Stamford University Bangladesh

Education for tomorrow's World a greener world

DMKD FINAL PROJECT

Course Code : CSI 382

Course Title : Data Mining and Knowledge Discovery Sessional

Submitted By : Sumaiya Akter(CSE06307402)

Submitted To : Md. Towhidul Islam Robin (Senior Lecturer)

[Date : 12/05/2021]

Student Information Analysis

& Predict Their Lunch Quality

1.**About Data:** This dataset has been collect from https://www.kaggle.com/spscientist/students-performance-in-exams/tasks?taskId=2743 its about Predicting student lunch quality with the demographic and socioeconomic information. I will classify the lunch quality between Standard or free/reduced.

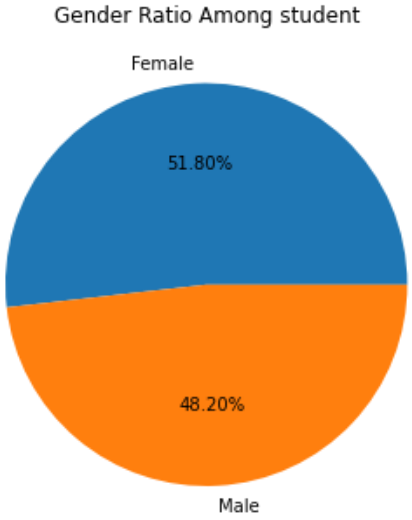
In this data set there are 8 feature they are describing below;

gender : this feature define whether the student is male or female
race/ethnicity :this feature distribute the student among group A', 'group B', 'group C', 'group D', 'group E
parental_level_of_education : this define the educational level of parents and distribute them among these categories associates degree', 'bachelor degree', 'high school', 'master degree', 'some college', 'some high school']
lunch: this feature describe the lunch quality of student whether it is Standard or free/reduced.
test preparation course this feature describe the test preparation quality of student whether it is complete or not complete .
math score: Marks secured by the students in math
reading_score : Marks secured by the students in reading
writing_score : Marks secured by the students in writing

In this data set there is 1000 records are gathered by student.

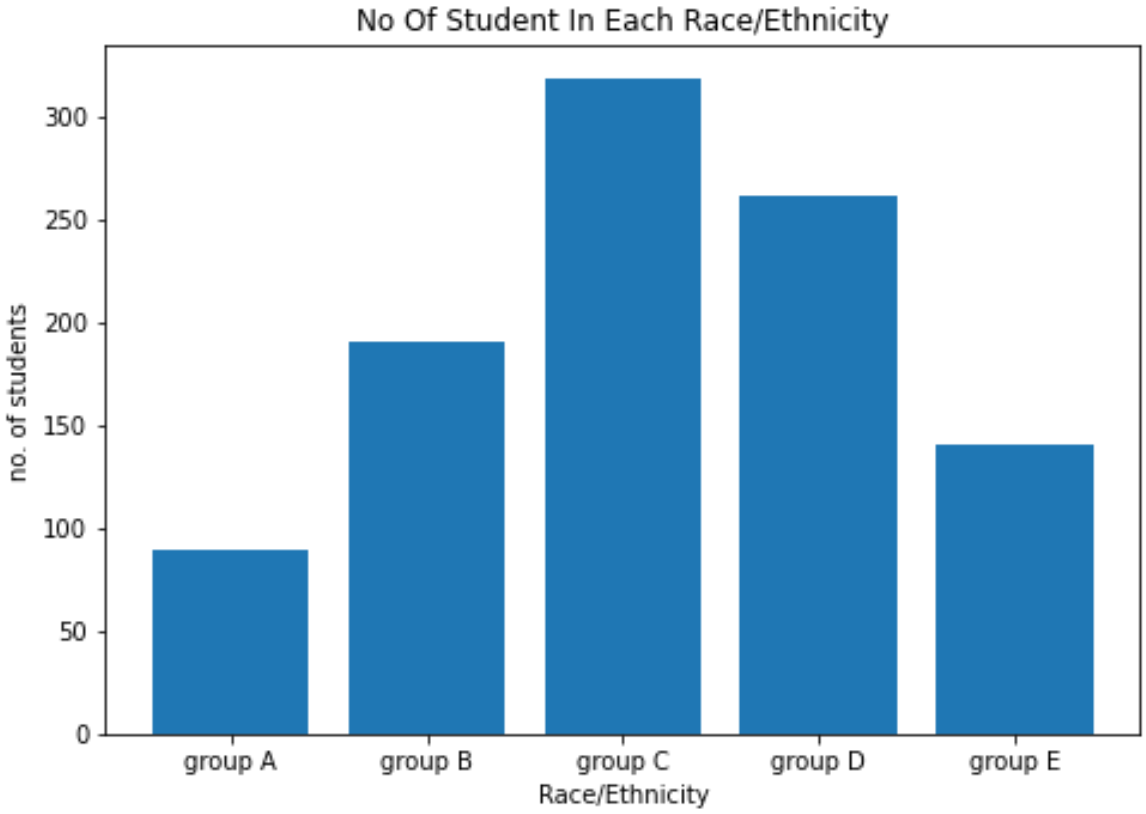
2.Dataset Properties:

Gender**:** Among Thousand student there are 518 female & 482 male student.

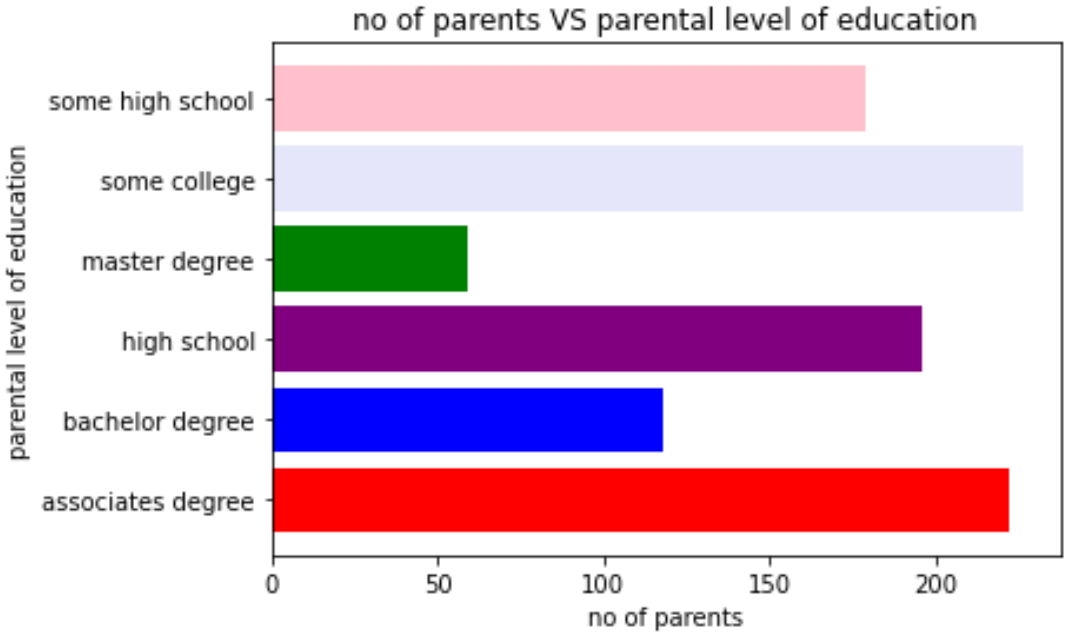


So there is more female student than male student.

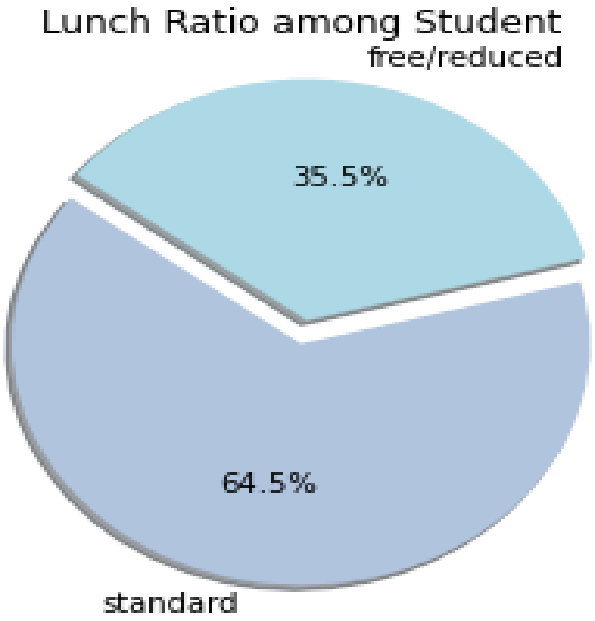
Race/Ethnicity: Most of the student are from group C > group D> group B > group E> group A>



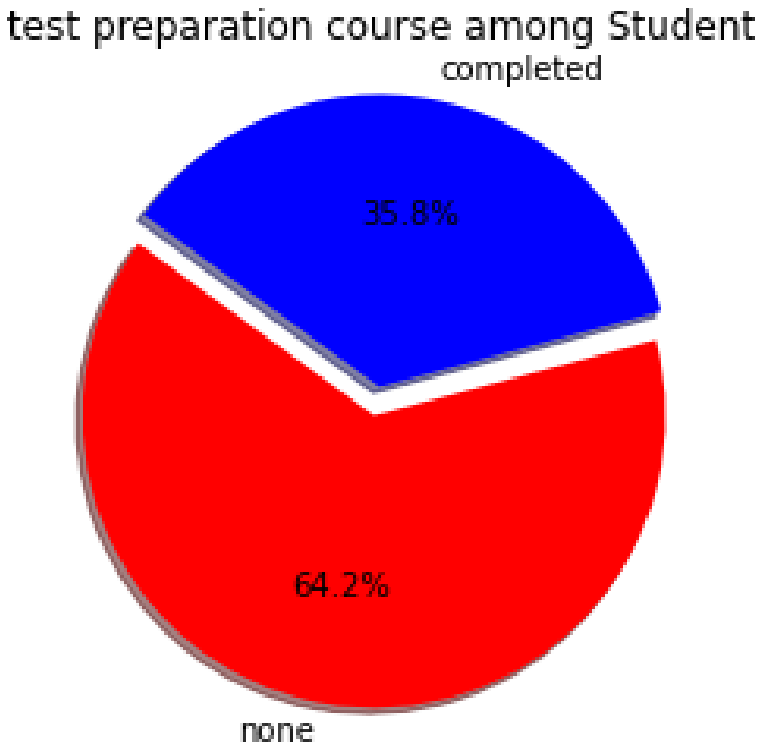
parental_level_of_education: Most of the parents are from 'some college'>associates degree',> 'high school'>"some high school'>bachelor degree'>'master degree',



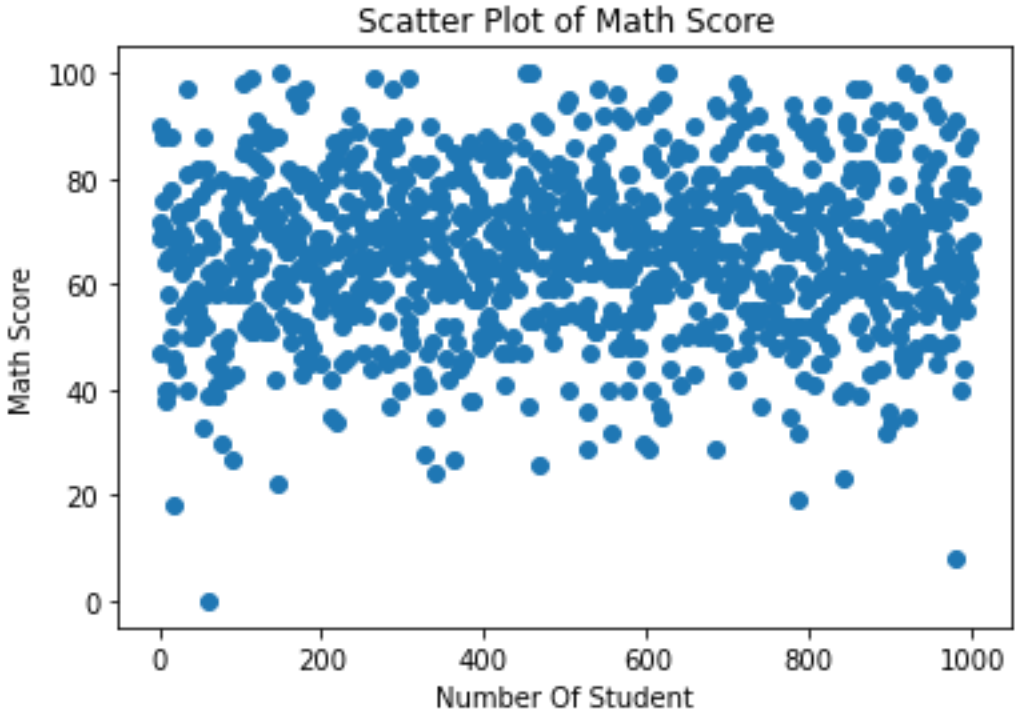
Lunch Ratio : More student get Standard lunch than the free lunch.



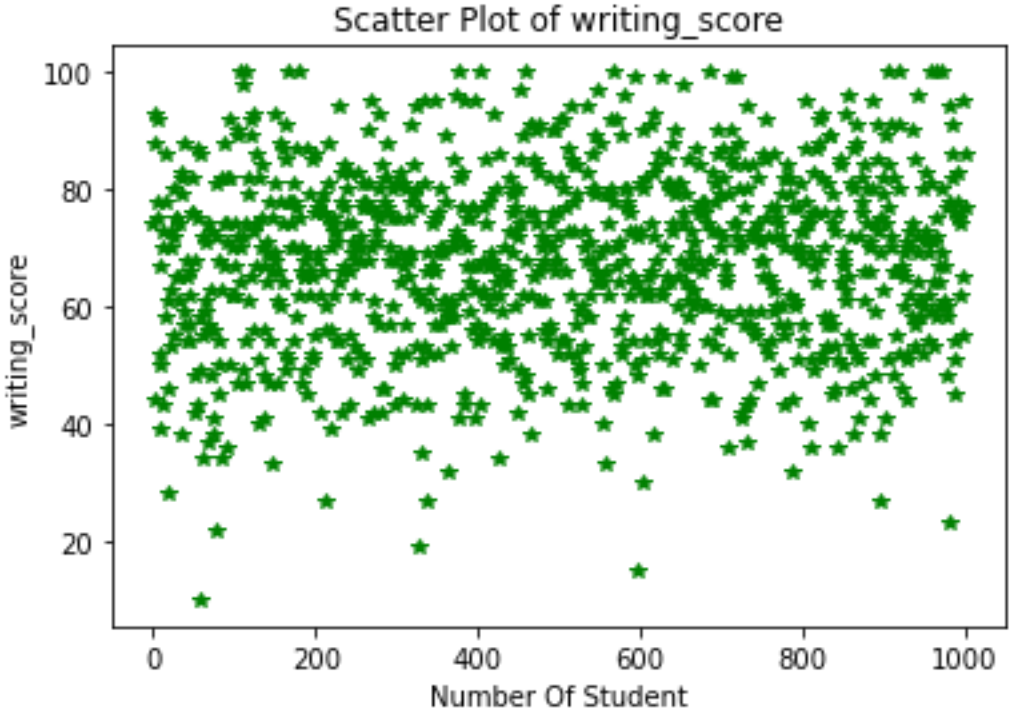
test preparation course: Only 35% student take test preparation.



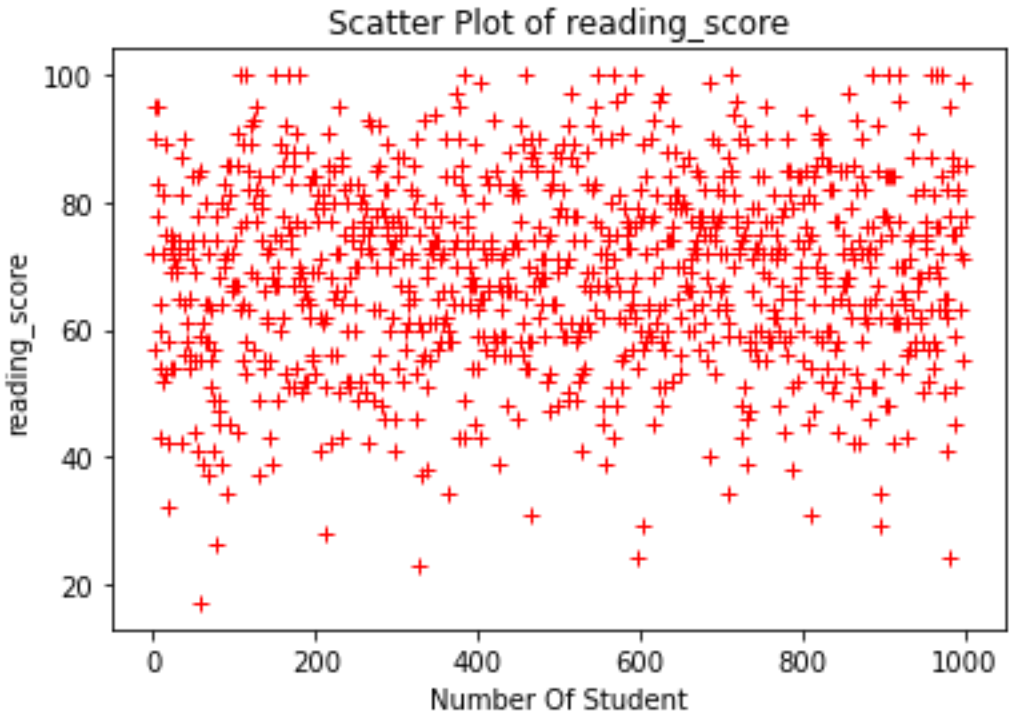
Math score: Secured By Student



Writing score: Secured By Student



Reading score: Secured By Student



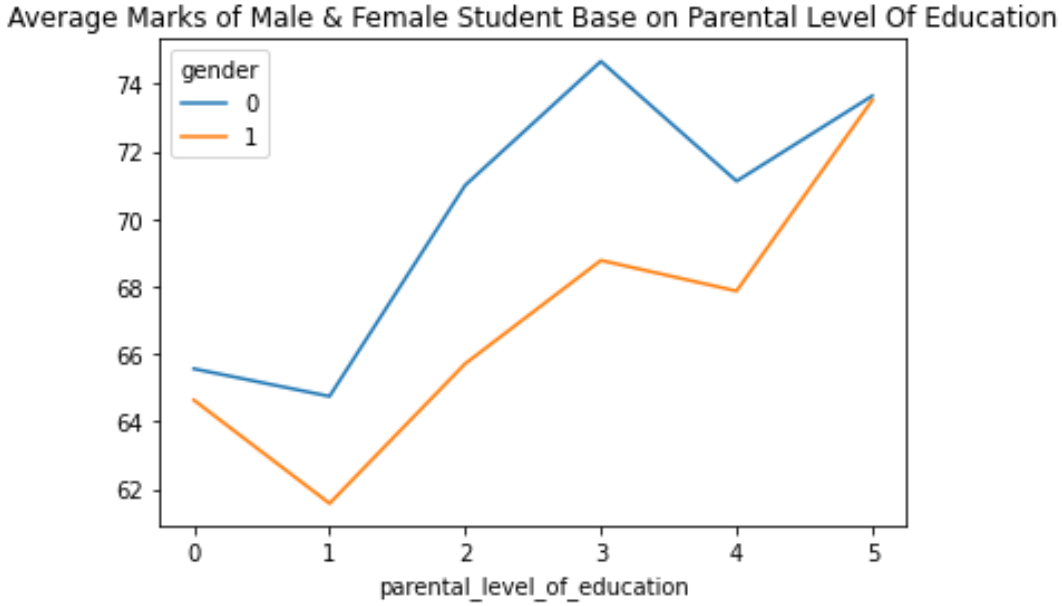
3. Preprocessing Of Data:

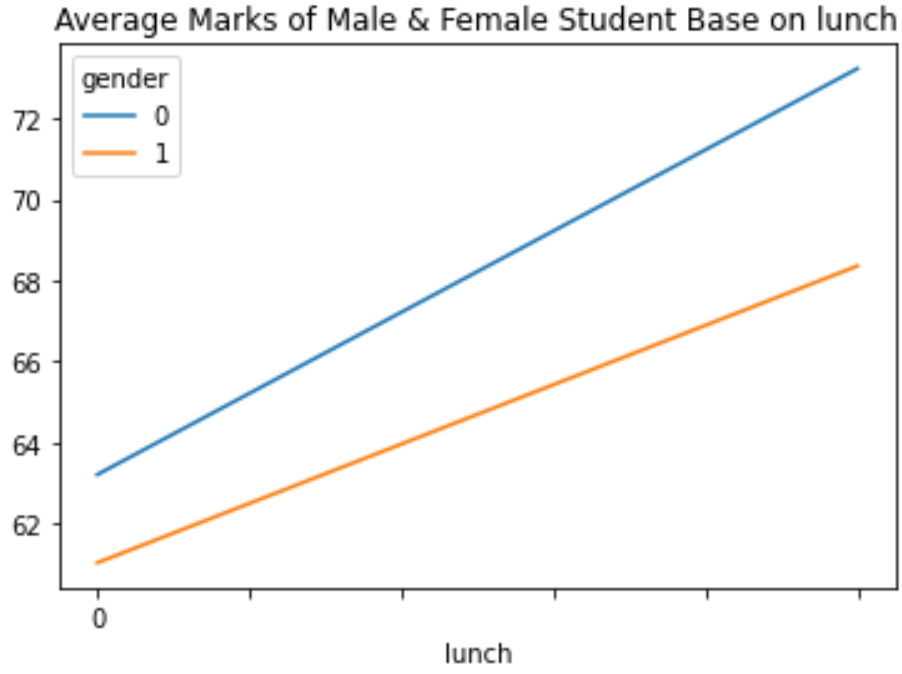
In this step data has been checked for duplication. For analyzing purpose categorical value of some column has been replace by the numerical value.

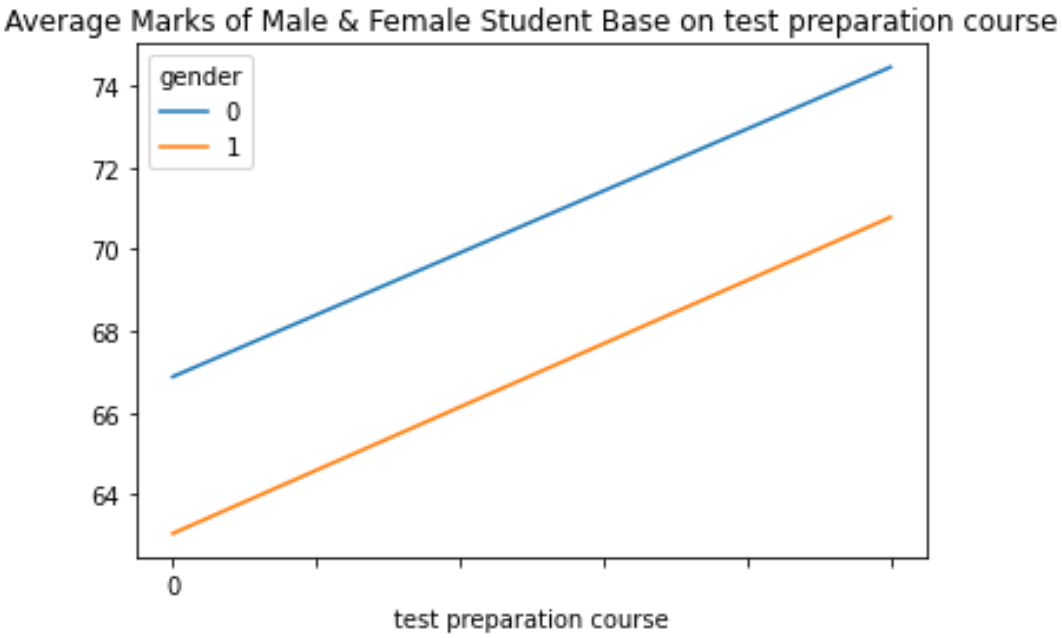
	gender	race/ethnicity	parental_level_of_education	lunch	test preparation course	math_score	reading_score	writing_score
0	0	1	3	1	0	72	72	74
1	0	2	2	1	1	69	90	88
2	0	1	5	1	0	90	95	93
3	1	0	4	0	0	47	57	44
4	1	2	2	1	0	76	78	75
...	...	...	...	...	...	...	...	...
995	0	4	5	1	1	88	99	95
996	1	2	1	0	0	62	55	55
997	0	2	1	0	1	59	71	65
998	0	3	2	1	1	68	78	77
999	0	3	2	0	0	77	86	86

6. **Analyze dataset to uncover hidden information.:**

Here I find average marks of student. And then plot the “Average Marks of Male & Female Student Base on Parental Level Of Education”
'Average Marks of Male & Female Student Base on lunch'
Average Marks of Male & Female Student Base on test preparation course







7. Split data set: Here Split my data set into four different ratios as TRAIN/ TEST Ratio 90/10 75/25 50/50 30/70 & Do the further procedure.

8. Fit your data into ML models: Here we use different ML models

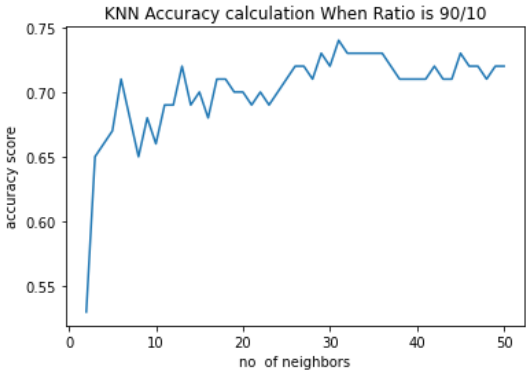
- KNN (with optimum value of K): I found the optimum value of k in different ratios

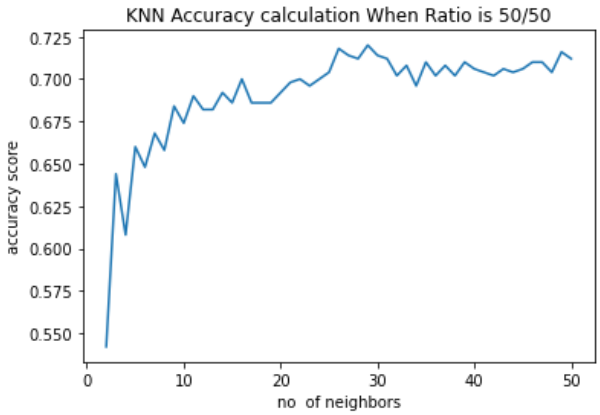
  90/10 - optimum value of K is 34

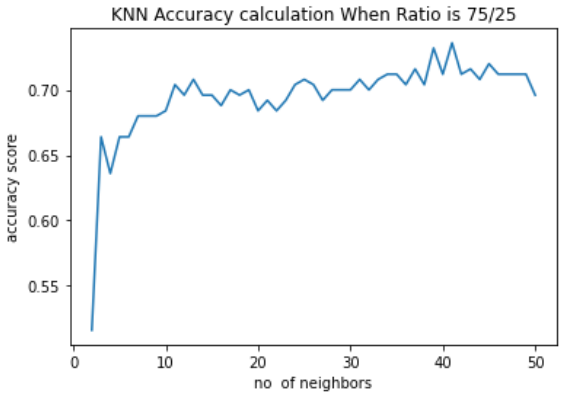
  75/25 - optimum value of K is 34

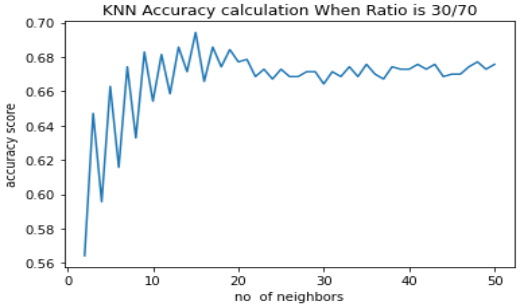
  50/50 - optimum value of K is 28

  30/70- optimum value of K is 13









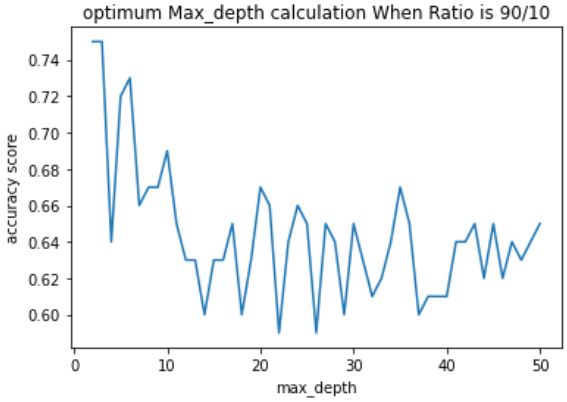
- Decision Tree (with optimum value of max depth) :): I found the optimum value of max depth in different ratios

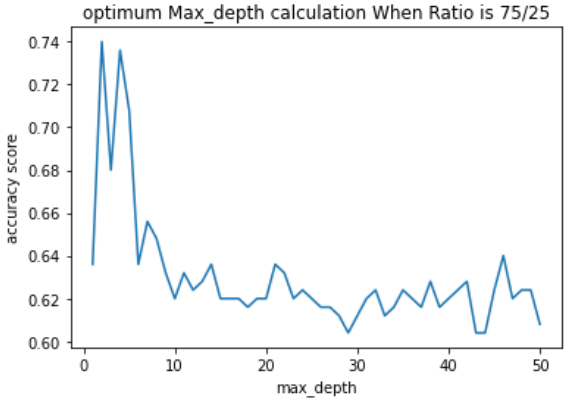
  90/10 - optimum value max depth of is 4

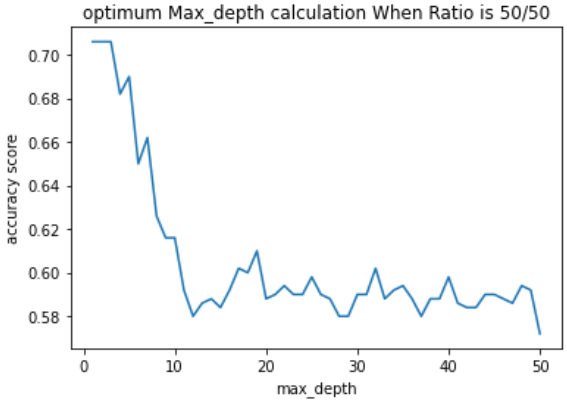
  75/25 - optimum value max depth of is 2

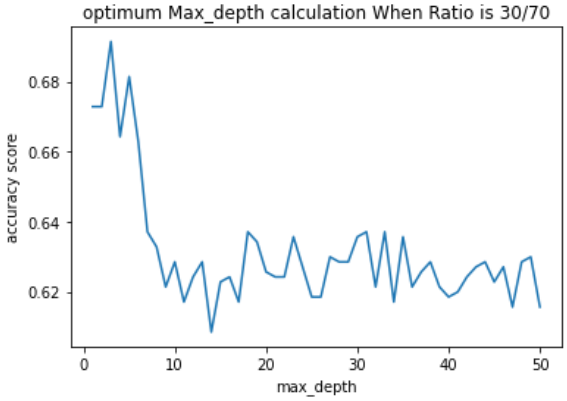
  50/50 - optimum value max depth of is 3

  30/70- optimum value max depth of is 4


optimum Max_depth calculation When Ratio is 90/10
accuracy score
max_depth


optimum Max_depth calculation When Ratio is 75/25
accuracy score
max_depth

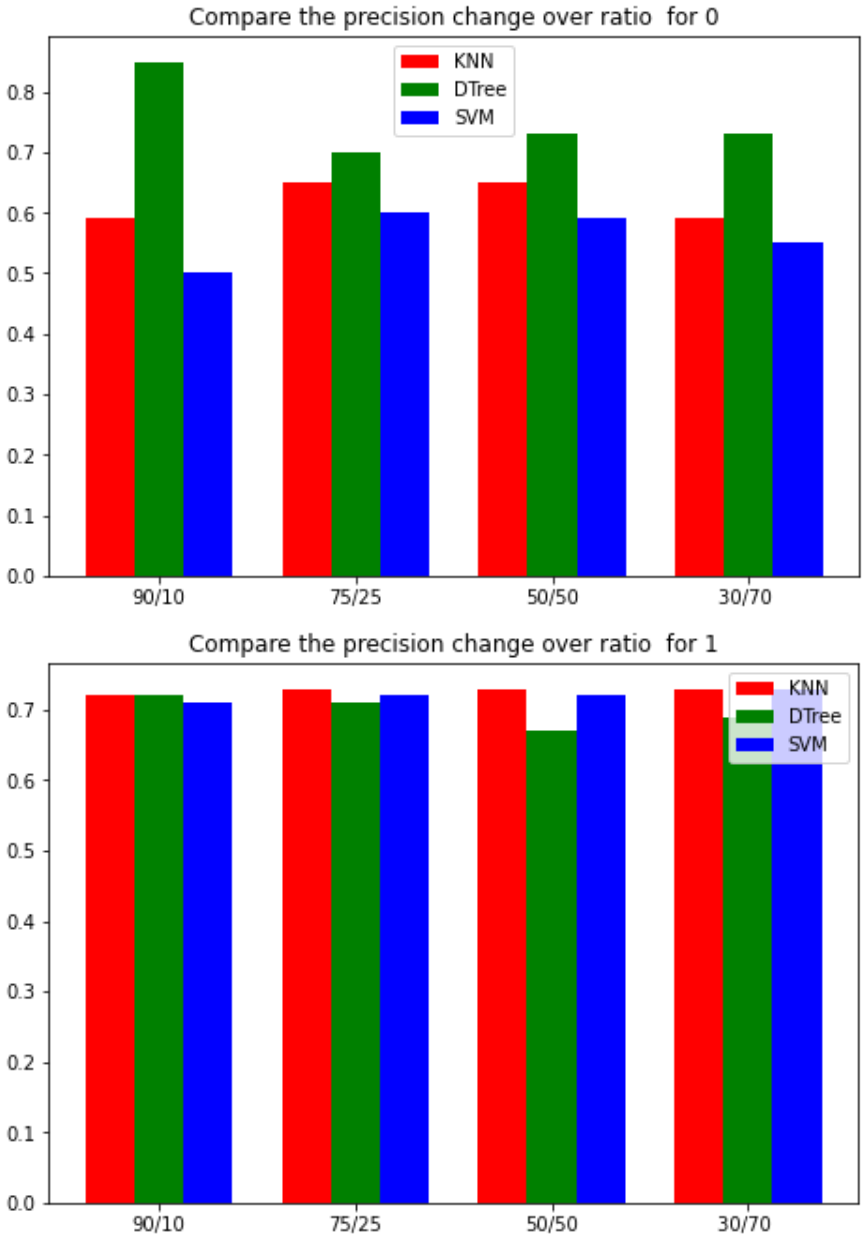

optimum Max_depth calculation When Ratio is 50/50
accuracy score
max_depth


optimum Max_depth calculation When Ratio is 30/70
accuracy score
max_depth

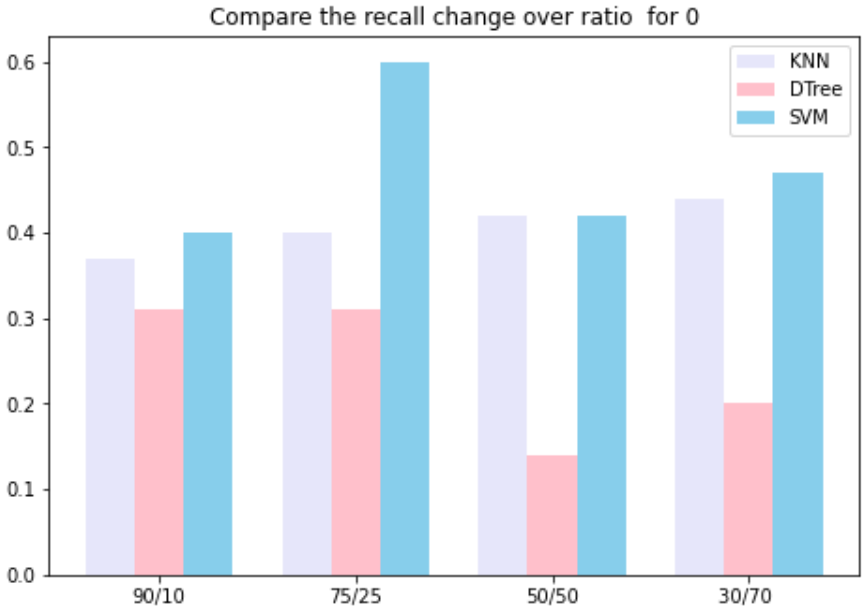
- SVM : I use SVM Supervised ML model for fitting the data.

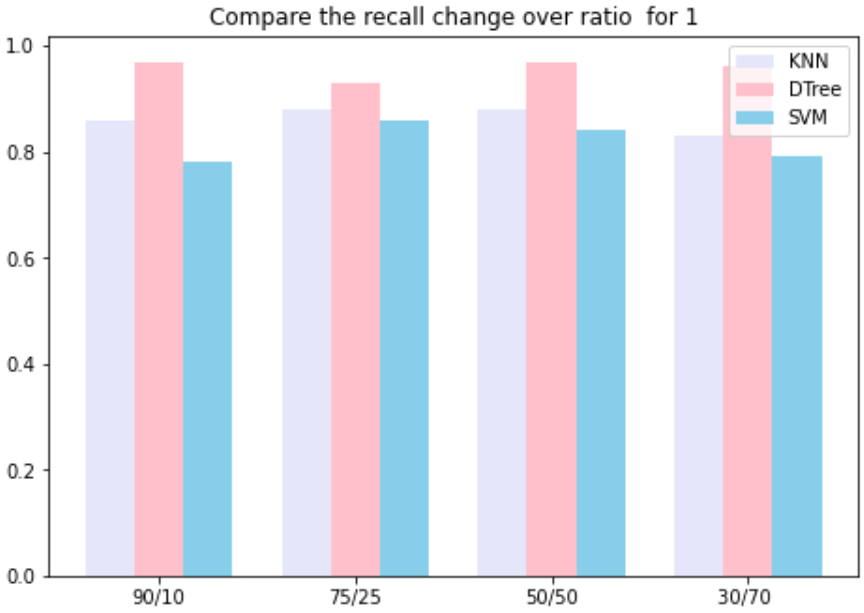
9. **Compare the accuracy, precision, recall, and f1 score for different algorithm and different ratios:**

Precision Comparison:

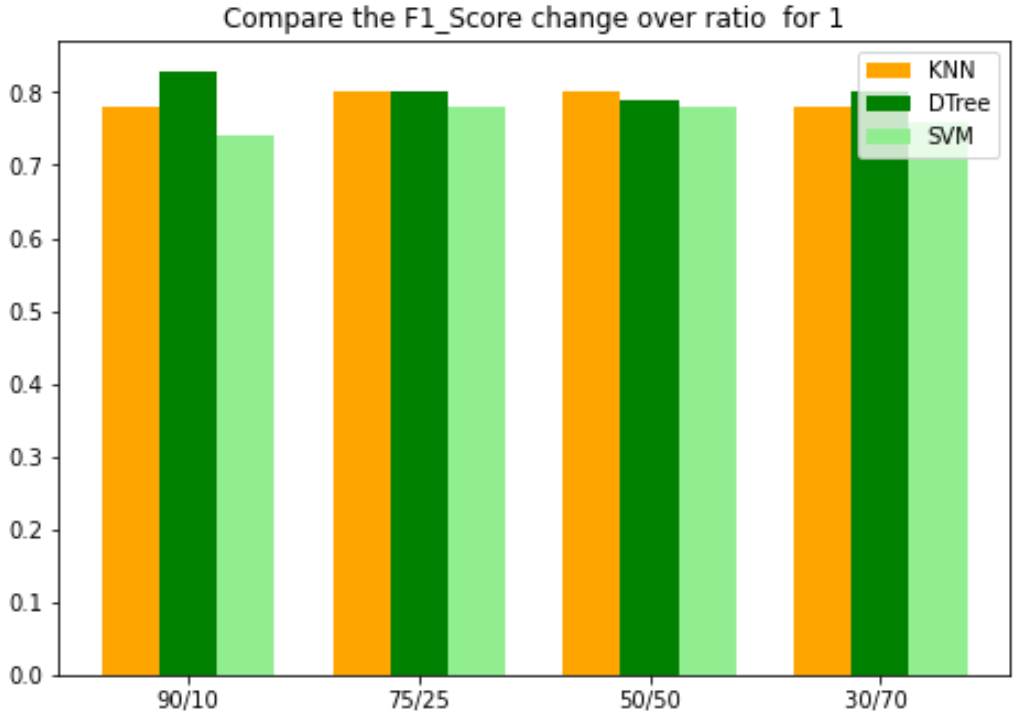


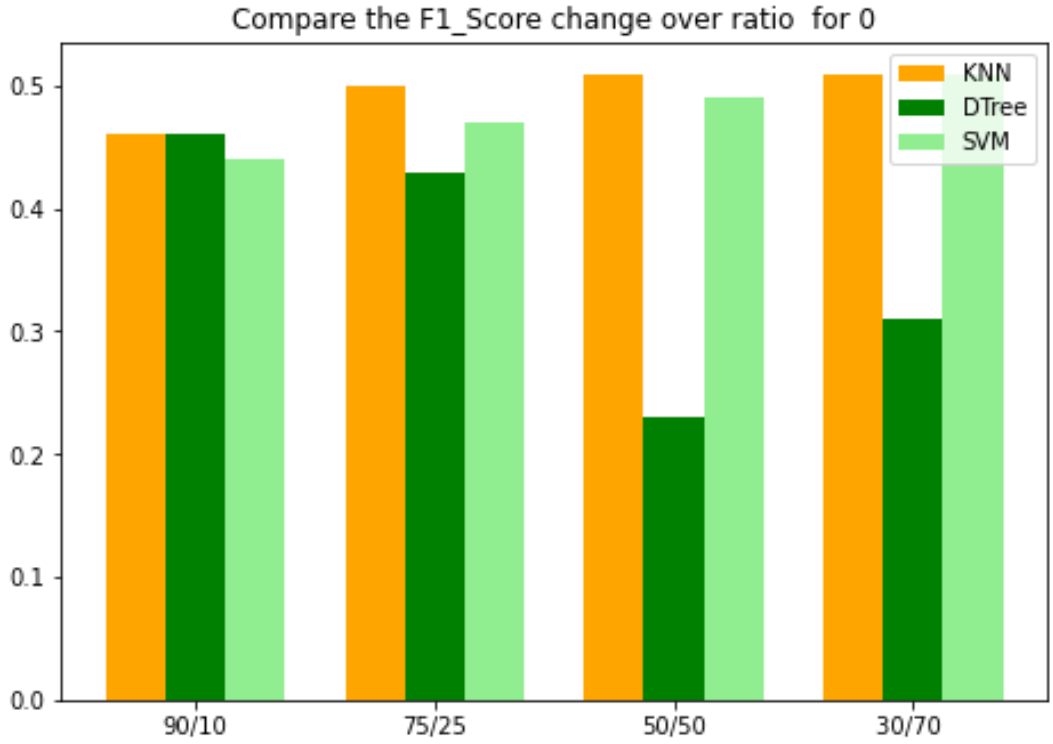
Recall Comparison:



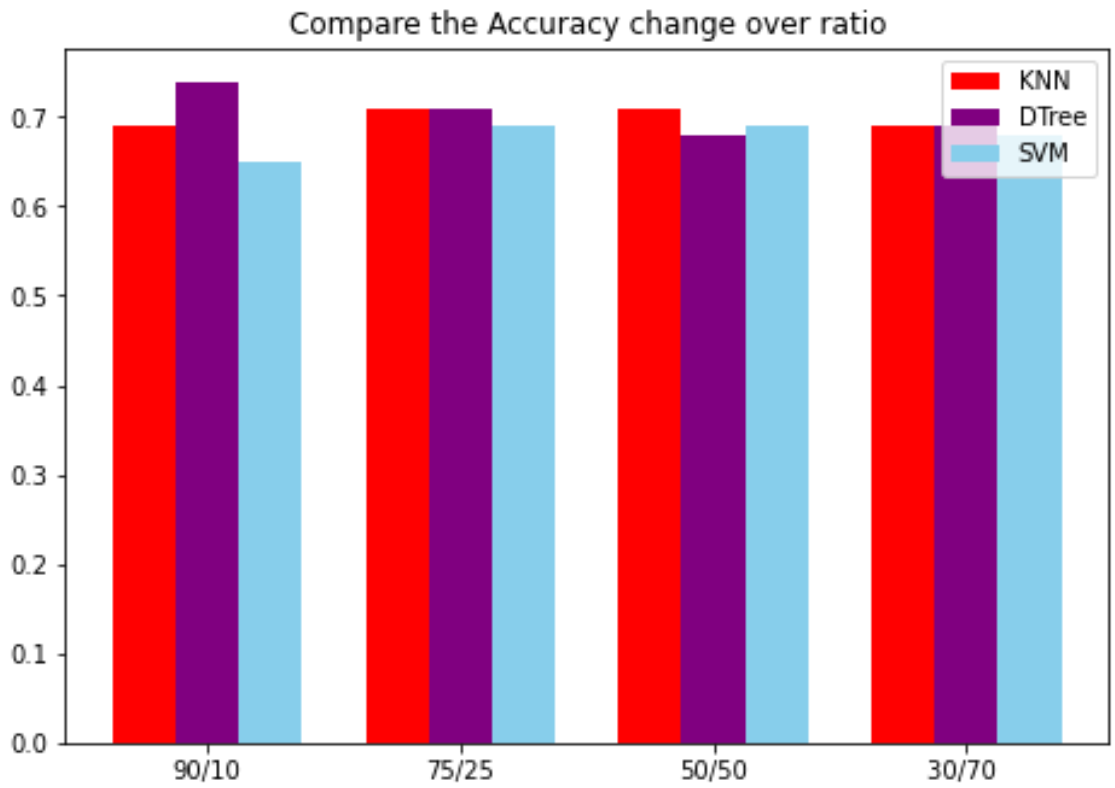


F1_Score Comparison:





Accuracy Comparison: observing this comparison we see DTree have highest accuracy in 90/10 ratio. KNN have highest accuracy in 50/50 ratio. SVM have highest accuracy in 50/50 ratio.



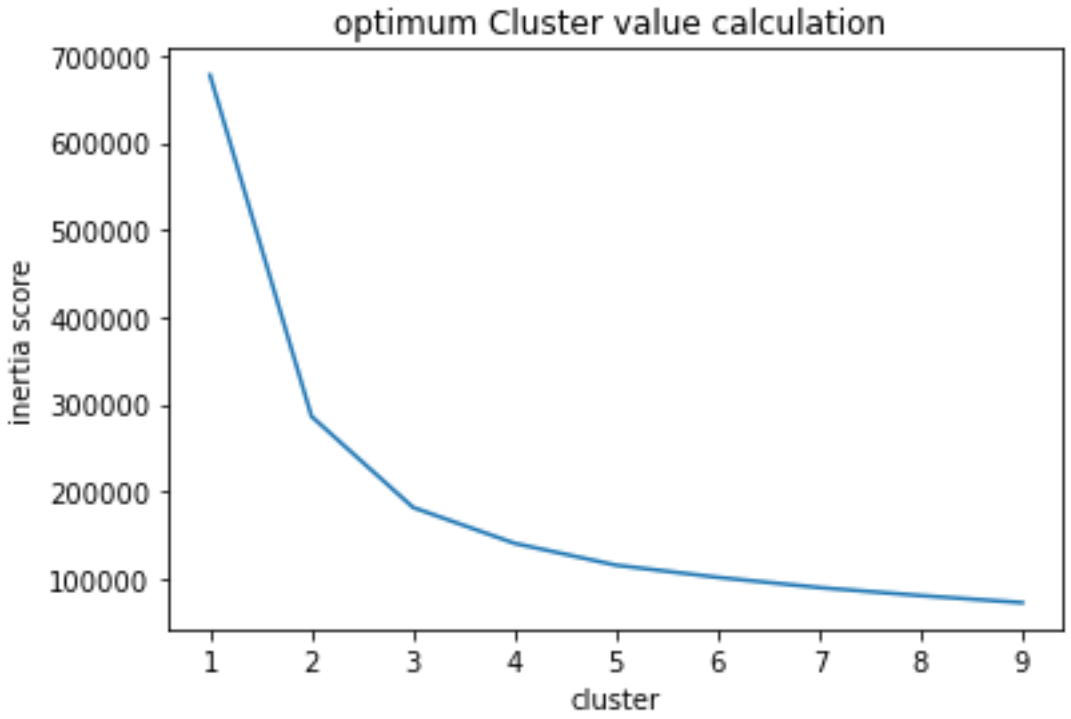
10.Fit the data set into KMean algorithm.

In this section I separate the target variable lunch from the orginal data set and fit the value in KMEAN

	gender	race/ethnicity	parental_level_of_education	test preparation course	math_score	reading_score	writing_score
0	0	1	3	0	72	72	74
1	0	2	2	1	69	90	88
2	0	1	5	0	90	95	93
3	1	0	4	0	47	57	44
4	1	2	2	0	76	78	75
...	...	...	...	...	...	...	...
995	0	4	5	1	88	99	95
996	1	2	1	0	62	55	55
997	0	2	1	1	59	71	65
998	0	3	2	1	68	78	77
999	0	3	2	0	77	86	86

1000 rows × 7 columns

11. Find the optimum cluster value and cluster quality using the value of intertia.: Here I find the optimum cluster value 2 for the given data set ,



12. Conclusion: By analyzing this dataset and apply the model I can predict the lunch quality of the student both in supervised & Unsupervised learning. Among three supervised learning decision Tree works better in 90/10 ratio and perform worst in 50/50 train /test split ratio. & KNN works better in 50/50 and worst in 90/10.the accuracy of SVM was not as good as two others but its work almost same in 75/25,50/50& 30/70 ratio.