Mining Relationships between US Petroleum Production and Oil/ Gas Price

Sanchit Singhal
Milind Siddhanti

Problem Definition / Motivation

- Volatile industry
- Growing US dominance in global oil markets
- Draw inferences between drilling performance per region and global oil and natural gas prices
- Help to :
    - Regulate petroleum firms by government
    - Direct investment by corporations
    - Develop economic relationships between states

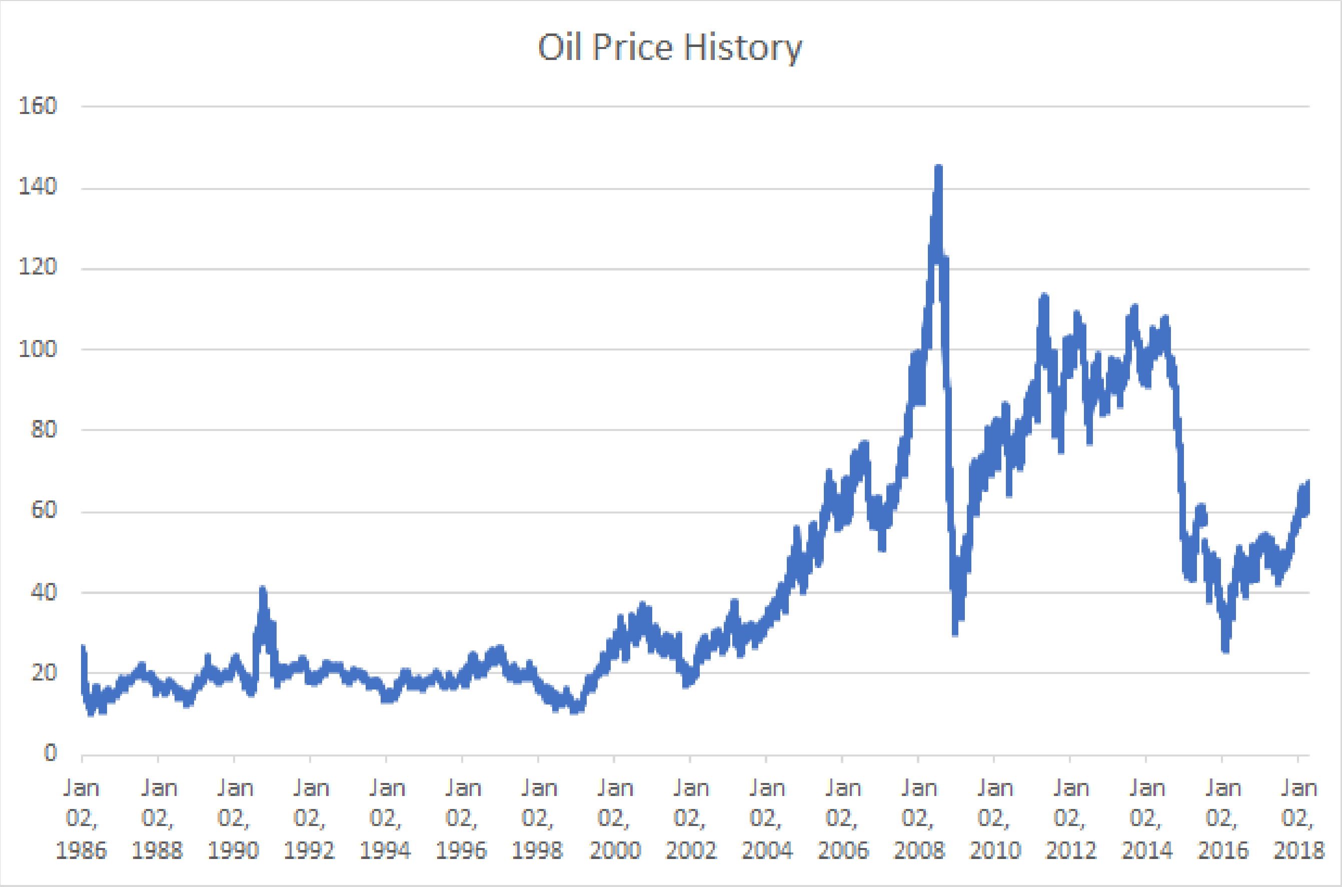
Datasets

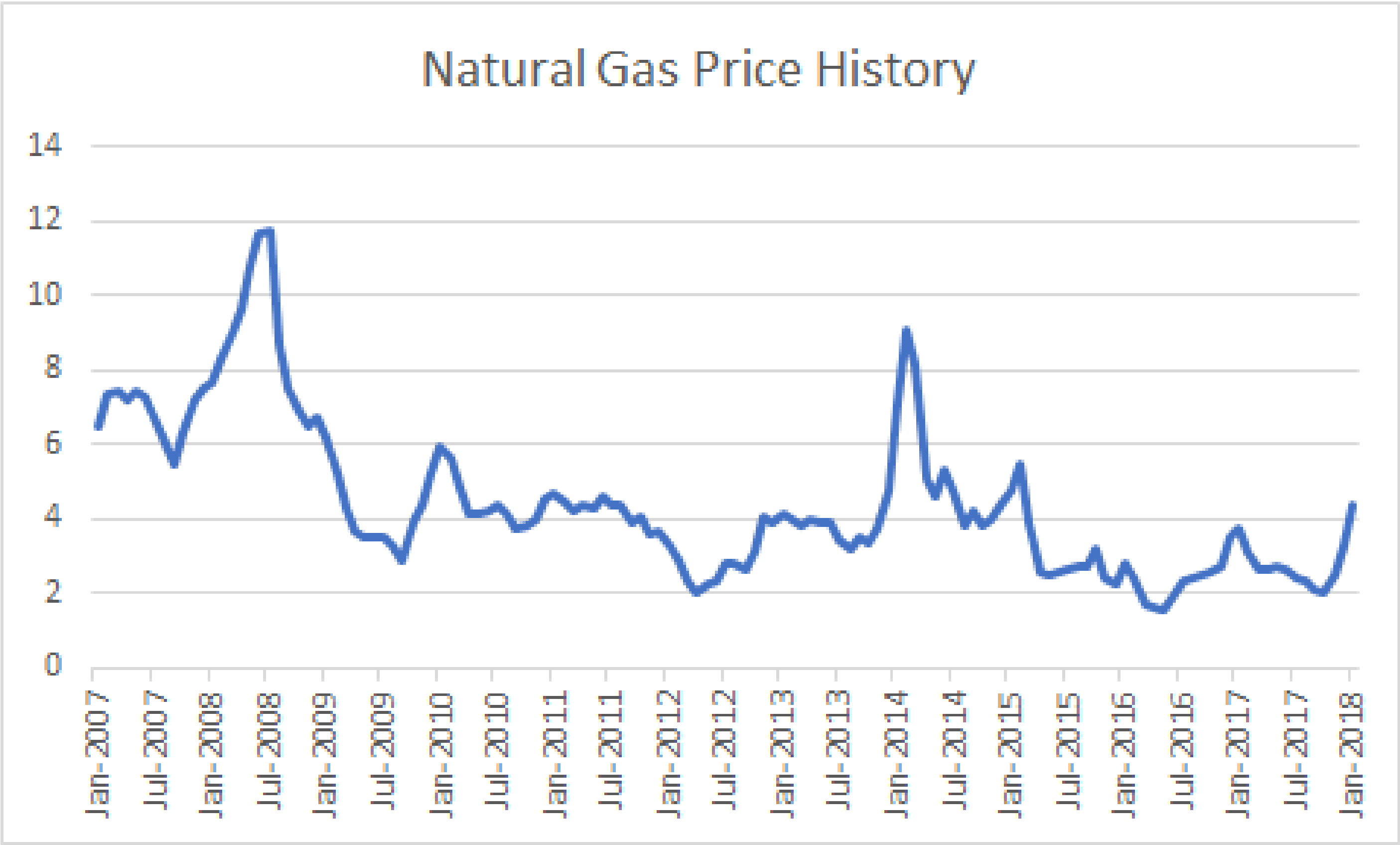
- Source: US Energy Information Administration (EIA)
- Files = 4 csvs (oil prices per day, gas prices per month, DPR, DUC)
- Available Features
    - Date/Month (Jan 07 - Jan 18)
    - Rig Count
    - Production per rig (oil/natural gas)
    - Total Production (oil/natural gas)
    - Prices (oil/natural gas)
    - Region
    - Completed and Uncompleted rig rate
    - Petroleum Imports
    - Petroleum Exports

Data Preprocessing

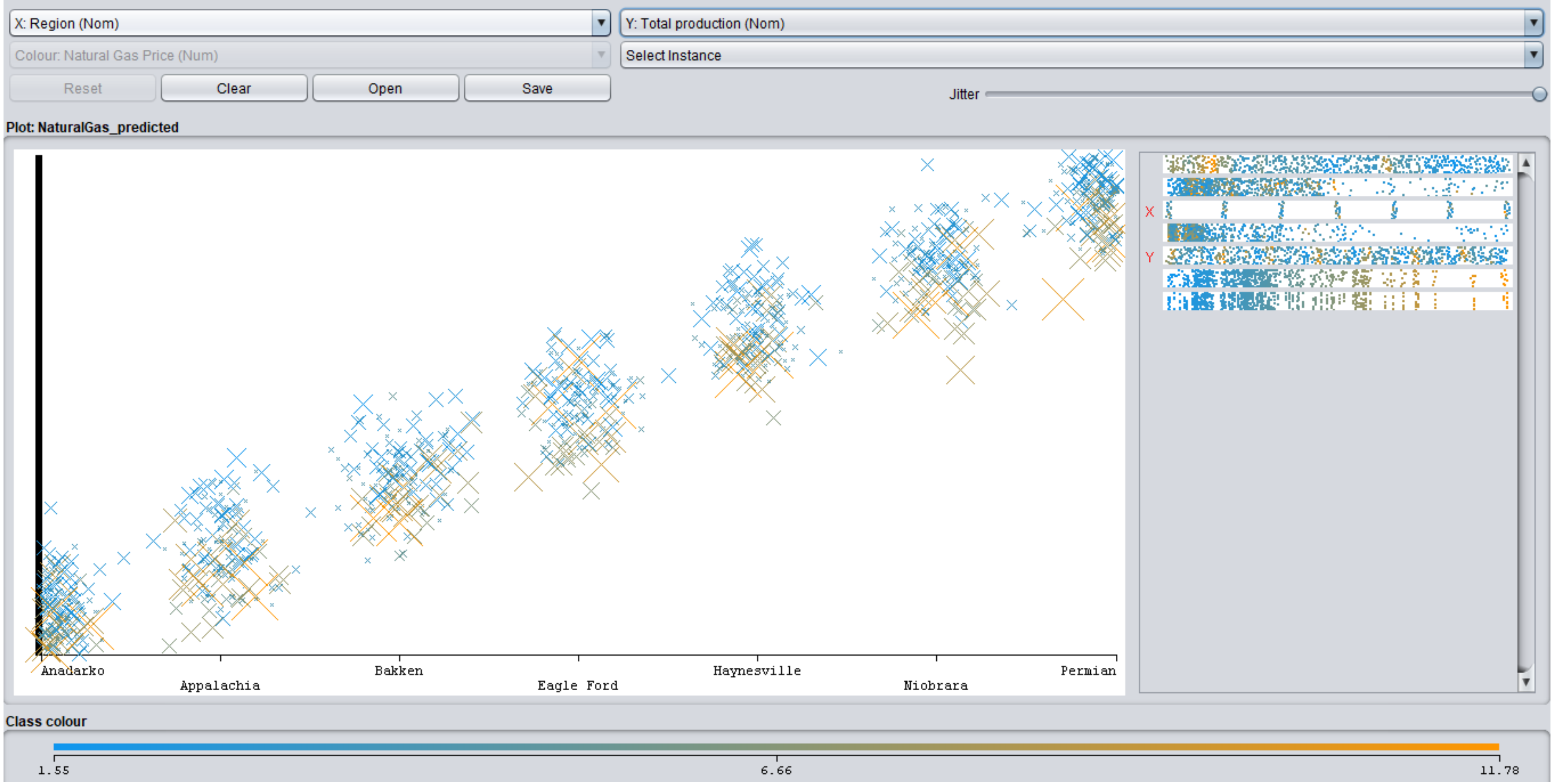
- Split Datasets into Oil and Natural gas sets
- Removed missing values
- Merged Region information as a feature
- Added Price information
- Selected last day price of month as bin value
- Discrete numerical values
  - Oil Price into 5 bins
  - Natural Gas into 4 bins

Data Exploration of Price Fluctuations


Oil Price History
160
140
120
100
80
60
40
20
0
Jan 02, 1986
Jan 02, 1988
Jan 02, 1990
Jan 02, 1992
Jan 02, 1994
Jan 02, 1996
Jan 02, 1998
Jan 02, 2000
Jan 02, 2002
Jan 02, 2004
Jan 02, 2006
Jan 02, 2008
Jan 02, 2010
Jan 02, 2012
Jan 02, 2014
Jan 02, 2016
Jan 02, 2018



Natural Gas Price History



X: Region (Nom)

Y: Total production (Nom)

Colour: Natural Gas Price (Num)

Plot: NaturalGas_predicted

Class colour

Data Modeling to Understand Natural Gas Production in the US

Simple K Means on Natural Gas

Features: Month, Rig count, Production per rig, Total production, Region

Model: Simple K Means(Clustering)

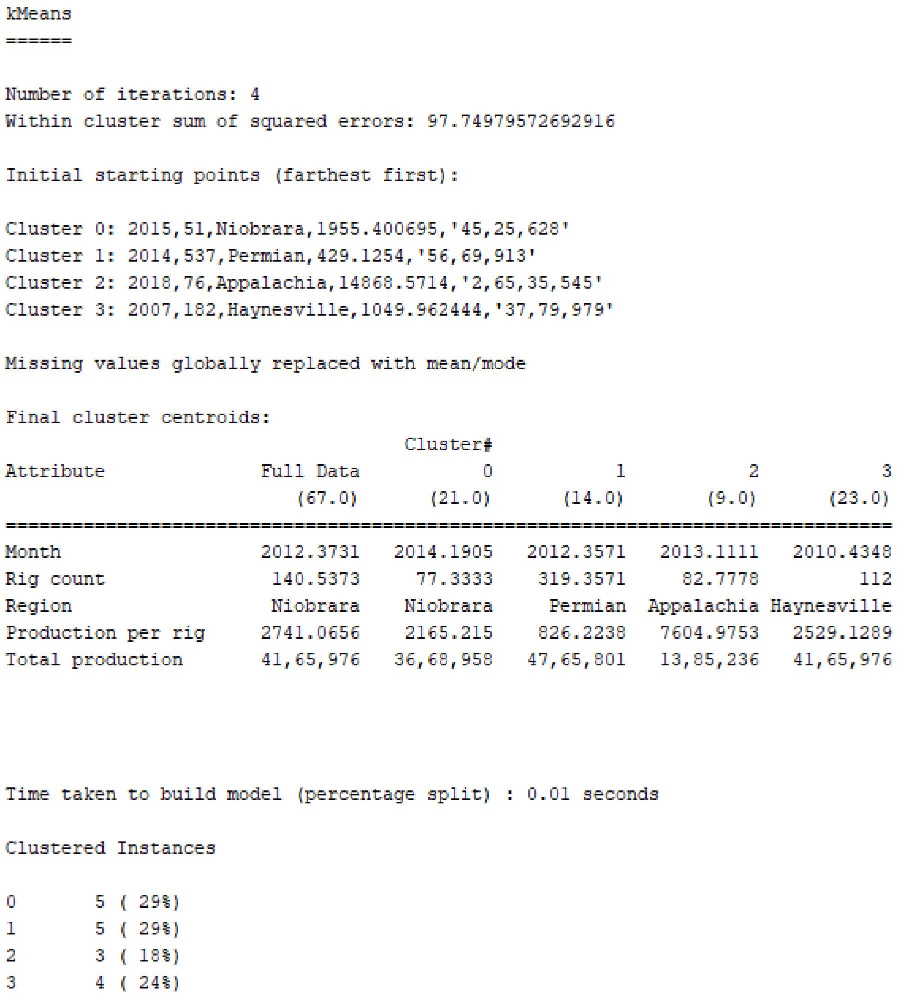
Properties:
Euclidean distance

Max iterations = 200

No of clusters = 4

Initialization method = Farthest first

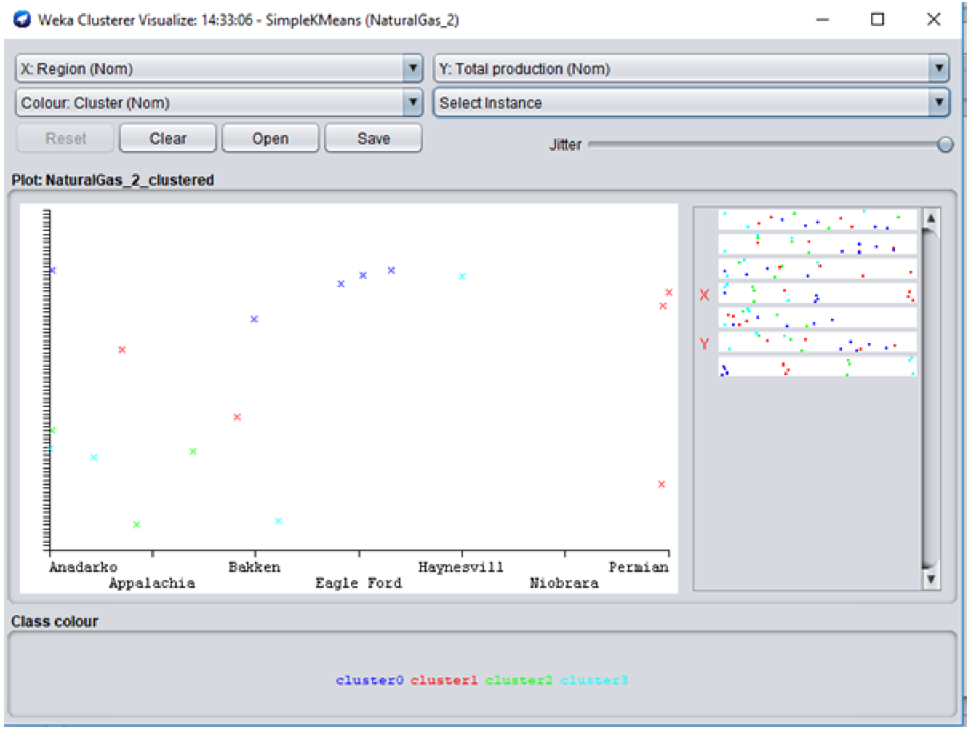
Percentage split = 80



```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 97.74979572692916

Initial starting points (farthest first):

Cluster 0: 2015,51,Niobrara,1955.400695,'45,25,628'
Cluster 1: 2014,537,Permian,429.1254,'56,69,913'
Cluster 2: 2018,76,Appalachia,14868.5714,'2,65,35,545'
Cluster 3: 2007,182,Haynesville,1049.962444,'37,79,979'

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                Full Data            0            1            2            3
                            (67.0)       (21.0)       (14.0)        (9.0)       (23.0)
==========================================================================================
Month                    2012.3731    2014.1905    2012.3571    2013.1111    2010.4348
Rig count                 140.5373      77.3333     319.3571      82.7778          112
Region                    Niobrara     Niobrara      Permian   Appalachia  Haynesville
Production per rig       2741.0656     2165.215     826.2238    7604.9753    2529.1289
Total production         41,65,976    36,68,958    47,65,801    13,85,236    41,65,976




Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0        5 ( 29%)
1        5 ( 29%)
2        3 ( 18%)
3        4 ( 24%)
```

Evaluation of Model

Classifier Visualization



Weka Clusterer Visualize: 14:33:06 - SimpleKMeans (NaturalGas_2)

X: Region (Nom)

Y: Total production (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: NaturalGas_2_clustered

Class colour

Hierarchical Clustering on Natural Gas

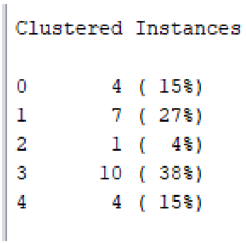
Features: Month, Rig count, Production per rig, Total production, Region

Model: Hierarchical (Clustering)

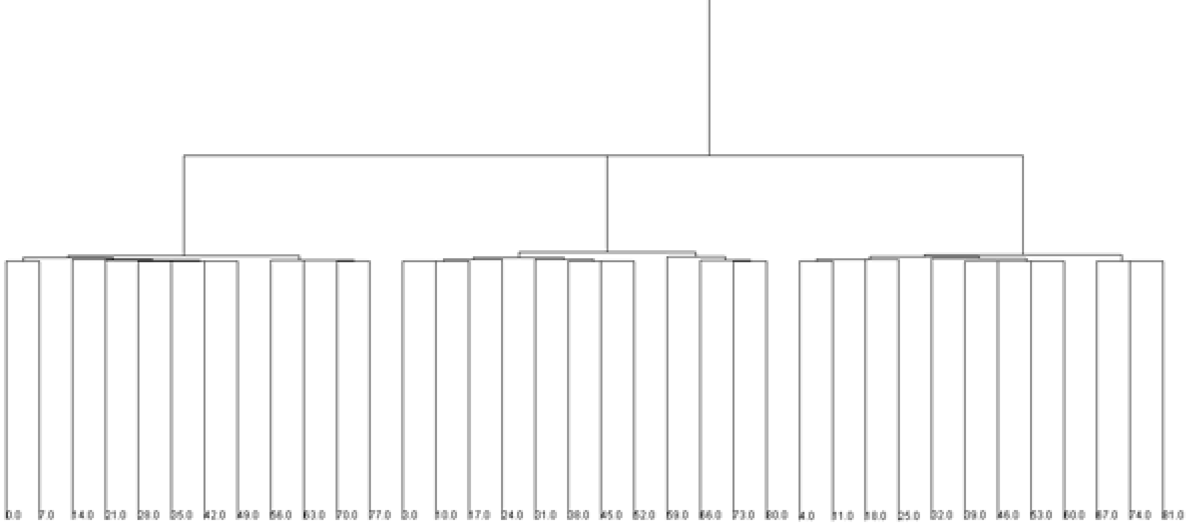
Properties:

Euclidean Distance

No of clusters = 5



```
Clustered Instances

0        4 ( 15%)
1        7 ( 27%)
2        1 (  4%)
3       10 ( 38%)
4        4 ( 15%)
```



Hierarchical Clustering on Natural Gas

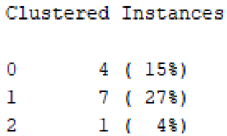
Features: Month, Rig count, Production per rig, Total production, Region

Model: Hierarchical (Clustering)

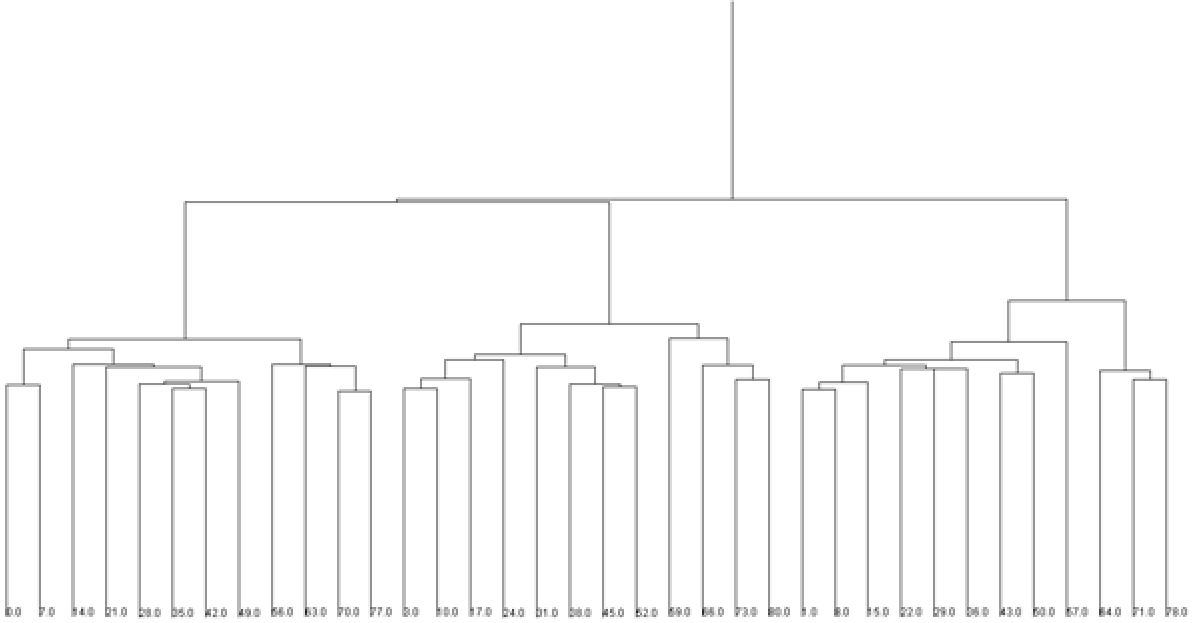
Properties:

Manhattan Distance

No of clusters = 5



```
Clustered Instances

0       4 ( 15%)
1       7 ( 27%)
2       1 (  4%)
```



Density-Based Clusters on Natural Gas

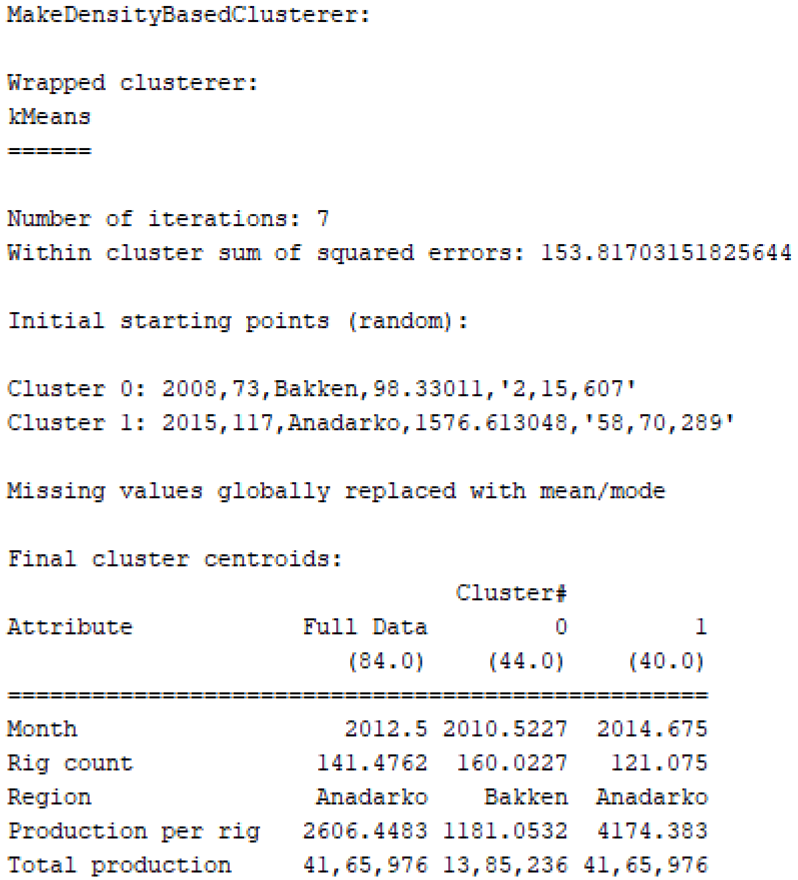
Features: Month, Rig count, Production per rig, Total productio, Region

Model: Density Based (Clustering)

Properties

No of clusters = 2

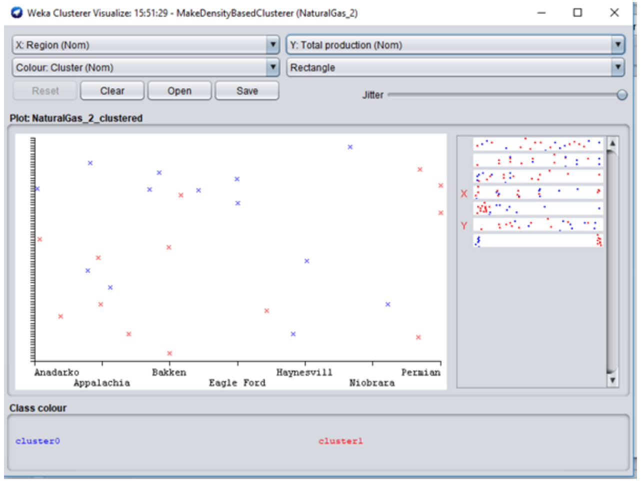
Percentage split = 70



```
MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 153.81703151825644

Initial starting points (random):

Cluster 0: 2008,73,Bakken,98.33011,'2,15,607'
Cluster 1: 2015,117,Anadarko,1576.613048,'58,70,289'

Missing values globally replaced with mean/mode

Final cluster centroids:
                                     Cluster#
Attribute           Full Data          0          1
                       (84.0)     (44.0)     (40.0)
===================================================
Month                  2012.5  2010.5227   2014.675
Rig count            141.4762   160.0227    121.075
Region               Anadarko     Bakken   Anadarko
Production per rig  2606.4483  1181.0532   4174.383
Total production    41,65,976  13,85,236  41,65,976
```

Evaluation of Model

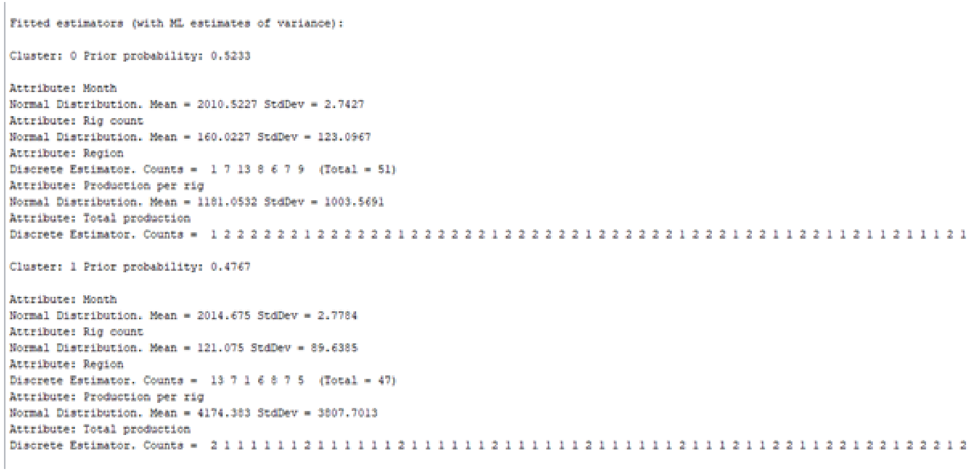
```
=== Model and evaluation on test split ===
MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 102.73881628823784

Initial starting points (random):

Cluster 0: 2007,173,Anadarko,860.851255,'41,65,976'
Cluster 1: 2014,537,Permian,429.1254,'56,69,913'

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute            Full Data          0          1
                        (58.0)     (30.0)     (28.0)
===================================================
Month                2012.2759  2014.0333  2010.3929
Rig count             144.8276       77.3   217.1786
Region                Niobrara   Niobrara    Permian
Production per rig   2416.4488  3946.0605   777.5791
Total production     41,65,976  41,65,976  13,85,236
```



Weka Clusterer Visualize: 15:51:29 - MakeDensityBasedClusterer (NaturalGas_2)

X: Region (Nom)
Y: Total production (Nom)
Colour: Cluster (Nom)
Rectangle

Plot: NaturalGas_2_clustered



```
Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.5233

Attribute: Month
Normal Distribution. Mean = 2010.5227 StdDev = 2.7427
Attribute: Rig count
Normal Distribution. Mean = 160.0227 StdDev = 123.0967
Attribute: Region
Discrete Estimator. Counts =  1 7 13 8 6 7 9  (Total = 51)
Attribute: Production per rig
Normal Distribution. Mean = 1181.0532 StdDev = 1003.5691
Attribute: Total production
Discrete Estimator. Counts =  1 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 1 2 2 1 1 2 2 1 1 2 1 1 2 1 1 1 2 1

Cluster: 1 Prior probability: 0.4767

Attribute: Month
Normal Distribution. Mean = 2014.675 StdDev = 2.7784
Attribute: Rig count
Normal Distribution. Mean = 121.075 StdDev = 89.6385
Attribute: Region
Discrete Estimator. Counts =  13 7 1 6 8 7 5  (Total = 47)
Attribute: Production per rig
Normal Distribution. Mean = 4174.383 StdDev = 3807.7013
Attribute: Total production
Discrete Estimator. Counts =  2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1 2 2 1 1 2 2 1 2 2 1 2 2 2 1 2
```

Farthest First Clusters on Natural Gas

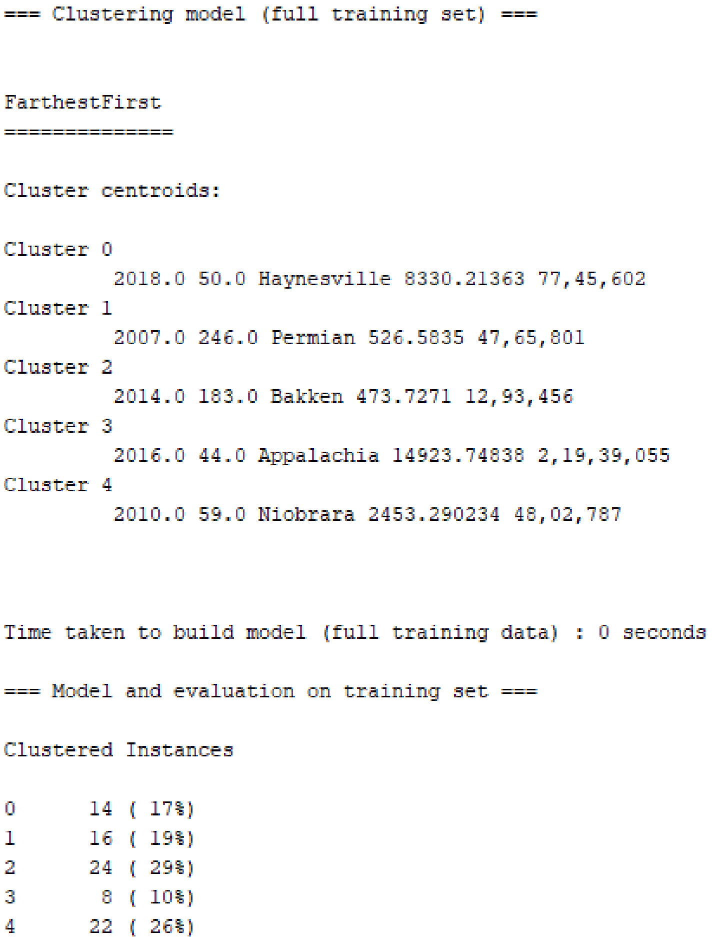
Features: Month, Rig count, Production per rig, Total production, Region

Model: Density Based (Clustering)

Properties

No of clusters = 5

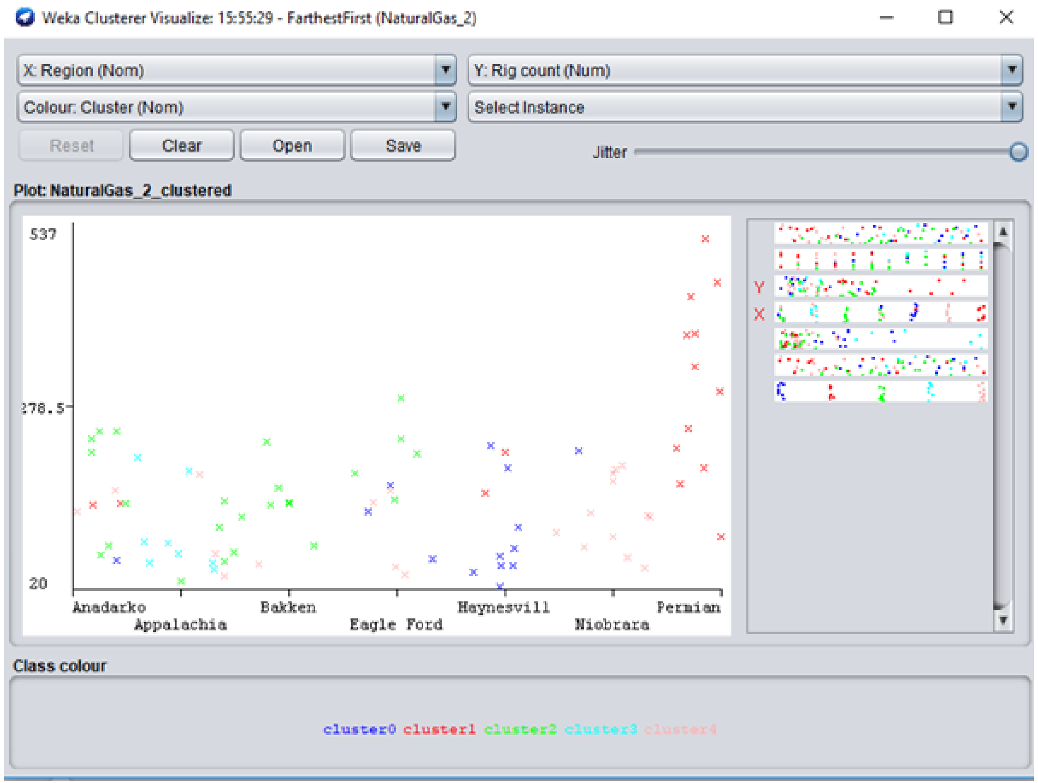
Percentage split = 70



```
=== Clustering model (full training set) ===


FarthestFirst
==============

Cluster centroids:

Cluster 0
		2018.0 50.0 Haynesville 8330.21363 77,45,602
Cluster 1
		2007.0 246.0 Permian 526.5835 47,65,801
Cluster 2
		2014.0 183.0 Bakken 473.7271 12,93,456
Cluster 3
		2016.0 44.0 Appalachia 14923.74838 2,19,39,055
Cluster 4
		2010.0 59.0 Niobrara 2453.290234 48,02,787



Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0      14 ( 17%)
1      16 ( 19%)
2      24 ( 29%)
3       8 ( 10%)
4      22 ( 26%)
```

Evaluation of Model

Classifier Visualization



Weka Clusterer Visualize: 15:55:29 - FarthestFirst (NaturalGas_2)

X: Region (Nom)

Y: Rig count (Num)

Colour: Cluster (Nom)

Select Instance

Plot: NaturalGas_2_clustered

Class colour

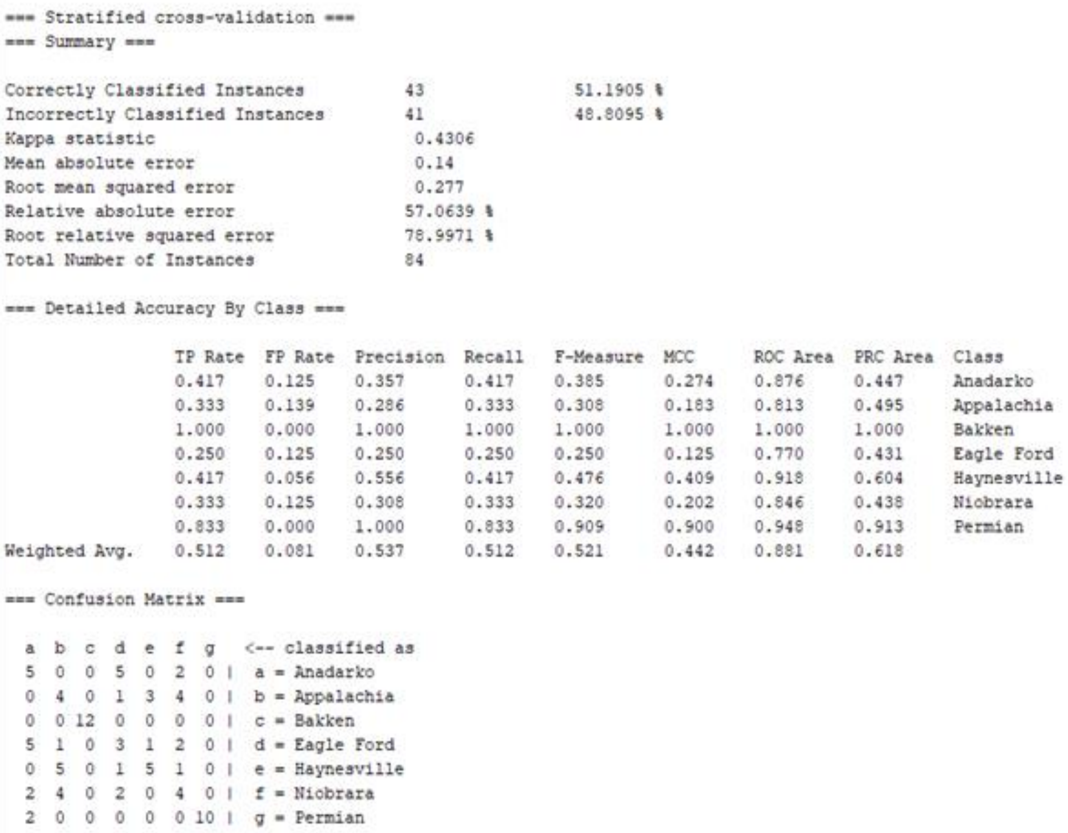
Logistic Regression on Natural Gas

Features: Month, Rig count, Production per rig, Total production

Response: Region

Model: Logistic Regression (Classification)

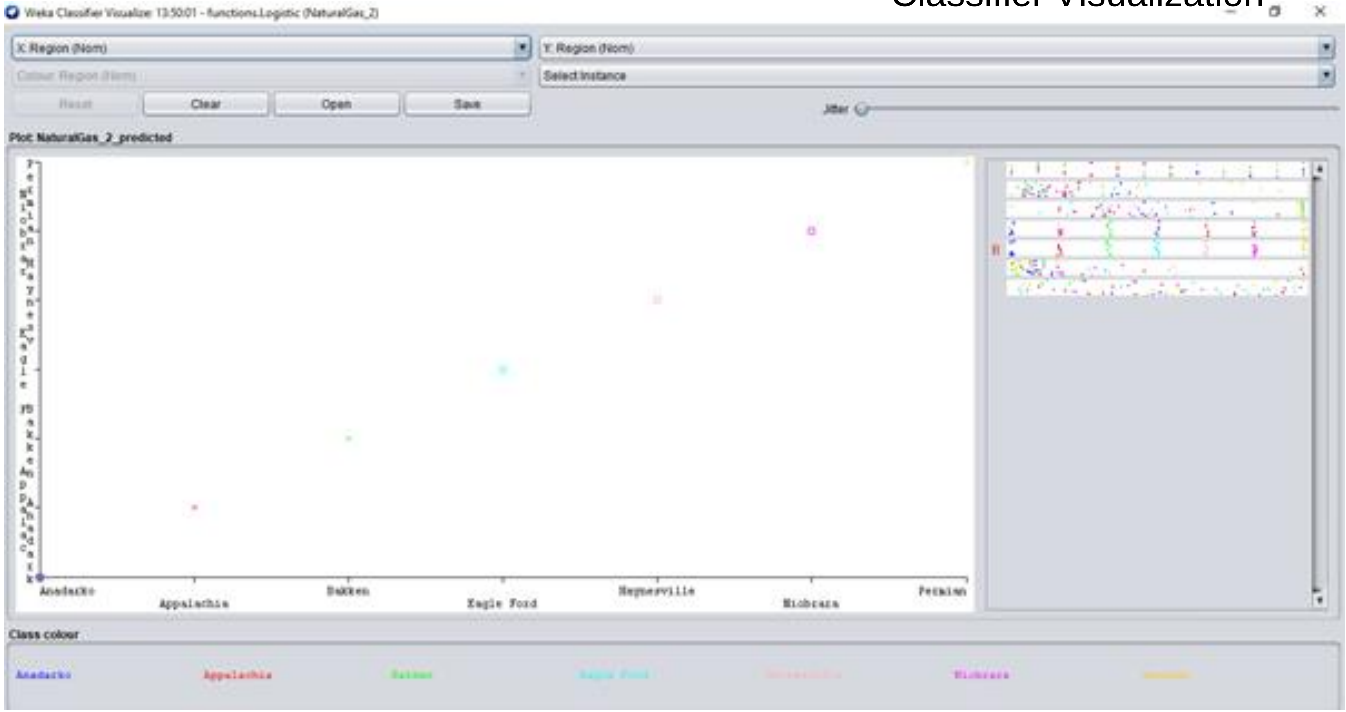
10-fold cross-validation



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

Correctly Classified Instances          43               51.1905 %
Incorrectly Classified Instances        41               48.8095 %
Kappa statistic                          0.4306
Mean absolute error                      0.14
Root mean squared error                  0.277
Relative absolute error                 57.0639 %
Root relative squared error             78.9971 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.417    0.125    0.357      0.417    0.385      0.274    0.876     0.447     Anadarko
                 0.333    0.139    0.286      0.333    0.308      0.183    0.813     0.495     Appalachia
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Bakken
                 0.250    0.125    0.250      0.250    0.250      0.125    0.770     0.431     Eagle Ford
                 0.417    0.056    0.556      0.417    0.476      0.409    0.918     0.604     Haynesville
                 0.333    0.125    0.308      0.333    0.320      0.202    0.846     0.438     Niobrara
                 0.833    0.000    1.000      0.833    0.909      0.900    0.948     0.913     Permian
Weighted Avg.    0.512    0.081    0.537      0.512    0.521      0.442    0.881     0.618

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  5  0  0  5  0  2  0 |  a = Anadarko
  0  4  0  1  3  4  0 |  b = Appalachia
  0  0 12  0  0  0  0 |  c = Bakken
  5  1  0  3  1  2  0 |  d = Eagle Ford
  0  5  0  1  5  1  0 |  e = Haynesville
  2  4  0  2  0  4  0 |  f = Niobrara
  2  0  0  0  0  0 10 |  g = Permian
```

Evaluation of Model

Classifier Visualization



Weka Classifier Visualize: 13:50:01 - functions.Logistic (NaturalGas_2)

X: Region (Nom)

Y: Region (Nom)

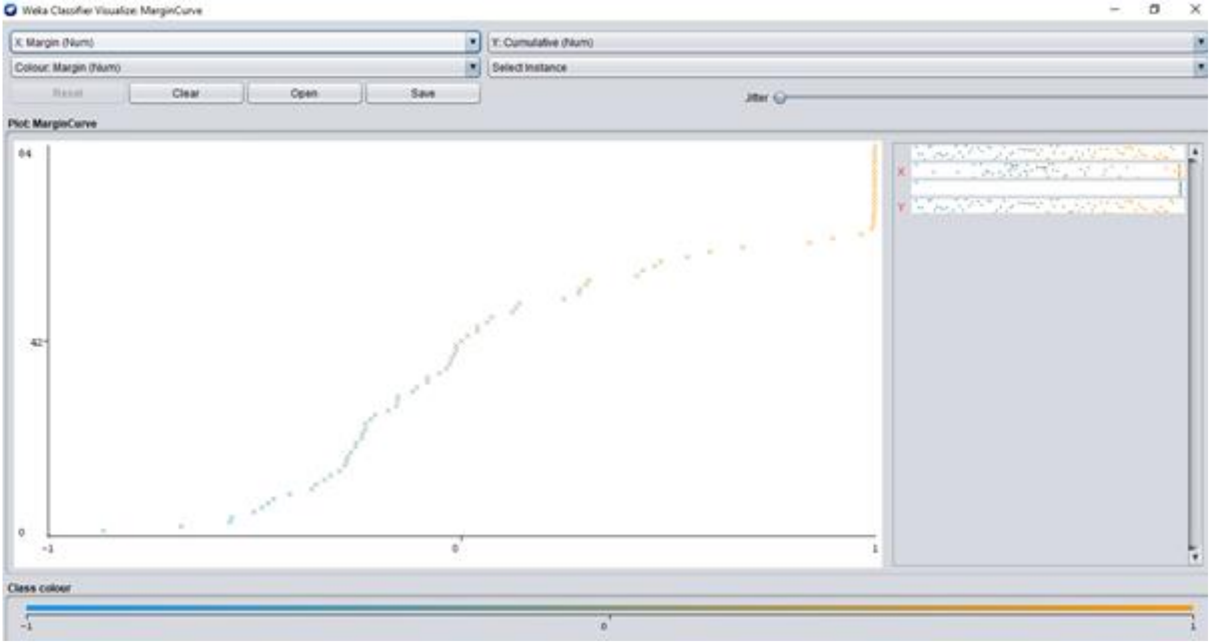
Colour: Region (Nom)

Select Instance

Plot: NaturalGas_2_predicted

Evaluation of Model

Margin Curve



Weka Classifier Visualize: MarginCurve

X: Margin (Num)

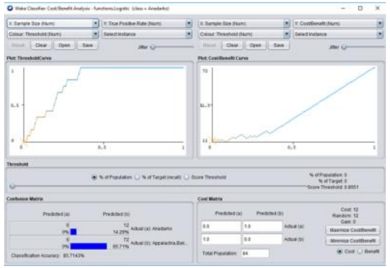
Y: Cumulative (Num)

Colour: Margin (Num)

Select Instance

Region-wise Threshold Curve and Cost-Benefit Analysis

Anadarko



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.8551

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Anadarko
0 0%	72 85.71%	Actual (b): Appalachia,Bak...

Classification Accuracy: 85.7143%

Cost Matrix

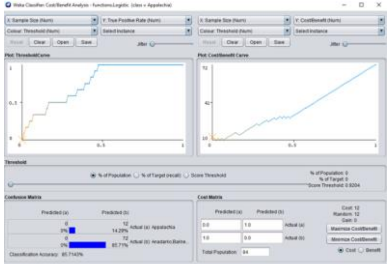
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

Appalachia



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.8204

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Appalachia
0 0%	72 85.71%	Actual (b): Anadarko,Bakke...

Classification Accuracy: 85.7143%

Cost Matrix

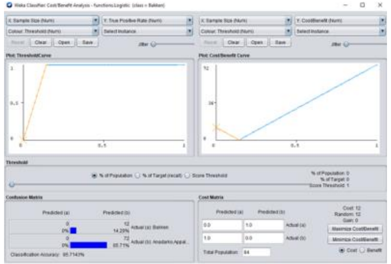
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

Bakken



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Bakken
0 0%	72 85.71%	Actual (b): Anadarko,Appal...

Classification Accuracy: 85.7143%

Cost Matrix

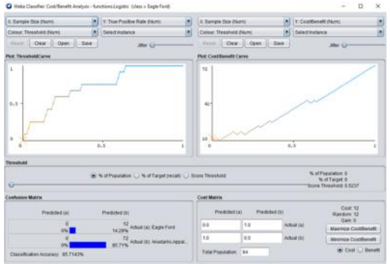
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

Eagle Ford



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.5237

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Eagle Ford
0 0%	72 85.71%	Actual (b): Anadarko,Appal...

Classification Accuracy: 85.7143%

Cost Matrix

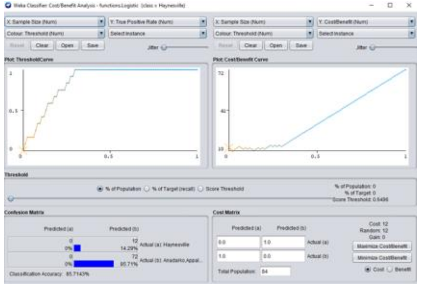
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

Haynesville



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.5496

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Haynesville
0 0%	72 85.71%	Actual (b): Anadarko,Appal...

Classification Accuracy: 85.7143%

Cost Matrix

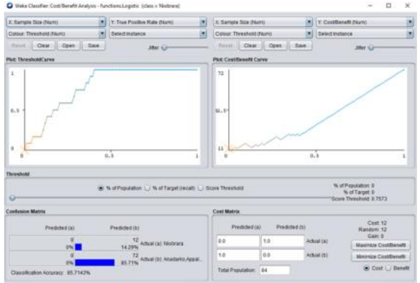
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

Niobrara



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.7573

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Niobrara
0 0%	72 85.71%	Actual (b): Anadarko,Appal...

Classification Accuracy: 85.7143%

Cost Matrix

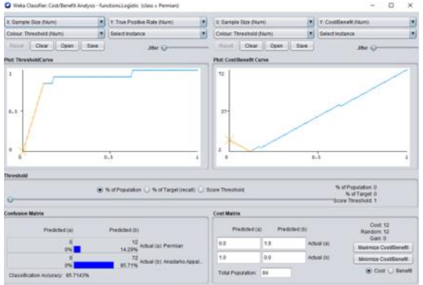
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

Permian



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	12 14.29%	Actual (a): Permian
0 0%	72 85.71%	Actual (b): Anadarko,Appal...

Classification Accuracy: 85.7143%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 84

Cost: 12
Random: 12
Gain: 0

● Cost ○ Benefit

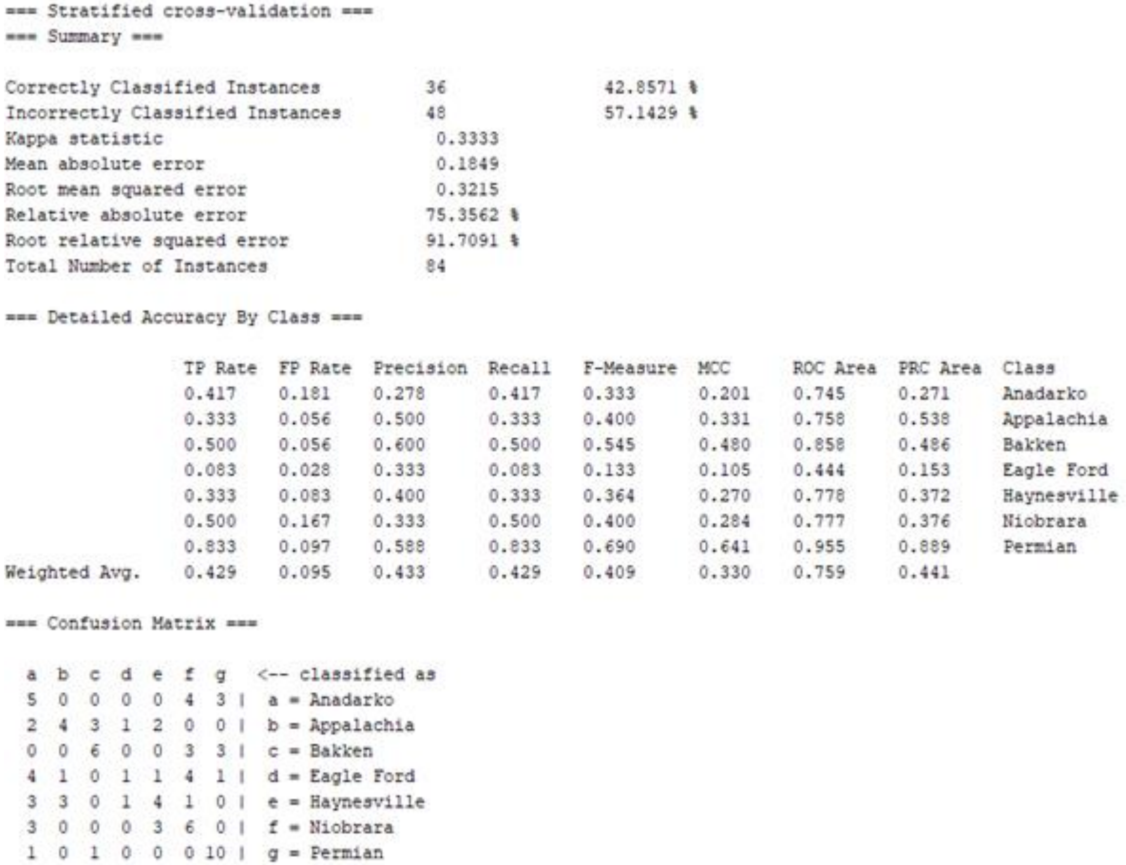
Naïve Bayes on Natural Gas

Features: Month, Rig count, Production per rig, Total production

Response: Region

Model: Naive Bayes(Classification)

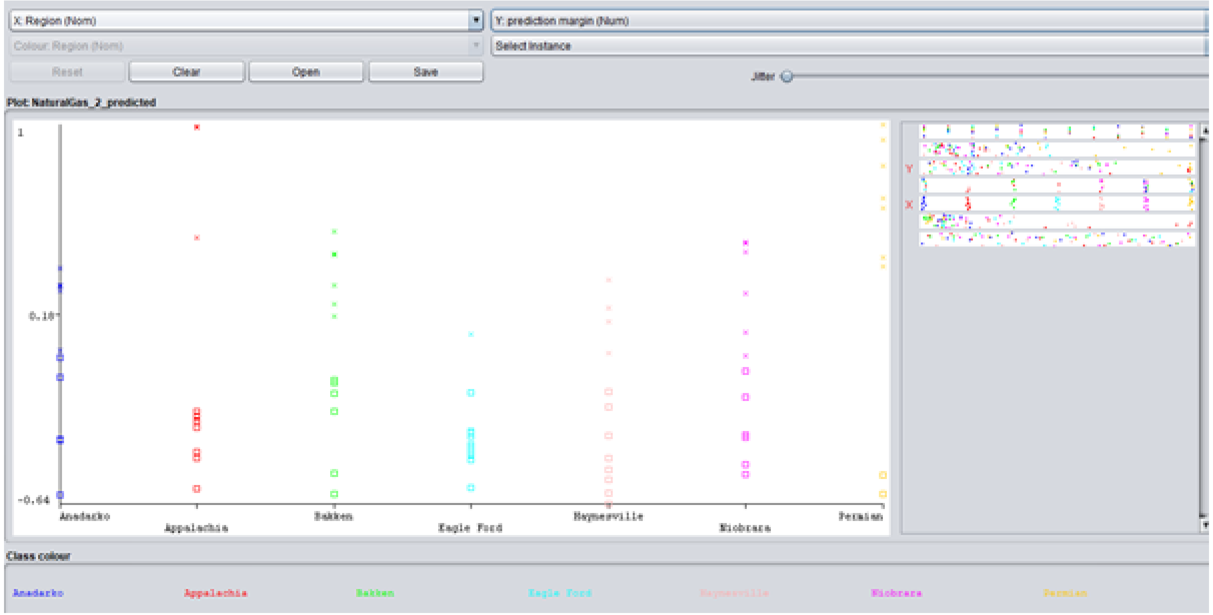
10-fold cross-validation



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

Correctly Classified Instances          36               42.8571 %
Incorrectly Classified Instances        48               57.1429 %
Kappa statistic                          0.3333
Mean absolute error                      0.1849
Root mean squared error                  0.3215
Relative absolute error                 75.3562 %
Root relative squared error             91.7091 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.417    0.181    0.278      0.417    0.333      0.201    0.745     0.271     Anadarko
                 0.333    0.056    0.500      0.333    0.400      0.331    0.758     0.538     Appalachia
                 0.500    0.056    0.600      0.500    0.545      0.480    0.858     0.486     Bakken
                 0.083    0.028    0.333      0.083    0.133      0.105    0.444     0.153     Eagle Ford
                 0.333    0.083    0.400      0.333    0.364      0.270    0.778     0.372     Haynesville
                 0.500    0.167    0.333      0.500    0.400      0.284    0.777     0.376     Niobrara
                 0.833    0.097    0.588      0.833    0.690      0.641    0.955     0.889     Permian
Weighted Avg.    0.429    0.095    0.433      0.429    0.409      0.330    0.759     0.441

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  5  0  0  0  0  4  3 |  a = Anadarko
  2  4  3  1  2  0  0 |  b = Appalachia
  0  0  6  0  0  3  3 |  c = Bakken
  4  1  0  1  1  4  1 |  d = Eagle Ford
  3  3  0  1  4  1  0 |  e = Haynesville
  3  0  0  0  3  6  0 |  f = Niobrara
  1  0  1  0  0  0 10 |  g = Permian
```

Evaluation of Model

Classifier Visualization



X: Region (Nom)

Y: prediction margin (Num)

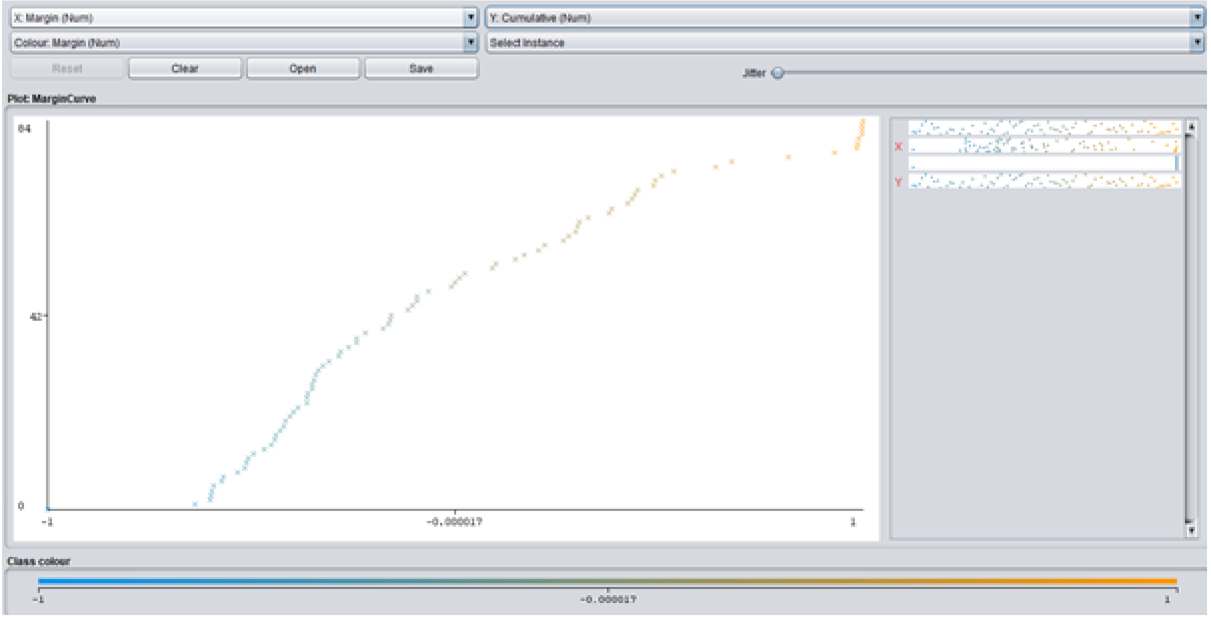
Colour: Region (Nom)

Select Instance

Plot: NaturalGas_2_predicted

Evaluation of Model

Margin Curve



X: Margin (Num)

Y: Cumulative (Num)

Colour: Margin (Num)

Select Instance

Plot: MarginCurve

Decision Tree on Natural Gas

Features: Month, Rig count, Production per rig, Total production

Response: Region

Model: Decision Tree (Classification)

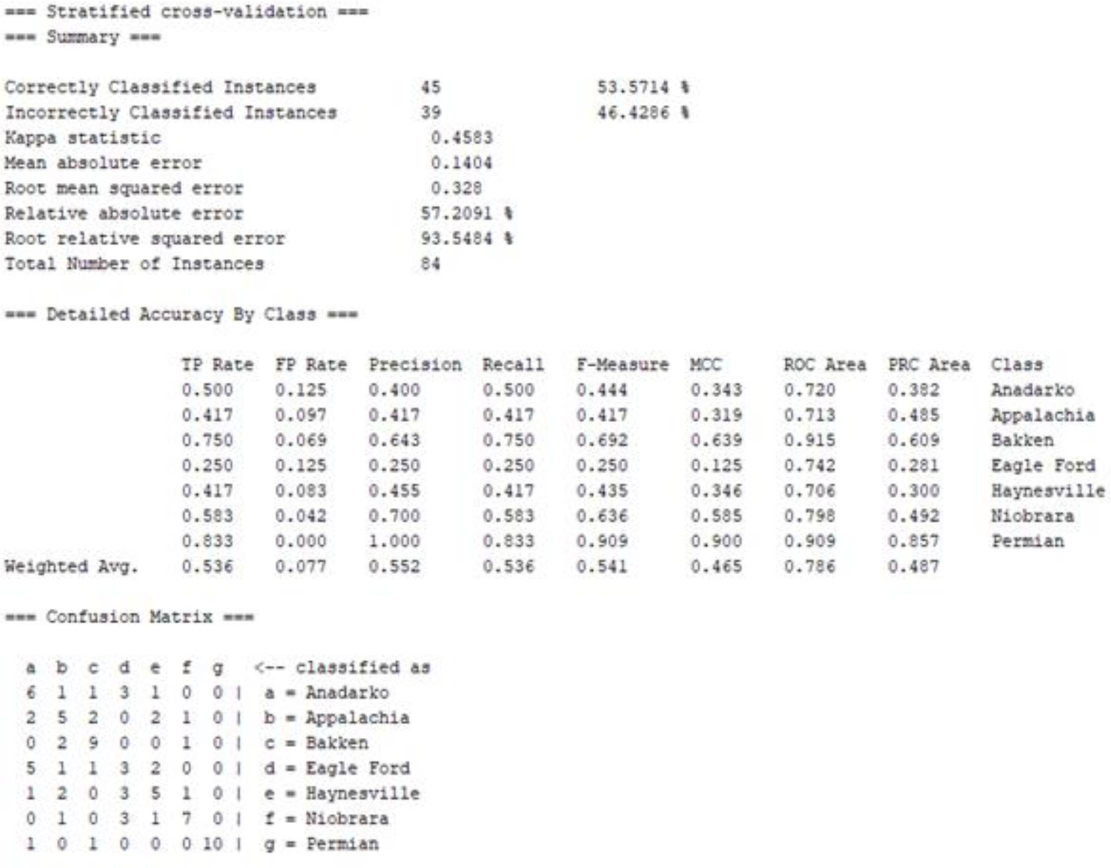
Filter: J48

Properties:

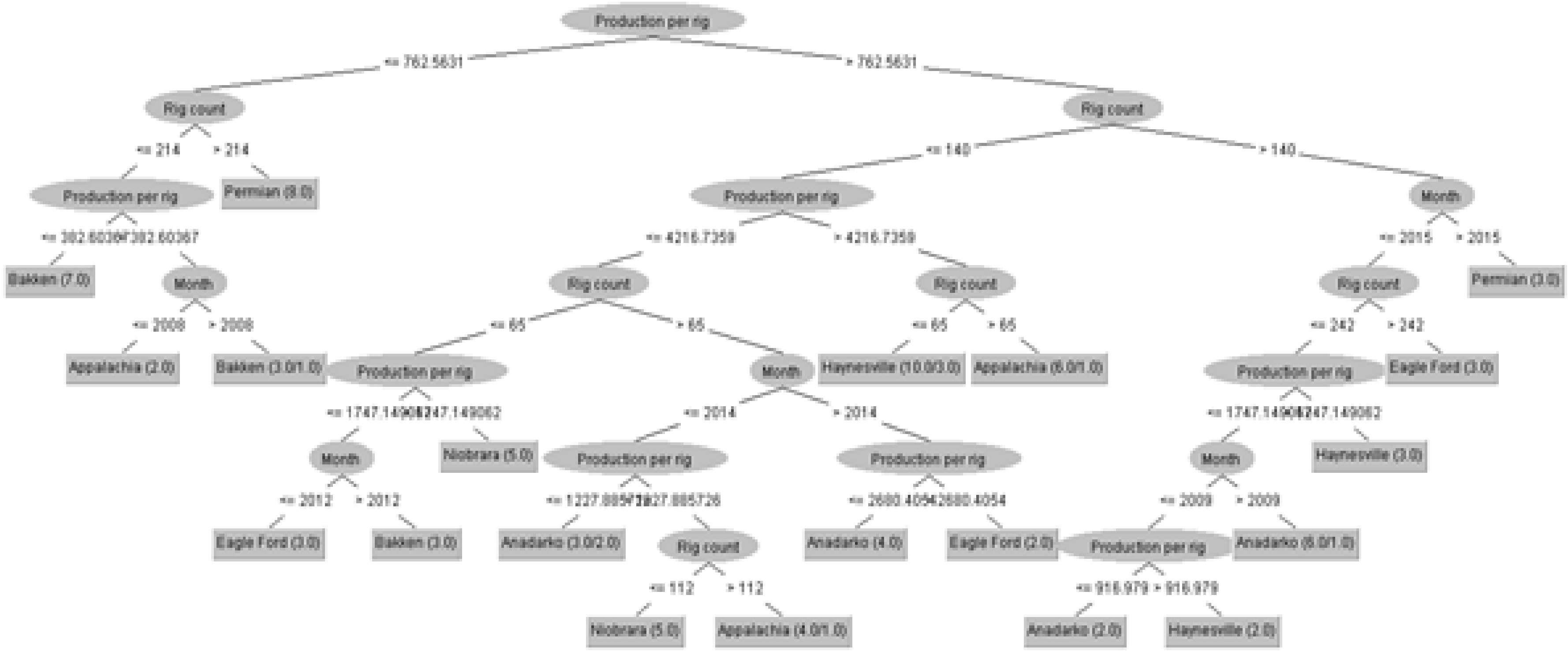
confidenceFactor = 0.25

numFolds = 3

unpruned = false



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

Correctly Classified Instances          45               53.5714 %
Incorrectly Classified Instances        39               46.4286 %
Kappa statistic                          0.4583
Mean absolute error                      0.1404
Root mean squared error                  0.328
Relative absolute error                 57.2091 %
Root relative squared error             93.5484 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.125    0.400      0.500    0.444      0.343    0.720     0.382     Anadarko
                 0.417    0.097    0.417      0.417    0.417      0.319    0.713     0.485     Appalachia
                 0.750    0.069    0.643      0.750    0.692      0.639    0.915     0.609     Bakken
                 0.250    0.125    0.250      0.250    0.250      0.125    0.742     0.281     Eagle Ford
                 0.417    0.083    0.455      0.417    0.435      0.346    0.706     0.300     Haynesville
                 0.583    0.042    0.700      0.583    0.636      0.585    0.798     0.492     Niobrara
                 0.833    0.000    1.000      0.833    0.909      0.900    0.909     0.857     Permian
Weighted Avg.    0.536    0.077    0.552      0.536    0.541      0.465    0.786     0.487

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  6  1  1  3  1  0  0 |  a = Anadarko
  2  5  2  0  2  1  0 |  b = Appalachia
  0  2  9  0  0  1  0 |  c = Bakken
  5  1  1  3  2  0  0 |  d = Eagle Ford
  1  2  0  3  5  1  0 |  e = Haynesville
  0  1  0  3  1  7  0 |  f = Niobrara
  1  0  1  0  0  0 10 |  g = Permian
```



Production per rig
<= 762.5631
> 762.5631
Rig count
<= 214
> 214
Production per rig
Permian (8.0)
<= 392.60367
> 392.60367
Bakken (7.0)
Month
<= 2008
> 2008
Appalachia (2.0)
Bakken (3.0/1.0)
Rig count
<= 140
> 140
Production per rig
<= 4216.7359
> 4216.7359
Rig count
<= 65
> 65
Production per rig
<= 1747.149062
> 1747.149062
Month
<= 2012
> 2012
Eagle Ford (3.0)
Bakken (3.0)
Niobrara (5.0)
Month
<= 2014
> 2014
Production per rig
<= 1227.885726
> 1227.885726
Anadarko (3.0/2.0)
Rig count
<= 112
> 112
Niobrara (5.0)
Appalachia (4.0/1.0)
Production per rig
<= 2680.4054
> 2680.4054
Anadarko (4.0)
Eagle Ford (2.0)
Rig count
<= 65
> 65
Haynesville (10.0/3.0)
Appalachia (6.0/1.0)
Month
<= 2015
> 2015
Rig count
Permian (3.0)
<= 242
> 242
Production per rig
Eagle Ford (3.0)
<= 1747.149062
> 1747.149062
Month
Haynesville (3.0)
<= 2009
> 2009
Production per rig
Anadarko (6.0/1.0)
<= 916.979
> 916.979
Anadarko (2.0)
Haynesville (2.0)

Apriori Rule Association on Natural Gas

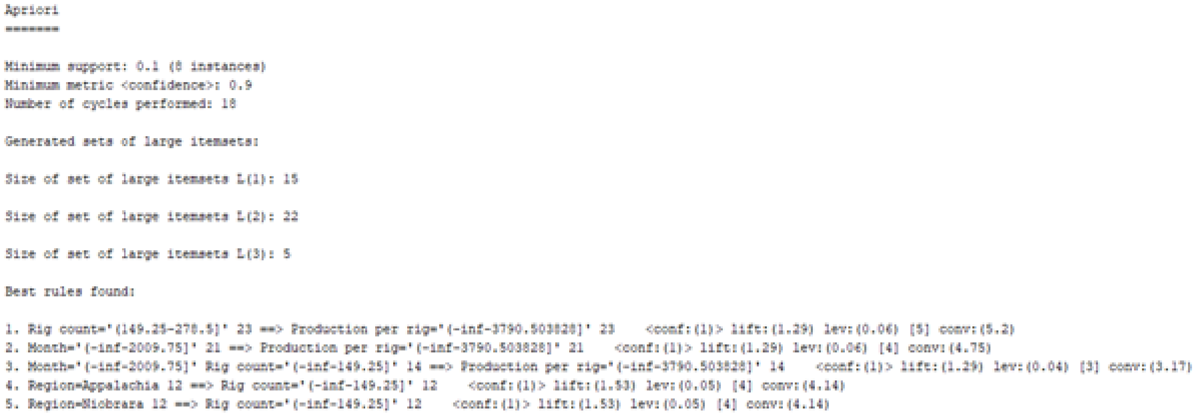
Features: Month, Rig count, Production per rig, Total production, Region

Model: Density Based (Clustering)

Properties

No of Rules = 5

Min Metric = 0.9



```
Apriori
=======

Minimum support: 0.1 (8 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 22

Size of set of large itemsets L(3): 5

Best rules found:

1. Rig count='(149.25-278.5]' 23 ==> Production per rig='(-inf-3790.503828]' 23    <conf:(1)> lift:(1.29) lev:(0.06) [5] conv:(5.2)
2. Month='(-inf-2009.75]' 21 ==> Production per rig='(-inf-3790.503828]' 21    <conf:(1)> lift:(1.29) lev:(0.06) [4] conv:(4.75)
3. Month='(-inf-2009.75]' Rig count='(-inf-149.25]' 14 ==> Production per rig='(-inf-3790.503828]' 14    <conf:(1)> lift:(1.29) lev:(0.04) [3] conv:(3.17)
4. Region=Appalachia 12 ==> Rig count='(-inf-149.25]' 12    <conf:(1)> lift:(1.53) lev:(0.05) [4] conv:(4.14)
5. Region=Niobrara 12 ==> Rig count='(-inf-149.25]' 12    <conf:(1)> lift:(1.53) lev:(0.05) [4] conv:(4.14)
```

Modeling Relationship between Natural Gas Prices and US Petroleum Production

Simple K Means on Natural Gas vs. Price

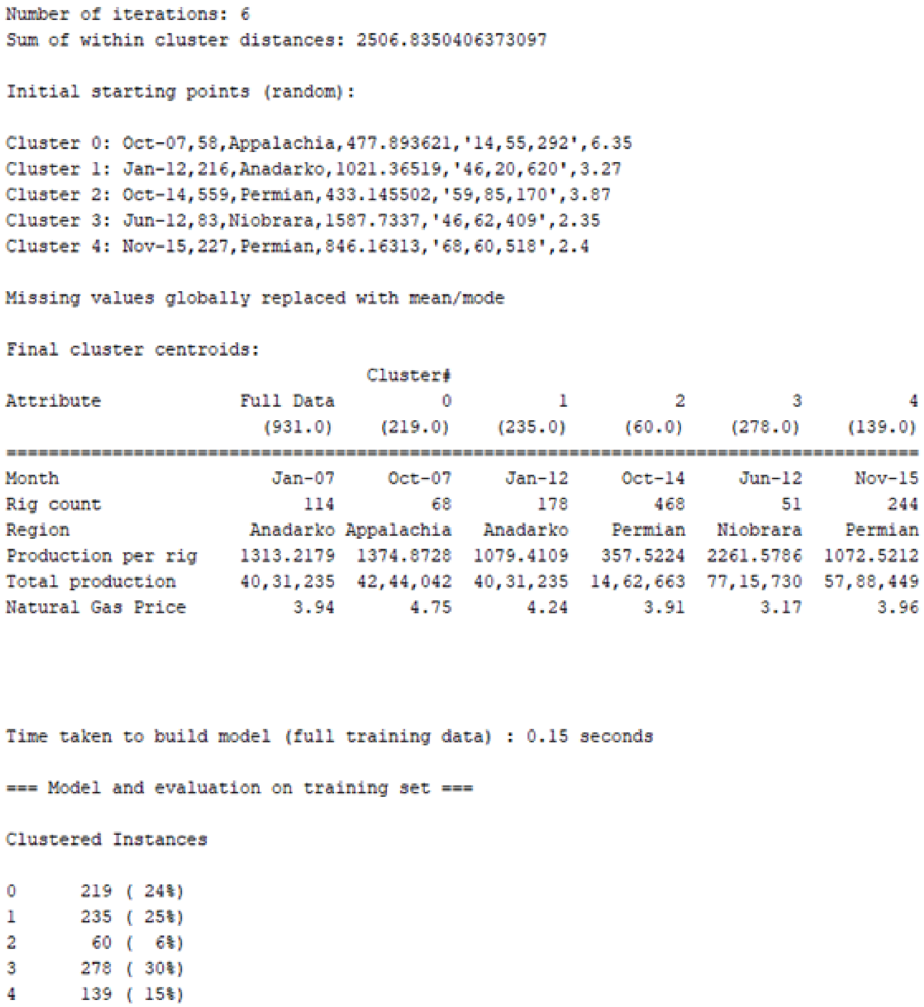
Features: Month, Rig count, Production per rig, Total production, Region, Natural Gas Price

Model: Simple K Means(Clustering)

Properties:
Manhattan distance

Max iterations = 200

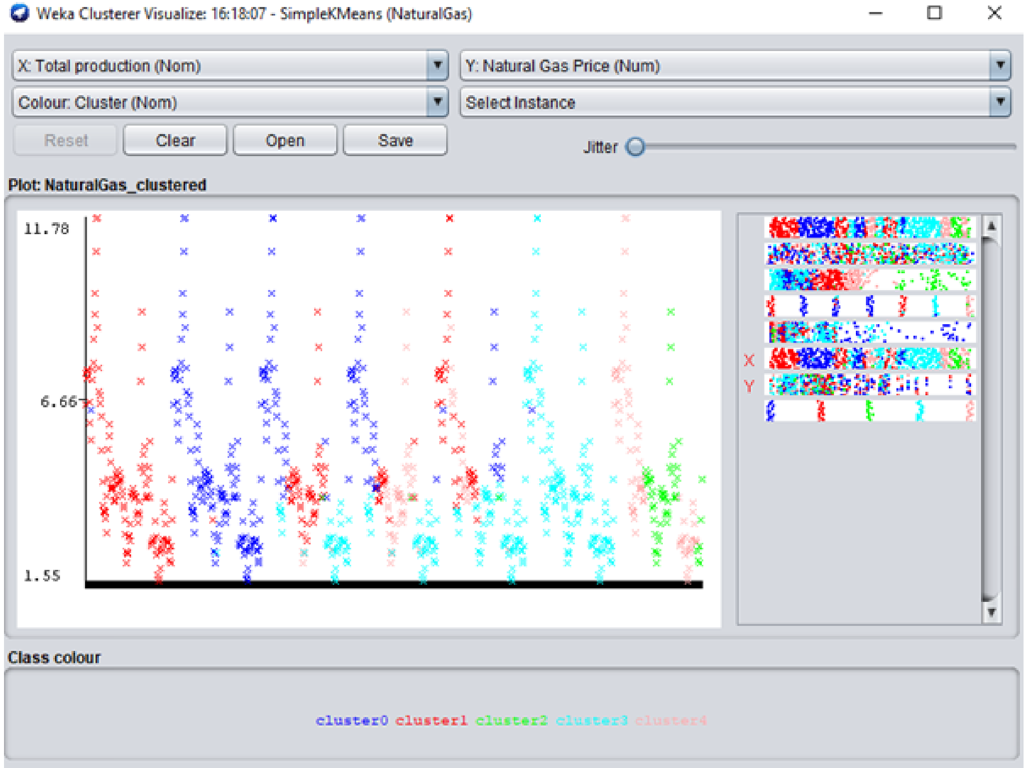
No of clusters = 5



```
Number of iterations: 6
Sum of within cluster distances: 2506.8350406373097

Initial starting points (random):

Cluster 0: Oct-07,58,Appalachia,477.893621,'14,55,292',6.35
Cluster 1: Jan-12,216,Anadarko,1021.36519,'46,20,620',3.27
Cluster 2: Oct-14,559,Permian,433.145502,'59,85,170',3.87
Cluster 3: Jun-12,83,Niobrara,1587.7337,'46,62,409',2.35
Cluster 4: Nov-15,227,Permian,846.16313,'68,60,518',2.4

Missing values globally replaced with mean/mode

Final cluster centroids:
                                  Cluster#
Attribute             Full Data          0          1          2          3          4
                        (931.0)    (219.0)    (235.0)     (60.0)    (278.0)    (139.0)
======================================================================================
Month                    Jan-07     Oct-07     Jan-12     Oct-14     Jun-12     Nov-15
Rig count                   114         68        178        468         51        244
Region                 Anadarko Appalachia   Anadarko    Permian   Niobrara    Permian
Production per rig    1313.2179  1374.8728  1079.4109   357.5224  2261.5786  1072.5212
Total production      40,31,235  42,44,042  40,31,235  14,62,663  77,15,730  57,88,449
Natural Gas Price          3.94       4.75       4.24       3.91       3.17       3.96




Time taken to build model (full training data) : 0.15 seconds

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

Clustered Instances

0      219 ( 24%)
1      235 ( 25%)
2       60 (  6%)
3      278 ( 30%)
4      139 ( 15%)
```

Evaluation of Model

Classifier Visualization



Weka Clusterer Visualize: 16:18:07 - SimpleKMeans (NaturalGas)

X: Total production (Nom)

Y: Natural Gas Price (Num)

Colour: Cluster (Nom)

Select Instance

Plot: NaturalGas_clustered

Class colour

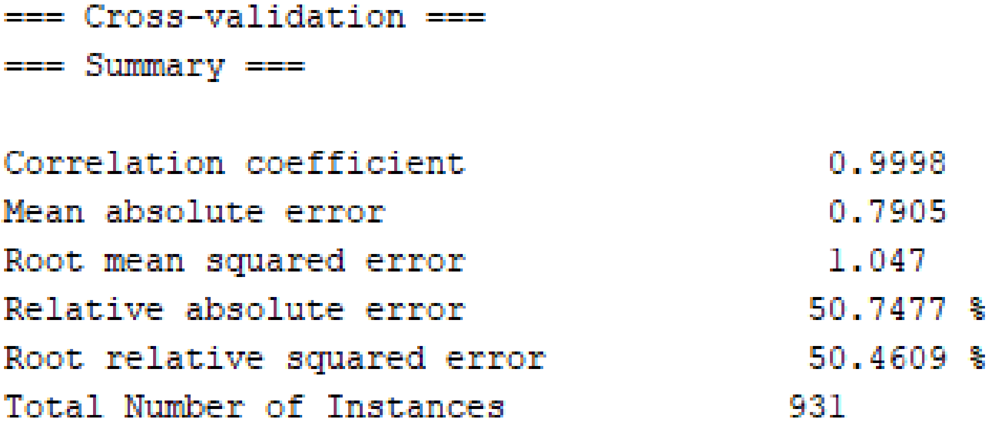
Linear Regression on Natural Gas vs. Price

Features: Month, Rig count, Production per rig, Total production, Region

Response: Natural Gas Price

Model: Linear Regression(Classification)

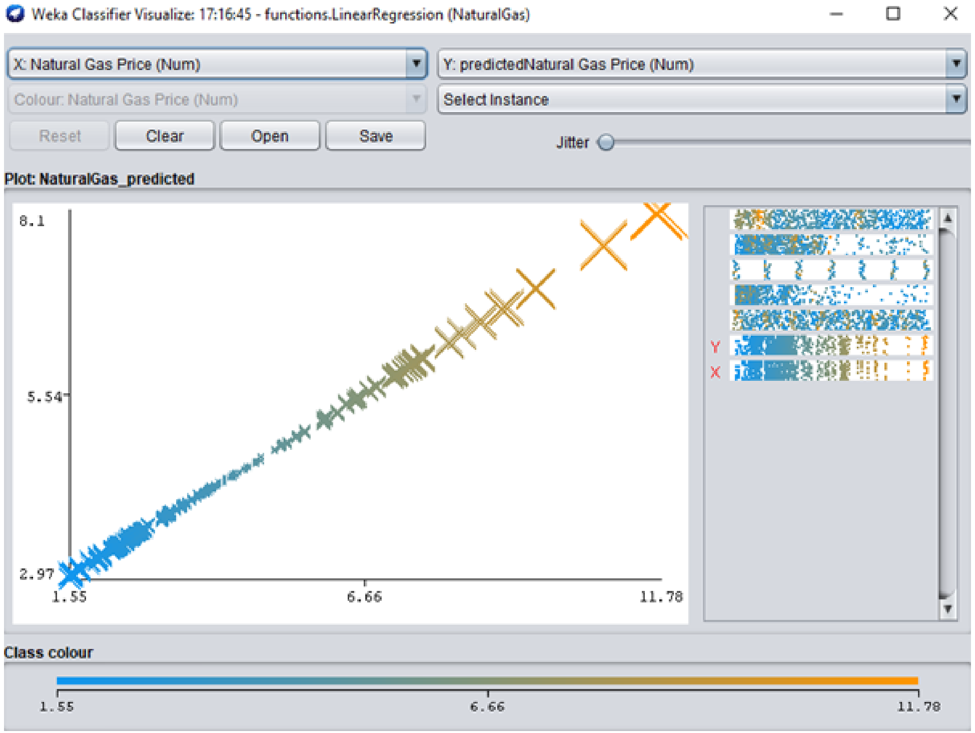
Properties:
Selection Process = M5 method



```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9998
Mean absolute error                      0.7905
Root mean squared error                  1.047
Relative absolute error                 50.7477 %
Root relative squared error             50.4609 %
Total Number of Instances              931
```

Evaluation of Model

Classifier Visualization



Weka Classifier Visualize: 17:16:45 - functions.LinearRegression (NaturalGas)
X: Natural Gas Price (Num)
Y: predictedNatural Gas Price (Num)
Colour: Natural Gas Price (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: NaturalGas_predicted
8.1
5.54
2.97
1.55
6.66
11.78
Y
X
Class colour
1.55
6.66
11.78

Random Forest on Natural Gas vs. Price

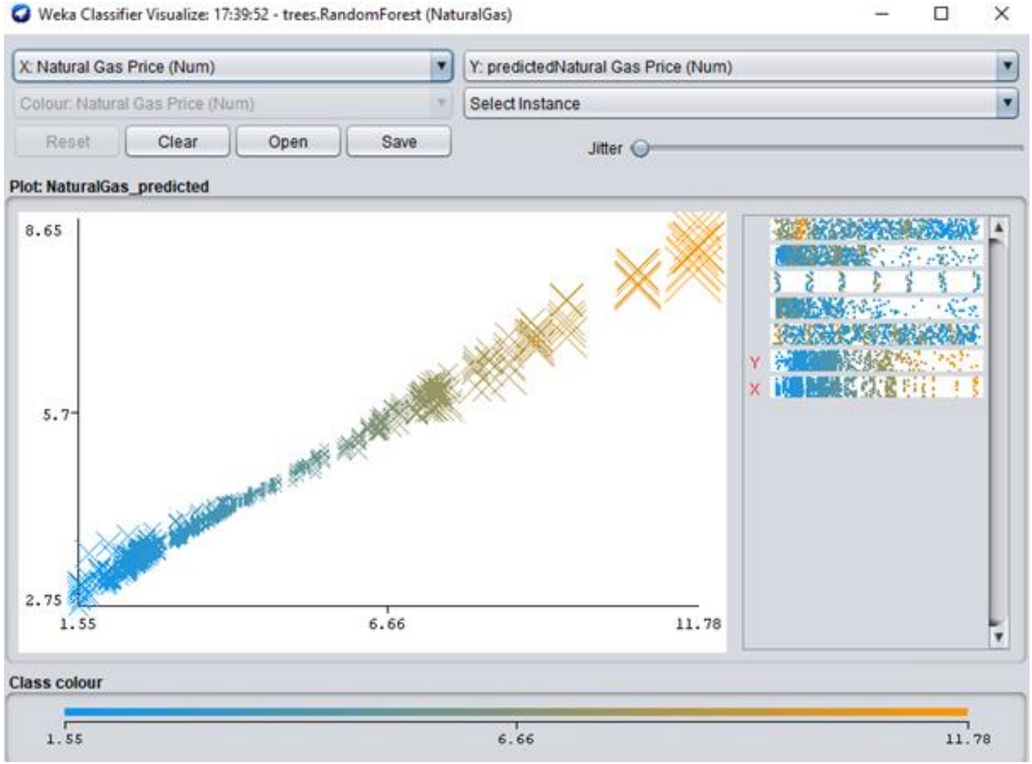
Features: Month, Rig count, Production per rig, Total production, Region

Response: Natural Gas Price

Model: Random Forest (Classification)

Properties:
Attribute importance is true

Batch and bagging size = 100


Weka Classifier Visualize: 17:39:52 - trees.RandomForest (NaturalGas)
X: Natural Gas Price (Num)
Y: predictedNatural Gas Price (Num)
Colour: Natural Gas Price (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: NaturalGas_predicted
8.65
5.7
2.75
1.55
6.66
11.78
Y
X
Class colour
1.55
6.66
11.78

Apriori Rule Association on Natural Gas vs. Price

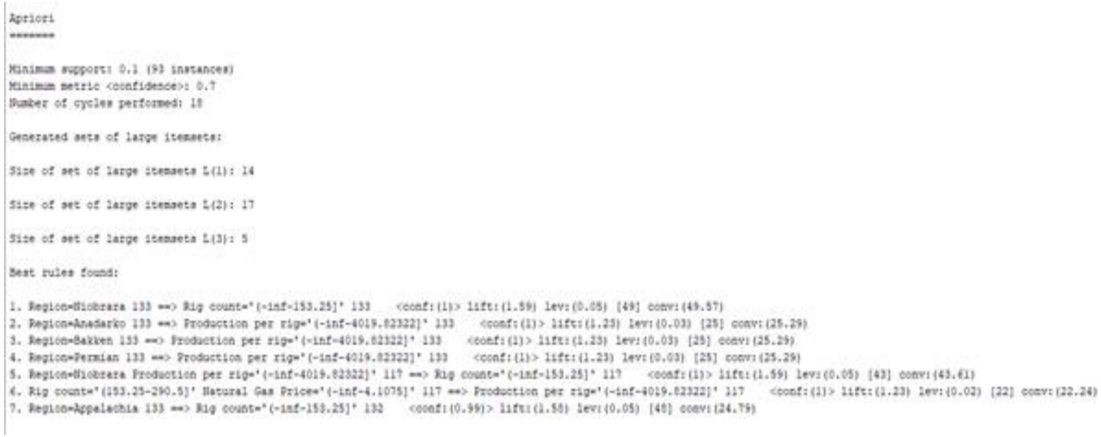
Features: Month, Rig count, Production per rig, Total production, Region, Natural Gas Price

Model: Apriori(Association)

Properties:
No of Rules = 7

Metric Type = confidence

Min metric = 0.7



```
Apriori
=======

Minimum support: 0.1 (93 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 5

Best rules found:

1. Region=Niobrara 133 ==> Rig count='(-inf-153.25]' 133    <conf:(1)> lift:(1.59) lev:(0.05) [49] conv:(49.57)
2. Region=Anadarko 133 ==> Production per rig='(-inf-4019.82322]' 133    <conf:(1)> lift:(1.23) lev:(0.03) [25] conv:(25.29)
3. Region=Bakken 133 ==> Production per rig='(-inf-4019.82322]' 133    <conf:(1)> lift:(1.23) lev:(0.03) [25] conv:(25.29)
4. Region=Permian 133 ==> Production per rig='(-inf-4019.82322]' 133    <conf:(1)> lift:(1.23) lev:(0.03) [25] conv:(25.29)
5. Region=Niobrara Production per rig='(-inf-4019.82322]' 117 ==> Rig count='(-inf-153.25]' 117    <conf:(1)> lift:(1.59) lev:(0.05) [43] conv:(43.61)
6. Rig count='(153.25-290.5]' Natural Gas Price='(-inf-4.1075]' 117 ==> Production per rig='(-inf-4019.82322]' 117    <conf:(1)> lift:(1.23) lev:(0.02) [22] conv:(22.24)
7. Region=Appalachia 133 ==> Rig count='(-inf-153.25]' 132    <conf:(0.99)> lift:(1.58) lev:(0.05) [48] conv:(24.79)
```

Modeling Relationship between Oil Prices and US Petroleum Production

Simple K Means on Oil vs. Price

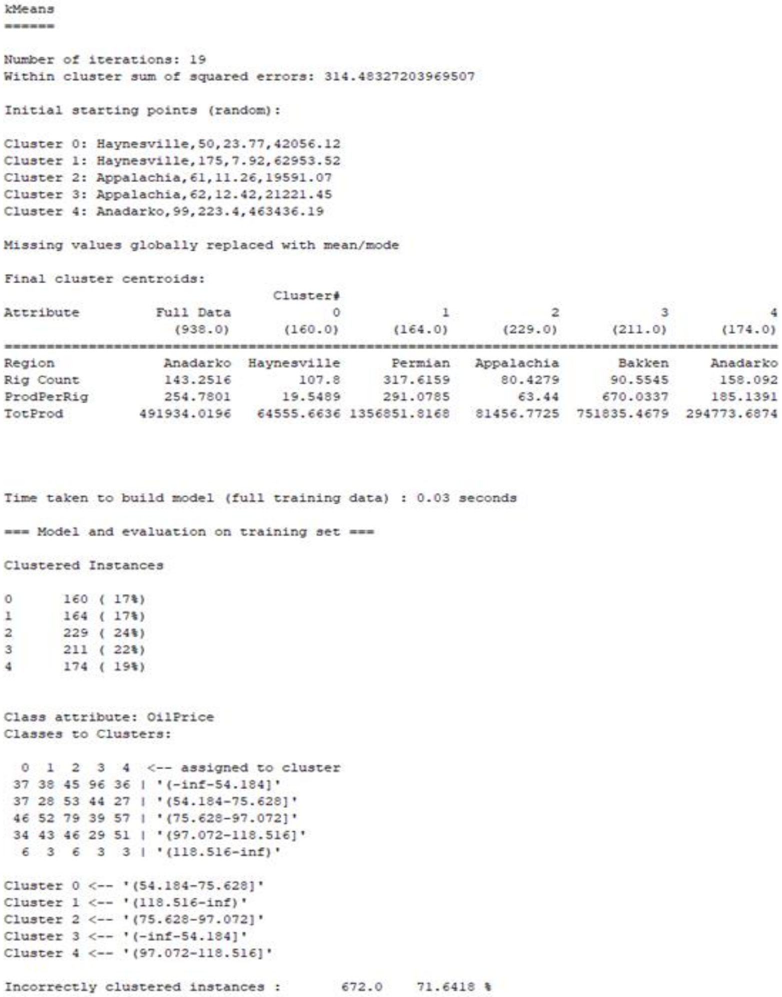
Features: Month, Rig count, Production per rig, Total production, Region, Oil Price

Model: Simple K Means(Clustering)

Properties:
Manhattan distance

Max iterations = 200

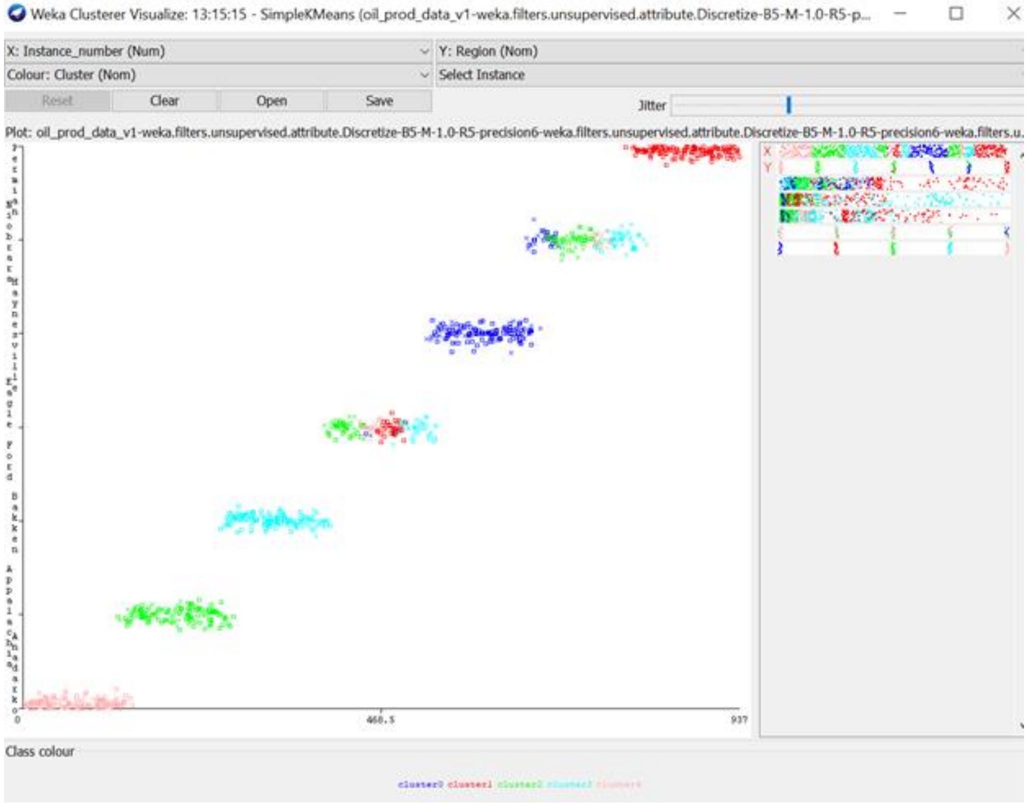
No of clusters = 5



```
kMeans
======

Number of iterations: 19
Within cluster sum of squared errors: 314.48327203969507

Initial starting points (random):

Cluster 0: Haynesville,50,23.77,42056.12
Cluster 1: Haynesville,175,7.92,62953.52
Cluster 2: Appalachia,61,11.26,19591.07
Cluster 3: Appalachia,62,12.42,21221.45
Cluster 4: Anadarko,99,223.4,463436.19

Missing values globally replaced with mean/mode

Final cluster centroids:
                             Cluster#
Attribute       Full Data           0            1            2            3            4
                  (938.0)     (160.0)      (164.0)      (229.0)      (211.0)      (174.0)
==========================================================================================
Region           Anadarko Haynesville      Permian   Appalachia       Bakken     Anadarko
Rig Count        143.2516       107.8     317.6159      80.4279      90.5545      158.092
ProdPerRig       254.7801     19.5489     291.0785        63.44     670.0337     185.1391
TotProd       491934.0196  64555.6636 1356851.8168   81456.7725  751835.4679  294773.6874



Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0       160 ( 17%)
1       164 ( 17%)
2       229 ( 24%)
3       211 ( 22%)
4       174 ( 19%)


Class attribute: OilPrice
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 37 38 45 96 36 | '(-inf-54.184]'
 37 28 53 44 27 | '(54.184-75.628]'
 46 52 79 39 57 | '(75.628-97.072]'
 34 43 46 29 51 | '(97.072-118.516]'
  6  3  6  3  3 | '(118.516-inf)'

Cluster 0 <-- '(54.184-75.628]'
Cluster 1 <-- '(118.516-inf)'
Cluster 2 <-- '(75.628-97.072]'
Cluster 3 <-- '(-inf-54.184]'
Cluster 4 <-- '(97.072-118.516]'

Incorrectly clustered instances :       672.0    71.6418 %
```

Evaluation of Model

Classifier Visualization



Weka Clusterer Visualize: 13:15:15 - SimpleKMeans (oil_prod_data_v1-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-p...

X: Instance_number (Num)

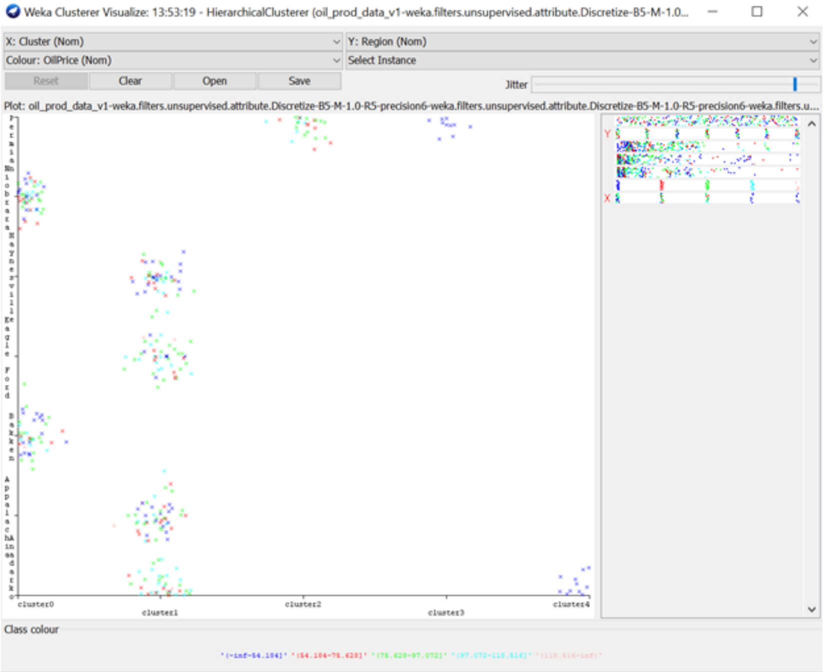
Y: Region (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: oil_prod_data_v1-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.u...

Hierarchical Clustering on Oil vs Price



Weka Clusterer Visualize: 13:53:19 - HierarchicalClusterer (oil_prod_data_v1-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0...

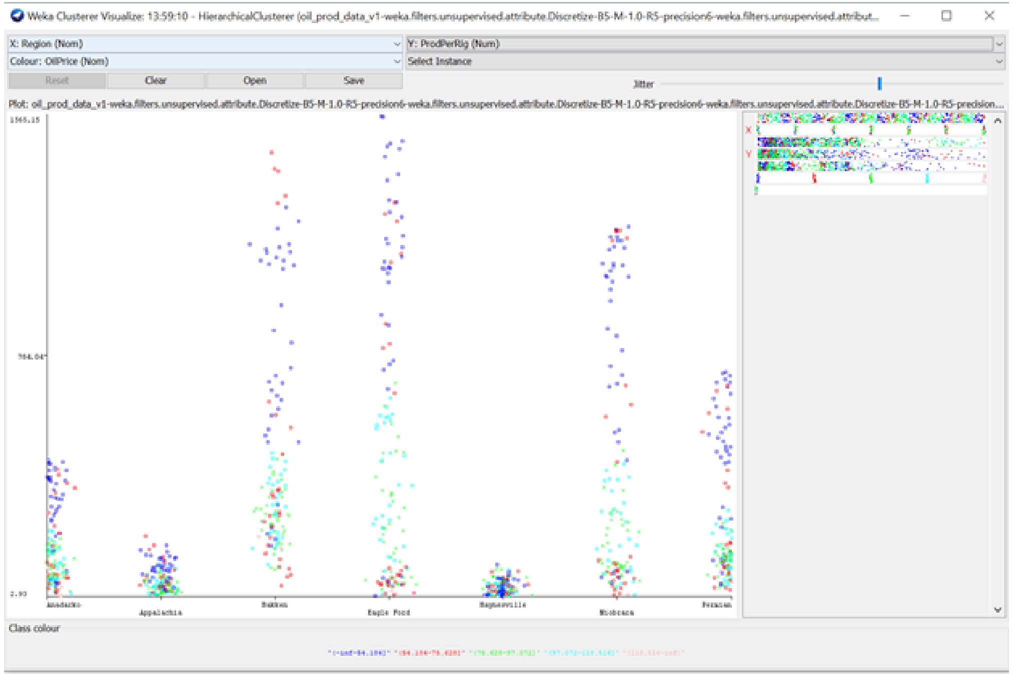
X: Cluster (Nom)

Y: Region (Nom)

Colour: OilPrice (Nom)

Select Instance

Plot: oil_prod_data_v1-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.u...



Weka Clusterer Visualize: 13:59:10 - HierarchicalClusterer (oil_prod_data_v1-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.unsupervised.attribut...

X: Region (Nom)

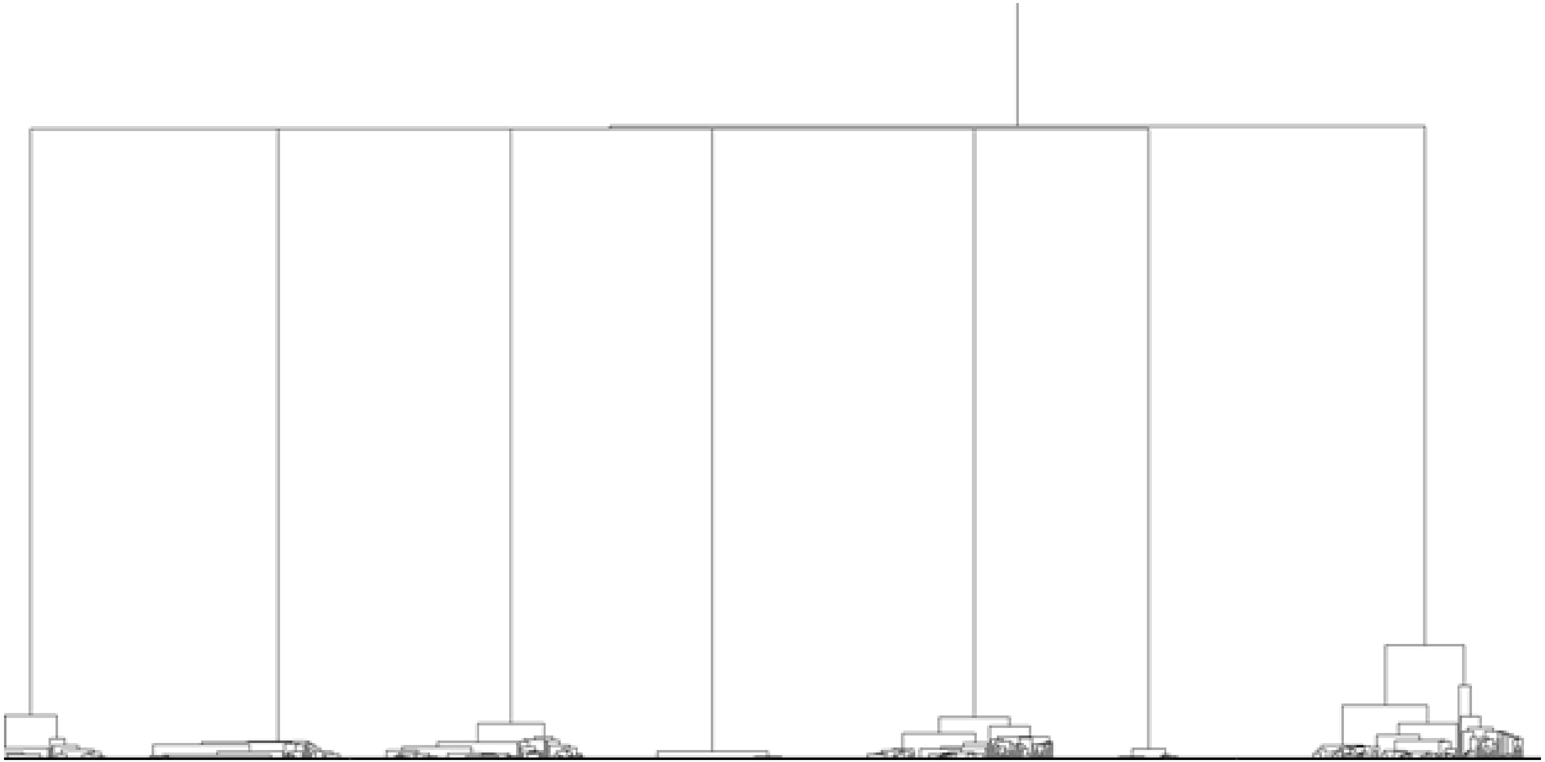
Y: ProdPerRig (Num)

Colour: OilPrice (Nom)

Select Instance

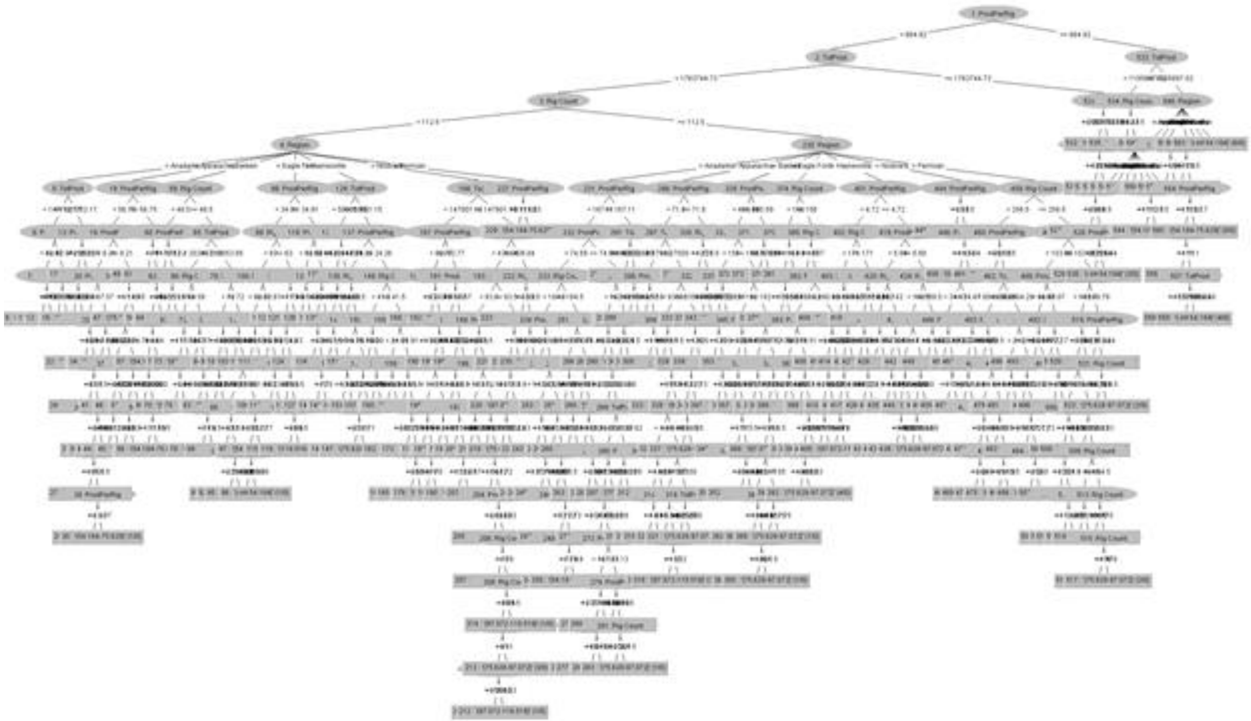
Plot: oil_prod_data_v1-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision6-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-R5-precision...

Hierarchical Clustering on Oil vs Price



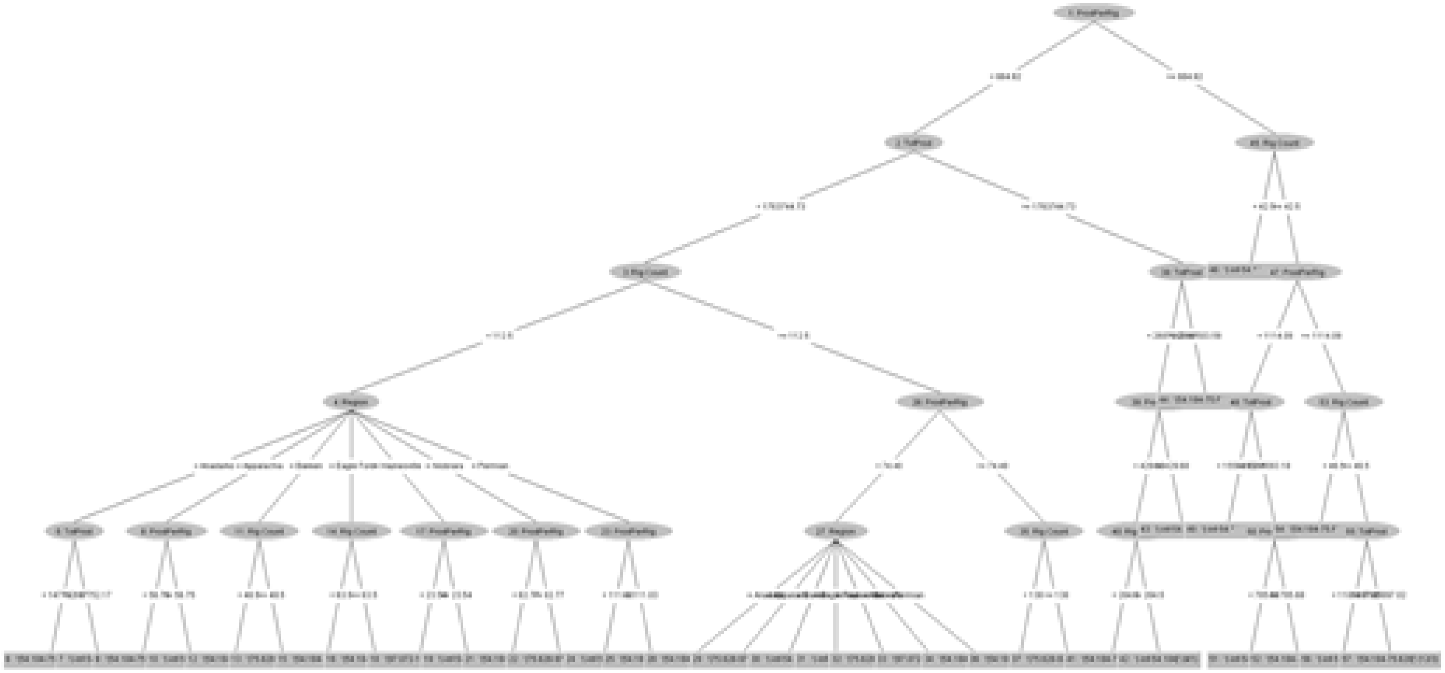
Decision Tree on Oil vs Price

Un-pruned tree

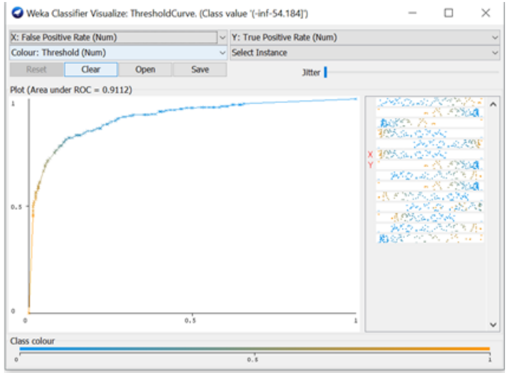


Decision Tree on Oil vs Price

Pruned tree (max depth=5)



Evaluation of Model



Weka Classifier Visualize: ThresholdCurve. (Class value '(-inf-54.184]')

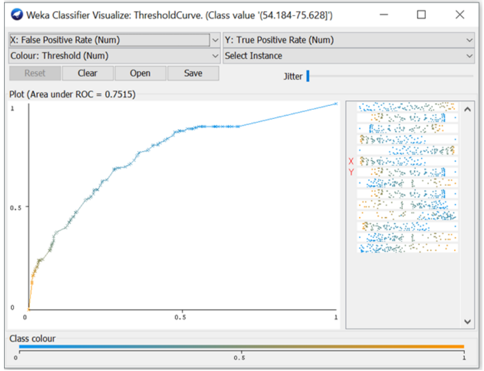
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9112)



Weka Classifier Visualize: ThresholdCurve. (Class value '(54.184-75.628]')

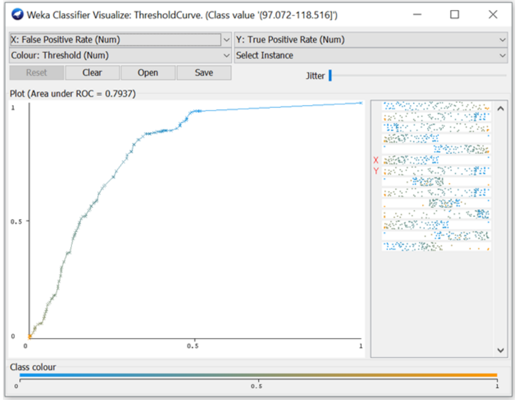
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7515)



Weka Classifier Visualize: ThresholdCurve. (Class value '(97.072-118.516]')

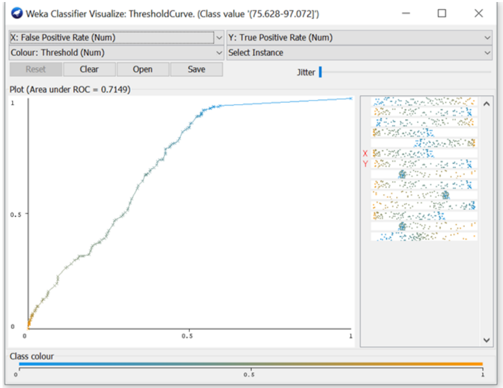
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7937)



Weka Classifier Visualize: ThresholdCurve. (Class value '(75.628-97.072]')

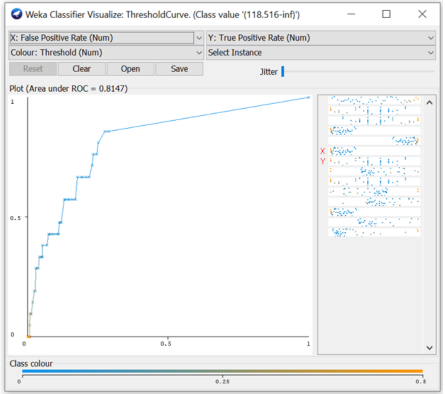
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7149)



Weka Classifier Visualize: ThresholdCurve. (Class value '(118.516-inf)')

X: False Positive Rate (Num)

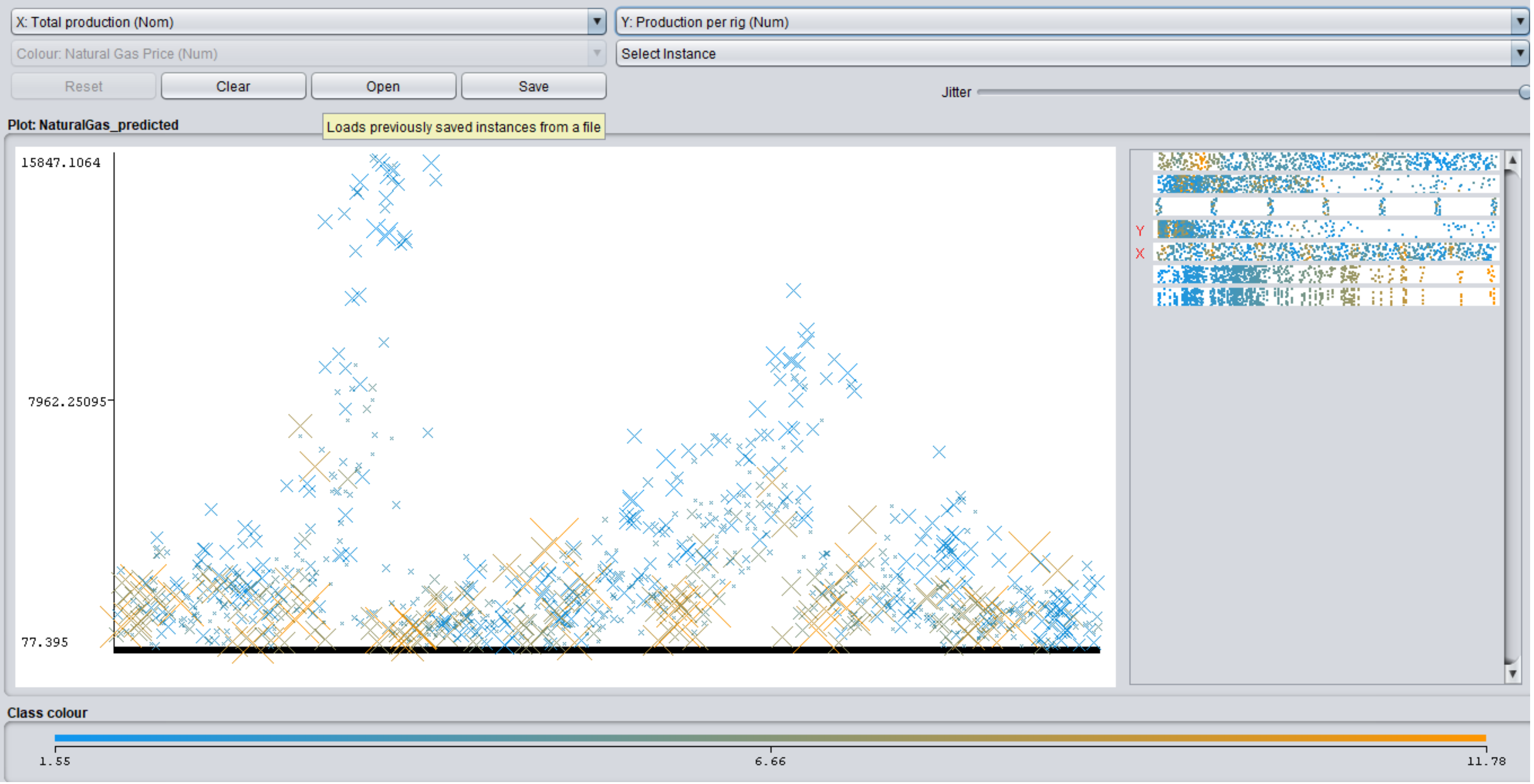
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.8147)

Results



X: Total production (Nom)

Y: Production per rig (Num)

Colour: Natural Gas Price (Num)

Plot: NaturalGas_predicted

Loads previously saved instances from a file

Key Findings

- Rig counts are highly correlated with region
    - Obvious, but data talks!
- Rig counts + Natural Gas Price is a good predictor of production per rig
    - Although not our objective, interesting association
- Low Oil Prices = easier to classifying (model performance)
    - Potentially because of number of data points
- Regions are so different : Should analyse individually
- Production per rig is the best predictor of oil price

- Model building takeaway: pruning is extremely helpful!
- Model building takeaway: limiting features can drive more insights

Further Research Topics

- Impact of Oil Quality on Production

- Global Natural Gas Production vs Prices

- Petroleum Production vs Energy Consumption

Q & A