
DWM EXPERIMENT NO. : 05 (A) **Roll no :50**

Problem definition :

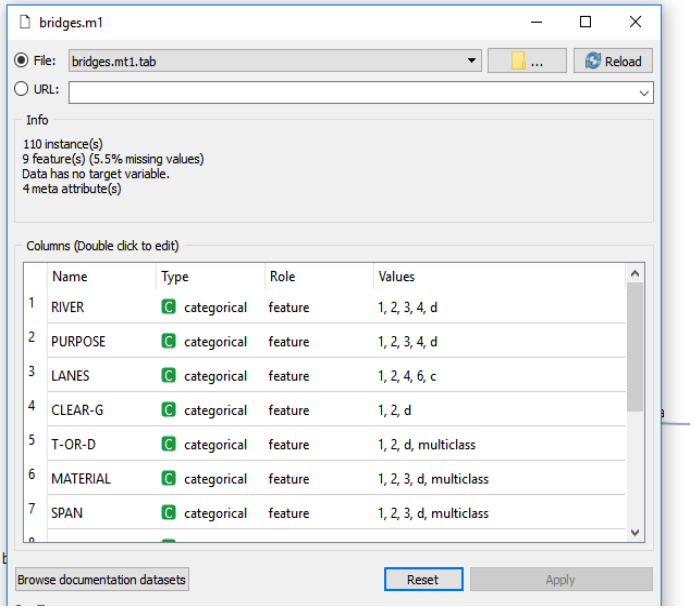
To perform ETL process for building a data warehouse using ORANGE tool.

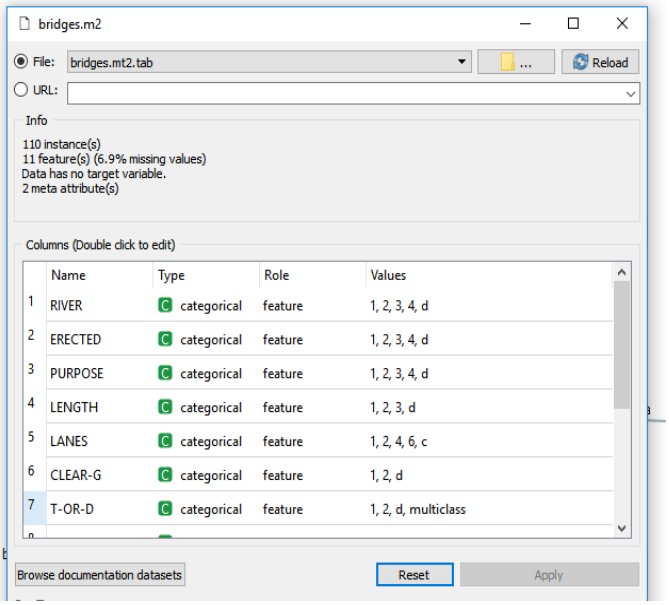
---

1. Consider bridges.mt1.tab and bridges.mt2.tab files as input and do as directed

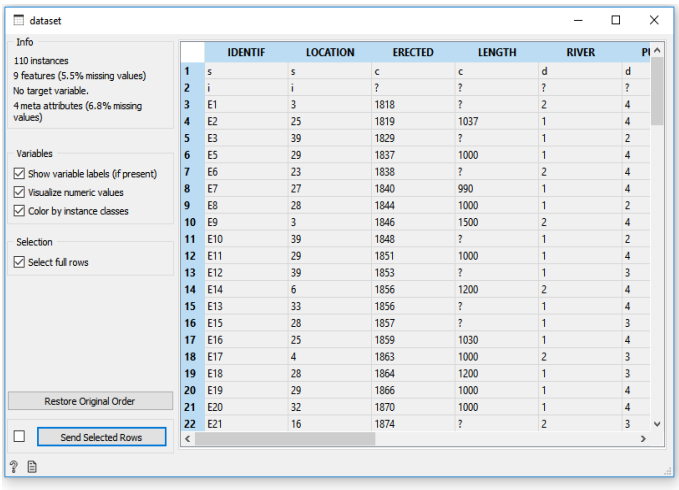
- Concatenate the files.
- Apply preprocessing for missing values
- Save the new file
- Open the new file and view it.

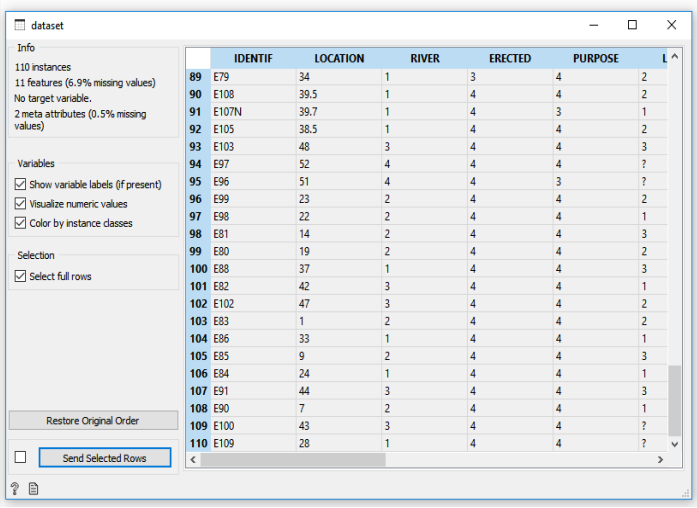
1. Loading the .tab files.



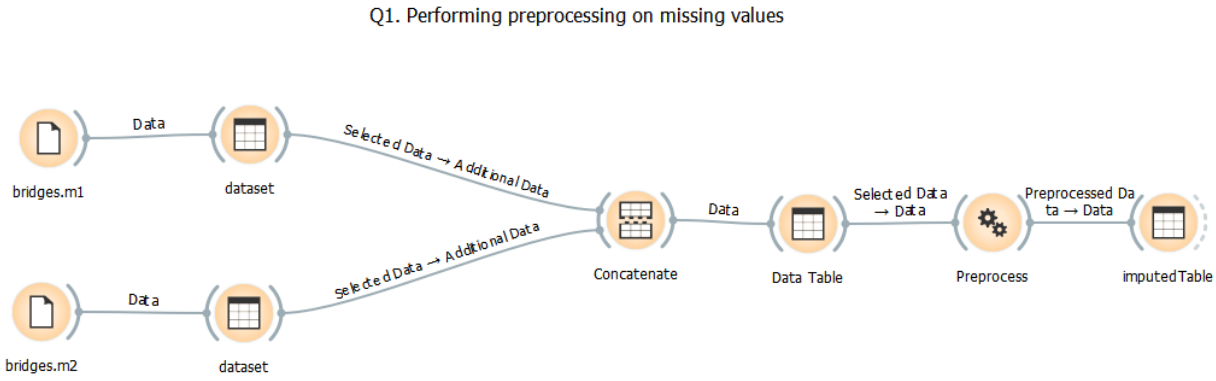


2. Checking datasets of files.

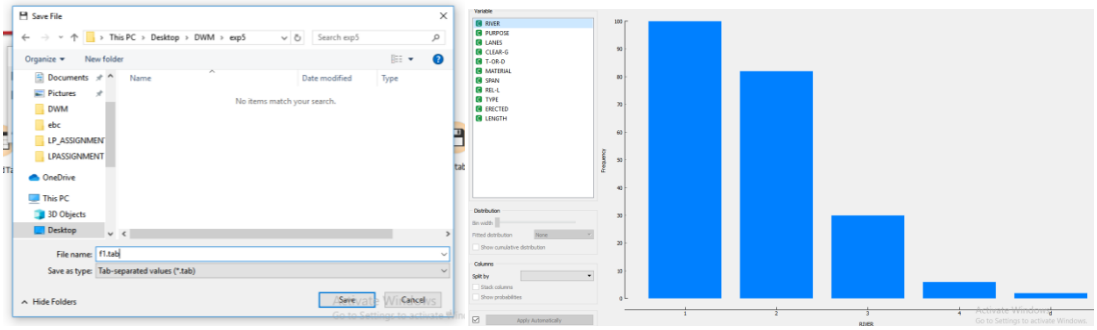




3. Structure of the tools to perform preprocessing.



4. Saving the result to a tab file & visualizing the result set.

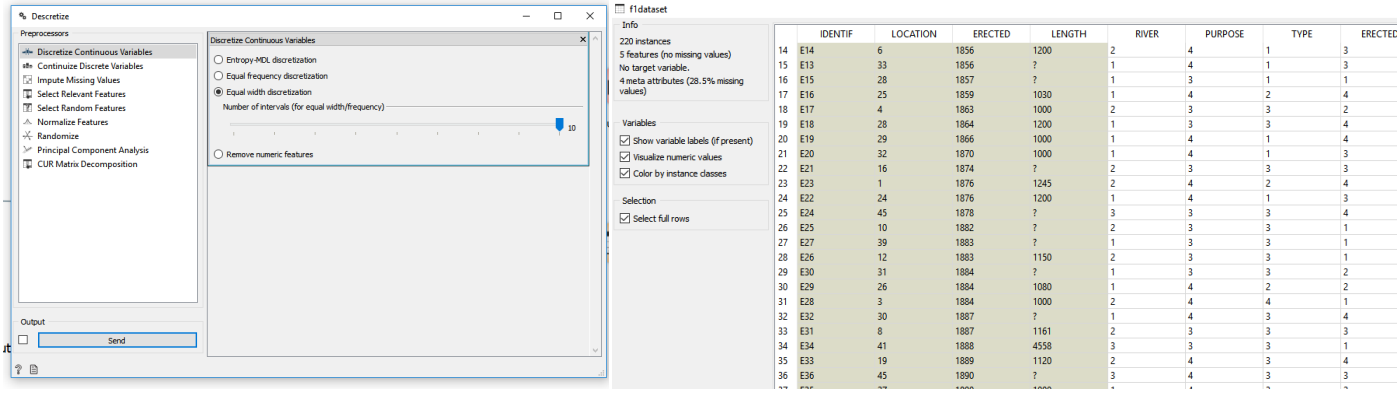


---

2. Use the above file to Concatenate the files.

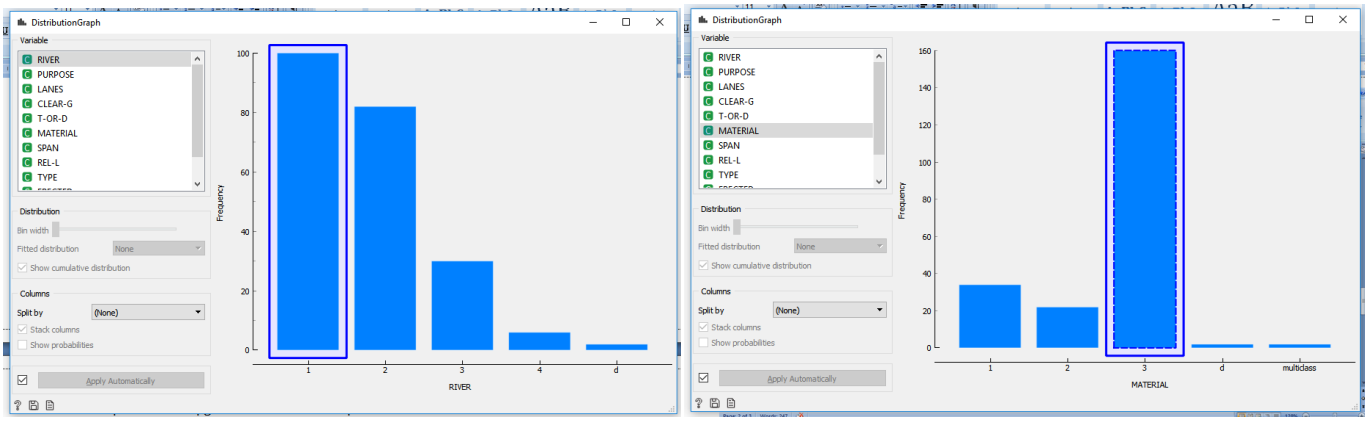
- Discretize the variables and find the type of river that is maximum in number, which type of material is majorly used?
- Consider only erected, type, river, purpose and length columns and save it as file-F1.tab

1. Discretizing the variable.



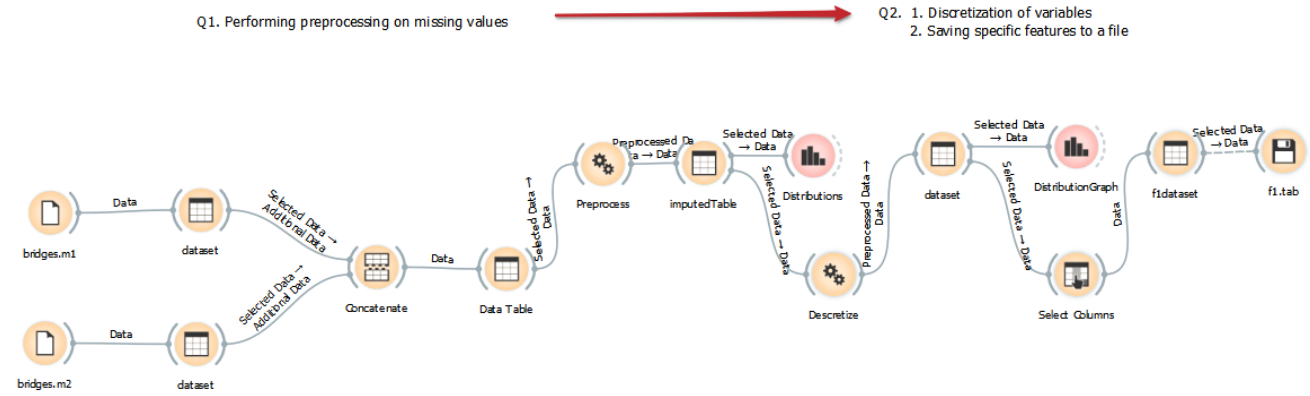
	IDENTIF	LOCATION	ERECTED	LENGTH	RIVER	PURPOSE	TYPE	ERECTED
14	E14	6	1856	1200	2	4	1	3
15	E13	33	1856	?	1	4	1	3
16	E15	28	1857	?	1	3	1	1
17	E16	25	1859	1030	1	4	2	4
18	E17	4	1863	1000	2	3	3	2
19	E18	28	1864	1200	1	3	3	4
20	E19	29	1866	1000	1	4	1	4
21	E20	32	1870	1000	1	4	1	3
22	E21	16	1874	?	2	3	3	3
23	E23	1	1876	1245	2	4	2	4
24	E22	24	1876	1200	1	4	1	3
25	E24	45	1878	?	3	3	3	4
26	E25	10	1882	?	2	3	3	1
27	E27	39	1883	?	1	3	3	1
28	E26	12	1883	1150	2	3	3	1
29	E30	31	1884	?	1	3	3	2
30	E29	26	1884	1080	1	4	2	2
31	E28	3	1884	1000	2	4	4	1
32	E32	30	1887	?	1	4	3	4
33	E31	8	1887	1161	2	3	3	3
34	E34	41	1888	4558	3	3	3	1
35	E33	19	1889	1120	2	4	3	4
36	E36	45	1890	?	3	4	3	3

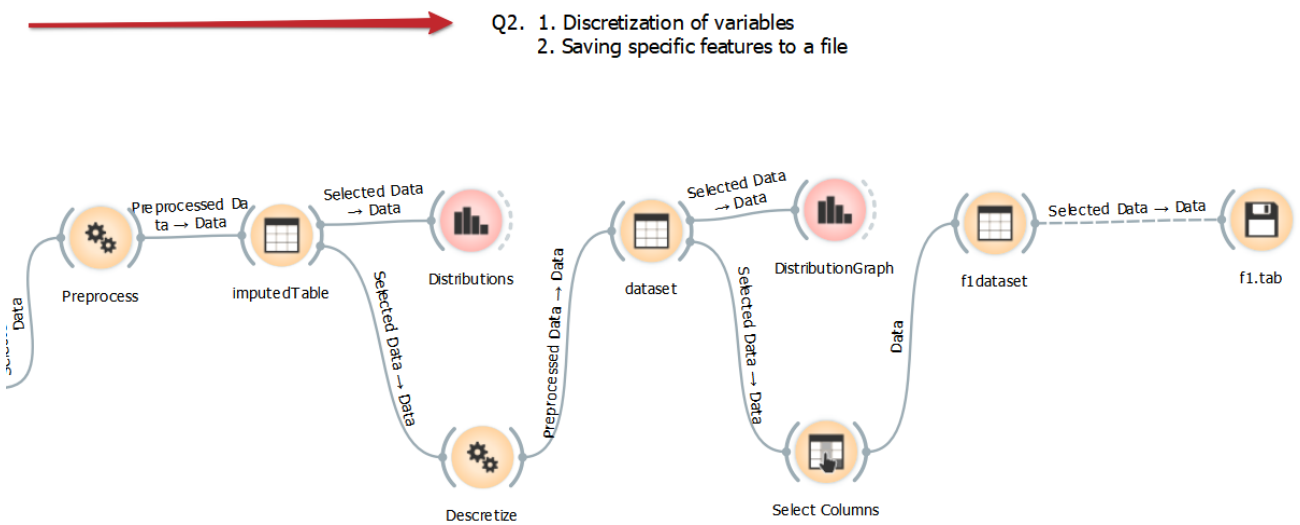
2. Checking the maximum value in result of River and Material feature.



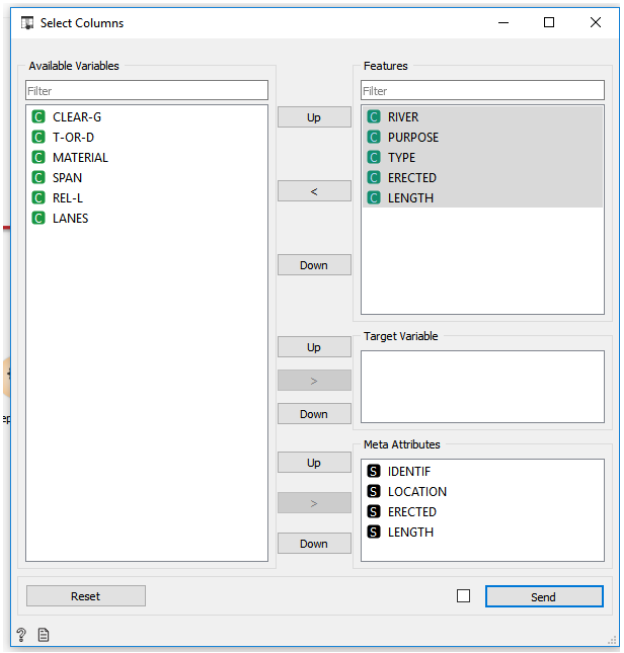
- The type of river that is maximum in number is River “A”.
- Material used majorly is WOOD.

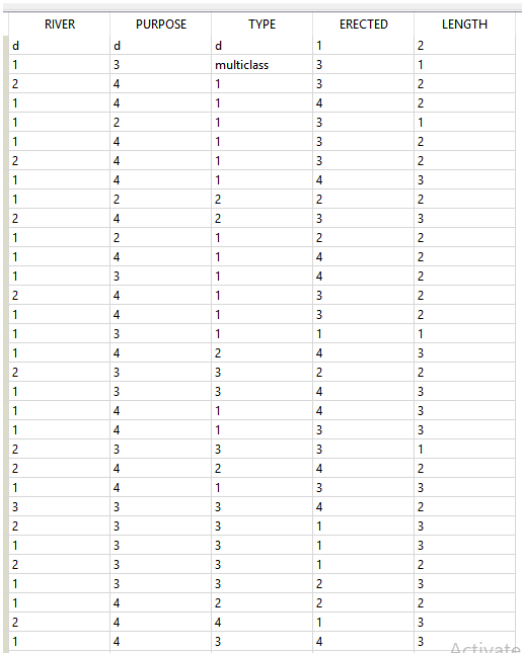
3. Structure of the tools in this case.





4. Choosing the specific features from all the features.





RIVER	PURPOSE	TYPE	ERECTED	LENGTH
d	d	d	1	2
1	3	multiclass	3	1
2	4	1	3	2
1	4	1	4	2
1	2	1	3	1
1	4	1	3	2
2	4	1	3	2
1	4	1	4	3
1	2	2	2	2
2	4	2	3	3
1	2	1	2	2
1	4	1	4	2
1	3	1	4	2
2	4	1	3	2
1	4	1	3	2
1	3	1	1	1
1	4	2	4	3
2	3	3	2	2
1	3	3	4	3
1	4	1	4	3
1	4	1	3	3
2	3	3	3	1
2	4	2	4	2
1	4	1	3	3
3	3	3	4	2
2	3	3	1	3
1	3	3	1	3
2	3	3	1	2
1	3	3	2	3
1	4	2	2	2
2	4	4	1	3
1	4	3	4	3

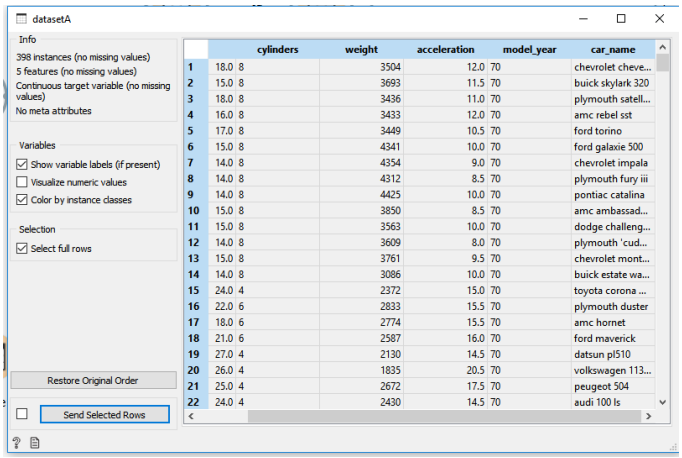
---

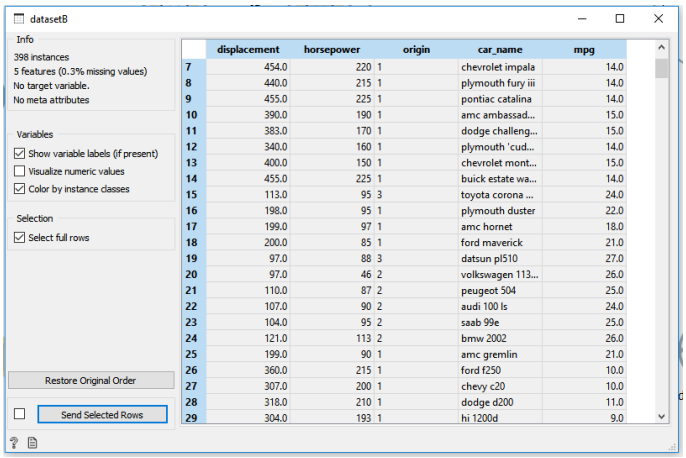
3. Do as directed

- Import auto-mpg.tab file. Make 2 parts of the file columns.
  - cylinders, weight, acceleration, model_year and car_name.
  - mpg, origin, horsepower, displacement and car_name.

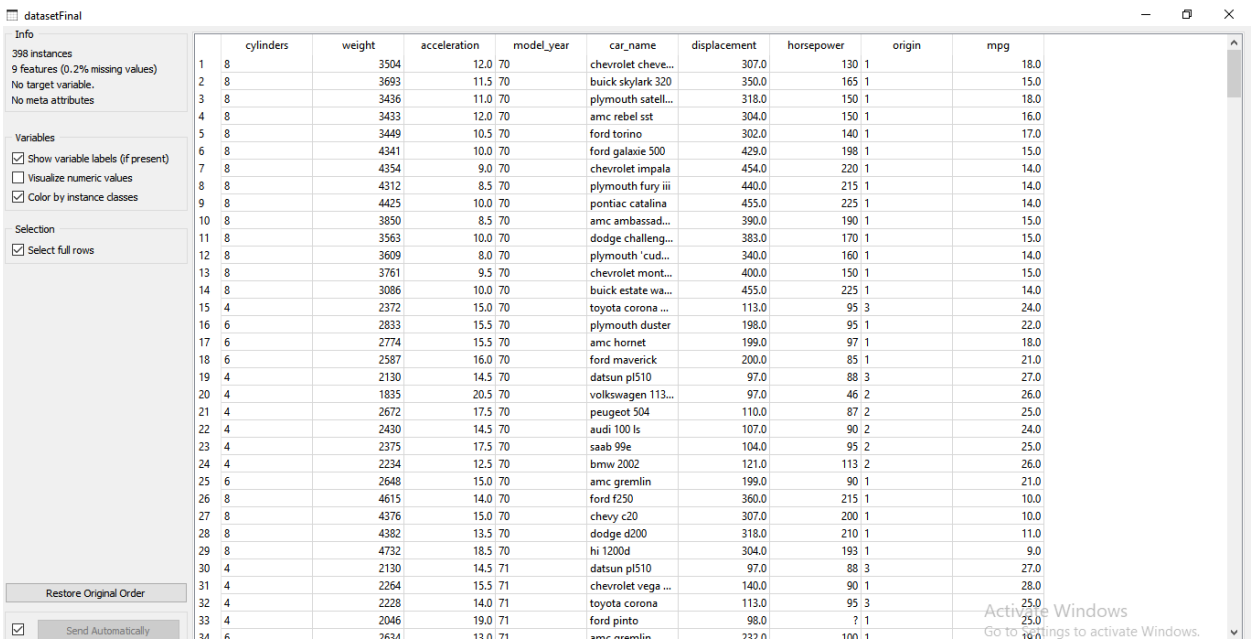
- Consider these 2 parts and merge to get the original file.

1. Selecting specific features from the file.



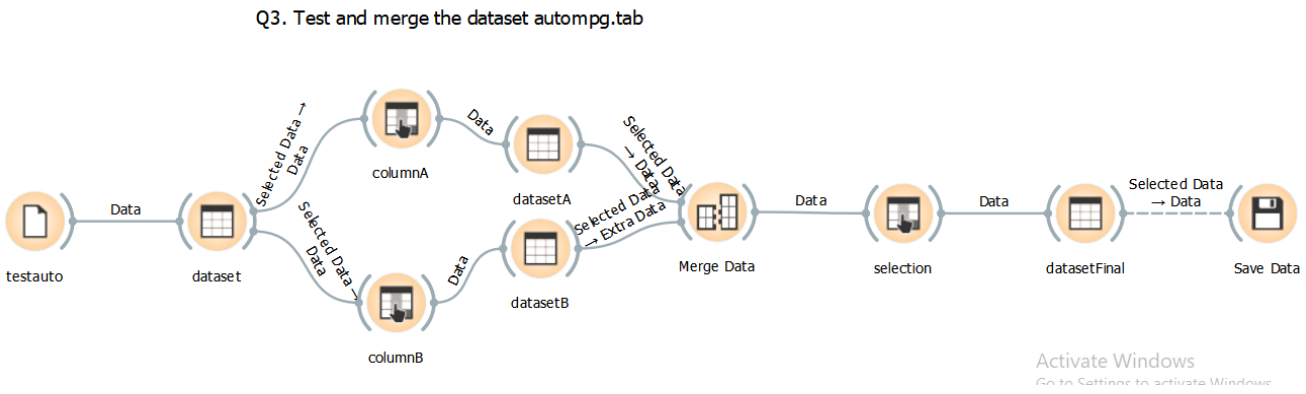


2. Merging the files.



	cylinders	weight	acceleration	model_year	car_name	displacement	horsepower	origin	mpg
1	8	3504	12.0	70	chevrolet cheve...	307.0	130	1	18.0
2	8	3693	11.5	70	buick skylark 320	350.0	165	1	15.0
3	8	3436	11.0	70	plymouth satell...	318.0	150	1	18.0
4	8	3433	12.0	70	amc rebel sst	304.0	150	1	16.0
5	8	3449	10.5	70	ford torino	302.0	140	1	17.0
6	8	4341	10.0	70	ford galaxie 500	429.0	198	1	15.0
7	8	4354	9.0	70	chevrolet impala	454.0	220	1	14.0
8	8	4312	8.5	70	plymouth fury iii	440.0	215	1	14.0
9	8	4425	10.0	70	pontiac catalina	455.0	225	1	14.0
10	8	3850	8.5	70	amc ambassad...	390.0	190	1	15.0
11	8	3563	10.0	70	dodge challeng...	383.0	170	1	15.0
12	8	3609	8.0	70	plymouth 'cud...	340.0	160	1	14.0
13	8	3761	9.5	70	chevrolet mont...	400.0	150	1	15.0
14	8	3086	10.0	70	buick estate wa...	455.0	225	1	14.0
15	4	2372	15.0	70	toyota corona ...	113.0	95	3	24.0
16	6	2833	15.5	70	plymouth duster	198.0	95	1	22.0
17	6	2774	15.5	70	amc hornet	199.0	97	1	18.0
18	6	2587	16.0	70	ford maverick	200.0	85	1	21.0
19	4	2130	14.5	70	datsun pl510	97.0	88	3	27.0
20	4	1835	20.5	70	volkswagen 113...	97.0	46	2	26.0
21	4	2672	17.5	70	peugeot 504	110.0	87	2	25.0
22	4	2430	14.5	70	audi 100 ls	107.0	90	2	24.0
23	4	2375	17.5	70	saab 99e	104.0	95	2	25.0
24	4	2234	12.5	70	bmw 2002	121.0	113	2	26.0
25	6	2648	15.0	70	amc gremlin	199.0	90	1	21.0
26	8	4615	14.0	70	ford f250	360.0	215	1	10.0
27	8	4376	15.0	70	chevy c20	307.0	200	1	10.0
28	8	4382	13.5	70	dodge d200	318.0	210	1	11.0
29	8	4732	18.5	70	hi 1200d	304.0	193	1	9.0
30	4	2130	14.5	71	datsun pl510	97.0	88	3	27.0
31	4	2264	15.5	71	chevrolet vega ...	140.0	90	1	28.0
32	4	2228	14.0	71	toyota corona	113.0	95	3	25.0
33	4	2046	19.0	71	ford pinto	98.0	?	1	25.0

3. Structural arrangement of the tools in this scenario.



Answer the following questions:

1. In which year maximum cars were built?

   Ans. In 1970th year.

2. Write a python script to display first 3 rows of auto-mpg.

   Ans.

```
indexes = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
df = pd.DataFrame(auto-mpg , index=indexes)
```

```
print("First three rows of the data frame:")
print(df.iloc[:3])
```

output

```
a      23.6   4      140.0  ?      2905   14.3   80     1      ford mustang cobra
b      32.4   4      107.0  72     2290   17.0   80     3      honda accord
c      27.2   4      135.0  84     2490   15.7   81     1      plymouth reliant
```

3. Write a python script to display first 3 rows of auto-mpg and to find the maximum displacement, its index and the corresponding row.

Python Script:

```
Script
import Orange
data= Orange.data.Table("auto-mpg")
print(data.domain)
for d in data[:3]:
    print(d)
    p=[1]
for i in range(0,398):
     p.append(data[i]['displacement'])
#for i in range(0,3):
#     print(p)
print('max value is:')
print(max(p))

print('index of max value is:')
i=p.index(max(p))
print(i)
data= Orange.data.Table("auto-mpg")
print(data.domain)
for q in data[i-1]:
     print (q)
```

Output

```
[cylinders, displacement, horsepower, weight, acceleration, model_year,origin,car_name| mpg]
[8, 307.000, 130.000, 3504.000, 12.000, 70, 1, chevrolet chevelle malibu | 18.000]
[8, 350.000, 165.000, 3693.000, 11.500, 70, 1, buick skylark 320 | 15.000]
[8, 318.000, 150.000, 3436.000, 11.000, 70, 1, plymouth satellite | 18.000]
max value is:
455.000
index of max value is:
9
[cylinders, displacement, horsepower, weight, acceleration, model_year, origin, car_name |
mpg]
4.0
455.0
225.0
4425.0
10.0
0.0
```

0.0
241.0
14.0

4. Write a python script to find the maximum displacement, its index and the corresponding row.

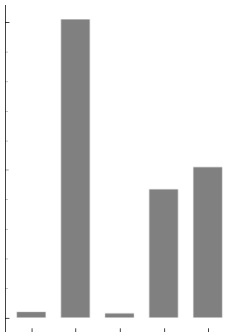
Ans.

```
#to find max displacement and index
import operator
index, value = max(enumerate(auto-mpg), key=operator.itemgetter(1))
```

```
12.0    8       455.0   225     4951    11.0    73      1       buick electra 225 custom
14.0    8       455.0   225     4425    10.0    70      1       pontiac catalina
14.0    8       455.0   225     3086    10.0    70      1       buick estate wagon (sw)
```

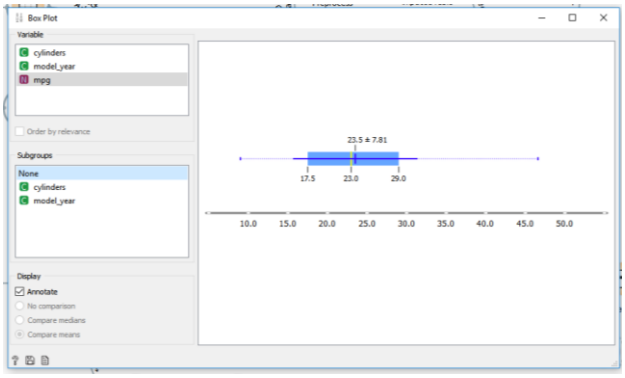
5. Conclude about the number of cylinders. Add grouping of model_year in box plot to make conclusions.

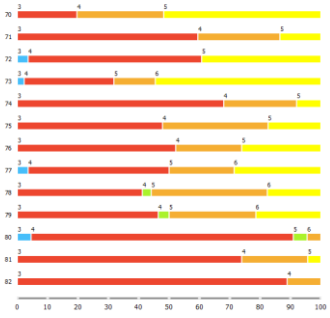
Ans.



It can be seen that for lower Cylinders, the mpg is far great. Hence we should definitely consider performing tests with Cylinder.

6. Show analysis using box plot.





DWM EXPERIMENT NO. : 05 (B) **Roll no :50**

Problem definition :

Introduction to the Weka machine learning toolkit.

---

1. Press the Explorer button on the main panel and load the weather dataset and answer the following questions

- **How many instances are there in the dataset?**
  14
- **State the names of the attributes along with their types and values.**

NAME	TYPE	VALUE
Outlook	Nominal	Sunny, overcast, rainy
Temperature	Numeric	Hot, mild, cool
Humidity	Numeric	High, normal
Windy	Nominal	True, false
Play	Nominal	Yes, no

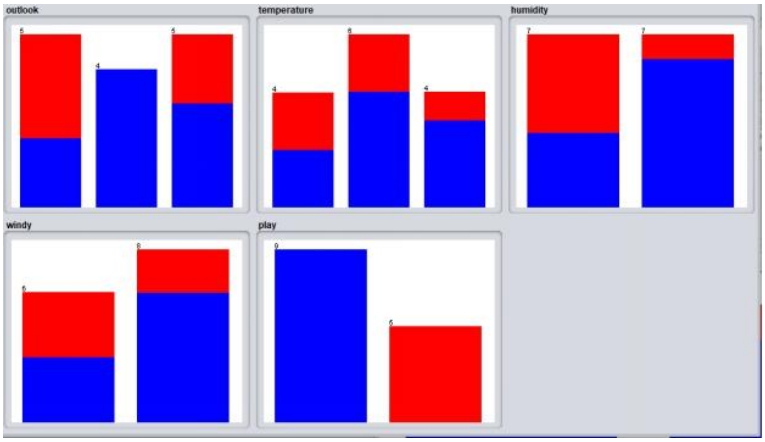
- **What is the class attribute?**
  A class attribute is a target attribute. It is the attribute for which all the hypotheses are validated. In this case, class attribute is **Play**.

- **In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?**
  On X axis, label is marked whereas on Y axis, count is marked.
  We can change the attributes on XY axes by changing the classes.

- **How will you determine how many instances of each class are present in the data?**
  Instance count is equal to the number of distinct values. In this dataset, its 14. We can count the distinct values.

- **What happens with the Visualize All button is pressed?**

  Histogram of all attributes get visible when visualize all is pressed.

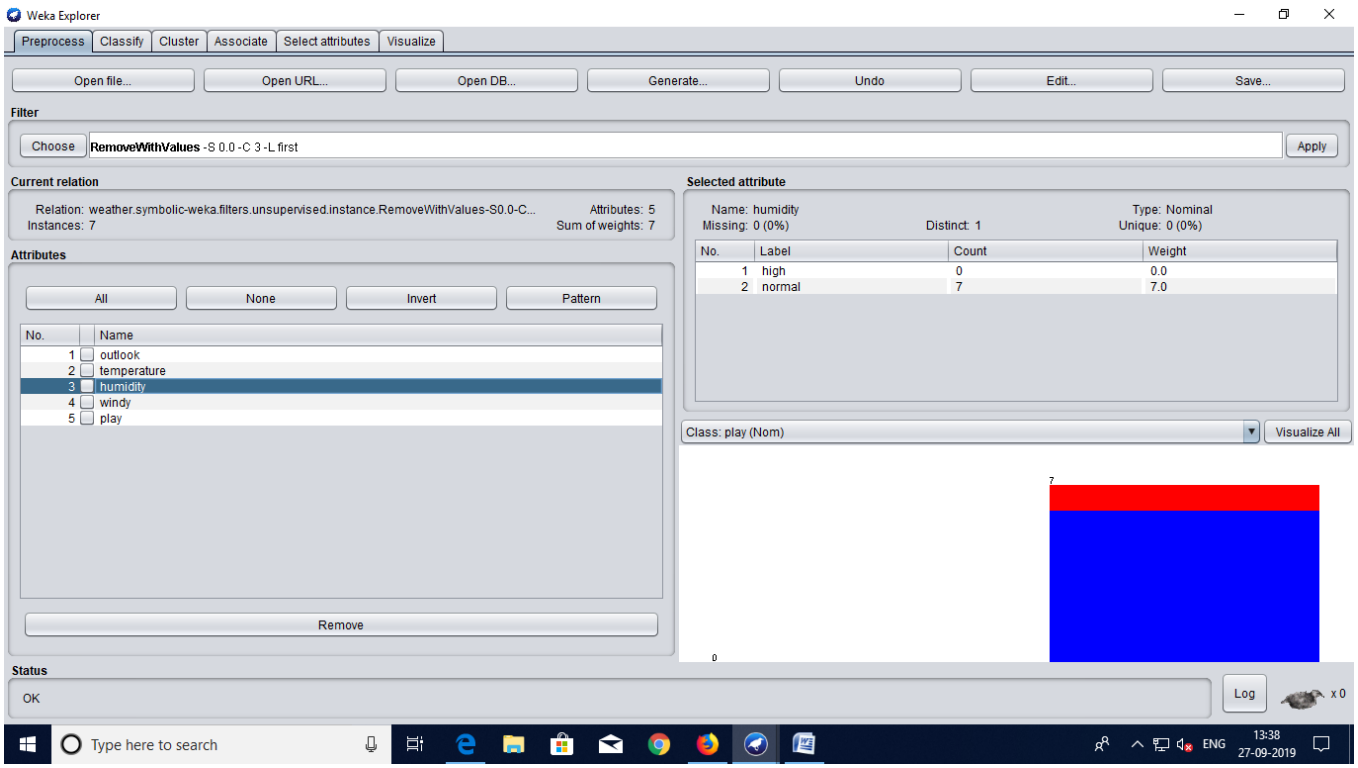


- **How will you view the instances in the dataset? How will you save the changes?** By clicking the edit button, the viewer tab gets opened. Clicking OK saves the changes of active thread.

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

2. Load the weather dataset and perform the following tasks:

1. Use the unsupervised filter RemoveWithValues to remove all instances where the attribute ‘humidity’ has the value ‘high’?



2. Undo the effect of the filter.

Answer the following questions:

- **What is meant by filtering in Weka?**
  Filtering is the process in weka which allows the data engineer to apply some attributes likeable to a criteria as per requirement to make some decision.
- **Which panel is used for filtering a dataset?**
  Filter panel.
- **What are the two main types of filters in Weka?**
  1. Supervised filter.
  2. Unsupervised filter.
- **What is the difference between the two types of filters? What is the difference between and attribute filter and an instance filter?**

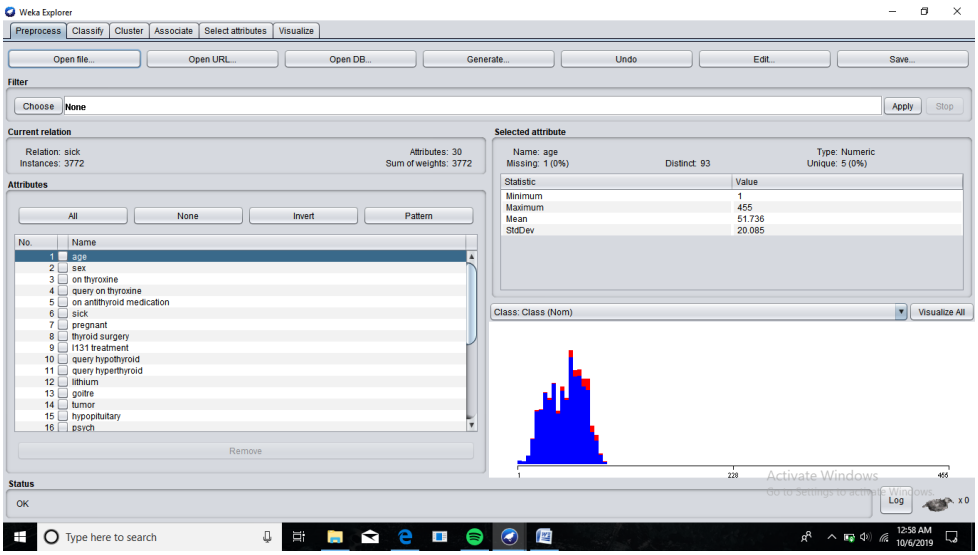
1. Supervised filters in Weka are filters that take the class distribution into account. If the data you are filtering is not classified or you don't want to use the classifications of the data points in the filter process, you'd want an "unsupervised filter".

2. An instance filter that creates a new attribute by applying a mathematical expression to existing attributes. An instance filter that adds an ID attribute to the dataset.A supervised attribute filter that can be used to select attributes. Converts the values of nominal and/or numeric attributes into class conditional probabilities. Changes the order of the classes so that the class values are no longer of in the order specified in the header.

Part A: Application of Discretization Filters

1. Perform the following tasks

1. Load the 'sick.arff' dataset ?



2. **How many instances does this dataset have?**
   3772 instances.

3. **How many attributes does it have?**
   30 attributes.

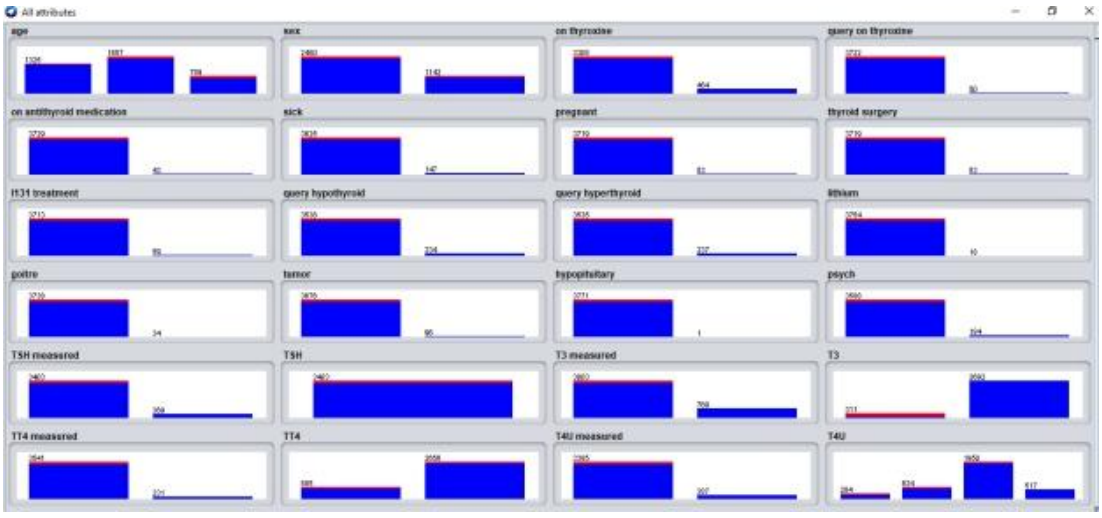
4. **Which is the class attribute and what are the characteristics of this attribute?**
   Health is the target or class attribute.
   Characteristics :
   1. It can contain two values, either sick or negative.
   2. It has a low cardinality.
   3. It has nominal datatype.

5. **How many attributes are numerics? What are the attribute indexes of the numerical attributes?**
   7 numeric attributes in all.
   1, 18,20,22,24,26 and 28 indexed attributes are numerical.

6. **Apply the Naive Bayes classifier. What is the accuracy of the classifier?**
   92.9745% is the accuracy.

2. Perform the following tasks:

1. **Load the 'sick.arff' dataset.**
2. **Apply the supervised discretization filter.**



3. **What is the effect of this filter on the attributes?**
   It discretizes a range of numeric attributes in the dataset into nominal attributes. The main benefit of this is that some classifiers can only take nominal attributes as input, not numeric attributes. Another advantage is that some classifiers that can take numeric attributes can achieve improved accuracy if the data is discretized prior to learning.

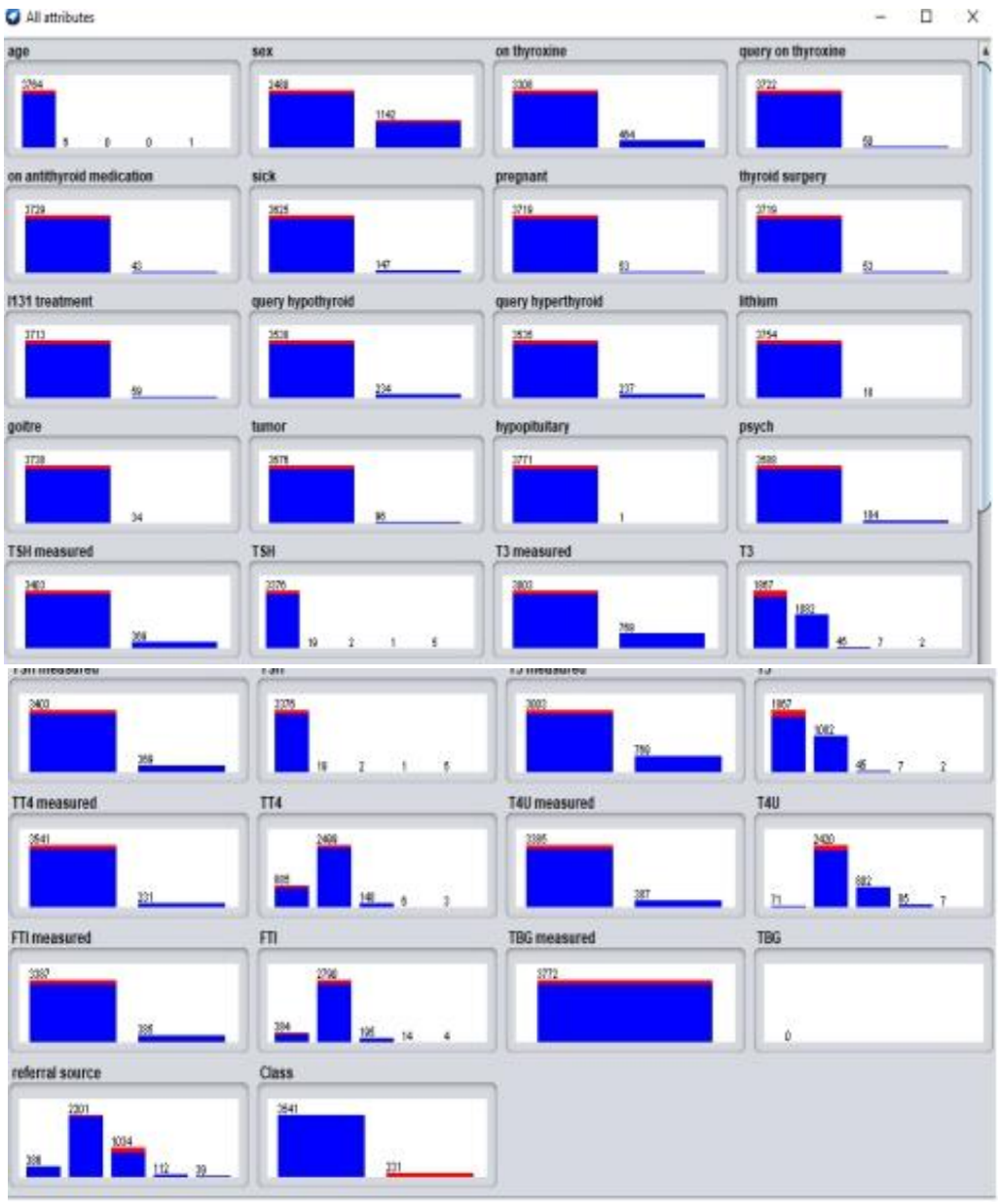
4. How many distinct ranges have been created for each attribute?

Age – 3, TSH-1,T3-2,TT4-2,T4U-4,FTI-1,TBG-1

5. Undo the filter applied in the previous step.

6. Apply the unsupervised discretization filter. Do this twice:

1. **In this step, set 'bins'=5**
2. **In this step, set 'bins'=10**
3. **What is the effect of the unsupervised filter filter on the datset?**



7. Run the the Naive Bayes classifier after apply the following filters

1. **Unsupervised discretized with 'bins'=5**

Correctly Classified Instances	3455	91.596 %
Incorrectly Classified Instances	317	8.404 %

3. Unsupervised discretized with 'bins'=10

Correctly Classified Instances	3654	96.8717 %
Incorrectly Classified Instances	118	3.1283 %

3. Unsupervised discretized with 'bins''=20.

Correctly Classified Instances	3662	97.0838 %
Incorrectly Classified Instances	110	2.9162 %

8. Compare the accuracy of the following cases

1. Naive Bayes without discretization filters

Correctly Classified Instances	3493	92.6034 %
Incorrectly Classified Instances	279	7.3966 %

2. Naive Bayes with a supervised discretization filter

Correctly Classified Instances	3662	97.0838 %
Incorrectly Classified Instances	110	2.9162 %

3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.

1. Unsupervised discretized with 'bins'=5

Correctly Classified Instances	3455	91.596 %
Incorrectly Classified Instances	317	8.404 %

2. Unsupervised discretized with 'bins'=10

Correctly Classified Instances	3654	96.8717 %
Incorrectly Classified Instances	118	3.1283 %

3. Unsupervised discretized with 'bins''=20.

Correctly Classified Instances	3662	97.0838 %
Incorrectly Classified Instances	110	2.9162 %

Part II: Attribute Selection

1. Perform the following tasks:

1. Load the 'mushroom.arff' dataset

2. Run the J48, 1Bk, and the Naive Bayes classifiers.

J48 Correctly Classified Instances 8124 100 %

1Bk Correctly Classified Instances 8124 100 %

Incorrectly Classified Instances 0 0 %

3. What is the accuracy of each of these classifiers?

J48 Correctly Classified Instances 8124 100 %

IBK Correctly Classified Instances 8124 100 %

Incorrectly Classified Instances 0 0 %

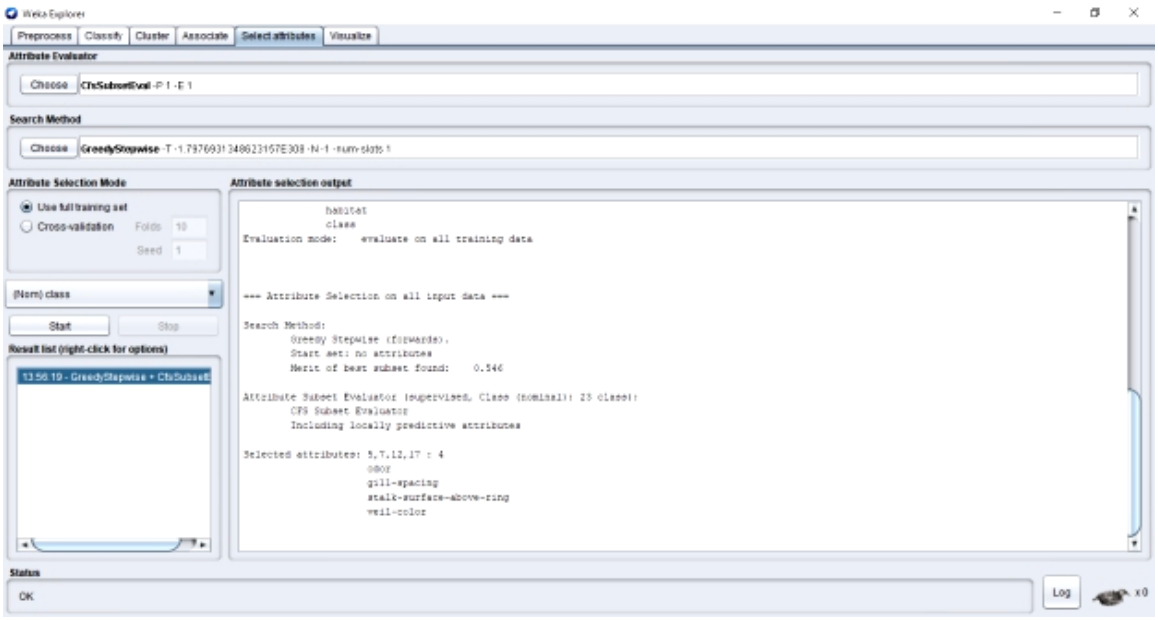
2. Perform the following tasks:

1. Go to the 'Select Attributes' panel

2. Set attribute evaluator to CFSSubsetEval

3. Set the search method to 'Greedy Stepwise'

4. Analyze the results window



5. Record the attribute numbers of the most important attributes

5,7,12,17

- odor
- gill-spacing
- stalk-surface-above-ring
- veil-color

6. Run the meta classifier AttributeSelectedClassifier using the following:

1. CFSSubsetEval

2. GreedStepwise

3. J48, 1Bk, and NaiveBayes

Results are same as above

7. Record the accuracy of the classifiers

8. What are the benefits of attribute selection?

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modelling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

Part C

1. Perform the following tasks:

1. Load the 'vote.arff' dataset.
2. Run the J48, 1Bk, and Naive Bayes classifiers.
3. Record the accuracies.

J48

Correctly Classified Instances	419	96.3218 %
Incorrectly Classified Instances	16	3.6782 %

1Bk

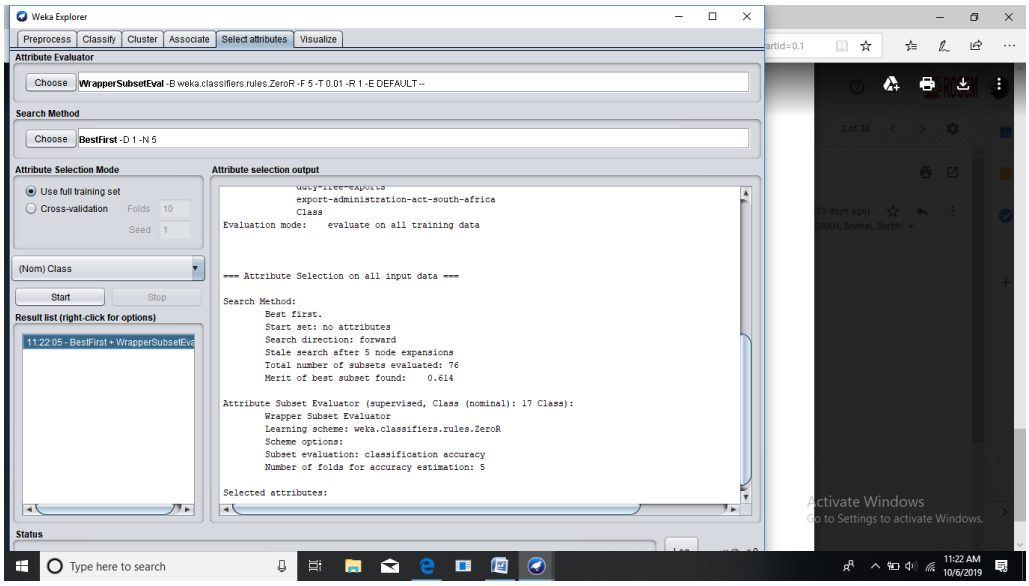
Correctly Classified Instances	434	99.7701 %
Incorrectly Classified Instances	1	0.2299 %

Naïve Bayes

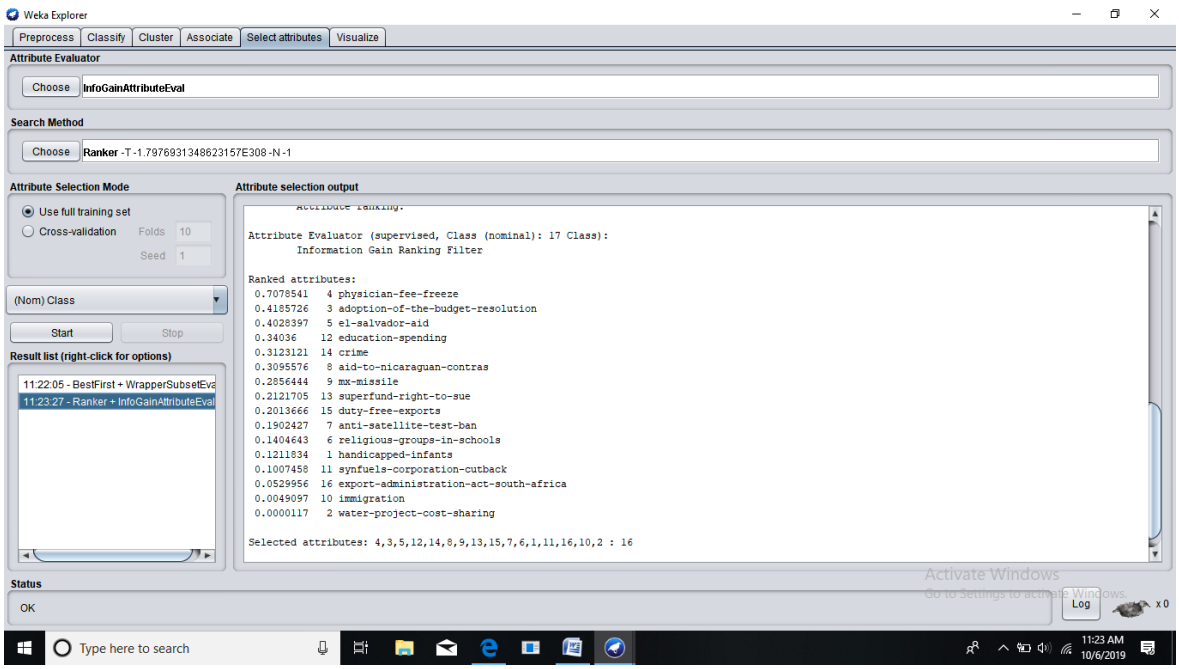
Correctly Classified Instances	393	90.3448 %
Incorrectly Classified Instances	42	9.6552 %

2. Perform the following tasks:

1. Go to the 'Select Attributes' panel
2. Set attribute evaluator to 'WrapperSubsetEval



3. Set search method to ''RankSearch'
4. Set attribute evaluator to 'InfoGainAttributeEval'
5. Analyze the results



6. Run the metaclassifier AttributeSelectedClassifier using the following:

1. WrapperSubsetEval

Correctly Classified Instances	416	95.6322 %
Incorrectly Classified Instances	19	4.3678 %

2. RankSearch

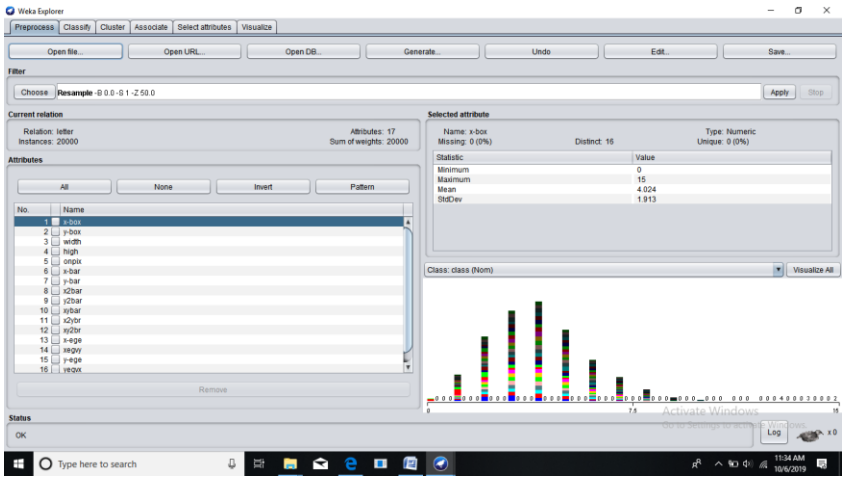
Correctly Classified Instances	416	95.6322 %
Incorrectly Classified Instances	19	4.3678 %

3. InfoGainAttributeEval

Correctly Classified Instances	416	95.6322 %
Incorrectly Classified Instances	19	4.3678 %

7. Sampling

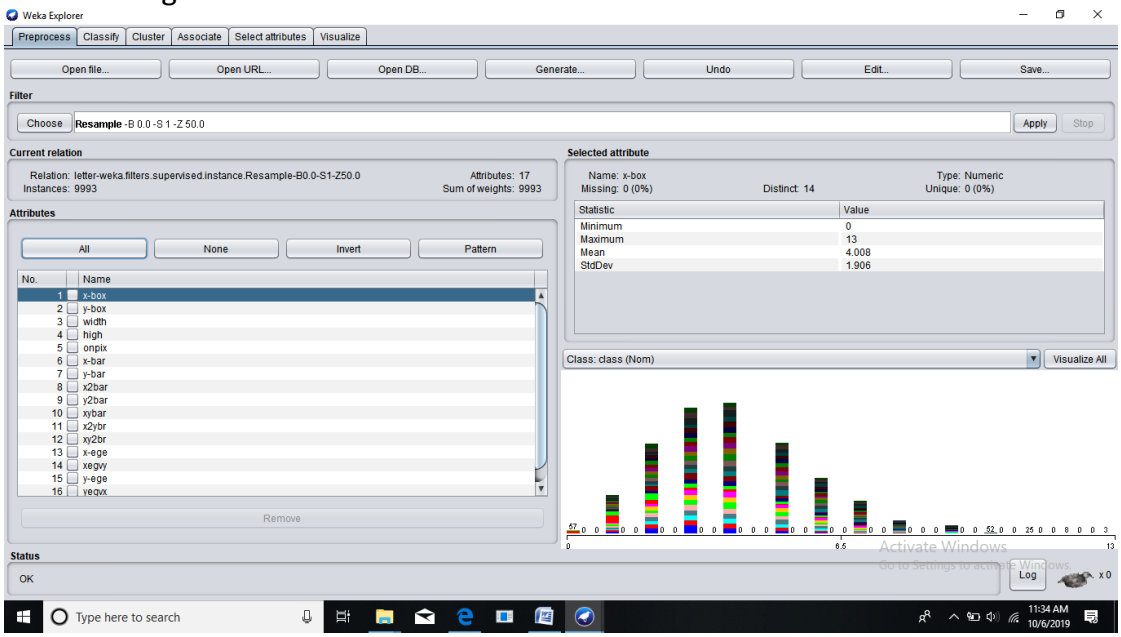
1. **Load the 'letter.arff' dataset**
2. **Take any attribute and record the min, max, mean, and standard deviation of the attribute**



3. **Apply the Resample filter with 'sampleSizePercent' set to 50 percent**

4. **What is the size of the filtered dataset. Observe the min, max, mean, and standard deviation Of the attribute that was selected in step 2. What is the percentage change in the values?**

0.36% change resulted in Standard deviation values.



4. **Give the benefit of sampling a large dataset.**

Sampling can be particularly useful with data sets that are too large to efficiently analyze in full, for example, in data science applications or surveys. Identifying and analyzing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population.