GENERAL SIR JOHN KOTELAWALA DEFENCE UNIVERSITY

BACHELOR OF SCIENCE IN INFORMATION TECHNOLOGY

PROJECT DOCUMENT

DATA MINING AND DATA WAREHOUSING (IT7013)

Project Details	
Student No:	**Student Name**
O/70239	**Capt. APL Madushanka**
D/IT/18/0004	**UDATP Bandara**
D/IT/18/0005	**KPMI Ramanayake**
D/IT/18/0043	**MCK Bandara**
D/IT/18/0068	**KN Jayasinghe**
Project Name	**Data mining and data warehousing mini project on student performance**
Commencement Date	**26-01-2021**
Submission Date	**21-03-2021**
Supervisor Name	**Mr. Dinesh Asanka**

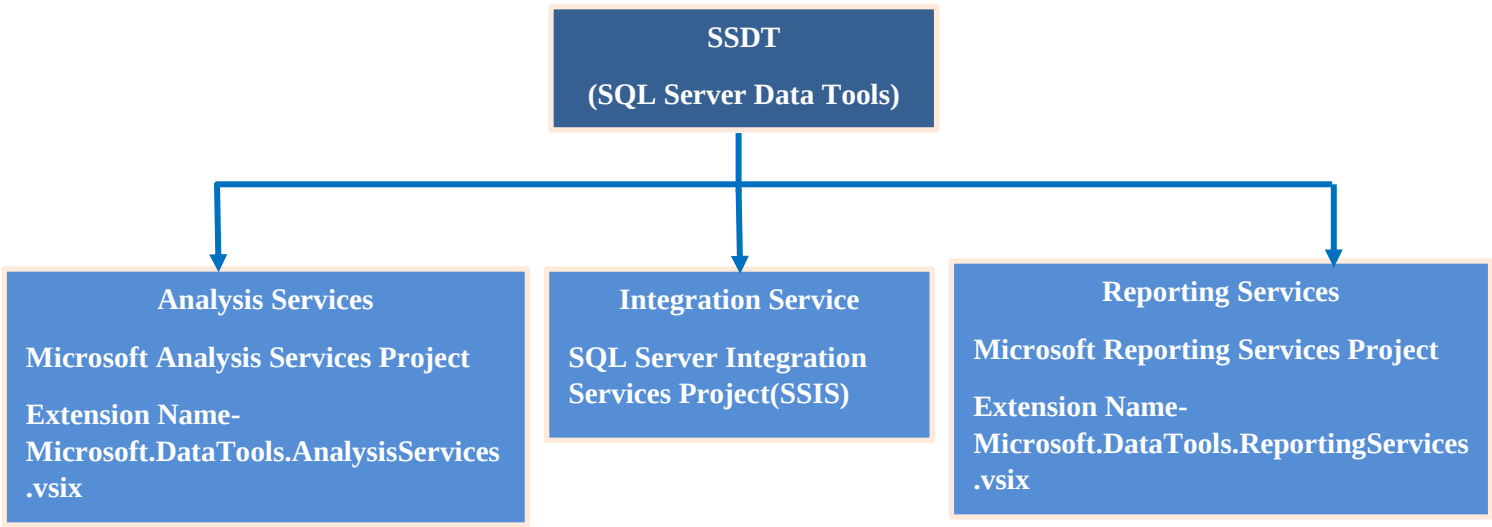
Contents

1. Introduction of the Mini project

Data warehouse can be defined as a databases collection. These databases consist of transformed cleaned data and they are subjective to improve decision making. Main objective of student performance project is to have an analysis on performance of the student depending on the gender, ethnicity, parental educational level, lunch and preparation for the examination. With this mini project, data warehouse is integrated with the information effectively. These information are stored in an integrated repository. This will be helpful in analysis.

2. Technologies Used

1. Microsoft SQL Server 2019



2. Microsoft Visual Studio 2019
3. Weka

3. Data Resource

https://www.kaggle.com/roshansharma/student-performance-analysis

4. About Data set

This dataset consists of the marks obtained by high school students of United States for math test, reading test and writing test. Gender, Race/Ethnicity, Parental level of education, lunch and preparation for the test are included in the dataset. The owner of the dataset is Mr. Roshan Sharma and it was published in 2019.

5. V's in our dataset

There are main 5 Vs in big data concept. They are,

1. Volume
2. Variety
3. Veracity
4. Velocity
5. Value

Volume

Volume is the size of our data set in our student performance data set is 1000 records.

Variety

Variety can be defined as different types of data. We can have 04 types of varieties such as structural variety, semantic variety, media variety and availability variety.

We have a structured data set. This dataset is in text form and presented as an Excel sheet. When think about the availability, it is intermittent.

Veracity

Veracity can be defined as the accuracy of the data. Here we cannot guarantee that this data set in 100% accurate. Because inaccuracy comes in two ways as either at the time of recording data and at time of recording data can be correct, but modifications are not recorded. So, in this instance we can't find whether inaccuracy has happened or not.

Velocity

Velocity is the rate of change. This means velocity is all about rate of change of data or how rapid the data is changed. Velocity has 03 aspects as speed of creating data, speed of storing data, and speed of analyzing data. In our case data do not change rapidly in all three aspects as in supermarkets. Because our data set consists of marks of the students. So, they won't change.

Value

Value can be our output. If we can manage above mentioned 04 Vs then we can get value out of our data. We can have a value of our dataset. There is a worthiness for our data set as we can have an analysis and predictions using it.

6. Types of questions that we are going to answer.

With a data set we can have types of analysis. That means we can find solutions for 04 types of problems such as Descriptive, Diagnostic, Predictive and Prescriptive.

Descriptive Analysis

By using our data set we are going to analyze the performance of the students. That means they have already answered for examinations and depending on the results we are analyzing the performance. This is what we are looking statistically about what has happened in the past. Therefore, we are doing a Descriptive Analysis.

Predictive Analysis.

We also can have a prediction on performance of the student in year examination by considering patterns and trends with the use of machine learning. These analytics can be known as Predictive Analysis.

Diagnostic Analysis.

As the next step we can take these data to find answers for the problems of why the performance of a particular student get lowered though their parents come from good educational background. Why female students get more marks than male students. This kind of analysis are called Diagnostic Analysis. But in our mini project we are not finding solutions for diagnostic problems.

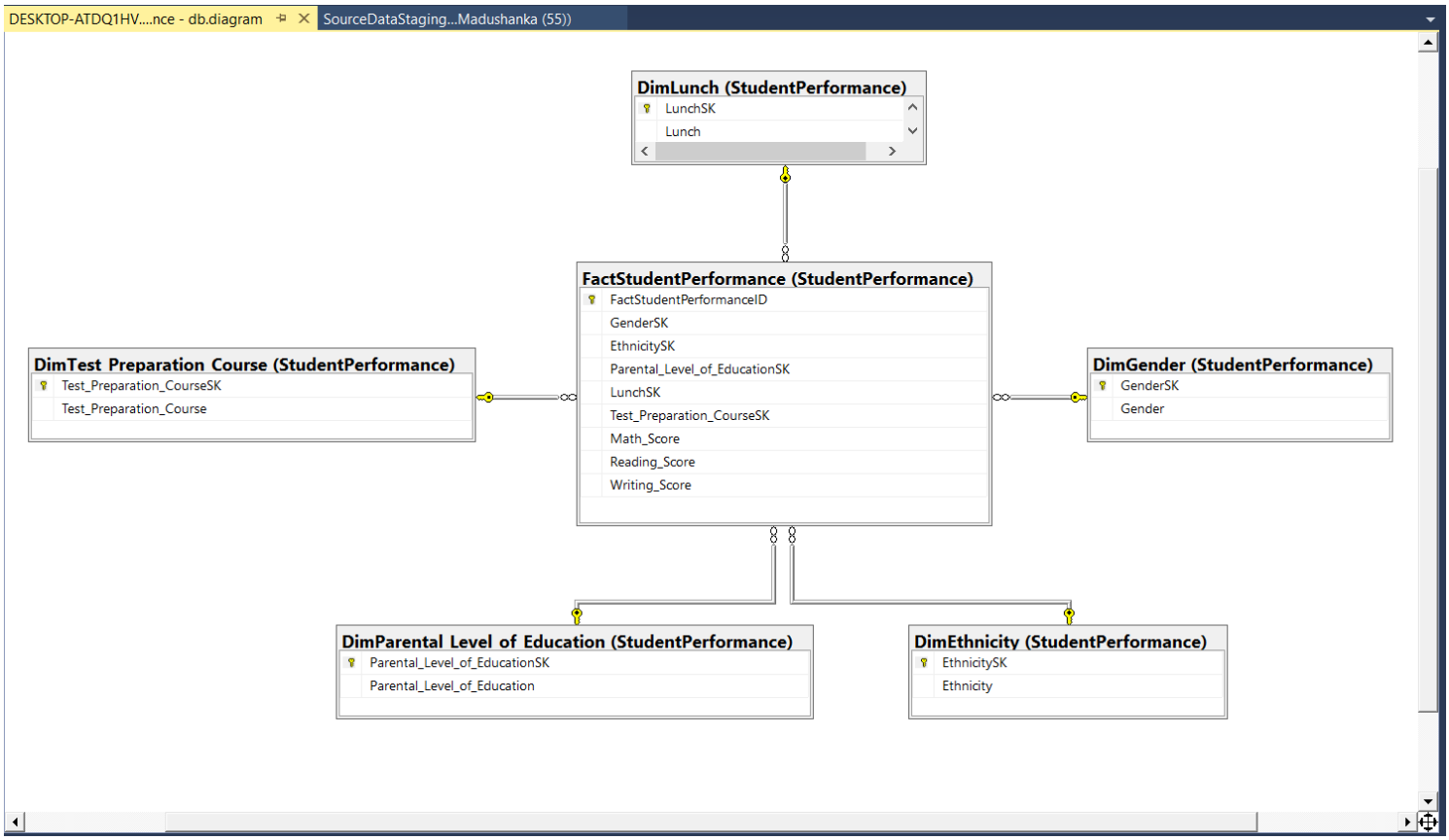
Prescriptive analytics.

Predictive data will be considered and gives hints for future. This analysis comes under Prescriptive analytics. But in our mini project we are not finding solutions for prescriptive problems.

<u>Problems</u>

- How gender affects on performance of the students?
- How parents' educational level effects on score of the student?
- Does prior preparation affect the score of the students?
- Does lunch affect the score of the student?

7. Data Warehouse design



8. ETL Design


Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Execution Resu...
ETL Data From StudentPerformance CSV to Staging
Load Gender Dimention
Update GenderID into Stagging Table
Load Ethnicity Dimension
Update EthnicityID into Stagging Table
Load Parental_Level_of_Education Dimention
Update Parental_Level_of_EducationID into Stagging Table
Load Lunch Dimention
Update LunchID into Stagging Table
Load Test_Preparation_Course Dimension
Update Test_Preparation_CourseID into Stagging Table
Load FactStudentPerformance Table

9. Generated Reports from the data warehouse

1. SQL Server

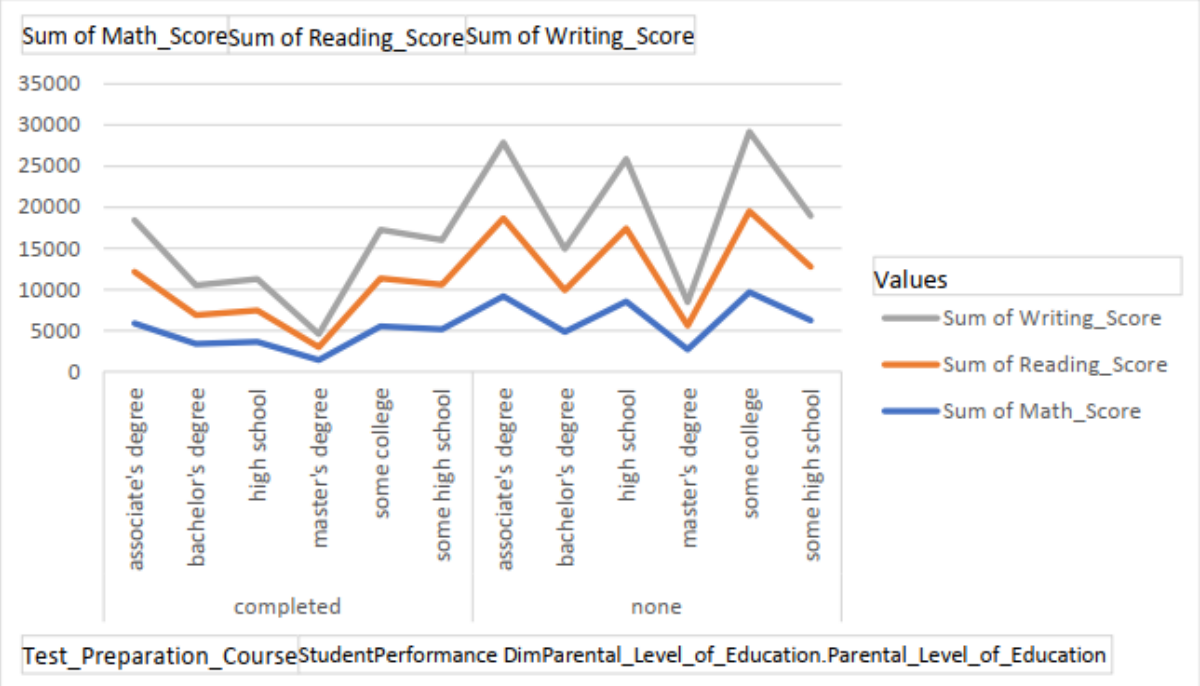
Results | Messages

	Ethnicity	Gender	Math_Score	Reading_Score	Writing_Score
1	group B	female	72	72	74
2	group C	female	69	90	88
3	group B	female	90	95	93
4	group A	male	47	57	44
5	group C	male	76	78	75
6	group B	female	71	83	78
7	group B	female	88	95	92
8	group B	male	40	43	39
9	group D	male	64	64	67
10	group B	female	38	60	50
11	group C	male	58	54	52
12	group D	male	40	52	43
13	group B	female	65	81	73
14	group A	male	78	72	70
15	group A	female	50	53	58
16	group C	female	69	75	78
17	group C	male	88	89	86
18	group B	female	18	32	28
19	group C	male	46	42	46
20	group C	female	54	58	61
21	group D	male	66	69	63
22	group B	female	65	75	70
23	group D	male	44	54	53
24	group C	female	69	73	73
25	group D	male	74	71	80
26	group A	male	73	74	72
27	group B	male	69	54	55
28	group C	female	67	69	75
29	group C	male	70	70	65
30	group D	female	62	70	75
31	group D	female	69	74	74

2. Excel Reports

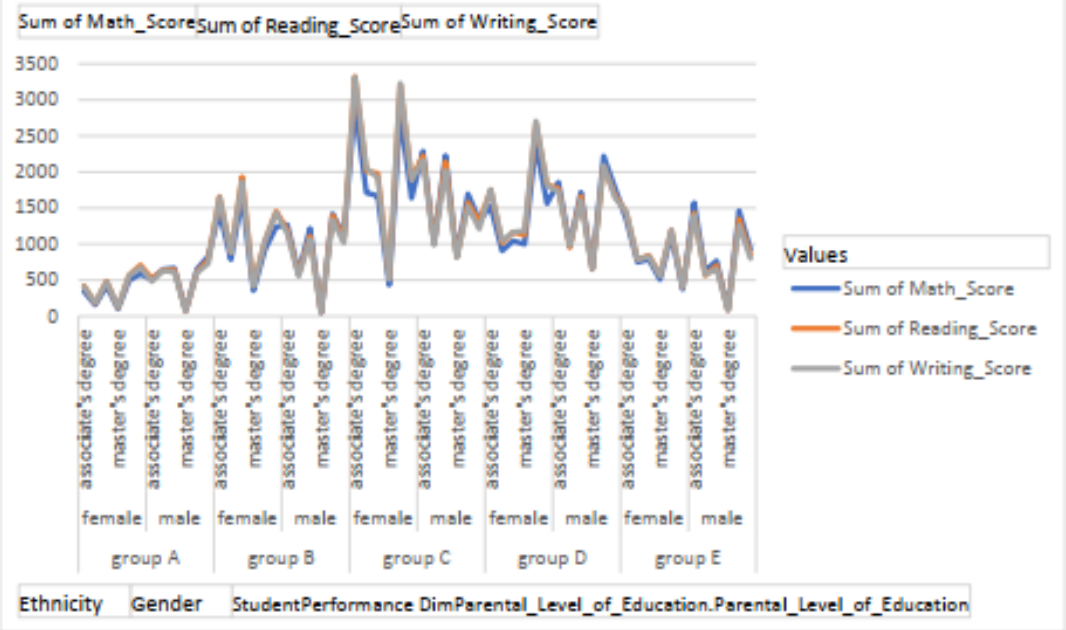
Educated Parents' Children Preparation Vs Scores

Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
completed	**24951**	**26454**	**26642**
associate's degree	5890	6246	6299
bachelor's degree	3371	3530	3620
high school	3640	3799	3811
master's degree	1412	1565	1602
some college	5502	5851	5892
some high school	5136	5463	5418
none	**41138**	**42715**	**41412**
associate's degree	9180	9500	9218
bachelor's degree	4817	5084	5039
high school	8539	8883	8429
master's degree	2703	2882	2863
some college	9669	9847	9666
some high school	6230	6519	6197
Grand Total	**66089**	**69169**	**68054**



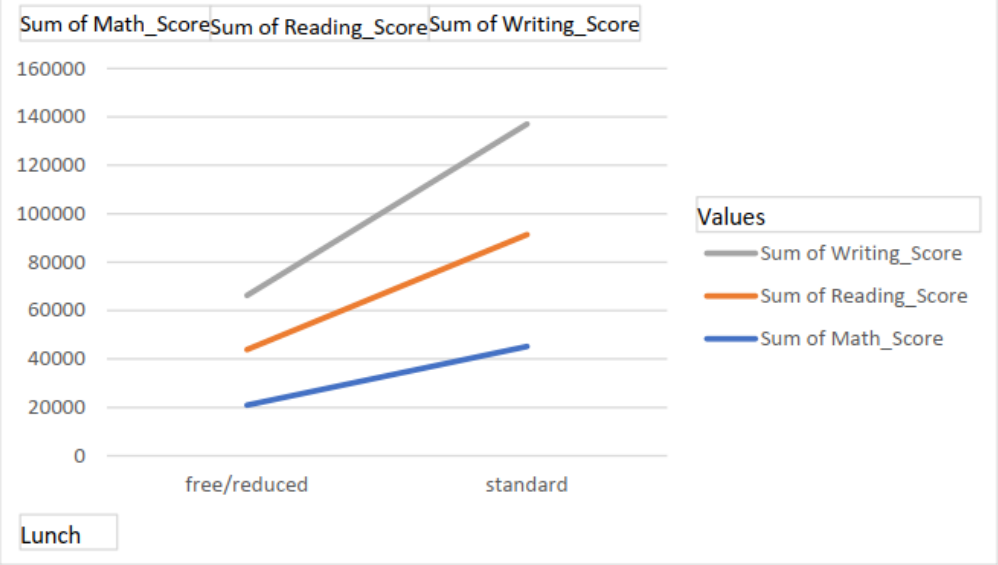
Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
group A	**5485**	**5756**	**5578**
female	**2107**	**2484**	**2443**
associate's degree	345	421	409
bachelor's degree	155	180	185
high school	423	489	477
master's degree	100	120	131
some college	497	566	565
some high school	587	708	676
male	**3378**	**3272**	**3135**
associate's degree	509	518	481
bachelor's degree	651	637	635
high school	665	643	612
master's degree	73	74	72
some college	653	618	605
some high school	827	782	730
group B	**12056**	**12797**	**12464**
female	**6386**	**7392**	**7285**
associate's degree	1441	1660	1649
bachelor's degree	785	880	879
high school	1652	1933	1871
master's degree	354	428	411
some college	917	1035	1039
some high school	1237	1456	1436
male	**5670**	**5405**	**5179**
associate's degree	1269	1193	1149
bachelor's degree	601	579	554
high school	1218	1113	1069
master's degree	49	53	52
some college	1421	1398	1336
some high school	1112	1069	1019
group C	**20564**	**22044**	**21637**
female	**11166**	**12950**	**12920**
associate's degree	2915	3324	3317
bachelor's degree	1711	2012	2051
high school	1669	1979	1921
master's degree	434	521	509
some college	2799	3213	3231
some high school	1638	1901	1891
male	**9398**	**9094**	**8717**
associate's degree	2290	2224	2164
bachelor's degree	1015	1015	985
high school	2229	2144	2025
master's degree	840	819	812
some college	1695	1577	1521
some high school	1329	1315	1210
group D	**17649**	**18348**	**18378**
female	**8417**	**9552**	**9678**
associate's degree	1521	1751	1760
bachelor's degree	907	1014	1043

high school	1045	1168	1169
master's degree	999	1124	1184
some college	2384	2671	2708
some high school	1561	1824	1814
male	**9232**	**8796**	**8700**
associate's degree	1859	1776	1733
bachelor's degree	985	950	970
high school	1721	1666	1610
master's degree	669	651	650
some college	2221	2078	2096
some high school	1777	1675	1641
group E	**10335**	**10224**	**9997**
female	**4886**	**5233**	**5212**
associate's degree	1347	1442	1449
bachelor's degree	748	783	780
high school	790	842	831
master's degree	507	572	560
some college	1121	1194	1195
some high school	373	400	397
male	**5449**	**4991**	**4785**
associate's degree	1574	1437	1406
bachelor's degree	630	564	577
high school	767	705	655
master's degree	90	85	84
some college	1463	1348	1262
some high school	925	852	801
Grand Total	**66089**	**69169**	**68054**



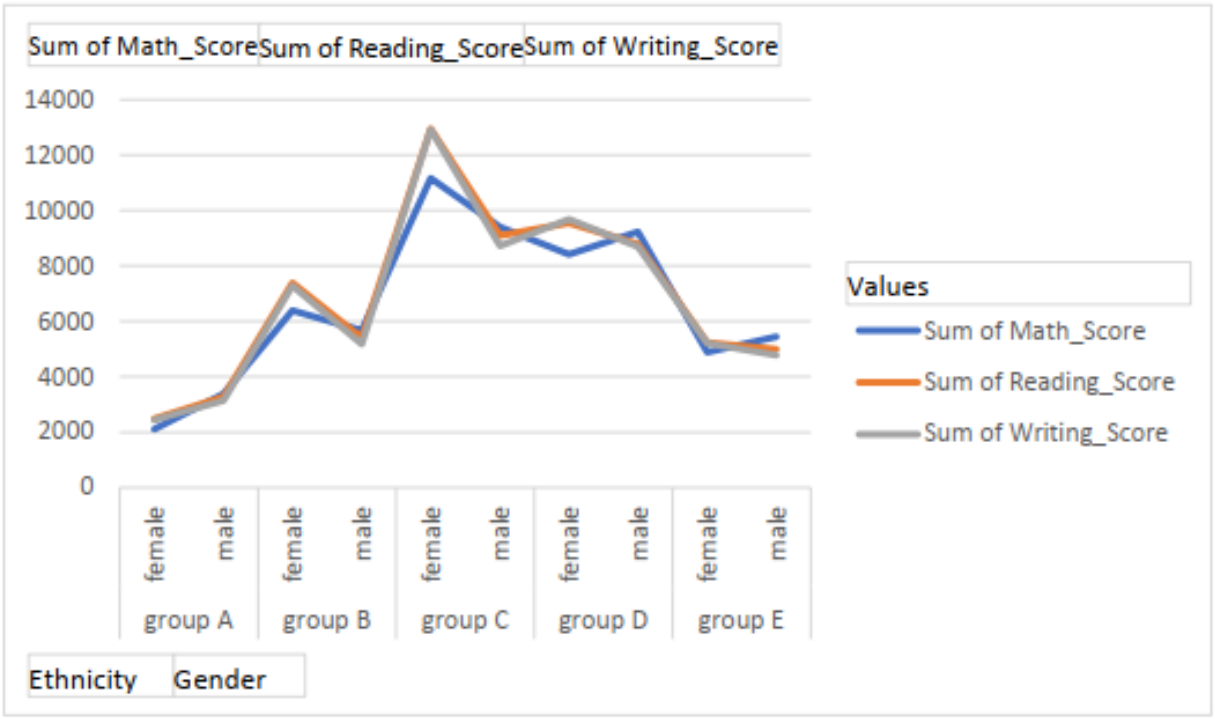
Luch vs Performance

Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
free/reduced	20917	22952	22373
standard	45172	46217	45681
Grand Total	**66089**	**69169**	**68054**



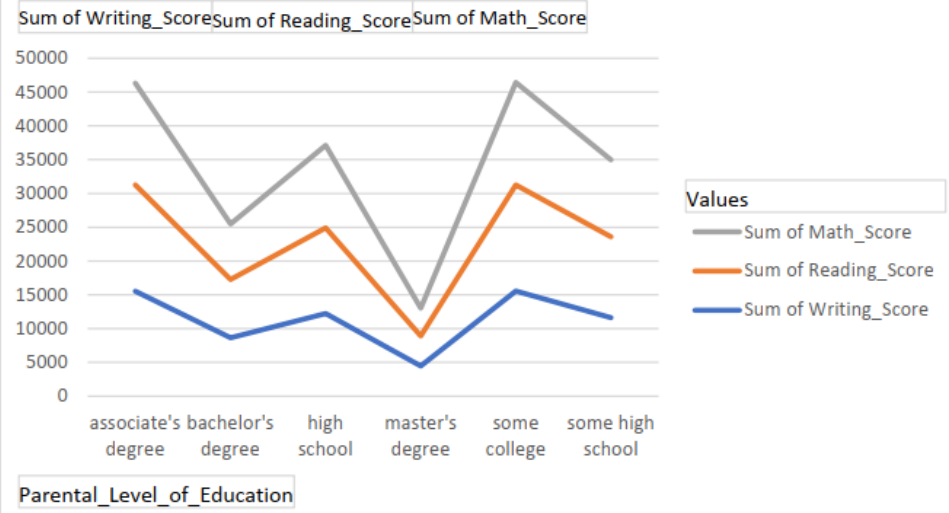
Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
female	32962	37611	37538
male	33127	31558	30516
Grand Total	**66089**	**69169**	**68054**

Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
group A	**5485**	**5756**	**5578**
female	2107	2484	2443
male	3378	3272	3135
group B	**12056**	**12797**	**12464**
female	6386	7392	7285
male	5670	5405	5179
group C	**20564**	**22044**	**21637**
female	11166	12950	12920
male	9398	9094	8717
group D	**17649**	**18348**	**18378**
female	8417	9552	9678
male	9232	8796	8700
group E	**10335**	**10224**	**9997**
female	4886	5233	5212
male	5449	4991	4785
Grand Total	**66089**	**69169**	**68054**



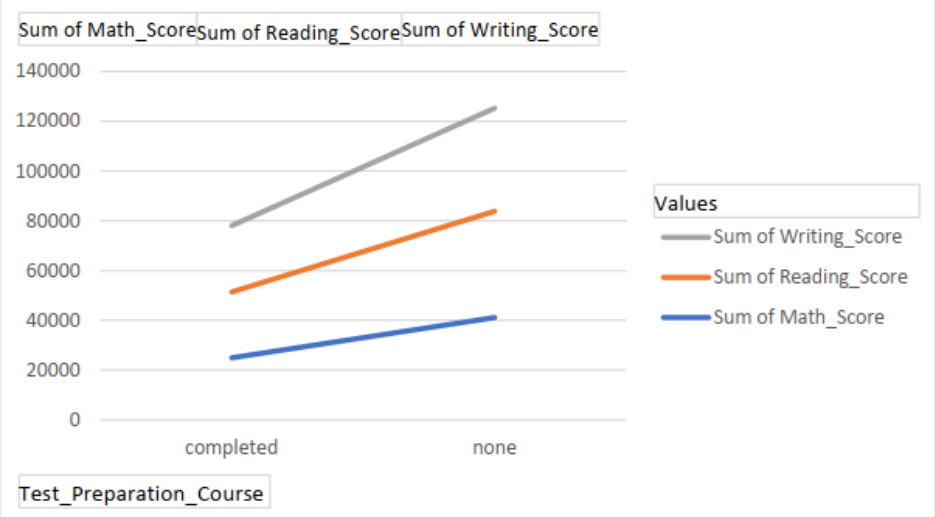
Parental Qualification Vs Childrens' Scores

Row Labels	Sum of Writing_Score	Sum of Reading_Score	Sum of Math_Score
associate's degree	15517	15746	15070
bachelor's degree	8659	8614	8188
high school	12240	12682	12179
master's degree	4465	4447	4115
some college	15558	15698	15171
some high school	11615	11982	11366
Grand Total	**68054**	**69169**	**66089**



Preparation Vs Score

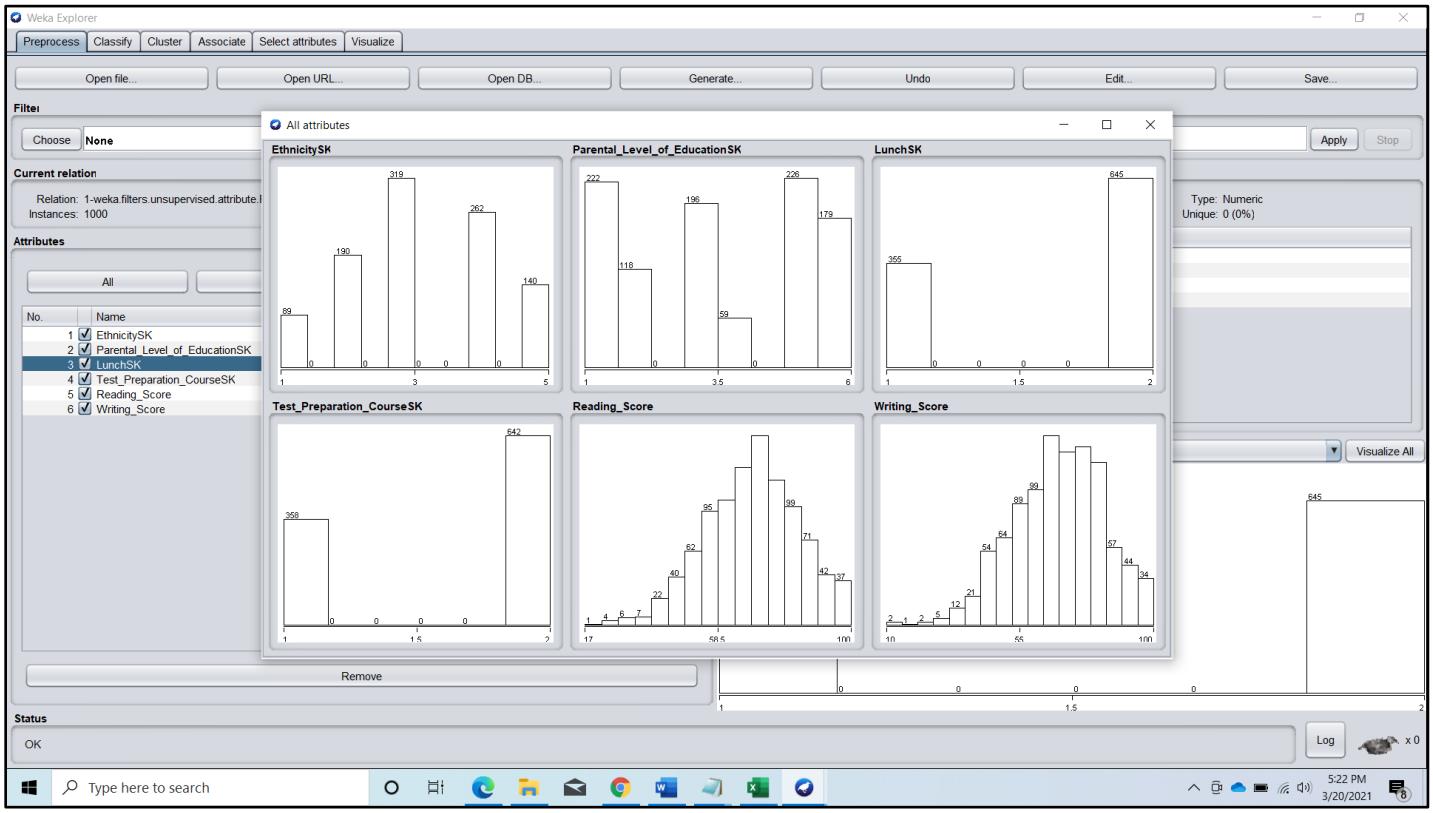
Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
completed	24951	26454	26642
none	41138	42715	41412
Grand Total	**66089**	**69169**	**68054**

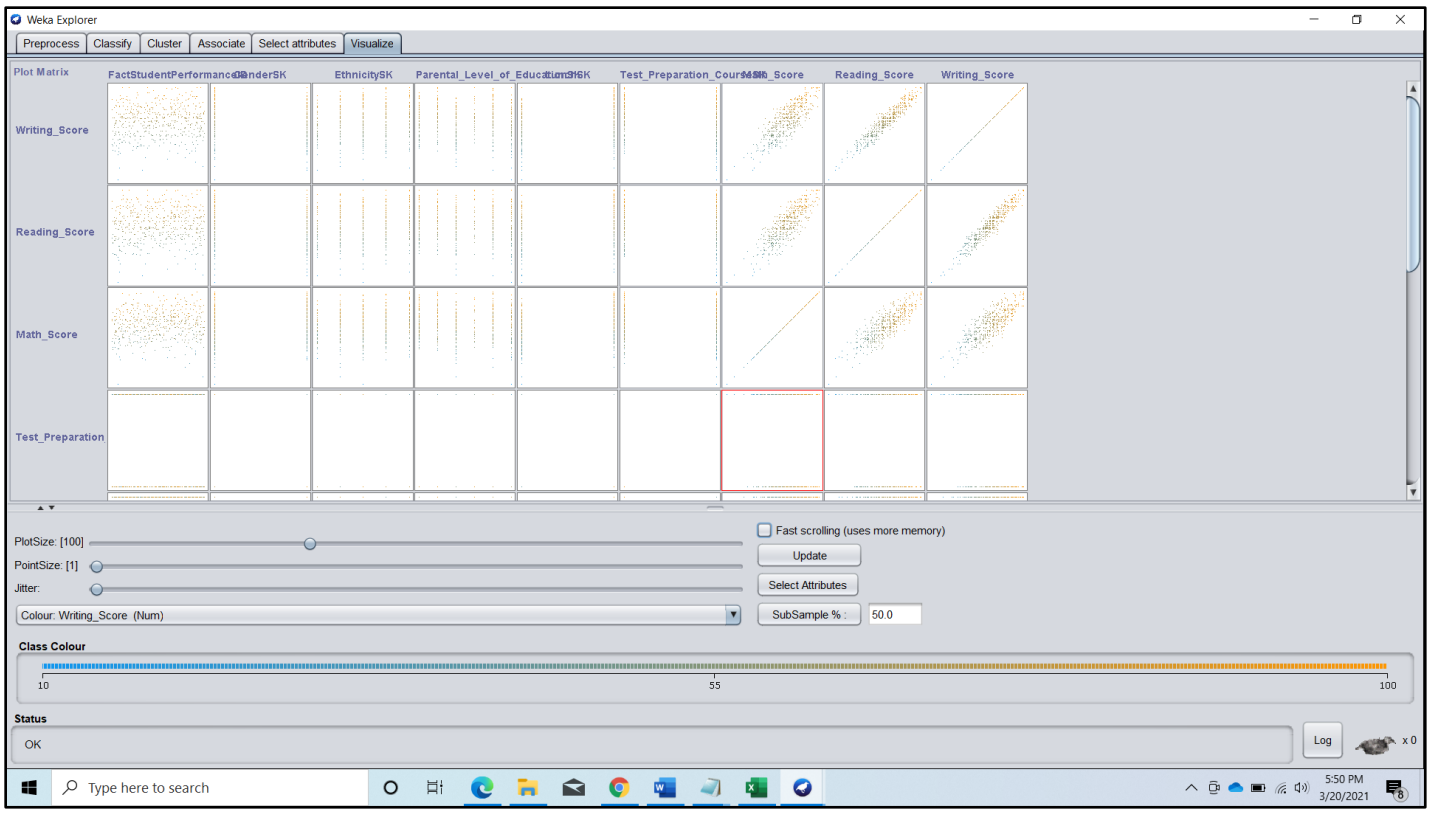


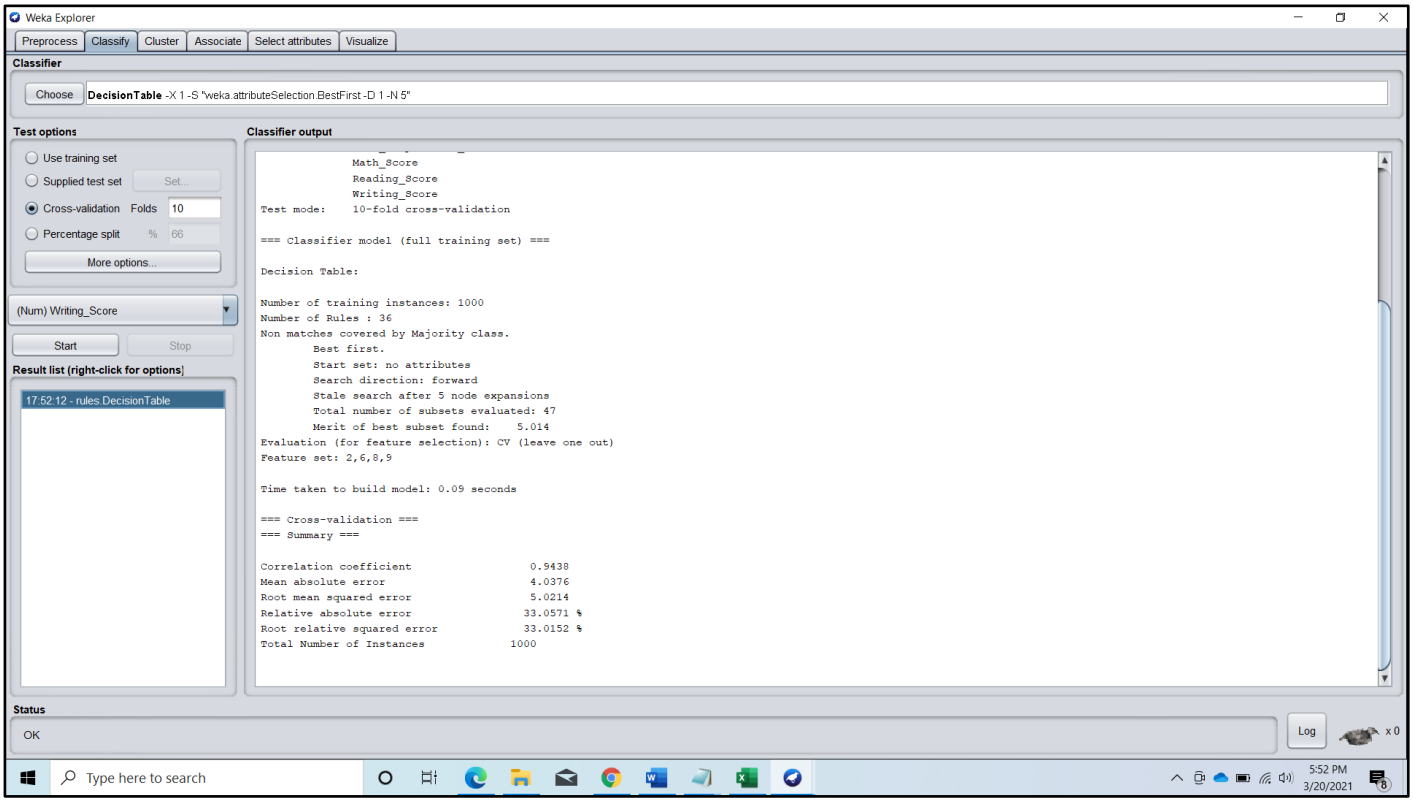
Scores Vs Gender

Row Labels	Sum of Math_Score	Sum of Reading_Score	Sum of Writing_Score
female	32962	37611	37538
male	33127	31558	30516
Grand Total	**66089**	**69169**	**68054**

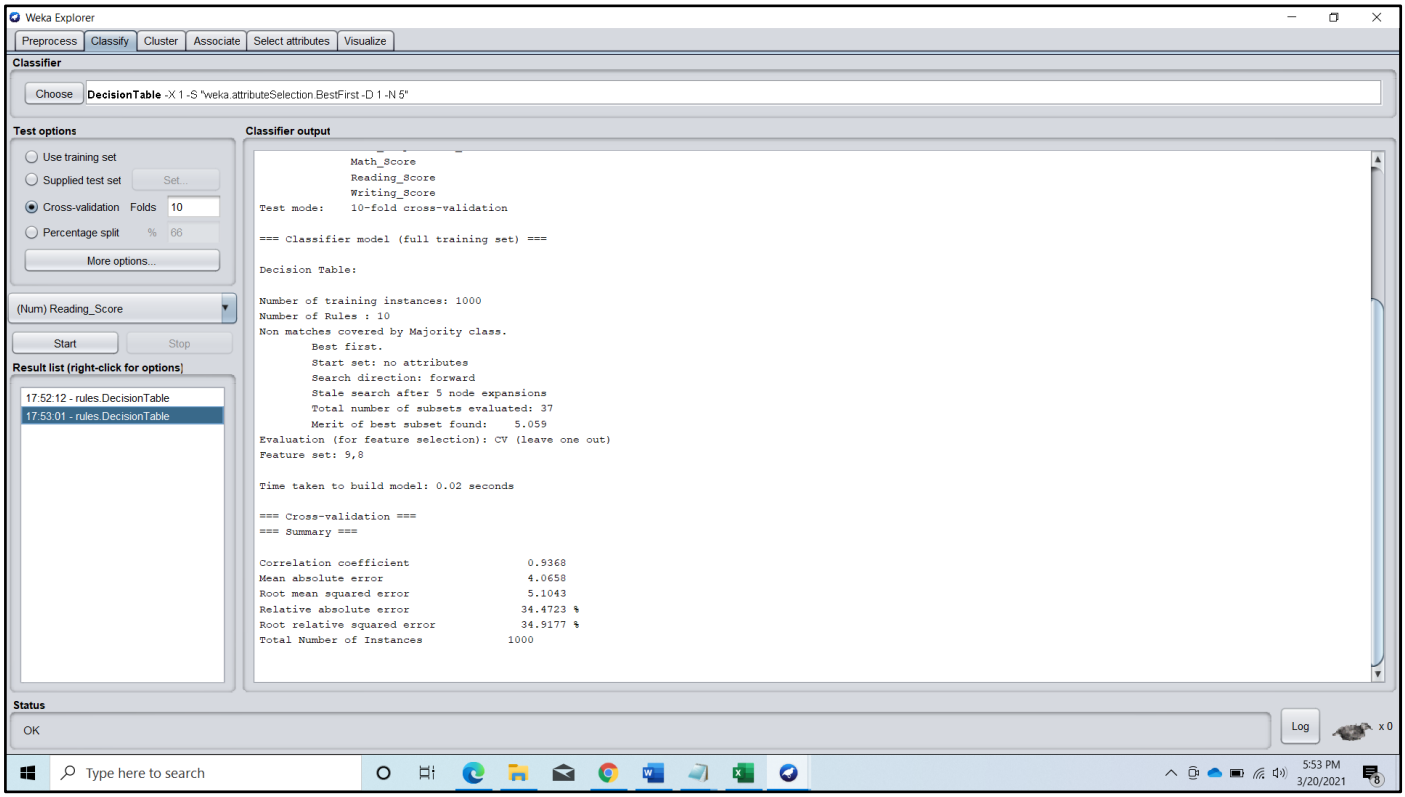
3. Weka

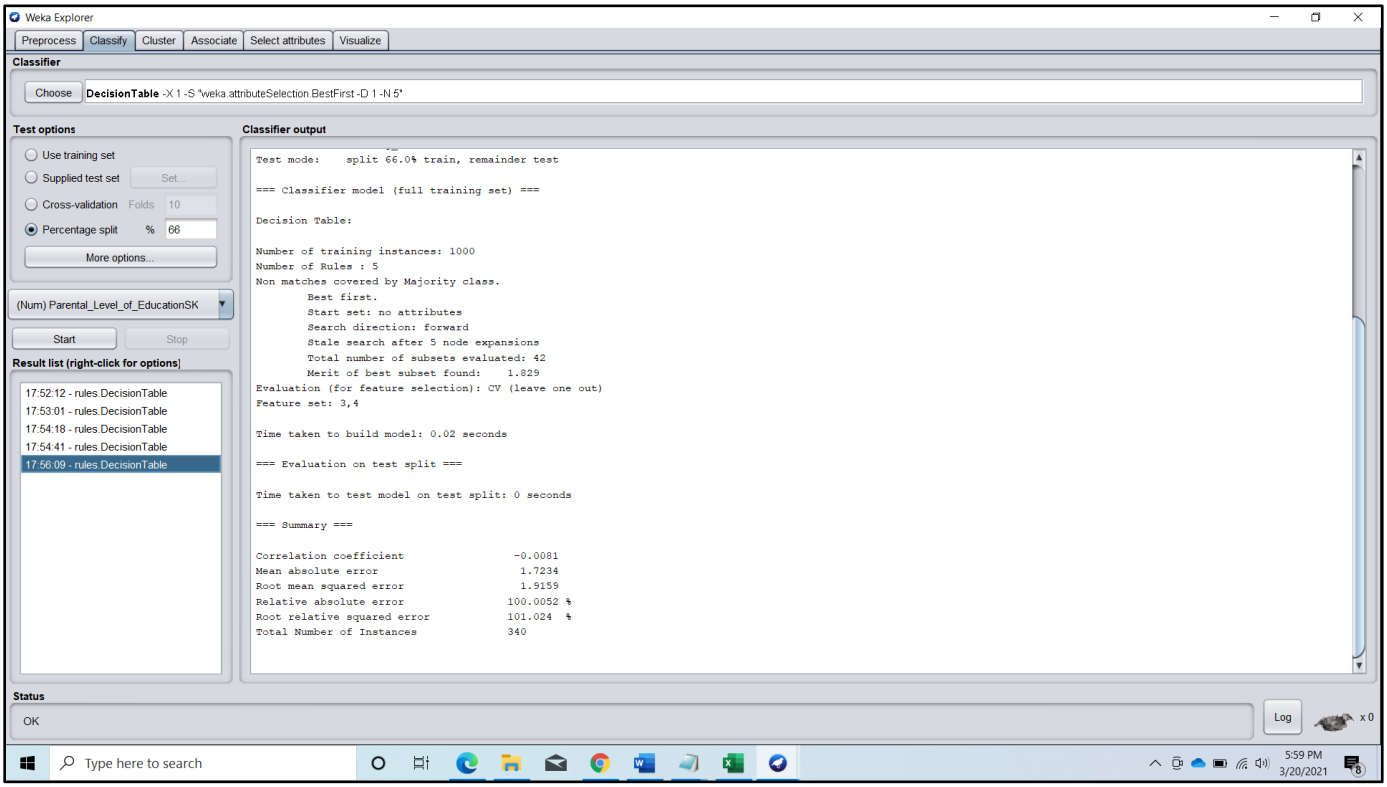



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
Writing_Score
Reading_Score
Math_Score
Test_Preparation
PlotSize: [100]
PointSize: [1]
Jitter:
Colour: Writing_Score (Num)
Fast scrolling (uses more memory)
Update
Select Attributes
SubSample %:
50.0
Class Colour
10
55
100
Status
OK
Log
x 0
Type here to search
5:50 PM
3/20/2021

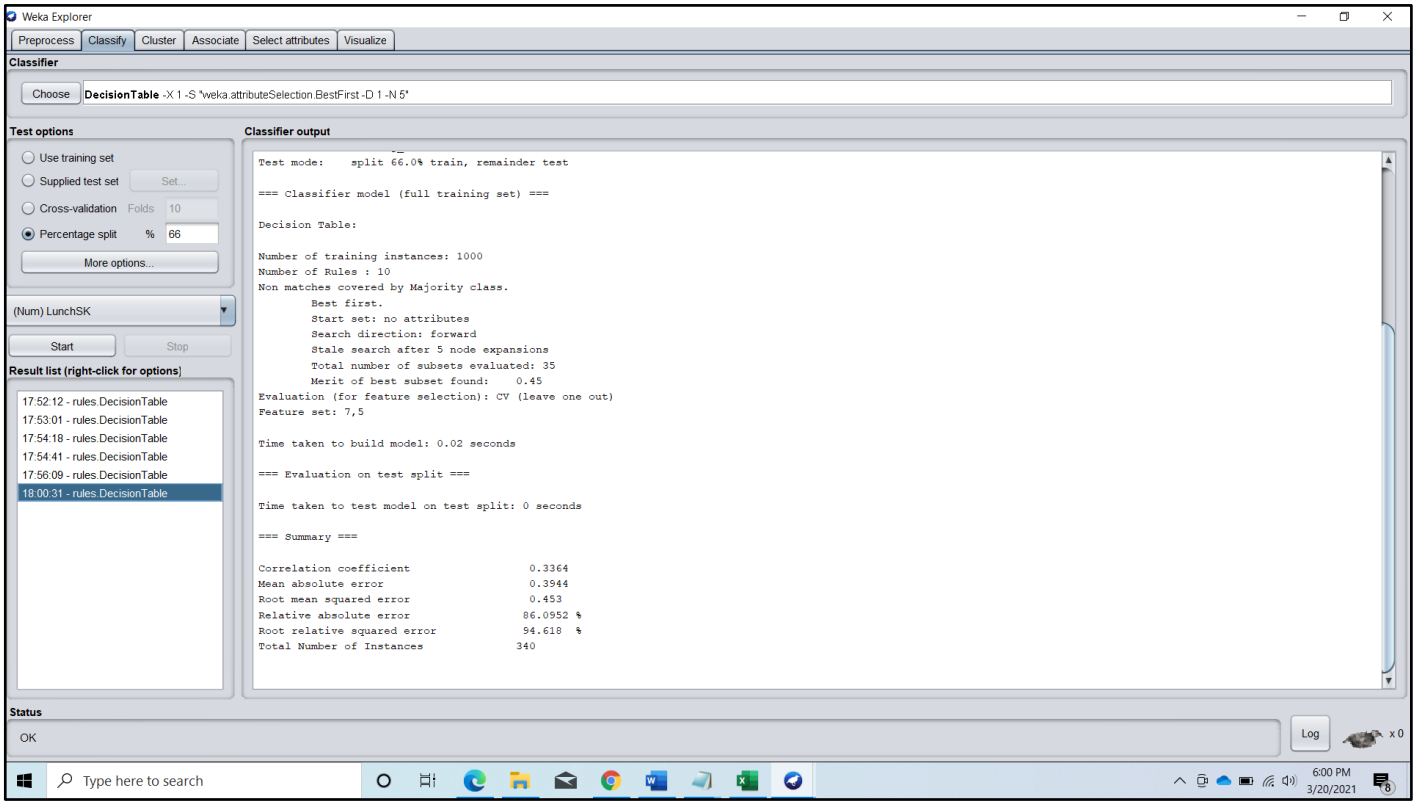

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Writing_Score
Start
Stop
Result list (right-click for options)
17:52:12 - rules.DecisionTable
Classifier output
Math_Score
Reading_Score
Writing_Score
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Decision Table:
Number of training instances: 1000
Number of Rules : 36
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 47
Merit of best subset found: 5.014
Evaluation (for feature selection): CV (leave one out)
Feature set: 2,6,8,9
Time taken to build model: 0.09 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.9438
Mean absolute error 4.0376
Root mean squared error 5.0214
Relative absolute error 33.0571 %
Root relative squared error 33.0152 %
Total Number of Instances 1000
Status
OK
Log
x 0
Type here to search
5:52 PM
3/20/2021

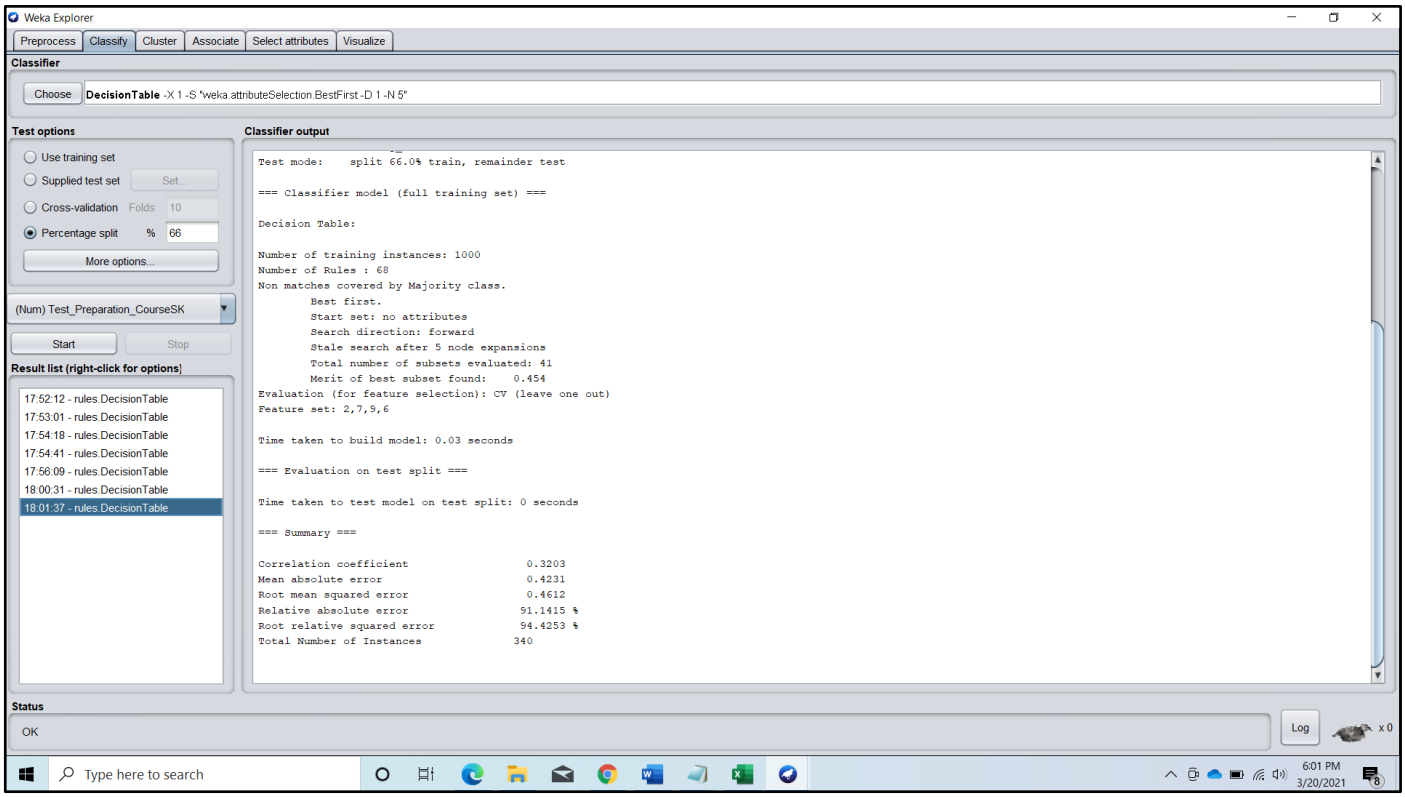
Decision Tree



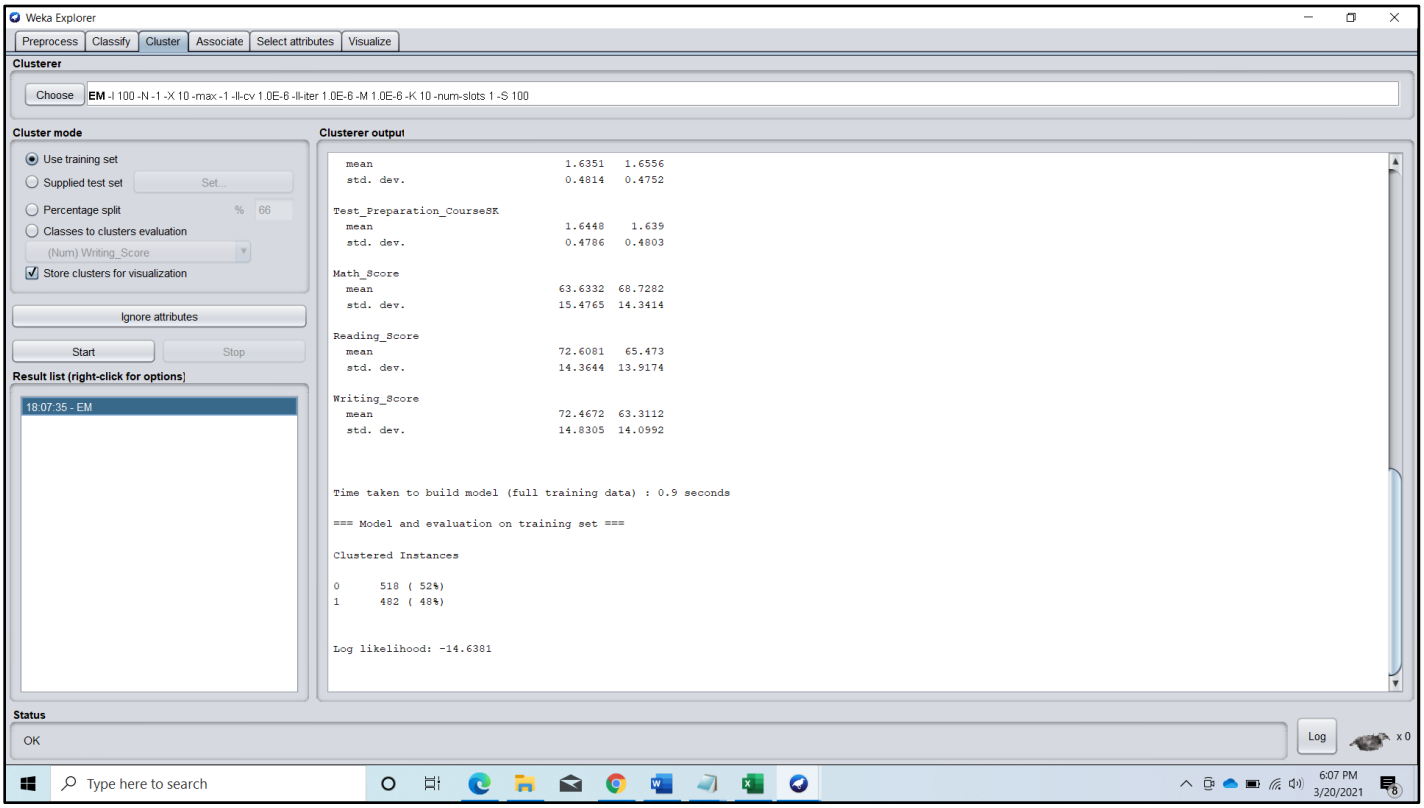


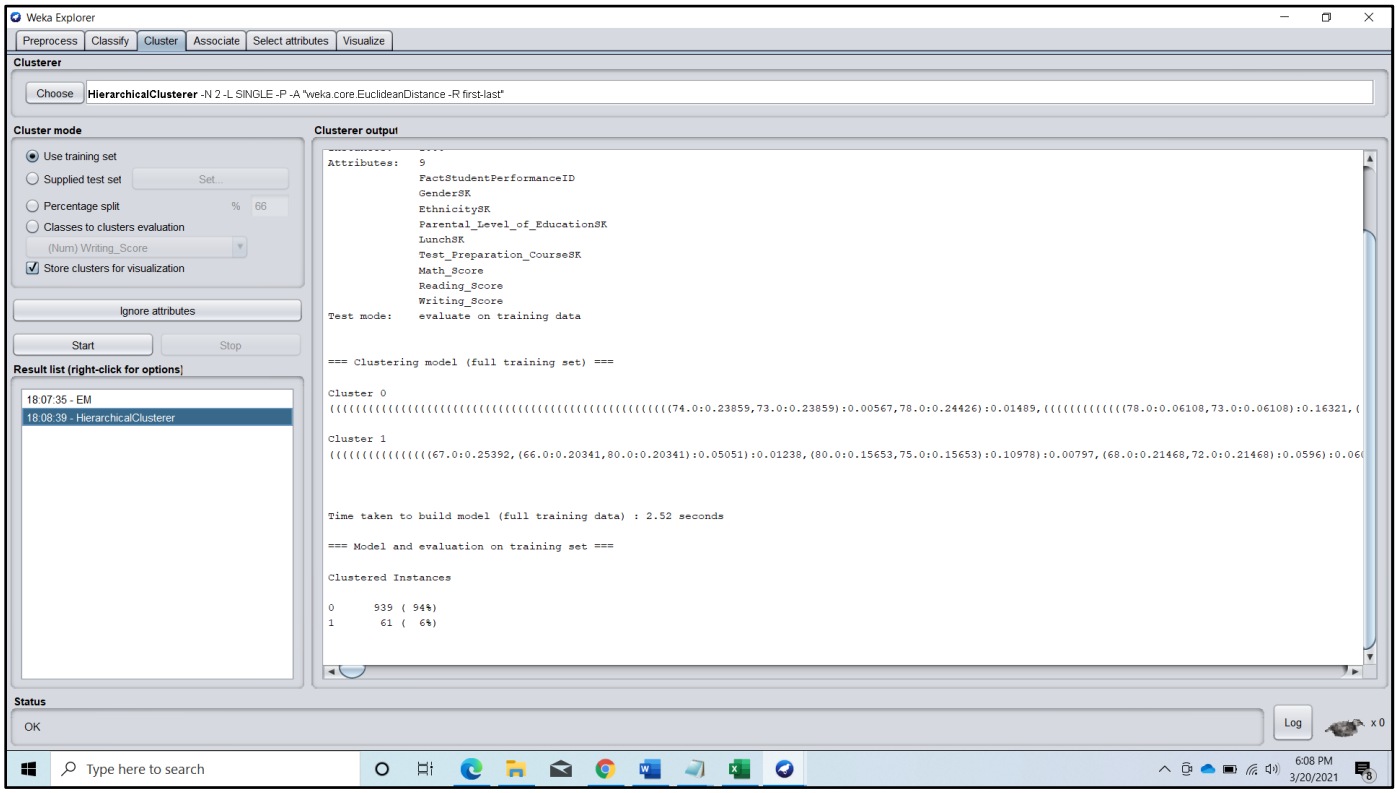
Percentage Split

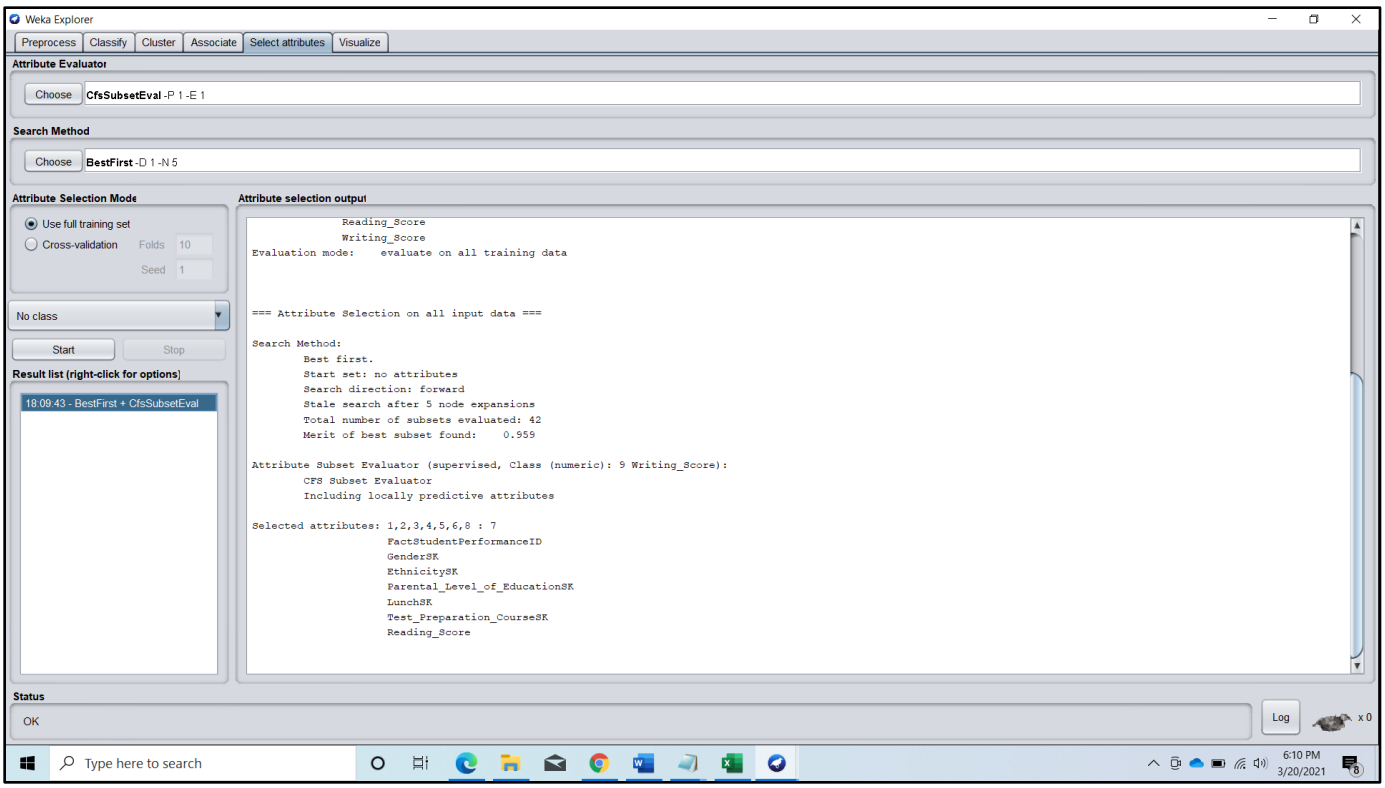




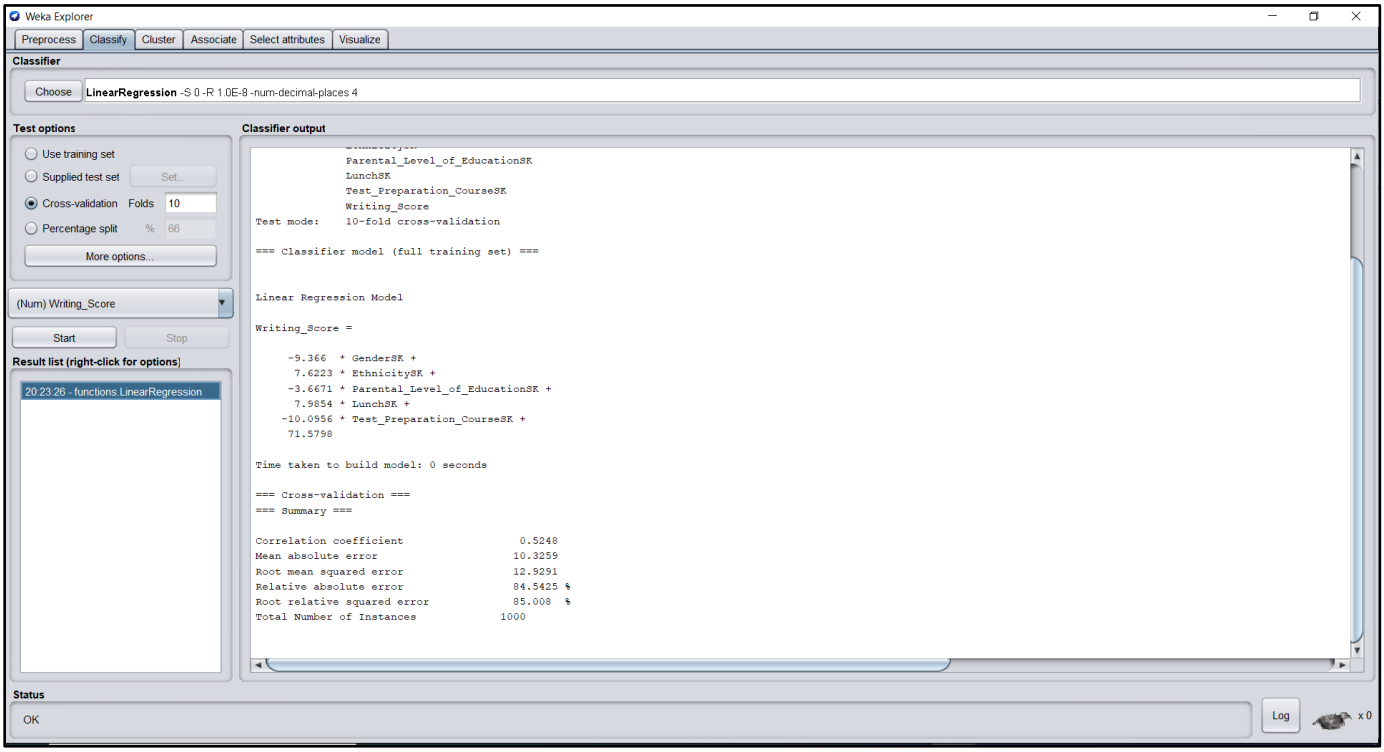
Cluster


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) Writing_Score
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
18:07:35 - EM
Clusterer output
mean 1.6351 1.6556
std. dev. 0.4814 0.4752
Test_Preparation_CourseSK
mean 1.6448 1.639
std. dev. 0.4786 0.4803
Math_Score
mean 63.6332 68.7282
std. dev. 15.4765 14.3414
Reading_Score
mean 72.6081 65.473
std. dev. 14.3644 13.9174
Writing_Score
mean 72.4672 63.3112
std. dev. 14.8305 14.0992
Time taken to build model (full training data) : 0.9 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 518 (52%)
1 482 (48%)
Log likelihood: -14.6381
Status
OK
Log
x 0
Type here to search
6:07 PM
3/20/2021

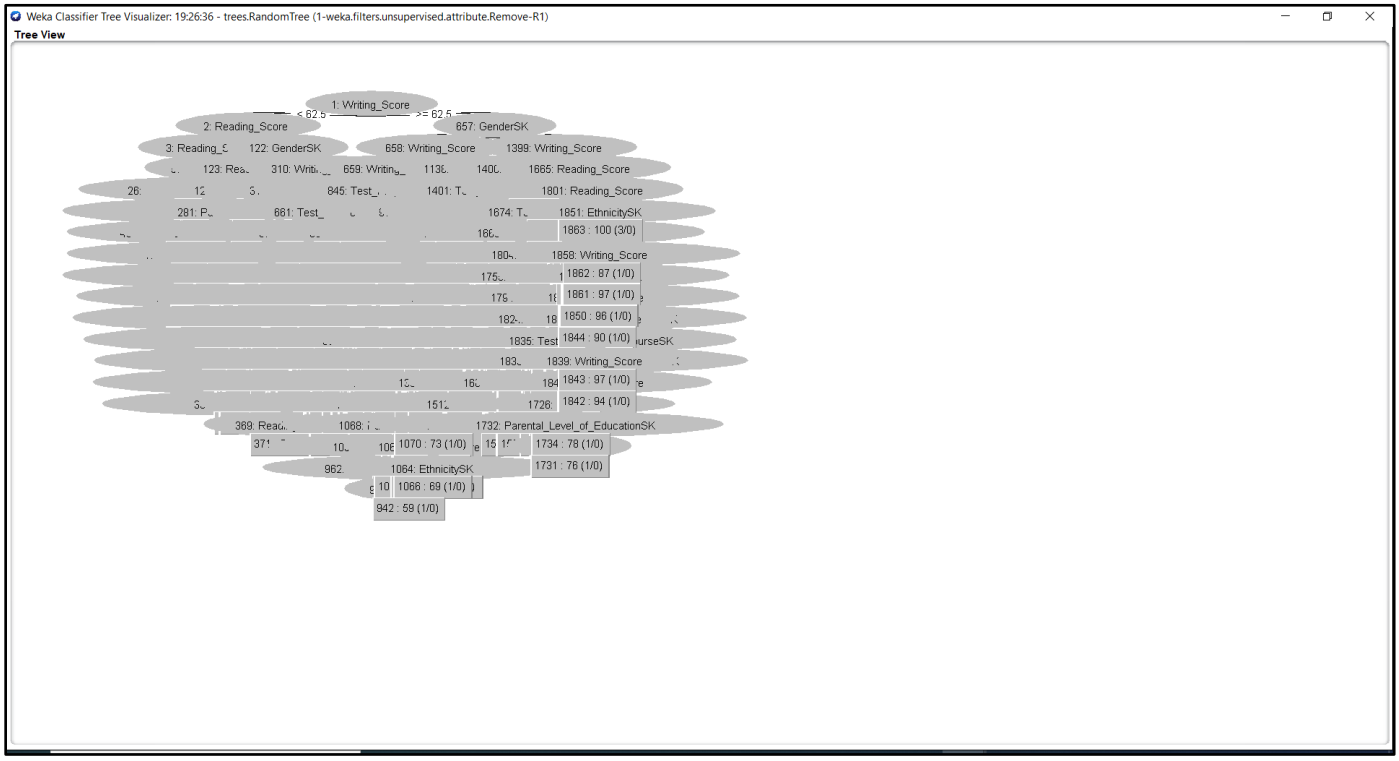

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) Writing_Score
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
18:07:35 - EM
18:08:39 - HierarchicalClusterer
Clusterer output
Attributes: 9
FactStudentPerformanceID
GenderSK
EthnicitySK
Parental_Level_of_EducationSK
LunchSK
Test_Preparation_CourseSK
Math_Score
Reading_Score
Writing_Score
Test mode: evaluate on training data
=== Clustering model (full training set) ===
Cluster 0
Cluster 1
Time taken to build model (full training data) : 2.52 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 939 (94%)
1 61 (6%)
Status
OK
Log
x 0
Type here to search
6:08 PM
3/20/2021

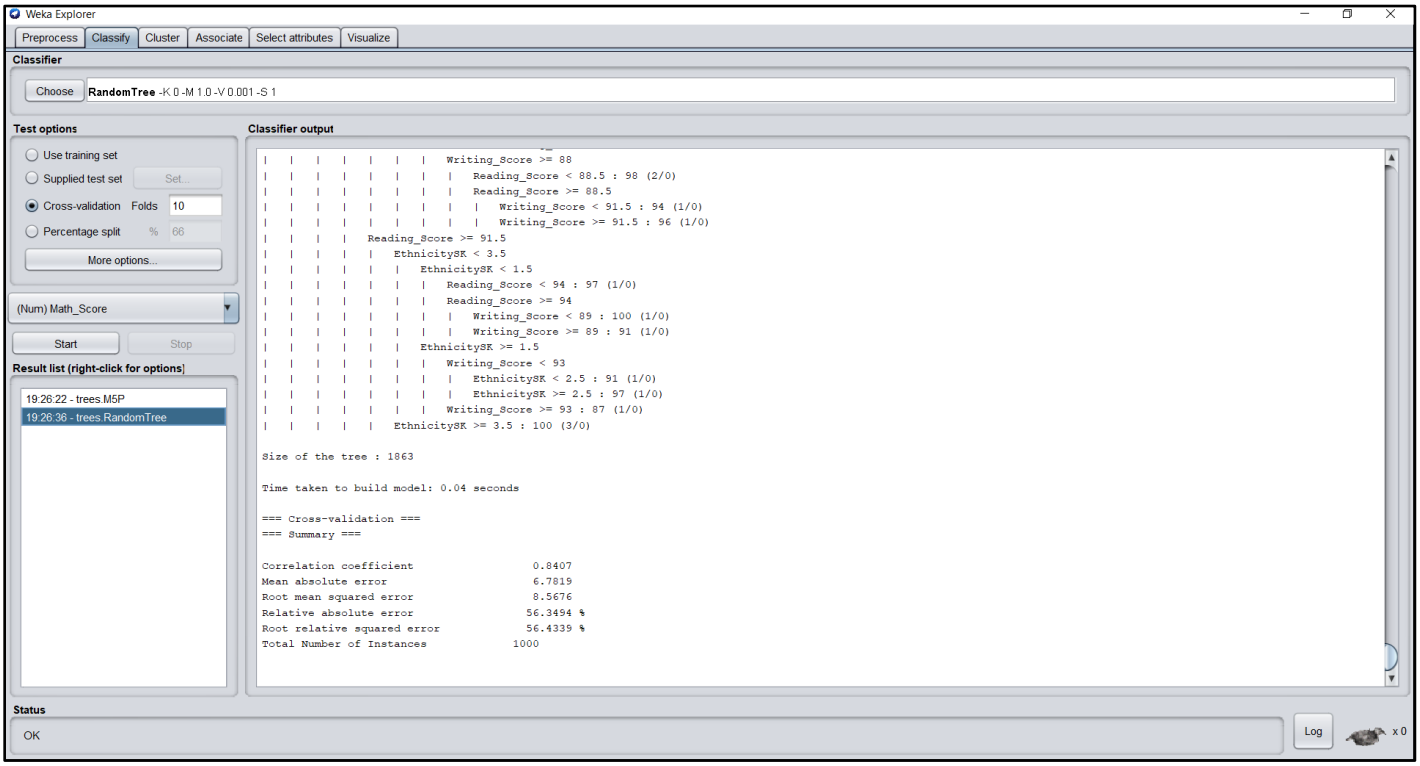

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval -P 1 -E 1
Search Method
Choose
BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
No class
Start
Stop
Result list (right-click for options)
18:09:43 - BestFirst + CfsSubsetEval
Attribute selection output
Reading_Score
Writing_Score
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: 42
Merit of best subset found: 0.959
Attribute Subset Evaluator (supervised, Class (numeric): 9 Writing_Score):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 1,2,3,4,5,6,8 : 7
FactStudentPerformanceID
GenderSK
EthnicitySK
Parental_Level_of_EducationSK
LunchSK
Test_Preparation_CourseSK
Reading_Score
Status
OK
Log
x 0
Type here to search
6:10 PM
3/20/2021

Regression

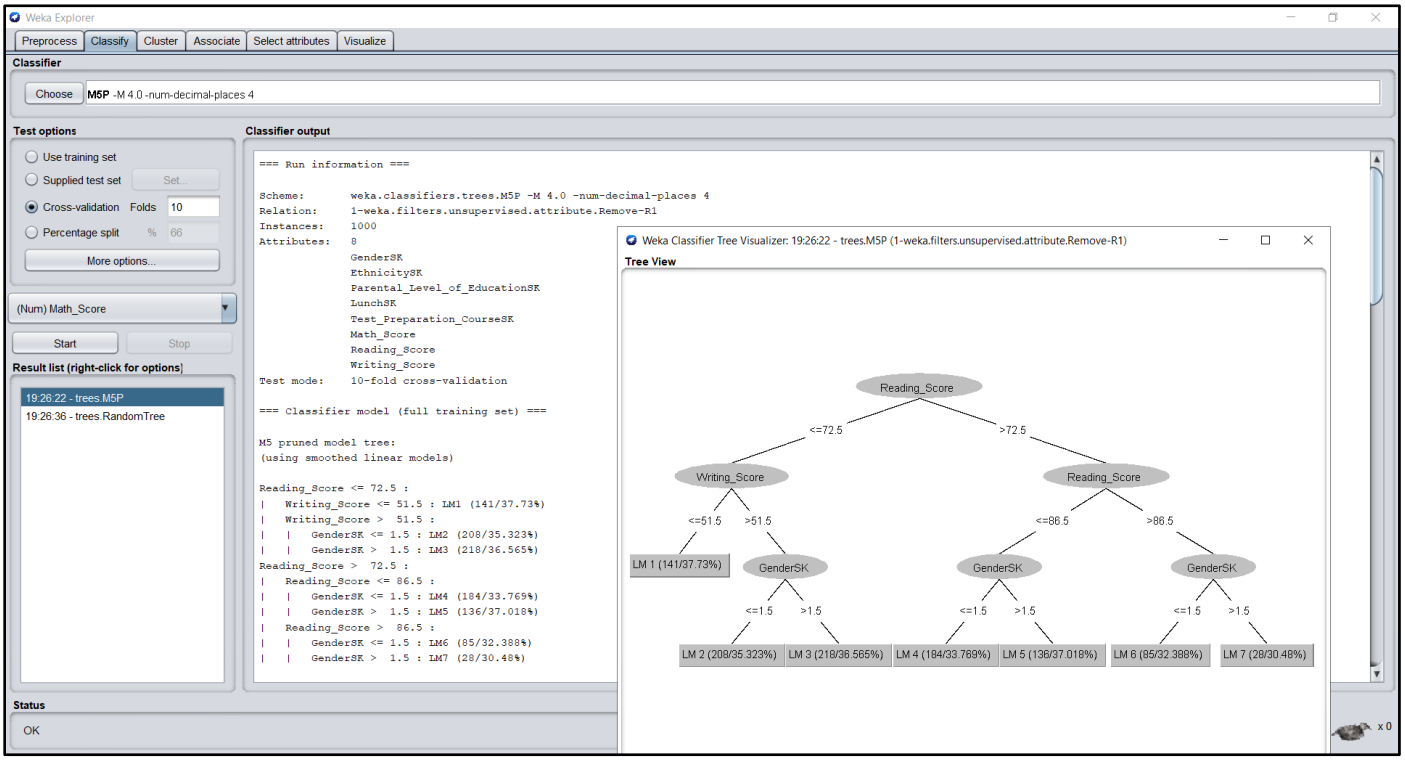

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Writing_Score
Start
Stop
Result list (right-click for options)
20:23:26 - functions.LinearRegression
Classifier output
Parental_Level_of_EducationSK
LunchSK
Test_Preparation_CourseSK
Writing_Score
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Linear Regression Model
Writing_Score =
-9.366 * GenderSK +
7.6223 * EthnicitySK +
-3.6671 * Parental_Level_of_EducationSK +
7.9854 * LunchSK +
-10.0956 * Test_Preparation_CourseSK +
71.5798
Time taken to build model: 0 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.5248
Mean absolute error 10.3259
Root mean squared error 12.9291
Relative absolute error 84.5425 %
Root relative squared error 85.008 %
Total Number of Instances 1000
Status
OK
Log
x 0

Random Tree


Weka Classifier Tree Visualizer: 19:26:36 - trees.RandomTree (1-weka.filters.unsupervised.attribute.Remove-R1)
Tree View
1: Writing_Score
< 62.5
>= 62.5
2: Reading_Score
657: GenderSK
3: Reading_S
122: GenderSK
658: Writing_Score
1399: Writing_Score
123: Rea
310: Writi
659: Writing_
1130
1400
1665: Reading_Score
28
12
845: Test_
1401: T
1801: Reading_Score
281: P
661: Test_
1674: T
1851: EthnicitySK
1863 : 100 (3/0)
1858: Writing_Score
1882 : 87 (1/0)
1861 : 97 (1/0)
1850 : 96 (1/0)
1844 : 90 (1/0)
1835: Test
1839: Writing_Score
1843 : 97 (1/0)
1842 : 94 (1/0)
1726
1732: Parental_Level_of_EducationSK
389: Read
1068
1070 : 73 (1/0)
1734 : 78 (1/0)
1731 : 76 (1/0)
962
1064: EthnicitySK
1066 : 69 (1/0)
942 : 59 (1/0)


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Math_Score
Start
Stop
Result list (right-click for options)
19:26:22 - trees.M5P
19:26:36 - trees.RandomTree
Classifier output
| | | | | | | Writing_Score >= 88
| | | | | | | | Reading_Score < 88.5 : 98 (2/0)
| | | | | | | | Reading_Score >= 88.5
| | | | | | | | | Writing_Score < 91.5 : 94 (1/0)
| | | | | | | | | Writing_Score >= 91.5 : 96 (1/0)
| | | | Reading_Score >= 91.5
| | | | | EthnicitySK < 3.5
| | | | | | EthnicitySK < 1.5
| | | | | | | Reading_Score < 94 : 97 (1/0)
| | | | | | | Reading_Score >= 94
| | | | | | | | Writing_Score < 89 : 100 (1/0)
| | | | | | | | Writing_Score >= 89 : 91 (1/0)
| | | | | | EthnicitySK >= 1.5
| | | | | | | Writing_Score < 93
| | | | | | | | EthnicitySK < 2.5 : 91 (1/0)
| | | | | | | | EthnicitySK >= 2.5 : 97 (1/0)
| | | | | | | Writing_Score >= 93 : 87 (1/0)
| | | | | EthnicitySK >= 3.5 : 100 (3/0)
Size of the tree : 1863
Time taken to build model: 0.04 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.8407
Mean absolute error 6.7819
Root mean squared error 8.5676
Relative absolute error 56.3494 %
Root relative squared error 56.4339 %
Total Number of Instances 1000
Status
OK
Log
x 0

M5P


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
M5P -M 4.0 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Math_Score
Start
Stop
Result list (right-click for options)
19:26:22 - trees.M5P
19:26:36 - trees.RandomTree
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.M5P -M 4.0 -num-decimal-places 4
Relation: 1-weka.filters.unsupervised.attribute.Remove-R1
Instances: 1000
Attributes: 8
GenderSK
EthnicitySK
Parental_Level_of_EducationSK
LunchSK
Test_Preparation_CourseSK
Math_Score
Reading_Score
Writing_Score
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
M5 pruned model tree:
(using smoothed linear models)
Reading_Score <= 72.5 :
| Writing_Score <= 51.5 : LM1 (141/37.73%)
| Writing_Score > 51.5 :
| | GenderSK <= 1.5 : LM2 (208/35.323%)
| | GenderSK > 1.5 : LM3 (218/36.565%)
Reading_Score > 72.5 :
| Reading_Score <= 86.5 :
| | GenderSK <= 1.5 : LM4 (184/33.769%)
| | GenderSK > 1.5 : LM5 (136/37.018%)
| Reading_Score > 86.5 :
| | GenderSK <= 1.5 : LM6 (85/32.388%)
| | GenderSK > 1.5 : LM7 (28/30.48%)
Weka Classifier Tree Visualizer: 19:26:22 - trees.M5P (1-weka.filters.unsupervised.attribute.Remove-R1)
Tree View
Reading_Score
<=72.5
>72.5
Writing_Score
Reading_Score
<=51.5
>51.5
<=86.5
>86.5
LM 1 (141/37.73%)
GenderSK
GenderSK
GenderSK
<=1.5
>1.5
<=1.5
>1.5
<=1.5
>1.5
LM 2 (208/35.323%)
LM 3 (218/36.565%)
LM 4 (184/33.769%)
LM 5 (136/37.018%)
LM 6 (85/32.388%)
LM 7 (28/30.48%)
Status
OK
x 0

10. Data Mining techniques

There are number of data mining techniques such as predicting, data cleaning, association, classification, regression and tracking patterns. Among mentioned classification techniques we have used,

Clustering

Clustering gives the meaning of grouping. Grouping helps in analysis. In our data set we have grouped our data according to a similarity or a behavior. We have used cluster technique with the use of Weka software. Under clustering we have used canopy cluster, EM cluster and hierarchical clusters.

Association

Association is the combination of two or more groups. Here, we have used combined two fields and analysis were done. Apriori is the default algorithm used in association data mining technique.

Classification

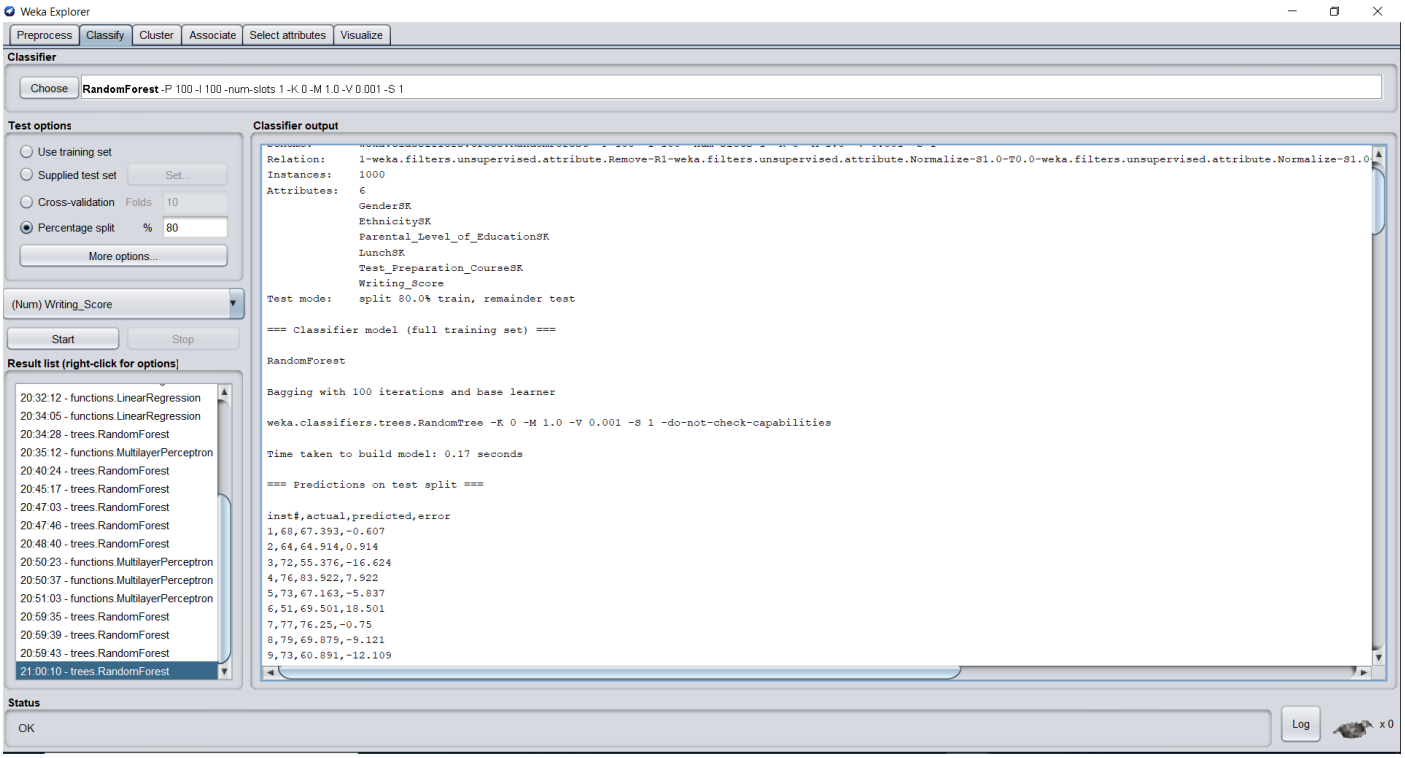
Classification helps in classifying data in different classes. We have used classification as a data mining technique for easy analysis.

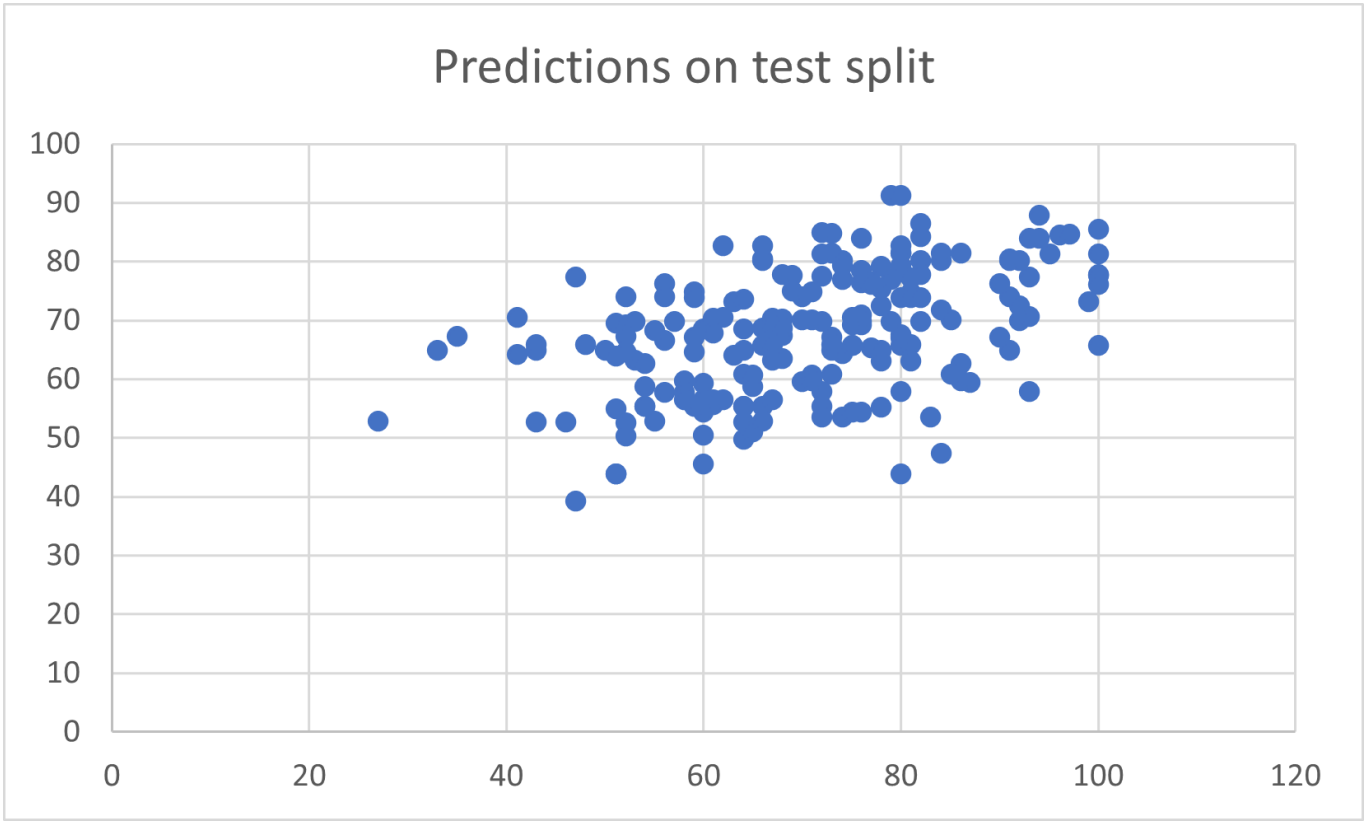
Percentage split and cross validation is used with decision trees.

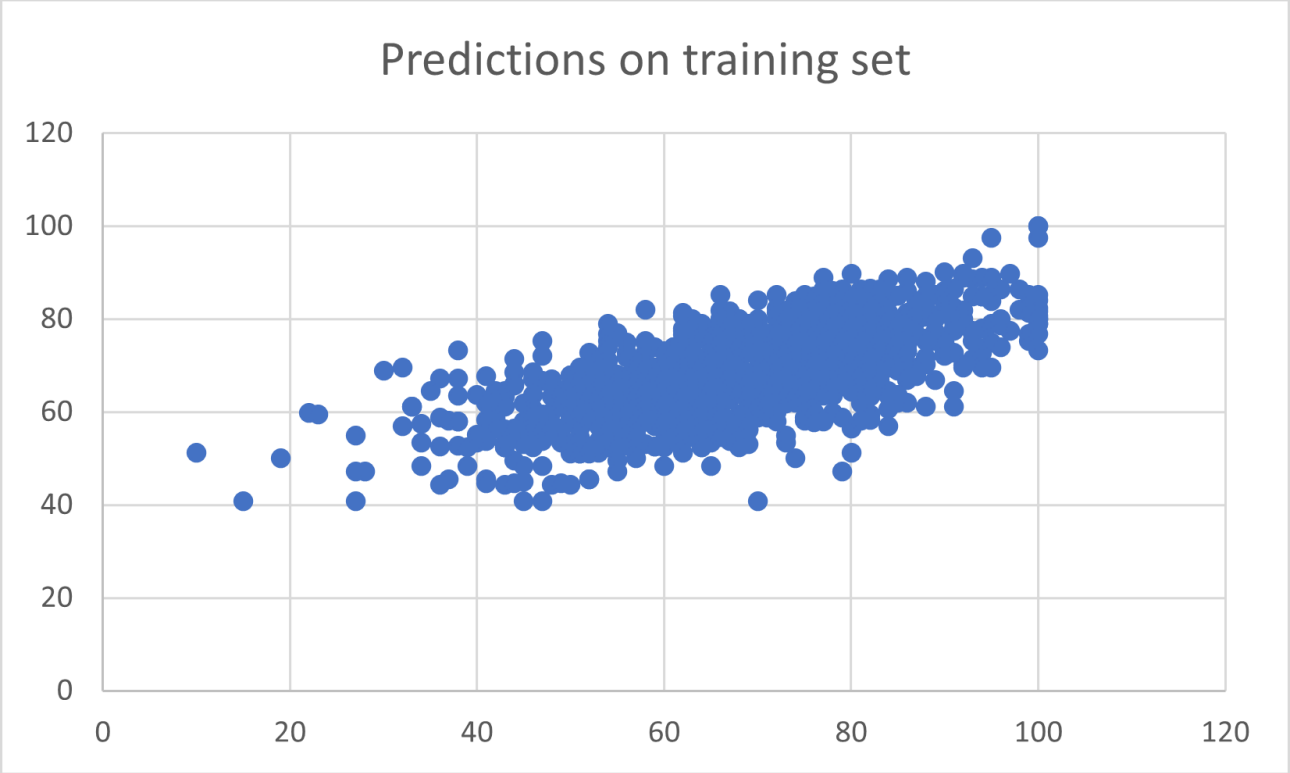
11. Predictions

There are some predictions which can be generated through the output results.

- When we compare the scores of math, reding and writng with gender, female students have scored more than male students.
- Students who didn't get the lunch or get less lunch will receive less marks, but students who get standard lunch will score more.
- Students who prepare for the examination will get high score.
- Female students of ethnicity group C will score high score than other ethnicity group female students and all male students.
- Ethnicity Group A female student will get less marks for math.
- Ethnicity Group C female students whose parent's have associate degrees will score more for writing.
- Students performance does not depend on Parent's education level.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
80
More options...
(Num) Writing_Score
Start
Stop
Result list (right-click for options)
20:32:12 - functions.LinearRegression
20:34:05 - functions.LinearRegression
20:34:28 - trees.RandomForest
20:35:12 - functions.MultilayerPerceptron
20:40:24 - trees.RandomForest
20:45:17 - trees.RandomForest
20:47:03 - trees.RandomForest
20:47:46 - trees.RandomForest
20:48:40 - trees.RandomForest
20:50:23 - functions.MultilayerPerceptron
20:50:37 - functions.MultilayerPerceptron
20:51:03 - functions.MultilayerPerceptron
20:59:35 - trees.RandomForest
20:59:39 - trees.RandomForest
20:59:43 - trees.RandomForest
21:00:10 - trees.RandomForest
Classifier output
Relation: 1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsupervised.attribute.Normalize-S1.0
Instances: 1000
Attributes: 6
GenderSK
EthnicitySK
Parental_Level_of_EducationSK
LunchSK
Test_Preparation_CourseSK
Writing_Score
Test mode: split 80.0% train, remainder test
=== Classifier model (full training set) ===
RandomForest
Bagging with 100 iterations and base learner
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
Time taken to build model: 0.17 seconds
=== Predictions on test split ===
inst#,actual,predicted,error
1,68,67.393,-0.607
2,64,64.914,0.914
3,72,55.376,-16.624
4,76,83.922,7.922
5,73,67.163,-5.837
6,51,69.501,18.501
7,77,76.25,-0.75
8,79,69.879,-9.121
9,73,60.891,-12.109
Status
OK
Log
x 0


Predictions on test split
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120



12. Conclusion

Dataset of student performance is mined using three data mining techniques clustering, association and classification. Analysis reports are designed using Microsoft SQL server, Microsoft Excel and Weka software. With the use of generated reports some predictions are done for the performance of the students. These predicts are based on the gender, Ethnicity group, educational level of their parents, test preparation and the lunch with the performance of the students.