# Data Mining Project Assignment Shark Attack Analysis

# **Abstract**

The state of North Carolina has seen the most attacks in one year in the 80 years for which the Global Shark Attack File keeps records. The highest previous total was five attacks in 2011. In this summer the shark problem is prominent. In this study we have analyzed diverse data sets related to sharks and have discovered various patterns and factors that caused a hike in shark attacks. We have found patterns which have caused shark attack which can be categorized as general meteorological, oceanographic data (like ocean current, sea level), moon phases and increase in human population or in tourist population. The most remarkable one related to ocean current, this years (2015) weakened Gulf Stream which brought more warm water near to coast.

# **Problem Definition**

If you live in North Carolina you know there has been a “shark problem” this summer at the shore. What is causing the problem? Is it really a problem? Is there a way to predict when/where shark attacks are more likely to occur? Is North Carolina really that unusual? What other questions are useful to explore?

What all aspects have trigger a spike in shark attacks? Acquire as much as knowledge about sharks feeding behavior, migratory patterns. What was different in this year? Is there a pattern of changes in ocean current? How much impact has moon phase and tidal currents had made? Is there a relation to this year’s weakened Gulf Stream and Labrador Currents? Increasing flood is it a result of variation in ocean current and increased sea level?

This study will gather all possible data related to above questions and will approach data mining methodology with help of Weka and understand, analyze and draw conclusions for the problem of increasing shark attacks in North Carolina.

This data mining project will consists of the data analysis of recent shark attacks in North Carolina coast and understand the reason of hike in attacks against various factors which would have triggered it.

Based on our domain knowledge we have selected below aspects of the shark attacks:

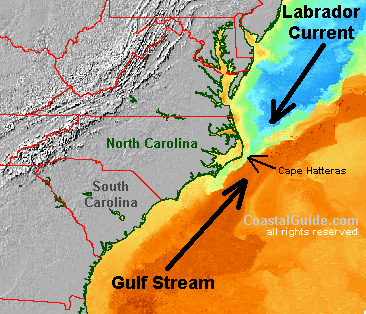
1. Ocean currents
2. Ocean Tide
3. Wind speed and Direction
4. Salinity
5. Sea Pressure
6. Ocean temperature
7. Precipitation rate
8. Shark prey population
9. Moon phase

We will collects the data from above and merge with Shark attack files based on the date of attack and place of attack. Against those parameters we will preprocess, cleanse, transform, and de-normalize the data.

We will use classification, association rules, clustering algorithms to process the data and then create results against most relevant factors from above list. Once the results are analyzed we will select most accurate results and prepare report of the reasons of shark attacks.

# **Domain Knowledge**

Knowledge of common types of shark species found in North Carolina. Data is taken for all cities of NC and coastal areas including shark attacked areas like Oak Island, Surf City, Cape Hatteras and Ocracoke Island. On further thorough study we chose to focus our analysis on finding the causes of shark attack in the North Carolina coastal areas. And as our analysis pointed to prominent influence of ocean currents, we choose to collect data set and focus on observations near to Cape Hatteras which is meeting point of the warmer Gulf stream current and colder Labrador current.



Source: http://elephantsandredwoods.blogspot.com/2011/08/irene-irene.html

We understand shark has the greatest electrical sensitivity of any animal. Sharks find prey hidden in sand by detecting the electric fields they produce. Ocean currents moving in the magnetic field of the Earth also generate electric fields that sharks use for orientation and navigation.

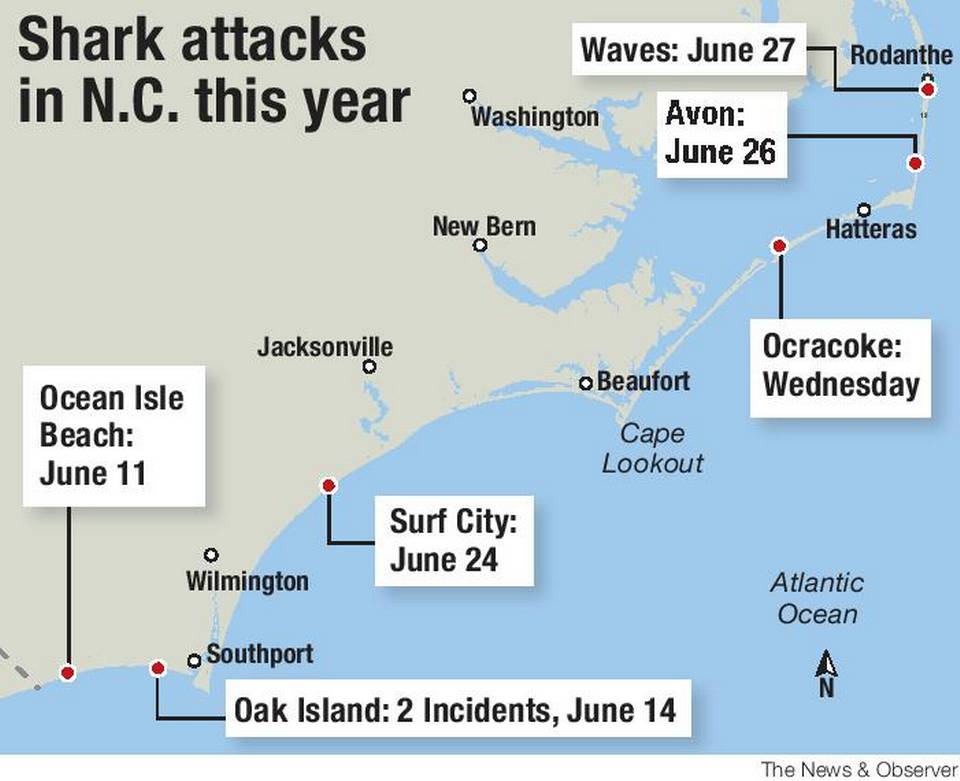
Sharks navigate mainly by electro-reception and a sensitivity to pressure and water temperature (mechanism called Magnetoception). Located at the front of a shark's head is an extremely sensitive system of bio-electrical sensors which we believe helps them to navigate in response to the ocean's currents and the earth's electro-magnetic field. One part of this system, the ampullae of Lorenzini detect faint electrical currents given off by the earth's magnetic field -- or by other living creatures. This seem to function like an internal compass, helping sharks to orient themselves in relation to north-south (the strongest electro-magnetic fields) and east-west (weaker) electrical fields.

The above behavior of sharks have prompted to learn how they detect direction of migration with the help of ocean currents. And data analysis pointed to variation in ocean currents this year which had increased influence on moving the migrating sharks closer to the coast especially on areas near to Cape Hatteras which is the merging point of Gulf Stream and Labrador currents.

Studied ocean current flow and impact of Gulf Stream, Labrador currents and this year’s weakened current flow. Analyzed ice berg movement changes in Artic region which was rare and this has implied variation in Labrador Current. Also changes in Gulf Stream which have raised sea level in North Carolina which has concreted evidence of weakened stream and ocean current variation.

As warmer water attracts more sharks, sharks follow their food, and the currents are bringing more shark food to the shore. Warmer water is brought to shore by Gulf Stream and it doesn’t have actual boundaries like a river or stream. Since it runs basically on top of and mixing together with the Atlantic Ocean, the Gulf Stream's current can meander close to North Carolina or further away depending on the season and the prevailing weather conditions. During the winter months the current may stay approximately 20 miles off our coast but during the summer because of the way our location sticks out into the Atlantic, the Gulf Stream and its warming effects can come extremely close to Hatteras Inlet.

This years increased ocean temperature showed perfect condition for bait fish and their predator sharks. This was strengthened by the pattern of attack locations.



Source: http://www.charlotteobserver.com/news/local/article26025094.html/

In addition to these we have analyzed common and recurring aspects of general meteorological, oceanographic data (like ocean current, sea level), moon phases and increase in human population.

Another feature we have taken to consideration is the prey population of the Sharks. We have taken into account the following Blue crab population, Herring migration along with Gulf Stream etc.

Learning of causes of both provoked and unprovoked shark attacks. Understand types of unprovoked attacks and impacts.

Explosion of tourists in North Carolina beaches and related stats.

Weather impacts like recent high temperature, less inflow of water which increased salinity.

Other aspects like elevated water temperature and water currents which distributed warm water.

Fish population stats which is part of shark food and hike of those in May-June months.

### Domain knowledge Sources

http://sharkattackfile.net/incidentlog.htm

http://www.data.gov/climate/

http://www.st.nmfs.noaa.gov/ecosystems/climate/activities/oceanadapt

<http://csi.northcarolina.edu/content/research/nccoos.htm>

<http://opendap.co-ops.nos.noaa.gov/ioos-dif-sos/>

<http://www.cbc.ca/news/canada/newfoundland-labrador/icebergs-moving-farther-east-this-year-says-ice-patrol-commander-1.3041817>

[http://www.newsobserver.com/news/politics-government/state-politics/article10208936.html#storylink=cpy](http://www.google.com/url?q=http%3A%2F%2Fwww.newsobserver.com%2Fnews%2Fpolitics-government%2Fstate-politics%2Farticle10208936.html%23storylink%3Dcpy&sa=D&sntz=1&usg=AFQjCNEjAPu8A2c_nya5aufxuXXexfQCuA)

[http://www.wral.com/revised-nc-sea-level-report-predicts-rise-along-entire-coast/14553253/#AsSVw9rQdb38vIKm.99](http://www.google.com/url?q=http%3A%2F%2Fwww.wral.com%2Frevised-nc-sea-level-report-predicts-rise-along-entire-coast%2F14553253%2F%23AsSVw9rQdb38vIKm.99&sa=D&sntz=1&usg=AFQjCNEIOM4keCXdaloJqlq5DnR476rSfQ)

[http://science.nasa.gov/earth-science/oceanography/physical-ocean/currents/](http://www.google.com/url?q=http%3A%2F%2Fscience.nasa.gov%2Fearth-science%2Foceanography%2Fphysical-ocean%2Fcurrents%2F&sa=D&sntz=1&usg=AFQjCNFZAEdA3Uw52NpvFTtFi8QFi4pInA)

<http://www.sea-angel-charters.com/gulf_stream.html>

<http://www.uncpress.unc.edu/browse/page/556>

**Input**

## Data Sources

Shark Attack File

<http://sharkattackfile.net/incidentlog.htm>

Latest News

Temperature Overall & Ocean Temperature

Hourly Rainfall data

<https://www.ncdc.noaa.gov/cdo-web/datasets>

Ocean water current

<https://www.nodc.noaa.gov/tsdb/oc_madcp.html>

Fish population

<http://www.st.nmfs.noaa.gov/commercial-fisheries/commercial-landings/>

Moon phase data

<http://www.almanac.com/moon/calendar/NC/Oak%20Island/2015-06>

Water salinity

<https://www.nodc.noaa.gov/General/salinity.html>

## Meta-data

Shark attack file

Date

Location

Time – Time of attack)

NOAA climate data set (includes temperature, precipitation)

PRCP - Precipitation (tenths of mm)

TOBS - Temperature at the time of observation (tenths of degrees C)

Blue Crab population

Date

Pounds

Global Temperature and Salinity Profile Data

Salinity

Ocean Current

Sea Level (in terms of MLLW - mean lower low water), which is the average height of the lowest tide recorded at a tide station each day during the recording period.

Wave height

Wind speed

Wind direction

Moon Phase

Illumination percentage – categorized into 4 phases

**Data Preprocessing**

**Data Cleansing**

1. Fill missing data

* Latest data in Shark Attack File was missing therefore referred the latest news and added required information. For example: 4th July 2015 attack details were added.
* On joining shark attack class attribute was added.

1. Identify irrelevant, outliers and smoothen noisy data

* Shark attack file had irrelevant data, those attributes which has not much relevance like Name, Investigator, Country, type of injury was removed.
* In Climate data set those outliers which are not impacting were identified by running clustering (SimpleKMeans) on the data.

1. Correct inconsistent data

Date field was in different format for almost all the database. Hence we converted into standard format by shell scripting.

1. Date field was converted to a uniform format of YYYYMMDD

**Data integration (denormalization)**

1. Merge multiple database

Used Google Refine / Open Refine for joining multiple files.

It was merged based on date attribute and attack location.

Merge database of Shark Attack File and Climatic Data File.

Merge database of Shark Attack File and Oceanographic Data.

Merge database of Shark Attack File and crab population.

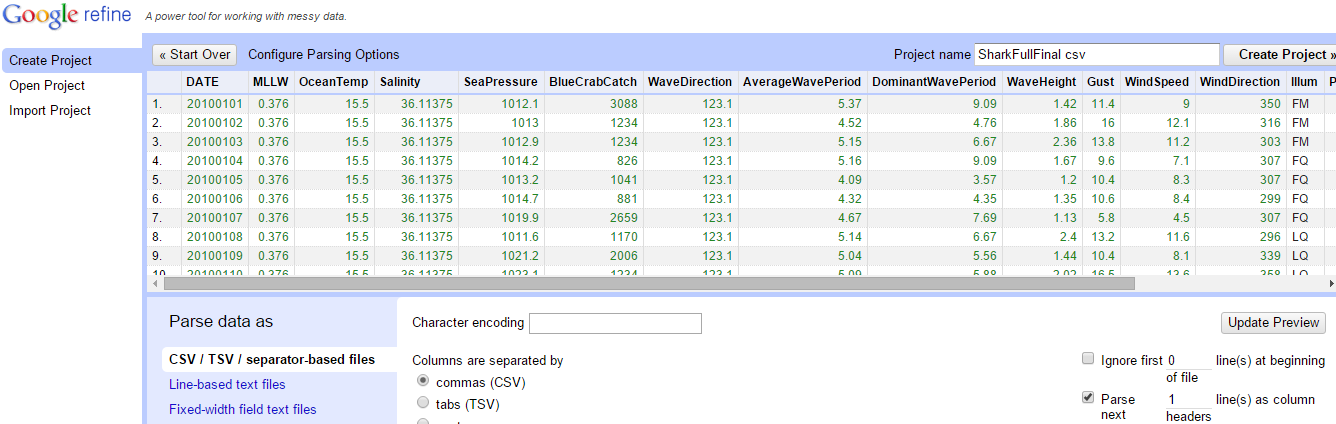
Merge database of Shark Attack File and moon phase.

1. Check for inconsistency or data value conflicts

Load CSV file in Weka and check for consistency issues or invalid data format issues.

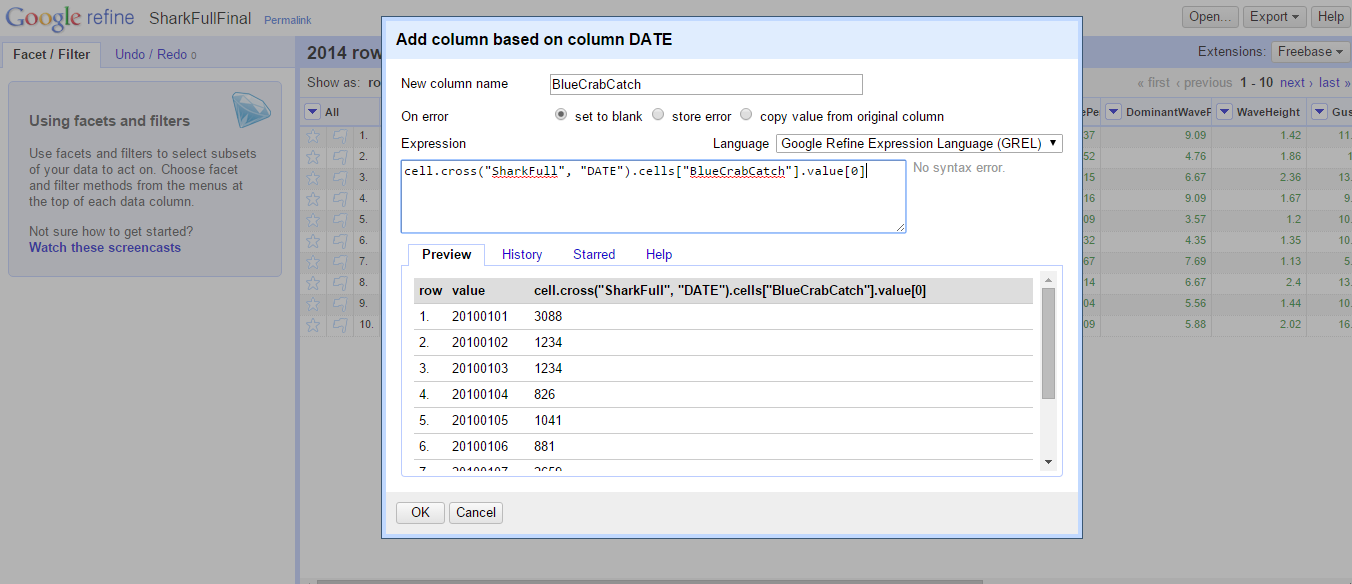
Fixed issues related to same column names in Climate Dataset.

We have used google refine for denormalizing. This ensured that join was based on the correct attribute and flawless. For this selected DATE field as key column. Below is the example of creating project and joining the blue crab data set with Shark Attack Data.



Using Google Refine Expression Language (GREL):

cell.cross("SharkFull", "DATE").cells["BlueCrabCatch"].value[0]



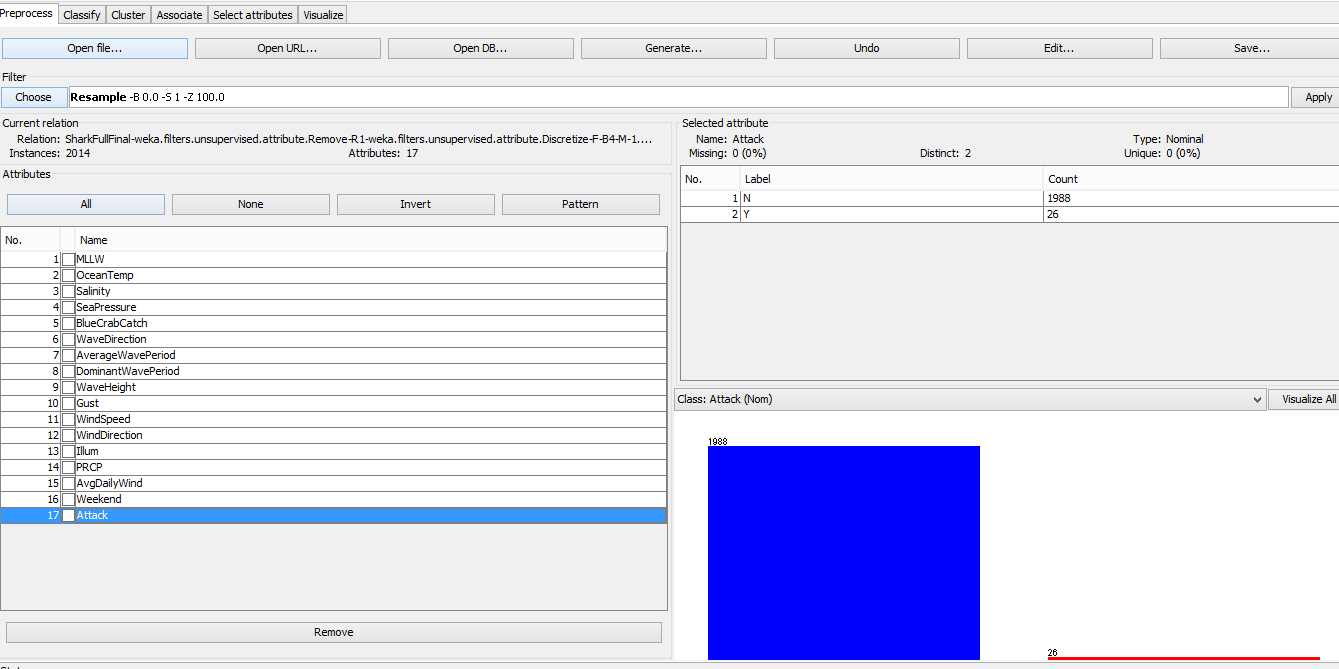
De-normalized the columns which are related. In the climate dataset attributes related to precipitation was consolidated and kept as single attribute.

**Data Transformation**

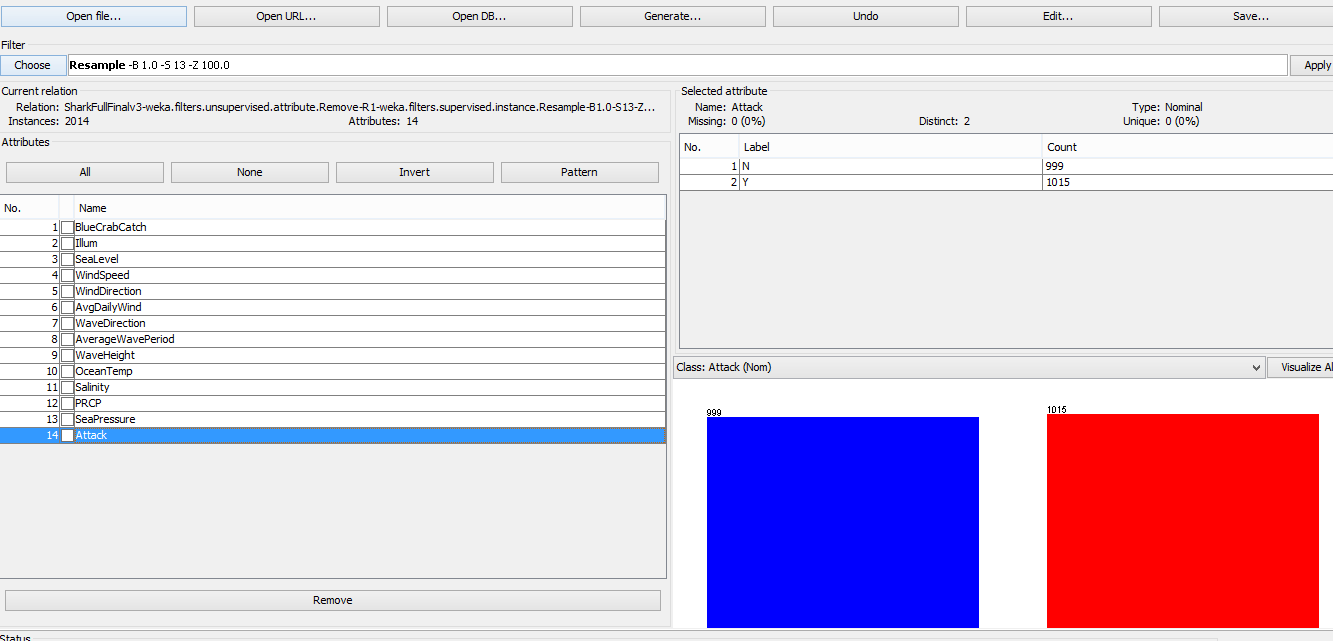
Once database merges, the data was transformed to calibrate which group of databases will give the best result. For instance, oceanographic data from our source NOAA was an hourly data. This was transformed into daily data. For this, we have taken time limit of the Shark attacks and then took the mean of instances in the time frame and then applied it.

With the help of Weka with integrated data set ran classifiers and got overfitting results. And could not run Apriori due to presence of numerical in the data set.

For data transformation used Weka and did discretization. For discretizing the data we used filters – “weka.filters.unsupervised.attribute.Discretize” as this would be suitable for our data. Here we changed number of bins to 4 and set useEqualFrequency=True. As a result we got 4 nominal groups which was good for classfiers.

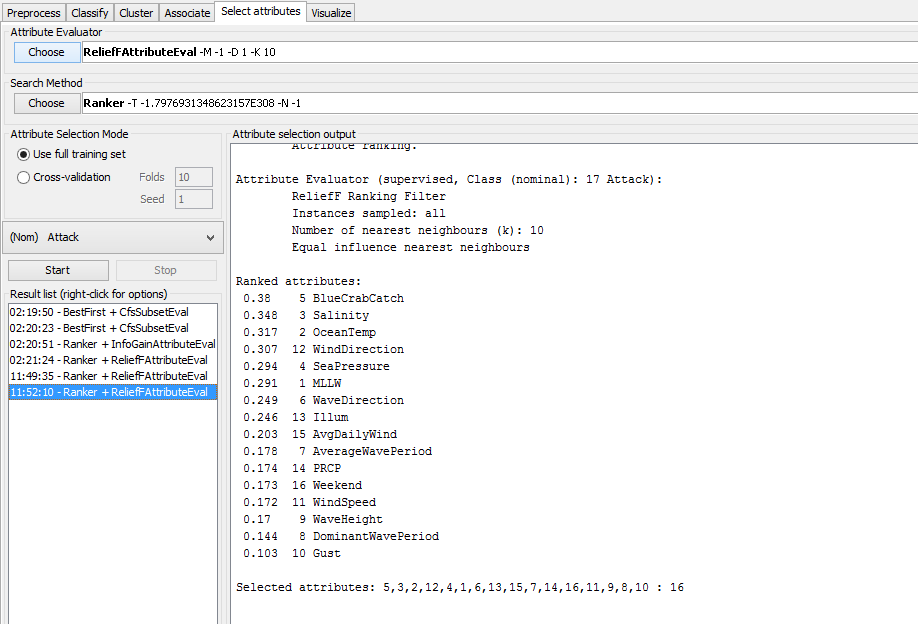


Thereby found that there is high amount of unbiased data as we have taken daily data and number of shark attacks are less, the class value of Attack=N was more baised. For normalizing this we used another filter “weka.filters.supervised.instance.Resample” and here we set biasToUniformClass=1 with seed as 13 for clearing the bias towards days without attack.



**Feature Selection**

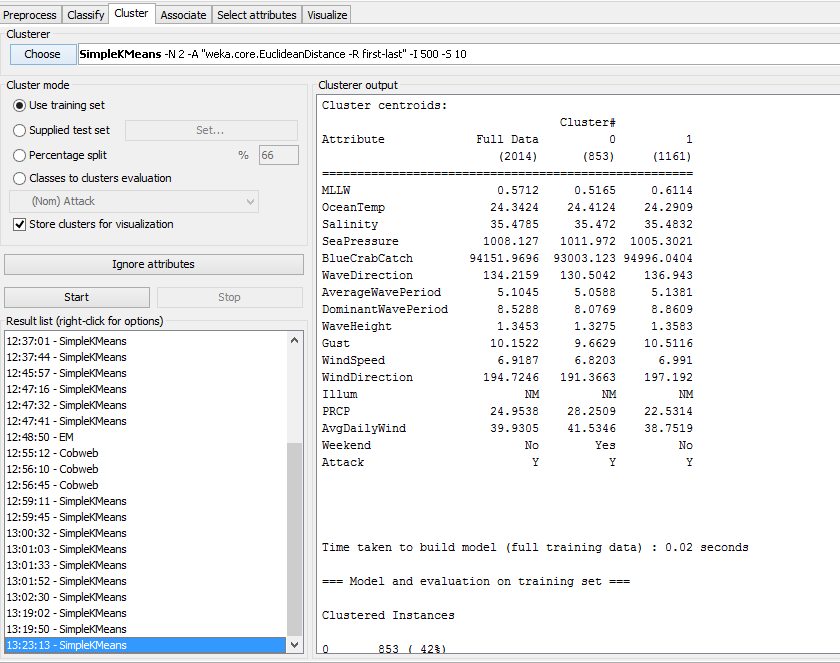
We used Weka “Select Attribute” tab with attribute evaluator as “weka.attributeSelection.ReliefFAttributeEval” and search method as “weka.attributeSelection.Ranker”. There by we got below results and found the attributes with least rank. Then removed them and ran classifier. It has increased the accuracy of the classifier.



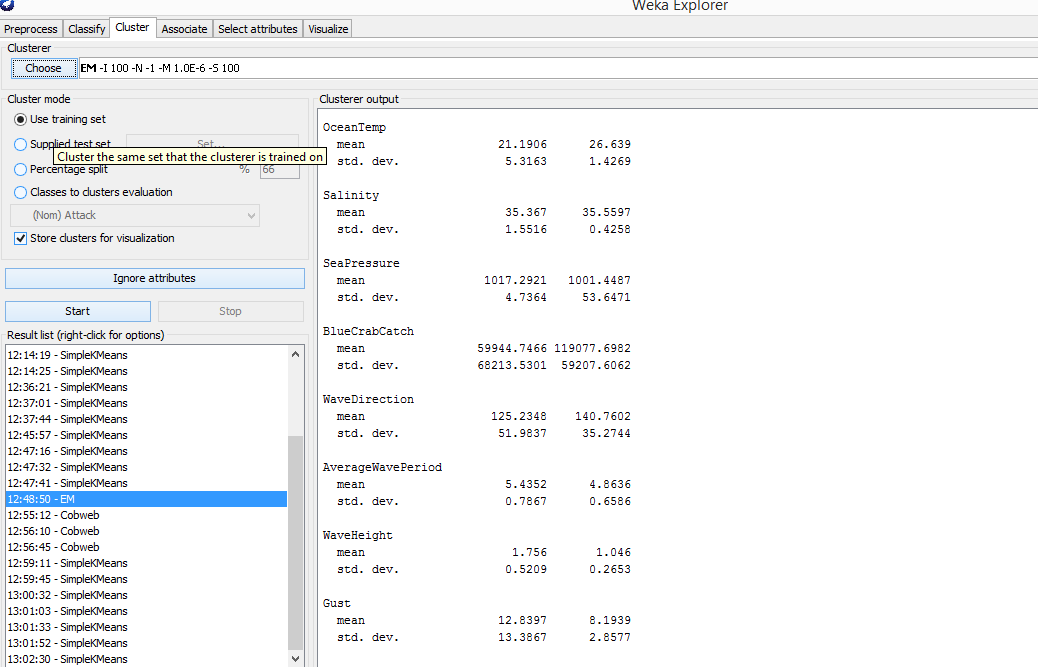
**Clustering**

We have used K-means clustering in Weka to cluster the data. The result window shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. SimpleKMeans algorithm automatically handles a mixture of categorical and numerical attributes.

Firstly, we tried with default values of the SimpleKMeans and we got following results:



Centroids were used to categorize the clusters. For example cluster 1 shows the segment where MLLW is 0.6114, ocean temp is 24.29 etc has attack said as Y.

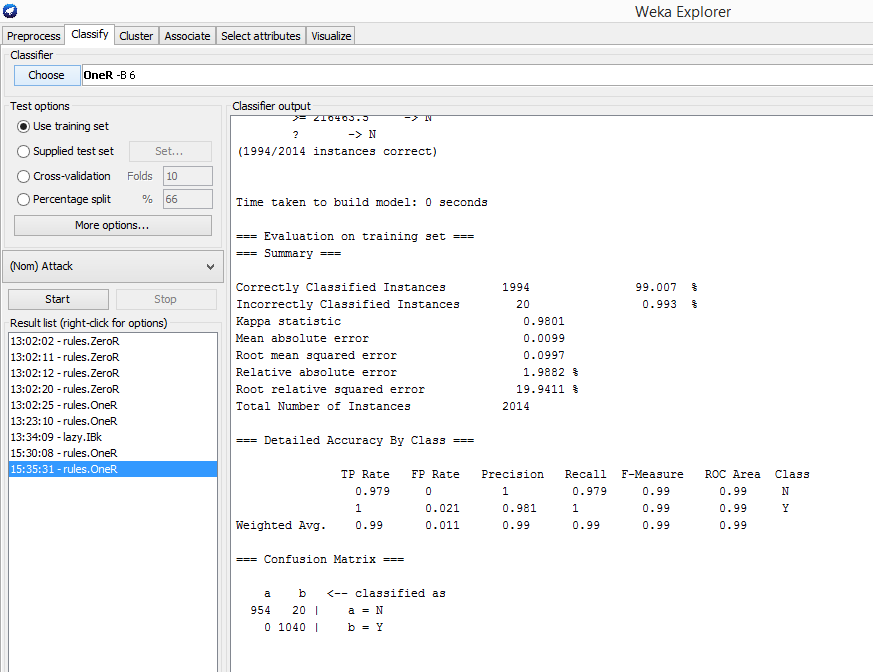


# **Processing**

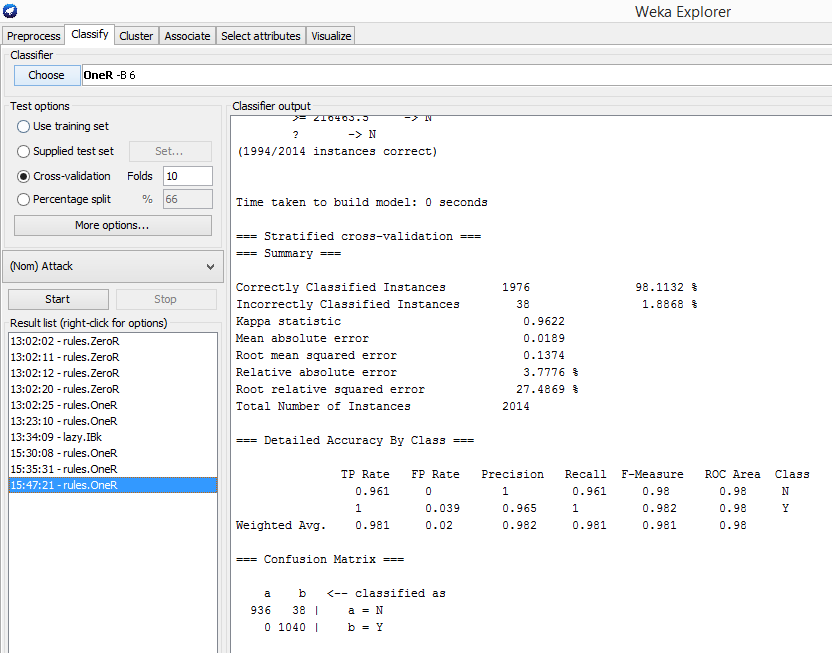
## Algorithms

### Rules - 1R

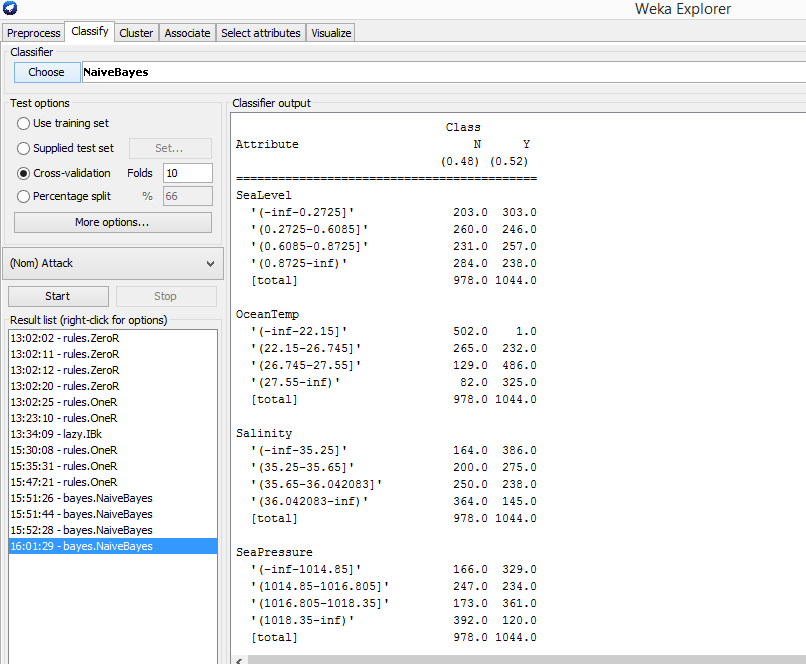
We have started with 1R which is simple classification by using training set



With cross-validation we got slightly less accurate results



### Naïve Bayes



Naïve Bayes gave information regarding most probable values in that particular vale range of the attribute. For instance we found less precipitation better probabilities of attack from below results:-

PRCP

'(-inf-1.5]' 657.0 700.0

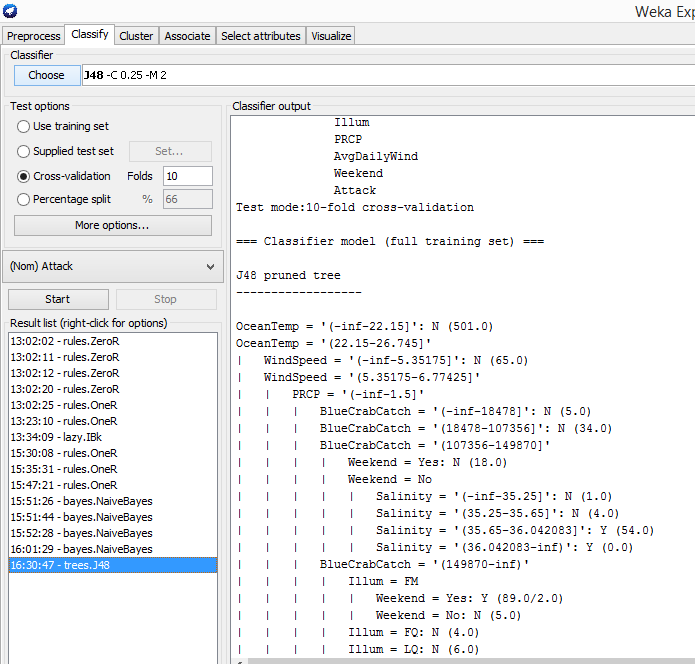
'(1.5-11.5]' 76.0 148.0

'(11.5-72.5]' 114.0 109.0

'(72.5-inf)' 131.0 87.0

[total] 978.0 1044.0

### Decision Trees - J48



From this we got following results:-

PRCP = '(-inf-1.5]'

| | | BlueCrabCatch = '(149870-inf)'

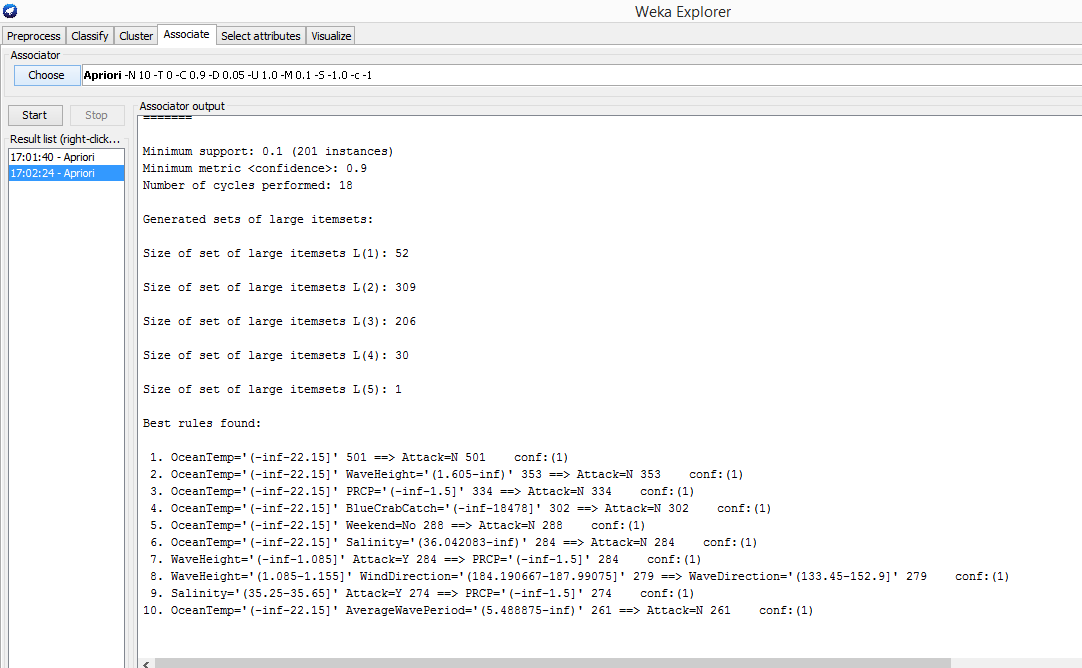
| | | | Illum = FM

| | | | | Weekend = Yes: Y (89.0/2.0)

This means Precipitation is less, BlueCrabCatch is high , Illum is that of Full Moon, Weekend is Y , attack will be more.

### Association Rules – Apriori

For getting Association rules, we ran Apriori and found out associations.



Here have interesting results like:

OceanTemp='(-inf-22.15]' 501 ==> Attack=N 501 <conf:(1)>

Salinity='(35.25-35.65]' Attack=Y 274 ==> PRCP='(-inf-1.5]' 274 <conf:(1)>

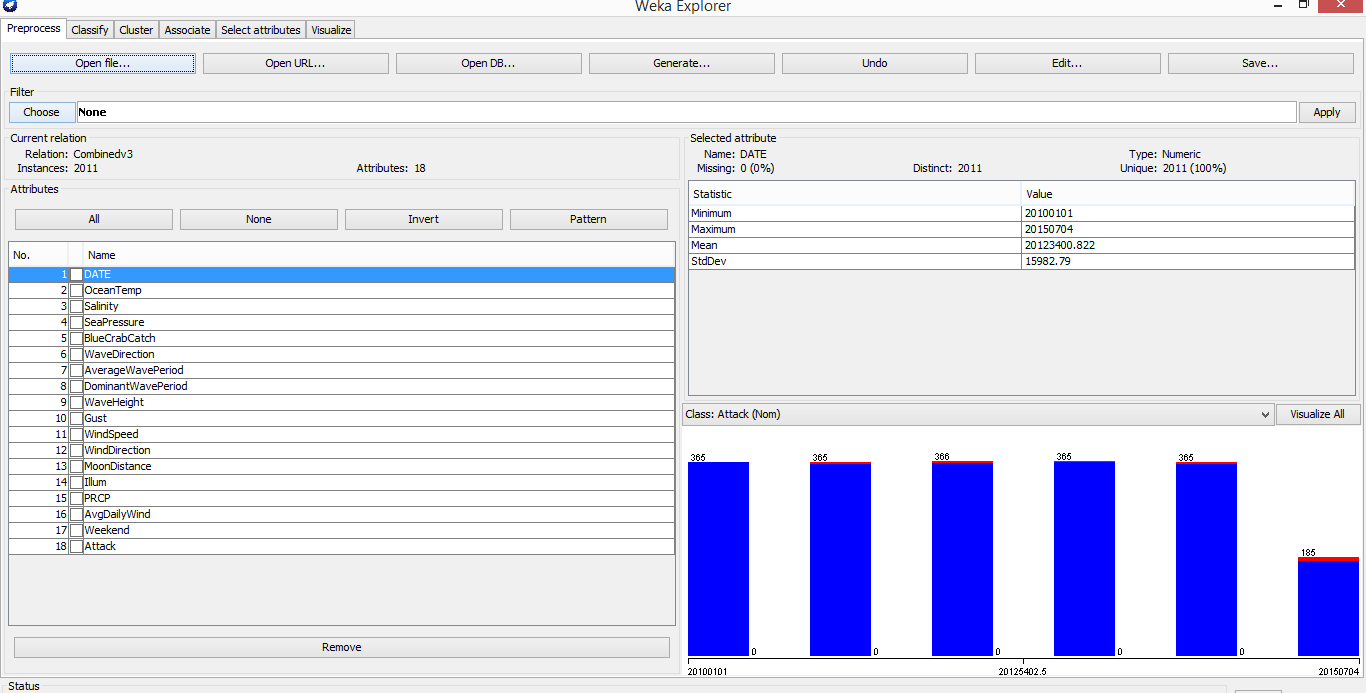
WaveHeight='(-inf-1.085]' Attack=Y 284 ==> PRCP='(-inf-1.5]' 284 conf:(1)

### Clustering – Simple K Means and EM

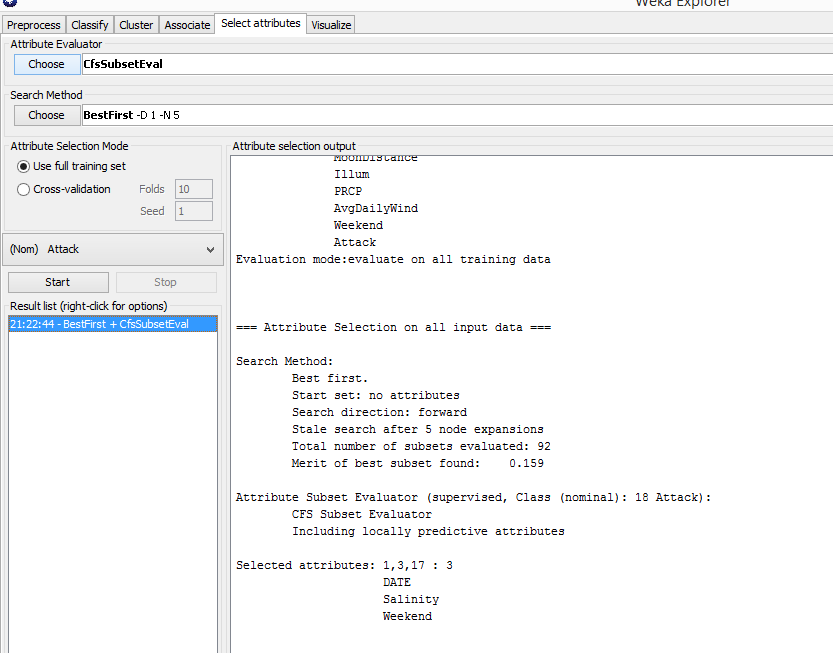
Mentioned earlier

### Iterative Evidence

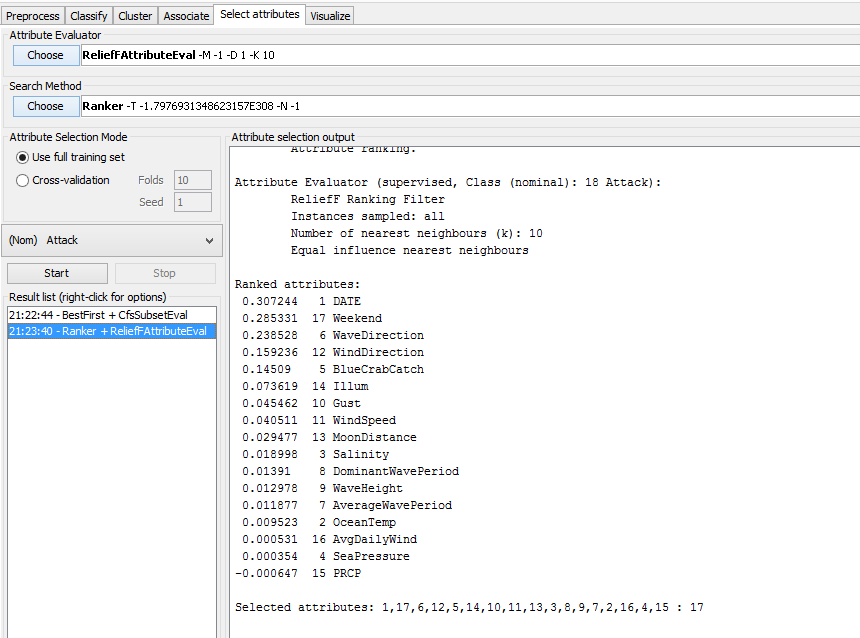
Initially 18 attributes



Select Attributes to find good features

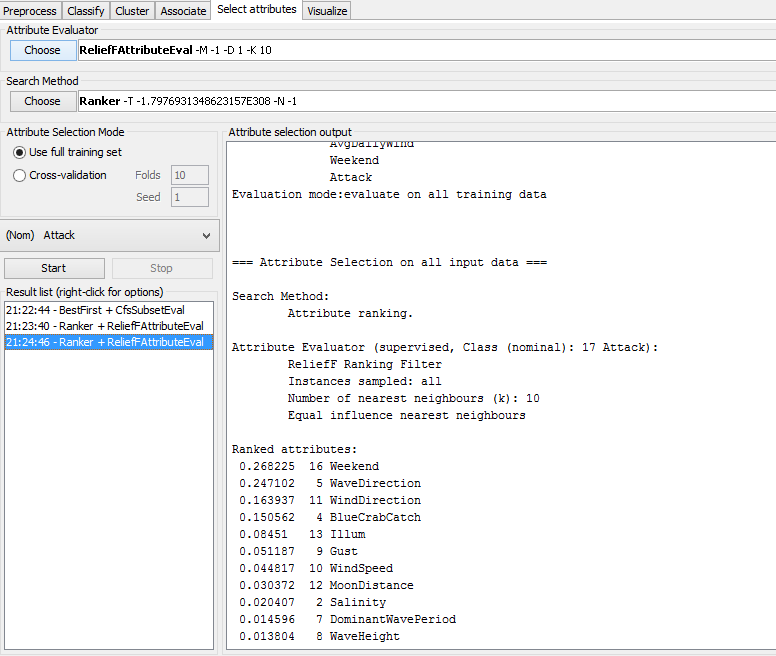


So relief attr selection



This was initial ranking

Removed DATE and ran relief algo again



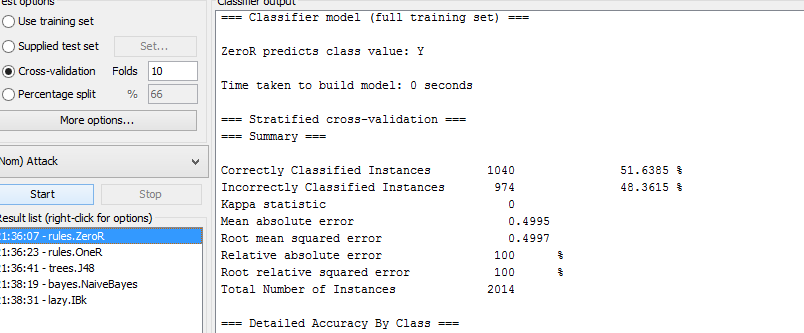
Results are better now

On analyzing data found lots of missing values. Like this we continued and selected features and optimized the values.

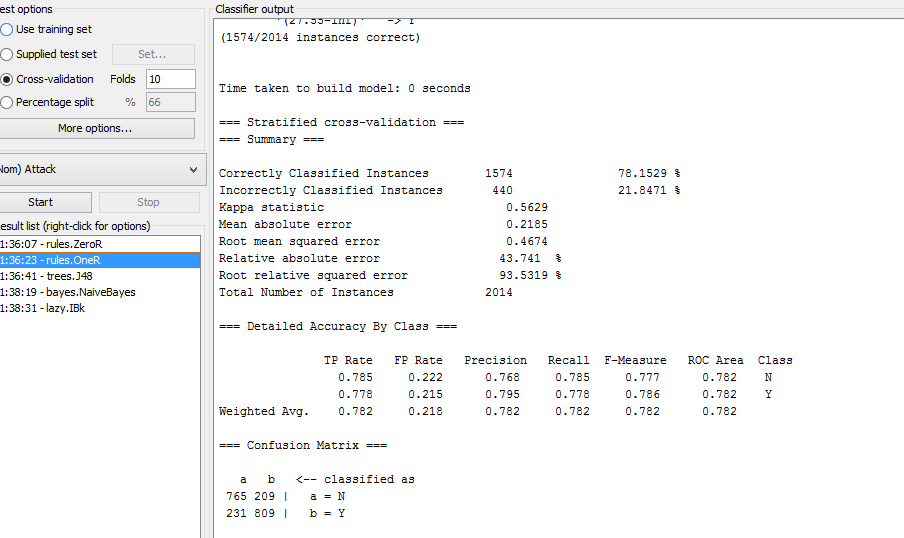
While running algorithms we faced issued as mentioned above like too much bias towards Attack=N.

Hence we did Resample and Discretize with 4 bin and equal frequency

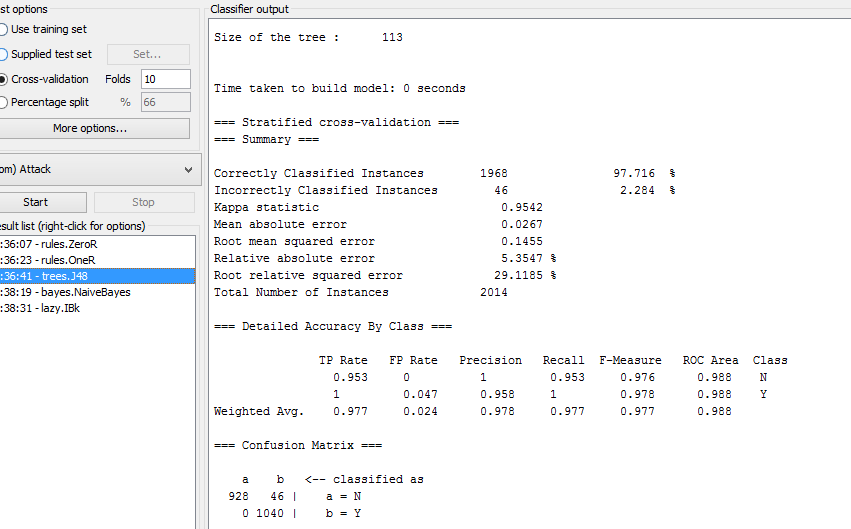
ZeroR:



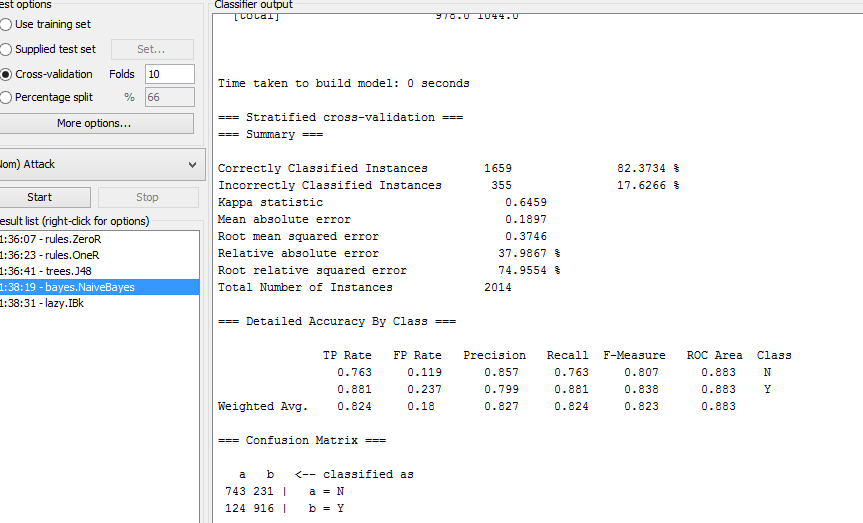
OneR:



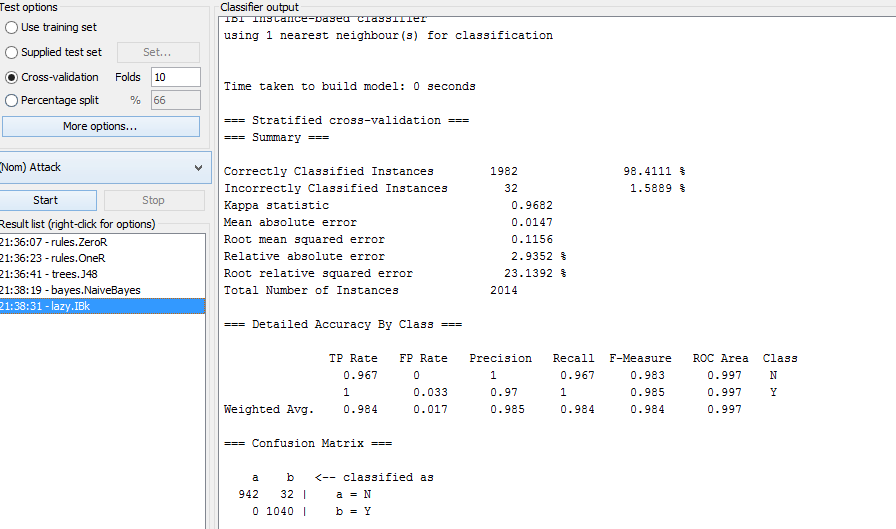
J48



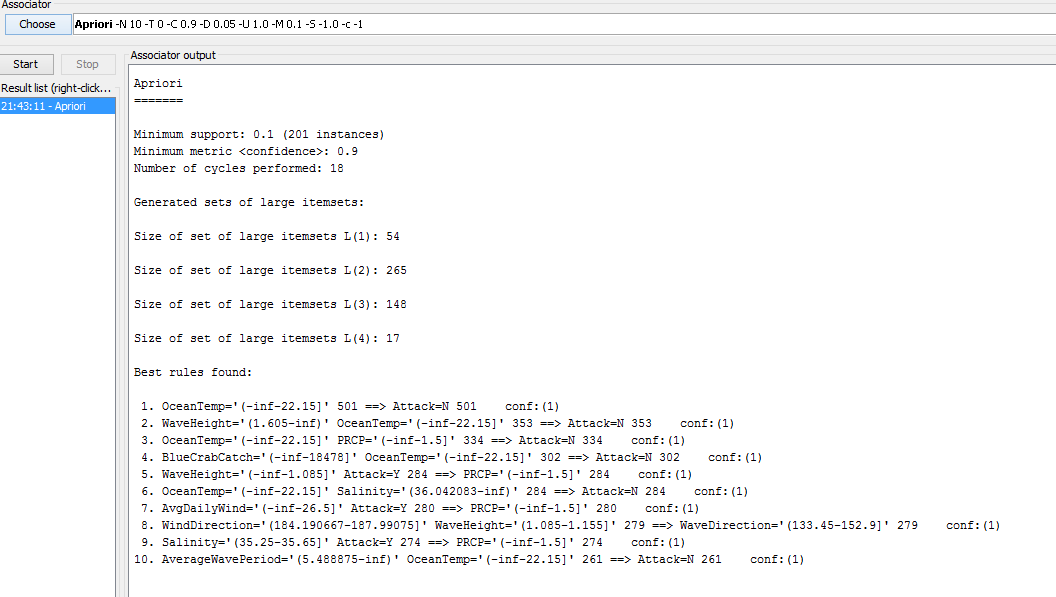
NaiveBayes



IBk



Apriori



EM

Scheme:weka.clusterers.EM -I 100 -N 4 -M 1.0E-6 -S 100

Relation: SharkFullFinalv3-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B4-M-1.0-Rfirst-last

Instances: 2014

Attributes: 14

BlueCrabCatch

Illum

SeaLevel

WindSpeed

WindDirection

AvgDailyWind

WaveDirection

AverageWavePeriod

WaveHeight

OceanTemp

Salinity

PRCP

SeaPressure

Attack

Test mode:evaluate on training data

=== Model and evaluation on training set ===

EM

==

Number of clusters: 4

Cluster

Attribute 0 1 2 3

(0.16) (0.44) (0.15) (0.25)

=================================================================

BlueCrabCatch

'(-inf-18478]' 2.6662 332.2041 85.7083 91.4214

'(18478-107356]' 134.2091 271.3516 2.044 98.3953

'(107356-149870]' 147.2427 85.7998 179.7799 93.1776

'(149870-inf)' 51.1145 197.8785 41.5791 215.4279

[total] 335.2324 887.234 309.1114 498.4222

Illum

FM 156.0079 270.9509 139.4645 163.5767

FQ 6.3715 134.3801 86.6579 13.5905

LQ 47.3903 129.3996 42.8511 11.359

NM 125.4627 352.5034 40.1379 309.8961

[total] 335.2324 887.234 309.1114 498.4222

SeaLevel

'(-inf-0.2725]' 67.5426 176.337 139.7307 124.3897

'(0.2725-0.6085]' 97.4274 224.3149 43.1038 143.1539

'(0.6085-0.8725]' 167.1087 174.7274 40.2577 107.9062

'(0.8725-inf)' 3.1538 311.8547 86.0192 122.9723

[total] 335.2324 887.234 309.1114 498.4222

WindSpeed

'(-inf-5.35175]' 39.7567 136.7534 1.0001 363.4898

'(5.35175-6.77425]' 238.2887 206.1094 1.0057 131.5963

'(6.77425-8.973875]' 56.1677 315.6221 92.1538 1.0564

'(8.973875-inf)' 1.0193 228.7492 214.9518 2.2797

[total] 335.2324 887.234 309.1114 498.4222

WindDirection

'(-inf-184.190667]' 1.2652 304.2402 38.6441 43.8505

'(184.190667-187.99075]' 68.5751 3.7962 136.8197 349.8091

'(187.99075-226.5]' 257.9665 277.4632 1.2085 5.3618

'(226.5-inf)' 7.4257 301.7344 132.4392 99.4008

[total] 335.2324 887.234 309.1114 498.4222

AvgDailyWind

'(-inf-26.5]' 36.93 187.8026 1.0128 279.2547

'(26.5-39.5]' 13.5238 307.3384 1.1506 195.9872

'(39.5-47.5]' 199.5706 185.5316 97.4469 12.4508

'(47.5-inf)' 85.208 206.5614 209.5011 10.7295

[total] 335.2324 887.234 309.1114 498.4222

WaveDirection

'(-inf-104.1]' 38.2121 312.065 39.1543 87.5686

'(104.1-133.45]' 3.9619 390.9069 1.0053 1.1259

'(133.45-152.9]' 67.8568 124.7062 1.2127 405.2243

'(152.9-inf)' 225.2017 59.5558 267.7391 4.5034

[total] 335.2324 887.234 309.1114 498.4222

AverageWavePeriod

'(-inf-4.675]' 197.5385 108.4101 84.8162 143.2352

'(4.675-5.264125]' 132.9615 112.651 184.6397 2.7478

'(5.264125-5.488875]' 1.0362 306.5406 1.0002 230.423

'(5.488875-inf)' 3.6962 359.6323 38.6553 122.0162

[total] 335.2324 887.234 309.1114 498.4222

WaveHeight

'(-inf-1.085]' 155.3701 140.2045 1.7746 174.6509

'(1.085-1.155]' 177.6267 18.7517 1.0078 321.6137

'(1.155-1.605]' 1.2324 299.0156 219.5946 1.1574

'(1.605-inf)' 1.0032 429.2622 86.7344 1.0002

[total] 335.2324 887.234 309.1114 498.4222

OceanTemp

'(-inf-22.15]' 1.0759 499.555 3.3572 1.0119

'(22.15-26.745]' 116.3173 244.9203 1.0328 136.7297

'(26.745-27.55]' 132.8779 101.277 86.9436 295.9015

'(27.55-inf)' 84.9613 41.4817 217.7779 64.7791

[total] 335.2324 887.234 309.1114 498.4222

Salinity

'(-inf-35.25]' 55.729 154.241 260.8175 81.2125

'(35.25-35.65]' 62.4709 117.9491 2.6619 293.9182

'(35.65-36.042083]' 216.031 260.9164 3.9927 9.0598

'(36.042083-inf)' 1.0014 354.1275 41.6394 114.2317

[total] 335.2324 887.234 309.1114 498.4222

PRCP

'(-inf-1.5]' 253.5387 619.1741 93.8285 392.4586

'(1.5-11.5]' 1.0676 67.2152 81.7848 75.9324

'(11.5-72.5]' 72.1668 92.4621 43.1507 17.2204

'(72.5-inf)' 8.4593 108.3826 90.3473 12.8108

[total] 335.2324 887.234 309.1114 498.4222

SeaPressure

'(-inf-1014.85]' 66.3885 145.7832 126.7673 158.061

'(1014.85-1016.805]' 3.6237 183.6923 43.2348 252.4492

'(1016.805-1018.35]' 225.7938 134.5367 89.1781 86.4915

'(1018.35-inf)' 39.4264 423.2219 49.9312 1.4205

[total] 335.2324 887.234 309.1114 498.4222

Attack

N 50.4077 813.6277 11.5186 102.446

Y 282.8247 71.6063 295.5928 393.9762

[total] 333.2324 885.234 307.1114 496.4222

Time taken to build model (full training data) : 0.61 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 331 ( 16%)

1 882 ( 44%)

2 305 ( 15%)

3 496 ( 25%)

Log likelihood: -15.00635

## **RESULTS:**

### Blue crab catch was more and is related to warmer ocean temperature, high illumination and less precipitation.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: SharkFullFinalv3-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B4-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R3-9,13

Instances: 2014

Attributes: 6

BlueCrabCatch

Illum

OceanTemp

Salinity

PRCP

Attack

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

OceanTemp = '(-inf-22.15]': N (501.0)

OceanTemp = '(22.15-26.745]'

| BlueCrabCatch = '(-inf-18478]': N (20.0)

| BlueCrabCatch = '(18478-107356]': N (99.0)

| BlueCrabCatch = '(107356-149870]'

| | Salinity = '(-inf-35.25]': N (8.0)

| | Salinity = '(35.25-35.65]': N (31.0)

| | Salinity = '(35.65-36.042083]'

| | | Illum = FM: N (6.0)

| | | Illum = FQ: N (1.0)

| | | Illum = LQ: N (3.0)

| | | Illum = NM

| | | | PRCP = '(-inf-1.5]': Y (60.0/6.0)

| | | | PRCP = '(1.5-11.5]': N (1.0)

| | | | PRCP = '(11.5-72.5]': N (1.0)

| | | | PRCP = '(72.5-inf)': N (2.0)

| | Salinity = '(36.042083-inf)': N (5.0)

**| BlueCrabCatch = '(149870-inf)'**

**| | Illum = FM**

| | | Salinity = '(-inf-35.25]': N (5.0)

**| | | Salinity = '(35.25-35.65]': Y (53.0/7.0)**

| | | Salinity = '(35.65-36.042083]': Y (47.0/6.0)

| | | Salinity = '(36.042083-inf)': N (3.0)

| | Illum = FQ: N (14.0)

| | Illum = LQ: N (14.0)

| | Illum = NM

| | | Salinity = '(-inf-35.25]': Y (91.0/1.0)

| | | Salinity = '(35.25-35.65]': N (21.0)

| | | Salinity = '(35.65-36.042083]': N (6.0)

| | | Salinity = '(36.042083-inf)': N (4.0)

OceanTemp = '(26.745-27.55]'

| PRCP = '(-inf-1.5]'

| | Salinity = '(-inf-35.25]'

| | | Illum = FM: N (2.0)

| | | Illum = FQ: Y (40.0/1.0)

| | | Illum = LQ: N (2.0)

| | | Illum = NM: N (4.0)

| | Salinity = '(35.25-35.65]'

| | | Illum = FM

| | | | BlueCrabCatch = '(-inf-18478]': Y (84.0/2.0)

| | | | BlueCrabCatch = '(18478-107356]': N (6.0)

| | | | BlueCrabCatch = '(107356-149870]': Y (38.0/1.0)

| | | | BlueCrabCatch = '(149870-inf)': N (3.0)

| | | Illum = FQ: N (8.0)

| | | Illum = LQ: N (2.0)

| | | Illum = NM

| | | | BlueCrabCatch = '(-inf-18478]': N (4.0)

| | | | BlueCrabCatch = '(18478-107356]': Y (76.0/1.0)

| | | | BlueCrabCatch = '(107356-149870]': N (3.0)

| | | | BlueCrabCatch = '(149870-inf)': Y (39.0/5.0)

| | Salinity = '(35.65-36.042083]'

| | | Illum = FM: N (12.0)

| | | Illum = FQ: N (4.0)

| | | Illum = LQ: N (6.0)

| | | Illum = NM

| | | | BlueCrabCatch = '(-inf-18478]': Y (0.0)

| | | | BlueCrabCatch = '(18478-107356]': Y (53.0/3.0)

| | | | BlueCrabCatch = '(107356-149870]': N (4.0)

**| | | | BlueCrabCatch = '(149870-inf)': Y (39.0/7.0)**

| | Salinity = '(36.042083-inf)': N (8.0)

| PRCP = '(1.5-11.5]'

| | Salinity = '(-inf-35.25]': Y (43.0)

| | Salinity = '(35.25-35.65]': N (5.0)

| | Salinity = '(35.65-36.042083]': N (2.0)

| | Salinity = '(36.042083-inf)': Y (66.0)

| PRCP = '(11.5-72.5]'

| | Illum = FM: Y (29.0/2.0)

| | Illum = FQ: N (3.0)

| | Illum = LQ: N (1.0)

| | Illum = NM: N (7.0)

| PRCP = '(72.5-inf)': N (20.0)

OceanTemp = '(27.55-inf)'

| Salinity = '(-inf-35.25]'

| | BlueCrabCatch = '(-inf-18478]': Y (39.0/1.0)

| | BlueCrabCatch = '(18478-107356]'

| | | PRCP = '(-inf-1.5]': N (4.0)

| | | PRCP = '(1.5-11.5]': Y (0.0)

| | | PRCP = '(11.5-72.5]': Y (43.0/3.0)

| | | PRCP = '(72.5-inf)': Y (0.0)

| | BlueCrabCatch = '(107356-149870]'

| | | Illum = FM: Y (139.0/4.0)

| | | Illum = FQ: N (1.0)

| | | Illum = LQ: N (1.0)

| | | Illum = NM: N (5.0)

| | BlueCrabCatch = '(149870-inf)': N (11.0)

| Salinity = '(35.25-35.65]': N (16.0)

| Salinity = '(35.65-36.042083]'

| | Illum = FM: Y (33.0)

| | Illum = FQ: N (5.0)

| | Illum = LQ: N (6.0)

| | Illum = NM: N (1.0)

| Salinity = '(36.042083-inf)'

| | Illum = FM: N (11.0)

| | Illum = FQ: N (4.0)

| | Illum = LQ: Y (41.0/1.0)

**| | Illum = NM**

**| | | PRCP = '(-inf-1.5]': Y (40.0/2.0)**

| | | PRCP = '(1.5-11.5]': N (3.0)

| | | PRCP = '(11.5-72.5]': N (2.0)

| | | PRCP = '(72.5-inf)': Y (0.0)

Number of Leaves : 76

Size of the tree : 101

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 1944 96.5243 %

Incorrectly Classified Instances 70 3.4757 %

Kappa statistic 0.9303

Mean absolute error 0.0562

Root mean squared error 0.1761

Relative absolute error 11.2482 %

Root relative squared error 35.2297 %

Total Number of Instances 2014

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.928 0 1 0.928 0.963 0.979 N

1 0.072 0.937 1 0.967 0.979 Y

Weighted Avg. 0.965 0.037 0.967 0.965 0.965 0.979

=== Confusion Matrix ===

a b <-- classified as

904 70 | a = N

0 1040 | b = Y

### With new moon and full moon attacks are more.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: SharkFullFinalv3-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B4-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1,3-13

Instances: 2014

Attributes: 2

Illum

Attack

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

**Illum = FM: Y (726.0/325.0)**

Illum = FQ: N (237.0/82.0)

Illum = LQ: N (227.0/80.0)

**Illum = NM: Y (824.0/347.0)**

Number of Leaves : 4

Size of the tree : 5

EM

=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100

Relation: SharkFullFinalv3-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B4-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.Remove-R1,3-13

Instances: 2014

Attributes: 2

Illum

Attack

Test mode:evaluate on training data

=== Model and evaluation on training set ===

EM

==

Number of clusters selected by cross validation: 3

Cluster

Attribute 0 1 2

(0.48) (0.25) (0.27)

=======================================

Illum

**FM 712.9185 7.1091 8.9725**

FQ 128.5622 10.0543 101.3835

LQ 125.0541 10.0712 94.8747

**NM 7.9965 475.7572 343.2463**

[total] 974.5313 502.9917 548.477

Attack

N 431.4569 7.9326 537.6105

Y 541.0744 493.0591 8.8664

[total] 972.5313 500.9917 546.477

Time taken to build model (full training data) : 0.83 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 888 ( 44%)

1 477 ( 24%)

2 649 ( 32%)

### Lower Low water in combination in sudden rise of temperature marks inflow of Gulf Stream and this increases chance of shark attack.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: SharkFullFinalv3-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-2,4-9,11-13-weka.filters.supervised.instance.Resample-B1.0-S13-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B8-M-1.0-Rfirst-last

Instances: 2014

Attributes: 3

SeaLevel

OceanTemp

Attack

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

OceanTemp = '(-inf-17.75]': N (249.0)

OceanTemp = '(17.75-21.75]': N (253.0)

OceanTemp = '(21.75-24.85]'

| SeaLevel = '(-inf--0.1775]': N (5.0)

| SeaLevel = '(-0.1775-0.2905]': N (24.0)

| SeaLevel = '(0.2905-0.489]': N (27.0)

| SeaLevel = '(0.489-0.6415]': N (18.0)

| SeaLevel = '(0.6415-0.7315]': N (17.0)

| SeaLevel = '(0.7315-0.8875]': N (13.0)

| SeaLevel = '(0.8875-1.1825]': N (69.0/27.0)

| SeaLevel = '(1.1825-inf)': Y (78.0/33.0)

OceanTemp = '(24.85-26.745]'

| SeaLevel = '(-inf--0.1775]': Y (45.0/1.0)

| SeaLevel = '(-0.1775-0.2905]': Y (56.0/15.0)

| SeaLevel = '(0.2905-0.489]': N (29.0)

| SeaLevel = '(0.489-0.6415]': N (12.0)

| SeaLevel = '(0.6415-0.7315]': Y (49.0/4.0)

| SeaLevel = '(0.7315-0.8875]': N (18.0)

| SeaLevel = '(0.8875-1.1825]': N (23.0)

| SeaLevel = '(1.1825-inf)': N (14.0)

OceanTemp = '(26.745-27.25]'

| SeaLevel = '(-inf--0.1775]': Y (81.0)

| SeaLevel = '(-0.1775-0.2905]': N (9.0)

| SeaLevel = '(0.2905-0.489]': N (7.0)

| SeaLevel = '(0.489-0.6415]': Y (37.0/5.0)

| SeaLevel = '(0.6415-0.7315]': Y (47.0/11.0)

| SeaLevel = '(0.7315-0.8875]': N (16.0)

| SeaLevel = '(0.8875-1.1825]': N (16.0)

| SeaLevel = '(1.1825-inf)': Y (59.0/10.0)

OceanTemp = '(27.25-27.55]'

| SeaLevel = '(-inf--0.1775]': Y (0.0)

| SeaLevel = '(-0.1775-0.2905]': N (4.0)

| SeaLevel = '(0.2905-0.489]': N (3.0)

| SeaLevel = '(0.489-0.6415]': Y (84.0/5.0)

| SeaLevel = '(0.6415-0.7315]': Y (78.0/9.0)

| SeaLevel = '(0.7315-0.8875]': Y (120.0/4.0)

| SeaLevel = '(0.8875-1.1825]': N (14.0)

| SeaLevel = '(1.1825-inf)': Y (53.0/6.0)

OceanTemp = '(27.55-27.95]'

| SeaLevel = '(-inf--0.1775]': Y (46.0/1.0)

| SeaLevel = '(-0.1775-0.2905]': Y (40.0/8.0)

| SeaLevel = '(0.2905-0.489]': Y (39.0/1.0)

| SeaLevel = '(0.489-0.6415]': N (6.0)

| SeaLevel = '(0.6415-0.7315]': N (6.0)

| SeaLevel = '(0.7315-0.8875]': Y (41.0/7.0)

| SeaLevel = '(0.8875-1.1825]': N (14.0)

| SeaLevel = '(1.1825-inf)': N (1.0)

OceanTemp = '(27.95-inf)'

| SeaLevel = '(-inf--0.1775]': Y (80.0)

| SeaLevel = '(-0.1775-0.2905]': N (4.0)

| SeaLevel = '(0.2905-0.489]': Y (45.0/9.0)

| SeaLevel = '(0.489-0.6415]': N (6.0)

| SeaLevel = '(0.6415-0.7315]': N (6.0)

| SeaLevel = '(0.7315-0.8875]': N (5.0)

**| SeaLevel = '(0.8875-1.1825]': Y (48.0/9.0)**

| SeaLevel = '(1.1825-inf)': Y (0.0)

Number of Leaves : 50

Size of the tree : 57

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 1838 91.2612 %

Incorrectly Classified Instances 176 8.7388 %

### When wave direction is from South East and wind direction is South and wind speed is maximum and ocean temperature is high we get maximum attacks.

**Wind Direction and Degrees**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Cardinal Direction** | **Degree Direction** | | http://climate.umn.edu/snow_fence/Images/Graphics/pix.gif | http://climate.umn.edu/snow_fence/Images/Graphics/pix.gif | | N | 348.75 - 11.25 | | NNE | 11.25 - 33.75 | | NE | 33.75 - 56.25 | | ENE | 56.25 - 78.75 | | E | 78.75 - 101.25 | | ESE | 101.25 - 123.75 | | SE | 123.75 - 146.25 | | SSE | 146.25 - 168.75 | | S | 168.75 - 191.25 | | SSW | 191.25 - 213.75 | | SW | 213.75 - 236.25 | | WSW | 236.25 - 258.75 | | W | 258.75 - 281.25 | | WNW | 281.25 - 303.75 | | NW | 303.75 - 326.25 | | NNW | 326.25 - 348.75 | | http://climate.umn.edu/snow_fence/Images/Wind/wind_bln1.gif |

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes

Relation: SharkFullFinalv3-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-3,6,8,11-13-weka.filters.supervised.instance.Resample-B1.0-S13-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B6-M-1.0-Rfirst-last

Instances: 2014

Attributes: 6

WindSpeed

WindDirection

WaveDirection

WaveHeight

OceanTemp

Attack

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute N Y

(0.5) (0.5)

============================================

WindSpeed

'(-inf-4.65]' 151.0 189.0

'(4.65-5.64875]' 129.0 194.0

'(5.64875-6.72425]' 163.0 113.0

'(6.72425-7.55]' 172.0 195.0

'(7.55-9.204167]' 216.0 121.0

'(9.204167-inf)' 174.0 209.0

[total] 1005.0 1021.0

WindDirection

'(-inf-174.10425]' 195.0 121.0

**'(174.10425-185.690667]' 92.0 323.0**

'(185.690667-200.497333]' 171.0 149.0

'(200.497333-213.895125]' 103.0 223.0

'(213.895125-248.61675]' 256.0 75.0

'(248.61675-inf)' 188.0 130.0

[total] 1005.0 1021.0

WaveDirection

'(-inf-98.5]' 252.0 121.0

'(98.5-114]' 272.0 62.0

'(114-133.45]' 142.0 50.0

**'(133.45-138.65]' 129.0 308.0**

'(138.65-169.55]' 143.0 232.0

'(169.55-inf)' 67.0 248.0

[total] 1005.0 1021.0

WaveHeight

'(-inf-0.845]' 98.0 238.0

'(0.845-1.135]' 163.0 119.0

'(1.135-1.155]' 58.0 300.0

'(1.155-1.535]' 219.0 116.0

'(1.535-1.885]' 233.0 112.0

'(1.885-inf)' 234.0 136.0

[total] 1005.0 1021.0

OceanTemp

'(-inf-18.75]' 329.0 1.0

'(18.75-24.45]' 303.0 28.0

'(24.45-26.65]' 165.0 176.0

'(26.65-27.35]' 88.0 268.0

'(27.35-27.75]' 63.0 277.0

'(27.75-inf)' 57.0 271.0

[total] 1005.0 1021.0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 1657 82.2741 %

Incorrectly Classified Instances 357 17.7259 %

Kappa statistic 0.6453

Mean absolute error 0.2005

Root mean squared error 0.3591

Relative absolute error 40.1003 %

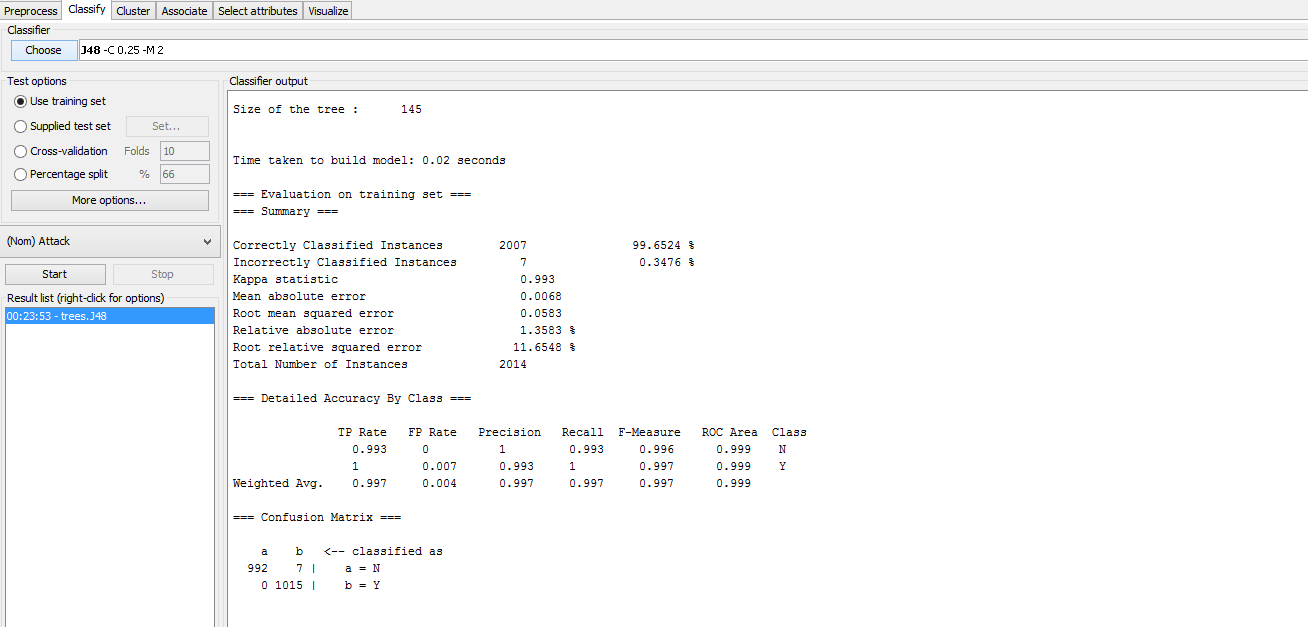
Root relative squared error 71.8249 %

Total Number of Instances 2014

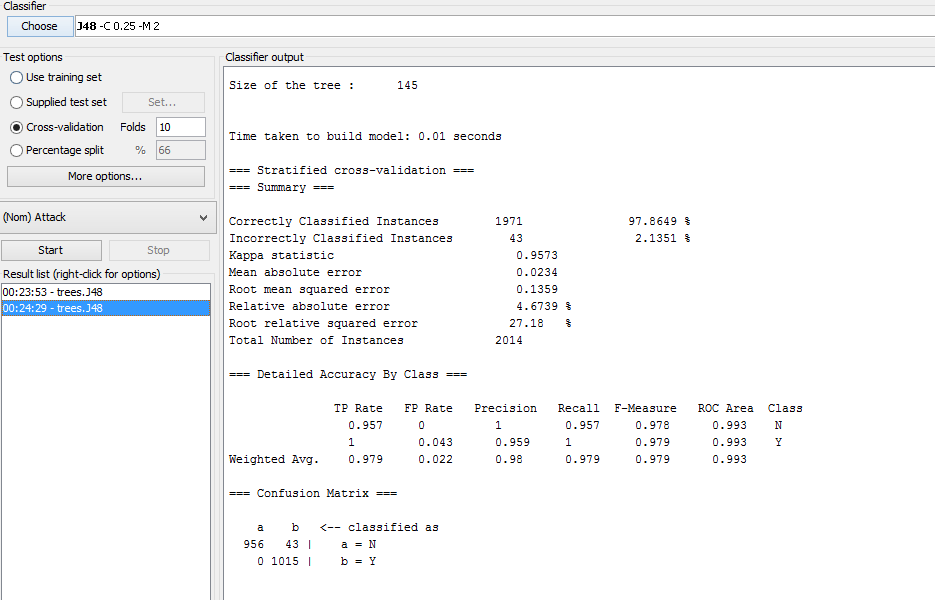
### Description of Training Methodology

For having a good measure on classifier’s performance we used different training methodology. To start with we used the whole dataset as training data.

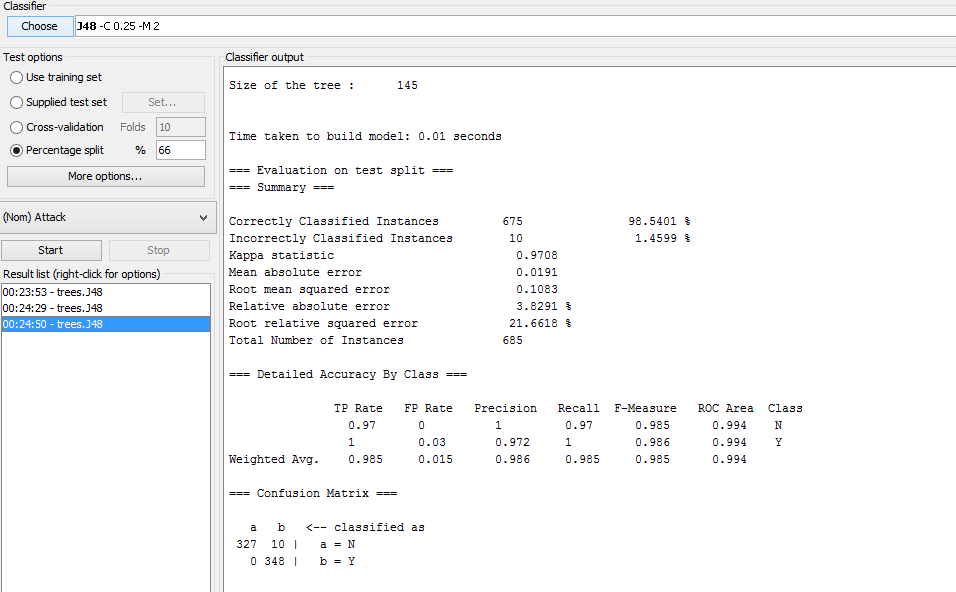
**Whole dataset as Training dataset:** In Weka from Classify tab, we selected test option ‘Use training set’. Ran the classifier on dataset without resampling, the classifier predicted an overly optimistic performance and resulted in overfitting. Another anomaly in data was the high proportion of ‘No’ instances in the training dataset; Over 98% of the samples in the training dataset were No attacks. To resolve this anomaly we resampled the dataset and used it as training data. Following screenshot shows the performance of classifier when whole dataset used as training data:



**Cross Validation:** For better results we used 10 fold cross validation. This methodology gave a better estimate of error and performance of classifier. Following is the screenshot displaying the performance of classifier on 10 fold cross validation.



**Percentage Split:** To have a good measure of future performance we split the dataset into training data and testing data (With default of 66% training and 34% test data). Data different from training dataset was used to test the model. In Weka we selected classify tab, from Test options selected ‘Percentage Split’ with split of 66% for training data. Following is the screenshot displaying the performance of classifier on percentage split.



**Supplied test set:** After training the data on available dataset to predict the class we used the test option supplied test set. A separate test file for instances whose class needs to be predicted is created and supplied to the trained model. For doing this in Weka from Classify tab we saved the trained model and then selected the test option- Supplied test set. Set the test file for prediction, now click on More Options button and select options ‘Output Model’, ‘Output per-class stats’, ‘Output Confusion Matrix’, ‘Store Predictions for visualization’. Start the classifier.

We have kept test data as below

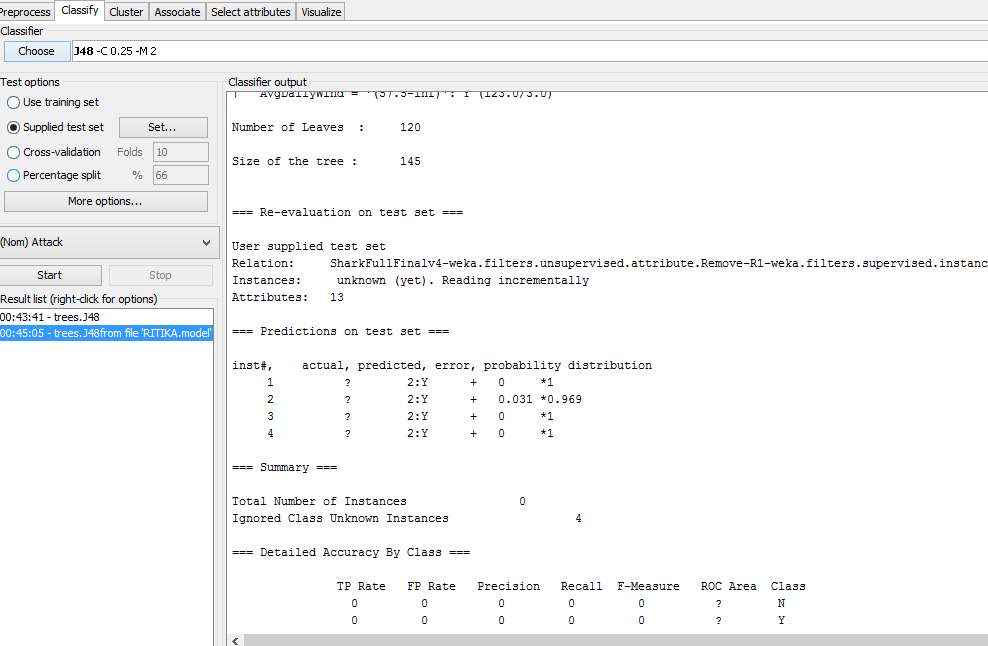
@data

'\'(177629.5-inf)\'',NM,'\'(0.7695-1.1005]\'','\'(5.64875-6.72425]\'','\'(248.61675-inf)\'','\'(45.5-57.5]\'','\'(-inf-98.5]\'','\'(-inf-4.19]\'','\'(-inf-0.845]\'','\'(26.65-27.35]\'','\'(35.91-36.103375]\'','\'(-inf-1.5]\'',**?**

'\'(177629.5-inf)\'',NM,'\'(0.0625-0.4185]\'','\'(-inf-4.65]\'','\'(-inf-174.10425]\'','\'(25.5-30.5]\'','\'(133.45-138.65]\'','\'(5.43-5.668]\'','\'(0.845-1.135]\'','\'(27.75-inf)\'','\'(35.91-36.103375]\'','\'(-inf-1.5]\'',**?**

'\'(103559-128576]\'',FM,'\'(0.6345-0.7695]\'','\'(6.72425-7.55]\'','\'(200.497333-213.895125]\'','\'(57.5-inf)\'','\'(138.65-169.55]\'','\'(-inf-4.19]\'','\'(1.135-1.155]\'','\'(26.65-27.35]\'','\'(35.593333-35.91]\'','\'(26.5-54.5]\'',**?**

'\'(177629.5-inf)\'',NM,'\'(1.1005-inf)\'','\'(9.204167-inf)\'','\'(-inf-174.10425]\'','\'(45.5-57.5]\'','\'(114-133.45]\'','\'(5.668-inf)\'','\'(1.885-inf)\'','\'(26.65-27.35]\'','\'(35.593333-35.91]\'','\'(-inf-1.5]\'',**?**



# Conclusions

Our goal was to understand why in the year of 2015 North Carolina coast have more attacks. We have found of good amount of reasoning for this with domain knowledge and data mining. We understood sharks migration behavior and magnetoception which are impacted by ocean current variation. This years ocean current change with warmer Gulf Stream coming closer to the stream and inflow of bait fish closer to shore.

From data mining we have got below results:-

1. Blue crab catch was more and is related to warmer ocean temperature, nearer moon and less precipitation.
2. With new moon and full moon attacks are more.
3. Lower Low water in combination in sudden rise of temperature marks inflow of Gulf Stream and this increases chance of shark attack.
4. When wave direction is from South East and wind direction is south and wind speed is maximum and ocean temperature is high we get maximum attacks.

From April 17 2015 the Gulf Stream which was further weakened due to global warming and variations start arriving to north Carolina coast. Along with this bait fish groups like Herring also starts arriving in the front of the Gulf Stream. This is evident from the rise in temperature and increase in catches of Blue Crab catch and wave and wind direction. This has hiked Shark attack in those condition.

## Future Work

More ocean current information from observatory near to Cape Hatteras which is the merging point of Gulf Stream and Labrador Current will give deeper picture and predict incoming Gulf Stream nearer to shore in coming summer. This will help us to predict future conditions of shark attack.

## Work Citation

<http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/>

(For Data Preprocessing)

[http://www.slideshare.net/sabarigr/data-preprocessing?next\_slideshow=1](http://www.google.com/url?q=http%3A%2F%2Fwww.slideshare.net%2Fsabarigr%2Fdata-preprocessing%3Fnext_slideshow%3D1&sa=D&sntz=1&usg=AFQjCNGLSU6WehN-HNXPq-HBZfjgldVQqA)

[http://www.slideshare.net/jasonrodrigues/data-preprocessing-5609305](http://www.google.com/url?q=http%3A%2F%2Fwww.slideshare.net%2Fjasonrodrigues%2Fdata-preprocessing-5609305&sa=D&sntz=1&usg=AFQjCNHrB1I_PZNmkpjeV0UgjxB_bGy2ag)

(For evaluation measures)

[http://iasri.res.in/ebook/win\_school\_aa/notes/Evaluation\_Measures.pdf](http://www.google.com/url?q=http%3A%2F%2Fiasri.res.in%2Febook%2Fwin_school_aa%2Fnotes%2FEvaluation_Measures.pdf&sa=D&sntz=1&usg=AFQjCNFgu46S-5yBypEHU4-79L3Ps0OPyg)

Shark

<http://oceanofk.org/tag/Tagmigrate/bhowmigrate.html>