Northumbria University
NEWCASTLE

Principles of Data Science Assessment

DATASET NAME: STUDENTS GRADE
TOTAL WORDS: 4194

NAME: NIDA ALYAS
STUDENT ID: 21040872

Contents

Introduction

Education is vital in the society, many education authorities in the world working hard to improve this area. The educational frameworks need, at this particular time, unconventional ways to improve the quality to accomplish the best outcomes and decrease the failures. Predict the student's performance in educational assessments is a critical problem to solve, with many factors which contributes to the outcomes achieved by a student. However, more accurate and effective prediction of students pass rates and identify the factors most firmly associated with high grades being achieved may allow us a better understanding to be acquired and best practice to be developed. Covid pandemic that has disturbed life all over the world in 2020, the educational frameworks have been impacted in numerous ways; different studies show that student's performance has decreased from that point forward, which highlights the need to address this issue more significantly. This study will focus to build a model to classify and capable of predict either the student will pass or fail on the basis of numerous factors. Also, identify the most important features that are associated with a student's pass or fail the assessment.

Data Acquisition

The student grade dataset consists of 395 observations and 32 attributes with different datatypes including categorical, numerical, nominal and Boolean. This dataset presents data on various factors that can affect the student performance. The aim of the study is build a model to classify and capable of predict either the student will pass or fail on the basis of numerous factors. Also, identify the most important features that are associated with a student's pass or fail the assessment.

Each attribute description and datatypes is given below:

- **School:** School attended and datatype is Categorical
- **Sex:** Sex Student's and datatype is Categorical
- **Age:** Student's age and datatype is Numeric
- **Address:** Student's home address type (u = urban, r = rural) and datatype is Categorical
- **Famsize:** Student's family size (LE3 = less or equal to 3, GT3 =greater than 3) and datatype is Categorical.
- **Pstatus:** Parent's cohabitation status (t = together, a = apart) and datatype is Categorical.
- **Medu**: Mother's education Mother's education level (1 = no qualifications, 2 = school-level qualifications, 3 = further education, 4 = higher education) and datatype is Categorical
- **Fedu:** Father's education father's education level (1 = no qualifications, 2 = school-level qualifications, 3 = further education, 4 = higher education) and datatype is Categorical.
- **Mjob:** Mother's job and datatype is Nominal
- **Fjob:** Father's job and datatype is Nominal
- **Reason:** Reason for choosing the school and datatype is Nominal
- **Guardian:** Guardian Student's and datatype is Nominal
- **Traveltime:** Time taken to travel to school (1 = <15 mins, 2 = 15-30 mins, 3 = 30mins-1 hour, 4 = > 1 hour) and datatype is Categorical
- **Studytime:** Study time Weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 – 5 to 10 hours, or 4 - >10 hours) and datatype is Categorical
- **Failures:** Number of previous assessment failures and datatype is Numeric
- **Schoolsup:** Extra educational support and datatype is Boolean

- **Famsup:** Family educational support and datatype is Boolean
- **Paid:** Student has extra paid for classes and datatype is Boolean
- **Activities:** Student engages in extra-curricular activities and datatype is Boolean
- **Nursery:** Attended nursery school and datatype is Boolean
- **Higher:** Higher Wants to attend higher education and datatype is Boolean
- **Internet:** Internet Has internet access at home and datatype is Boolean
- **Romantic:** Romantic Is involved in a romantic relationship and datatype is Boolean
- **Famrel:** Quality of family relationships (1 = very bad, 5 = very high) and datatype is Categorical
- **Freetime:** Free time after school (1 = very low, 5 = very high) and datatype is Numeric
- **Gout:** Going out with friends (1 = very low, 5 = very high) and datatype is Numeric
- **Dalc:** How much alcohol is consumed on an average weekday (1 = very low, 5 = very high) and datatype is Numeric
- **Walc:** How much alcohol is consumed on an average weekend day (1 = very low, 5 = very high) and datatype is Numeric
- **Health:** overall health status (1 = very bad, 5 = very good) and datatype is Numeric
- **Absences:** Number of school absences and datatype is Numeric
- **Pass:** Whether the student passed the assessment (1 = yes, 0 = no) and datatype is Boolean

```
> str(grade)
'data.frame':   395 obs. of  31 variables:
 $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
 $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
 $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
 $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
 $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
 $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
 $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
 $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
 $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
 $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
 $ Pass      : int  0 0 0 1 0 1 1 0 1 1 ...
```

Fig. 1. student grade dataset

Data Pre-processing

In the step, we have to load the mushroom dataset into the R environment called Rstudio One of the essential phase of any machine learning task is ensuring that the data is in the optimal state for processing, prior to building any models. Few standard tasks will be implemented for data preparation as follows:

- Data Cleaning: Identify errors, corrections and no trivial/redundant data in our dataset
- Data Transformation: Change the scaling of attributes.
- Data Normalization: Scaling/normalizing the values
- Feature Selection: Select important and most relevant features

Data Cleaning

The initial task in data preprocessing is data cleaning. Discover how to transform messy data into clean by identifying and handling outliers and missing values using statistical and modeling techniques. We can ensure that dataset has no missing data, redundant records, and outliers through coding in Rstudio.

Remove Missing values

To check the missingness of data from all attributes of the student grade dataset, we used colSums(is.na()) function. This function shows the missing values in each feature. Fig.2 demonstrates that there is no missing value in the given dataset.

```
> #Checking missing values
> colSums(is.na(grade))
    school        sex        age    address    famsize    Pstatus       Medu       Fedu       Mjob
         0          0          0          0          0          0          0          0          0
      Fjob     reason   guardian traveltime  studytime   failures  schoolsup     famsup       paid
         0          0          0          0          0          0          0          0          0
activities    nursery     higher   internet   romantic     famrel   freetime      goout       Dalc
         0          0          0          0          0          0          0          0          0
      Walc     health   absences       Pass
         0          0          0          0
> |
```

Fig. 2. Missing values in student grade dataset

Remove Duplicate values

Duplicate values analysis test (Fig. 3.) showing that no duplicate values are in our dataset. All the attributes contains only distinct values.

```
> #Checking for duplicated data.
> sum(duplicated(grade))
[1] 0
> |
```

Fig. 3. Duplicate values in student grade dataset

Outliers

Outliers identification is also important tasks, usually extreme values can be outliers in dataset. When only categorical data i.e. gender: male or female, there's no chance of an outlier detection but we have combination of different datatypes. We will check the insights of data with some descriptive statistics, and specifically with the minimum and maximum which is basic way to find outliers.

```
> summary(grade)
 school   sex          age        address famsize   Pstatus      Medu            Fedu            Mjob           Fjob           reason      guardian
 GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000   Min.   :0.000   at_home : 59   at_home : 20   course    :145   father: 90
 MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000   1st Qu.:2.000   health  : 34   health  : 18   home      :109   mother:273
                  Median :17.0                             Median :3.000   Median :2.000   other   :141   other   :217   other     : 36   other : 32
                  Mean   :16.7                             Mean   :2.749   Mean   :2.522   services:103   services:111   reputation:105
                  3rd Qu.:18.0                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58   teacher : 29
                  Max.   :22.0                             Max.   :4.000   Max.   :4.000
   traveltime      studytime        failures      schoolsup famsup     paid     activities nursery   higher    internet  romantic      famrel
 Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no :214   no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242   yes:181   yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000
 Median :1.000   Median :2.000   Median :0.0000                                                                                     Median :4.000
 Mean   :1.448   Mean   :2.035   Mean   :0.3342                                                                                     Mean   :3.944
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                                                                                     3rd Qu.:5.000
 Max.   :4.000   Max.   :4.000   Max.   :3.0000                                                                                     Max.   :5.000
    freetime        goout           Dalc            Walc           health         absences          Pass            train
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.0
 Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median :0.05333   Median :1.0000   Median :1.0
 Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554   Mean   :0.07612   Mean   :0.5291   Mean   :0.8
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:0.10667   3rd Qu.:1.0000   3rd Qu.:1.0
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0
>
```

Fig. 4. Descriptive statistics

From the summary shown in Fig.4, there seems to be no observations extreme or higher than all other observations.

Data Transformation

In data transformation, we will transform each string observation to numerical for training. By visualizing the dataset, we can extract the categorical variables that we need to map into numeric values. The initial view of dataset is given in Fig. 5.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nurs
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes
GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3	2	0	yes	yes	no	yes	yes
GP	F	15	R	GT3	T	2	4	services	health	course	mother	1	3	0	yes	yes	yes	yes	yes
GP	M	16	U	LE3	A	3	4	services	other	home	mother	1	2	0	yes	yes	no	yes	yes
GP	F	15	R	GT3	T	3	4	services	health	course	mother	1	3	0	yes	yes	yes	yes	yes
GP	F	15	R	GT3	T	2	2	at_home	other	reputation	mother	1	1	0	yes	yes	yes	yes	yes
GP	M	15	U	GT3	T	2	2	services	services	course	father	1	1	0	yes	yes	no	no	yes
GP	F	16	U	LE3	T	2	2	other	at_home	course	father	2	2	1	yes	no	no	yes	yes
GP	F	15	U	LE3	A	4	3	other	other	course	mother	1	2	0	yes	yes	yes	yes	yes
GP	F	15	U	GT3	T	4	4	services	teacher	other	father	1	2	1	yes	yes	no	yes	no
GP	F	15	U	GT3	T	4	4	services	services	course	mother	1	1	0	yes	yes	yes	no	yes
GP	M	15	U	LE3	T	1	2	other	at_home	home	father	1	2	0	yes	yes	no	yes	yes
GP	F	16	U	GT3	T	1	1	services	services	course	father	4	1	0	yes	yes	no	yes	no
GP	F	16	U	LE3	T	1	2	other	services	reputation	father	1	2	0	yes	no	no	yes	yes
GP	F	16	U	GT3	T	4	3	teacher	health	home	mother	1	3	0	yes	yes	yes	yes	yes
GP	F	15	U	LE3	T	4	3	services	services	reputation	father	1	2	0	yes	no	no	yes	yes
GP	F	16	U	GT3	T	3	1	services	other	course	mother	1	4	0	yes	yes	yes	no	yes
GP	F	15	R	LE3	T	2	2	health	services	reputation	mother	2	2	0	yes	yes	yes	no	yes
GP	F	15	R	GT3	T	1	1	other	other	reputation	mother	1	2	2	yes	yes	no	no	no
GP	F	16	U	GT3	T	3	3	other	services	home	mother	1	2	0	yes	yes	yes	yes	yes
GP	M	17	U	GT3	T	2	1	other	other	home	mother	2	1	3	yes	yes	no	yes	yes
GP	M	15	U	GT3	T	2	3	other	services	course	father	1	1	0	yes	yes	yes	yes	no
GP	M	15	U	GT3	T	2	3	other	other	home	mother	1	3	0	yes	no	yes	no	no
GP	F	16	U	LE3	T	3	1	other	other	home	father	1	2	0	yes	yes	no	no	yes
GP	F	15	R	GT3	T	1	1	at_home	other	home	mother	2	4	1	yes	yes	yes	yes	yes
GP	M	16	R	GT3	T	4	3	services	other	reputation	mother	2	1	0	yes	yes	no	yes	no

Fig. 5. Actual dataset with different datatypes

After transforming categorical values in the numerical data, the transformed data is given in Fig. 6.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	pai
1	1	18	2	1	1	4	4	1	5	1	2	2	2	0	1	0	
1	1	17	2	1	2	1	1	1	3	1	1	1	2	0	0	1	
1	1	15	2	2	2	1	1	1	3	3	2	1	2	3	1	0	
1	1	15	2	1	2	4	2	2	4	2	2	1	3	0	0	1	
1	1	16	2	1	2	3	3	3	3	2	1	1	2	0	0	1	
1	2	16	2	2	2	4	3	4	3	4	2	1	2	0	0	1	
1	2	16	2	2	2	2	2	3	3	2	2	1	2	0	0	0	
1	1	17	2	1	1	4	4	3	5	2	2	2	2	0	1	1	
1	2	15	2	2	1	3	2	4	3	2	2	1	2	0	0	1	
1	2	15	2	1	2	3	4	3	3	2	2	1	2	0	0	1	
1	1	15	2	1	2	4	4	5	2	4	2	1	2	0	0	1	
1	1	15	2	1	2	2	1	4	3	4	1	3	3	0	0	1	
1	2	15	2	2	2	4	4	2	4	1	1	1	1	0	0	1	
1	2	15	2	1	2	4	3	5	3	1	2	2	2	0	0	1	
1	2	15	2	1	1	2	2	3	3	2	3	1	3	0	0	1	
1	1	16	2	1	2	4	4	2	3	2	2	1	1	0	0	1	
1	1	16	2	1	2	4	4	4	4	4	2	1	3	0	0	1	
1	1	16	2	1	2	3	3	3	3	4	2	3	2	0	1	1	
1	2	17	2	1	2	3	2	4	4	1	2	1	1	3	0	1	
1	2	16	2	2	2	4	3	2	3	2	1	1	1	0	0	0	
1	2	15	2	1	2	4	3	5	3	4	2	1	2	0	0	0	
1	2	15	2	1	2	4	4	2	2	3	1	1	1	0	0	1	
1	2	16	2	2	2	4	2	5	3	1	2	1	2	0	0	0	
1	2	16	2	2	2	2	2	3	3	4	2	2	2	0	0	1	

Fig. 6. Transformed dataset with numerical data

Data Normalization

Feature scaling or normalization is a method which can be used to normalize data, it scale the range of independent variables. In data processing step, it is known as data scaling or data normalization. It will help us to achieve cost and time effective model quick convergence. The process requires to take each attribute, and convert through minmax normalization except the binary attributes because there is no reason to scale 1 and 0. Data after normalization has shown in Fig. 7.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid
0	0	0.4285714	1	0	0	1.00	1.00	0.00	1.00	0.0000000	0.5	0.3333333	0.3333333	0.0000000	1	0	
0	0	0.2857143	1	0	1	0.25	0.25	0.00	0.50	0.0000000	0.0	0.0000000	0.3333333	0.0000000	0	1	
0	0	0.0000000	1	1	1	0.25	0.25	0.00	0.50	0.6666667	0.5	0.0000000	0.3333333	1.0000000	1	0	
0	0	0.0000000	1	0	1	1.00	0.50	0.25	0.75	0.3333333	0.5	0.0000000	0.6666667	0.0000000	0	1	
0	0	0.1428571	1	0	1	0.75	0.75	0.50	0.50	0.3333333	0.0	0.0000000	0.3333333	0.0000000	0	1	
0	1	0.1428571	1	1	1	1.00	0.75	0.75	0.50	1.0000000	0.5	0.0000000	0.3333333	0.0000000	0	1	
0	1	0.1428571	1	1	1	0.50	0.50	0.50	0.50	0.3333333	0.5	0.0000000	0.3333333	0.0000000	0	0	
0	0	0.2857143	1	0	0	1.00	1.00	0.50	1.00	0.3333333	0.5	0.3333333	0.3333333	0.0000000	1	1	
0	1	0.0000000	1	1	0	0.75	0.50	0.75	0.50	0.3333333	0.5	0.0000000	0.3333333	0.0000000	0	1	
0	1	0.0000000	1	0	1	0.75	1.00	0.50	0.50	0.3333333	0.5	0.0000000	0.3333333	0.0000000	0	1	
0	0	0.0000000	1	0	1	1.00	1.00	0.25	1.00	0.0000000	0.5	0.0000000	0.3333333	0.0000000	0	1	
0	0	0.0000000	1	0	1	0.50	0.25	0.75	0.50	1.0000000	0.0	0.6666667	0.6666667	0.0000000	0	1	
0	1	0.0000000	1	1	1	1.00	1.00	0.25	0.75	0.0000000	0.0	0.0000000	0.0000000	0.0000000	0	1	
0	1	0.0000000	1	0	1	1.00	0.75	1.00	0.50	0.0000000	0.5	0.3333333	0.3333333	0.0000000	0	1	
0	1	0.0000000	1	0	0	0.50	0.50	0.50	0.50	0.3333333	1.0	0.0000000	0.6666667	0.0000000	0	1	
0	0	0.1428571	1	0	1	1.00	1.00	0.25	0.50	0.3333333	0.5	0.0000000	0.0000000	0.0000000	0	1	
0	0	0.1428571	1	0	1	1.00	1.00	0.75	0.75	1.0000000	0.5	0.0000000	0.6666667	0.0000000	0	1	
0	0	0.1428571	1	0	1	0.75	0.75	0.50	0.50	1.0000000	0.5	0.6666667	0.3333333	0.0000000	1	1	
0	1	0.2857143	1	0	1	0.75	0.50	0.75	0.75	0.0000000	0.5	0.0000000	0.0000000	1.0000000	0	1	
0	1	0.1428571	1	1	1	1.00	0.75	0.25	0.50	0.3333333	0.0	0.0000000	0.0000000	0.0000000	0	0	
0	1	0.0000000	1	0	1	1.00	0.75	1.00	0.50	1.0000000	0.5	0.0000000	0.3333333	0.0000000	0	0	
0	1	0.0000000	1	0	1	1.00	1.00	0.25	0.25	0.6666667	0.0	0.0000000	0.0000000	0.0000000	0	1	
0	1	0.1428571	1	1	1	1.00	0.50	1.00	0.50	0.0000000	0.5	0.0000000	0.3333333	0.0000000	0	0	
0	1	0.1428571	1	1	1	0.50	0.50	0.50	0.50	1.0000000	0.5	0.3333333	0.3333333	0.0000000	0	1	

Fig. 6. Transformed dataset with numerical data

Finally, we have complete data pre-processing. Moving forward for next step is to perform Exploratory Data Analysis (EDA). Through visualization we can get insights of data and select important features.

Exploratory Data Analysis (EDA)

After data preprocessing, we will perform EDA in the next step through dataset visualization and non-visualized techniques as well. Firstly we will visualize each features and go in to deeper after this we will understand the impactful features to predict student's performances. We can be able to find useful patterns, insights, trends and correlations using EDA that might not possible to detect. In EDA, we will be able to understand the dataset by using it in a visual context using R libraries such as: ggplot2. There are various ways and methods to visualize a dataset. In this study we will implement to:

- Univariate analysis through plotting histogram to check the distributions so that we can visualize the number of observations that are in each particular feature of dataset. E.g. Higher education histogram or distribution indicates that our dataset consists of more than 350 students who want to attend higher education, while there are around 30 students who do not want to go for higher studies.
- Plot “correlation matrix” to remove multi collinearity and check the correlation of different attributes with each other as well as with class attribute.

Univariate analysis

In this step, we are going to plot distributions of each features and extract the best demographic as well as social, and school conditions which impact the students’ performance. Let’s check the class distribution first, we can analyze through Fig. 7 that the target attribute is balanced and showing normal distribution. So, we have to balance the target attribute, depending on the Machine Learning (ML) algorithms, we have same probability for both classes.

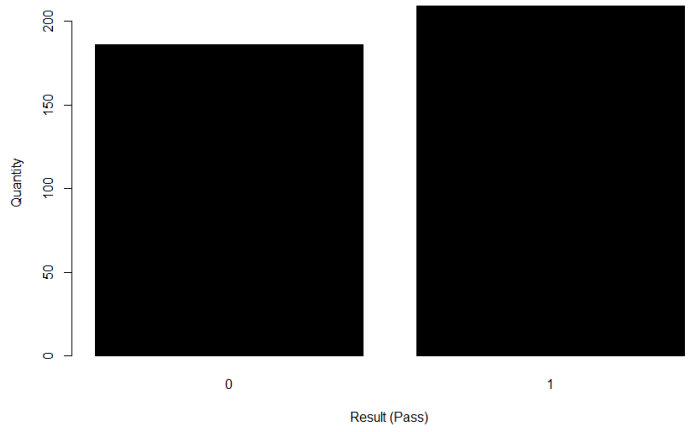

Fig. 7. Distribution of class

Therefore, class distribution is normal and the class attribute is normal. 206 students are passed (1) while 186 are failed (0) in assessment.

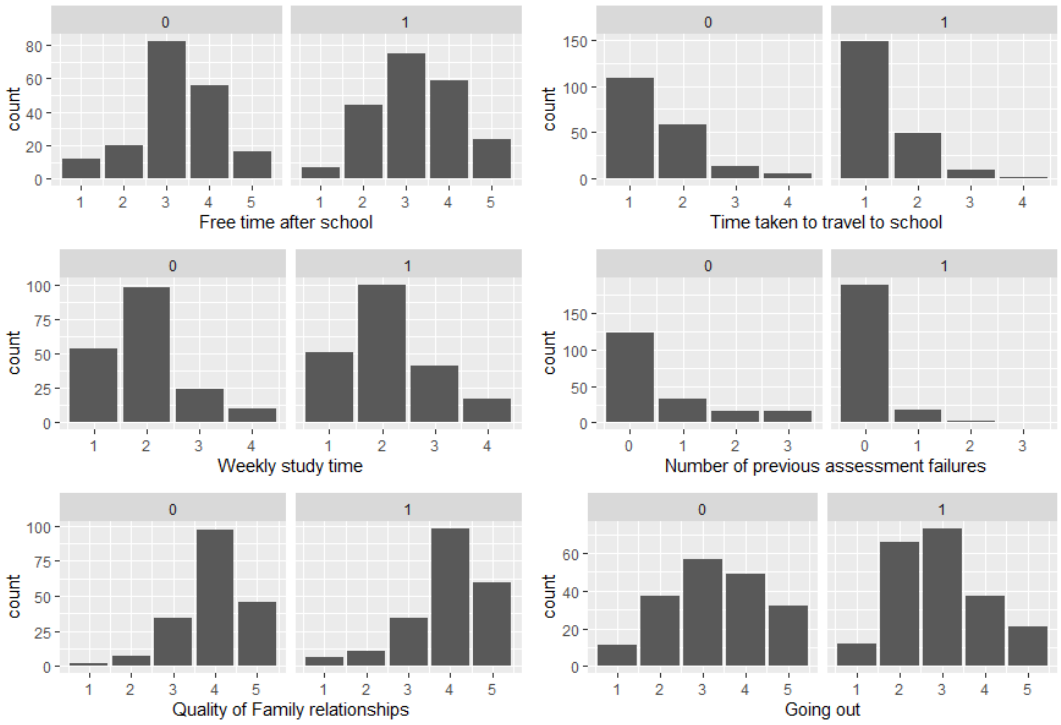

Fig. 8. Distribution of data

Fig.8. showing the distribution of different attributes, going out, free time, and weekly study time are showing normal distributions, family relationships are right skewed and rest 2 features are left skewed. More students are spending average free time after their school.

- For the correlation b/w the previous failures and the class, there is a strong relationship, where the more students failed in the previous class, the less chances to pass. The students who have less previous failures more likely to pass the assessment.
- All the student with quality family relationships have almost same ratio to pass or fail.
- It seems that more students who pass the assessment spending less hours to go out compared to the students who got fail.
- Most of students who passed the assessment studying 5-10 hours weekly.

Students who are going with friends count in dataset is showing normal distribution. Most of the student's relationships quality is high. All this data is shown in Fig. 8.

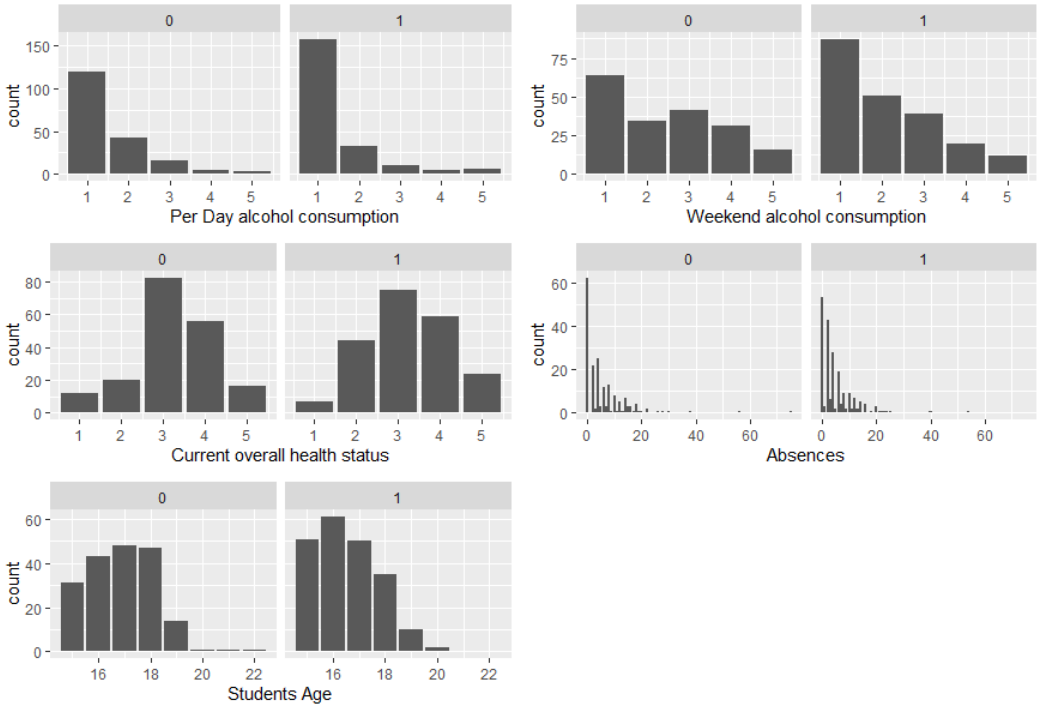

Fig. 9. Distribution of data

From fig.9 we can analyze the distribution of alcohol consumption, health, absences, and student age.

- In terms of age, most of the students who are 14 to 18 passed the assessment, but considerable students have failed in the same age group.
- For weekly alcohol consumption it does not showing a strong impact as even students with low alcohol consumption also failed but if consume les alcohol then more chances to pass. Similar trend in per day alcohol consumption, student with very less amount of alcohol also fail but higher chances to pass.
- Mostly, student who fails the assessments have bad health, with good health more likely to pass.

In Fig. 10, attributes that are related to the student's school GP are more in number, and a very few students belongs to MS. It doesn't make any difference because the students from any school showing almost same ratio to fail or pass. Likewise student address distribution in both rural and urban, more students belongs to urban areas and more students from urban areas pass but major number of students fail as well.

Fig.10 shows that more students' parents are together but T is almost equally distributed in in both classes 1 and 0. Student's family size is not showing tendency to associate with only pass or fail. Mother with higher level of education had very good impact on students' performance similarly father with higher education have positive affect for students. Family size doesn't had an impact on student performance. Student who has Greater than 3 family members pass as well as fail, similar trend for those who has small family.

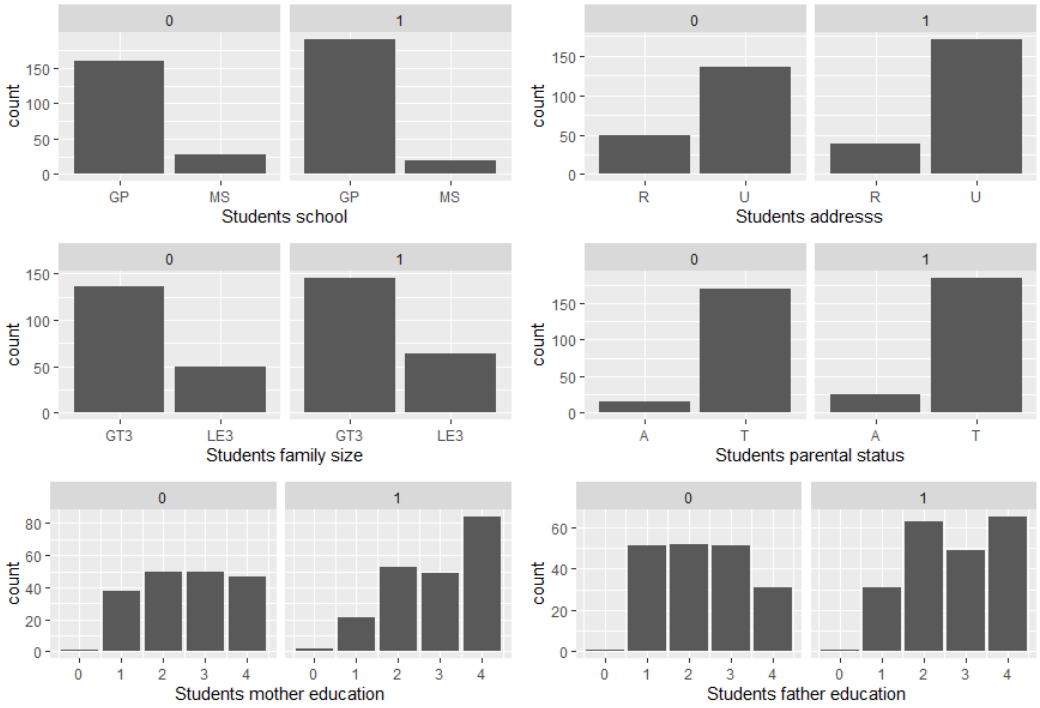

Fig. 10. Distribution of data

In Fig. 11, If Mother's job is grouped into five categories – House_wife, Employment sectors (Health, Education, Other services), and other. The classification into other and services seems very important as these both categories are highly crucial and showing positive impact on student's performance and a mother who is at home should have more time to focus the children but it is showing negative correlation if mother is at home show more failures. In addition, the distinction of health and education is also important and classified more pass students. Trend for students is almost who fail or pass with father's job in other and services category.

Students who are paying for extra classes does not making any difference on pass or fail ratio. Similarly for students with family support are not biased to one class. More students are not taking extra educational support from school, while few students who are taking support are more likely to fail.

Reason for choosing a school have not any impact on students' performance, student choosing on the basis of reputation of school more likely to pass.

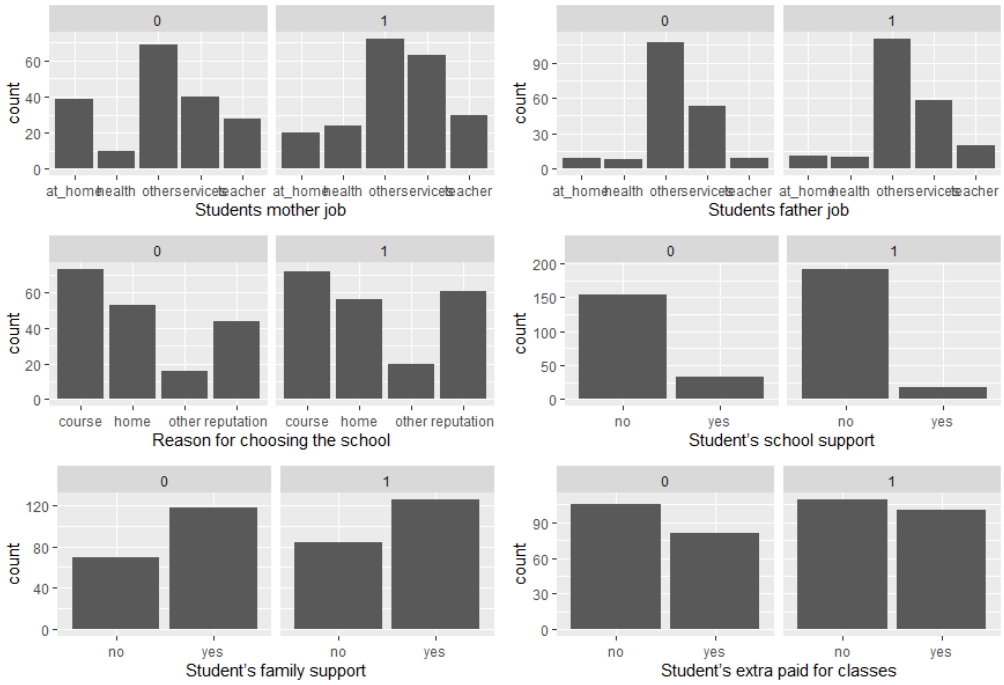

Fig. 11. Distribution of data

Internet accessibility has a slight role in increasing students' performance, as seen in the in Fig. 12. More students passed the exam who want to take higher. The number of the students who are in romantic relationship are less than the students who are not. From the Fig. 12, the distribution of the class, whether in romantic relationship or not, is somehow similar. For passed students, it tends to be owned by more students who are not in any romantic relationship.

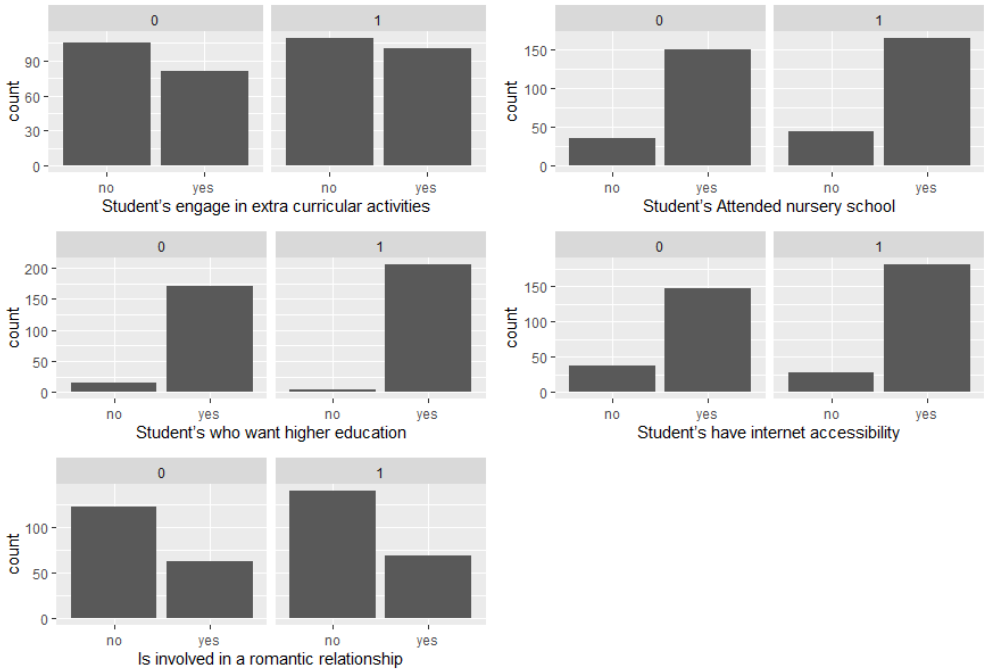

Fig. 11. Distribution of data

Students who wants to continue to higher studies can trigger students to be more active in education so more likely the students pass the assessment. There is general summary of after this analysis that most of the students pass if they:

- Do not go out frequently
- Is not in any romantic relationship
- Mother have higher education
- Who to tend to go for higher education
- Mother is from services or other profession
- Less number of absences in classes
- Have Internet accessibility
- Good health

Feature Selection

After Exploratory Data Analysis (EDA), we performed the Chi-Square test on the student grade dataset to choose the significant attributes which have more dependency with the class attribute. Dataset may contain several irrelevant and inappropriate attributes, which will effect accurate classification. There is another problem called high dimensionality in which included numerous features and characteristics which can influence the performance of student performance such as demographics, social, family and educational background. Feature selection will help to reduce dimensions from the dataset. An important and relevant features have more association with the class attribute. The top 12 important with more association features are selected then arranged from high relevancy to low association value. Table 1. contains the top 12 relevant features along with their relevancy score.

Features	X-Squared	df	p-value
School	2.3126	1	0.1283
Sex	1.7517	1	0.1857
Age	11.49	7	0.1186
Address	3.8142	1	0.05082
Famsize	0.50079	1	0.4792
Pstatus	0.86024	1	0.3537
Medu	14.489	4	0.005886
Fedu	16.729	4	0.002182
Mjob	15.867	4	0.003203
Fjob	3.5341	4	0.4727
Reason	1.9537	3	0.5821
Guardian	2.1194	2	0.3466
Travel time	7.7536	3	0.05139
Study time	5.0447	3	0.1685
Failures	42.628	3	2.952e-09
Schoolsup	6.5056	1	0.01075
Famup	0.27746	1	0.5984
Paid	0.56957	1	0.4504

Activities	0.00089177	1	0.9762
Nursery	0.16801	1	0.6819
Higher	7.8199	1	0.005167
Internet	3.0107	1	0.08272
romantic	0.0053739	1	0.9416
famrel	3.4154	4	0.4909
freetime	11.004	4	0.02652
goout	12.839	4	0.01209
Dalc	8.0289	4	0.09052
Walc	8.5872	4	0.07229
health	4.8557	4	0.3024
absences	36.094	33	0.3259

Table.1 Chi-Square Test

After performing the Chi-Square Test, the correlation of 12 selected attributes with target variable in students grade dataset is as follows: failures (42.628), absences (36.094), fedu (16.729), mjob (15.867), goout (12.839), medu (14.489), freetime (11.004), age (11.49), walc (8.5872), dalc (11.49), higher (7.8199), traveltime (7.7536).

Multivariate and Multicollinearity analysis

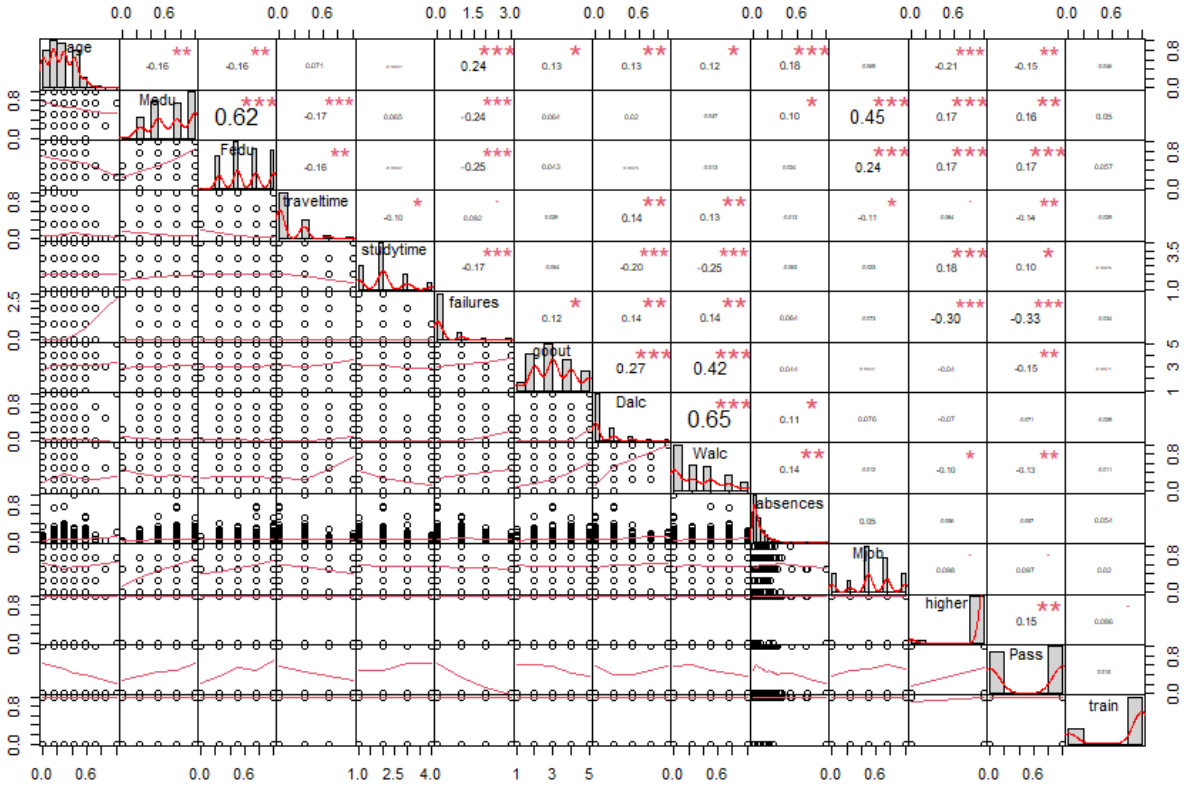

Fig. 12. Correlation Matrix

The 12 features with highest squared value among all other variables are selected for multivariate analysis and multicollinearity analysis. In this multivariate and Multicollinearity analysis, we need to considered correlation of two independent features with each other and with the class variable as well. The relationship of different attributes with each other is not clear. We will plot the correlation matrix to find out how strongly the features relate to each other and eliminate multicollinearity. After plotting correlation matrix using all features with response variables, shown in Fig. 12. There is neglect able or no multicollinearity in the dataset. Multicollinearity occurs when the independent features are highly correlated with each other. Only Mjob and Fjob is showing strong correlation with 0.62 and another is between Dalc and Walc which is 0.65, but this is what we can neglect.

Classification Models

Our task is to building a model which will classify and capable of predicting the value of class variable, we have to extract most important features also which are strongly associated with a student pass or fail an assessment. Classification is basically a technique to identify the class label of the data to which it actually belongs. After data pre-processing, exploratory data analysis and feature selection we can build classification model to classify and prediction either students pass or fail using our important selected features. The classifiers we will use are decision trees, random forest, and SVM and eventually we will compare the performance of all models. Before training the model, we will split the dataset into training and test sets. The training set will be used to train and learn the model while test set will use to evaluate the performance. Hence 80% dataset will be used for training and rest 20% for testing.

All the Machine Learning (ML) models implemented on the entire dataset as well as selected important features. The actual classification of whole dataset of pass (1) or fail (0) the assessment is as follows: 206 students are passed (1) while 186 are failed (0) in assessment.

Decision Tree

A Decision Tree (DT) uses contains nodes and branches which is a tree similar to a graph. All the nodes (either parent or child) arranged in sequence. Root node represents the whole dataset and usually on the top of the tree. The root node is decides by calculating the entropy. DT works like a flowchart but it is not cyclic. DT is robust and easy to understand algorithm for classification. For student's grade dataset, we are performing classification of pass and fail of students and using DT technique because it is simple, and easy to interpret. Another reason to prefer this model is the gathered information through trees which make the training better and make the decision more effective. We prepared test data already which will be unseen data for model to evaluate the performance of model which is trained using training dataset. When implemented on test dataset, it will predict the student performance on the basis of previous learning and generate class column, which will compared with actual values. For this purpose we used predict function given in Fig. 16.

```
#Decision tree
grade_tree <- rpart(Pass~., data = trainset1, method = 'class')
rpart.plot(grade_tree, extra = 106)
test_data <- testset1[-13]
tree_pred <- predict(grade_tree, newdata = test_data, type = 'class')
table(predicted = tree_pred, actual = testset1$Pass)
mean(tree_pred==testset1$Pass)
```

Fig.16 Building Decision Tree Model

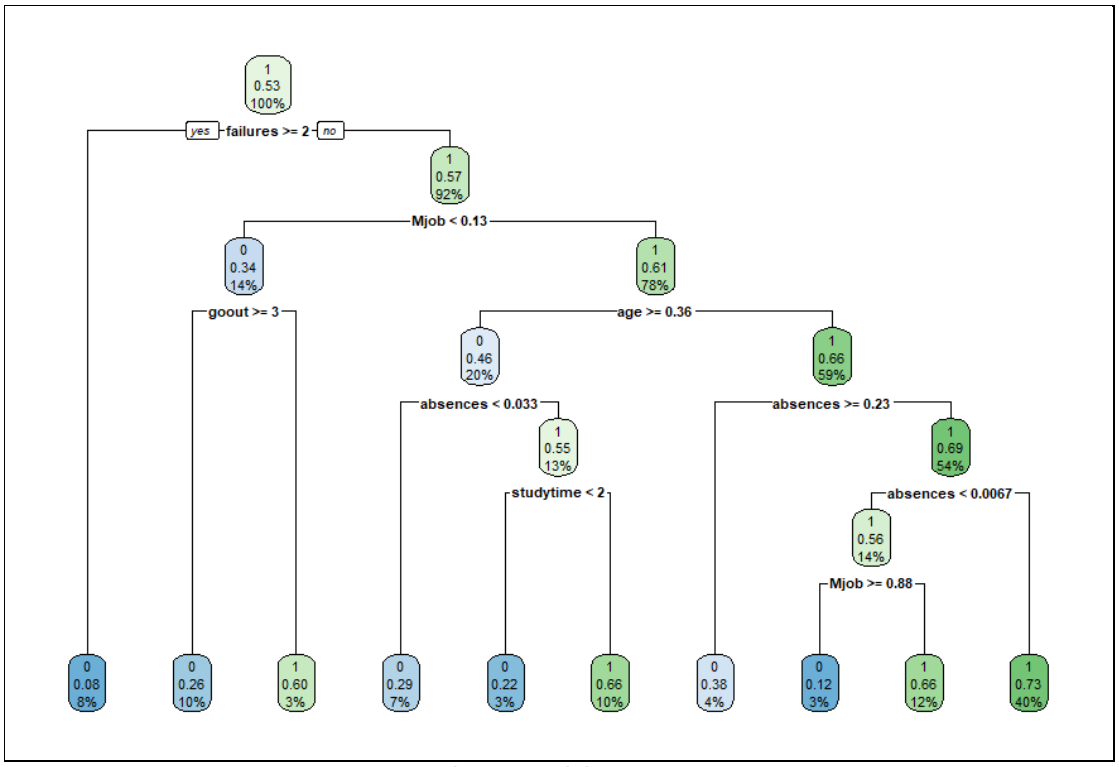

Fig.17 Decision Tree

Random Forest

Random Forest (RF) is collection of tress and an ensemble method. It is also supervised Machine Learning technique, used for classification. The forest created through randomly generated decision trees and used bagging method to train. Random forest normally used to get better accuracy. This achieve by creating several decision trees then combine them to get more accurate prediction.

The models evaluation using different evaluation parameters provide feedback about the important features selected by different methods. We used 50 number of decision trees, this is what we can change as well. We used RF to check feature importance as well by using ForestVarImp.

```
#random forest
forest_grade <- cforest(Pass~., data = trainset1, control = cforest_unbiased(mtry = 10, ntree = 50))
rf_prob <- predict(forest_grade, newdata = test_data, type = "response")
rf_pred <- ifelse(rf_prob>0.5, 1, 0)
table(predicted = rf_pred, actual = testset1$Pass)
mean(rf_pred==testset1$Pass)
ForestVarImp <- varimp(forest_grade)
barplot(ForestVarImp)
```

Fig.18 Random Forest

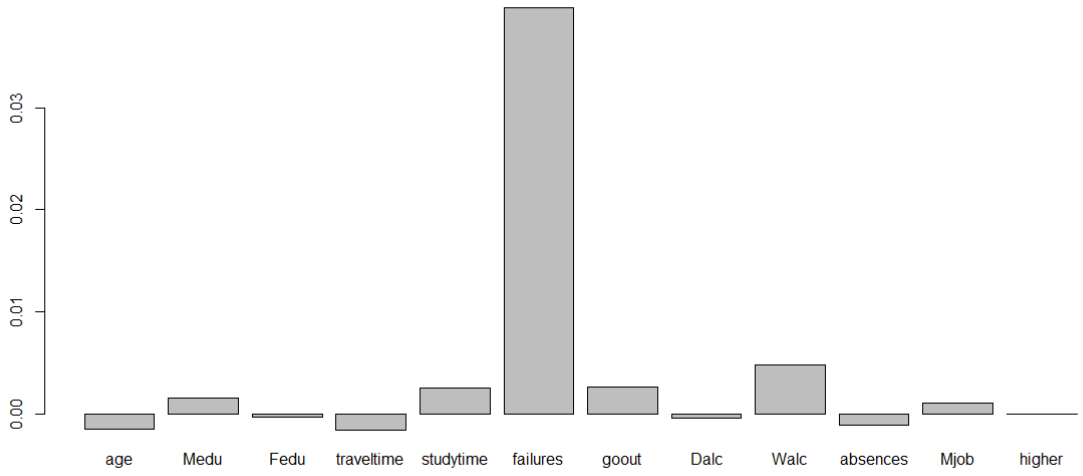

Fig.19 Important Feature Plot

From the Fig. 19 we can see failure is plotted as most important feature in the classification of student grades and we can recall the chi-square analysis which gave high x square value for failure attribute. We also plotted the important features using same function for entire dataset and the results were as given by chi-square which shows the reliability of selected features after EDA, correlation matrix, chi-square and random forest features important plot.

Support Vector Machine (SVM)

In this step, we will implement the popular classification algorithm Support Vector Machine (SVM). It is useful for classification and regression tasks. The goal is to draw the best hyperplane to decide the class. SVM chooses the data points called vectors which help to identify the hyperplane. These data points are known as support vectors. So it will use hyperplane or decision boundary to separate either student will pass or fail. The code for SVM in Rstudio is given in Fig. 20.

```
#SVM
svm_trainset <- trainset1
str(svm_trainset)
svm_trainset$Pass <- as.factor(svm_trainset$Pass)
svm_grade <- svm(Pass~., data = svm_trainset)
svm_pred <- predict(svm_grade, newdata = test_data, type =
                      "response")
table(predicted = svm_pred, actual = testset1$Pass)
mean(svm_pred==testset$Pass)
```

Fig. 20 Support Vector Machine

Model Evaluation

The performances of all classification models implemented on student's grade dataset evaluated using different ordinary evaluation parameters.

Confusion Matrix

First of all, confusion matrix used to assess the models which compared the actual values and predicted values by models. Confusion matrix used 4 different combinations as follows:

- True Positive (TP): If actual is positive and predicted positive.
- True Negative (TN): If actual is negative and predicted as negative.
- False Positive (FP): If predicted value is positive, but actually it was negative.
- False Negative (FN): If predicted value is negative, but actually it was positive.

It will provided us table which helped us to evaluate the model and know exactly about actual and predicted values.

Accuracy

To calculate the accuracy of the model mean function is used or model evaluation, other parameter used is accuracy. The model accuracy can be calculated by using sum of TP and TN, divide by the sum of all values in confusion. Overall accuracy of models were calculated using the formula given in Fig. 21.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig.21 Accuracy

F1 Score, Precision and Recall

Furthermore, we will use F1 score, precision and recall will calculated, these will be calculated by using values of confusion matrix and determined by using the formulas given Fig. 22.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Fig.29 Precision and Recall

Results and Discussion

In the first experiment, we implemented 3 different algorithms (decision tree, random forest, and Support Vector Machine) without selecting important features. Experiment performed by choosing entire dataset and results are reported in Table. 2. Highest accuracy achieved by SVM (0.66) which is high as compared to Decision Tree and Random Forest. However, second experiment performed by using selective and important features, compared these algorithms by using different evaluation parameters. As shown in Table 3, the best performance shown by Random forest when we are reducing dimensionality and implementing it on selected features. Results of best performed algorithm is also given in Fig. 30.

Models evaluation with entire dataset					
Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**MCC**
Decision Tree	0.53	0.45	0.47	0.46	-0.024
Random Forest	0.59	0.62	0.73	0.67	0.163
SVM	0.66	0.67	0.78	0.72	0.30

Table. 2 Models evaluation with entire dataset

```
> #random forest
> forest_grade <- cforest(Pass~., data = trainset1, control = cforest_unbiased(mtry = 10, ntree = 50))
> rf_prob <- predict(forest_grade, newdata = test_data, type = "response")
> rf_pred <- ifelse(rf_prob>0.5, 1, 0)
> table(predicted = rf_pred, actual = testset1$Pass)
         actual
predicted  0  1
        0 18  9
        1 17 35
> mean(rf_pred==testset1$Pass)
[1] 0.6708861
```

Fig.30 Highest accuracy by Random Forest

Models evaluation with important features					
Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 score**	**MCC**
Decision Tree	0.59	0.64	0.64	0.6	0.179
Random Forest	0.63	0.67	0.80	0.72	0.32
SVM	0.65	0.84	0.64	0.73	0.27

Table. 2 Models evaluation with important features

Conclusion

This study conduct an analysis using Machine Learning techniques to classify and predict the students' performance in assessment either they pass or fail on the basis of their educational, demographic and social features. Three popular classification algorithms (decision tree, random forest, and support vector machine) were implemented and extensive experiments were performed using entire dataset and with selected important features. Then comparative analysis using ordinary evaluation parameters such as accuracy, precision, recall, and f1 score. Explore dataset using EDA and then important features selected using correlation matrix and chi-square test which improved the classification performance. The chi-square method as well as the random forest variable important function indicates same important features. After performing the Chi-Square Test, 12 important variables were selected those are strongly associated with students' performance such as failures, absences, fedu, mjob, goout, medu, freetime, age, walc, dalc, higher, and traveltime. Overall, better classification accuracy achieved by selecting important features and implementing Random Forest. In the future, other feature selection techniques such as Principle Component Analysis (PCA) and Genetic Algorithm (GA) can be used to improve the accuracy. In addition, Neural Networks and deep learning techniques can also be utilized on this datasets for better classification and effective results.