COMP 309
Assignment two

Name: Siwen Feng
ID: 300363512
Dataset: Fish stocks

Business understanding

General description:

The fish resource is one of the more important resources for human beings. Fish is also the most common food for human and its industry has been considered as a very profitable business to exploit. Nowadays, with the increasing amount of fish consumption, overfishing has been engaged more attention globally. This assignment aims to find out the actual influence brought to fish stocks by human's fishing activities based on a fish stock dataset. The results will help us to decide whether we should believe the News headline. Hence, it could potentially inspire us to take actions and protect our fishing stocks.

Business objectives:

- Is there any evidence of fish stocks collapsing in NZ waters?
- How does the change in fish stocks affect the New Zealand marine environment?
- What criteria does the government use to formulate fish protection policies?

Situation assessment:

Until 2016, the fishing industry is taking account of 0.18% New Zealand GDP which is approximately $459 million. The related industry like seafood processing also takes 0.2% of NZ GDP. Those industries offered 43,640 jobs. Meanwhile, the total exports value increased by 43% since 2003. However, the fish resource is limited and can be affected by many factors (climate, pollution, and human activities). New Zealand government claimed that they put massive efforts to make aquaculture industry sustainable. Therefore, the media holds a negative view of fishing scenarios. Hence, I will use and analyze the authorized dataset to make a fair conclusion regarding real fishing scenarios.

Data mining goal:

- Investigating the characteristics of the dataset
- Summarize and find the future trend of the dataset.

Project plan:

1. Collect initial data
2. Modify and pre-processing raw data.
3. Analyze data using machine learning methods.
4. Compare business goals.
5. Get a conclusion.

Data understanding

Initial data:

I collected initial data from Data NZ (https://www.data.govt.nz/). The dataset called "Fish monetary stock account 06-18" which contains the variation of fish stocks over the last 12 years. This dataset also has a few attributes like catchment per species, TACC per species, etc. I chose this dataset because it provides attributes related to fish stocks and the size of the dataset is reasonable. We can also find 12 years of statistics which could contribute to the accuracy of the prediction.

Data description:

species	year	variable	units	magnitude	source	data_value

The initial dataset consists of 7 attributes and they are species, year, variable, units, magnitude, source, and data-value respectively. Data values contain the amount of the variable. Species indicates the various range of the fish. TACC stands for the total allowable commercial catchment which is the estimated maximum amount of fishing per species. The catchment represents actual fishing amount. Based on observation, the attributes above are highly relevant to overfishing.

Data verification:

The information contained in dataset is useful but complex and hard to understand. The variable contains the important values with inconsistent units. For example, the asset value is money and use dollars as the unit. TACC is the number of the allowable catchment which uses ton as the unit. Thus, when I used weka to load data it will lead to an inaccurate result. Then, the dataset is of high complexity and high noise which results in a medium quality dataset.

For solving the problems above, we can modify the data from three aspects

- Reformating the data and make it clearer to follow.
- Getting rid of meaningless attributes.
- Add functional features to strengthen the characters.

Data preparation

Reformat data:

variable
Asset value
Catch
TACC

Before

The variable attribute contains three important values. To represent data clearly, I choose to change values to attributes. Hence, those important values can have their column to display data.

species	year	asset value - $.M	catchment -tonnes	TACC - tonnes

After

Eliminate meaningless attributes:

units	magnitude	source
Dollars	Millions	Environmental Accounts
Dollars	Millions	Environmental Accounts
Dollars	Millions	Environmental Accounts

Those attributes in the initial dataset are related to the variable. In the previous step, we have reformated the dataset and change items in the variable to attributes. Thus, those attributes can be removed from the new dataset.

Add valuable attributes:

Is under TACC	TACC changes/year	Catchment changes/year
No	0	1593.1
No	0	3035.7
Yes	0	-6066.4
Yes	0	735.9
Yes	0	-1765.4
Yes	0	926.3

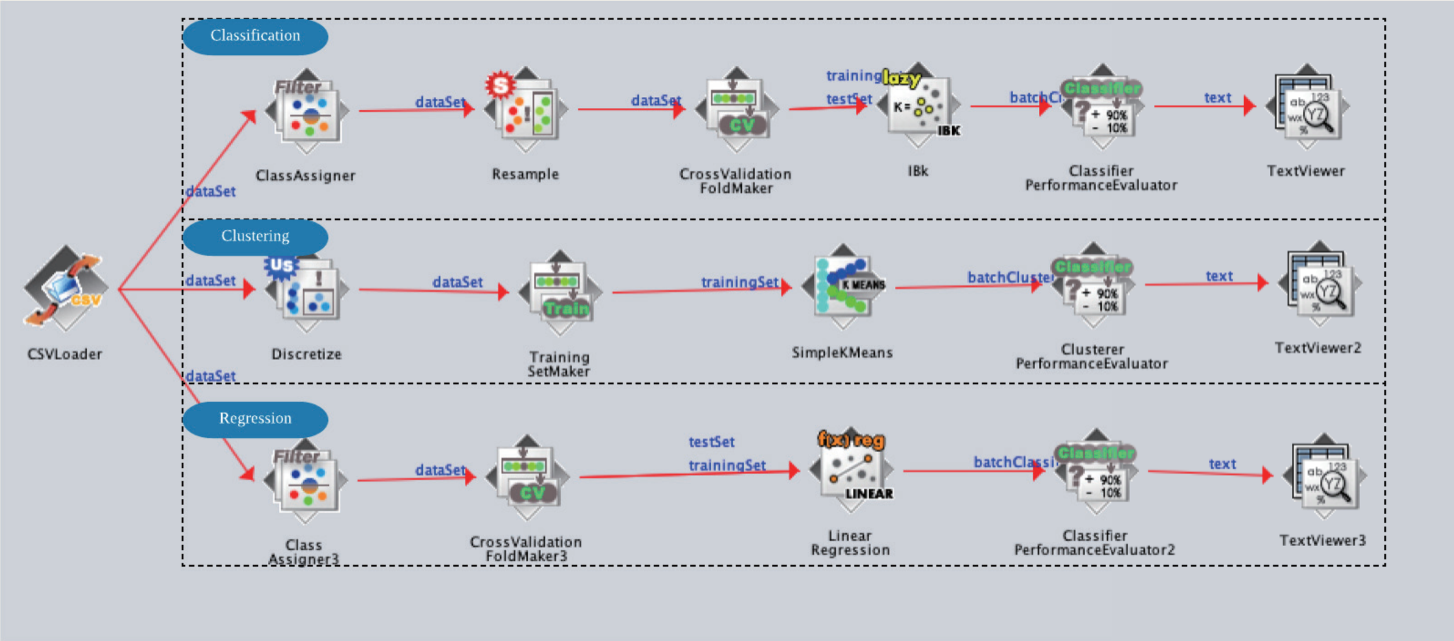
To make the dataset have stronger characteristics, I added three useful attributes. They are the changes between actual catchment and TACC, changes between TACC and previous year TACC, and the difference between this year's catchment and previous year's catchment respectively. The added attributes can help us analyze the data from different directions. As a consequence, we can get an obvious and clear trend of the dataset after manipulating it.

Dataset display:

species	year	asset value - $.M	catchment -tonnes	TACC - tonnes	Is under TACC	TACC changes/year	Catchment changes/year
Silver Warehou	2006	72	11138	10380.2	No	0	1593.1
Silver Warehou	2007	85.7	14173.7	10380.2	No	0	3035.7
Silver Warehou	2008	82.7	8107.3	10380.2	Yes	0	-6066.4
Silver Warehou	2009	86.4	8843.2	10380.2	Yes	0	735.9
Silver Warehou	2010	89	7077.8	10380.2	Yes	0	-1765.4
Silver Warehou	2011	100.7	8004.1	10380.2	Yes	0	926.3
Silver Warehou	2012	83.6	7130.3	10380.2	Yes	0	-873.8
Silver Warehou	2013	75.3	8663.1	10380.2	Yes	0	1532.8
Silver Warehou	2014	85.2	7988.1	10380.2	Yes	0	-675
Silver Warehou	2015	96.7	9052.6	10380.2	Yes	0	1064.5
Silver Warehou	2016	124.6	7514.9	10380.2	Yes	0	-1537.7
Silver Warehou	2017	143.1	8670.7	10380.2	Yes	0	1155.8
Silver Warehou	2018	171.1	8652.8	10380.2	Yes	0	-17.9
Blue Cod	2006	56.7	2187.4	2681.5	Yes	0	-264.9
Blue Cod	2007	46.2	2419.8	2681.5	Yes	0	232.4
Blue Cod	2008	41.3	2316	2681.5	Yes	0	-103.8
Blue Cod	2009	39.6	2418.2	2681.5	Yes	0	102.2
Blue Cod	2010	46.3	2162.5	2681.5	Yes	0	-255.7
Blue Cod	2011	48	2342.6	2681.5	Yes	0	180.1
Blue Cod	2012	45.8	2216.5	2331.6	Yes	-349.9	-126.1
Blue Cod	2013	53.8	2193.5	2331.6	Yes	0	-23
Blue Cod	2014	63.8	2176.1	2331.6	Yes	0	-17.4
Blue Cod	2015	56.5	2207.4	2331.6	Yes	0	31.3
Blue Cod	2016	130.2	2105.7	2331.6	Yes	0	-101.7
Blue Cod	2017	81.4	2155.1	2331.6	Yes	0	49.4
Blue Cod	2018	148	2045.3	2331.6	Yes	0	-109.8
Southern Blue Whiting	2006	71.1	30277.6	35648	Yes	0	8658
Southern Blue Whiting	2007	54.9	25363.4	30648	Yes	-5000	-4914.2
Southern Blue Whiting	2008	63.2	25586.6	30648	Yes	0	223.2

Modeling

Pipeline simulation:



In the modeling part, I implemented three techniques, classification, clustering and regression. Three machine learning methods are specializing in different fields. In the following part, I will explain the results of those three techniques based on the given dataset. Meantime, I will also address the difference between the three techniques.

Classification:

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

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

Correctly Classified Instances         230               99.1379 %
Incorrectly Classified Instances         2                0.8621 %
Kappa statistic                          0.9828
Mean absolute error                      0.0109
Root mean squared error                  0.0928
Relative absolute error                  2.175  %
Root relative squared error             18.5645 %
Total Number of Instances              232

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.017    0.983      1.000    0.991      0.983    0.988     0.970     No
                 0.983    0.000    1.000      0.983    0.991      0.983    0.988     0.993     Yes
Weighted Avg.    0.991    0.009    0.992      0.991    0.991      0.983    0.988     0.981

=== Confusion Matrix ===

   a   b   <-- classified as
 116   0 |   a = No
   2 114 |   b = Yes
```

From observation, we can find that classification use percentage to represent the result. IBK algorithm delivers a good performance regarding this dataset (99.13%). However, this is not an ideal and appropriate form to illustrate this dataset. Because we are aiming to see the future trend of the fishing industry. The classification is very effective to label the existing instances. Even though the performance is very high but it does not match my data mining goal. Thus, alternative options should be considered.

Clustering:

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 124.78471602708888

Initial starting points (random):

Cluster 0: Scampi,2008,188.9,668.2,1291,Yes,0,-169.5
Cluster 1: 'Orange Roughy',2016,362.5,7811.1,8736,Yes,0,-405.3
Cluster 2: 'Southern Blue Whiting',2012,82.7,38439.5,43408,Yes,-1440,-268.6

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                   Cluster#
Attribute                          Full Data              0                1                     2
                                     (130.0)         (54.0)           (50.0)                (26.0)
==================================================================================================
species                       Silver Warehou         Scampi    Orange Roughy Southern Blue Whiting
year                                    2012      2009.6667          2014.52                  2012
asset value - $.M                     420.03       281.2352           521.62              512.9308
catchment -tonnes                 19869.6746      4743.1167         6815.446            76390.6577
TACC - tonnes                     22492.9677      6125.5963         8235.252            83905.4231
Is under TACC                            Yes            Yes              Yes                   Yes
TACC changes/year                   436.0338        -4.9463          -124.31                2429.5
Catchment changes/year              176.5062      -127.2519          -12.836             1171.5077




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

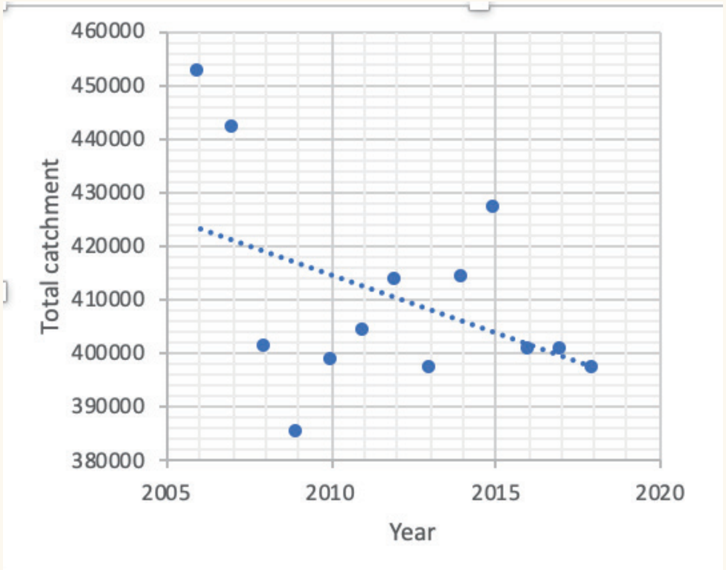
```
=== Model and evaluation on training set ===

Clustered Instances

0       54 ( 42%)
1       50 ( 38%)
2       26 ( 20%)
```

The result indicates that the clustering method is incompatible with a numeric dataset. However, the given dataset only has two nominal features. The rest of attributes are all numeric attributes. The objective of the task is not grouping the data as well. The eventual goal of implementing fish stocks dataset is to predict the risks of overfishing in the future. Cluster analysis is not helpful to forecast the future trend. Thus, clustering is not an appropriate form for representing my dataset as well.

Regression:

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


Linear Regression Model

Catchment changes/year =

   4153.7453 * species=Ling,Paua,Blue Cod,Scampi,Snapper,Rock Lobster,Orange Roughy,Silver Warehou,Hoki +
   1738.7195 * species=Paua,Blue Cod,Scampi,Snapper,Rock Lobster,Orange Roughy,Silver Warehou,Hoki +
  -1695.5132 * species=Snapper,Rock Lobster,Orange Roughy,Silver Warehou,Hoki +
   1431.2998 * species=Rock Lobster,Orange Roughy,Silver Warehou,Hoki +
  -8707.5648 * species=Hoki +
      0.0605 * TACC - tonnes +
   1601.7515 * Is under TACC=No +
      1.0978 * TACC changes/year +
  -5987.2548

Time taken to build model: 0.01 seconds

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

Correlation coefficient                  0.875
Mean absolute error                    982.8675
Root mean squared error               2224.5674
Relative absolute error                 41.1049 %
Root relative squared error             48.4902 %
Total Number of Instances              232
```

Regression uses a statistic model to find relationships among variables. Regression is used for predicting and forecasting the future trend of variables. This is matching with my data mining goal - to gather characters of my dataset and trend of the dataset shows. I have done a regression regarding Catchment. Catchment's change is the select class. The results show strong positive correlations (0.875). The diagram helps us have a better understanding of the result given above.



As we can see from the diagram on the left, this dataset showed a strong negative correlation. This means that with time goes by, the fishing amount declined. This regression analysis reveals a different result which is advertised by media.

It also explained why the regression result has a high mean absolute error. After observing the diagram, we can find that the value is quite spread out. But the trend of the correlation is obvious. We still can find a strong correlation on it. For predicting the future, the regression analysis is the best form to achieve my data mining goal.

Different aspects of techniques:

Classification is the process of categorizing instances into different classes. In the classification process, we should have distinct classes or label the algorithm which can help decide which categories the item should go, related to my dataset. The only nominal attribute is whether catchment is under TACC. At the same time, classification is a supervised learning method. K-nearest neighbor ask us to define a K value which represents the K numbers neighbors. In my dataset, 10 species have little relationships. In the meantime, the numeric attributes are highly random and it may cause inaccurate results.

Cluster analysis is the process of grouping a set of unlabeled objects or instances into a few groups. Instances in one group all have similar characters. Clustering is normally done as an unsupervised learning technique. I used k-means clustering method to achieve my goal on this dataset. For my dataset, all attributes have clear relationships with other features. The potential class attributes should be taken considered. Hence, there is no way to define a good K value to implement K-mean clustering.

Regression is another supervised learning method which relies on mathematics models to calculate the relationships among variables. Linear Regression uses dependent and independent variables to predict future possible value. For my dataset, I inputted catchment as my variable and it gave me a strong correlation, which indicates the strong relastions among the variables. However, the value of dataset is big. The absolute mean errors haven been raised. But regression is still the best prediction analysis method among the three techniques.

Evaluation

Review business understanding:

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

After analyzing the dataset, the regression of actual catchment helps me to understand the real fishing scenarios. The catchment has been slightly declined. The prediction of future trend is also decreasing. This result conveys us with a positive message that the NZ fishing Industry has a small chance to collapse. Nevertheless, we can never say the result we got is accurate. The variation in fish stocks can be affected by many factors. The conclusion I made is only validated based on this vdataset. For considering other aspects, we should also get more information.

Further two questions:

- How does the change in fish stocks affect the New Zealand marine environment?

Ecosystems are interrelated. This question can explore the chain reaction brought about by changes in one aspect. Fisheries are important for the marine environment. Will the change in fish stocks dramatically impact on the marine environment? It is worth to investigate.

- What criteria does the government use to formulate fish protection policies?

When the government decides on fish protection policies, what are the prioritized indicators to push them to make the decision? It is very interesting to explore. In part 2 , I will merge or append two foreign data sets to my original dataset to gather more information and find the answer to this question.

Business understanding(Part 2)

- What criteria does the government use to formulate fish protection policies?

Project plan:

1. Describe and explain new supporting datasets.
2. Merge datasets
3. Dimensionality reduction
4. Analyze results
5. Give the conclusion

Data understanding(Part 2)

In part one, we were focusing on fish stocks. The dataset gives us a clear indication of the future trend brought by catchment. For achieving the new business objective - understanding the motivation behind why the New Zealand government set those regulations. We should merge a few support dataset.

In fish stock status published by fishery New Zealand, there are three indexes related to fishing. These are soft limit, hard limit, and the management target of average fishing. They are represented as percentages. I used the 2009-2015 soft limit and hard limit data as my support datasets to observe how those limits affect the government's behavior.

Why soft limit and hard limt?

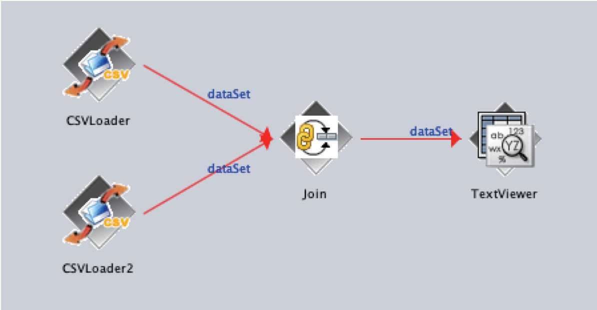
Soft limit and hard limit directly reflect the real situation of one marine species. If the fish stock is above both hard and soft limit, it means the fish remain in a healthy quantity, vice versa. Thus, those two indicators help us to make the decision should the government adjust downwardly the allowable catchment. If TACC and soft and hard limits have a strong relationship, we can say, these two indexes can be the guide when the government trying to create new fishing protection policy. In the following part, I will merge these two datasets with my original data and modify it.

Verification data:

- Merge data
- Deal with missing values
- Dimensionality reduction

Data preparation(Part 2)

merge data:



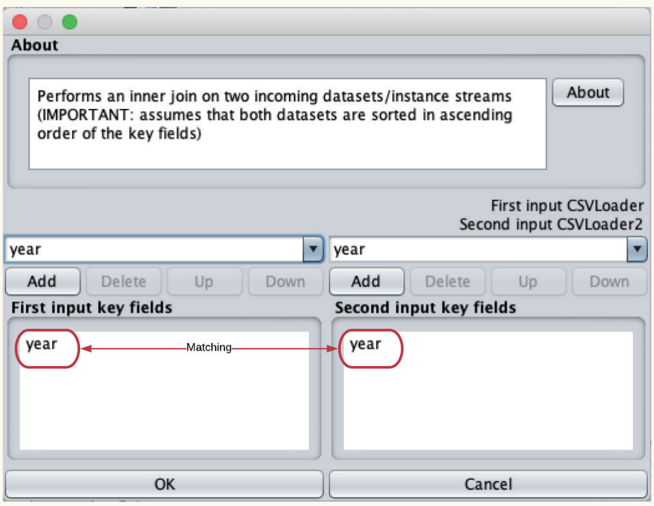
Weka has a function called "Join". The function helps us to merge two datasets. My original dataset shared the same years with soft and hard limits. Thus, I used years as my joint key to making three datasets together.

species	year	asset value - $.M	catchment -tonnes	TACC - tonnes	Is under TACC	TACC changes/year	Catchment changes/year
Silver Warehou	2006	72	11138	10380.2	No	0	1593.1
Silver Warehou	2007	85.7	14173.7	10380.2	No	0	3035.7
Silver Warehou	2008	82.7	8107.3	10380.2	Yes	0	-6066.4
Silver Warehou	2009	86.4	8843.2	10380.2	Yes	0	735.9
Silver Warehou	2010	89	7077.8	10380.2	Yes	0	-1765.4
Silver Warehou	2011	100.7	8004.1	10380.2	Yes	0	926.3
Silver Warehou	2012	83.6	7130.3	10380.2	Yes	0	-873.8
Silver Warehou	2013	75.3	8663.1	10380.2	Yes	0	1532.8
Silver Warehou	2014	85.2	7988.1	10380.2	Yes	0	-675
Silver Warehou	2015	96.7	9052.6	10380.2	Yes	0	1064.5
Silver Warehou	2016	124.6	7514.9	10380.2	Yes	0	-1537.7
Silver Warehou	2017	143.1	8670.7	10380.2	Yes	0	1155.8
Silver Warehou	2018	171.1	8652.8	10380.2	Yes	0	-17.9
Blue Cod	2006	56.7	2187.4	2681.5	Yes	0	-264.9
Blue Cod	2007	46.2	2419.8	2681.5	Yes	0	232.4
Blue Cod	2008	41.3	2316	2681.5	Yes	0	-103.8
Blue Cod	2009	39.6	2418.2	2681.5	Yes	0	102.2
Blue Cod	2010	46.3	2162.5	2681.5	Yes	0	-255.7
Blue Cod	2011	48	2342.6	2681.5	Yes	0	180.1
Blue Cod	2012	45.8	2216.5	2331.6	Yes	-349.9	-126.1
Blue Cod	2013	53.8	2193.5	2331.6	Yes	0	-23
Blue Cod	2014	63.8	2176.1	2331.6	Yes	0	-17.4
Blue Cod	2015	56.5	2207.4	2331.6	Yes	0	31.3
Blue Cod	2016	130.2	2105.7	2331.6	Yes	0	-101.7
Blue Cod	2017	81.4	2155.1	2331.6	Yes	0	49.4
Blue Cod	2018	148	2045.3	2331.6	Yes	0	-109.8
Southern Blue Whiting	2006	71.1	30277.6	35648	Yes	0	8658
Southern Blue Whiting	2007	54.9	25363.4	30648	Yes	-5000	-4914.2
Southern Blue Whiting	2008	63.2	25586.6	30648	Yes	0	223.2

Before

species	year	asset value - $.M	catchment -tonnes	TACC - tonnes	Is under TACC	TACC changes/year	Catchment changes/year	landings_from_stocks_above_soft_limit	landings_from_stocks_above_hard_limit
ilver Wareho	2006	72	11138	10380.2	No	0	1593.1	?	?
ilver Wareho	2007	85.7	14173.7	10380.2	No	0	3035.7	?	?
ilver Wareho	2008	82.7	8107.3	10380.2	Yes	0	-6066.4	?	?
ilver Wareho	2009	86.4	8843.2	10380.2	Yes	0	735.9	94	99.5
ilver Wareho	2010	89	7077.8	10380.2	Yes	0	-1765.4	94.8	99.1
ilver Wareho	2011	100.7	8004.1	10380.2	Yes	0	926.3	95.1	97.1
ilver Wareho	2012	83.6	7130.3	10380.2	Yes	0	-873.8	96.6	99.5
ilver Wareho	2013	75.3	8663.1	10380.2	Yes	0	1532.8	96.1	99.5
ilver Wareho	2014	85.2	7988.1	10380.2	Yes	0	-675	96.4	99.6
ilver Wareho	2015	96.7	9052.6	10380.2	Yes	0	1064.5	96.8	99.6
ilver Wareho	2016	124.6	7514.9	10380.2	Yes	0	-1537.7	?	?
ilver Wareho	2017	143.1	8670.7	10380.2	Yes	0	1155.8	?	?
ilver Wareho	2018	171.1	8652.8	10380.2	Yes	0	-17.9	?	?
Blue Cod	2006	56.7	2187.4	2681.5	Yes	0	-264.9	?	?
Blue Cod	2007	46.2	2419.8	2681.5	Yes	0	232.4	?	?
Blue Cod	2008	41.3	2316	2681.5	Yes	0	-103.8	?	?
Blue Cod	2009	39.6	2418.2	2681.5	Yes	0	102.2	94	99.5
Blue Cod	2010	46.3	2162.5	2681.5	Yes	0	-255.7	94.8	99.1
Blue Cod	2011	48	2342.6	2681.5	Yes	0	180.1	95.1	97.1
Blue Cod	2012	45.8	2216.5	2331.6	Yes	-349.9	-126.1	96.6	99.5
Blue Cod	2013	53.8	2193.5	2331.6	Yes	0	-23	96.1	99.5
Blue Cod	2014	63.8	2176.1	2331.6	Yes	0	-17.4	96.4	99.6
Blue Cod	2015	56.5	2207.4	2331.6	Yes	0	31.3	96.8	99.6
Blue Cod	2016	130.2	2105.7	2331.6	Yes	0	-101.7	?	?
Blue Cod	2017	81.4	2155.1	2331.6	Yes	0	49.4	?	?
Blue Cod	2018	148	2045.3	2331.6	Yes	0	-109.8	?	?
hern Blue Wh	2006	71.1	30277.6	35648	Yes	0	8658	?	?
hern Blue Wh	2007	54.9	25363.4	30648	Yes	-5000	-4914.2	?	?
hern Blue Wh	2008	63.2	25586.6	30648	Yes	0	223.2	?	?
hern Blue Wh	2009	77.7	31887.4	36948	Yes	6300	6300.8	94	99.5
hern Blue Wh	2010	76.7	39540.1	41848	Yes	4900	7652.7	94.8	99.1
hern Blue Wh	2011	88.7	38708.1	44848	Yes	3000	-832	95.1	97.1
hern Blue Wh	2012	82.7	38439.5	43408	Yes	-1440	-268.6	96.6	99.5
hern Blue Wh	2013	85.8	29906.1	43408	Yes	0	-8533.4	96.1	99.5
hern Blue Wh	2014	92.3	33454.8	43408	Yes	0	3548.7	96.4	99.6
hern Blue Wh	2015	125.2	31866.5	53208	Yes	9800	-1588.3	96.8	99.6
hern Blue Wh	2016	178	24733.4	49288	Yes	-3920	-7133.1	?	?
hern Blue Wh	2017	172.2	22587.8	49288	Yes	0	-2145.6	?	?
hern Blue Wh	2018	172.6	21045.9	48815	Yes	-473	-1541.9	?	?
Ling	2006	217.4	14177.8	21977.1	Yes	0	-3008.3	?	?
Ling	2007	238.4	16099.3	21977.1	Yes	0	1921.5	?	?
Ling	2008	233.4	16262.7	21977.1	Yes	0	163.4	?	?

After



However, it raises another problem for the dataset. Because my original dataset is from the 2006-2018. Soft and hard limit datasets only offered data from 2009-2015. The new merged dataset has to contain some missing values. I will deal with missing values in the next steps.

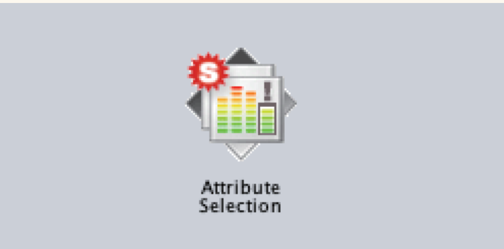
Replace missing values:

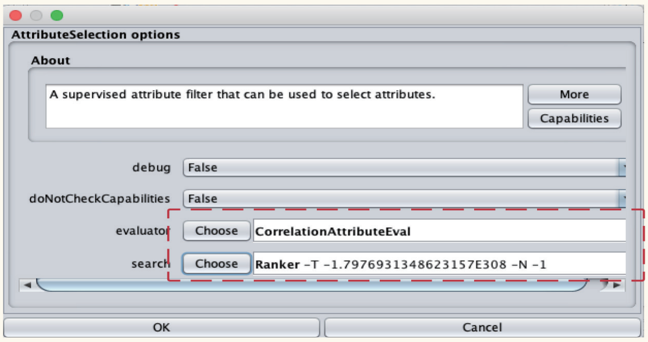
Weka has a function called "replaceMissingValue". I used this function to fill up the missing value in my dataset. The function fills up the gaps by using the mean value.

species	year	asset value - $.M	catchment -tonnes	TACC - tonnes	Is under TACC	TACC changes/year	Catchment changes/year	landings_from_stocks_above_soft_limit	landings_from_stocks_above_hard_limit
lver Wareh	2006	72	11138	10380.2	No	0	1593.1	95.69	99.13
lver Wareh	2007	85.7	14173.7	10380.2	No	0	3035.7	95.69	99.13
lver Wareh	2008	82.7	8107.3	10380.2	Yes	0	-6066.4	95.69	99.13
lver Wareh	2009	86.4	8843.2	10380.2	Yes	0	735.9	94	99.5
lver Wareh	2010	89	7077.8	10380.2	Yes	0	-1765.4	94.8	99.1
lver Wareh	2011	100.7	8004.1	10380.2	Yes	0	926.3	95.1	97.1
lver Wareh	2012	83.6	7130.3	10380.2	Yes	0	-873.8	96.6	99.5
lver Wareh	2013	75.3	8663.1	10380.2	Yes	0	1532.8	96.1	99.5
lver Wareh	2014	85.2	7988.1	10380.2	Yes	0	-675	96.4	99.6
lver Wareh	2015	96.7	9052.6	10380.2	Yes	0	1064.5	96.8	99.6
lver Wareh	2016	124.6	7514.9	10380.2	Yes	0	-1537.7	95.69	99.13
lver Wareh	2017	143.1	8670.7	10380.2	Yes	0	1155.8	95.69	99.13
lver Wareh	2018	171.1	8652.8	10380.2	Yes	0	-17.9	95.69	99.13
Blue Cod	2006	56.7	2187.4	2681.5	Yes	0	-264.9	95.69	99.13
Blue Cod	2007	46.2	2419.8	2681.5	Yes	0	232.4	95.69	99.13
Blue Cod	2008	41.3	2316	2681.5	Yes	0	-103.8	95.69	99.13
Blue Cod	2009	39.6	2418.2	2681.5	Yes	0	102.2	94	99.5
Blue Cod	2010	46.3	2162.5	2681.5	Yes	0	-255.7	94.8	99.1
Blue Cod	2011	48	2342.6	2681.5	Yes	0	180.1	95.1	97.1
Blue Cod	2012	45.8	2216.5	2331.6	Yes	-349.9	-126.1	96.6	99.5
Blue Cod	2013	53.8	2193.5	2331.6	Yes	0	-23	96.1	99.5
Blue Cod	2014	63.8	2176.1	2331.6	Yes	0	-17.4	96.4	99.6
Blue Cod	2015	56.5	2207.4	2331.6	Yes	0	31.3	96.8	99.6
Blue Cod	2016	130.2	2105.7	2331.6	Yes	0	-101.7	95.69	99.13
Blue Cod	2017	81.4	2155.1	2331.6	Yes	0	49.4	95.69	99.13
Blue Cod	2018	148	2045.3	2331.6	Yes	0	-109.8	95.69	99.13
hern Blue W	2006	71.1	30277.6	35648	Yes	0	8658	95.69	99.13
hern Blue W	2007	54.9	25363.4	30648	Yes	-5000	-4914.2	95.69	99.13
hern Blue W	2008	63.2	25586.6	30648	Yes	0	223.2	95.69	99.13
hern Blue W	2009	77.7	31887.4	36948	Yes	6300	6300.8	94	99.5
hern Blue W	2010	76.7	39540.1	41848	Yes	4900	7652.7	94.8	99.1
hern Blue W	2011	88.7	38708.1	44848	Yes	3000	-832	95.1	97.1
hern Blue W	2012	82.7	38439.5	43408	Yes	-1440	-268.6	96.6	99.5
hern Blue W	2013	85.8	29906.1	43408	Yes	0	-8533.4	96.1	99.5
hern Blue W	2014	92.3	33454.8	43408	Yes	0	3548.7	96.4	99.6
hern Blue W	2015	125.2	31866.5	53208	Yes	9800	-1588.3	96.8	99.6
hern Blue W	2016	178	24733.4	49288	Yes	-3920	-7133.1	95.69	99.13
hern Blue W	2017	172.2	22587.8	49288	Yes	0	-2145.6	95.69	99.13
hern Blue W	2018	172.6	21045.9	48815	Yes	-473	-1541.9	95.69	99.13

After

Dimensionality reduction(clear noise):






```
Attribute Evaluator (supervised, Class (numeric): 9 landings_from_stocks_above_soft_limit):
        Correlation Ranking Filter
Ranked attributes:
 0.36005720220225745      2 year
 0.3358275259861406      10 landings_from_stocks_above_hard_limit
 0.07431965898867963      3 asset value - $.M
 0.04316911467866966      5 TACC - tonnes
 0.04009763345516277      4 catchment -tonnes
 0.03767561125990619      6 Is under TACC
 0.02936224164019883      8 Catchment changes/year
 0.02768274994076389      7 TACC changes/year
 0.0000000000000000969    1 species
```

Attributes ranker is one of the dimensionality reduction methods. The method gives us a general idea of what the important attributes are for this dataset. The diagram above shows the least related four attributes. Thus, I will remove those four attributes and make a new dataset as a subset of the original dataset. In the following, I will compare two result before and after

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


Linear Regression Model

landings_from_stocks_above_soft_limit =

     -1.0971 * species=Hoki,Blue Cod,Southern Blue Whiting,Rock Lobster +
      1.0927 * species=Blue Cod,Southern Blue Whiting,Rock Lobster +
      0.0649 * year +
      0      * catchment -tonnes +
     -0      * TACC - tonnes +
      0.3385 * landings_from_stocks_above_hard_limit +
    -68.5107

Time taken to build model: 0 seconds

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

Correlation coefficient                  0.396
Mean absolute error                      0.5286
Root mean squared error                  0.6639
Relative absolute error                106.4149 %
Root relative squared error             91.4166 %
Total Number of Instances              130
```

Before

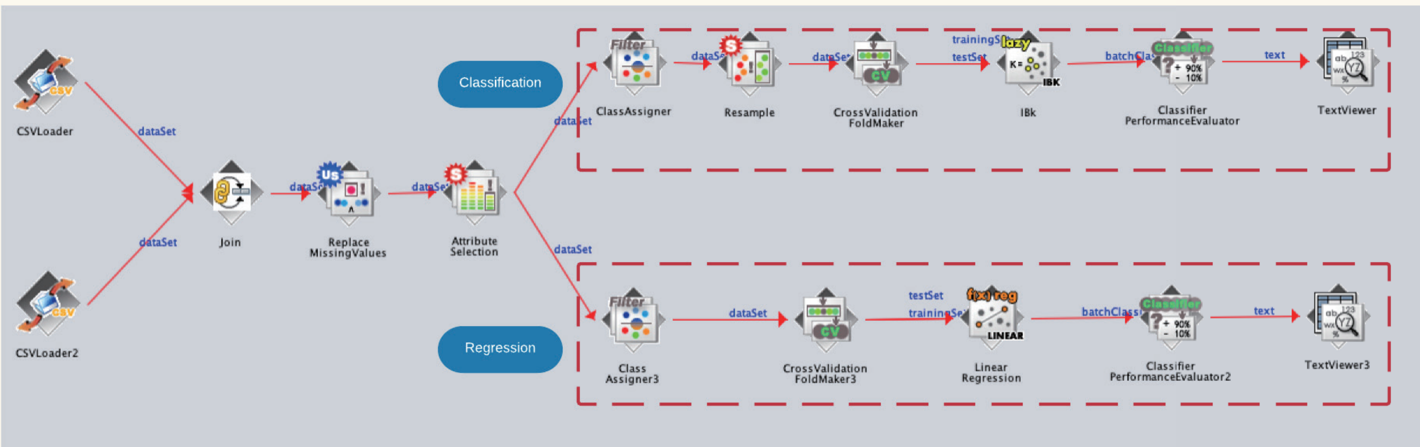
```
=== Classifier model (full training set) ===


Linear Regression Model

landings_from_stocks_above_soft_limit =

      0.0648 * year +
      0      * catchment -tonnes +
     -0      * TACC - tonnes +
      0.3426 * landings_from_stocks_above_hard_limit +
    -68.5585

Time taken to build model: 0 seconds

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

Correlation coefficient                  0.4165
Mean absolute error                      0.5236
Root mean squared error                  0.6552
Relative absolute error                105.4045 %
Root relative squared error             90.2178 %
Total Number of Instances              130
```

After

Even though the correlation is not that strong in this case, we can still see the increasing performance after the attributes were removed. This also helps us investigate the importance of the attributes.

Modeling(Part 2)

Pipeline simulation:



After merging the values, I also used attributes selection to decide what attributes should stay in the dataset. The result also implemented as a subset of the merged data.

Classification:

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

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

Correctly Classified Instances         129               99.2308 %
Incorrectly Classified Instances         1                0.7692 %
Kappa statistic                          0.9781
Mean absolute error                      0.016
Root mean squared error                  0.0874
Relative absolute error                  4.4665 %
Root relative squared error             20.7359 %
Total Number of Instances              130

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.967    0.000    1.000      0.967    0.983      0.978    0.983     0.974     '(-inf-95.4]'
                 1.000    0.033    0.990      1.000    0.995      0.978    0.983     0.990     '(95.4-inf)'
Weighted Avg.    0.992    0.026    0.992      0.992    0.992      0.978    0.983     0.986

=== Confusion Matrix ===

   a   b   <-- classified as
  29   1 |   a = '(-inf-95.4]'
   0 100 |   b = '(95.4-inf)'
```

IBK received an excellent performance on this dataset. However, finding relationships between two attributes and predicting the future trend is my priority. Thus, we are not going to discuss this result here.

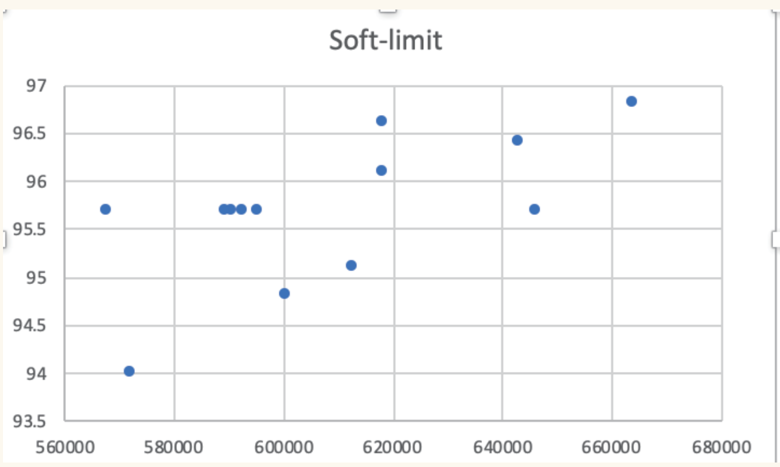
Important features:

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


Linear Regression Model

landings_from_stocks_above_soft_limit =

      0.0648 * year +
      0      * catchment -tonnes +
     -0      * TACC - tonnes +
      0.3426 * landings_from_stocks_above_hard_limit +
    -68.5585

Time taken to build model: 0 seconds

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

Correlation coefficient                  0.4165
Mean absolute error                      0.5236
Root mean squared error                  0.6552
Relative absolute error                105.4045 %
Root relative squared error             90.2178 %
Total Number of Instances              130
```

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

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 landings_from_stocks_above_soft_limit):
	Correlation Ranking Filter
Ranked attributes:
 0.4981  6 landings_from_stocks_above_hard_limit
 0.2928  1 year
 0.0911  2 asset value - $.M
 0.0289  4 TACC - tonnes
 0.0219  3 catchment -tonnes
```

Based on the observation, the Both regression analysis and weight analysis show TACC and soft limt are not highly related. The diagram on the left indicate that the statistics speard out and we don't have enough information to say it is a strong pattern. The regression analysis shows

Evaluation



Both regression analysis and weight analysis show TACC and soft limit are not highly related. The diagram on the left indicates that the statistics spread out and we don't have enough information to say it is a strong pattern. The regression analysis shows 0.41 correlation coefficient which is lower than 0.5. The weighted analysis also tells us TACC and catchment have zero influence on the soft limit. These pieces of evidence prove that soft limit and hard limit may only show the fish stocks. The government may use other techniques to determine fishing protection policies.

Appendix

Part one dataset:

https://www.stats.govt.nz/assets/Uploads/Environmental-economic-accounts-2019/Download-data/fish-monetary-stock-account-1996-2018.csv

Part two dataset:

https://data.mfe.govt.nz/table/53467-performance-of-assessed-fish-stock-in-relation-to-the-soft-limit-200915/

https://data.mfe.govt.nz/table/53469-performance-of-assessed-fish-stock-in-relation-to-the-hard-limit-200915/