Machine Learning Tools and Techniques
Assignment 2: Real-World Data Handling, Modelling and Visualisation

Name: Karu Skipper
ID: 300417869

2.1 Part 1: Core: Evidence related to fish stocks in New Zealand [40 marks]

“Whatever happens, over-fishing today will lead to collapses tomorrow, ... ” Jun 14 2019 Stuff

- Firstly, what evidence can be found to support (or dismiss) the newspaper headlines?
- Secondly, what features can be considered as the underlying cause for any increase (or decrease) in fish numbers (stocks)?
- Thirdly, what trends and predictions can be made from the data?
- Finally, what findings would you communicate, based on your analysis [also consider the consequences of publicising the information found]?

Overview of all datasets I have investigated before selecting the highlighted dataset for analysis

Data sets name	Links to the datasets
Number of trawl tows	Link
Estimated fish and invertebrate bycatch	Link
Marine Protected Areas	Link
Sea Temperature	Link

a)

Dataset : *Estimated fish and invertebrate bycatch*
Note: *Bycatch is the amount of unwanted fish caught per tonne*
Tool: *WEKA*
Description: *The unintended catch of marine species other than the target species puts pressure on the populations of marine species by removing individuals or potentially modifying ecosystems.*
Coverage: *1991–2012, New Zealand waters*

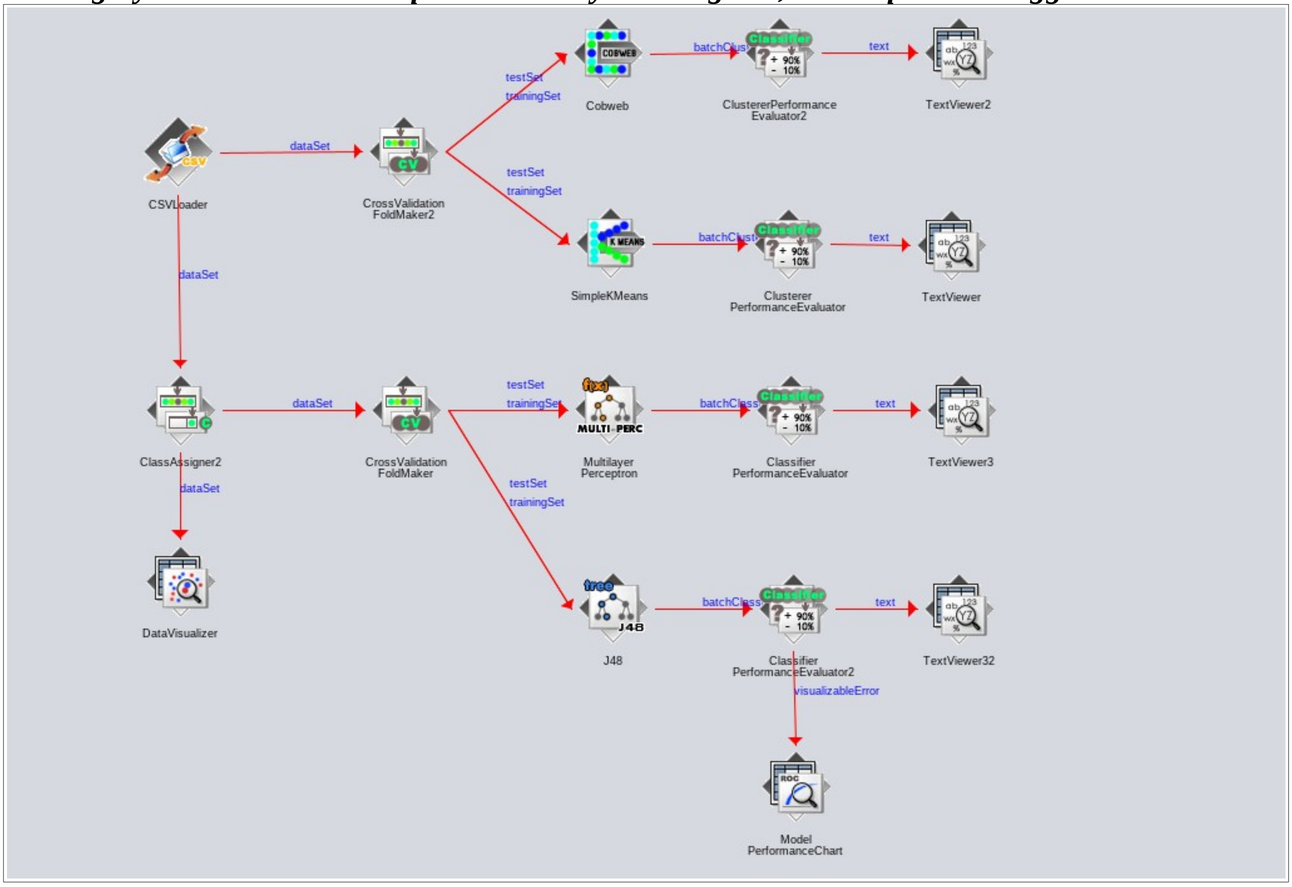
Manipulating the data:

Removing features:
For the “*Estimated fish and invertebrate bycatch*” dataset I have removed unnecessary attributes that are not relevent to our investigation. This included ‘*feature 5*’ which only included two instances through out the dataset. This feature I was unable to find the reasoning for, however because of it’s low appearances it was rendered useless and removed from the investigation.

I also removed ‘*estimated type*’, which described the type of scale the fish had. This was also removed because fish scale does not directly affect bycatch amount. Finally, I removed rows of information that contained (no_data) as it may affect the techniques we apply to it.

Merging features:
Because this is the first question I have not merged the datasets together yet, I have only removed features that are unnecessary to our results.

I will be using a pipeline to analyse my solutions, comparing and contrasting between different learning algorithms in WEKA. Below I have added extra algorithms to test my results thoroughly. From these techniques I will only be using two, as the question suggests.



Determining the methods to use on the dataset:

Classification methods are used for categorical data eg.(yes/no or positive/negative). This may perform poorly because we are expecting to identify correlations between the attributes, not classify them into categories/classes. The method will classify the data into the ‘fish’ categories, returning an accuracy and matrix of how it performed. This may help us in our investigation as we can predict the bycatch amount relative to the next year.

Clustering is quite different in comparison to classification techniques: there there are no outputs in the data at all. So in clustering, the idea is to assign different classes to different data points depending on how they group together. Clustering is called an “unsupervised” learning method, because there are no examples of “correct” labelling that can be used for training. It is also used for finding similarities/patterns in observed data, which relates well to how we want to analyse the dataset. This should perform well on the dataset as we are more concerned on how the bycatch rate performed over a period of time, in relation to fish category.

To have a good comparison, I have chosen two very different techniques to analyse my data with. Firstly, I will run a *MLP* on the dataset which is a classification technique. Secondly, I will use a clustering approach and use *K-Means* to analyse my data. This should help give me a clearer understanding of why a technique may perform better than the other. Especially with a graphical output.

b)

Results:

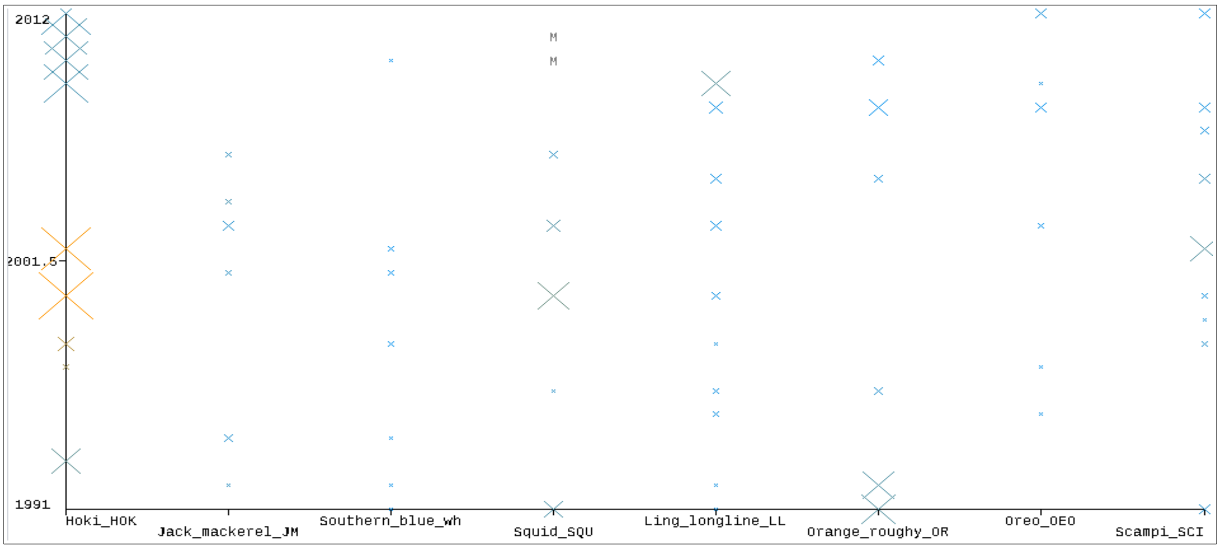
Multi Layer Perceptron in the Pipeline - Classification:

```
Correctly Classified Instances          65               38.0117 %
Incorrectly Classified Instances       106               61.9883 %
Kappa statistic                          0.2843
Mean absolute error                      0.1652
Root mean squared error                  0.2885
Relative absolute error                 75.8379 %
Root relative squared error             87.3886 %
Total Number of Instances              171
```

Simple K-Means in the Pipeline - Clustering:

Visualise option in WEKA :

Note: Usually years would be displayed on the x axis but due to the fish names taking up space it was easier to display on the y.



Bycatch Amount vs Years

```
Final cluster centroids:
=============================================================
Attribute                Full Data          0              1
                          (171.0)        (83.0)         (88.0)
=============================================================
year                     2001.3041      1997.1928      2005.1818
Bycatch_tonnes           9871.5556     13577.1205      6376.5341
Fishery              Jack_mackerel_JMA   Hoki_HOK   Jack_mackerel_JMA
```

Summary:

Simple K-Means:
From running *Simple K-Means clustering* on the dataset, I found some interesting assumptions in the *visualisaltion graph*. Here I am comparing the year vs the fishery (fish type). The greater the X on the graph, the higher the fish bycatch count. *This is interesting a we can see what fish populations have been affected from bycatch over time.*

Let’s take the first category ‘Hoki’ as an example of the fish population decreasing relative to bycatch. If we analyse it closely, we can see large amount of Hoki being bycaught around 1992, then shortly decreasing in count over the next few years. This gives the Hoki population a chance to repopulate, then as we climb towards 1998 the bycatch of this species increases again. This maxs around 1999 – 2001 where there is a significant increase in bycatch of this species.

After this point, the fish population again comes back and is again but due to previous decreases in bycatch population. We can predict that there will be a significant decrease in the Hoki population considering the current bycatch population of the species. This is similar in most fish populations in this data, meaning that this is supporting the fact we are bycatching to the point where we have significant population decreases.

We can also see from the text output that we have two final clusters, clustering relative to attributes with simlarities on Year, Bycatch, Fishery.

Multi-Layer-Perceptron:
The MLP correctly classified 38.0117% of the data correctly, which is extremely low for classification. This confirms that our method of using an MLP is not a good method of making assumptions when we have few attributes to base this off. It is hard to summarise these results based off the poor results but should produce better results when we have more attributes to help our predictions. If we can do this, we will be able to predict which fish category is being affected next. To conclude these results, bycatch amount and year does not help us identify what category of fish is being affected.

c)

How does the results differ from Classification vs Clustering?

Classification:
In classification, the data is categorised under different labels according to some parameters and then the labels are predicted for the data. Multi-Layer-Perceptron in this case is a supervised learning technique and doesn’t perform as well as clustering techniques when testing on this dataset.

This is because the low amount of input nodes with a high amount of output nodes (fish categories). The attributes we are using (Bycatch Amount and Year) didn’t directly support the classification of our data, thus resulting in a low classification accuracy (38.0117 %). To improve our results, a good idea would be to merge relevant attributes to help support this algorithm.

Clustering:
Clustering is the task of partitioning the dataset into groups, called clusters. The goal is to split up the data in such a way that points within single cluster are very similar and points in different clusters are different. It determines grouping among unlabeled data. In this case our results differed quite a lot compared to classification techniques.

Clustering is great to visualise results with where MLP is good text results. I was able to analyse relations between features a lot easier than MLP as the results for MLP were just accuaracy of what the output buffer returned.

In the dataset 'Fishing bycatch' we have 3 main attributes; year, fishery and bycatch amount. Because we are trying to make predictions and analyse correlations, it was much more reasonable to use the clustering approach. Refering back to our dataset, we could see how the results clustered and also projected as a graphical approach. Which is far more suitable for identifying associations of attributes, rather than MLP which is more of text-orientated result outputter.

d)

"Whatever happens, over-fishing today will lead to collapses tomorrow, ... " Jun 14 2019 Stuff

Is there any evidence of fish stocks collapsing in NZ waters?
From the evidence I have gathered we can see that their is a problem with fish stocks being negatively affected. This could be due to many reasons but not neccessarily the '*Estimated fish and invertebrate bycatch'* dataset I have analysed. Although there is evidence to suggest that some fish populations are due to decrease in the next few years. It may be important to use this in further investigation later, with other attributes to support this claim.

Does bycatch amount suggest that their is a population decrease in fish?
Their is some evidence in the fish bycatch trend that indicate drastic changes in fish population. It could be possible that we are bycatching species less because their is less fish to catch. Although this is true, their could be other variables affecting the fish population. For example, less bycatch could be due to temperature change or even pollution. Further investigation would be needed to draw proper conclusion. Combining relevant datasets would be a good start to find a proper correlation between attributes.

Is there evidence to suggest other variables are causing a decrease in the fish population?
Although we are only analysing Bycatch amount in this first section of the assignment, we should further investigate how other variables may be affecting the fish population. Bycatch it's self can possible contribute to the fish population but so can other variables such as Temperature. It would be important to check these variables in comparison to the data we already have.

Part 2: Completion: Feature importance to Fish stocks in New Zealand [40 marks]

a)

Business Case:
Is there evidence to suggest other variables are causing a decrease in the fish population?

This is an interesting question as we can go into further investigation of the fish poluation, relative to the dataset I have already analysed. We can *investigate into other attributes that may be influencing and causing the bycatch* to increase or decrease. This could include the location, time of year or weather conditions. For example a fish population may be greater/less in other parts of New Zealand at different times of year due to temperature. This means we have to run further technique to prove or disprove the reasoning behind the fish stock increase or decrease.

This means we have to find the main features that could be affecting the bycatch amount. I have listed below attributes to investigate:

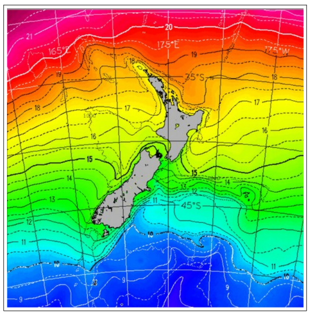
- Location
- Time of the year
- Weather conditions
- Temperature
- Trawling

From this selection I have chosen a *Sea Temperature* dataset, that averages the yearly temperature of NZ sea waters. This should be converted to monthly data to increase our instance count and also accurarcy of temperature. This is specifically important as seasons are a main impact on the temperate of the water. If we can categorise the months into seasons we would have more of an accurate result to analyse.

I have also chosen *Number of trawl tows per year* as my secondary dataset to compliment the bycatch amount dataset. This should give me a clearer understanding of bycatch amount comparitive to the number of trawls each year.

Why was the data gathered? What did the acquisition hope to achieve?

Temperature:
The ocean waters surrounding New Zealand vary in temperature from North to South. They interact with heat and moisture in the atmosphere and affect our weather. The sea surface temperature anomaly provides an indication of the heat change in the ocean. Long-term changes and short-term variability in *sea-surface temperatures can affect marine processes, habitats, and species. Some species may find it hard to survive in changing environmental conditions.*



Visual Representation of average Temperature around NZ (Provided with the dataset)

Trawling:
Seabed trawling is the practice of towing fishing nets near or along the ocean floor. *The towing process can physically damage seabed (benthic) habitats and species.* It can also stir up sediment from the seabed, creating sediment plumes that can smother sensitive species and change light conditions. This can affect marine species (eg by limiting their ability to generate energy through photosynthesis).

Describe why the dataset(s) chosen are appropriate to the business understanding case.

Temperature:
It's likely that temperature (Global warming) could also be affecting the fish population, making it an important dataset to investigate with bycatch amount. From this, it is likely we can draw further conclusions to support our buiness question. This will most likely be shown through similarities, trends and correlations between other attributes. If we relate this back to the dataset, we are investigating if their is a relation between the temperature, bycatch amount and the year. This way

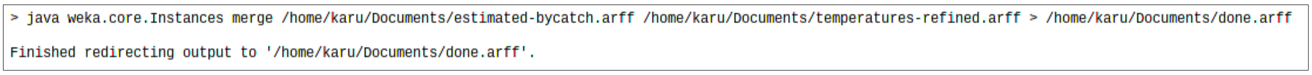
we can further analyse the fish bycatch rate in relation to the temperature each year. **This should help us make a more justified conclusion on whether temperature is a key factor in fish population decrease.**

Trawling:

Trawling is likely to be a main factor to distrupting fish species and living areas. I have added this dataset because I believe it contains valuable attributes such as: ‘number of trawls each year’ that should help us in our investigation. Comparing this to the bycatch amount, we would see if the number of trawls affects the amount of fish bycaught each year. This is complemented with the temperature dataset above.

b)

I used the WEKA CLI technique as described in lectures to merge my datasets into one. This resulted in a dataset that included the temperature, trawl amount and bycatch of marine species by year. In doing this I had to make slight altercations to the temperature dataset which didnt match the marine bycatch instance amount correctly. This involved removing some instances to match the correct amount, which included data reduction techniques.



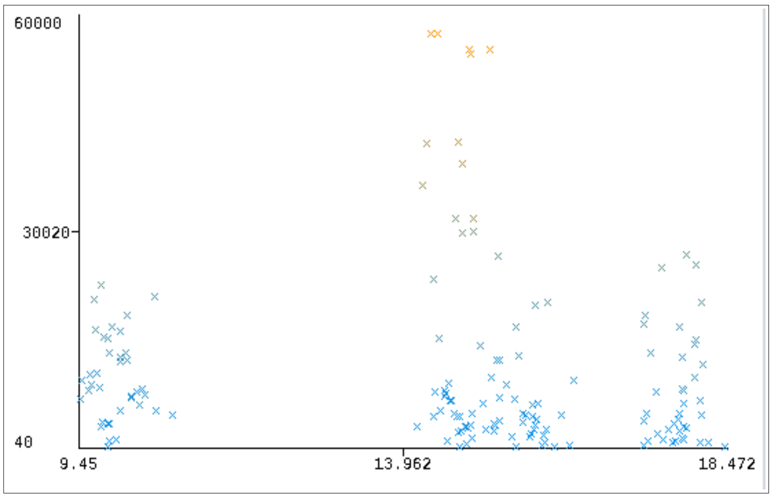
```
> java weka.core.Instances merge /home/karu/Documents/estimated-bycatch.arff /home/karu/Documents/temperatures-refined.arff > /home/karu/Documents/done.arff

Finished redirecting output to '/home/karu/Documents/done.arff'.
```

c)

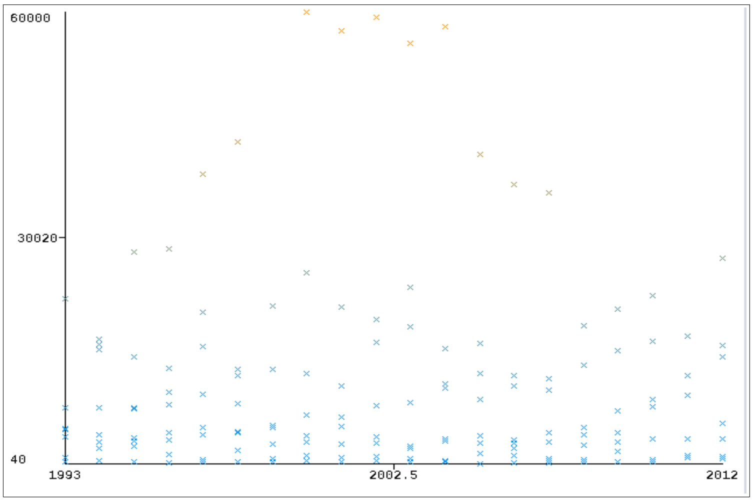
Removal of irrelevant features was necessary in producing suitable data to analyse. For the Temperature dataset I removed the *sample size and standard deviation* as we wern’t using these features to support our question of temperature affecting fish populations. Along with the refined bycatch dataset, merging the dataset was quite easy. The only conflict was the instance amount being different and the years not corresponding properly.

This was resolved through keeping neccessary data from the bycatch and removing redundant instances that wernt necessary to our research. eg. (years from 1991 – 1993 didnt include any temperature features at this time, so it was removed). This was also the case for instances past 2013 which were deleted to match the correct years of the sample.

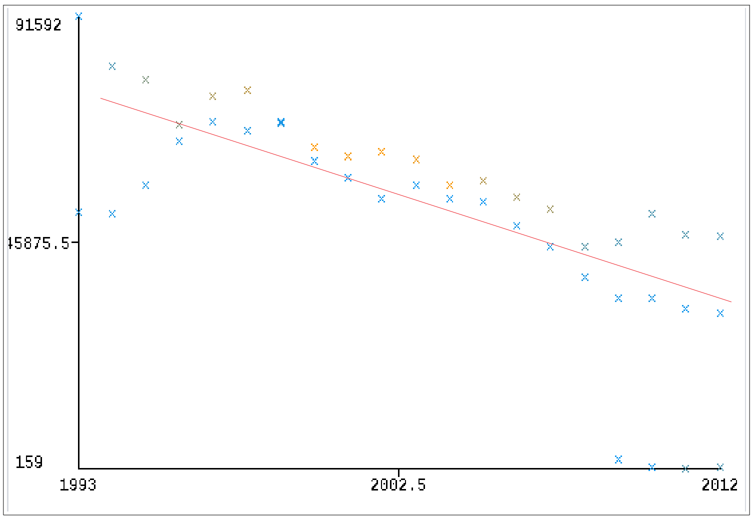
Finally, the last technique involved replicating data to suit the right instance count before merging. Each year contained 8 fish with a bycatch count, vs the temperature dataset which contained 4 locations with corresponding temperature means. To account for this, I duplicated the data to 8 locations to match the years and fish categories. After this was done I merged the dataset using the WEKA CLI. Once I had completed these steps, I ran the new dataset through my pipeline and analysed the results:



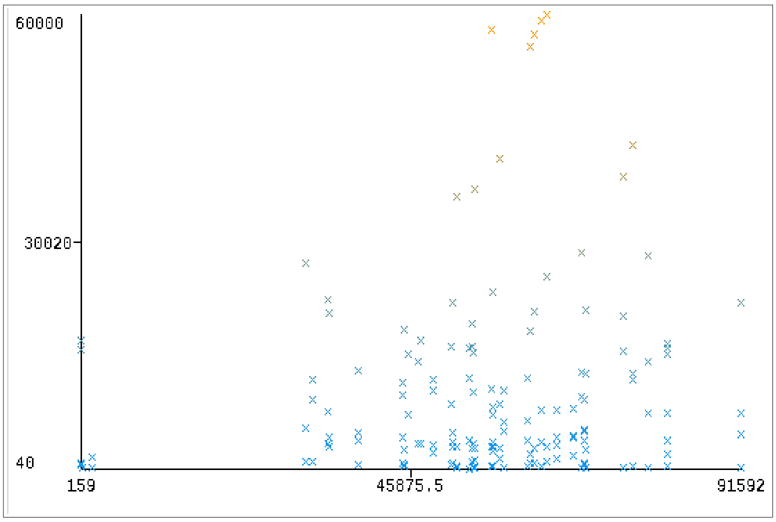
Temperature vs Bycatch Amount
(Orange = high bycatch, blue = low)



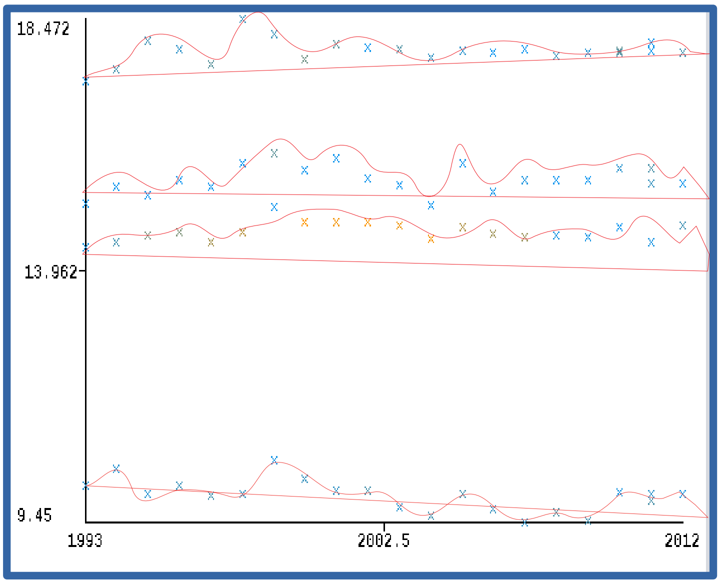
Year vs Bycatch amount
(Orange = high bycatch, blue = low)



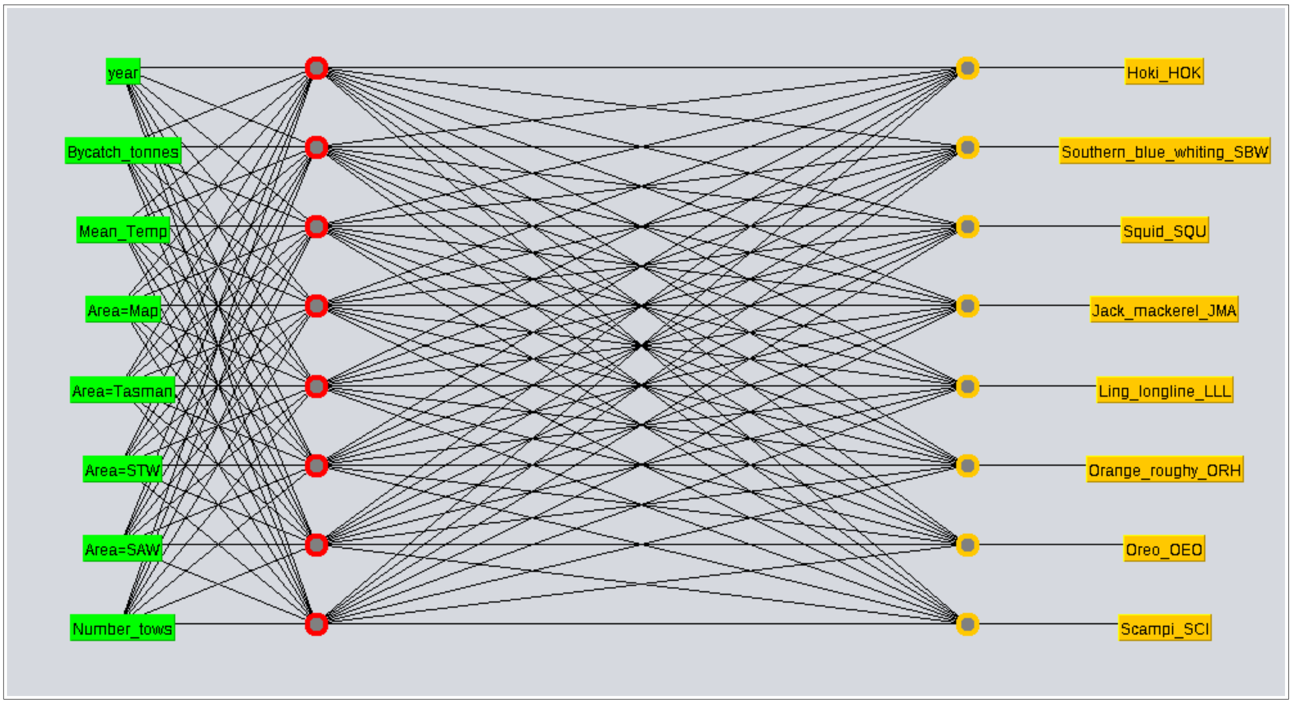
Year vs Trawl Amount
(Orange = higher temp, blue = lower)



Number of Tows vs Bycatch
(Orange = high bycatch, blue = low)



Years vs Temperature
(Orange = high bycatch, blue = low)



Mulitlayer-Perceptron

```
Correctly Classified Instances          87               56.129  %
Incorrectly Classified Instances        68               43.871  %
Kappa statistic                          0.4982
Mean absolute error                      0.1269
Root mean squared error                  0.2804
Relative absolute error                 58.0271 %
Root relative squared error             84.7979 %
Total Number of Instances              155
```

Clustering results:
One major result to note is the direct decrease correlation in trawl amount over the past 10 - 20 years. If we apply the temperature class colour to this, we can see there was a larger trawl amount around the mid-higher range.

This means we have begun to trawl less over the past few years, regardless of temperature fluxuations tha also affect the fish. This means that trawling and bycatch may not directly affect the fish population in NZ. Although this is true now, it is likely this was significantly different when our trawl amount and bycatch rate was at its greatest.

Relating back to our question of ***‘Is there evidence to suggest other variables are causing a decrease in the fish population?’*** If we consider temperature as a possible variable, we can see there isn’t much of a correlation or trend between the bycatch amount and temperature. Although this is true, there is a bell shaped trend with bycatch and year. Indicating that we were once bycatching a large amount but due to to population decrease there a less fish to be caught. This is unlikely though as we have decreased the amount in which we trawl, meaning there is less bycatch amount.

From this we can see that bycatch amount is significantly lower in cold temperatures, as well as hot. Compared to around the mean of 14 degrees, which is in the area MAP. This area is described in the description as the average temperature around NZ waters, which is what we should focus on the most.

MLP results:
The classification accuracy has increased by about 20%. We can predict around 56% of the time what category of fish we are predicting based off the attributes we have fed our MLP. This means we could start to make realistic predictions about the future fish populations of these categories. This however would only apply if we had a higher classification accuracy, as the result is too innaccurate to make assumptions.

To conclude, I would support the business case that there are other variables involved with the affected fish population. To further investigate, I would research water pollution and include temperature also.