DATA MINING- PROJECT

Submitted to: DR. MD. MAHBUB CHOWDHURY MISHU

Submitted By: MD. SAMIUL HAUQUE CHOWDHURY

MARCH 10, 2021

Section: A

Project Title: Application of KNN to a data-set on weka and analyzing the accuracy by applying different algorithms.

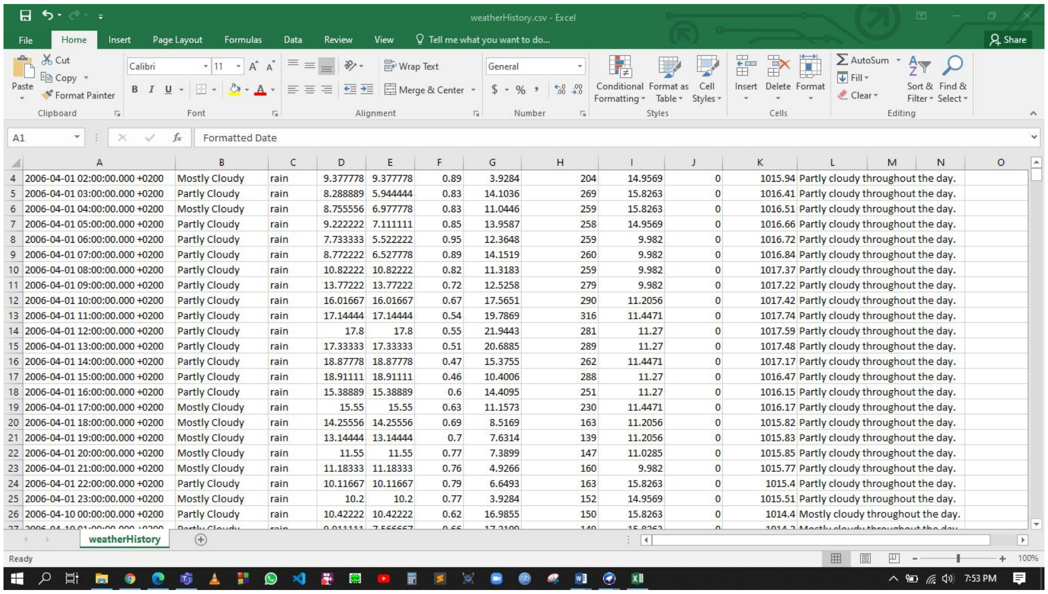
Introduction:

One of the most common data mining concept is classification techniques. The objective of classification is to specifically predict the target class for each case in the data. In this project, I used a widely utilized classification procedure called **"K-nearest neighbors (KNN)"** which is also known as lazy learner algorithm. By utilizing K-NN, it will recognize the category or class of a specific dataset. In this project, I used a weather data set and name of the data set is therHistory.csv**".** It has some data records of an area that recorded temperature, apparent temperature, humidity, wind speed, wind bearing, visibility, and pressure type data and by analyzing we need to predict next day weather.

Methods of KNN:

1. Assigning the value of K in KNN K=1, 3, 5... (I used K=1)
2. Now Applying Euclidean distance d= sqrt(x2-x1)^2-(y2-y1)^2
3. Now we need to sort the values of distances.
4. Now we need to assign the new data points to that category for which number of the data the number of neighbor is maximum.
5. Among these k neighbors, we need to count the points in each category.

DataSet:



A1 | Formatted Date

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
4	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204	14.9569	0	1015.94	Partly cloudy throughout the day.			
5	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269	15.8263	0	1016.41	Partly cloudy throughout the day.			
6	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259	15.8263	0	1016.51	Partly cloudy throughout the day.			
7	2006-04-01 05:00:00.000 +0200	Partly Cloudy	rain	9.222222	7.111111	0.85	13.9587	258	14.9569	0	1016.66	Partly cloudy throughout the day.			
8	2006-04-01 06:00:00.000 +0200	Partly Cloudy	rain	7.733333	5.522222	0.95	12.3648	259	9.982	0	1016.72	Partly cloudy throughout the day.			
9	2006-04-01 07:00:00.000 +0200	Partly Cloudy	rain	8.772222	6.527778	0.89	14.1519	260	9.982	0	1016.84	Partly cloudy throughout the day.			
10	2006-04-01 08:00:00.000 +0200	Partly Cloudy	rain	10.82222	10.82222	0.82	11.3183	259	9.982	0	1017.37	Partly cloudy throughout the day.			
11	2006-04-01 09:00:00.000 +0200	Partly Cloudy	rain	13.77222	13.77222	0.72	12.5258	279	9.982	0	1017.22	Partly cloudy throughout the day.			
12	2006-04-01 10:00:00.000 +0200	Partly Cloudy	rain	16.01667	16.01667	0.67	17.5651	290	11.2056	0	1017.42	Partly cloudy throughout the day.			
13	2006-04-01 11:00:00.000 +0200	Partly Cloudy	rain	17.14444	17.14444	0.54	19.7869	316	11.4471	0	1017.74	Partly cloudy throughout the day.			
14	2006-04-01 12:00:00.000 +0200	Partly Cloudy	rain	17.8	17.8	0.55	21.9443	281	11.27	0	1017.59	Partly cloudy throughout the day.			
15	2006-04-01 13:00:00.000 +0200	Partly Cloudy	rain	17.33333	17.33333	0.51	20.6885	289	11.27	0	1017.48	Partly cloudy throughout the day.			
16	2006-04-01 14:00:00.000 +0200	Partly Cloudy	rain	18.87778	18.87778	0.47	15.3755	262	11.4471	0	1017.17	Partly cloudy throughout the day.			
17	2006-04-01 15:00:00.000 +0200	Partly Cloudy	rain	18.91111	18.91111	0.46	10.4006	288	11.27	0	1016.47	Partly cloudy throughout the day.			
18	2006-04-01 16:00:00.000 +0200	Partly Cloudy	rain	15.38889	15.38889	0.6	14.4095	251	11.27	0	1016.15	Partly cloudy throughout the day.			
19	2006-04-01 17:00:00.000 +0200	Mostly Cloudy	rain	15.55	15.55	0.63	11.1573	230	11.4471	0	1016.17	Partly cloudy throughout the day.			
20	2006-04-01 18:00:00.000 +0200	Mostly Cloudy	rain	14.25556	14.25556	0.69	8.5169	163	11.2056	0	1015.82	Partly cloudy throughout the day.			
21	2006-04-01 19:00:00.000 +0200	Mostly Cloudy	rain	13.14444	13.14444	0.7	7.6314	139	11.2056	0	1015.83	Partly cloudy throughout the day.			
22	2006-04-01 20:00:00.000 +0200	Mostly Cloudy	rain	11.55	11.55	0.77	7.3899	147	11.0285	0	1015.85	Partly cloudy throughout the day.			
23	2006-04-01 21:00:00.000 +0200	Mostly Cloudy	rain	11.18333	11.18333	0.76	4.9266	160	9.982	0	1015.77	Partly cloudy throughout the day.			
24	2006-04-01 22:00:00.000 +0200	Partly Cloudy	rain	10.11667	10.11667	0.79	6.6493	163	15.8263	0	1015.4	Partly cloudy throughout the day.			
25	2006-04-01 23:00:00.000 +0200	Mostly Cloudy	rain	10.2	10.2	0.77	3.9284	152	14.9569	0	1015.51	Partly cloudy throughout the day.			
26	2006-04-10 00:00:00.000 +0200	Partly Cloudy	rain	10.42222	10.42222	0.62	16.9855	150	15.8263	0	1014.4	Mostly cloudy throughout the day.			

Dataset Description:

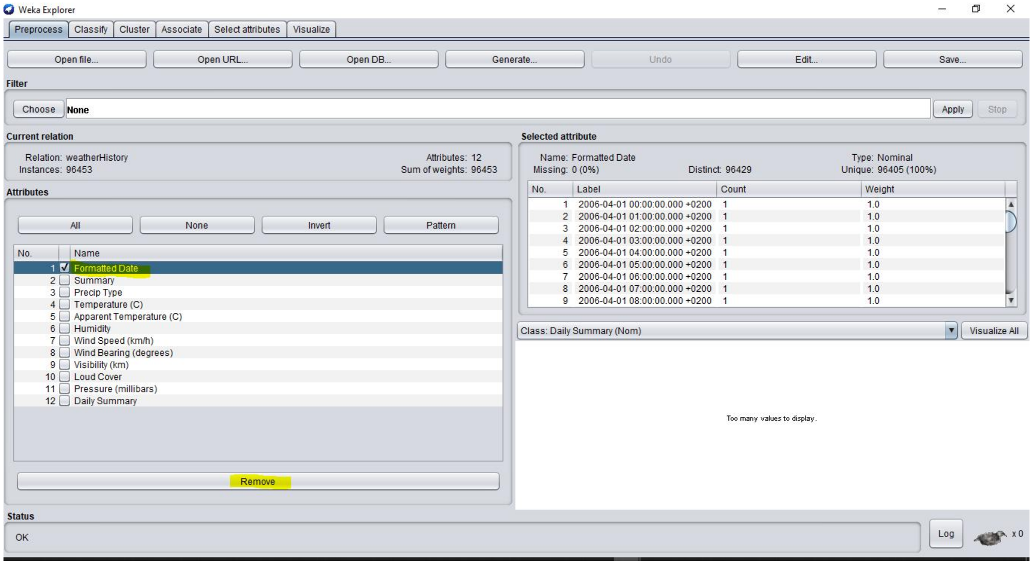
Name of this dataset is “weatherHistory.csv”. This dataset contain 12 attributes and 96453 instances. It has 12 columns and 96454 rows.

Procedure in weka:

First I need to upload my dataset to weka from the open file section. Now next procedure is to delete less important attribute to increase significant of accuracy.

1. Deleting Unnecessary attributes:

Here “formatted date” is an unnecessary attribute that has no significant relation with the entire dataset. That’s why I have deleted this attribute by remove function.



Filter

None

Current relation

Relation: weatherHistory
Instances: 96453
Attributes: 12
Sum of weights: 96453

Attributes

No.		Name
1	☑	Formatted Date
2	□	Summary
3	□	Precip Type
4	□	Temperature (C)
5	□	Apparent Temperature (C)
6	□	Humidity
7	□	Wind Speed (km/h)
8	□	Wind Bearing (degrees)
9	□	Visibility (km)
10	□	Loud Cover
11	□	Pressure (millibars)
12	□	Daily Summary

Selected attribute

Name: Formatted Date
Missing: 0 (0%)
Distinct: 96429
Type: Nominal
Unique: 96405 (100%)

No.	Label	Count	Weight
1	2006-04-01 00:00:00.000 +0200	1	1.0
2	2006-04-01 01:00:00.000 +0200	1	1.0
3	2006-04-01 02:00:00.000 +0200	1	1.0
4	2006-04-01 03:00:00.000 +0200	1	1.0
5	2006-04-01 04:00:00.000 +0200	1	1.0
6	2006-04-01 05:00:00.000 +0200	1	1.0
7	2006-04-01 06:00:00.000 +0200	1	1.0
8	2006-04-01 07:00:00.000 +0200	1	1.0
9	2006-04-01 08:00:00.000 +0200	1	1.0

Class: Daily Summary (Nom)

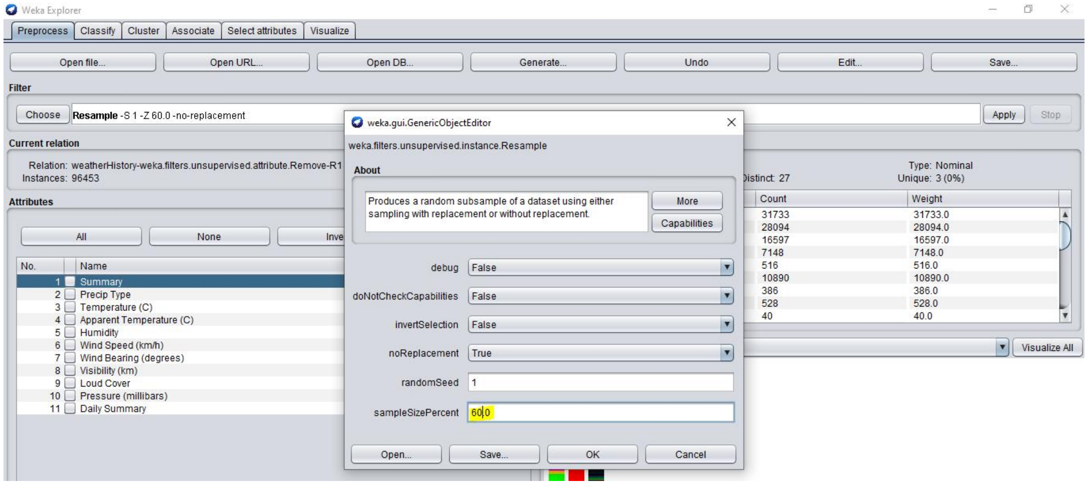
Too many values to display.

Status

OK

2. Splitting data set:

Now we need to divide data into two set one is training set another is test set.



Filter: Resample -S 1 -Z 60.0 -no-replacement

Current relation
Relation: weatherHistory-weka.filters.unsupervised.attribute.Remove-R1
Instances: 96453

Distinct: 27 Type: Nominal
Unique: 3 (0%)

No.	Name
1	Summary
2	Precip Type
3	Temperature (C)
4	Apparent Temperature (C)
5	Humidity
6	Wind Speed (km/h)
7	Wind Bearing (degrees)
8	Visibility (km)
9	Loud Cover
10	Pressure (millibars)
11	Daily Summary

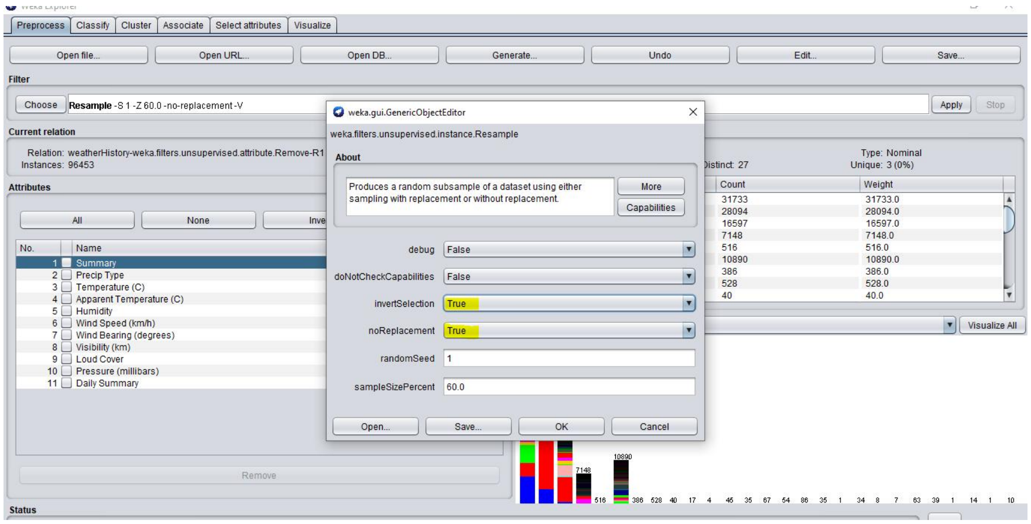
Count	Weight
31733	31733.0
28094	28094.0
16597	16597.0
7148	7148.0
516	516.0
10890	10890.0
386	386.0
528	528.0
40	40.0

weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.Resample
About
Produces a random subsample of a dataset using either sampling with replacement or without replacement.

debug	False
doNotCheckCapabilities	False
invertSelection	False
noReplacement	True
randomSeed	1
sampleSizePercent	60.0

Here first one is we have divided into training set by taking 60% of data set.

And the next picture we have saved for test set by taking the 40% of the rest.



Filter: Resample -S 1 -Z 60.0 -no-replacement -V

Current relation
Relation: weatherHistory-weka.filters.unsupervised.attribute.Remove-R1
Instances: 96453

Distinct: 27 Type: Nominal
Unique: 3 (0%)

No.	Name
1	Summary
2	Precip Type
3	Temperature (C)
4	Apparent Temperature (C)
5	Humidity
6	Wind Speed (km/h)
7	Wind Beaing (degrees)
8	Visibility (km)
9	Loud Cover
10	Pressure (millibars)
11	Daily Summary

Count	Weight
31733	31733.0
28094	28094.0
16597	16597.0
7148	7148.0
516	516.0
10890	10890.0
386	386.0
528	528.0
40	40.0

weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.Resample
About
Produces a random subsample of a dataset using either sampling with replacement or without replacement.

debug	False
doNotCheckCapabilities	False
invertSelection	True
noReplacement	True
randomSeed	1
sampleSizePercent	60.0

3. Applying Algorithms:

Now we need to apply KNN, Naïve Bayes and decision tree algorithms to our train set and test set.

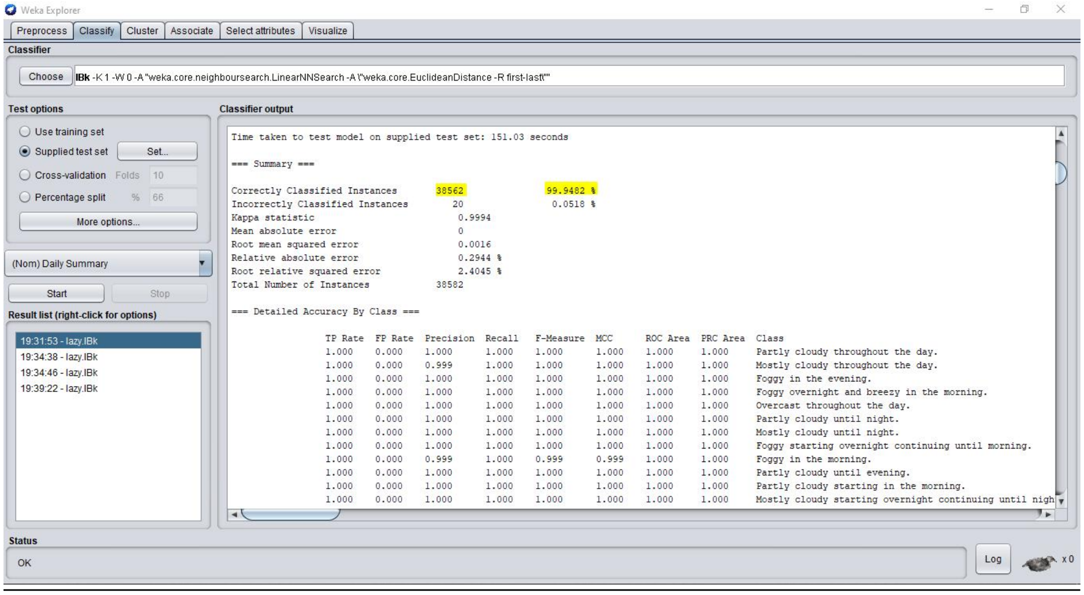
For KNN: classify>choose>lazy>IBk

For Naïve Bayes: classify>choose> bayes >NaiveBayes

For KNN: classify>choose>trees>j48

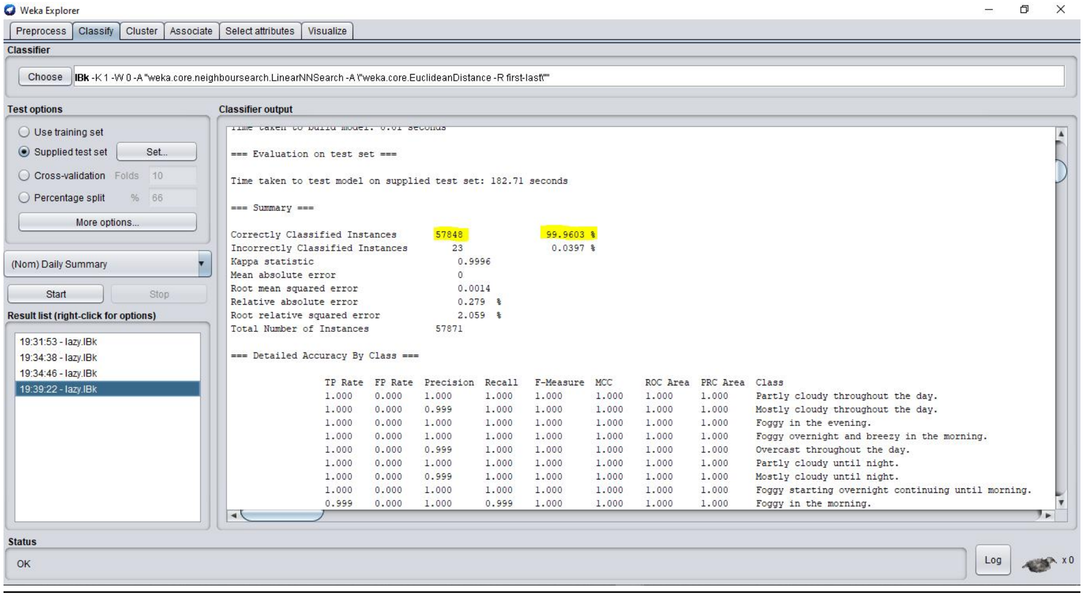
Here are some screenshots of test and train set accuracy result in different algorithms.

KNN (Test set and Train set):



```
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
```

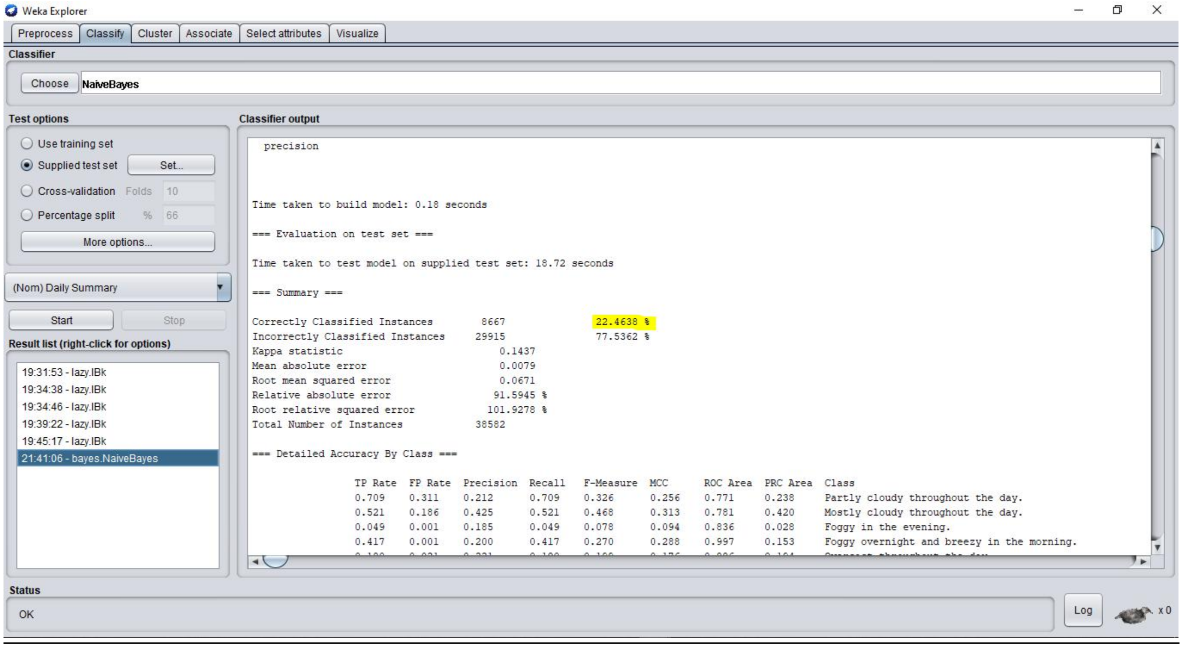
```
Time taken to test model on supplied test set: 151.03 seconds

=== Summary ===

Correctly Classified Instances       38562               99.9482 %
Incorrectly Classified Instances        20                0.0518 %
Kappa statistic                          0.9994
Mean absolute error                      0
Root mean squared error                  0.0016
Relative absolute error                  0.2944 %
Root relative squared error              2.4045 %
Total Number of Instances            38582

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Partly cloudy throughout the day.
                 1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     Mostly cloudy throughout the day.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Foggy in the evening.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Foggy overnight and breezy in the morning.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Overcast throughout the day.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Partly cloudy until night.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Mostly cloudy until night.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Foggy starting overnight continuing until morning.
                 1.000    0.000    0.999      1.000    0.999      0.999    1.000     1.000     Foggy in the morning.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Partly cloudy until evening.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Partly cloudy starting in the morning.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Mostly cloudy starting overnight continuing until nigh
```



```
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
```

```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 182.71 seconds

=== Summary ===

Correctly Classified Instances       57848               99.9603 %
Incorrectly Classified Instances        23                0.0397 %
Kappa statistic                          0.9996
Mean absolute error                      0
Root mean squared error                  0.0014
Relative absolute error                  0.279  %
Root relative squared error              2.059  %
Total Number of Instances            57871

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Partly cloudy throughout the day.
                 1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     Mostly cloudy throughout the day.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Foggy in the evening.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Foggy overnight and breezy in the morning.
                 1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     Overcast throughout the day.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Partly cloudy until night.
                 1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     Mostly cloudy until night.
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Foggy starting overnight continuing until morning.
                 0.999    0.000    1.000      0.999    1.000      1.000    1.000     1.000     Foggy in the morning.
```

Naïve Bayes (train set and test set):



Classifier: NaiveBayes

Test options:
○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66
More options...

(Nom) Daily Summary

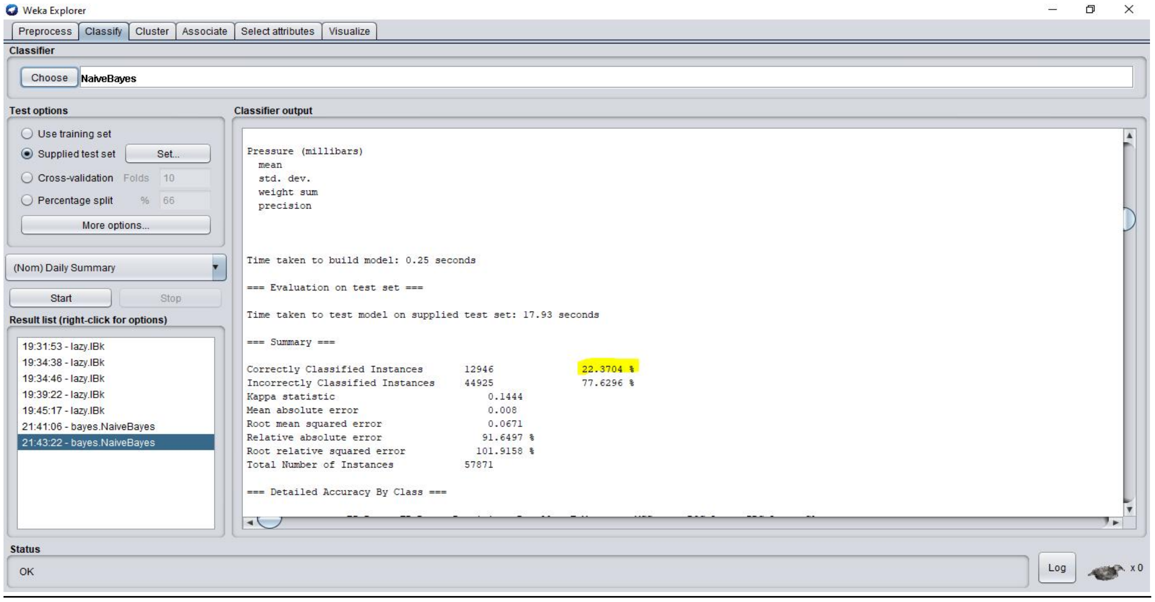
Result list (right-click for options)

```
19:31:53 - lazy.IBk
19:34:38 - lazy.IBk
19:34:46 - lazy.IBk
19:39:22 - lazy.IBk
19:45:17 - lazy.IBk
21:41:06 - bayes.NaiveBayes
```

```
  precision



Time taken to build model: 0.18 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 18.72 seconds

=== Summary ===

Correctly Classified Instances        8667               22.4638 %
Incorrectly Classified Instances     29915               77.5362 %
Kappa statistic                          0.1437
Mean absolute error                      0.0079
Root mean squared error                  0.0671
Relative absolute error                 91.5945 %
Root relative squared error            101.9278 %
Total Number of Instances            38582

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.709    0.311    0.212      0.709    0.326      0.256    0.771     0.238     Partly cloudy throughout the day.
                 0.521    0.186    0.425      0.521    0.468      0.313    0.781     0.420     Mostly cloudy throughout the day.
                 0.049    0.001    0.185      0.049    0.078      0.094    0.836     0.028     Foggy in the evening.
                 0.417    0.001    0.200      0.417    0.270      0.288    0.997     0.153     Foggy overnight and breezy in the morning.
```

Status

OK



Classifier: NaiveBayes

Test options:
○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66
More options...

(Nom) Daily Summary

Result list (right-click for options)

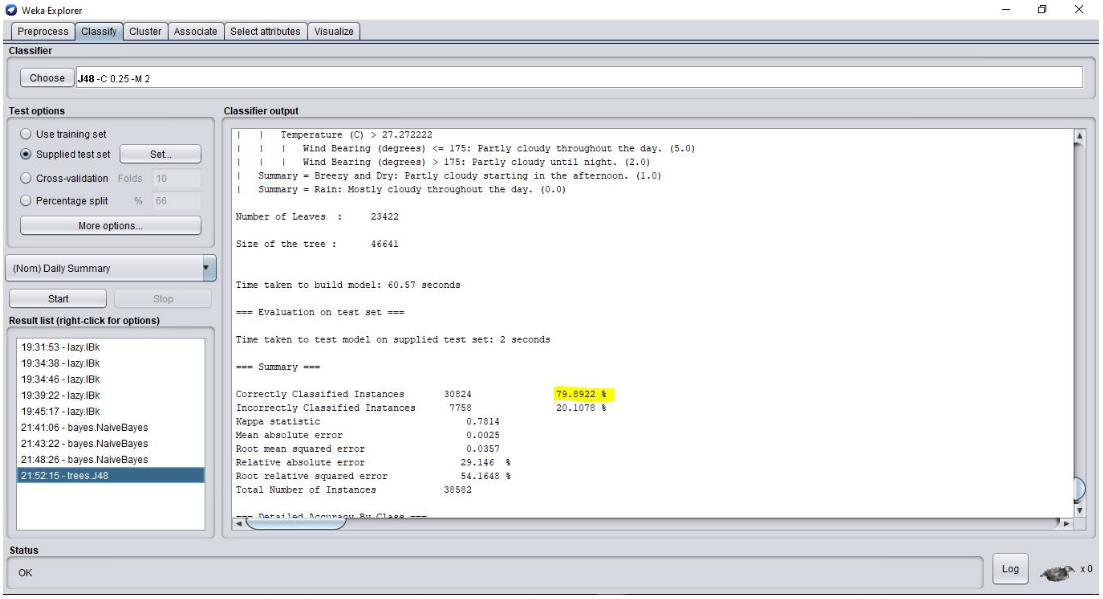
```
19:31:53 - lazy.IBk
19:34:38 - lazy.IBk
19:34:46 - lazy.IBk
19:39:22 - lazy.IBk
19:45:17 - lazy.IBk
21:41:06 - bayes.NaiveBayes
21:43:22 - bayes.NaiveBayes
```

```

Pressure (millibars)
  mean
  std. dev.
  weight sum
  precision



Time taken to build model: 0.25 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 17.93 seconds

=== Summary ===

Correctly Classified Instances       12946               22.3704 %
Incorrectly Classified Instances     44925               77.6296 %
Kappa statistic                          0.1444
Mean absolute error                      0.008
Root mean squared error                  0.0671
Relative absolute error                 91.6497 %
Root relative squared error            101.9158 %
Total Number of Instances            57871

=== Detailed Accuracy By Class ===
```

Status

OK

Decision Tree (Training Set and Test Set):



J48 -C 0.25 -M 2

○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Daily Summary

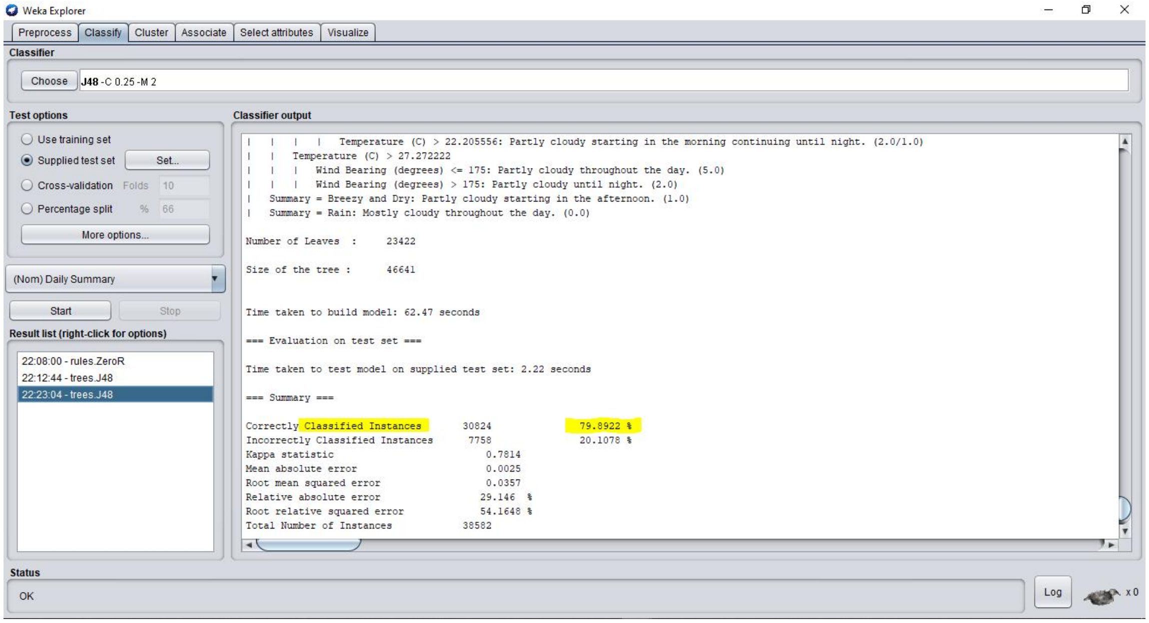
Result list (right-click for options)

```
19:31:53 - lazy.IBk
19:34:38 - lazy.IBk
19:34:46 - lazy.IBk
19:39:22 - lazy.IBk
19:45:17 - lazy.IBk
21:41:06 - bayes.NaiveBayes
21:43:22 - bayes.NaiveBayes
21:48:26 - bayes.NaiveBayes
21:52:15 - trees.J48
```

```
|   |   Temperature (C) > 27.272222
|   |   |   Wind Bearing (degrees) <= 175: Partly cloudy throughout the day. (5.0)
|   |   |   Wind Bearing (degrees) > 175: Partly cloudy until night. (2.0)
|   Summary = Breezy and Dry: Partly cloudy starting in the afternoon. (1.0)
|   Summary = Rain: Mostly cloudy throughout the day. (0.0)

Number of Leaves  :     23422

Size of the tree :      46641


Time taken to build model: 60.57 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 2 seconds

=== Summary ===

Correctly Classified Instances       30824               79.8922 %
Incorrectly Classified Instances      7758               20.1078 %
Kappa statistic                          0.7814
Mean absolute error                      0.0025
Root mean squared error                  0.0357
Relative absolute error                 29.146  %
Root relative squared error             54.1648 %
Total Number of Instances            38582

=== Detailed Accuracy By Class ===
```

Status

OK



J48 -C 0.25 -M 2

○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Daily Summary

Result list (right-click for options)

```
22:08:00 - rules.ZeroR
22:12:44 - trees.J48
22:23:04 - trees.J48
```

```
|   |   |   |   Temperature (C) > 22.205556: Partly cloudy starting in the morning continuing until night. (2.0/1.0)
|   |   Temperature (C) > 27.272222
|   |   |   Wind Bearing (degrees) <= 175: Partly cloudy throughout the day. (5.0)
|   |   |   Wind Bearing (degrees) > 175: Partly cloudy until night. (2.0)
|   Summary = Breezy and Dry: Partly cloudy starting in the afternoon. (1.0)
|   Summary = Rain: Mostly cloudy throughout the day. (0.0)

Number of Leaves  :     23422

Size of the tree :      46641


Time taken to build model: 62.47 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 2.22 seconds

=== Summary ===

Correctly Classified Instances       30824               79.8922 %
Incorrectly Classified Instances      7758               20.1078 %
Kappa statistic                          0.7814
Mean absolute error                      0.0025
Root mean squared error                  0.0357
Relative absolute error                 29.146  %
Root relative squared error             54.1648 %
Total Number of Instances            38582
```

Status

OK

Results:

C = Correctly Classified Instances

In = Incorrectly Classified Instances

By analyzing the data table of accuracy of different algorithms we can say that KNN is best algorithm for this specific problem. Accuracy of KNN is 99.96% and 99.94% for training set and test set where accuracy of Decision Tree is 79.08% and 88.44%. At the mean time the accuracy of Naïve Bayes for this data set is only 22.37% that means it can correctly classify only 22.37% instances. So the accuracy of Naïve Bayes is very poor for this data set.

	KNN	Naïve Bayes	Decision Tree
Training set	[C: 99.96%, In: 0.03%]	[C:22.37%, In:77.56%]	[C:79.89%, In:20.10%]
Test Set	[C:99.94%, In: 0.05%]	[C:22.46%, In:77.54%]	[C:88.44%, In:19.55%]

Why KNN is best for this dataset:

According to my results that I got from weka I can say that KNN is definitely suitable for this type of dataset.

According my data set there are 12 columns and our targeted column name is Daily Summary which gives us what is the weather of today. By analyzing this data set we need to find what will be the weather of tomorrow. This problem is a classification problem because it has numeric and discrete values and it has multiple classes. So it is a classification and supervised algorithm. Here we can't predict what our output will be so our dataset need unsupervised algorithms. Our data set is also unlabeled dataset.

I have executed this data set with k=1. So we only use nearest neighbor to define the category. As I have separated this data set into two part that are training set and test and after executing this data set I got 99.96% and 99.94% accuracy respectively. The second best accuracy value I got from decision tree. Decision tree uses several decision tree for several output and it need to re-execute when two output results are same. That's why Decision tree is not a suitable for this type of dataset. On the other hand Naïve Bayes is for basically sentiment analysis and text classification so it is not a superior algorithm for this data set and we got very poor accuracy for Naïve Bayes that is correct instances are only 22.37% where incorrect instances are 77.54%.

When we implemented KNN it works by searching the distances (Euclidean distance) between a query and all the possible examples of dataset and after that we assign a value for K (I used k=1) then it select nearest label. Our data set need to measure the distance among of **temperature**, **wind speed**, **pressure**, **visibility**, **Precip type**, **wind bearing** to give a proper decision that whether the weather of next day will be Partly cloudy throughout the day or Mostly cloudy throughout the day or other state. That's why KNN give us the best accuracy.

To recapitulate, based on above decision and analyzation undoubtedly we can say that KNN is best for this dataset.

Data-Set Reference: https://www.kaggle.com/abhishek20182/performing-analysis-of-meteorological-data