2.1

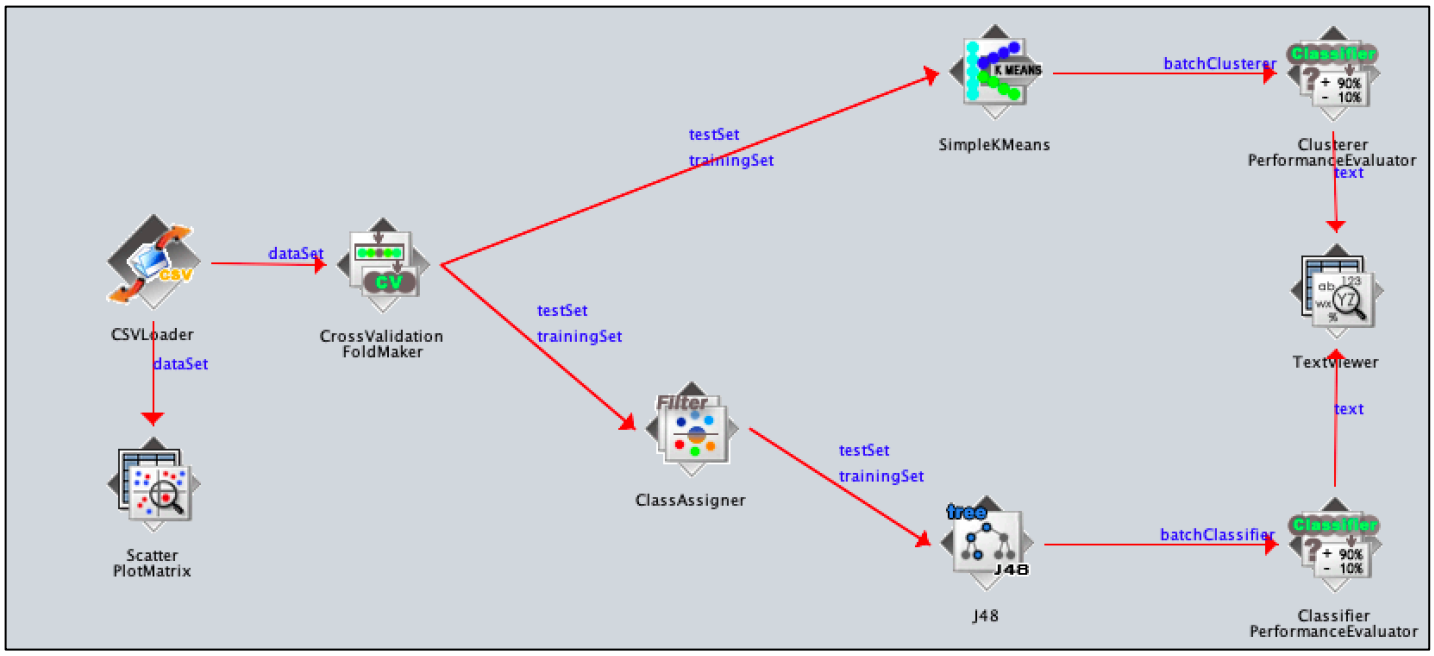
I have selected a data set which gives instances of protected species bycatch in New Zealand Fisheries between 2002 and 2017. It is publicly available on https://psc.dragonfly.co.nz/latest/.

Before I began to analyse the data set I needed to do significant data preparation (preprocessing), namely:

- I removed all apostrophes which caused an I/O error when trying to load the file into WEKA
- I removed all instances of bird/seal as they heavily imbalanced the class and I wish to focus on other species which are affected by trawling such as aquatic mammals and other animals which are primarily sea dwelling. This will also allow the Machine learning techniques to identify key patterns more easily in comparison to a data set with many different classes and instances.
- Because there were so many different types of protected species which were bycatch, I decided to group instances into their animal families to have fewer classes. This would provide better clustering and classification analysis as it will be easier to discern the results with fewer classes. For example, Hectors Dolphin and the Common Dolphin are now simply classified as a Dolphin.

I have chosen to use classification and clustering as the most appropriate techniques to analyse the given dataset. This is because clustering will be a useful tool to analyse the given dataset as it will help us identify natural groupings which are not apparent prima facie. Furthermore classification will help us predict the animal bycatch and consequentially may allow us identify and infer interesting trends in the dataset. I have not chosen regression as an appropriate technique because the class is not a numeric type. Because of this inherent characteristic of the dataset it cannot be effectively analysed by regression. Perhaps after I merge the current data set with a complementary data set regression might be a useful tool.

Pipeline for initial analysis 1.1

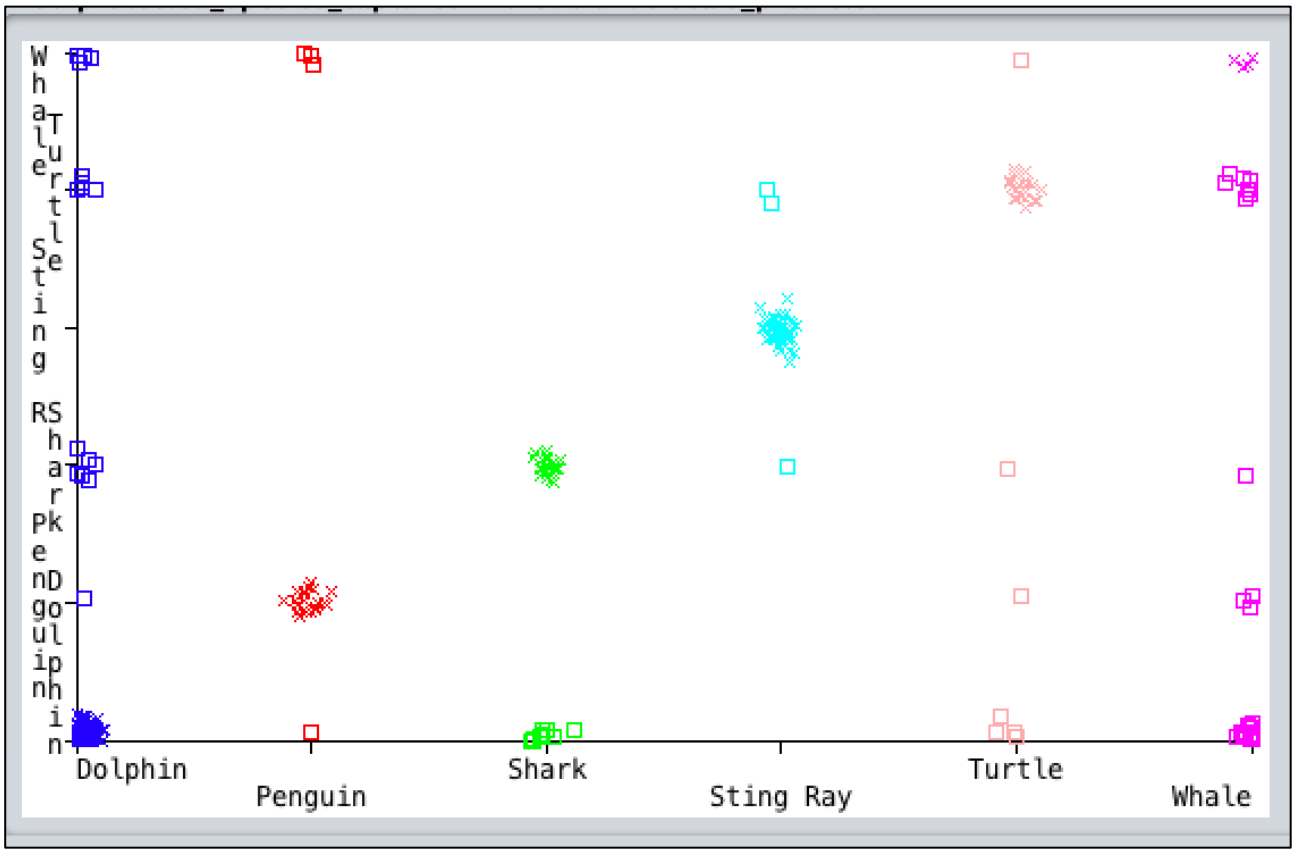


Results for J48 - Classification (without dimensionality reduction)

1.2 General results for classification:

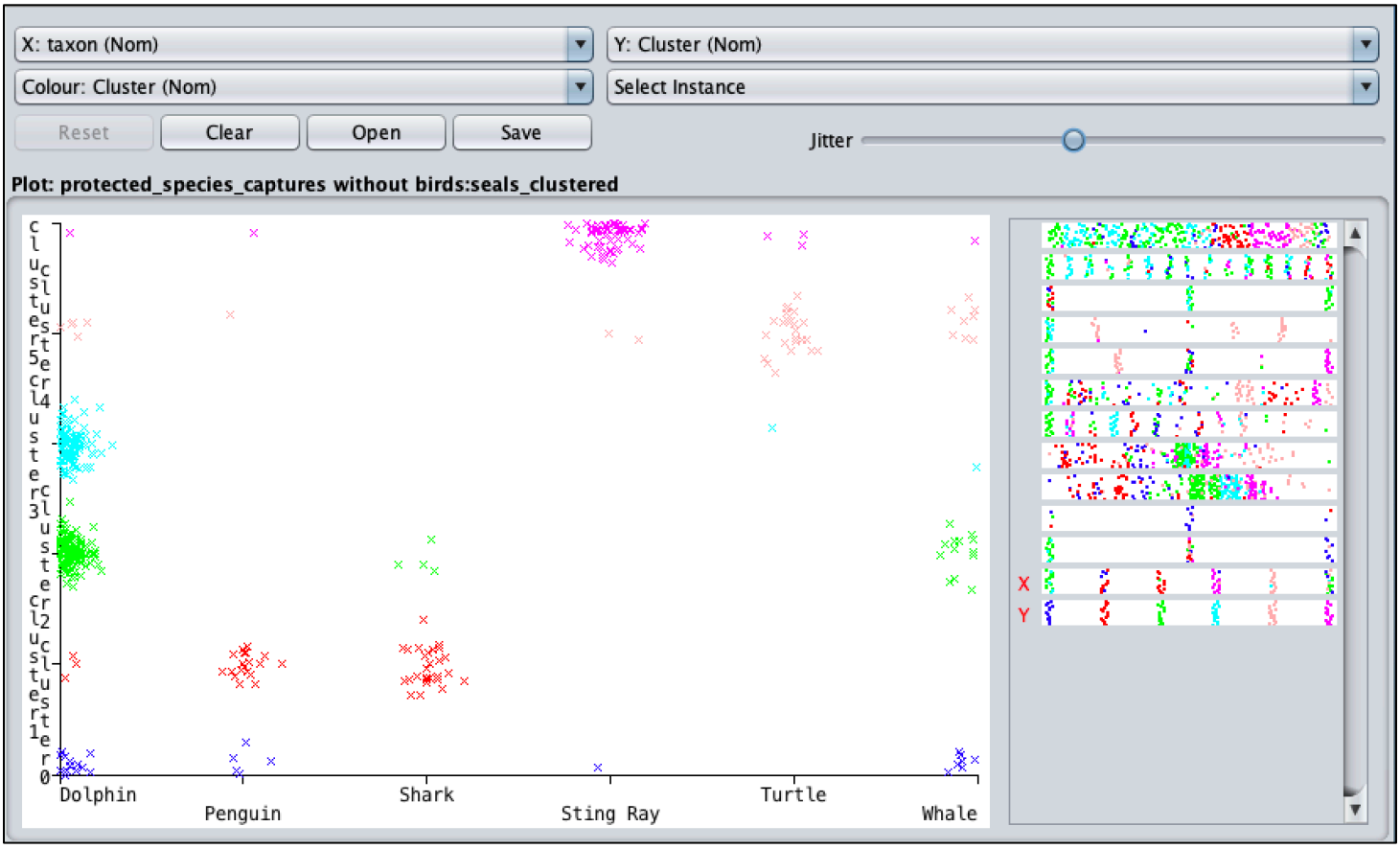
```
Time taken to build model: 0.04 seconds

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

Correctly Classified Instances         370               84.6682 %
Incorrectly Classified Instances        67               15.3318 %
Kappa statistic                          0.7471
Mean absolute error                      0.0657
Root mean squared error                  0.2058
Relative absolute error                 31.6216 %
Root relative squared error             64.0356 %
Total Number of Instances              437

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.933     0.159     0.891      0.933     0.912      0.783     0.900     0.876     Dolphin
               0.846     0.012     0.815      0.846     0.830      0.819     0.939     0.814     Penguin
               0.719     0.025     0.697      0.719     0.708      0.684     0.884     0.657     Shark
               0.953     0.000     1.000      0.953     0.976      0.972     0.975     0.960     Sting Ray
               0.759     0.037     0.595      0.759     0.667      0.645     0.854     0.502     Turtle
               0.129     0.020     0.333      0.129     0.186      0.172     0.685     0.217     Whale
Weighted Avg.  0.847     0.099     0.829      0.847     0.834      0.753     0.894     0.797

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 238   1   7   0   5   4 |   a = Dolphin
   1  22   0   0   0   3 |   b = Penguin
   9   0  23   0   0   0 |   c = Shark
   0   0   1  61   2   0 |   d = Sting Ray
   4   1   1   0  22   1 |   e = Turtle
  15   3   1   0   8   4 |   f = Whale
```

1.3 Visualisation of classifier errors

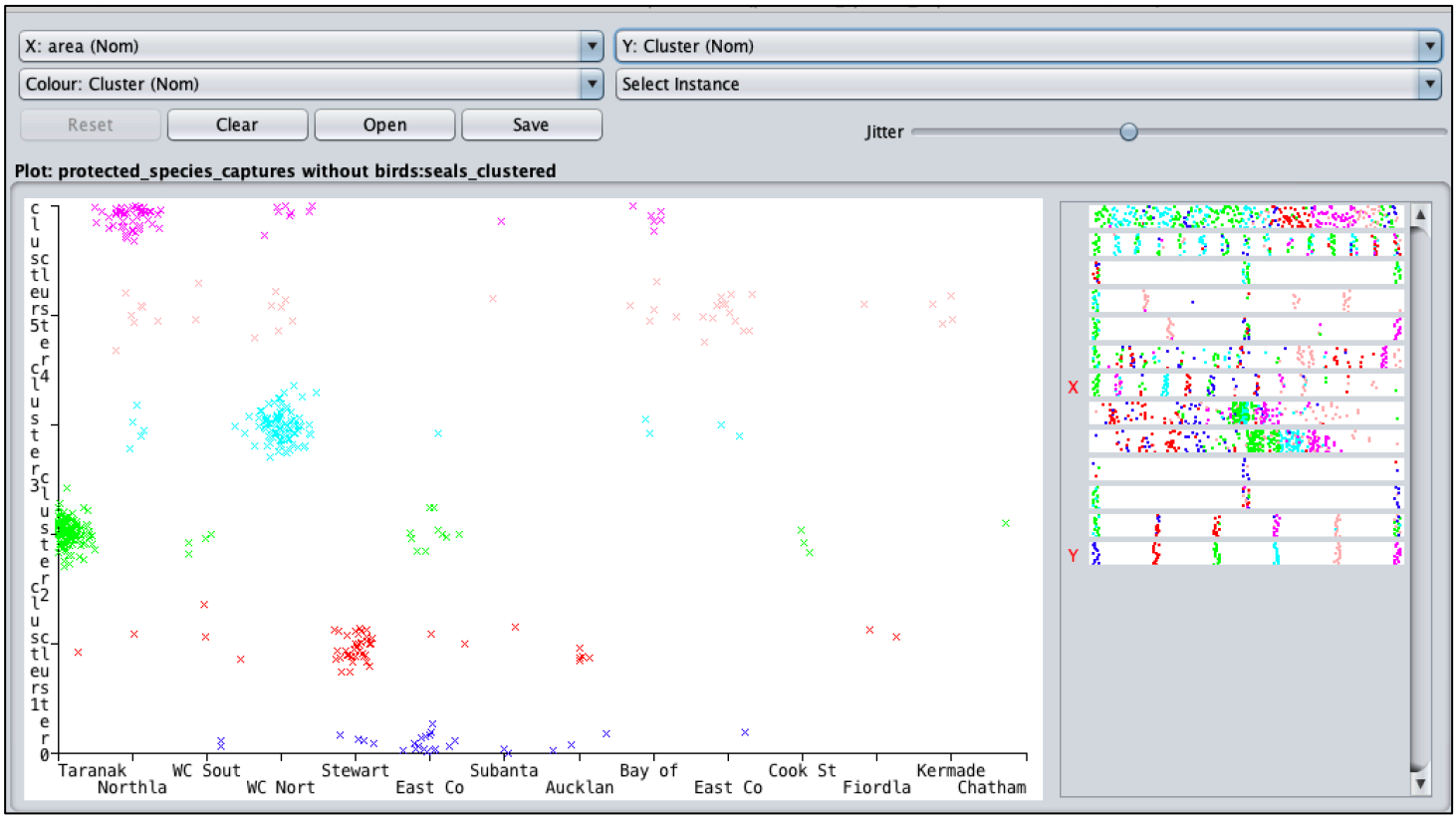


Results for SimpleKMeans – Clustering (K value equal to 5)

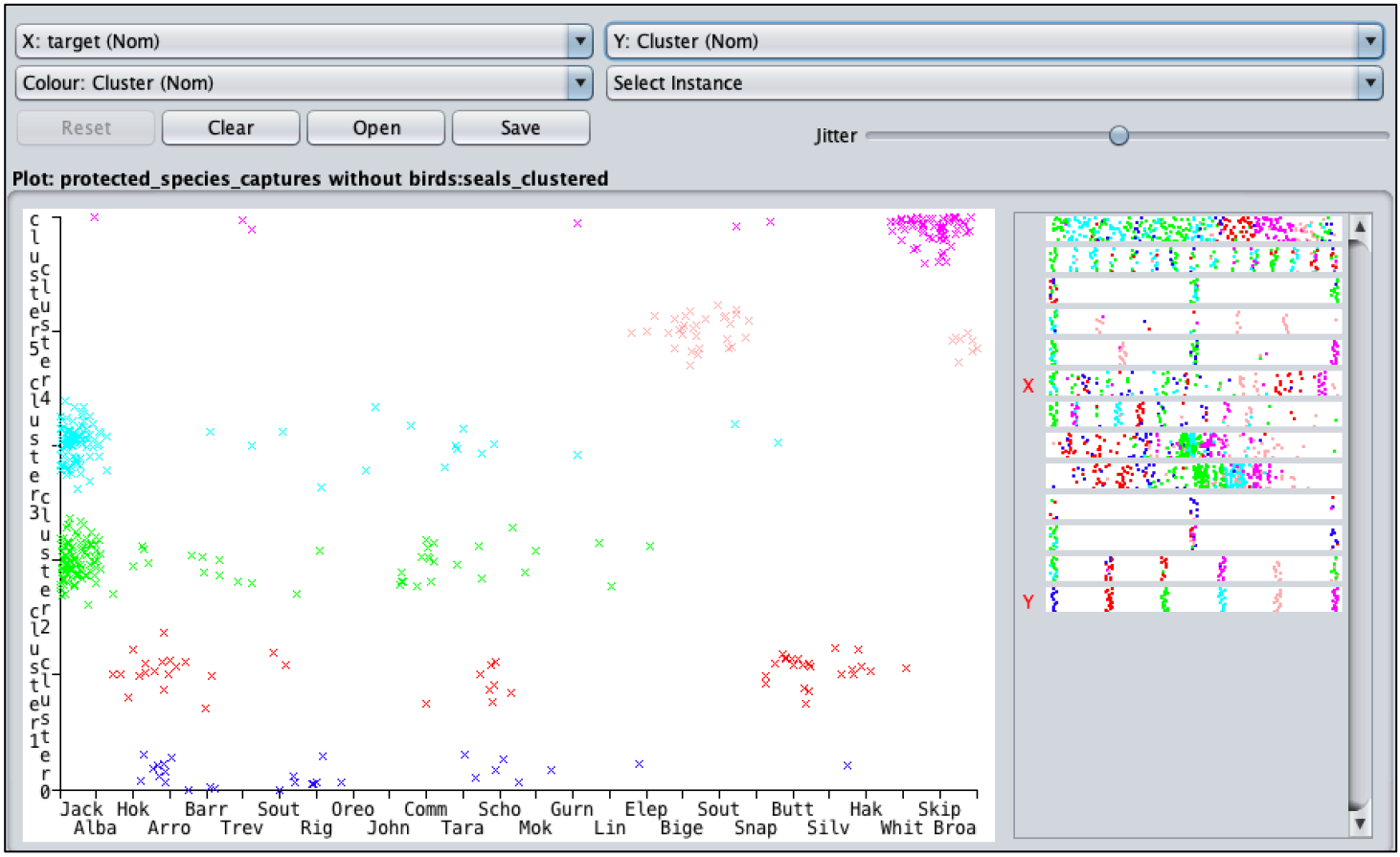
1.4 Clustering of instances in relation to their taxon



1.5 Clustering of instances in relation to the area they were caught as bycatch



1.6 Clustering of instances in relation to the target fish the trawlers were aiming for.



1.6 General results

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 839.7061819147734
```

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

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

Clustered Instances

0       28 (  6%)
1       50 ( 11%)
2      143 ( 33%)
3      108 ( 25%)
4       41 (  9%)
5       67 ( 15%)
```

Discussion of results

Classification – J48

I would have wished to display the J48 classification results through a tree representation. However, because I have done little data manipulation the tree is undiscernible (80 leaves) and therefore provides no value to analyse at the present moment. After we apply some feature selection and other data manipulation techniques it will likely give some value to evaluate.

With regard to 1.2 we can see that the J48 model correctly classifies the taxon **84.7%** of the time. This is quite a good performance percentage, however I would be skeptical as to how it would perform with new data as I expect that overfitting may be at play and is something I will have to consider in the future. Furthermore because the data is unbalanced in terms of class instances I find it useful to look to the **F-measure – 0.834** in 1.2 to evaluate the model because it takes both recall and precision into consideration.

We can look to the confusion matrix in 1.2 and the visualization of classifier errors in 1.3 to understand where the incorrectly classified instances are coming from. We can see that the taxon 'whale' is struggling to be classified correctly and is often being classified as a dolphin (false positive). I would deduce that this is the result of similar attribute values for dolphins and whales, which I would conclude as logical. This is because dolphins and whales share the same habitat in many cases in New Zealand, which is making it hard for the machine learning technique to distinguish them without more data. This is one of the characteristics of the dataset we must take into consideration to analyse it.

By holistically looking at the dataset's classification performance we can see that there is a definite correlation between the given attributes and the taxon. We now need to refine and potentially at some attributes to optimize classification of the taxon.

Clustering - SimpleKMeans

The success of clustering is of usually measured subjectively, so I will look at the results holistically to see whether it gives us any valuable information about the dataset.

With regard to 1.4 we can see that the clusters given by SimpleKMeans are somewhat related to the instances taxon. For example we can see that Sting Rays (purple) are mostly ascribed to one cluster which implies that there is a natural grouping for the class based on the inherent properties of the dataset. This is interesting to note because it implies that Sting Rays are all similar instances in relation to being bycatch. Because of this we can infer that Sting Rays as bycatch must have similar attributes. This inference is backed up by both 1.5 and 1.6 where we can that the cluster ascribed to String Rays is also strongly clustered in relation to attributes **target fish** and **area they were caught.** We can also say this of many of the other taxon.

In 1.4 with regard to the instances of the dolphin taxon we can see that there is two distinct groupings (blue and green). This interestingly allows us in infer that there is two natural groupings of dolphins which are caught as bycatch. So what is the difference between these two natural groupings is it because the dolphins are different species or live in different areas. If we have regard to 1.5 it leads us to the point of difference between these two clusters – they are dolphins caught as bycatch in different areas. So we can see that there are two natural groupings within the same class, which likely is one of the reasons why the classification technique J48 may have struggled to correctly classify some dolphin.

We can also see that the cluster sum of squared errors is relatively high in 1.6, which in turn may make our deduction above less valuable. However on the whole the SimpleKMeans clustering technique has provided us some useful insights into the natural groupings of the data set and some characteristics which the classification technique was unable to pick up on.

How Classification and Clustering are different

The primary difference between classification and clustering is that they correspond to supervised learning and unsupervised learning respectively. A supervised learning algorithm like J48 takes a known set of input data in the form of attributes and their corresponding values alongside the known outputs for each instance. In relation to this particular dataset the attributes such as area, longitude etc. act as the input data and the known responses to the data are the 'taxon'. Together through classification they train a model to predict categorical responses. In the present case the classification method is using the attributes and known responses to predict the taxon of the particular instance.

Because the classification model is able to see the known responses in relation to the prediction made by the model we get results such as false positives and true positives. This is unlike any clustering algorithm as they have no regard to the known responses of any given input(s).

The clustering algorithm SimpleKMeans is an example of unsupervised learning. Unlike a supervised learning technique, unsupervised learning finds natural patterns or groupings from only input data (no classes). With regard to the current data set all attributes including the 'taxon' which is considered a known response in supervised learning are inputs. This allows us to draw inferences from the data sets which might be less apparent through classification. An example of this would be that through clustering we have been able to infer that there are two distinct groups of dolphins. Classification disallows us to find similar inferences because it has labelled responses (class – taxon).

Questions

1. Is there any evidence of fish stocks collapsing in NZ waters?

2. Is it likely that any protected species which is caught as bycatch from trawling will die as a result of being bycatch?

3. Is there any evidence to suggest that trawling and its various catch methods are threatening any New Zealand protected species such as dolphins and whales?

I believe the two questions I have created above essentially evaluating the performance of the trawling industries from the perspective of New Zealand protected species. More specifically question 2 is asking whether trawling is a inhumane and redundant practice. This may give rise to New Zealand fisheries considering that its methods should be updated in order to mitigate deaths of bycatch. Whereas question 3 takes a look at the bigger picture of trawling in New Zealand in relation to New Zealand's protected species. Because of this it will shed light on whether any particular or type of trawl should be allowed to continue if it can be shown that it is detrimental to any protected species.

2.2

Selection and explanation as to why the question is interesting

3. Is there any evidence to suggest that trawling and its various catch methods are threatening any New Zealand protected species such as dolphins and whales?

I have chosen to address **question 3** as I believe it has many different avenues in relation to the data set which can be explored and this consequentially makes it very interesting. The question allows us to explore how any one of the attributes affects any particular species. For example does any target fish or area of trawlers disproportionately affect any given species. Through this we will be able to conclude (hopefully) which practices of trawlers are detrimental or not to a species as a whole. Furthermore because we have information as to where the trawling took place we may be able to find conclusions for the effects of trawling for a particular sub populations of any given species. Because of all the aspects of the dataset we can explore in relation to this question I ultimately find that this question is very interesting.

Why was the data collected and what did it hope to achieve (Business aspects)

The dataset regarding protected species bycatch in New Zealand fisheries was likely collected in order to monitor the impacts of trawl fisheries on protected species. It was collected by the department of conservation's observers on fishing vessels for the Ministry of primary industries. The department of conservation may have collected this data in order to help the government shape new fishing quotas with regard to how much trawling can occur. Perhaps the department of conservation may also use this information to create fishing exclusion zones (marine reserves) when necessary, if they notice that many protected species are dying as a result of over fishing at any particular time. Essentially the department of conservation would have created this dataset to ensure all protected species population are sustainable concurrently with trawl fisheries in New Zealand.

Why the chosen datasets(s) are appropriate given the business understanding

I have chosen **fishing effort (number of trawls) by year (1990-2014)**[1] and the value of **seafood exports by year (2007-2014)**[2] as a complementary data set to the original dataset. The dataset can be found publicly and have been collected for the Ministry for the Environment.

The dataset of seafood exports by year was collected in order to report on the value of the industry this in turn would help the government understand the marine economy. This is an appropriate complementary dataset because it will add a monetary dimension to the dataset in the sense that the amount of bycatch or type of bycatch may be influenced by how valuable the industry is and how much money there is too be made. No fisherman wishes to catch any protected species, however it is unavoidable and therefore it might be something they would be less inclined to mitigate if there is more money to be made.

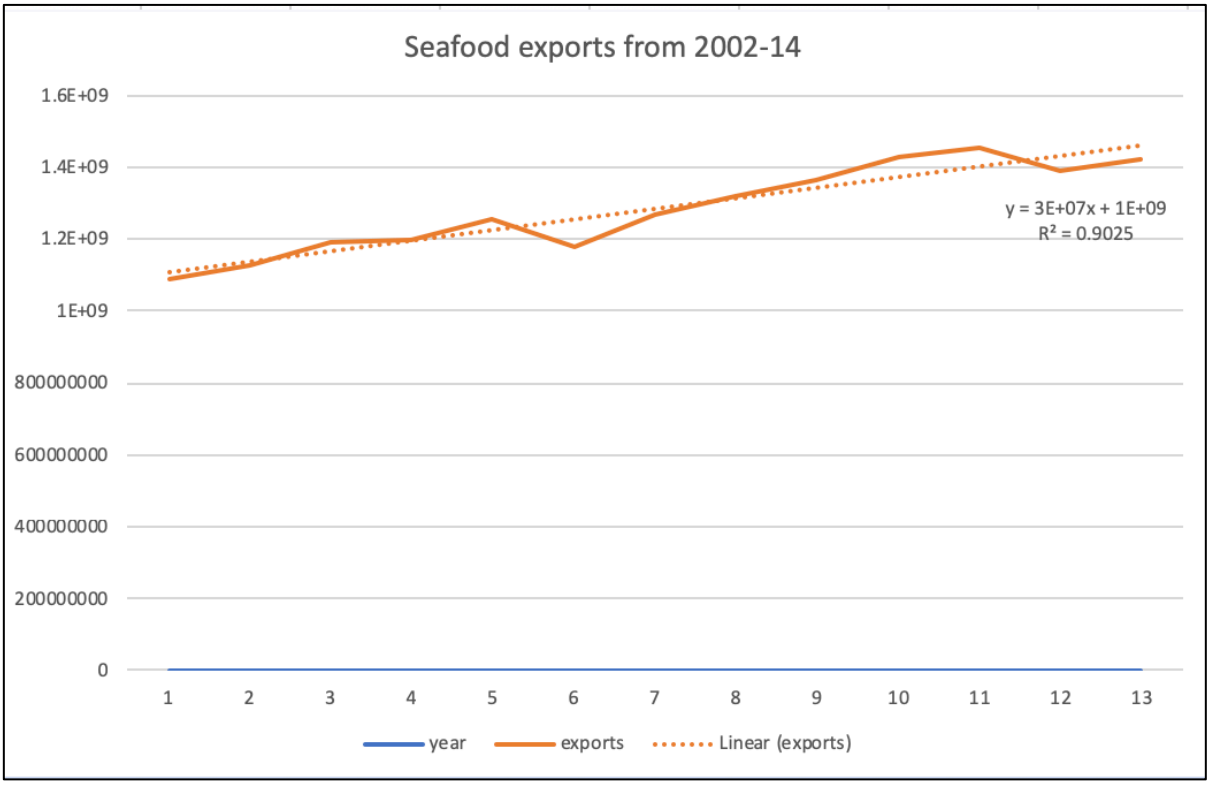
The dataset for fishing efforts (number of trawls) by year was collected because towing a fishing net along the bottom of the ocean floor can physically damage ocean habitats and species. Because of this it is important for the government to keep track of how much trawling is actually occurring in New Zealand waters. It consequentially would allow the government to pull back on trawling when necessary if they believe too much is occurring for whatever reason. This is an appropriate complementary dataset as it may give us insight as to why some of the less usual protected species bycatch are apparent in any given year where there is more trawling.

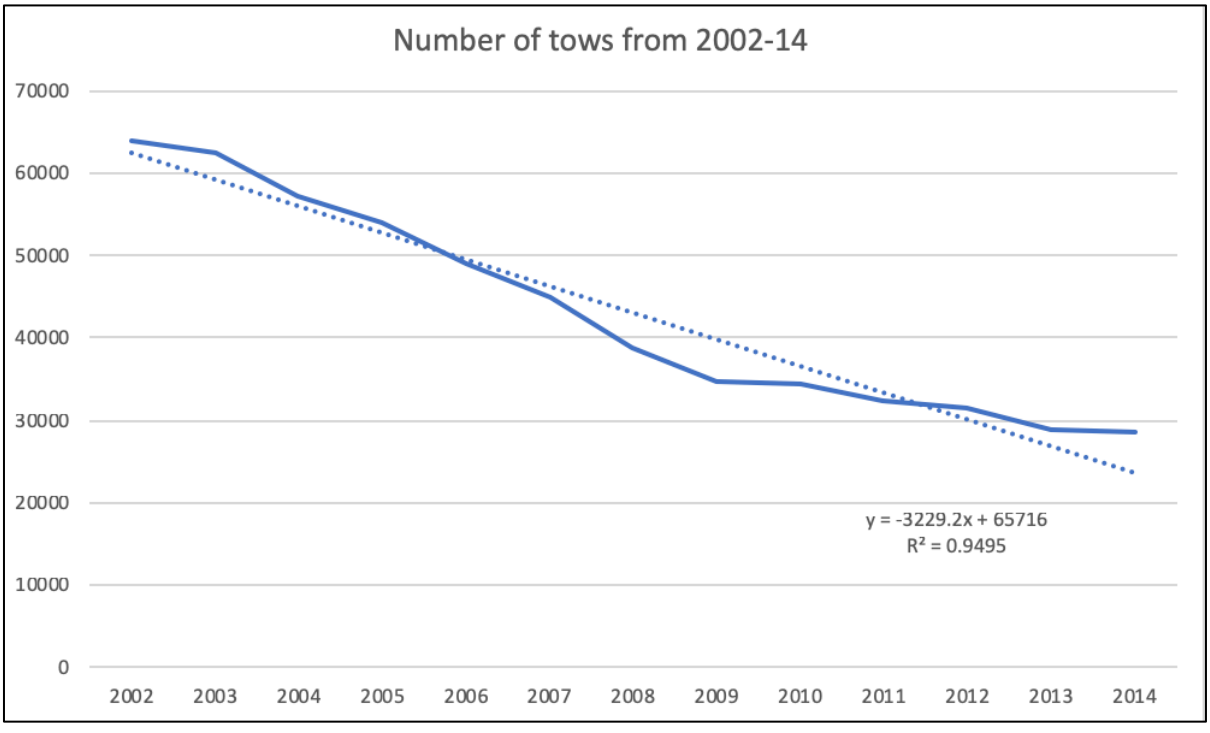
[1] https://data.mfe.govt.nz/table/52504-fishing-effort-number-of-trawl-tows-by-year-19902014/
[2] https://data.mfe.govt.nz/table/52526-seafood-export-values-200714/

Merging the data sets

Imputation

Because each of the complementary data sets have different date ranges to the original data set we are going to have to use imputation to find values which would provide us with use for the data set. Because both datasets give values (number of tows and export values) over time I decided to plot each respective dataset in a chart to see if I can predict future values using linear regression. Below are the respective charts created to help impute future values.





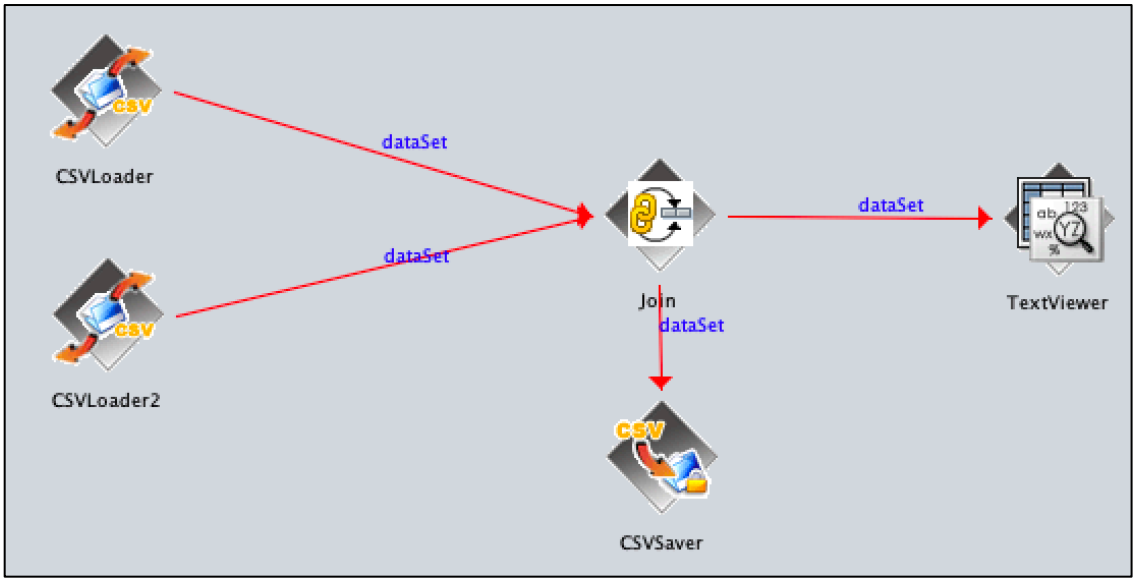
From the charts and information consequentially derived, I have great confidence in imputing the future values for seafood exports value and the number of trawls. This is because of the strong R squared values (0.9025 and 0.9495 respectively), which indicates that the trend line is very strongly fitted for the plotted points. I then used the formula given by the trend line (y = mx+c) to impute uncollected values from 2015-2017 for each complementary dataset.

Further imputation is necessary as there is only as many instances as there is years in the complementary datasets. If I were to merge the datasets right now it would cause errors as the number of instances do not align. Because of this I am going to have to align the values given by the complementary dataset and instances of bycatch by their common attribute value 'fishing year'. To line these up I ordered the protected species bycatch csv file by their year and then saw how many of each years instance in the complementary datasets is needed and dealt with it correspondingly. Each dataset now will have 438 instances which will correspond to a particular fishing year.

Merging

I merged the three datasets using a KnowledgeFlow pipeline on Weka displayed in 2.1.

2.1 KnowledgeFlow pipeline for merging two datasets



While merging I encountered a "invalid key attribute name" which simply required me to remove any underscores of full stops from the attribute list. After merging was complete I ended up with the merged dataset displayed below in 2.2.

2.2. dataset after merging protected species bycatch with seafood export value/trawling effort in a given year.

year	identificatior	capturemethod	method	target	area	latitude	longitude	excluded	status	taxon	year2	exports	year3	NumberTows
2002/2003	necropsy	net	trawl	'Jack mackerel'	Taranaki	173.241059	-39.770584	?	dead	Dolphin	2002/2003	1087227089	2002/2003	64011
2002/2003	necropsy	net	trawl	'Jack mackerel'	Taranaki	174.258571	-40.56399	?	dead	Dolphin	2002/2003	1087227089	2002/2003	64011
2002/2003	necropsy	net	trawl	'Jack mackerel'	Taranaki	174.26656	-40.53901	?	dead	Dolphin	2002/2003	1087227089	2002/2003	64011
2002/2003	observer	net	trawl	'Jack mackerel'	Taranaki	174.243453	-40.534747	?	dead	Dolphin	2002/2003	1087227089	2002/2003	64011
2002/2003	observer	net	trawl	'Jack mackerel'	Taranaki	174.091825	-40.375022	?	dead	Dolphin	2002/2003	1087227089	2002/2003	64011
2002/2003	observer	net	trawl	'Jack mackerel'	Taranaki	174.094808	-40.394272	?	dead	Dolphin	2002/2003	1087227089	2002/2003	64011

Dimensionality reduction techniques

General

Firstly, after merging the three datasets there are now three 'year' attributes. I have decided to delete two of them because they may give unexpected consequences if not removed or weaken the validity of my results. There are a few unnecessary instances (outliers or redundant instances), so I do have to remove many instances which may skew the results. The instances I do have to remove relate to the 'target fish' values as there are some instances which only show the respective value once or twice. For example I removed the instance(s) where the target fish are 'Oreos', 'Albacore Tuna', 'Southern Blue Whiting', 'Gurnard', 'Elephant fish', 'white Warehou' because they only give one instance each. I also had to remove outliers relating to the 'area' feature as there was only one instance of trawling bycatch in 'Chatham rise', 'subantarctic' and 'Fiordland'. I also removed one of the trawling methods value 'bottom long line' as it only had two instances.

Missing data

The attribute 'capturemethod' portrayed many missing values throughout the dataset. Because of this I decided I would replace the missing values with the most common value displayed, which was "net" by a significant amount. There is another attribute named 'excluded' with many missing values which has been given no description in the information for the dataset. It has so many missing values that I have decided to leave the missing values as I believe that it would make any results I deduce via classification/clustering/regression potentially void if I were to assign any values.

Redundant features to selected tools performance

To gain information as to which features are redundant, I used the Attribute evaluator Ranker method on Weka. This will give me information as to which features actually help predict the known responses or classes (taxon) of the new dataset. Below in 2.3 are the results from the Ranker method.

2.3 Ranker results for the features of the merged dataset

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 taxon):
        Information Gain Ranking Filter

Ranked attributes:
 1.31607    5 target
 1.000538   8 longitude
 0.960817   4 method
 0.887385   6 area
 0.855908   7 latitude
 0.490052   1 year
 0.48094   10 status
 0.355479  11 exports
 0.290202  12 NumberTows
 0.28589    2 identification
 0.274789   3 capturemethod
 0.003275   9 excluded

Selected attributes: 5,8,4,6,7,1,10,11,12,2,3,9 : 12
```

Because of the information results given in 2.3 I have decided to remove the following features:

- Identification
- Capture method
- Excluded

This is because I believe they are redundant in relation to classifying the correct taxon and will unlikely give any insightful results in relation to clustering. Furthermore they cannot be used to give any useful deduction prima facie in relation to protected species bycatch at the hands of trawling.

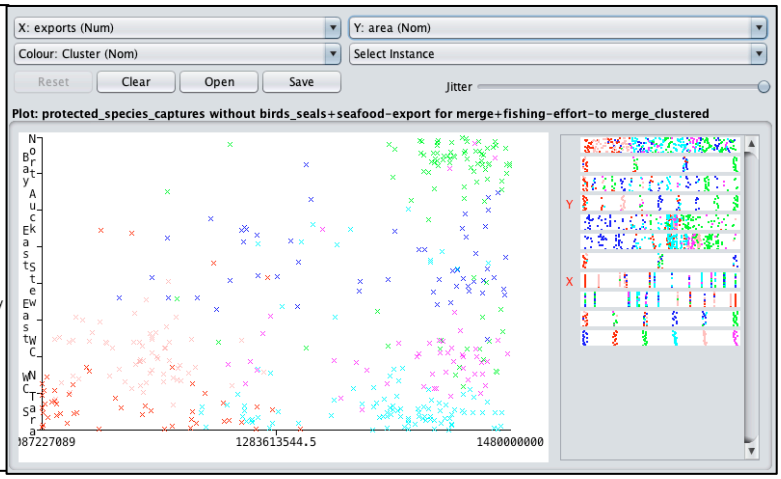
Interestingly the ranker method has labelled the year in which the bycatch was caught as useful for classifying its taxon. However, it must be noted that the reason that the bycatch was caught by trawlers is not because of the year itself, but instead what the trawl fisheries are doing in that particular year. For example dolphin's may be caught more as a result of the government giving quotas for certain areas which haven't been fished recently. For this reason I have decided to remove the 'year' feature as in its current capacity does not give any value or insight to why more protected species are being caught as bycatch.

After all this Dimensionality reduction I ran the J48 and SimpleKMeans pipeline found at 1.1 and actually found that classification accuracy increased to **87.2236%** and the clusters were **much more defined**. 2.6 displays what the final dataset looks like after dimensionality reduction.

2.4 Results after dimensionality reduction for J48.

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

Correctly Classified Instances         355               87.2236 %
Incorrectly Classified Instances        52               12.7764 %
Kappa statistic                          0.7926
Mean absolute error                      0.0536
Root mean squared error                  0.1944
Relative absolute error                 26.2183 %
Root relative squared error             60.9922 %
Total Number of Instances              407

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.913     0.096     0.932      0.913     0.922      0.813     0.925     0.927     Dolphin
               0.750     0.031     0.600      0.750     0.667      0.648     0.886     0.436     Turtle
               0.500     0.026     0.583      0.500     0.538      0.509     0.849     0.446     Whale
               0.864     0.008     0.864      0.864     0.864      0.856     0.921     0.673     Penguin
               0.793     0.029     0.676      0.793     0.730      0.710     0.871     0.477     Shark
               0.968     0.000     1.000      0.968     0.984      0.981     0.983     0.973     Sting Ray
Weighted Avg.  0.872     0.063     0.877      0.872     0.874      0.803     0.923     0.826

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 220   6   7   2   6   0 |   a = Dolphin
   3  18   2   0   1   0 |   b = Turtle
   7   4  14   0   3   0 |   c = Whale
   2   0   0  19   1   0 |   d = Penguin
   4   0   1   1  23   0 |   e = Shark
   0   2   0   0   0  61 |   f = Sting Ray
```

2.5 Results after dimensionality reduction for SKM



2.6 Final dataset after dimensionality reduction techniques

method	target	area	latitude	longitude	status	exports	NumberTow	taxon
trawl	'Jack macker	Taranaki	173.241059	-39.770584	dead	1087227089	64011	Dolphin
trawl	'Jack macker	Taranaki	174.258571	-40.56399	dead	1087227089	64011	Dolphin
trawl	'Jack macker	Taranaki	174.26656	-40.53901	dead	1087227089	64011	Dolphin
trawl	'Jack macker	Taranaki	174.243453	-40.534747	dead	1087227089	64011	Dolphin
trawl	'Jack macker	Taranaki	174.091825	-40.375022	dead	1087227089	64011	Dolphin
trawl	'Jack macker	Taranaki	174.094808	-40.394272	dead	1087227089	64011	Dolphin
trawl	'Jack macker	Taranaki	174.113414	-40.410567	dead	1087227089	64011	Dolphin

Analysis for what features are important for the selected tools output

Classification – J48

Now that the dataset is process the tree representation (2.7) given is discernable and useful to us to evaluate which features the model thinks are the most important for classifying the taxon.

2.7 Tree representation for J48

```
J48 pruned tree
------------------

method = trawl
|   latitude <= 172.844011
|   |   exports <= 1269510569: Dolphin (15.0/5.0)
|   |   exports > 1269510569: Shark (29.0/7.0)
|   latitude > 172.844011
|   |   status = dead
|   |   |   longitude <= -39.371666: Dolphin (90.0)
|   |   |   longitude > -39.371666
|   |   |   |   longitude <= -38.444064
|   |   |   |   |   exports <= 1269510569: Whale (6.0)
|   |   |   |   |   exports > 1269510569
|   |   |   |   |   |   longitude <= -38.516764: Dolphin (6.0)
|   |   |   |   |   |   longitude > -38.516764: Whale (5.0)
|   |   |   |   longitude > -38.444064: Dolphin (110.0)
|   |   status = alive: Dolphin (3.0/2.0)
|   |   status = decomposed: Whale (2.0)
method = surface longline
|   exports <= 1180165363: Whale (6.0/1.0)
|   exports > 1180165363: Turtle (29.0/9.0)
method = setnet
|   longitude <= -44.781955: Penguin (24.0/3.0)
|   longitude > -44.781955
|   |   longitude <= -39.078412: Dolphin (13.0)
|   |   longitude > -39.078412: Shark (8.0/3.0)
method = purse seine: Sting Ray (61.0)

Number of Leaves  :     15

Size of the tree :      26
```

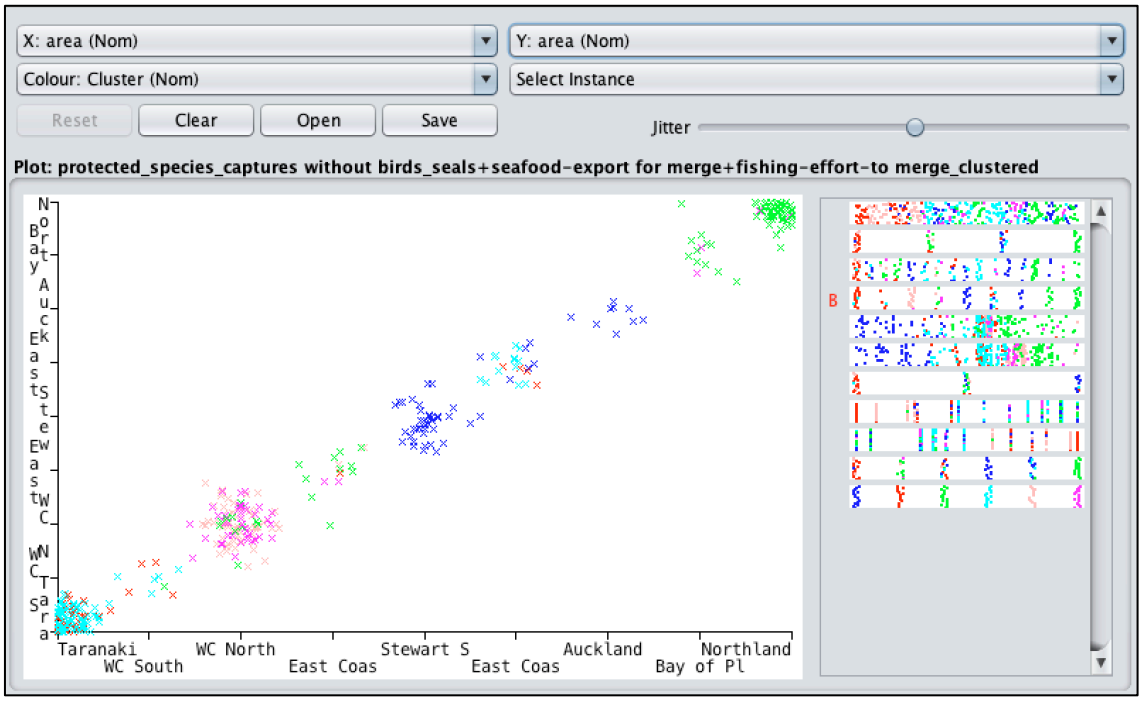
From this tree we can deduce that the method is a very important feature for the techniques output as it is the first question the model asks the instance it is trying to predict. Interestingly latitude is quite an important feature to the prediction model. This is almost certainly because of the machine learning technique picking up on the pattern that some animals are only found above or below certain points. This is indicative of the fact that some protected species like the colder water as opposed to the warmer water or visa versa. A similar point could be made about the longitude feature. This is because some animals may not like the currents which occur on the east coast as a result of being on the edge of the pacific rim and consequentially only stick to the west coast. These deductions above all lead to why some protected species or groups may be more likely to be caught as bycatch if they share the same wants as the fish species the trawlers are targeting.

Interestingly the J48 model has left of the ‘area’ and ‘target’ features from the tree representation. I believe this is because the many different areas and target species cannot be used to help advance the divide and conquer methodology used by J48 as their values are represented by strings as opposed to numbers.

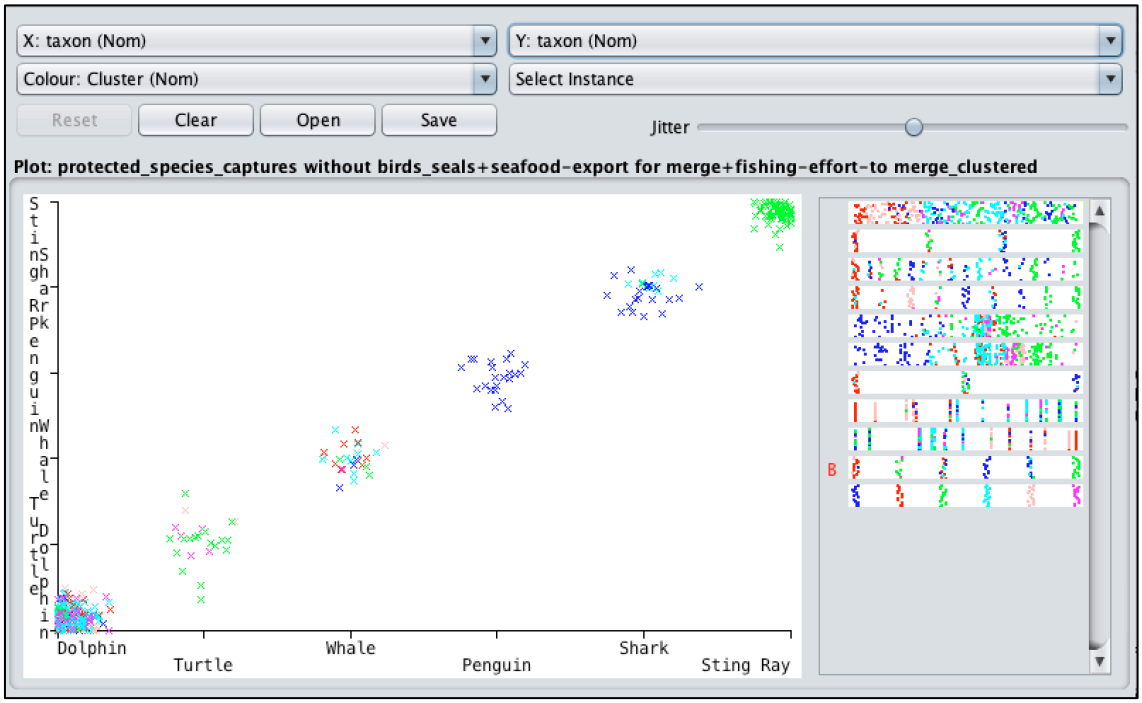
Clustering - SimpleKMeans

Interestingly the clustering algorithm thinks that the 'area' in which the bycatch is caught is a strong feature for finding the underlying natural groupings in the data. This is opposed to the classification technique J48 and might be because of the their differences as supervised and unsupervised models. Below in 2.8 is how we can see that the clustering algorithm gives a lot of weight to the 'area' feature as each different 'area' largely constitutes another respective cluster. We can compare this with 2.9 to see how the clustering algorithm SimpleKMeans values the feature 'area' as opposed to others.

2.8 Clusters represented by the individual colours in relation to the feature 'area'

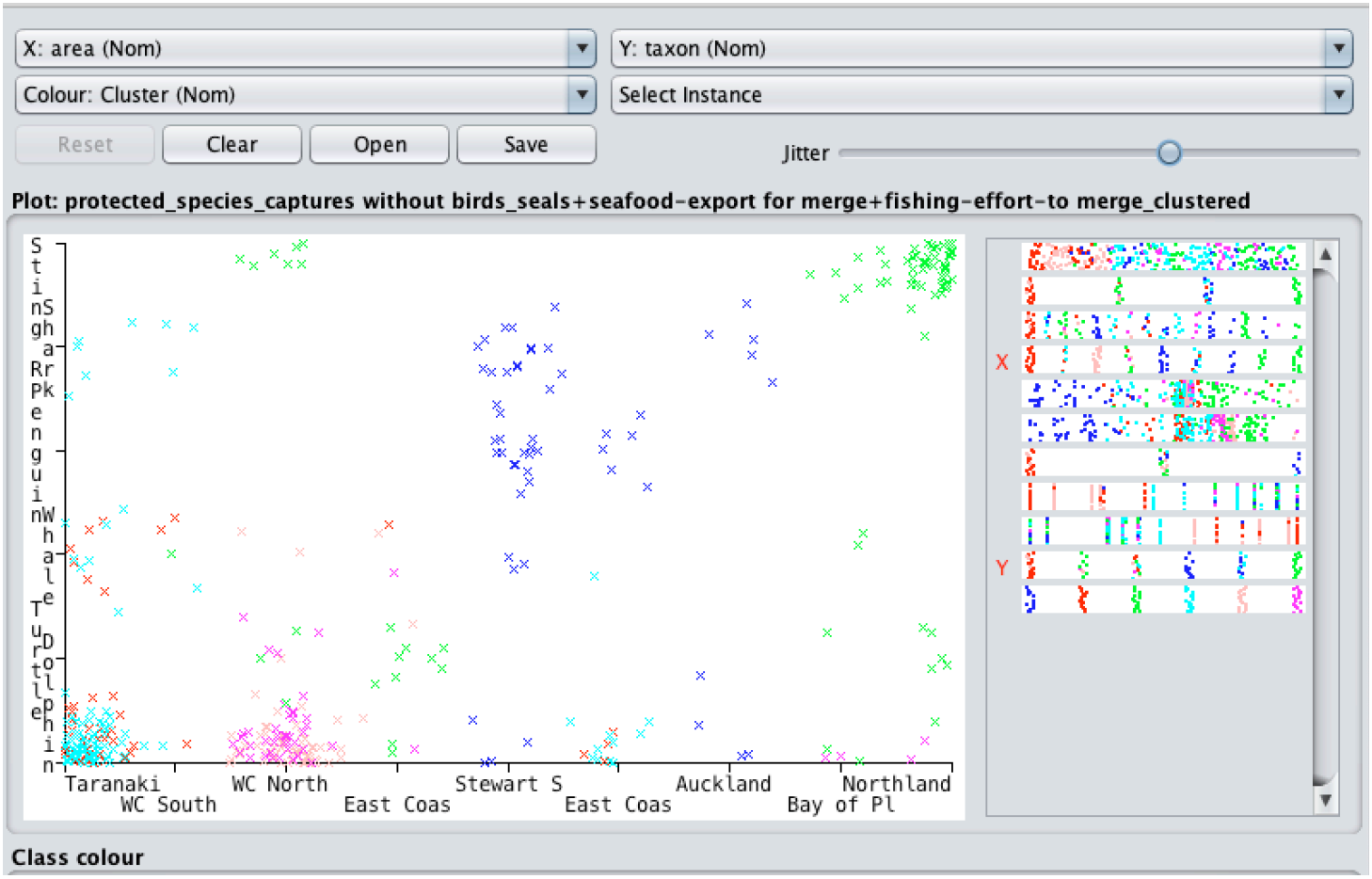


2.9 Clusters represented by the individual colours in relation to the feature 'taxon'



Now the question to ask is why does the SimpleKMeans model value the feature ‘area’? I would infer that it must be because there is a strong relationship with where the bycatch was caught and the other features values. For example in 2.10 it shows that some areas of New Zealand disproportionately affects some protected species as a likely result of there being more trawling going on in the area. Furthermore the trawlers may be targeting fish which are the food resource for the protected species in the area, so of course there will be more of them as bycatch.

2.10 Area in relation to the type of protected species which is bycatch



From this clustering assignment from SimpleKMeans we can see that trawling in Taranaki and the west coast of the North Island are affecting two separate natural groupings of dolphins identified by SKM (Pink and light blue) disproportionately to other species. This may be something to worry about if trawling in those region increases in the coming years. Dolphins as bycatch of trawling on the whole may not affect the entire species of dolphins, but may affect some sub populations in Taranaki and the West Coast of the North Island. Furthermore, we have to have regard to the nature of interrelationships between dolphins. Dolphin are affected similarly to humans when a loved one passes, so perhaps the loss of these dolphins has caused serious harm within some sub populations in the given areas where trawling bycatch is most prevalent.

2.3 – refer to the dolphin.pdf in the submission.