[Type the company name]

Data Analytics Project

(Analysis of SSLC data set)

Under the guidance of:

Prof. Chandrasekhar R

Group 11

{

Joydeep (MT2013062)

}

14

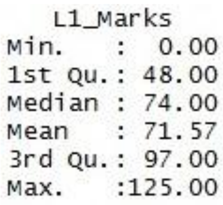
Contents

Section 1: Descriptive analytics of the data

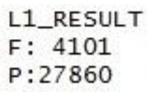
The data that was provided had 36 attributes and 33003 rows.

After having a detailed discussion we decided that we will not consider ant students data who have been absent in at least one of the exams. So After removing those rows we were left with 31962.

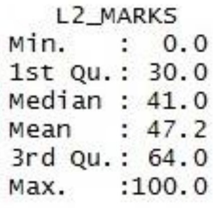
The descriptions are below:



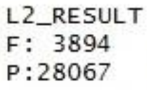
```
  L1_Marks
Min.   :  0.00
1st Qu.: 48.00
Median : 74.00
Mean   : 71.57
3rd Qu.: 97.00
Max.   :125.00
```



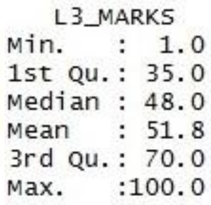
```
L1_RESULT
F: 4101
P:27860
```



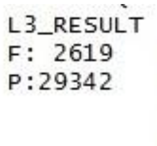
```
   L2_MARKS
Min.   :  0.0
1st Qu.: 30.0
Median : 41.0
Mean   : 47.2
3rd Qu.: 64.0
Max.   :100.0
```



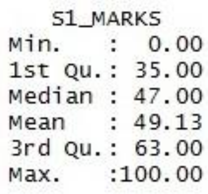
```
L2_RESULT
F: 3894
P:28067
```



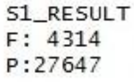
```
   L3_MARKS
Min.   :  1.0
1st Qu.: 35.0
Median : 48.0
Mean   : 51.8
3rd Qu.: 70.0
Max.   :100.0
```



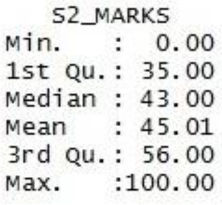
```
L3_RESULT
F: 2619
P:29342
```



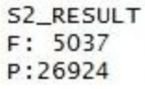
```
  S1_MARKS
Min.   :  0.00
1st Qu.: 35.00
Median : 47.00
Mean   : 49.13
3rd Qu.: 63.00
Max.   :100.00
```



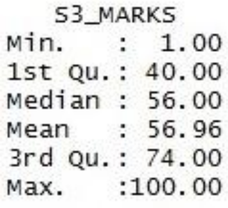
```
S1_RESULT
F: 4314
P:27647
```



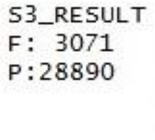
```
   S2_MARKS
Min.   :  0.00
1st Qu.: 35.00
Median : 43.00
Mean   : 45.01
3rd Qu.: 56.00
Max.   :100.00
```



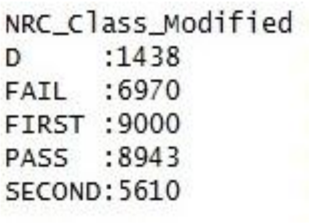
```
S2_RESULT
F: 5037
P:26924
```



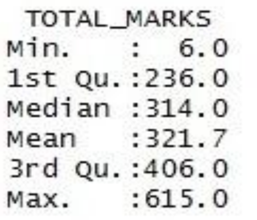
```
   S3_MARKS
Min.   :  1.00
1st Qu.: 40.00
Median : 56.00
Mean   : 56.96
3rd Qu.: 74.00
Max.   :100.00
```



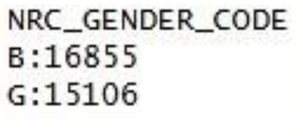
```
S3_RESULT
F: 3071
P:28890
```



```
NRC_Class_Modified
D     :1438
FAIL  :6970
FIRST :9000
PASS  :8943
SECOND:5610
```



```
  TOTAL_MARKS
Min.   :  6.0
1st Qu.:236.0
Median :314.0
Mean   :321.7
3rd Qu.:406.0
Max.   :615.0
```



```
NRC_GENDER_CODE
B:16855
G:15106
```

Section 2: Report of suggested eight experiments

Experiment 1

Objective: Discretization and Classification

Procedure: Using rpart

- We made a different class of marks like L1_CLASS, L2_CLASS, L3_CLASS, S1_CLASS, S2_CLASS, S3_CLASS in the csv file using excel.
- In data preparation we kept 70%of the entire data as training data and rest 30% as test data.
- We built a model using *rpart()* in rpart package. Using formula *NRC_CLASS~L1+L2+L3+S1+S2+S3*
- Then that model was used to predict *NRC_CLASS* for test data.
- Compared with actual NRC_CLASS with predicted one.

Using C5.0

- We made a different class of marks like L1_CLASS, L2_CLASS, L3_CLASS, S1_CLASS, S2_CLASS, S3_CLASS in the csv file using excel.
- In data preparation we kept 70%of the entire data as training data and rest 30% as test data.
- We built a model using *C5.0()* in C50 package. Using formula *NRC_CLASS~L1+L2+L3+S1+S2+S3*
- Then that model was used to predict *NRC_CLASS* for test data.
- Compared with actual NRC_CLASS with predicted one.
- Compared result and performance of both Algorithms.

Results Obtained:

```
Using Rpart
        Pred
true     D    FAIL FIRST PASS SECOND
  D       217    0   218    0      0
  FAIL      0 1966     3   99     34
  FIRST    47    0  2366   25    237
  PASS      0    0    63 2383    245
  SECOND    0    0   548  429    706

Variable importance
S2 L1 L2 S1 S3 L3 
22 20 19 14 12 12 
```

Using C5.0

```
        pred
true     D    FAIL FIRST PASS SECOND
  D       356    0    79    0      0
  FAIL      0 2088     4    5      5
  FIRST    60    1  2425    0    189
  PASS      0    0     0 2504    187
  SECOND    0    0   206  230   1247
```

```
Attribute usage:

100.00%  S2       81.67%  L2       79.17%  L3       77.67%  S1       75.46%  L1
62.33%   S3
```

Conclusions:

- Accuracy rate of rpart is 79.68% while C5.0 has accuracy of 89.92%. Hence C5.0 is better than rpart.
- Attribute usage/importance is comparable.
- Tree of both the algorithms are same till level 1.

Experiment 2

Objective: Regression and Classification

Procedure:

- In the data preparation step we made a new data frame containing L1, L2, L3, S1, S2, S3, TotalMarks.
- Computed z-scores of marks.
- Build different regression models with four, five independent variables and one dependent variable.
- Compared their accuracy with respect to p value.
- In data preparation we kept 70%of the entire data as training data and rest 30% as test data.
- Applied *Knn()* with k = 149.
- Checked accuracy of predicted result.

Results Obtained:

-
  ```
   (Intercept)   Estimate Std. Error  t value Pr(>|t|)
  L1:L2             1.363e-02  6.058e-03    2.250 0.024441 *
  L1:L3            -9.164e-04  6.238e-03   -0.147 0.883206
  L2:L3             4.620e-03  6.177e-03    0.748 0.454506
  L1:S1             3.951e-04  1.114e-04    3.545 0.000393 ***
  L2:S1             2.653e-04  1.060e-04    2.504 0.012285 *
  L3:S1            -2.533e-04  1.123e-04   -2.256 0.024070 *
  L1:S2            -1.793e-03  5.599e-03   -0.320 0.748757
  L2:S2             3.008e-03  5.706e-03    0.527 0.598132
  L3:S2             9.673e-03  6.812e-03    1.420 0.155603
  S1:S2            -3.597e-05  1.183e-04   -0.304 0.761158
  L1:S3                    NA         NA       NA       NA
  L2:S3                    NA         NA       NA       NA
  L3:S3             4.113e-03  5.081e-03    0.810 0.418213
  S1:S3                    NA         NA       NA       NA
  S2:S3                    NA         NA       NA       NA
  L1:L2:L3         -9.758e-04  6.043e-03   -0.161 0.871725
  L1:L2:S1         -1.208e-04  1.154e-04   -1.047 0.295001
  L1:L3:S1          1.100e-04  1.230e-04    0.894 0.371353
  L2:L3:S1         -1.023e-04  1.214e-04   -0.843 0.399379
  L1:L2:S2          1.179e-02  4.469e-03    2.639 0.008311 **
  L1:L3:S2         -4.795e-03  5.498e-03   -0.872 0.383153
  L2:L3:S2         -2.184e-03  5.550e-03   -0.394 0.693902
  L1:S1:S2          4.772e-05  1.084e-04    0.440 0.659743
  L2:S1:S2         -2.134e-05  1.036e-04   -0.206 0.836833
  ```

```
L3:S1:S2          -1.364e-04  1.342e-04   -1.016 0.309523
L1:L2:S3                  NA         NA       NA       NA
L1:L3:S3          -2.685e-03  4.599e-03   -0.584 0.559408
L2:L3:S3           7.132e-03  4.775e-03    1.494 0.135314
L1:S1:S3                  NA         NA       NA       NA
L2:S1:S3                  NA         NA       NA       NA
L3:S1:S3          -1.934e-04  1.017e-04   -1.901 0.057251 .
L1:S2:S3                  NA         NA       NA       NA
L2:S2:S3                  NA         NA       NA       NA
L3:S2:S3           3.506e-03  4.471e-03    0.784 0.433028
S1:S2:S3                  NA         NA       NA       NA
L1:L2:L3:S1       -5.531e-05  1.164e-04   -0.475 0.634508
L1:L2:L3:S2       -1.485e-03  3.493e-03   -0.425 0.670774
L1:L2:S1:S2       -2.395e-04  7.896e-05   -3.033 0.002425 **
L1:L3:S1:S2        1.080e-04  1.046e-04    1.033 0.301840
L2:L3:S1:S2        1.012e-04  1.024e-04    0.988 0.323064
L1:L2:L3:S3       -1.225e-03  3.356e-03   -0.365 0.715028
L1:L2:S1:S3               NA         NA       NA       NA
L1:L3:S1:S3        7.049e-05  9.091e-05    0.775 0.438141
L2:L3:S1:S3       -8.261e-05  9.260e-05   -0.892 0.372342
L1:L2:S2:S3               NA         NA       NA       NA
L1:L3:S2:S3       -4.304e-03  2.755e-03   -1.562 0.118287
L2:L3:S2:S3        2.888e-03  2.934e-03    0.984 0.324994
L1:S1:S2:S3               NA         NA       NA       NA
L2:S1:S2:S3               NA         NA       NA       NA
```

- Choosing L1,S1 marks

Conclusions:

- If we drop four variables L2,L3, S2 and S3 we can predict the class with 99% (using Knn) accuracy which was earlier 90% (using C5.0). Hence we can say that S1 and S3 are may not be required to predict the overall class of the student.

Experiment 3

Objective: Clustering and association rules

Procedure:

- Took L1, L2, L3, S1, S2, S3 marks, replaced the marks with their respective z-scores.
- Applied k-means algorithm and assigned cluster no to each data point.
- Replaced each z-score of L1, L2, L3, S1, S2, and S3 by their respective class and factored them.
- Applied Apriori algorithm to get association rules.
- Pruned them with class.

Results Obtained:

-
  ```
  entireDataCluster$size
  [1] 4658 8550 6475 8169 4109
  ```

- entireDataCluster$centers

```
      L1         L2            L3             S1             S2            S3
1 -1.4355413  -1.1347246   -1.17006378  -1.3244764330   -1.35110492  -1.3545179
2 -0.5981309  -0.5625785  -0.60126552  -0.5350373860   -0.51202986  -0.5987636
3  0.7480054   0.7140014   0.75890751   0.6819094620   0.58009531    0.6889489
4  0.2292555  -0.1300286  -0.05699718  -0.0008792643   0.01507892    0.1678569
5  1.2374416   1.5903231   1.49492849   1.5419322996   1.65295820    1.3620361
```

-
  ```
  lhs              rhs          support confidence     lift
  1 {clusters=1} => {L3=PASS} 0.2134789  0.7980117 1.879671
  2 {clusters=1} => {S2=PASS} 0.2247427  0.8401170 1.781159
  3 {clusters=1} => {L2=PASS} 0.2149182  0.8033918 1.696321
  4 {S1=PASS}    => {S2=PASS} 0.2872876  0.7302951 1.548322
  5 {S1=PASS}    => {L2=PASS} 0.2777135  0.7059572 1.490592
  6 {L3=PASS,
     S2=PASS}    => {L2=PASS} 0.2101311  0.7709792 1.627883
  ```

Conclusions:

- In cluster 1 most of the students have passed in L2(confidence = 80%),S2(confidence = 84%),L3(confidence = 80%).
- Those who have passed in S1 have passed in S2 (confidence = 73%)
- Those who have passed in S2 and L3 have passed in L2 (confidence = 77%)
- Those who have passed in S1 have passed in L2 (confidence = 70%)

Experiment 4

Objective: Confidence interval

Procedure:

- Wrote a sql quey to get total no of students and students passed and grouped them by district.
- Calculated pass percentage for each district.
- Calculated confidence interval.
- A new attribute is assigned based on the confidence interval obtained.
- Took top two and bottom two districts. Based on these districts a new data frame is made of students from these districts.
- Association rules generated on these data.
- Similar process for selecting top n bottom schools except selecting only those schools who have more than 15 students and then applying *arules()* to get association rules.

Results Obtained:

Top 3 rules of top 2 districts (PA = SIRSI, GA= UDUPI)

```
  lhs                    rhs                           support confidence      lift
1 {DIST_CODE=GA}      => {URBAN_RURAL=R}             0.5201923  0.8110945 1.1217264
2 {URBAN_RURAL=R}     => {NRC_MEDIUM=K}              0.6019231  0.8324468 1.1185332
3 {URBAN_RURAL=R}     => {CANDIDATE_TYPE=RF}         0.6750000  0.9335106 1.0029453
```

Top 3 rules of bottom two districts (QA= YADGIR, SS= BIDAR)

```
  lhs                    rhs                           support confidence      lift
1 {URBAN_RURAL=R}     => {NRC_MEDIUM=K}              0.5362022  0.8194154 1.0856328
2 {URBAN_RURAL=R}     => {CANDIDATE_TYPE=RF}         0.5184426  0.7922756 1.0051052
3 {DIST_CODE=SS}      => {NRC_PHYSICAL_CONDITION=N} 0.7247268  0.9971805 1.0012841
```

Top 3 rules of top 20 schools which have pass percent more than 81% and no of students more than 15.

```
1 {NRC_MEDIUM=E}              => {URBAN_RURAL=U}     0.4187817  0.8918919 1.237343
2 {NRC_GENDER_CODE=G}         => {URBAN_RURAL=U}     0.4086294  0.8518519 1.181794
3 {NRC_MEDIUM=K}              => {SCHOOL_TYPE=A}     0.3299492  0.7386364 1.119318
```

Top 3 rules of bottom 24 schools who have pass percent more than 81% and no of students more than 15

```
  lhs                    rhs                           support confidence      lift
1 {NRC_MEDIUM=K}      => {SCHOOL_TYPE=G}             0.5774648  0.8541667 1.1737903
2 {URBAN_RURAL=U}     => {NRC_PHYSICAL_CONDITION=N} 0.8122066  0.9971182 1.0018215
3 {NRC_MEDIUM=K}      => {NRC_PHYSICAL_CONDITION=N} 0.6737089  0.9965278 1.0012284
```

Conclusions:

- Udupi district is rural and it is one of the best performing districts.
- In the top performing districts which is in Rural area Most student opted for Kannada medium and their type is RF.
- In the worst performing district BIDAR Most students are Normal in physical condition.
- In top performing schools which are in urban areas, students have attempted the exam in English medium.
- In top performing schools which are in urban areas Girls have performed well.
- In top performing Govt. schools the medium of students is Kannada.

Experiment 5

Objective: Urban / Rural characterization.

Procedure:

- Made a new data frame that has *SCHOOL_TYPE, URBAN_RURAL, NRC_CASTE_CODE, NRC_GENDER_CODE, NRC_MEDIUM, NRC_PHYSICAL_CONDITION, CANDIDATE_TYPE.*
- Factored every attribute in the data frame.
- Applied *apriori()* to generate arules.
- Removed all redundant rules.
- Pruned on *URBAN_RURAL* attribute.
- Add marks L1, L2, L3, S1, S2, S3, Total marks and apply 2,3,4,5.

Results Obtained:

Just one rule

```
  lhs                 rhs               support confidence     lift
1 {NRC_MEDIUM=K} => {URBAN_RURAL=R} 0.493758  0.7126535 1.251215
```

After adding marks to the data frame

```
  lhs                            rhs              support confidence      lift
1 {NRC_MEDIUM=K}              => {URBAN_RURAL=R} 0.4937580  0.7126535 1.2512151
2 {L2_CLASS=PASS}             => {URBAN_RURAL=R} 0.3065611  0.6472881 1.1364522
3 {NRC_GENDER_CODE=B}         => {URBAN_RURAL=R} 0.3065611  0.5813112 1.0206156
4 {CANDIDATE_TYPE=RF}         => {URBAN_RURAL=R} 0.5204155  0.5743043 1.0083135
5 {NRC_PHYSICAL_CONDITION=N} => {URBAN_RURAL=R} 0.5686305  0.5698608 1.0005120
6 {NRC_CASTE_CODE=4}          => {URBAN_RURAL=R} 0.3865649  0.5421237 0.9518137
```

Conclusions:

- Most of the rural Schools are of **kannada** medium.
- Most of the students who have passed in L2 belong to rural areas.
- **Most of the students are male in rural areas.**
- **Most of the rural area students belong to general category.**

Experiment 6

Objective: Performance characteristics.

Procedure:

- In data preparation take those rows of students in which *NRC_CLASS* is either *FAIL* or *I.*
- Made a data frame containing *SCHOOL_TYPE, URBAN_RURAL, NRC_CASTE_CODE, NRC_GENDER_CODE, NRC_MEDIUM, NRC_PHYSICAL_CONDITION, CANDIDATE_TYPE.*
- Factor the attributes.
- Generate arules using *apriori()* .

- Remove the redundant rules.
- Prune then based on NRC_Class = D or NRC_Class = FAIL.
- Add marks L1,L2,L3,S1,S2,S3,Total marks and apply 2,3,4,5.

Results Obtained:

Arules generated based on *SCHOOL_TYPE, URBAN_RURAL, NRC_CASTE_CODE, NRC_GENDER_CODE, NRC_MEDIUM, NRC_PHYSICAL_CONDITION, CANDIDATE_TYPE.*

```
  lhs                            rhs                         support confidence      lift
1 {NRC_GENDER_CODE=B}        => {NRC_Class_Modified=FAIL} 0.5195052  0.8769323 1.0578547
2 {NRC_CASTE_CODE=4}         => {NRC_Class_Modified=FAIL} 0.5164129  0.7695853 0.9283605
3 {NRC_MEDIUM=K}             => {NRC_Class_Modified=FAIL} 0.6304710  0.9332746 1.1258211
```

```
  lhs                    rhs                         support confidence     lift
1 {NRC_MEDIUM=E,
   CANDIDATE_TYPE=RF} => {NRC_Class_Modified=D   } 0.1230971  0.5634186 3.294314
```

Conclusions:

- **Most of the boys have Failed in the examination**
- **Most of the general category students have failed the examination**
- **Most of the students who belong to kannada medium have failed**.

Experiment 7

Objective: Decision tree vis-à-vis A-rules

Procedure:

- Make a new data frame that has L1_CLASS,L2_CLASS, L3_CLASS, S1_CLASS,S2_CLASS, S3_CLASS, NRC_CLASS.
- Factor all attributes.
- Apply aprori() to get association rules.
- Remove redundant rules.
- In data preparation we kept 70%of the entire data as training data and rest 30% as test data.
- We built a model using *C5.0()* in C50 package. Using formula *NRC_CLASS~L1+L2+L3+S1+S2+S3*
- Then that model was used to predict *NRC_CLASS* for test data.
- Compared with actual NRC_CLASS with predicted one.

Results Obtained:

Association rules obtained are

```
  lhs                            rhs               support confidence     lift
1 {NRC_Class_Modified=PASS} => {S2_CLASS=PASS} 0.2560308  0.9150173 1.939958
2 {S1_CLASS=PASS}           => {S2_CLASS=PASS} 0.2872876  0.7302951 1.548322
3 {S1_CLASS=PASS}           => {L2_CLASS=PASS} 0.2777135  0.7059572 1.490592
4 {L3_CLASS=PASS}           => {S2_CLASS=PASS} 0.2725509  0.6419780 1.361079
5 {L3_CLASS=PASS}           => {L2_CLASS=PASS} 0.2900410  0.6831749 1.442489
6 {S2_CLASS=PASS}           => {L2_CLASS=PASS} 0.3167923  0.6716418 1.418137
```

Conclusions:

- **Rule 1,5 were not found in the decision tree**
- **Rule 2, 3, 6 were found in the tree.**
- **Rule 4 was found but in reversed order**

Experiment 8

Objective: Cross-cluster analysis

Procedure:

- We made a data frame of L1, L2, L3, S1, S2, S3, TotalMarks.
- Replaced all marks by its z-scores.
- Applied K-means algorithm with k = 5.
- Factored each attribute after gaining associating cluster no with each row.
- Apply aprori() to get association rules.
- Remove redundant rules.
- Prune them with cluster number.
- Make a data frame of L1, L2, L3, S1, S2, S3, TotalMarks having NRC_GENDER_CODE = B and do steps 1 to 7.
- Make a data frame of L1, L2, L3, S1, S2, S3, TotalMarks having NRC_GENDER_CODE = G and do steps 1 to 7.

Results Obtained:

For entire data set cluster

-
  ```
  entireDataCluster$size
  [1] 4605 4107 8211 8543 6495
  ```

-
  ```
  entireDataCluster$centers
            L1         L2          L3           S1           S2         S3 TOTAL_MARKS
  1 -1.4466198 -1.1418156 -1.17649974 -1.324220991 -1.348630651 -1.3589479 -1.49748642
  2  1.2467525  1.5916025  1.50098869  1.539574658  1.644385468  1.3635032  1.67854625
  3  0.2112958 -0.1129855 -0.04566548 -0.004001783  0.008531986  0.1455746  0.05067729
  4 -0.5997211 -0.5727148 -0.60881696 -0.541766560 -0.519214688 -0.5954637 -0.66224588
  5  0.7590053  0.6992727  0.74354169  0.683013842  0.588537024  0.7005052  0.80732730
  ```

-
  ```
  inspect(myrulesPrun)
  lhs                           rhs                support confidence     lift
  ```

```
1 {L3=-0.961163910927722} => {clusters=4} 0.05281437  0.7465723 2.793070
2 {S2=-0.554964416287521} => {clusters=4} 0.05221989  0.6100146 2.282182
3 {S2=-0.832055622478925} => {clusters=4} 0.06604925  0.6985440 2.613387
4 {L2=-0.736942039857381} => {clusters=4} 0.09001596  0.6386238 2.389214
5 {L2=-0.522653035016499} => {clusters=4} 0.06667501  0.4713559 1.763433
```

For boys data set cluster

-
  ```
  boysDataCluster$size
  [1] 3315 4293 4640 2023 2584
  ```

-
  ```
  boysDataCluster$centers
  L1          L2          L3          S1          S2          S3        TOTAL_MARKS
  0.8001718  0.70545098  0.73792629  0.69585734  0.63665804  0.7677068  0.83755568
  0.1981790 -0.07189008 -0.04642886  0.04526838  0.04620043  0.1628383  0.07349977
  -0.5873311 -0.53025936 -0.55206907 -0.51908482 -0.48437790 -0.5804761 -0.62626324
  1.3364514  1.65657664  1.60301422  1.59847567  1.68048915  1.4087958  1.75159902
  -1.3474369 -1.13033941 -1.13320583 -1.28725502 -1.33938697 -1.3160207 -1.44336492
  ```

-
  ```
  lhs                           rhs             support confidence     lift
  1 {S2=-0.732777268788576} => {clusters=3} 0.06816968  0.6963636 2.529571
  2 {S2=-0.456725091936009} => {clusters=3} 0.05588846  0.6132812 2.227771
  3 {L2=-0.664395029737087} => {clusters=3} 0.08062889  0.5862813 2.129692
  4 {L3=-0.616232824097381} => {clusters=3} 0.09652922  0.5353735 1.944767
  5 {L2=-0.445859387144984} => {clusters=3} 0.08484129  0.5312036 1.929620
  ```

For girls data set cluster

-
  ```
  girlsDataCluster$size
  [1] 3984 1907 3250 3992 1973
  ```

-
  ```
  girlsDataCluster$centers
            L1         L2          L3          S1          S2         S3 TOTAL_MARKS
  1  0.2029860 -0.1397034 -0.03670772 -0.04302937 -0.05879902  0.1131183  0.01989999
  2 -1.5889062 -1.1787867 -1.24067717 -1.36395830 -1.35516868 -1.4219704 -1.57906025
  3  0.7253572  0.6981804  0.75051688  0.67258259  0.57048108  0.6505855  0.79102388
  4 -0.6153050 -0.6267193 -0.67730960 -0.59963339 -0.57656137 -0.6224871 -0.71799159
  5  1.1759932  1.5394330  1.40742777  1.51056287  1.65541383  1.3338063  1.63577352
  ```

-
  ```
    lhs                           rhs             support confidence     lift
  1 {L3=-1.11122391213094}  => {clusters=4} 0.05428307  0.8241206 3.118529
  2 {S2=-0.954341797525885} => {clusters=4} 0.06394810  0.7040816 2.664293
  3 {L3=-0.89236987502739}  => {clusters=2} 0.05229710  0.4403567 3.488217
  4 {L3=-0.89236987502739}  => {clusters=4} 0.05395207  0.4542921 1.719072
  5 {L2=-0.610193535779823} => {clusters=4} 0.05183371  0.4281028 1.619970
  ```

Conclusions:

- **Data cluster centers have similar properties in entire data set, boys data set and girls data set.**
- **Cluster 4 in entire data set has resemblance with cluster 3 in male data set.**

Section 3: Additional activities carried out

Experiment 1

Objective: Apply SVM , apply PCA then SVM to compare Accuracy of PCA.

Procedure:

- We made a different marks like L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS, NRC_CLASS in the csv file using excel.
- Factor NRC_CLASS.
- In data preparation we kept 70%of the entire data as training data and rest 30% as test data.
- We built a model using SVM.
- Then that model was used to predict *NRC_CLASS* for test data.
- Compared with actual NRC_CLASS with predicted one.
- Then Apply PCA on the data set. Take the Projected value and apply SVM on it.
- Check Accuracy rate.

Results Obtained:

- Accuracy of SVM before PCA**.**

  ```
  FALSE        TRUE
  0.04195223 0.95804777
  ```

- Accuracy after taking 5 Principle components with 97.7% proportion of variance

  ```
       FALSE        TRUE
  0.04058001 0.95941999
  ```

Conclusions:

Before PCA the accuracy of SVM was 96.3%. After applying PCA and taking first five PC’s with 97.7% variance, the accuracy reduced to 95.94%. Hence we can reduce dimension from six to five with accuracy going down by 0.005% (Which is quite acceptable in this case)

Experiment 2

Objective: Analyzing those student who were absent in any of the examination.

Procedure:

- Cleaned data, took those data pint in which students were absent in at least one of the exams.

- Allocated class based for each subject marks.
- Generated association rules on entire data set pruned in terms of School type, Urban_Rural, Gender_code.
- Generated association rules specific to Boys and Girls

Results Obtained:

While generating rules on entire data set

- Pruning based on School type

```
  lhs                  rhs                           support confidence      lift
1 {SCHOOL_TYPE=G} => {NRC_MEDIUM=K}               0.3886114  0.8881279 1.0723956
2 {SCHOOL_TYPE=G} => {NRC_PHYSICAL_CONDITION=N} 0.4375624  1.0000000 1.0090726
3 {SCHOOL_TYPE=G} => {L2_CLASS=A}                 0.3266733  0.7465753 0.9898304
4 {SCHOOL_TYPE=G} => {S1_CLASS=A}                 0.3476523  0.7945205 0.9770455
```

- Pruning based on URBAN_RURAL

```
  lhs                  rhs                   support confidence     lift
1 {URBAN_RURAL=R} => {NRC_MEDIUM=K}       0.5024975  0.9212454 1.112384
2 {URBAN_RURAL=R} => {L1_CLASS=A}         0.3586414  0.6575092 1.020413
3 {URBAN_RURAL=R} => {S1_CLASS=A}         0.4525475  0.8296703 1.020270
4 {URBAN_RURAL=U} => {NRC_GENDER_CODE=B} 0.3166833  0.6967033 1.015138
5 {URBAN_RURAL=R} => {L2_CLASS=A}         0.4165834  0.7637363 1.012583
```

- Among the boys

```
  lhs                    rhs                           support confidence     lift
1 {NRC_GENDER_CODE=B} => {S2_CLASS=FAIL}             0.5834166  0.8500728 1.015421
2 {NRC_GENDER_CODE=B} => {NRC_MEDIUM=K}              0.5734266  0.8355167 1.008869
3 {NRC_GENDER_CODE=B} => {S1_CLASS=A}                0.5594406  0.8151383 1.002400
4 {NRC_GENDER_CODE=B} => {NRC_PHYSICAL_CONDITION=N} 0.6813187  0.9927220 1.001729
```

- Based on medium

```
   lhs                             rhs            support confidence      lift
1  {URBAN_RURAL=R}              => {NRC_MEDIUM=K} 0.5024975  0.9212454 1.1123844
2  {L2_CLASS=A}                 => {NRC_MEDIUM=K} 0.6393606  0.8476821 1.0235583
3  {S3_CLASS=A}                 => {NRC_MEDIUM=K} 0.6093906  0.8367627 1.0103733
4  {S1_CLASS=A}                 => {NRC_MEDIUM=K} 0.6803197  0.8366093 1.0101881
5  {L3_CLASS=A}                 => {NRC_MEDIUM=K} 0.5404595  0.8361669 1.0096539
6  {NRC_GENDER_CODE=B}          => {NRC_MEDIUM=K} 0.5734266  0.8355167 1.0088688
7  {S2_CLASS=FAIL}              => {NRC_MEDIUM=K} 0.6993007  0.8353222 1.0086339
8  {NRC_Class=FAIL}             => {NRC_MEDIUM=K} 0.8281718  0.8281718 1.0000000
9  {NRC_PHYSICAL_CONDITION=N} => {NRC_MEDIUM=K} 0.8201798  0.8276210 0.9993348
10 {L1_CLASS=A}                 => {NRC_MEDIUM=K} 0.5314685  0.8248062 0.9959361
```

- Without Pruning on attributes

```
lhs                      rhs                            support confidence     lift
1  {L3_CLASS=A}          => {S2_CLASS=FAIL}             0.6203796  0.9598145 1.146509
2  {S3_CLASS=A}          => {L2_CLASS=A}                0.6193806  0.8504801 1.127590
3  {NRC_MEDIUM=K}        => {S2_CLASS=FAIL}             0.6993007  0.8443908 1.008634
```

In girls data set

```
•    lhs                  rhs                support confidence     lift
•  1 {L1_CLASS=A}    => {L3_CLASS=A}     0.5445860  0.8507463 1.284300
•  2 {L3_CLASS=A}    => {S3_CLASS=A}     0.5955414  0.8990385 1.232743
•  3 {L1_CLASS=A}    => {S3_CLASS=A}     0.5541401  0.8656716 1.186991
•  4 {L1_CLASS=A}    => {S2_CLASS=FAIL} 0.6019108  0.9402985 1.162416
•  5 {L3_CLASS=A}    => {L2_CLASS=A}     0.6146497  0.9278846 1.156174
```

In boys data set

```
•    lhs              rhs              support confidence     lift
•  1 {L3_CLASS=A} => {S3_CLASS=A}   0.5807860  0.9088838 1.248806
•  2 {L3_CLASS=A} => {L2_CLASS=A}   0.5764192  0.9020501 1.232025
```

```
3  {L3_CLASS=A} => {S2_CLASS=FAIL} 0.6215429  0.9726651 1.144214
4  {S3_CLASS=A} => {L2_CLASS=A}    0.6069869  0.8340000 1.139082
5  {L1_CLASS=A} => {L2_CLASS=A}    0.5342067  0.8265766 1.128943
```

Conclusions:

- No rules based on District among the absentees.
- Absentees who are from government schools belong to kannada mediun(This rule was proved earlier for non-absentee data)
- Absentees who are from government schools were absent in L2 and S1.
- Absentees who are from Urban area are boys (s = 0.3 , c = 0.7)
- Absentees who are Boys have Failed in S2 (s = 0.58 , c = 0.85)
- Absentees who were Absent in L3 have failed in S2
- Girls Absentees who were absent in L1 were either absent in L3 or failed in S2.
- Boys Absentees who were absent in L3 were either absent in L3 or L2 or failed in S2.

---The End---