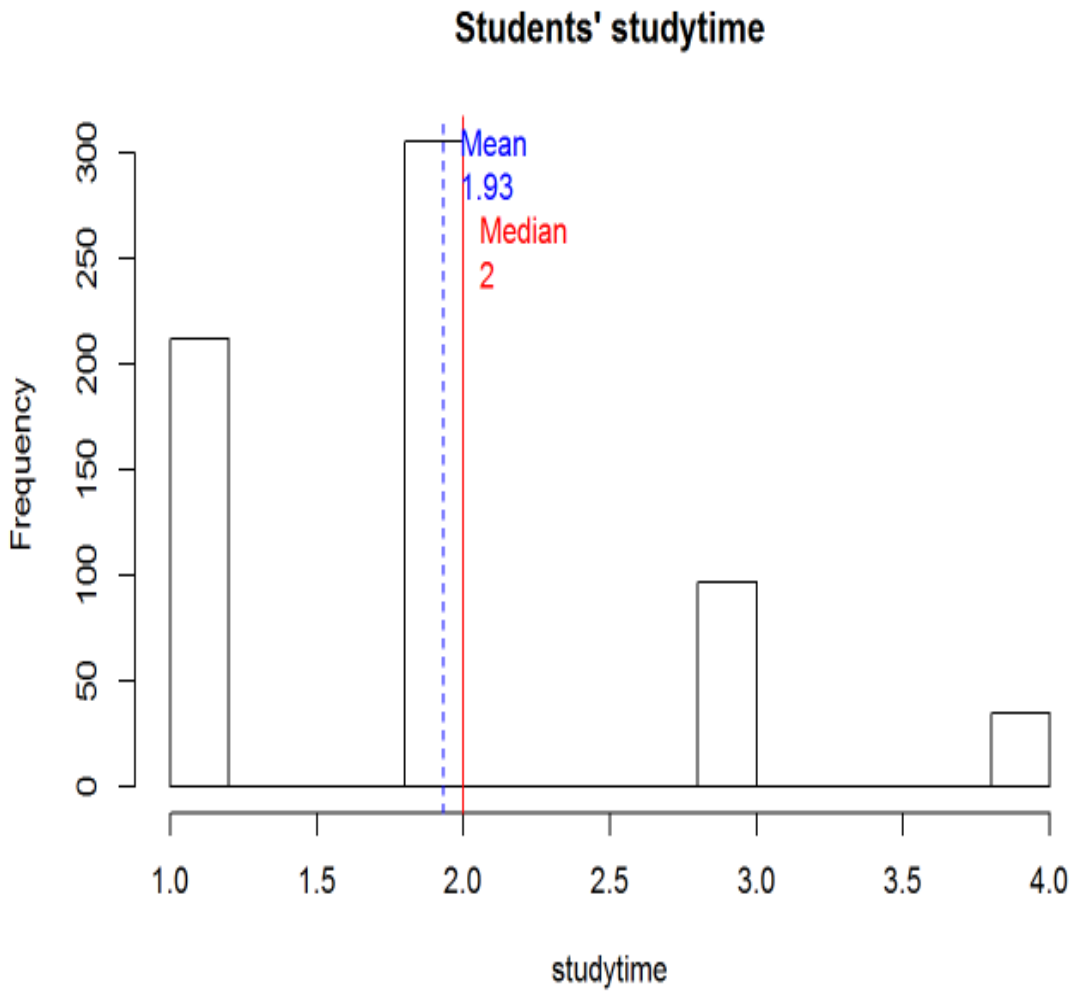
```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## student.por$reason    3     17   5.795   0.539  0.656
## Residuals           613   6587  10.745
## 32 observations deleted due to missingness
```

Without the outliers, there is no significant difference between students' absences for the different reasons why they chose their school, F(613) = 0.539, p = 0.656.

```
#-------------------------------------------------------------------------------
#Association between study time and grade

hist (x $ studytime,                    //Histogram between study time and frequency
    main = "Students' studytime",
    xlab = "studytime")
abline(v = median(x $ studytime),
      lty = 1, col = "red")
abline(v = mean(x $ studytime),
      lty = 2, col = "blue")
text(mean(x $ studytime, na.rm = T), 300,      //Mean of study time
    labels = paste("
Mean\n", round(mean(x $ studytime, na.rm = T), 2), sep = ""  ),
    adj = 0,
    pos = 4,
    col = "blue"
)
text(median(x $ studytime, na.rm = T), 250,    //Median of study time
    labels = paste("Median\n", round(median(x $ studytime, na.rm = T), 2), sep = ""  ),
    adj = 0,
    pos = 4,
    col = "red"
)


aov.study <- aov (formula = x $ G3 ~ x $ studytime)
summary (aov.study)
```



Carry out a one-way-anova to test if there is a significant association between students' study time and their final grades.

```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## student.por$studytime   1    422   422.0   43.06 1.09e-10 ***
## Residuals             647   6341     9.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant association between final grades and study time, $F(647) = 43.06$, $p = 1.09e\text{-}10$.

MODULE 5

Novelty 3:

Clustering the data into groups:

We want to divide all the tuples into 10 groups.And give that information to parents.When we give that information to the parents they can compare their own children behaviour with the other the children.They can also observe how student behaviour can predict the grade.They can also observe that the no of students are there in each category.By that they can understand that what measures are to be taken by the parents to make their children in that particular group.

For example lets us consider an example ,If the no of students whose alchool drinking is more and their grades are less have lare count it means that it is having more chance that their children can fall into that group.Then parents have to see the characteristics of that group and they have to make their children not to all into that group.

Final cluster centroids:

		Cluster#								
Attribute	Full Data	0	1	2	3	4	5	6	7	8	9
	(649.0)	(89.0)	(90.0)	(107.0)	(45.0)	(41.0)	(59.0)	(70.0)	(36.0)	(56.0)	(56.0)
school	GP	GP	GP	GP	MS	MS	GP	MS	GP	MS	GP
sex	F	M	F	F	F	F	F	F	F	F	M
age	17	16	16	17	18	17	16	17	15	18	16
address	U	U	U	U	U	U	U	R	U	U	U
famsize	GT3	GT3	GT3	GT3	GT3	GT3	GT3	GT3	GT3	GT3	GT3
Pstatus	T	T	T	T	T	T	T	T	T	T	T
Medu	2	4	4	3	1	4	1	1	2	2	2
Fedu	2	4	4	2	1	2	1	2	2	2	2
Mjob	other	teacher	other	other	at_home	teacher	at_home	other	services	other	other
Fjob	other	services	other	other	other	services	other	other	services	other	other
reason	course	course	reputation	home	course	course	course	course	reputation	course	course
guardian	mother	mother	mother	mother	mother	mother	mother	mother	father	mother	mother
traveltime	1	1	1	1	2	1	1	2	1	2	1
studytime	2	1	2	2	1	2	2	1	2	2	1
failures	0	0	0	0	0	0	0	0	0	0	0
schoolsup	no	no	no	no	no	no	no	no	no	no	no
famsup	yes	yes	yes	yes	no	yes	yes	no	yes	yes	no
paid	no	no	no	no	no	no	no	no	no	no	no
activities	no	yes	yes	no	yes	yes	no	no	yes	no	no
nursery	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
higher	yes	yes	yes	yes	no	yes	yes	yes	yes	yes	yes
internet	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	yes
romantic	no	no	no	no	yes	no	no	no	no	yes	no
famrel	4	4	4	4	5	5	4	4	5	4	4
freetime	3	3	4	3	5	3	3	4	4	3	3
goout	3	3	4	2	5	5	4	3	2	3	3
Dalc	1	1	1	1	1	1	1	1	1	1	1
Walc	1	4	1	1	1	1	1	2	1	1	1
health	5	5	4	5	5	5	5	5	5	5	5
absences	0	0	0	0	0	0	0	0	0	0	0
G1	10	10	14	12	8	14	11	10	10	9	11
G2	11	9	15	12	10	13	11	11	11	10	11
G3	11	11	16	13	9	14	11	11	12	10	11

Output

MODULE 6

Novelty 4:

Always the grade of a student depends on his own characteristics as well as his surroundings. The data set we initially considered have only student characteristics.So we created a new data set which includes the details of the student's friends grades and his surrounding people.We just prepared a dataset with 20 tuples and tried to predict grades.In R using svm we predicted the grade of the student.

details of dataset:

f1:grade of 1st friend
f2:grade of 2nd friend
f3:grade of 3rd friend

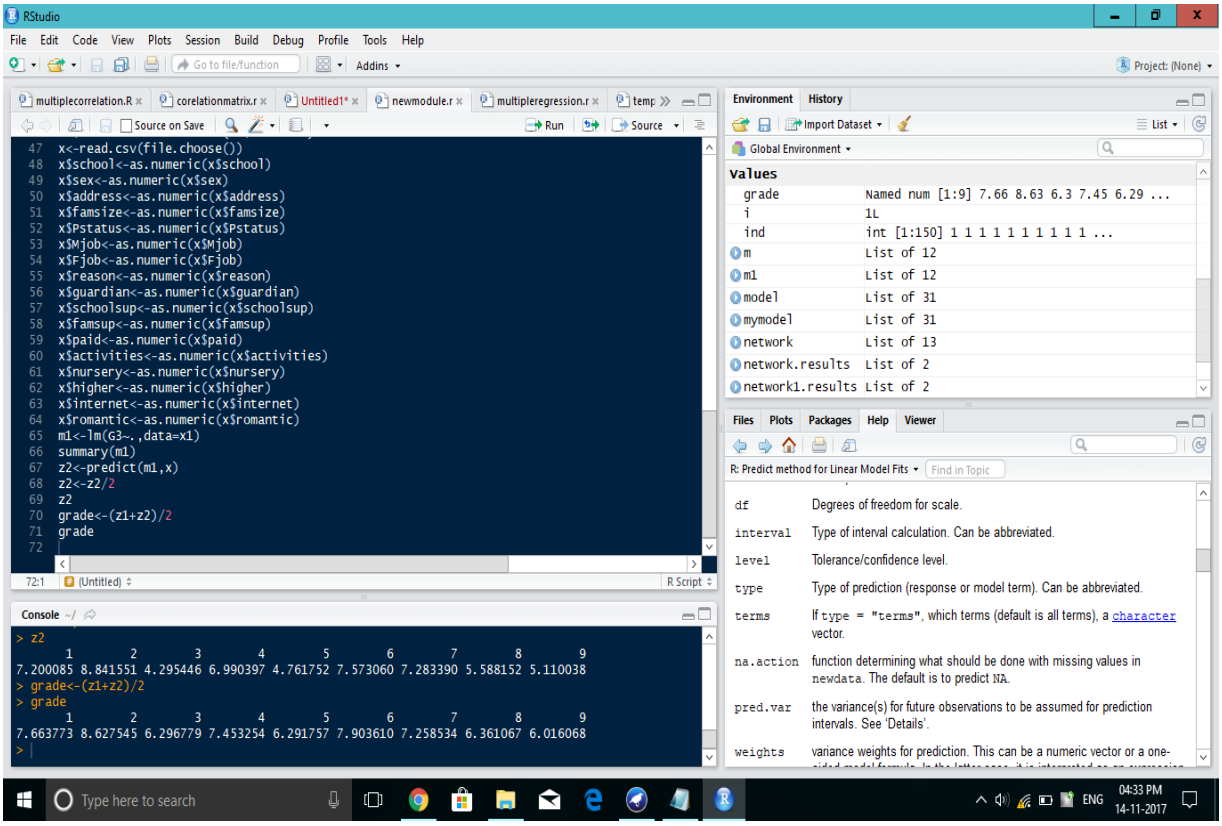
educates:no of educates in his surroundings
mncs:percentage of people doing jobs
bussiness:percentage of people doing bussiness
farmers:percentage of people doing farming aroun him
unemployment:percentage of people facing unemployment
grade:grade of the student(class label)

CODE:

```
x<-read.csv(file.choose()) //choosing the file containing the new data set
y<-read.csv(file.choose()) //choosing the file containing the new tuples to predict the grade.
m<-lm(grade~.,data=x)//lm is the function which is used to create a model using the data set and that model is stored in m with grade as class label
summary(m)//gives the summary of the model m
z1<-predict(m,y)//predict the grades for y dataset using m model
z1//prints the new grades
```

In the same way we did for the initial dataset and find the grade for test data set.which is stored in z2 variable.

```
x1<-read.csv(file.choose()) //choosing the file containing the new data set
y1<-read.csv(file.choose()) //choosing the file containing the new tuples to predict the grade.
m1<-lm(G3~.,data=x1)//lm is the function which is used to create a model using the data set and that model is stored in m1 with grade as class label
summary(m1)//gives the summary of the model m1
z2<-predict(m1,y1)//predict the grades for y dataset using m1 model
z2<-z2/2//scaled to 10 as intial values are for 20
z2//print the grades
grade<-(z1+z2)/2//final grade is average of two grades
grade//prints final grades
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
multiplecorrelation.R
corelationmatrix.r
Untitled1*
newmodule.r
multipleregression.r
temp
Source on Save
Run
Source
47 x<-read.csv(file.choose())
48 x$school<-as.numeric(x$school)
49 x$sex<-as.numeric(x$sex)
50 x$address<-as.numeric(x$address)
51 x$famsize<-as.numeric(x$famsize)
52 x$Pstatus<-as.numeric(x$Pstatus)
53 x$Mjob<-as.numeric(x$Mjob)
54 x$Fjob<-as.numeric(x$Fjob)
55 x$reason<-as.numeric(x$reason)
56 x$guardian<-as.numeric(x$guardian)
57 x$schoolsup<-as.numeric(x$schoolsup)
58 x$famsup<-as.numeric(x$famsup)
59 x$paid<-as.numeric(x$paid)
60 x$activities<-as.numeric(x$activities)
61 x$nursery<-as.numeric(x$nursery)
62 x$higher<-as.numeric(x$higher)
63 x$internet<-as.numeric(x$internet)
64 x$romantic<-as.numeric(x$romantic)
65 m1<-lm(G3~.,data=x1)
66 summary(m1)
67 z2<-predict(m1,x)
68 z2<-z2/2
69 z2
70 grade<-(z1+z2)/2
71 grade
72
72:1 (Untitled) R Script
Console ~/
> z2
1 2 3 4 5 6 7 8 9
7.200085 8.841551 4.295446 6.990397 4.761752 7.573060 7.283390 5.588152 5.110038
> grade<-(z1+z2)/2
> grade
1 2 3 4 5 6 7 8 9
7.663773 8.627545 6.296779 7.453254 6.291757 7.903610 7.258534 6.361067 6.016068
>
Environment History
Import Dataset
List
Global Environment
values
grade Named num [1:9] 7.66 8.63 6.3 7.45 6.29 ...
i 1L
ind int [1:150] 1 1 1 1 1 1 1 1 1 1 ...
m List of 12
m1 List of 12
model List of 31
mymodel List of 31
network List of 13
network.results List of 2
network1.results List of 2
Files Plots Packages Help Viewer
R: Predict method for Linear Model Fits
Find in Topic
df Degrees of freedom for scale.
interval Type of interval calculation. Can be abbreviated.
level Tolerance/confidence level.
type Type of prediction (response or model term). Can be abbreviated.
terms If type = "terms", which terms (default is all terms), a character vector.
na.action function determining what should be done with missing values in newdata. The default is to predict NA.
pred.var the variance(s) for future observations to be assumed for prediction intervals. See 'Details'.
weights variance weights for prediction. This can be a numeric vector or a one-
Type here to search
ENG
04:33 PM
14-11-2017

Output

5. Experiments and Evaluation:

- While predicting the class labels we used different algorithms like j48,random forest, SVM, naive bayes ,id3 .We opt the one which gives highest accuracy.
- For attribute subset selection we first constructed corelation matrix, But we cannot able to fix the threshold, So we cannot predict exactly .So we directly used an option in Weka.
- For clustering we also tried in R, using knn algorithm .But the clusters we get cannot give any best meaning.
- And again for clustering we cannot give nominal vales so we have to covert all the the values into numeric , As a result the output of clustering does not give any sense.

6. References:

- To learn R programming
  https://www.udemy.com/r-programming/
- IEEE for base papers
  http://ieeexplore.ieee.org
- Dataset
  https://data.world/

7. Timeline:

- We have gone through many websites to find a problem statement. After going through many websites, knowing many problem statements we got many ideas. We finally fixed it as prediction of student behaviour on 28-8-2017.

- We started searching many websites for dataset. We finally found best dataset in Data World website after two days.

- Next we searched for base papers for 4 days and prepared our own base paper for our project by 11-9-2017.

- In the next week we generated association rules in weka but we didnot gain much knowledge from that because we got hundreds of rules at support=0.9 and confidence=1.

- Next we took one week to learn R to improve our association rules.

- Parallelly we tried to predict grades of the students using different classification techniques in weka. After trying 5 to 6 algorithms we finally got 98% accuracy in predicting the grades, which took us two weeks.

- In the next week In R using template matching we found interesting rules and correlation between attributes.

- In the next two days we applied attribute subset selection and again predicted the grades. we finally got 100% accuracy .

- In the next week we apply simple k means algorithm and clustered the dataset into 10 groups.

- To extend our project we searched some more base papers and we found that from our dataset we can also predict whether student is addicted to alcohol or not. This took us 1 week.

- We thought that in prediction of grade, the student grade is not only dependent on his own habits but also on their surroundings .so we created our own dataset and try to predict the student grade again.

- Next we tried to predict the final grade of the new students using both the data models.

- We completed all the modules by 7-11-2017.
- We took 5 days to complete the project report and finally completed by 12-11-2017.

8. Conclusion:

For any country ,the future is dependent on the students. If we are able to predict the student behaviour and take necessary precautions for their growth then it will help the student as well as it may change the future of the society.Today in the present society it is very neessary to observe the student behaviour. I think this system will address this problem to some extent .This system will be very helpful for the educational administration as we predicted the grades they can predict the student grades priorly and can concentrate on them who have less grades.

9. Future works:

In the prediction of grade we have considered the details of the children and in novelty we have added an extra feature which also uses the surroundings of the children like friends grade and educates around student.

To globalize this system, we have to consider the details of the school in which he is studying .The details like faculty dealing the course, courses selected, environment in school etc.

We can also add another attribute like 'remarks' given by school and family which can be helpful in predicting student character.