

Data Classification Project

Team Members

Jyoti Sharma
Siyu Xu

March 21, 2020

Data Mining

Jae Yong Lee

Table of Contents

Data Classification Project.....	0
Data Classification Project.....	0
DATA MINING GOAL	3
DATASET INFORMATION.....	3
Dataset Description.....	3
Dataset Attributes	3
Class Attribute.....	6
Data Sources	6
DATA MING TOOLS OR ALGORITHMS INFORMATION.....	7
Attribute Selection Algorithms	7
Classifier Algorithms	7
DATA PREPROCESSING	9
Manual preprocessing	9
DATA PREPROCESSING	11
Attribute Reduction	11
1) OneR	11
2) Cfs Subset.....	13
3) Information Gain.....	14
4) Correlation.....	16
Manual Selection	17
Attribute Reduction	18
Mapping of Datasets created:.....	21
DATA MINING RESULT AND EVALUATION	22
Test Performed on Daily_1 Dataset.....	22
1. Naïve Bayes	22
2. SGD.....	23
3. J48	24
4. Random Forest.....	25
5. IBk (k = 10).....	26
Test Performed on Daily_2 Dataset.....	27

1. Naïve Bayes	27
2. SGD.....	28
3. J48	29
4. Random Forest.....	30
5. IBk (k = 10).....	31
Test Performed on Daily_3 Dataset.....	32
1. Naïve Bayes	32
2. SGD.....	33
3. J48	34
4. Random Forest.....	35
5. IBk (k = 10).....	36
Test Performed on Daily_4 Dataset.....	37
1. Naïve Bayes	37
2. SGD.....	38
3. J48	39
4. Random Forest.....	40
5. IBk (k = 10).....	41
Test Performed on Daily_5 Dataset.....	42
1. Naïve Bayes	42
2. SGD.....	43
3. J48	44
4. Random Forest.....	45
5. IBk (k = 10).....	46
Justification of Best Model.....	48
1. Analysis	48
2. Results.....	49
3. Conclusion.....	54
SUMMARY	55
Bibliography.....	56

DATA MINING GOAL

The goal of this Data mining classification project is to predict the occurrence of rain in San Francisco Bay area.

The classification process will be carried out on based on information about various factors such as humidity, max/min air temperature, air speed etc.

DATASET INFORMATION

Dataset Description

The Dataset used is a monthly report generated by CIMIS (California Irrigation Management Information system) from January 2018 to January 2020. The Dataset constitutes of **732 tuples** containing **33 attributes** each.

The attributes consist of data about air temperature, wind speed, wind direction, solar radiation, relative humidity, vapor pressure, and rainfall.

Dataset Attributes

Below table consists of description of all attributes in the dataset:

S. No.	Attribute Name	Attribute Details
1	Stn Id	Id of computerized automated weather stations part of CIMS network
2	Stn Name	Name of computerized automated weather stations part of CIMS network
3	CIMIS Region	Geographic Region where Observations are made
4	Date	Date on Which Observation was taken
5	Jul	Number of Observation made in current year including the current record
6	ETo (in)	evapotranspiration recorded in in
7	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for ETO recording

8	Precip (in)	Precipitation field value indicates if precipitation occurred or not
9	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Precipitation recording
10	Sol Rad (Ly/day)	Solar radiation intensity per day
11	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for solar radiation recording
12	Avg Vap Pres (mBars)	Average Vapor pressure
13	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Average Vapor pressure recording
14	Max Air Temp (F)	Maximum Air Temp
15	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Max Air Temp recording
16	Min Air Temp (F)	Minimum Air temperature
17	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Minimum Air temperature recording
18	Avg Air Temp (F)	Average Air temperature
19	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Average Air temperature recording
20	Max Rel Hum (%)	Maximum Relative Humidity measurement
21	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Max Rel Hum recording
22	Min Rel Hum (%)	Min Relative Humidity measurement

23	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Min Rel Hum recording
24	Avg Rel Hum (%)	Average Humidity
25	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Avg Rel Hum recording
26	Dew Point (F)	Dew point temperature is the temperature to which the atmosphere must be cooled, at constant pressure and water vapor content, in order to reach saturation
27	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Dew Point recording
28	Avg Wind Speed (mph)	Average windspeed
29	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Avg Wind Speed recording
30	Wind Run (miles)	Wind Run is the total distance the wind has traveled past the weather station within a given time period
31	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Wind Run recording
32	Avg Soil Temp (F)	Average Soil temperature in F
33	qc	Quality Control to indicate if data recorded is in normal range or far out of normal range or historical average for Avg Soil Temp recording

Class Attribute

The class attribute is our prediction goal and would be the 8th attribute in the dataset. The current values populated are zero or greater than zero, the goal will be to generate this class attribute values as 0 or 1 indicating if precipitation occurred on that day or not.

Data Sources

<https://cimis.water.ca.gov>

DATA MING TOOLS OR ALGORITHMS INFORMATION

Attribute Selection Algorithms

As a part of preprocessing and for generating unique datasets we have decided the below attribute selection algorithms.

1) OneRAttributeEval:

Evaluates the worth of an attribute by using the OneR classifier.

2) CfsSubsetEval:

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred (Weka).

3) InfoGainAttributeEval:

Evaluates the worth of an attribute by measuring the information gain with respect to the class (Weka).

$$\text{InfoGain(Class, Attribute)} = H(\text{Class}) - H(\text{Class} | \text{Attribute}) \text{ (Weka).}$$

4) CorrelationAttributeEval:

Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average (Weka).

5) Manual Selection:

Attributes are selected by our understanding of datasets.

Classifier Algorithms

1) Naïve Bayes:

In machine learning, Naïve Bayes classifiers are a family of simple “probabilistic classifiers” based on applying Bayes’ theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models (Wikipedia, 2020).

2) SGD:

SGD implements stochastic gradient descent for learning various linear models

(binary class SVM, binary class logistic regression, squared loss, Huber loss and epsilon-insensitive loss linear regression).

Globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes, so the coefficients in the output are based on the normalized data. For numeric class attributes, the squared, Huber or epsilon-insensitive loss function must be used. Epsilon-insensitive and Huber loss may require a much higher learning rate (Weka).

3) J48:

J48 (C4.5) is an algorithm used to generate a decision tree developed by Ross Quinlan, which can be used for classification. And for this reason, J48 is often referred to as a statistical classifier (Wikipedia, 2020).

4) Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set (Wikipedia, 2020).

5) IBk (k = 10):

IBk (Instance Based Learner) algorithm generates a prediction for a test instance just-in-time. It uses a distance measure to locate k “close” instances in the training data for each test instance and uses those selected instances to make a prediction (Weka).

DATA PREPROCESSING

Manual preprocessing

In the original dataset, it has 33 attributes, as a part of data pre-processing. We have removed some columns for which more than 95% values were missing.

Below are the steps used.

Firstly, we use RStudio to do the data preprocessing for our project:

1. Import our original data into RStudio.

```
> #import data
> data <- read.table(file = "daily.csv", header = TRUE, sep = ",") 
> head(data)
  Stn.Id Stn.Name    CIMIS.Region   Date Jul ETo..in. qc Precip..in. qc.1 Sol.Rad..Ly.day. qc.2
1      47 Brentwood San Francisco Bay 1/1/2018   1  0.05     0.00          190
2      47 Brentwood San Francisco Bay 1/2/2018   2  0.03     0.00          105
3      47 Brentwood San Francisco Bay 1/3/2018   3  0.01     0.27          77
4      47 Brentwood San Francisco Bay 1/4/2018   4  0.03     0.15         141
5      47 Brentwood San Francisco Bay 1/5/2018   5  0.01     0.02          80
6      47 Brentwood San Francisco Bay 1/6/2018   6  0.06     0.00         231
Avg.Vap.Pres..mBars. qc.3 Max.Air.Temp..F. qc.4 Min.Air.Temp..F. qc.5 Avg.Air.Temp..F. qc.6 Max.Rel.Hum....
1           8.3          66.5          38.3          51.1          90
2           8.6          60.1          39.8          49.1          85
3          10.0          55.0          38.5          48.5          99
4          13.7          67.3          49.0          55.0          99
5          14.2          63.0          47.8          54.5          99
6          11.5          65.5          45.4          55.0          98
  qc.7 Min.Rel.Hum.... qc.8 Avg.Rel.Hum.... qc.9 Dew.Point..F. qc.10 Avg.Wind.Speed..mph. qc.11
1            38             65            39.8            2.9
2            59             72            40.6            2.2
3            70             86            44.5            3.1
4            70             92            52.9            3.5
5            90             98            54.0            4.5
6            57             78            48.2            4.7
Wind.Run..miles. qc.12 Avg.Soil.Temp..F. qc.13
1            68.9            47.8
2            53.6            48.0
3            74.7            48.6
4            85.2            51.6
5           107.6            52.8
6           113.8            53.6
```

2. Remove all unnecessary columns, such as “Stn Id”, “Stn Name”, “CIMIS Region” (as entire dataset belongs to one weather station), and all “qc” columns (more than 95% values are missing).

```
> #delete unnecessary information
> data2 <- data[,c(-1:-5,-7,-9,-11,-13,-15,-17,-19,-21,-23,-25,-27,-29,-31,-33)]
```

3. Add one more column named “class”, which is based on the column of “Precip (in)”, if the value is larger than 0 it shows in “True”, otherwise, “False”.

```
> #add one more column to clarify class attribute
> attach(data2)
> data2$class = ifelse(Precip..in. > 0, 'True', 'False')
```

4. remove “Precip (in)”, in order to keep our class attribute independent.

```
> #remove the previous class attribute "Precip (in)"
> data3 <- data2[,-2]
```

After the preprocessing from RStudio, our dataset looks like:

```
> head(data3)
  ET0..in. Sol.Rad..Ly.day. Avg.Vap.Pres..mBars. Max.Air.Temp..F. Min.Air.Temp..F. Avg.Air.Temp..F.
1   0.05          190        8.3       66.5      38.3      51.1
2   0.03          105        8.6       60.1      39.8      49.1
3   0.01           77       10.0       55.0      38.5      48.5
4   0.03          141       13.7       67.3      49.0      55.0
5   0.01           80       14.2       63.0      47.8      54.5
6   0.06          231       11.5       65.5      45.4      55.0
  Max.Rel.Hum.... Min.Rel.Hum.... Avg.Rel.Hum.... Dew.Point..F. Avg.Wind.Speed..mph. Wind.Run..miles.
1            90          38         65       39.8       2.9      68.9
2            85          59         72       40.6       2.2      53.6
3            99          70         86       44.5       3.1      74.7
4            99          70         92       52.9       3.5      85.2
5            99          90         98       54.0       4.5     107.6
6            98          57         78       48.2       4.7     113.8
  Avg.Soil.Temp..F. class
1           47.8  False
2           48.0  False
3           48.6  True
4           51.6  True
5           52.8  True
6           53.6  False
```

Secondly, we use Weka to transform our dataset from the format of CSV to ARFF.

DATA PREPROCESSING

Attribute Reduction

1. Dataset creation via select attributes algorithms

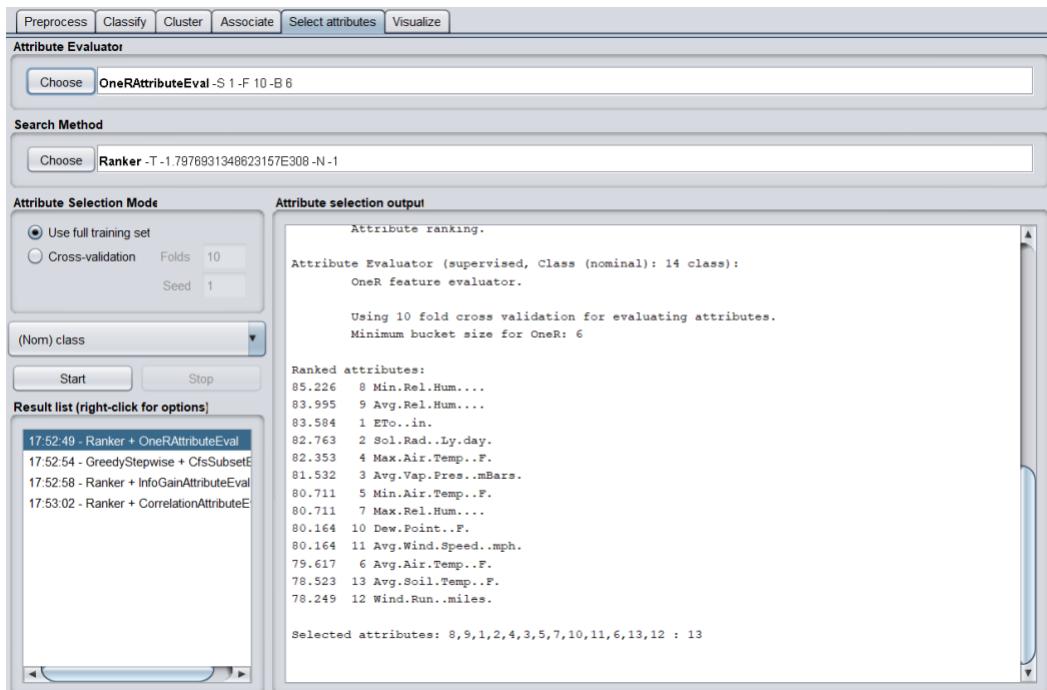
We have created first four datasets using select attributes algorithms.

Below is the mapping of dataset and the corresponding attribute evaluator and Search Method combinations.

Data set Names	Attribute Evaluator	Search Method
Daily_1	OneRAttributeEval	Ranker
Daily_2	CfsSubsetEval	GreedyStepwise
Daily_3	InfoGainAttributeEval	Ranker
Daily_4	CorrelationAttributeEval	Ranker

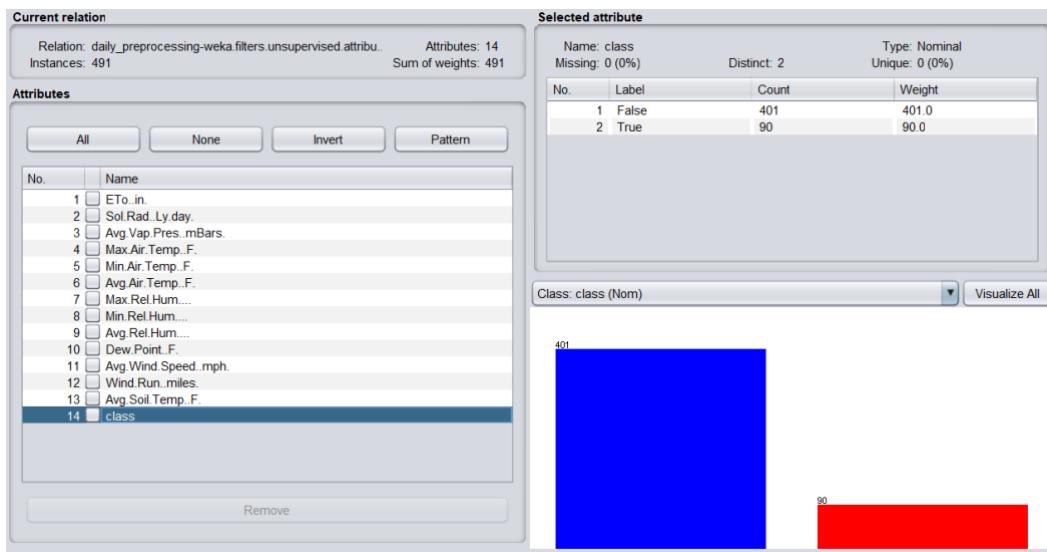
Below are the screenshots depicting steps for each attribute evaluator:

1) OneR

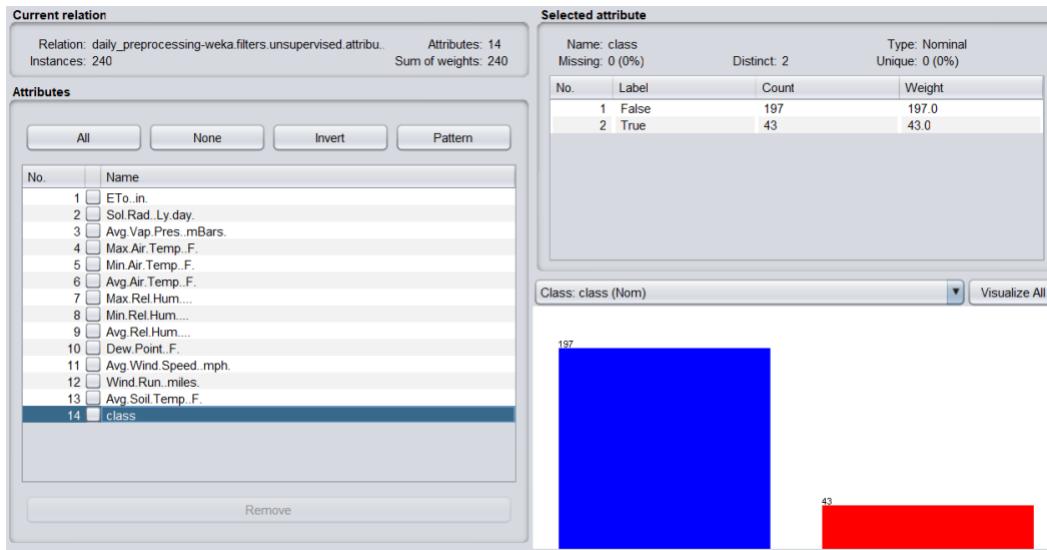


The class attribute is: 13.

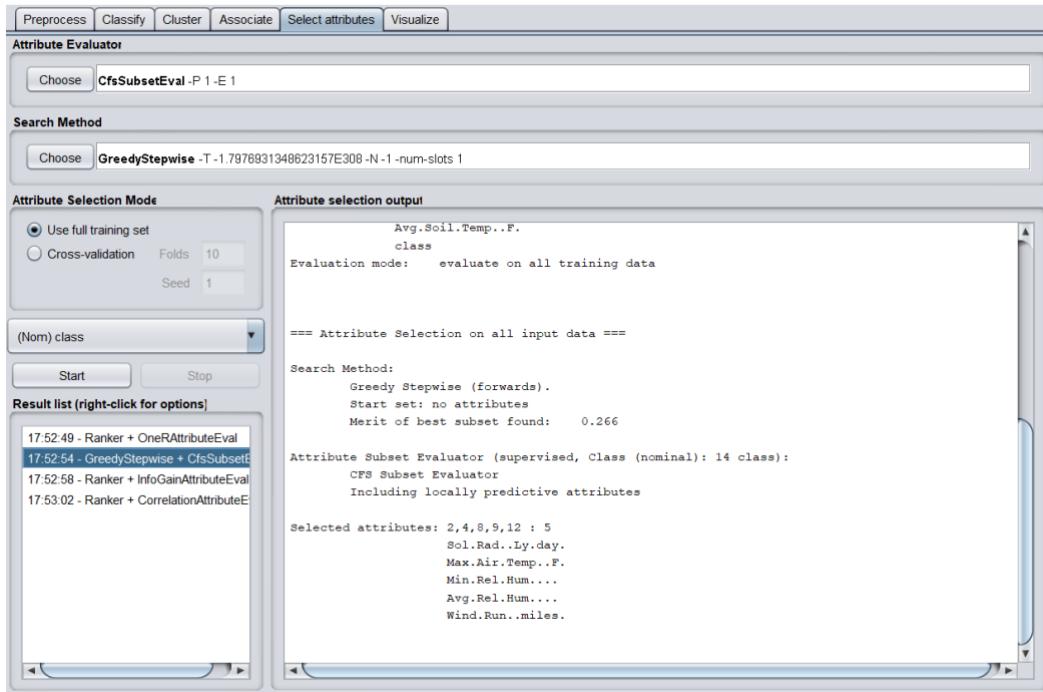
The training dataset with selected attributes:



The test dataset with selected attributes:

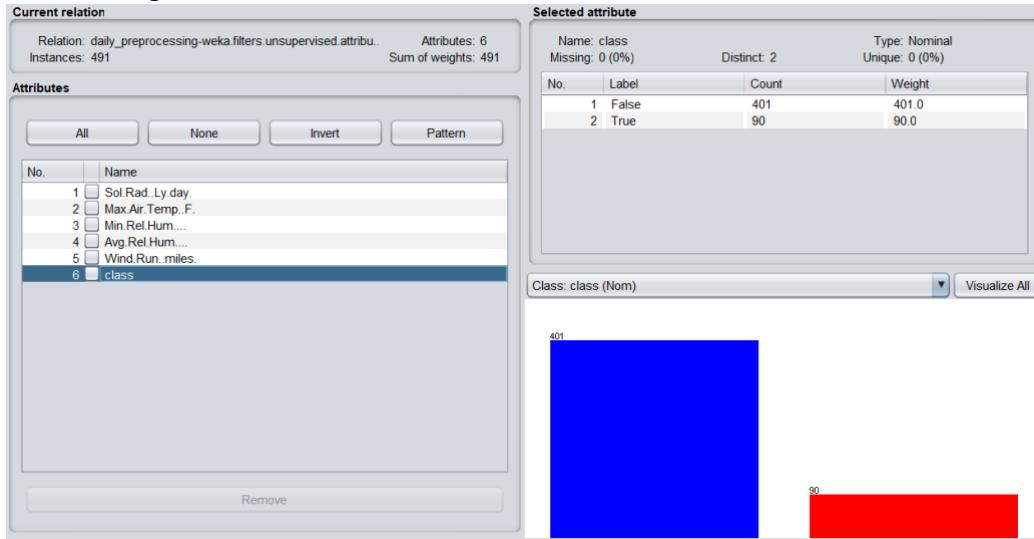


2) Cfs Subset

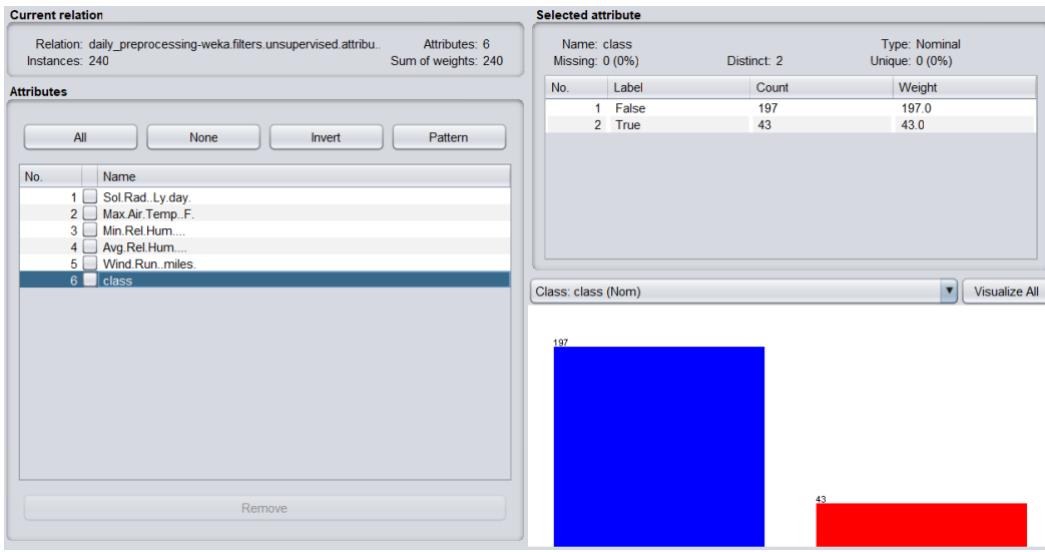


The class attribute is: 5.

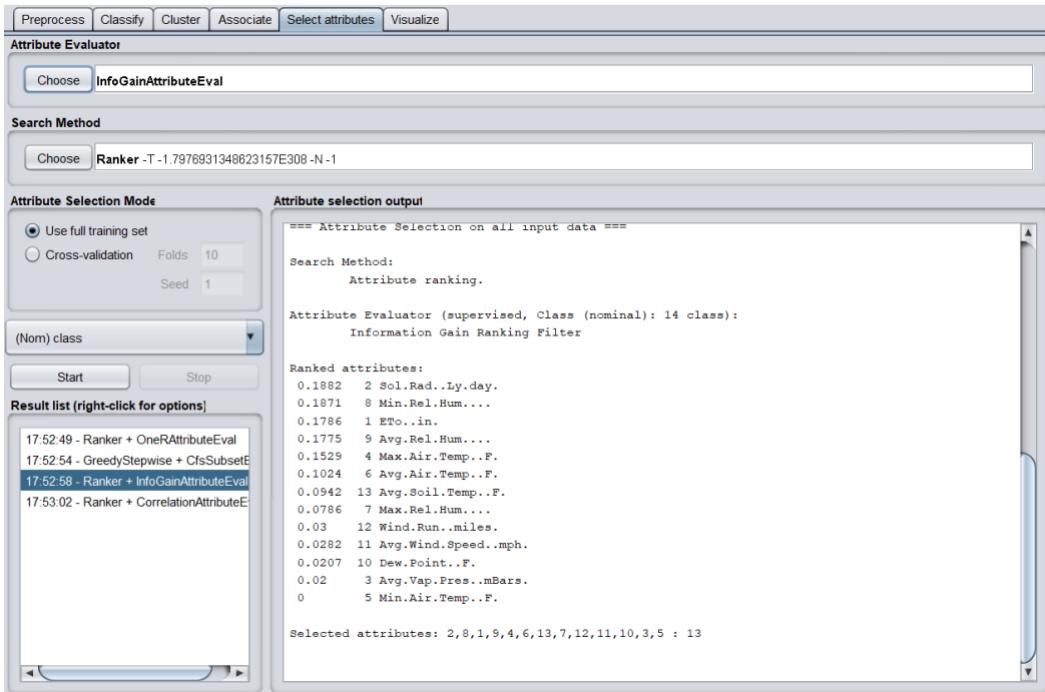
The training dataset with selected attributes:



The test dataset with selected attributes:

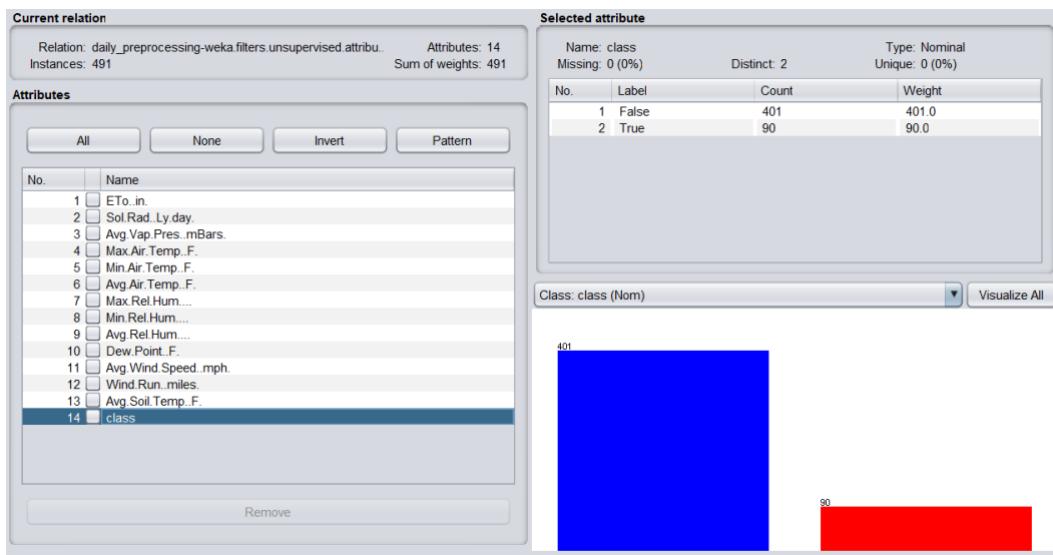


3) Information Gain

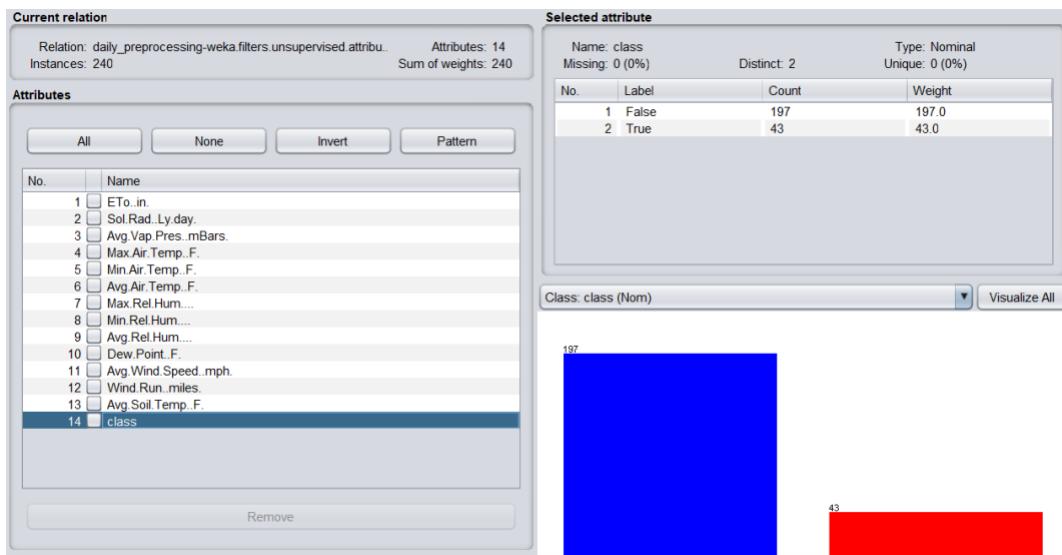


The class attribute is: 13.

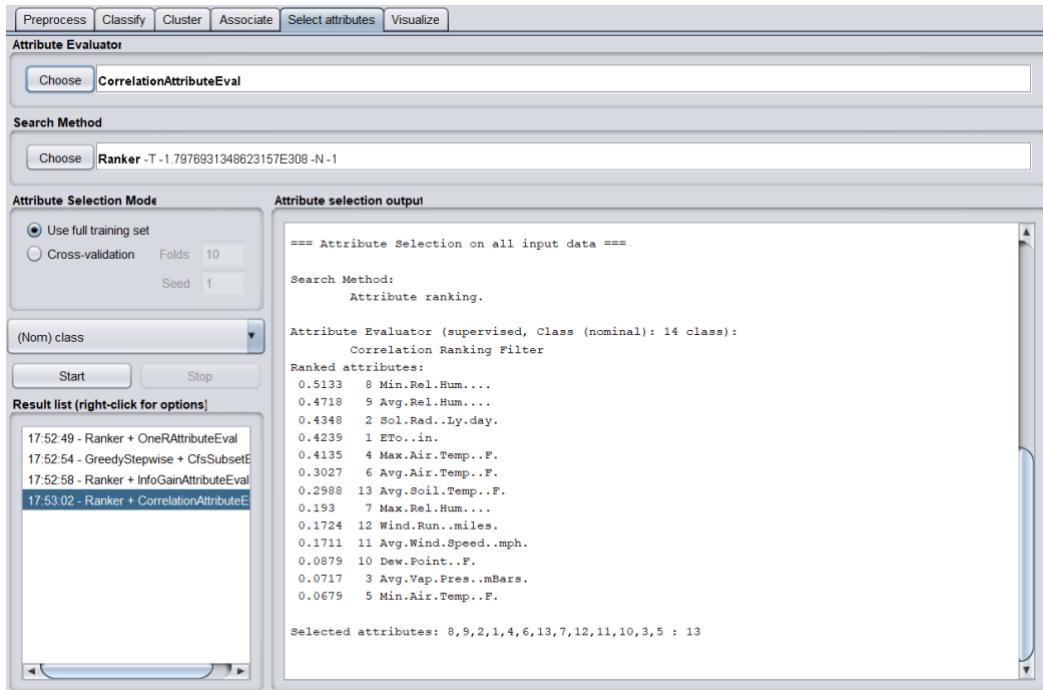
The training dataset with selected attributes:



The test dataset with selected attributes:

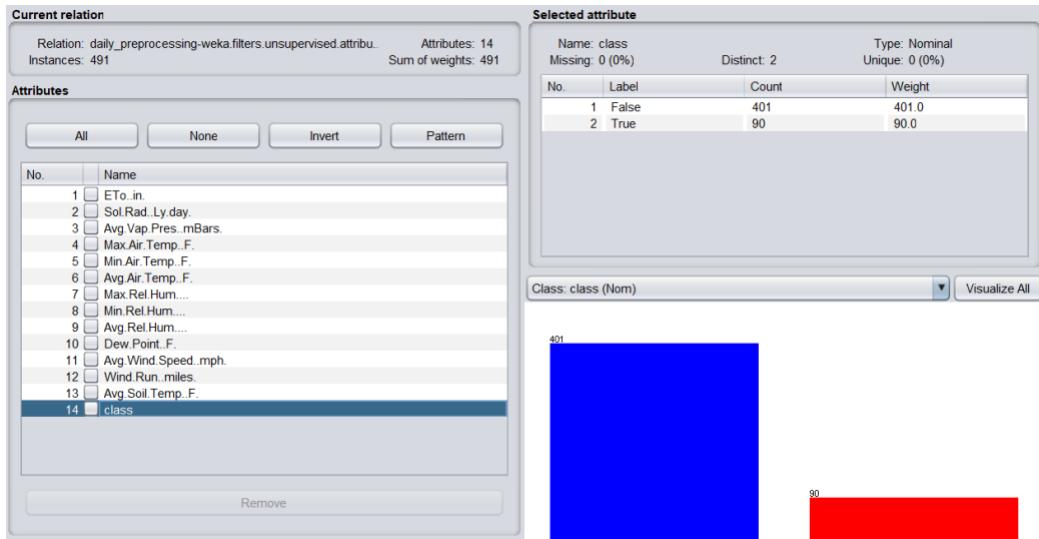


4) Correlation

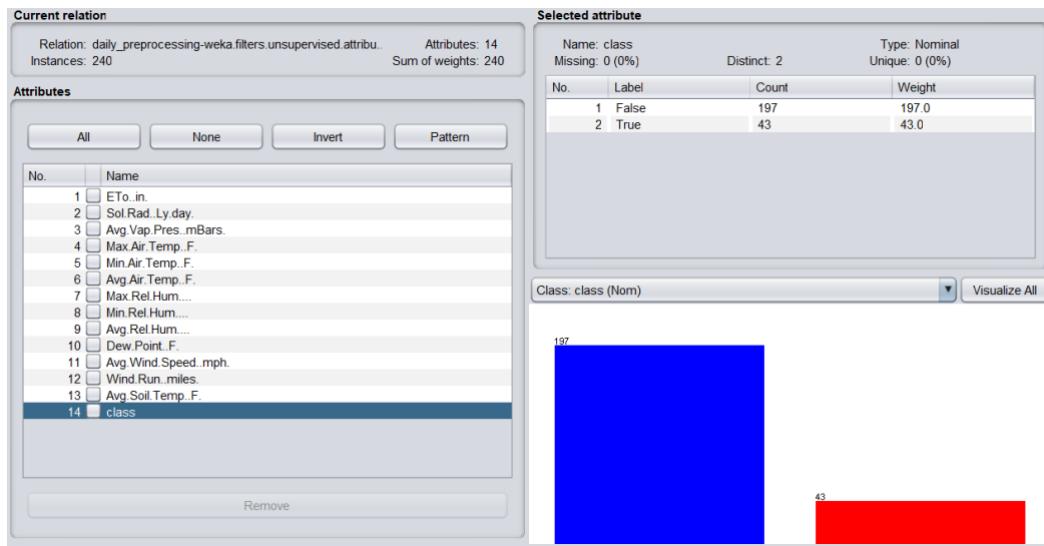


The class attribute is: 13.

The training dataset with selected attributes:



The test dataset with selected attributes:



2. Dataset creation manually

Manual Selection

Dataset Daily_5 is created by reducing attributes manually. Based on our understanding we have removed values:

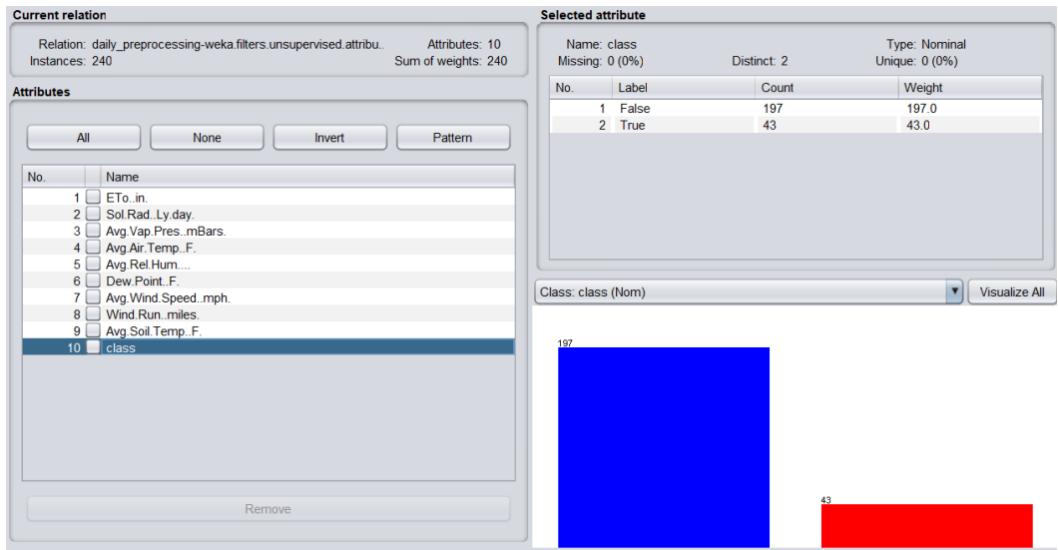
Max Rel Hum (%)	Min Rel Hum (%)	Max Air Temp (F)	Min Air Temp (F)

Because all of them have average values of these attributes already present in the dataset: **Avg Rel Hum (%)** and **Avg Air Temp (F)**.

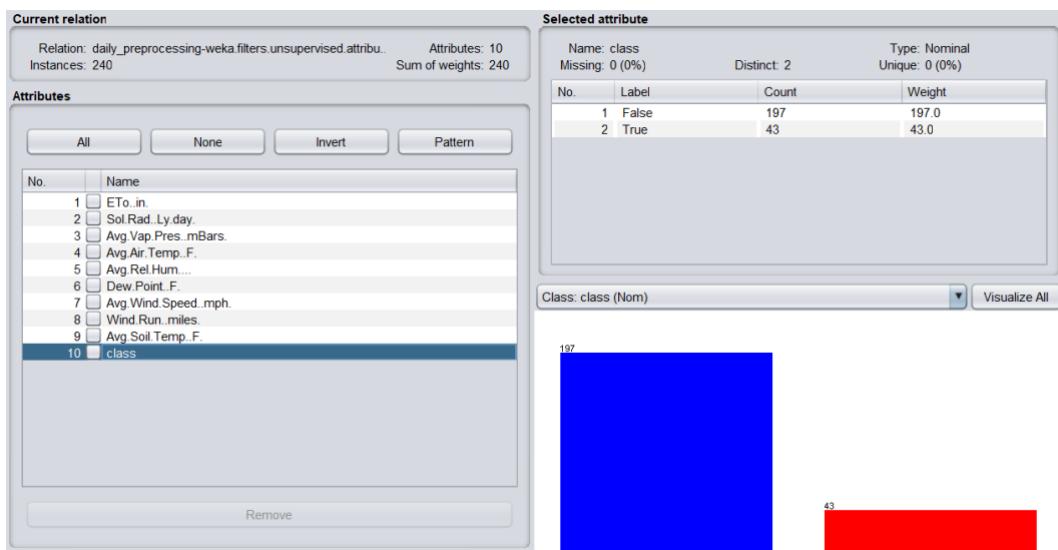
The attributes we keep are: **ETo (in)**, **Sol Rad (Ly/day)**, **Avg Vap Pres (mBars)**, **Avg Air Temp (F)**, **Avg Rel Hum (%)**, **Dew Point (F)**, **Avg Wind Speed (mph)**, **Wind Run (miles)**, **Avg Soil Temp (F)**, and class attribute **class**.

The class attribute is: 10.

The training dataset with selected attributes:



The test dataset with selected attributes:

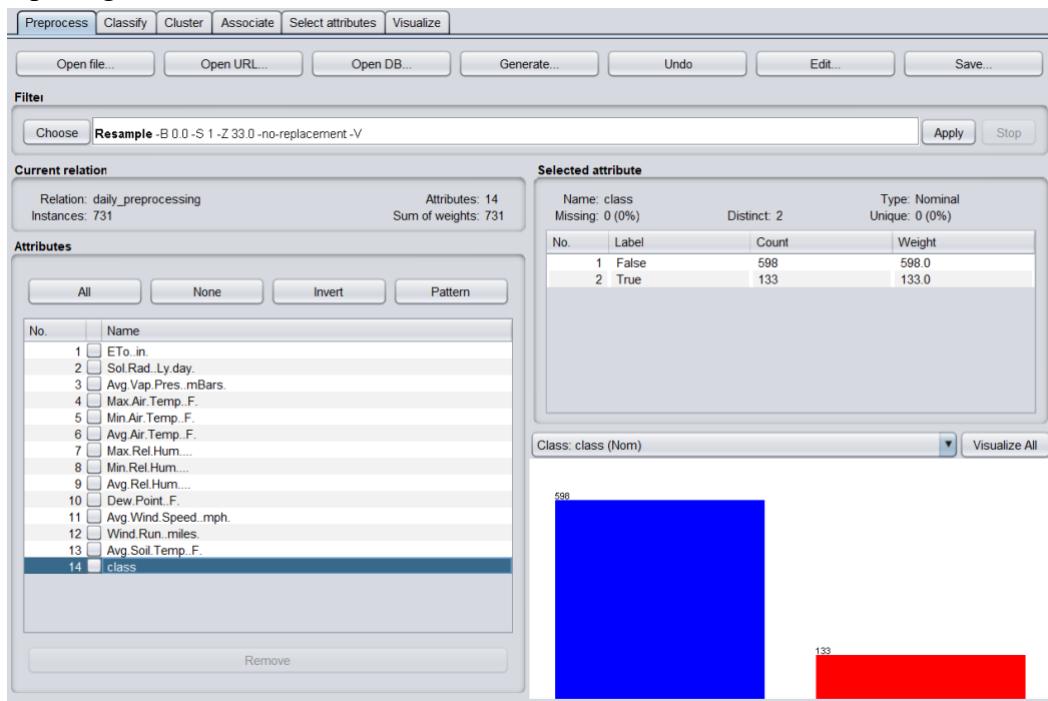


Attribute Reduction

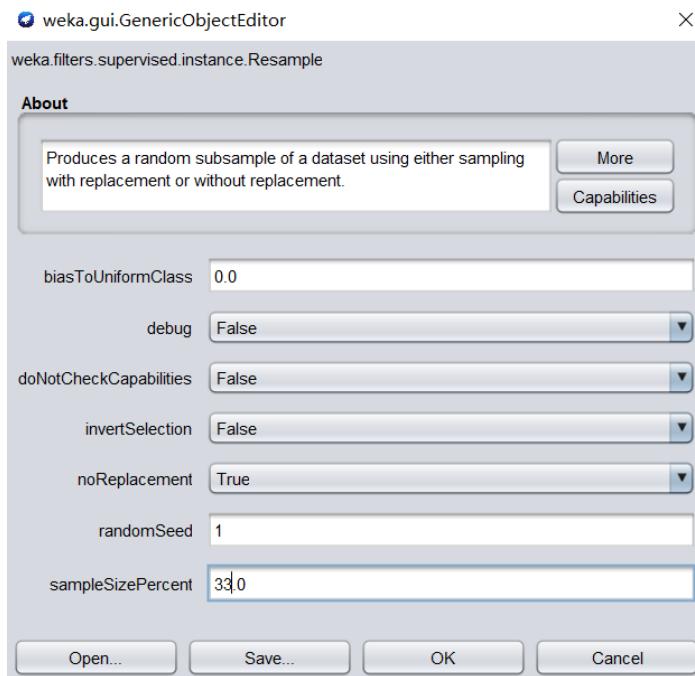
From the previous part we split all our dataset into training dataset and test dataset in a ratio of 67:33 using Resample attribute.

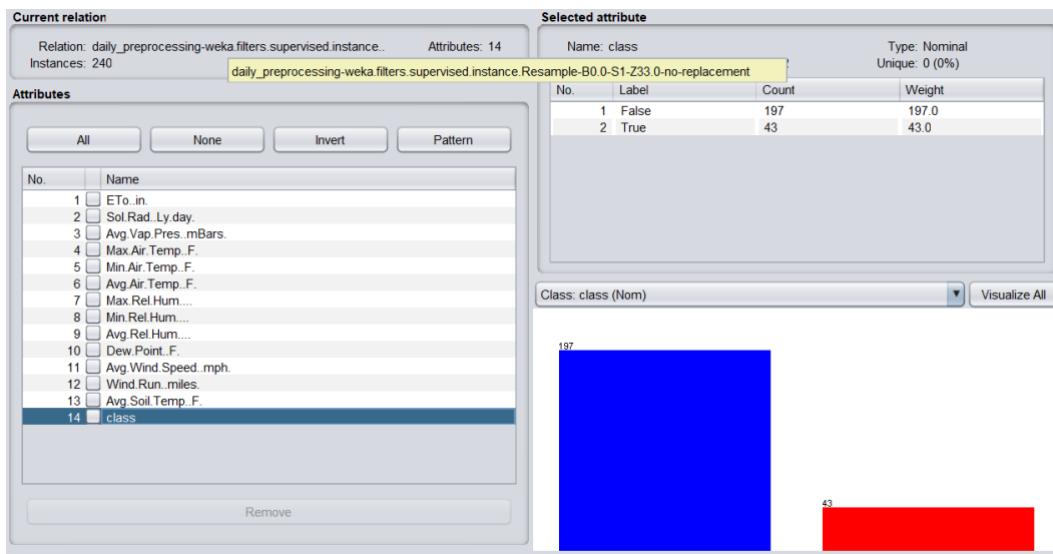
Screenshots of the dataset splitting process:

1. Opening the dataset in Weka

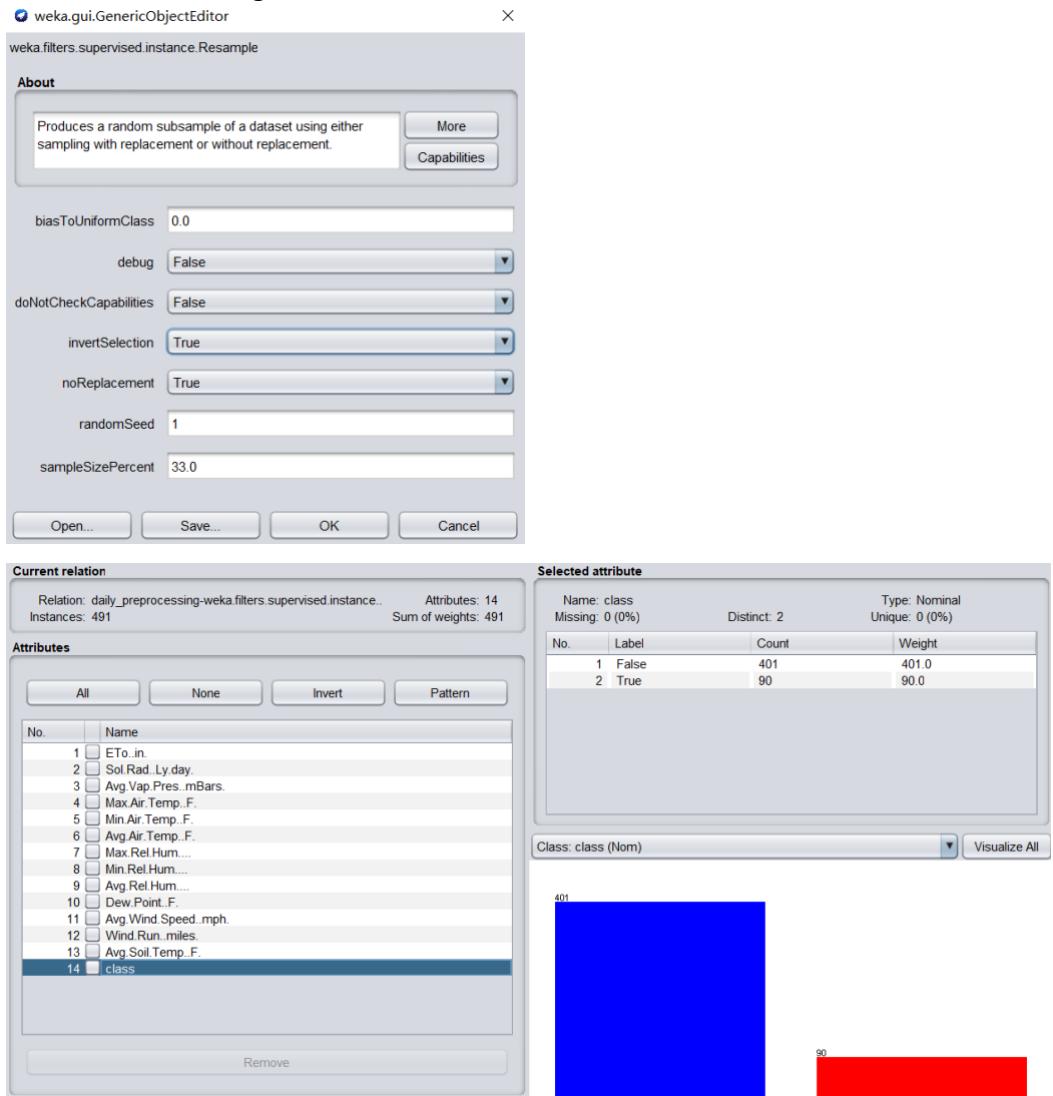


2. Creation of Test dataset

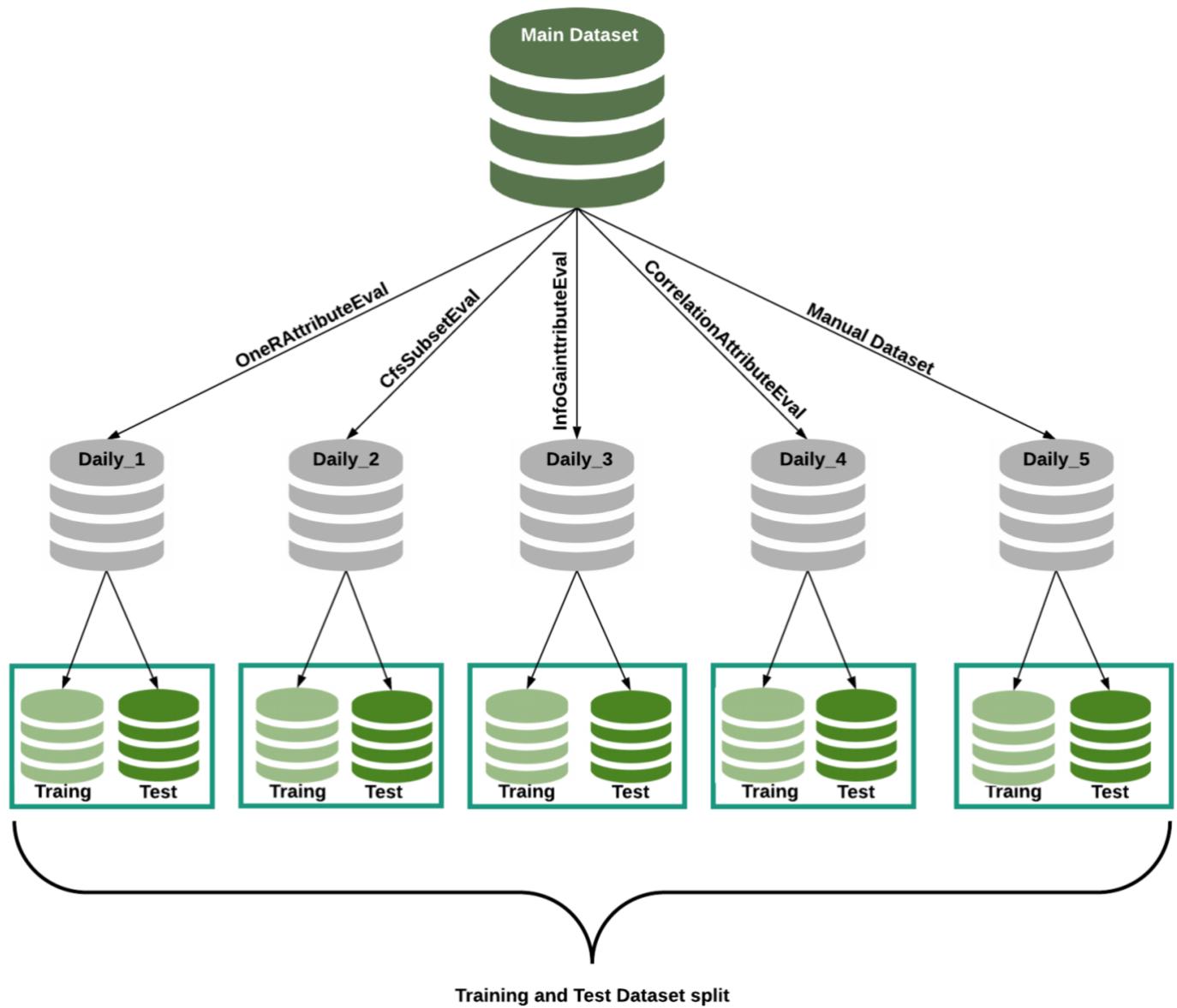




3. Creation of Training dataset



Mapping of Datasets created:



DATA MINING RESULT AND EVALUATION

To figure out the best model for classification process, we will be performing classification process using below classifiers:

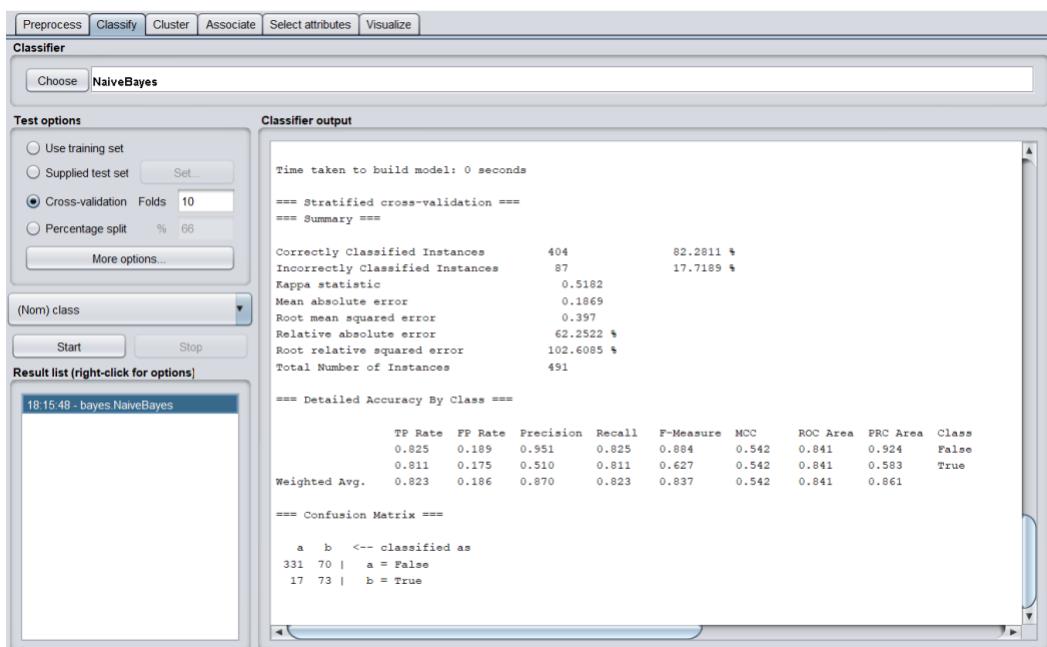
NaiveBayes
SGD
J48
Random Forest
IBK(K=10)

We will be building models using Training datasets and testing them using the test datasets created after splitting the datasets in 67:33 ratio.

Test Performed on Daily_1 Dataset

1. Naïve Bayes

Build the model with training dataset:



Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'NaiveBayes' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

== Evaluation on test set ==
Time taken to test model on supplied test set: 0 seconds

== Summary ==
Correctly Classified Instances      188          78.3333 %
Incorrectly Classified Instances   52           21.6667 %
Kappa statistic                   0.4294
Mean absolute error               0.2281
Root mean squared error          0.4438
Relative absolute error           76.628 %
Root relative squared error     115.7144 %
Total Number of Instances        240

== Detailed Accuracy By Class ==
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.787    0.233    0.939    0.787    0.856    0.459    0.807    0.924    False
      0.767    0.213    0.440    0.767    0.559    0.459    0.807    0.592    True
Weighted Avg.                      0.783    0.229    0.850    0.783    0.803    0.459    0.807    0.865

== Confusion Matrix ==
      a      b  <-- classified as
155  42 |  a = False
 10  33 |  b = True

```

2. SGD

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'SGD -F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1' is chosen. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the following evaluation results:

```

Time taken to build model: 0.06 seconds

== Stratified cross-validation ==
== Summary ==
Correctly Classified Instances      430          87.5764 %
Incorrectly Classified Instances   61           12.4236 %
Kappa statistic                   0.5481
Mean absolute error               0.1242
Root mean squared error          0.3525
Relative absolute error           41.369 %
Root relative squared error     91.098 %
Total Number of Instances        491

== Detailed Accuracy By Class ==
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.948    0.444    0.905    0.948    0.926    0.554    0.752    0.900    False
      0.556    0.052    0.704    0.556    0.621    0.554    0.752    0.473    True
Weighted Avg.                      0.876    0.373    0.868    0.876    0.870    0.554    0.752    0.822

== Confusion Matrix ==
      a      b  <-- classified as
380  21 |  a = False
 40  50 |  b = True

```

Test it with test dataset:

```

Choose SGD -F 0-L 0.01-R 1.0E-4-E 500-C 0.001-S 1

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

(Nom) class Start Stop

Result list (right-click for options)
18:15:48 - bayes.NaiveBayes
18:16:29 - bayes.NaiveBayes
18:16:59 - functions.SGD
18:17:27 - functions.SGD

Classifier output
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0 seconds

==== Summary ====
Correctly Classified Instances 210 87.5 %
Incorrectly Classified Instances 30 12.5 %
Kappa statistic 0.559
Mean absolute error 0.125
Root mean squared error 0.3536
Relative absolute error 42.0015 %
Root relative squared error 92.184 %
Total Number of Instances 240

==== Detailed Accuracy By Class ====
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.934 0.395 0.915 0.934 0.925 0.560 0.769 0.909 False
0.605 0.066 0.667 0.605 0.634 0.560 0.769 0.474 True
Weighted Avg. 0.875 0.336 0.871 0.875 0.873 0.560 0.769 0.831

==== Confusion Matrix ====
a b <-- classified as
184 13 | a = False
17 26 | b = True

```

3. J48

Build the model with training dataset:

```

Choose J48 -C 0.25 -M 2

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

(Nom) class Start Stop

Result list (right-click for options)
18:15:48 - bayes.NaiveBayes
18:16:29 - bayes.NaiveBayes
18:16:59 - functions.SGD
18:17:27 - functions.SGD
18:17:54 - trees.J48

Classifier output
Time taken to build model: 0.03 seconds

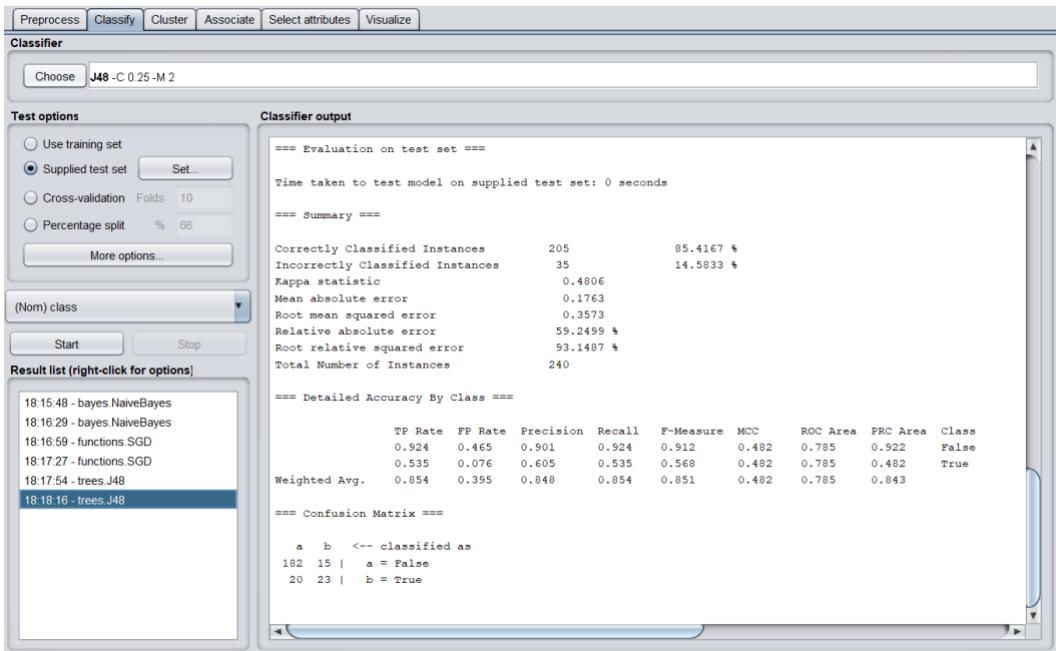
==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances 439 89.4094 %
Incorrectly Classified Instances 52 10.5906 %
Kappa statistic 0.6092
Mean absolute error 0.1386
Root mean squared error 0.309
Relative absolute error 46.1694 %
Root relative squared error 79.8498 %
Total Number of Instances 491

==== Detailed Accuracy By Class ====
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.963 0.411 0.913 0.963 0.937 0.610 0.780 0.899 False
0.589 0.037 0.779 0.589 0.671 0.618 0.780 0.569 True
Weighted Avg. 0.894 0.343 0.888 0.894 0.888 0.618 0.780 0.839

==== Confusion Matrix ====
a b <-- classified as
386 15 | a = False
37 53 | b = True

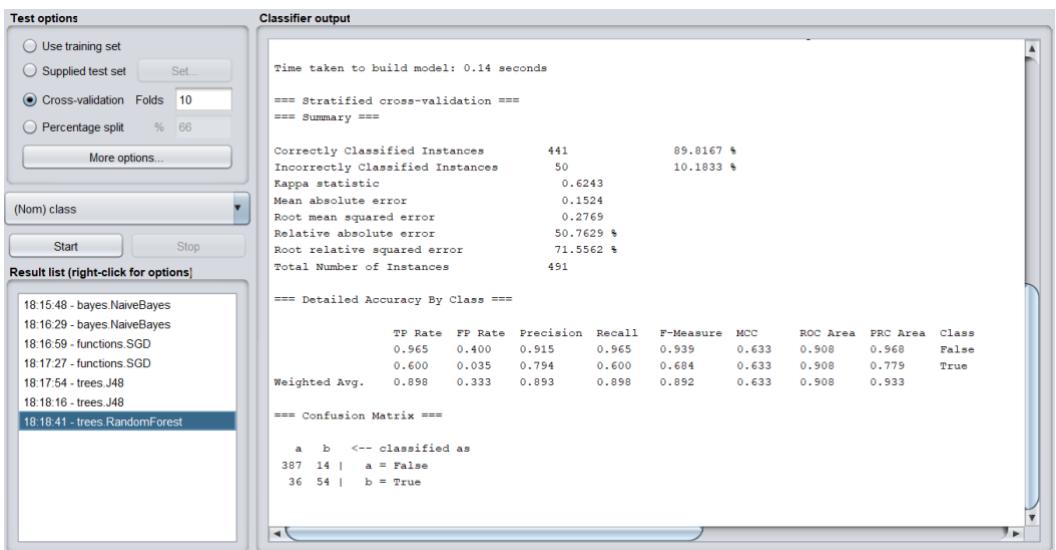
```

Test it with test dataset:

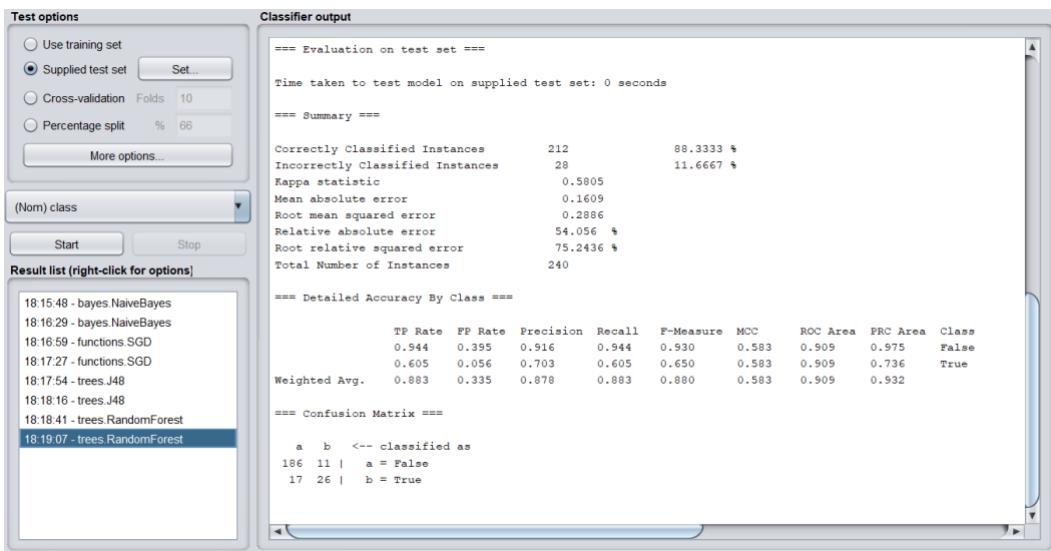


4. Random Forest

Build the model with training dataset:

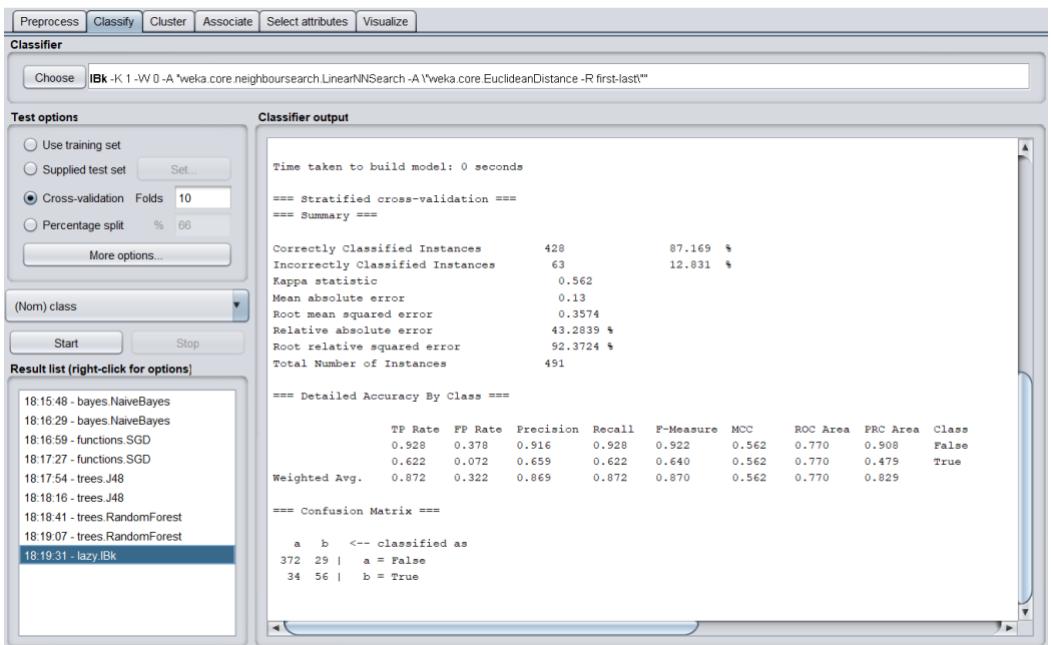


Test it with test dataset:



5. IBk (k = 10)

Build the model with training dataset:



Test it with test dataset:

Classifier output

```

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

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

  (Nom) class
  Start Stop

Result list (right-click for options)
  18:15:48 - bayes.NaiveBayes
  18:16:29 - bayes.NaiveBayes
  18:16:59 - functions SGD
  18:17:27 - functions SGD
  18:17:54 - trees.J48
  18:18:16 - trees.J48
  18:18:41 - trees.RandomForest
  18:19:07 - trees.RandomForest
  18:19:31 - lazy.IBk
18:19:47 - lazy.IBk

Classifier output
  === Evaluation on test set ===
  Time taken to test model on supplied test set: 0.01 seconds
  === Summary ===
  Correctly Classified Instances 200 83.3333 %
  Incorrectly Classified Instances 40 16.6667 %
  Kappa statistic 0.4627
  Mean absolute error 0.168
  Root mean squared error 0.4074
  Relative absolute error 56.4564 %
  Root relative squared error 106.2301 %
  Total Number of Instances 240
  === Detailed Accuracy By Class ===
    TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
    0.883 0.395 0.911 0.883 0.897 0.464 0.744 0.900 False
    0.605 0.117 0.531 0.605 0.565 0.464 0.744 0.392 True
  Weighted Avg. 0.833 0.345 0.843 0.833 0.837 0.464 0.744 0.809
  === Confusion Matrix ===
    a b <-- classified as
    174 23 | a = False
    17 26 | b = True
  
```

Test Performed on Daily_2 Dataset

1. Naïve Bayes

Build the model with training dataset:

Classifier output

```

Choose NaiveBayes

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

  (Nom) class
  Start Stop

Result list (right-click for options)
  18:20:39 - bayes.NaiveBayes

Classifier output
  Time taken to build model: 0 seconds
  === Stratified cross-validation ===
  === Summary ===
  Correctly Classified Instances 414 84.3177 %
  Incorrectly Classified Instances 77 15.6823 %
  Kappa statistic 0.5271
  Mean absolute error 0.1644
  Root mean squared error 0.3464
  Relative absolute error 54.7392 %
  Root relative squared error 89.5372 %
  Total Number of Instances 491
  === Detailed Accuracy By Class ===
    TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
    0.873 0.289 0.931 0.873 0.901 0.533 0.851 0.925 False
    0.711 0.127 0.557 0.711 0.624 0.533 0.851 0.587 True
  Weighted Avg. 0.843 0.259 0.862 0.843 0.850 0.533 0.851 0.863
  === Confusion Matrix ===
    a b <-- classified as
    350 51 | a = False
    26 64 | b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'NaiveBayes' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      196          81.6667 %
    Incorrectly Classified Instances   44           18.3333 %
    Kappa statistic                   0.4726
    Mean absolute error               0.1923
    Root mean squared error          0.3933
    Relative absolute error          64.6209 %
    Root relative squared error     102.5542 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.838    0.279    0.932      0.838    0.882      0.487   0.814      0.924      False
    0.721    0.162    0.492      0.721    0.585      0.487   0.814      0.603      True
    Weighted Avg.                     0.817    0.258    0.853      0.817    0.829      0.487   0.814      0.867

    === Confusion Matrix ===
    a   b   <-- classified as
    165 32 |  a = False
    12 31 |  b = True
  
```

2. SGD

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'SGD -F 0-L 0.01-R 1.0E-4-E 500-C 0.001-S 1' is chosen. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0.01 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      430          87.5764 %
    Incorrectly Classified Instances   61           12.4236 %
    Kappa statistic                   0.5114
    Mean absolute error               0.1242
    Root mean squared error          0.3525
    Relative absolute error          41.3699 %
    Root relative squared error     91.098 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.968    0.533    0.890      0.968    0.927      0.533   0.717      0.888      False
    0.467    0.032    0.764      0.467    0.579      0.533   0.717      0.454      True
    Weighted Avg.                     0.876    0.442    0.867      0.876    0.863      0.533   0.717      0.808

    === Confusion Matrix ===
    a   b   <-- classified as
    388 13 |  a = False
    48 42 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' section, 'SGD - F 0-L 0.01-R 1.0E-4-E 500-C 0.001-S 1' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      206      85.8333 %
    Incorrectly Classified Instances   34       14.1667 %
    Kappa statistic                   0.4906
    Mean absolute error               0.1417
    Root mean squared error           0.3764
    Relative absolute error            47.6017 %
    Root relative squared error       98.1374 %
    Total Number of Instances         240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.929    0.465    0.901     0.929    0.915     0.493    0.732    0.896    False
    0.535    0.071    0.622     0.535    0.575     0.493    0.732    0.416    True
    Weighted Avg.                      0.858    0.395    0.851     0.858    0.854     0.493    0.732    0.810

    === Confusion Matrix ===
    a   b   <-- classified as
    183 14 |  a = False
    20 23 |  b = True
  
```

3. J48

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' section, 'J48 - C 0.25-M 2' is chosen. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      432      87.9837 %
    Incorrectly Classified Instances   59       12.0163 %
    Kappa statistic                   0.5502
    Mean absolute error               0.1588
    Root mean squared error           0.3157
    Relative absolute error            52.8947 %
    Root relative squared error       81.5853 %
    Total Number of Instances         491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.958    0.467    0.901     0.958    0.929    0.560    0.822    0.922    False
    0.533    0.042    0.738     0.533    0.619    0.560    0.822    0.562    True
    Weighted Avg.                      0.880    0.389    0.872     0.880    0.872    0.560    0.822    0.856

    === Confusion Matrix ===
    a   b   <-- classified as
    384 17 |  a = False
    42 48 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'J48 - C 0.25 - M 2'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the evaluation results on the test set.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      208          86.6667 %
    Incorrectly Classified Instances   32           13.3333 %
    Kappa statistic                   0.5111
    Mean absolute error              0.1725
    Root mean squared error          0.3434
    Relative absolute error          57.9775 %
    Root relative squared error     89.547 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.939    0.465    0.902      0.939    0.920      0.515    0.797    0.926    False
    0.535    0.061    0.657      0.535    0.590      0.515    0.797    0.477    True
    Weighted Avg.                      0.867    0.393    0.858      0.867    0.861      0.515    0.797    0.846

    === Confusion Matrix ===

    a   b   <-- classified as
    185 12 |  a = False
    20  23 |  b = True
  
```

4. Random Forest

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' panel shows 'Cross-validation' selected with 'Folds 10'. The 'Classifier output' panel displays the evaluation results using stratified cross-validation.

```

    Time taken to build model: 0.05 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      442          90.0204 %
    Incorrectly Classified Instances   49           9.9796 %
    Kappa statistic                   0.6403
    Mean absolute error              0.1473
    Root mean squared error          0.2817
    Relative absolute error          49.0538 %
    Root relative squared error     72.8133 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.960    0.367    0.921      0.960    0.940      0.645    0.900    0.968    False
    0.633    0.040    0.781      0.633    0.659      0.645    0.900    0.771    True
    Weighted Avg.                      0.900    0.307    0.895      0.900    0.896      0.645    0.900    0.932

    === Confusion Matrix ===

    a   b   <-- classified as
    385 16 |  a = False
    33  57 |  b = True
  
```

Test it with test dataset:

Classifier output

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===

    Correctly Classified Instances      214          89.1667 %
    Incorrectly Classified Instances   26           10.8333 %
    Kappa statistic                   0.6178
    Mean absolute error              0.1604
    Root mean squared error          0.2967
    Relative absolute error           53.888 %
    Root relative squared error     77.3535 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.944    0.349    0.925     0.944    0.935     0.619    0.876    0.955    False
    0.651    0.056    0.718     0.651    0.683     0.619    0.876    0.704    True
    Weighted Avg.                    0.892    0.296     0.888     0.892    0.890     0.619    0.876    0.910

    === Confusion Matrix ===

    a   b   <-- classified as
    186 11 |  a = False
    15 28 |  b = True
  
```

5. IBk (k = 10)

Build the model with training dataset:

Classifier output

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===

    Correctly Classified Instances      424          86.3544 %
    Incorrectly Classified Instances   67           13.6456 %
    Kappa statistic                   0.5342
    Mean absolute error              0.1381
    Root mean squared error          0.3686
    Relative absolute error           45.9845 %
    Root relative squared error     95.2596 %
    Total Number of Instances        491

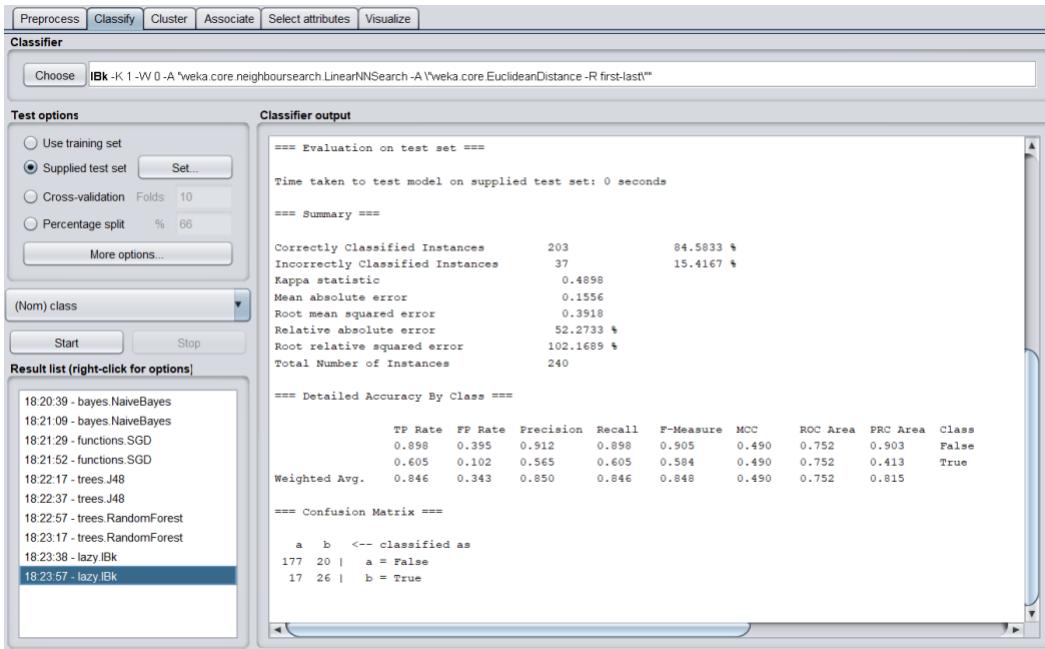
    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.923    0.400    0.911     0.923    0.917     0.535    0.761    0.904    False
    0.600    0.077    0.635     0.600    0.617     0.535    0.761    0.458    True
    Weighted Avg.                    0.864    0.341     0.861     0.864    0.862     0.535    0.761    0.823

    === Confusion Matrix ===

    a   b   <-- classified as
    370 31 |  a = False
    36 54 |  b = True
  
```

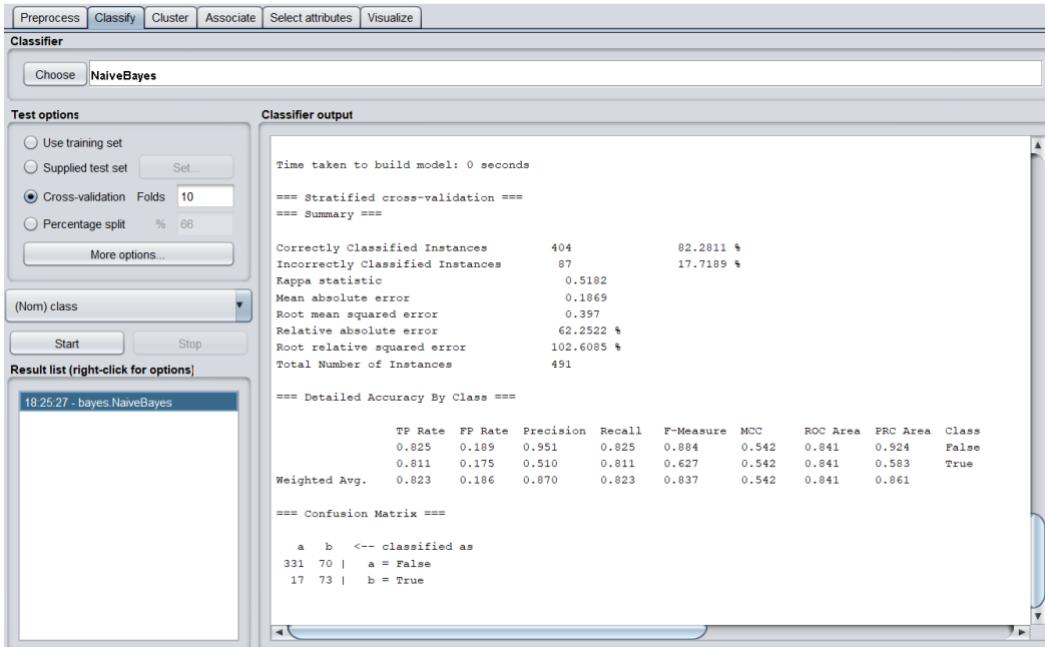
Test it with test dataset:



Test Performed on Daily_3 Dataset

1. Naïve Bayes

Build the model with training dataset:



Test it with test dataset:

The screenshot shows the Weka interface with the "Classify" tab selected. Under "Classifier", "NaiveBayes" is chosen. In the "Test options" panel, "Supplied test set" is selected. The "Classifier output" pane displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      188      78.3333 %
    Incorrectly Classified Instances   52       21.6667 %
    Kappa statistic                   0.4294
    Mean absolute error              0.2281
    Root mean squared error          0.4438
    Relative absolute error           76.628 %
    Root relative squared error      115.7144 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.787     0.233     0.939      0.787     0.856      0.459     0.807     0.924     False
    0.767     0.213     0.440      0.767     0.559      0.459     0.807     0.592     True
    Weighted Avg.                      0.783     0.229     0.850      0.783     0.803      0.459     0.807     0.865

    === Confusion Matrix ===
    a   b   <-- classified as
    155  42 |  a = False
    10   33 |  b = True
  
```

2. SGD

Build the model with training dataset:

The screenshot shows the Weka interface with the "Classify" tab selected. Under "Classifier", "SGD-F 0-L 0.01-R 1.0E-4-E 500-C 0.001-S 1" is chosen. In the "Test options" panel, "Cross-validation" is selected with "Folds" set to 10. The "Classifier output" pane displays the following evaluation results:

```

    Time taken to build model: 0.01 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      430      87.5764 %
    Incorrectly Classified Instances   61       12.4236 %
    Kappa statistic                   0.5481
    Mean absolute error              0.1242
    Root mean squared error          0.3525
    Relative absolute error           41.3699 %
    Root relative squared error      91.098 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.948     0.444     0.905      0.948     0.926      0.554     0.752     0.900     False
    0.556     0.052     0.704      0.556     0.621      0.554     0.752     0.473     True
    Weighted Avg.                      0.876     0.373     0.868      0.876     0.870      0.554     0.752     0.822

    === Confusion Matrix ===
    a   b   <-- classified as
    380  21 |  a = False
    40   50 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the "Classify" tab selected. The "Classifier" dropdown is set to "SGD -F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1". The "Test options" panel shows "Supplied test set" selected with 10 folds. The "Classifier output" pane displays the evaluation results on the test set.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===

    Correctly Classified Instances      210          87.5 %
    Incorrectly Classified Instances   30           12.5 %
    Kappa statistic                   0.559
    Mean absolute error               0.125
    Root mean squared error           0.3536
    Relative absolute error            42.0015 %
    Root relative squared error       52.184 %
    Total Number of Instances         240

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.934     0.395     0.915      0.934     0.925      0.560     0.769     0.909     False
    0.605     0.066     0.667      0.605     0.634      0.560     0.769     0.474     True

    Weighted Avg.                      0.875     0.336     0.871      0.875     0.873      0.560     0.769     0.831

    === Confusion Matrix ===

    a   b   <-- classified as
    184 13 |  a = False
    17  26 |  b = True
  
```

3. J48

Build the model with training dataset:

The screenshot shows the Weka interface with the "Classify" tab selected. The "Classifier" dropdown is set to "J48 -C 0.25 -M 2". The "Test options" panel shows "Cross-validation" selected with 10 folds. The "Classifier output" pane displays the evaluation results using stratified cross-validation.

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===

    Correctly Classified Instances      439          89.4094 %
    Incorrectly Classified Instances   52           10.5906 %
    Kappa statistic                   0.6092
    Mean absolute error               0.1386
    Root mean squared error           0.309
    Relative absolute error            46.1694 %
    Root relative squared error       79.8498 %
    Total Number of Instances         491

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.963     0.411     0.913      0.963     0.937     0.618     0.780     0.899     False
    0.589     0.037     0.779      0.589     0.671     0.618     0.780     0.569     True

    Weighted Avg.                      0.894     0.343     0.888      0.894     0.888     0.618     0.780     0.839

    === Confusion Matrix ===

    a   b   <-- classified as
    386 15 |  a = False
    37  53 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'J48 - C 0.25-M 2'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      205          85.4167 %
    Incorrectly Classified Instances   35           14.5833 %
    Kappa statistic                   0.4806
    Mean absolute error              0.1763
    Root mean squared error          0.3573
    Relative absolute error          59.2499 %
    Root relative squared error     93.1487 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.924    0.465    0.901      0.924    0.912      0.482   0.785     0.922     False
    0.535    0.076    0.605      0.535    0.568      0.482   0.785     0.482     True
    Weighted Avg.                     0.854    0.395    0.848      0.854    0.851      0.482   0.785     0.843

    === Confusion Matrix ===
    a   b   <-- classified as
    182 15 |  a = False
    20 23 |  b = True
  
```

4. Random Forest

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'RandomForest -P 100-I 100-num-slots 1-K 0-M 1.0-V 0.001-S 1'. The 'Test options' panel shows 'Cross-validation' selected with 'Folds 10'. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0.06 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      441          89.8167 %
    Incorrectly Classified Instances   50           10.1833 %
    Kappa statistic                   0.6243
    Mean absolute error              0.1524
    Root mean squared error          0.2769
    Relative absolute error          50.7629 %
    Root relative squared error     71.5562 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.965    0.400    0.915      0.965    0.939      0.633   0.908     0.968     False
    0.600    0.035    0.794      0.600    0.684      0.633   0.908     0.779     True
    Weighted Avg.                     0.898    0.333    0.893      0.898    0.892      0.633   0.908     0.933

    === Confusion Matrix ===
    a   b   <-- classified as
    387 14 |  a = False
    36 54 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' pane displays the evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      212      88.3333 %
    Incorrectly Classified Instances   28       11.6667 %
    Kappa statistic                   0.5805
    Mean absolute error               0.1609
    Root mean squared error          0.2886
    Relative absolute error           54.056 %
    Root relative squared error      75.2436 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
    0.944  0.395  0.916  0.944  0.930  0.583  0.909  0.975  False
    0.605  0.056  0.703  0.605  0.650  0.503  0.909  0.736  True

    Weighted Avg.                   0.883  0.335  0.878  0.883  0.880  0.583  0.909  0.932

    === Confusion Matrix ===
    a   b   <-- classified as
    186 11 |  a = False
    17 26 |  b = True
  
```

5. IBk (k = 10)

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'IBk -K 1 -W 0 -A weka.core.neighboursearch.LinearNNSearch -A \weka.core.EuclideanDistance -R first-last**'. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' pane displays the evaluation results:

```

    Time taken to build model: 0 seconds

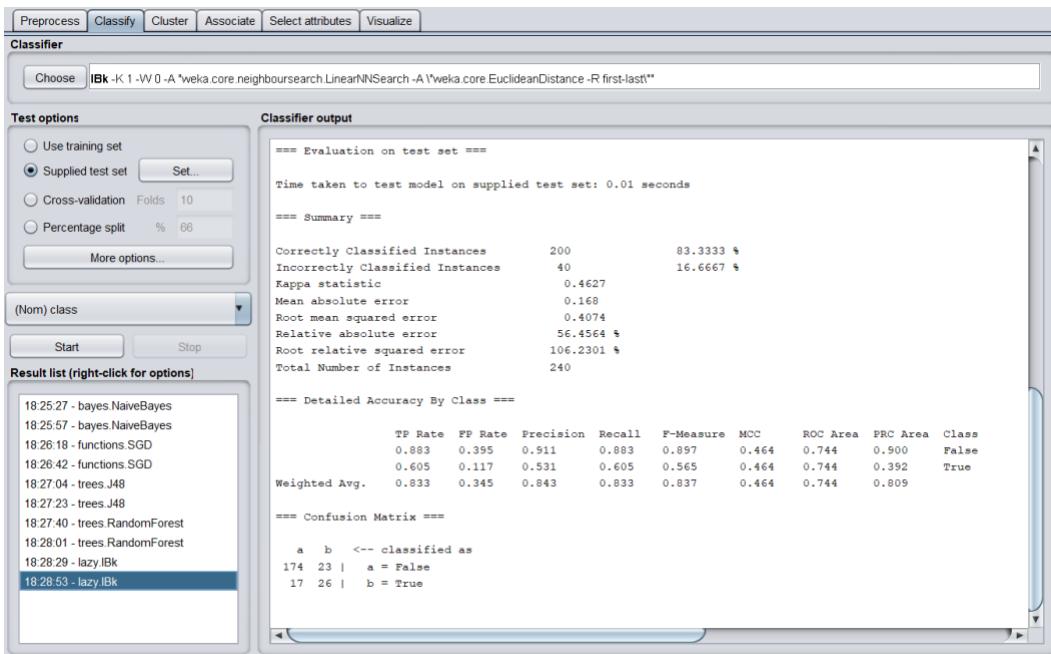
    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      428      87.169 %
    Incorrectly Classified Instances   63       12.831 %
    Kappa statistic                   0.562
    Mean absolute error               0.13
    Root mean squared error          0.3574
    Relative absolute error           43.2839 %
    Root relative squared error      92.3724 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
    0.928  0.378  0.916  0.928  0.922  0.562  0.770  0.908  False
    0.622  0.072  0.659  0.622  0.640  0.562  0.770  0.479  True

    Weighted Avg.                   0.872  0.322  0.869  0.872  0.870  0.562  0.770  0.829

    === Confusion Matrix ===
    a   b   <-- classified as
    372 29 |  a = False
    34 56 |  b = True
  
```

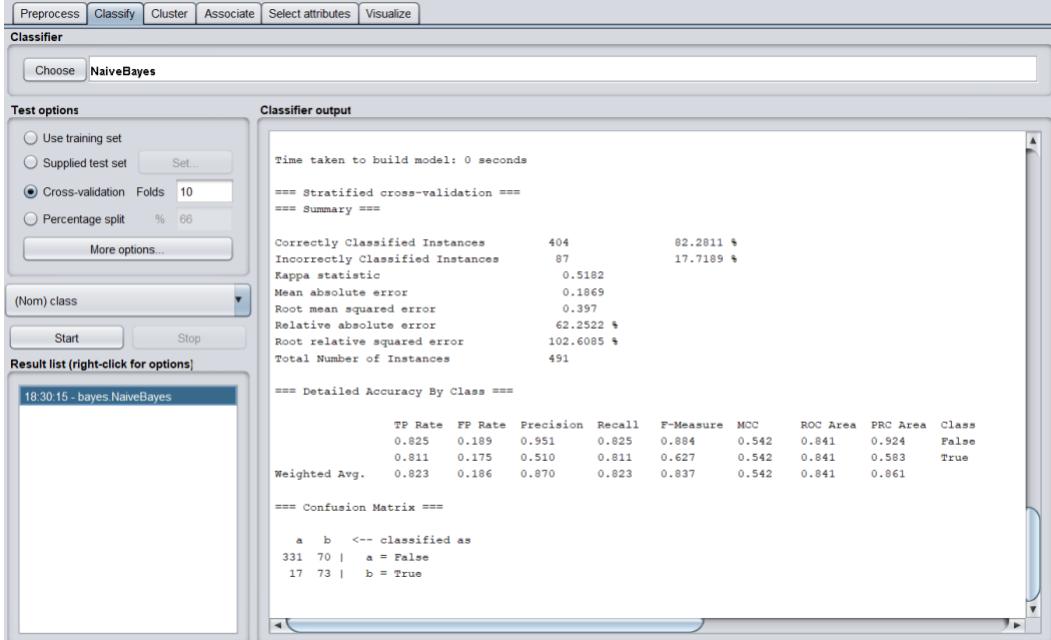
Test it with test dataset:



Test Performed on Daily_4 Dataset

1. Naïve Bayes

Build the model with training dataset:



Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'NaiveBayes' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      188      78.3333 %
    Incorrectly Classified Instances   52       21.6667 %
    Kappa statistic                   0.4294
    Mean absolute error              0.2281
    Root mean squared error          0.4438
    Relative absolute error           76.628 %
    Root relative squared error      115.7144 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.787    0.233    0.939     0.787    0.856     0.459   0.807    0.924    False
    0.767    0.213    0.440     0.767    0.559     0.459   0.807    0.592    True
    Weighted Avg.                     0.783    0.229    0.850     0.783    0.803     0.459   0.807    0.865

    === Confusion Matrix ===
    a   b   <-- classified as
    155 42 |  a = False
    10  33 |  b = True
  
```

2. SGD

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'SGD -F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1' is chosen. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0.01 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      430      87.5764 %
    Incorrectly Classified Instances   61       12.4236 %
    Kappa statistic                   0.5481
    Mean absolute error              0.1242
    Root mean squared error          0.3525
    Relative absolute error           41.3699 %
    Root relative squared error      91.098 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.948    0.444    0.905     0.948    0.926     0.554   0.752    0.900    False
    0.556    0.052    0.704     0.556    0.621     0.554   0.752    0.473    True
    Weighted Avg.                     0.876    0.373    0.868     0.876    0.870     0.554   0.752    0.822

    === Confusion Matrix ===
    a   b   <-- classified as
    380 21 |  a = False
    40  50 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'SGD' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      210      87.5 %
    Incorrectly Classified Instances   30       12.5 %
    Kappa statistic                   0.559
    Mean absolute error               0.125
    Root mean squared error          0.3536
    Relative absolute error           42.0015 %
    Root relative squared error      92.184 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Class
    0.934    0.395    0.915     0.934    0.925     0.560    0.769    0.909    False
    0.605    0.066    0.667     0.605    0.634     0.560    0.769    0.474    True
    Weighted Avg.                      0.875    0.336    0.871     0.875    0.873     0.560    0.769    0.831

    === Confusion Matrix ===
    a   b   <-- classified as
    184 13 |  a = False
    17 26 |  b = True
  
```

3. J48

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' panel, 'J48' is chosen. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      439      89.4094 %
    Incorrectly Classified Instances   52       10.5906 %
    Kappa statistic                   0.6092
    Mean absolute error               0.1386
    Root mean squared error          0.309
    Relative absolute error           46.1694 %
    Root relative squared error      79.8490 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Class
    0.963    0.411    0.913     0.963    0.937     0.618    0.780    0.899    False
    0.589    0.037    0.779     0.589    0.671     0.618    0.780    0.569    True
    Weighted Avg.                      0.894    0.343    0.888     0.894    0.888     0.618    0.780    0.839

    === Confusion Matrix ===
    a   b   <-- classified as
    386 15 |  a = False
    37 53 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'J48 - C 0.25 - M 2'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      205          85.4167 %
    Incorrectly Classified Instances   35           14.5833 %
    Kappa statistic                   0.4806
    Mean absolute error               0.1763
    Root mean squared error          0.3573
    Relative absolute error          59.2499 %
    Root relative squared error     93.1487 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.924    0.465    0.901      0.924    0.912      0.482   0.785     0.922     False
    0.535    0.076    0.605      0.535    0.568      0.482   0.785     0.482     True
    Weighted Avg.                     0.854    0.395    0.848      0.854    0.851      0.482   0.785     0.843

    === Confusion Matrix ===
    a   b   <-- classified as
    182 15 |  a = False
    20  23 |  b = True
  
```

4. Random Forest

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' panel shows 'Cross-validation' selected with 'Folds 10'. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0.06 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      441          89.8167 %
    Incorrectly Classified Instances   50           10.1833 %
    Kappa statistic                   0.6243
    Mean absolute error               0.1524
    Root mean squared error          0.2769
    Relative absolute error          50.7629 %
    Root relative squared error     71.5562 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.965    0.400    0.915      0.965    0.939      0.633   0.908     0.968     False
    0.600    0.035    0.794      0.600    0.684      0.633   0.908     0.779     True
    Weighted Avg.                     0.898    0.333    0.893      0.898    0.892      0.633   0.908     0.933

    === Confusion Matrix ===
    a   b   <-- classified as
    387 14 |  a = False
    36  54 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Choose' dropdown is set to 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' panel shows 'Supplied test set' selected with 10 folds. The 'Classifier output' pane displays the following summary statistics:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      212      88.3333 %
    Incorrectly Classified Instances   28       11.6667 %
    Kappa statistic                   0.5805
    Mean absolute error               0.1609
    Root mean squared error           0.2886
    Relative absolute error            54.056 %
    Root relative squared error       75.2436 %
    Total Number of Instances         240
  
```

Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.944	0.395	0.916	0.944	0.930	0.583	0.909	0.975	False	
0.605	0.056	0.703	0.605	0.650	0.583	0.909	0.736	True	
Weighted Avg.	0.883	0.335	0.878	0.883	0.880	0.583	0.909	0.932	

Confusion Matrix:

	a	b	<-- classified as
186	11	11	a = False
17	26	26	b = True

5. IBk (k = 10)

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Choose' dropdown is set to 'IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \weka.core.EuclideanDistance -R first-last"'. The 'Test options' panel shows 'Cross-validation' selected with 10 folds. The 'Classifier output' pane displays the following summary statistics:

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      428      87.169 %
    Incorrectly Classified Instances   63       12.831 %
    Kappa statistic                   0.562
    Mean absolute error               0.13
    Root mean squared error           0.3574
    Relative absolute error            43.2839 %
    Root relative squared error       92.3724 %
    Total Number of Instances         491
  
```

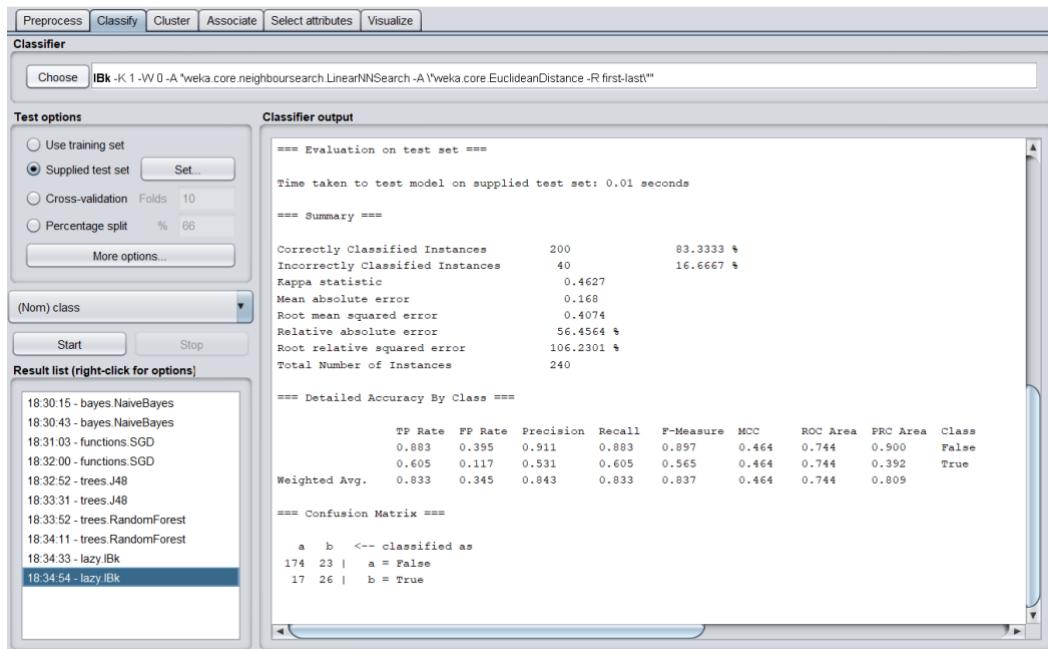
Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.928	0.378	0.916	0.928	0.922	0.562	0.770	0.908	0.929	False
0.622	0.072	0.659	0.622	0.640	0.562	0.770	0.479	0.521	True
Weighted Avg.	0.872	0.322	0.869	0.872	0.870	0.562	0.770	0.829	

Confusion Matrix:

	a	b	<-- classified as
372	29	29	a = False
34	56	56	b = True

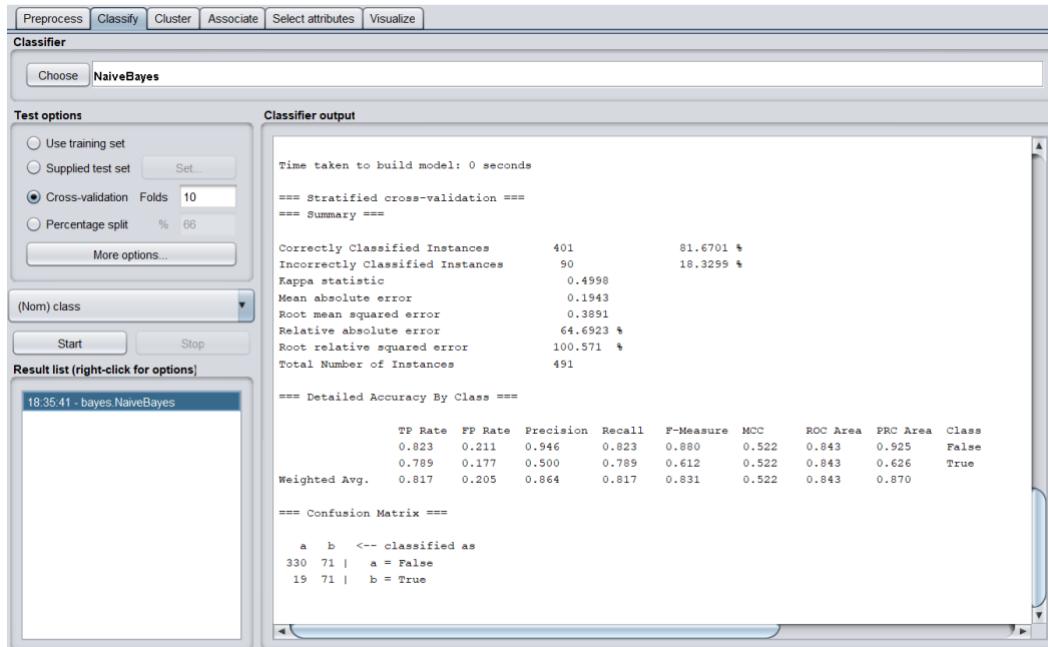
Test it with test dataset:



Test Performed on Daily_5 Dataset

1. Naïve Bayes

Build the model with training dataset:



Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. Under 'Classifier', 'NaiveBayes' is chosen. In the 'Test options' panel, 'Supplied test set' is selected. The 'Classifier output' pane displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===
    Correctly Classified Instances      186      77.5 %
    Incorrectly Classified Instances   54       22.5 %
    Kappa statistic                   0.4074
    Mean absolute error               0.2403
    Root mean squared error           0.4426
    Relative absolute error            80.7565 %
    Root relative squared error       115.3969 %
    Total Number of Instances         240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.782     0.256     0.933      0.782     0.851      0.435     0.797     0.919     False
    0.744     0.218     0.427      0.744     0.542      0.435     0.797     0.531     True
    Weighted Avg.                      0.775     0.249     0.843      0.775     0.796      0.435     0.797     0.850

    === Confusion Matrix ===
    a   b   <-- classified as
    154 43 |  a = False
    11 32 |  b = True
  
```

2. SGD

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. Under 'Classifier', 'SGD-F 0-L 0.01-R 1.0E-4-E 500-C 0.001-S 1' is chosen. In the 'Test options' panel, 'Cross-validation' is selected with 'Folds' set to 10. The 'Classifier output' pane displays the following evaluation results:

```

    Time taken to build model: 0.01 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      434      88.391 %
    Incorrectly Classified Instances   57       11.609 %
    Kappa statistic                   0.5569
    Mean absolute error               0.1161
    Root mean squared error           0.3407
    Relative absolute error            38.6571 %
    Root relative squared error       88.0605 %
    Total Number of Instances         491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.965     0.478     0.900      0.965     0.931      0.572     0.744     0.897     False
    0.522     0.035     0.770      0.522     0.623      0.572     0.744     0.490     True
    Weighted Avg.                      0.884     0.397     0.876      0.884     0.875      0.572     0.744     0.822

    === Confusion Matrix ===
    a   b   <-- classified as
    387 14 |  a = False
    43 47 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' section, 'SGD -F 0-L 0.01-R 1.0E-4-E 500-C 0.001-S 1' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the evaluation results on the test set.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0 seconds

    === Summary ===

    Correctly Classified Instances      209          87.0833 %
    Incorrectly Classified Instances   31           12.9167 %
    Kappa statistic                   0.5121
    Mean absolute error               0.1292
    Root mean squared error          0.3594
    Relative absolute error           43.4016 %
    Root relative squared error      93.7078 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.949    0.488    0.899     0.949    0.923     0.520   0.730     0.895     False
    0.512    0.051    0.688     0.512    0.587     0.520   0.730     0.439     True
    Weighted Avg.                     0.871    0.410    0.861     0.871    0.863     0.520   0.730     0.813

    === Confusion Matrix ===

    a   b   <-- classified as
    187 10 |  a = False
    21 22 |  b = True
  
```

3. J48

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. In the 'Classifier' section, 'J48 -C 0.25 -M 2' is chosen. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the evaluation results using stratified cross-validation.

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===

    Correctly Classified Instances      442          90.0204 %
    Incorrectly Classified Instances   49           9.9796 %
    Kappa statistic                   0.6436
    Mean absolute error               0.1375
    Root mean squared error          0.2996
    Relative absolute error           45.7982 %
    Root relative squared error      77.4406 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===

    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.958    0.356    0.923     0.958    0.940     0.648   0.811     0.914     False
    0.644    0.042    0.773     0.644    0.703     0.648   0.811     0.609     True
    Weighted Avg.                     0.900    0.298    0.896     0.900    0.897     0.648   0.811     0.858

    === Confusion Matrix ===

    a   b   <-- classified as
    384 17 |  a = False
    32 58 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The classifier chosen is 'J48 - C 0.25 - M 2'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the evaluation results on the test set.

```

==== Evaluation on test set ====
Time taken to test model on supplied test set: 0 seconds

==== Summary ====
Correctly Classified Instances      210          87.5   %
Incorrectly Classified Instances    30           12.5   %
Kappa statistic                      0.5417
Mean absolute error                  0.1582
Root mean squared error              0.3347
Relative absolute error              53.1535 %
Root relative squared error         87.2804 %
Total Number of Instances            240

==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.944   0.442    0.907    0.944   0.925    0.546  0.837   0.940   False
      0.558   0.056    0.686    0.558   0.615    0.546  0.837   0.527   True
Weighted Avg.                      0.875   0.373    0.868    0.875   0.870    0.546  0.837   0.866

==== Confusion Matrix ====
      a     b  <-- classified as
186  11 |  a = False
19   24 |  b = True
  
```

4. Random Forest

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The classifier chosen is 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' panel shows 'Cross-validation' selected with 'Folds 10'. The 'Classifier output' panel displays the evaluation results on the training set.

```

Time taken to build model: 0.07 seconds

==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances      440          89.613   %
Incorrectly Classified Instances    51           10.387   %
Kappa statistic                      0.6074
Mean absolute error                  0.1526
Root mean squared error              0.2808
Relative absolute error              50.8307 %
Root relative squared error         72.5709 %
Total Number of Instances            491

==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.970   0.433    0.909    0.970   0.938    0.621  0.903   0.968   False
      0.567   0.030    0.810    0.567   0.667    0.621  0.903   0.773   True
Weighted Avg.                      0.896   0.359    0.891    0.896   0.889    0.621  0.903   0.932

==== Confusion Matrix ====
      a     b  <-- classified as
389  12 |  a = False
39   51 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Choose' dropdown is set to 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the following evaluation results:

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0.01 seconds

    === Summary ===
    Correctly Classified Instances      211      87.9167 %
    Incorrectly Classified Instances   29       12.0833 %
    Kappa statistic                   0.5613
    Mean absolute error              0.1591
    Root mean squared error          0.2926
    Relative absolute error           53.468 %
    Root relative squared error     76.3023 %
    Total Number of Instances        240

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.944    0.419     0.912      0.944    0.928      0.564   0.896     0.964     False
    0.581    0.056     0.694      0.581    0.633      0.564   0.896     0.725     True
    Weighted Avg.                     0.879    0.354     0.873      0.879    0.875      0.564   0.896     0.921

    === Confusion Matrix ===
    a   b   <-- classified as
    186 11 |  a = False
    18 25 |  b = True
  
```

5. IBk (k = 10)

Build the model with training dataset:

The screenshot shows the Weka interface with the 'Classify' tab selected. The 'Choose' dropdown is set to 'IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last"''. The 'Test options' panel shows 'Cross-validation' with 'Folds 10' selected. The 'Classifier output' panel displays the following evaluation results:

```

    Time taken to build model: 0 seconds

    === Stratified cross-validation ===
    === Summary ===
    Correctly Classified Instances      429      87.3727 %
    Incorrectly Classified Instances   62       12.6273 %
    Kappa statistic                   0.5671
    Mean absolute error              0.128
    Root mean squared error          0.3546
    Relative absolute error           42.6088 %
    Root relative squared error     91.6364 %
    Total Number of Instances        491

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.930    0.378     0.916      0.930    0.923      0.568   0.771     0.908     False
    0.622    0.070     0.667      0.622    0.644      0.568   0.771     0.484     True
    Weighted Avg.                     0.874    0.321     0.871      0.874    0.872      0.568   0.771     0.831

    === Confusion Matrix ===
    a   b   <-- classified as
    373 28 |  a = False
    34 56 |  b = True
  
```

Test it with test dataset:

The screenshot shows the Weka interface with the Classifier tab selected. In the 'Choose' field, 'IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last"' is entered.

Test options:

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output:

```

==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds

==== Summary ====
Correctly Classified Instances      206      85.8333 %
Incorrectly Classified Instances    34      14.1667 %
Kappa statistic                   0.5353
Mean absolute error               0.1431
Root mean squared error          0.3756
Relative absolute error           48.0902 %
Root relative squared error      97.9395 %
Total Number of Instances         240

==== Detailed Accuracy By Class ====
      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
      0.904     0.349     0.922     0.904     0.913     0.536     0.777     0.912     False
      0.651     0.096     0.596     0.651     0.622     0.536     0.777     0.450     True
Weighted Avg.                     0.858     0.304     0.864     0.858     0.861     0.536     0.777     0.830

==== Confusion Matrix ====
      a   b   <-- classified as
178  19 |  a = False
  5  28 |  b = True
  
```

Result list (right-click for options):

- 18:35:41 - bayes.NaiveBayes
- 18:36:04 - bayes.NaiveBayes
- 18:36:22 - functions.SGD
- 18:36:44 - functions.SGD
- 18:37:03 - trees.J48
- 18:37:23 - trees.J48
- 18:37:41 - trees.RandomForest
- 18:38:08 - trees.RandomForest
- 18:38:31 - lazy.IBk
- 18:38:59 - lazy.IBk**

Justification of Best Model

1. Analysis

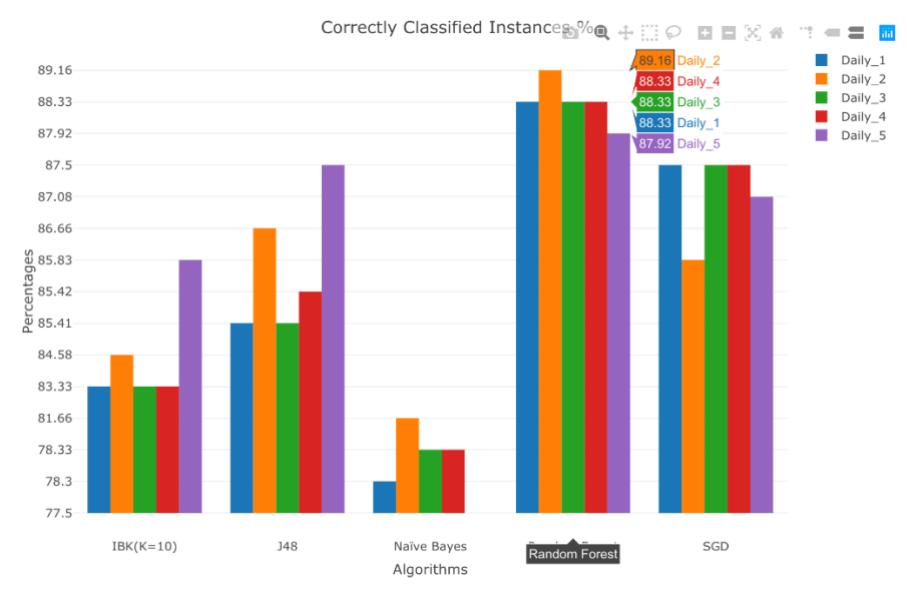
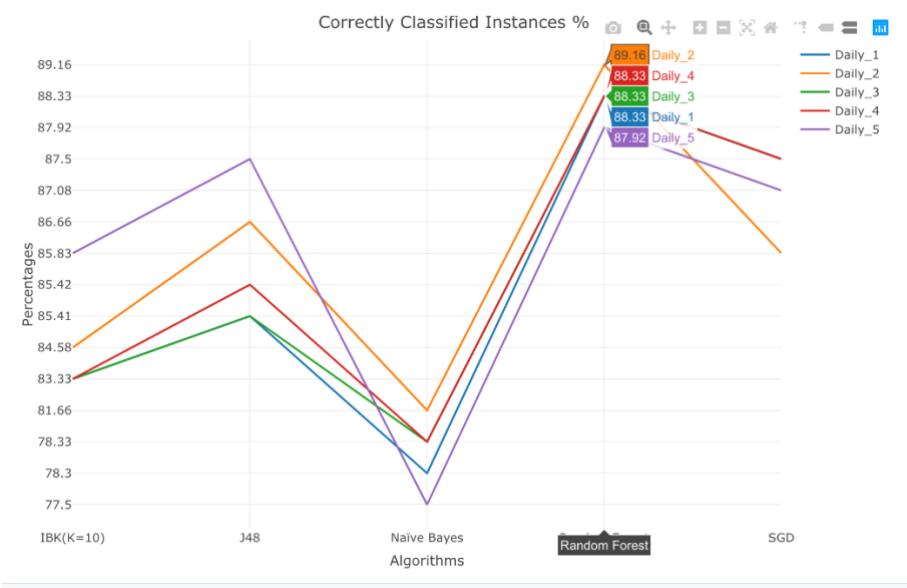
After performing the test on all the datasets created using all the classification model build, we decided to choose Accuracy(%), TP rate, FP Rate, ROC Area and F-Measure as the appropriate metrics for performing comparison. Also since Test statistics help us in deciding how useful a model is in classifying new instances correctly, we will be performing comparison on basis of metrics observed on test datasets.

Dataset	Classifier	Accuracy (%)	TP Rate	FP Rate	Roc Area	F-Measure
Daily_1	Naïve Bayes	78.33%	0.767	0.213	0.807	0.559
	SGD	87.50%	0.605	0.066	0.769	0.634
	J48	85.42%	0.535	0.076	0.785	0.568
	Random Forest	88.33%	0.605	0.056	0.909	0.65
	IBK(K=10)	83.33%	0.605	0.117	0.744	0.565
Daily_2	Naïve Bayes	81.67%	0.721	0.162	0.814	0.585
	SGD	85.83%	0.535	0.071	0.732	0.575
	J48	86.67%	0.535	0.061	0.797	0.59
	Random Forest	89.17%	0.651	0.056	0.876	0.683
	IBK(K=10)	84.58%	0.605	0.102	0.752	0.584
Daily_3	Naïve Bayes	78.33%	0.767	0.213	0.807	0.559
	SGD	87.50%	0.605	0.066	0.769	0.634
	J48	85.42%	0.535	0.076	0.785	0.568
	Random Forest	88.33%	0.605	0.056	0.909	0.65
	IBK(K=10)	83.33%	0.605	0.117	0.744	0.565
Daily_4	Naïve Bayes	78.33%	0.767	0.213	0.807	0.559
	SGD	87.50%	0.605	0.066	0.769	0.634
	J48	85.42%	0.535	0.076	0.785	0.568
	Random Forest	88.33%	0.605	0.056	0.909	0.65
	IBK(K=10)	83.33%	0.605	0.117	0.744	0.565
Daily_5	Naïve Bayes	77.50%	0.744	0.218	0.797	0.542
	SGD	87.08%	0.512	0.051	0.73	0.587
	J48	87.50%	0.558	0.056	0.837	0.615
	Random Forest	87.92%	0.581	0.056	0.896	0.633
	IBK(K=10)	85.83%	0.651	0.096	0.777	0.622

Metrics Observed on Test Dataset(For TRUE Class)

2. Results

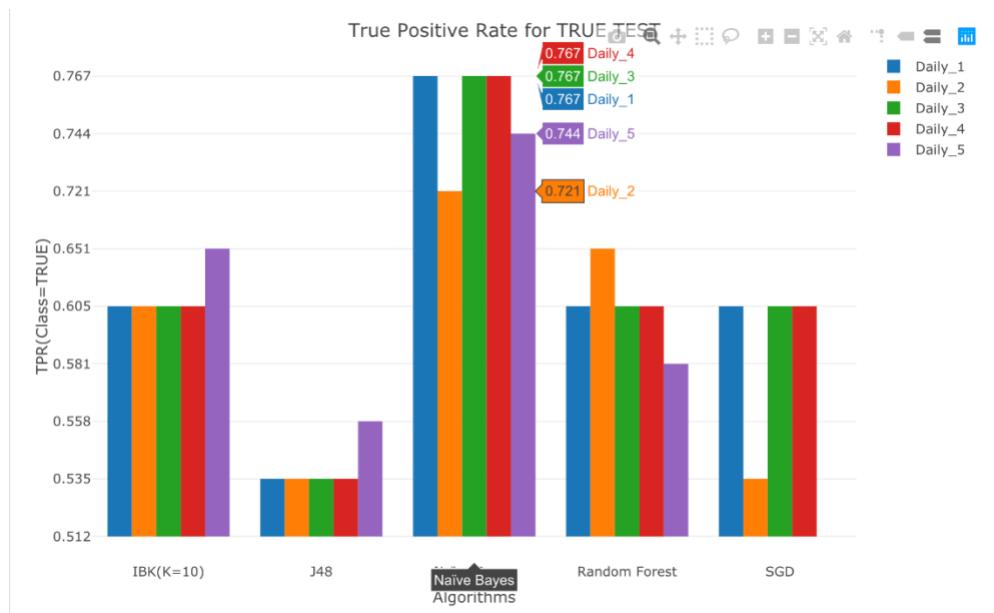
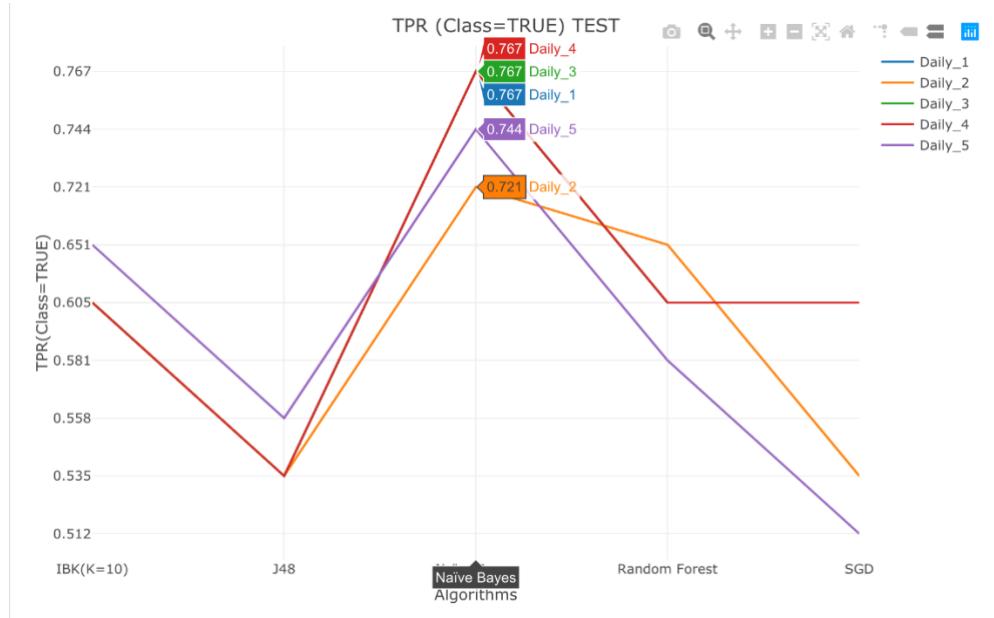
1) Correctly Classified Instances (Accuracy (%))



As we can see clearly that the test model of **Random Forest** tested on dataset **Daily_2** has the highest accuracy, which is 89.1667%.

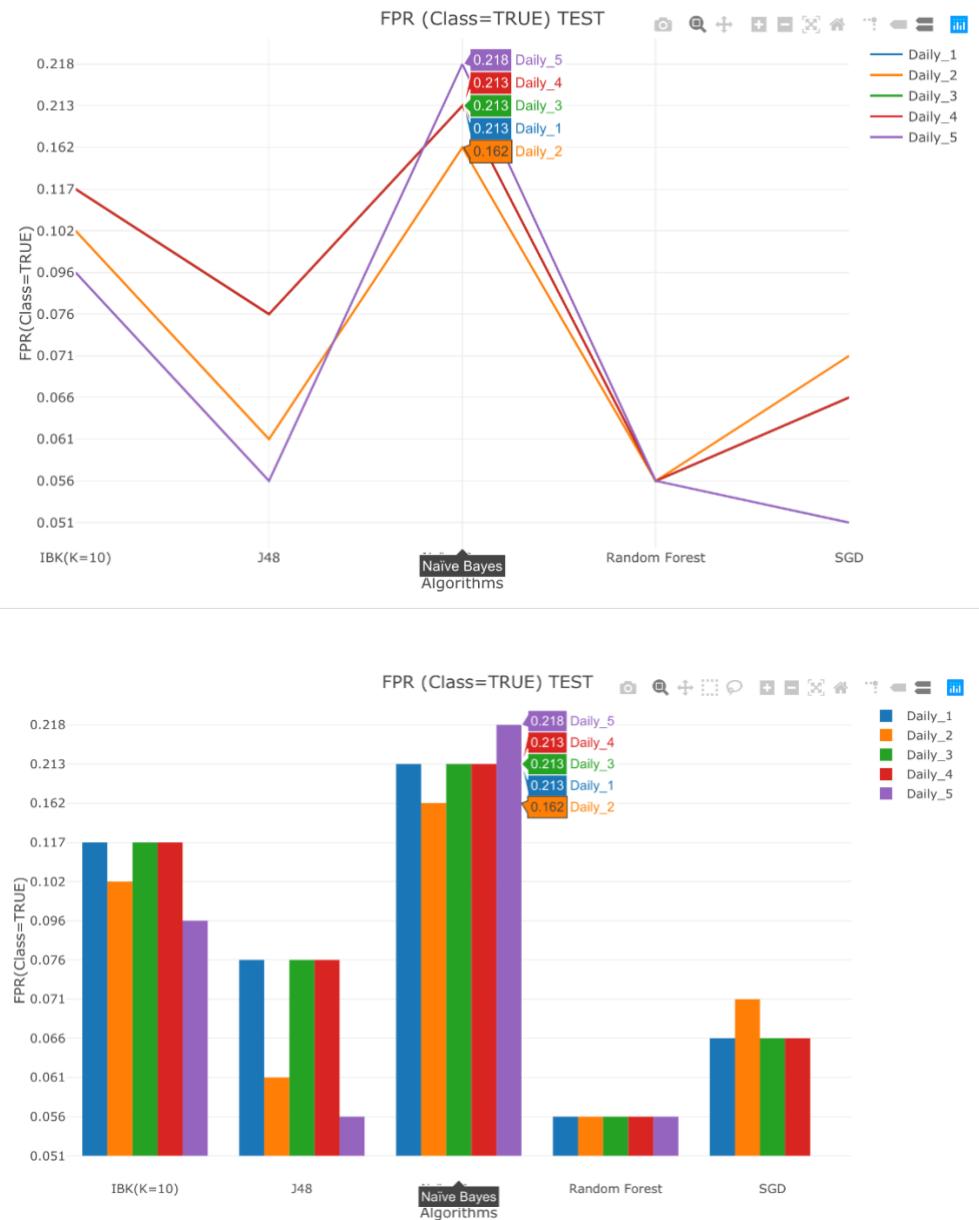
2) TP Rate for TRUE

For our model classifying the occurrence is crucial, that's why we will be comparing TP Rate for TRUE class values only.



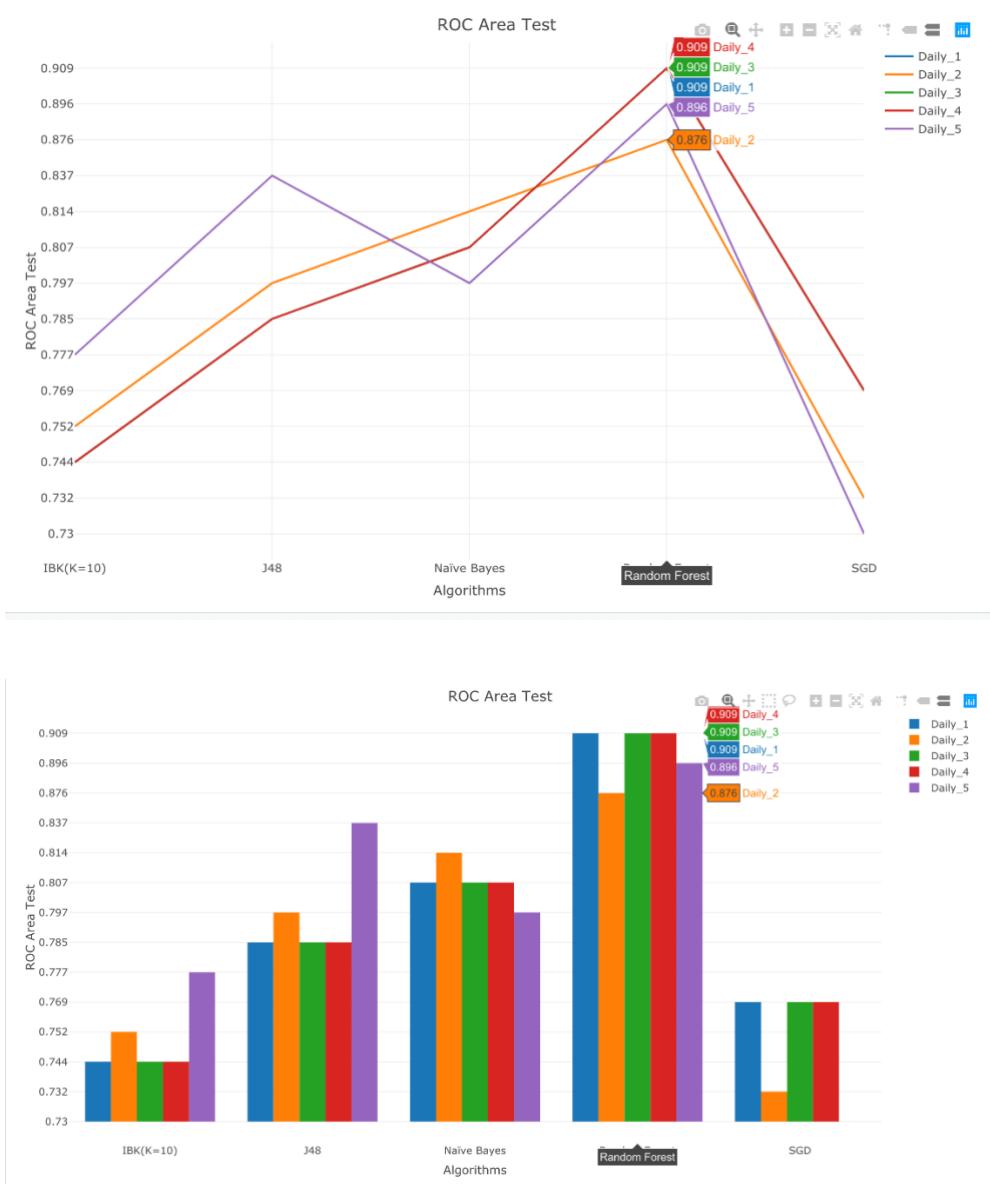
As we can see clearly that the test model of **Naïve Bayes** tested on datatset **Daily_1, Daily_3, and Daily_4** have the highest TP Rate for Class TRUE, which is 0.767.

3) FP Rate for TRUE



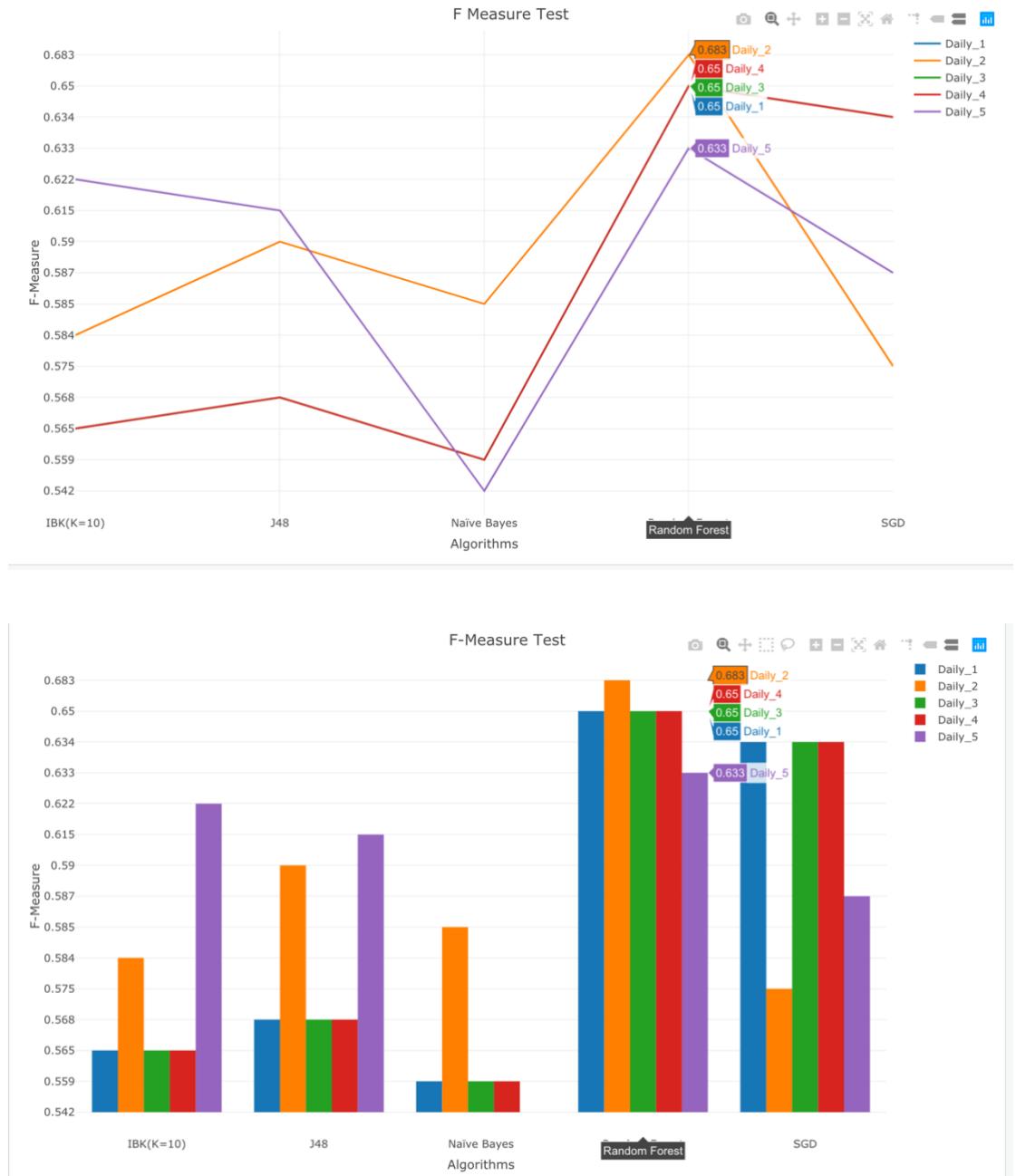
As we can see clearly that the test model of **Naïve Bayes** tested on **Daily_5** has the highest FP Rate for Class TRUE, which is 0.218.

4) Roc Area



As we can see clearly that the test model of **Random Forest** tested on **Daily_1, Daily_3, and Daily_4** have the largest Roc Area, which is 0.909.

5) F-Measure



As we can see clearly that the test model of **Random Forest** tested on **Daily_2** have the largest F-Measure which is 0.683.

3. Conclusion

After comparing all the metrics together, we think the **best classifier model** would be **Random Forest** built on **dataset Daily_2**. It has the highest accuracy 89.1667%, with TP Rate of 0.651 for class TRUE, TP Rate of 0.944 for class FALSE, FP Rate of 0.056 for class TRUE, FP Rate of 0.349 for class FALSE, 0.876 Roc Area and highest F-Measure 0.683.

SUMMARY

For this data mining project, we start with a raw dataset about weather information of Brentwood at San Francisco Bay area from January 1, 2018 to January 1, 2020. Our goal is to predict the occurrence of rain in this area.

We struggled a bit at which attributes should be eliminated manually in the dataset as a part of preprocessing, based on our testing, different columns reductions have different impact on our results. After comparing different reduction methods, we choose to only keep the attributes which are correlated with our data mining goal and remove the attributes having more than 95% of the values missing. We learned about our dataset and accordingly took the decision.

We also learned how to use the attribute selection algorithms and classification algorithms in Weka, as well as splitting dataset. And we understand the difference between “Use training...”, “Supplied test...”, and “Cross-validation...”. “Use training...” is not quite useful, as we observed significant difference in accuracy(%) , moreover it was unable to evaluate the metrics for TRUE class values. Validation dataset is predominately used to describe the evaluation of models when tuning hyperparameter and data preparation, and the Test dataset is predominately used to describe the evaluation of a final tuned model when comparing it to other final models. This is the reason why we choose the test dataset to process our model comparison.

For this project, both of group members had contributed a considerable amount of time. We generally split our project into two part, Jyoti takes care of the first three attribute select algorithms as well as its' classification algorithms, the dataset splitting explanation, and the model justification graphs. Siyu takes care of the last two attribute select algorithms as wells its' classification algorithms, the dataset preprocess, and wrapped all of them up and finish the model justification and conclusion. We have several versions of our report, and both of group members modified it and made it better.

Bibliography

CIMIS (2020). CIMIS Station Reports. Retrieved from <https://cimis.water.ca.gov>.

Wikipedia (February 29, 2020). C4.5 algorithm. Retrieved from
https://en.wikipedia.org/wiki/C4.5_algorithm.

Wikipedia (March 8, 2020). Naïve Bayes classifier. Retrieved from
https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

Wikipedia (March 12, 2020). Random forest. Retrieved from
https://en.wikipedia.org/wiki/Random_forest.

Wikipedia (Octover 26, 2019). Multilayer perceptron. Retrieved from
https://en.wikipedia.org/wiki/Multilayer_perceptron.