COMP309: Real World Data Handling, Modelling and Visualisation

Dominic Sonneveld – 300439310

12-08-19

“Whatever happens, over-fishing today will lead to collapses tomorrow”, an interesting and potentially daunting thought for the wellbeing of New Zealand marine life and the recreational fishermen of our country. Being a country surrounded by water, fishing is an integral part of who we are as Kiwi’s and in a cut throat effort to make money on things we take for granted the wellbeing of our marine life has been put on hold.

In an effort to outline the issues of over fishing in New Zealand the CRISP-DM process was undertaken on a set of data, NIWA Freshwater Fish Sites, retrieved from (Council, 2016). This dataset comprises of information regarding fishing stocks in the Greater Wellington region from 1922 to 2012. Notable attributes included in the dataset include, X and Y coordinates of each instance, the year and month the instance was recorded, the type of fish and its ID and the location it was caught. This dataset is appropriate for investigating the problem statement as it spans over 90 years containing a great number of instances allowing us to accurately classify and determine specific fishing trends that have occurred over the years and it allows us to make hypothesis’ about what this might mean for the future of fishing in New Zealand.

While the dataset is very useful due to its size and how long a timespan the data was recorded there were some impurities which had to be dealt with in order to apply proper machine learning techniques and algorithms such as bayesian networks and K – Nearest Neighbours without running into issues. The first issue I encountered was that there was some missing data in the number of fishes attribute. The missing data seemed to be random and was in no specific time frame and was not from any specific location or any one type of fish, so due to the importance of this attribute and the fact that there were so many other instances in the dataset, these impure instances were removed. These instances were removed as opposed to using other imputation techniques as to not ‘infect’ the rest of the dataset with some potentially false values. This did not pose a problem and the great number of instances in the dataset was still sufficient to show viable trends. Another issue about the dataset was some irrelevant information with regards to the problem statement. This data includes the X, Y coordinates of where the fish was caught and the object ID. These attributes were removed from the dataset as to increase efficiency and simplicity when analysing the data. The location the fish being caught was also removed despite the potentially interesting conclusions that could be made with this knowledge such as which body of water contained what types of fish and how the population of these fish may have been affected over the years. This attribute was removed purely for efficiency sake as the great number of streams, rivers, lakes, estuaries etcetera posed a problem for the WEKA software and the machine learning algorithms to deal with and resulted in any analysis taking far too long to worthwhile. In addition to these impurities there was also some data

that contained physically wrong values, like having a year or month as zero. There were only a small number of these instances so they were removed. The final adaptation to the dataset was reducing the number of different types of fish. The original dataset had over 50 different types of fish, this number was reduced by joining similar types of fish, for example, longfin eel, shortfin eel and common eel all simply became 'eel', and so forth with other types of common fish. This was done because when analysing the dataset, the names of the fish was the class and having that many classes resulted in some very slow analysing. After attempting to reduce the number of types of fish, there were still 27 which is still a large number of classes to analyse.

To begin analysing the adapted dataset was imported into the WEKA software where a visualisation of the entire data can be seen in **figure 1**.

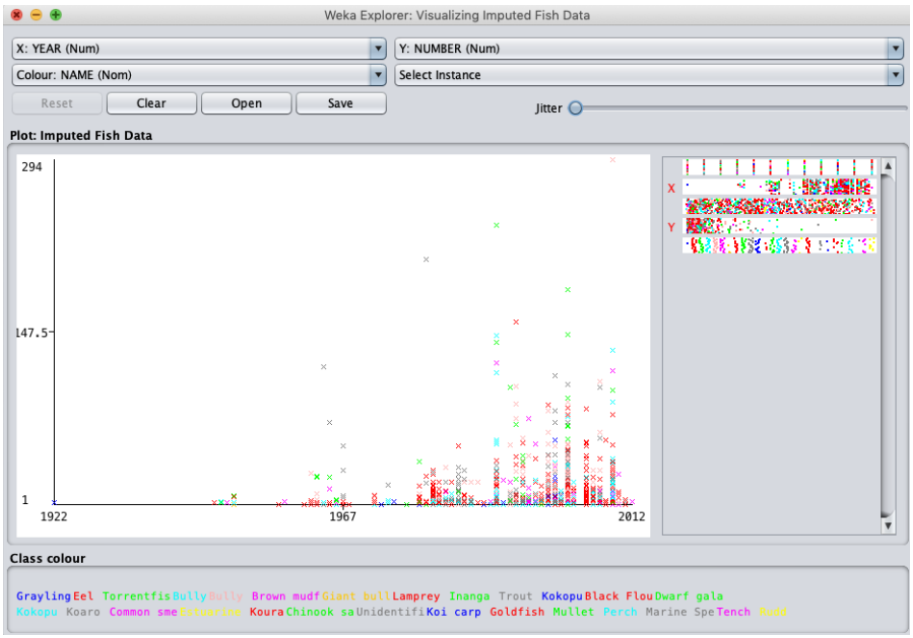

Figure 1: Number of fish caught in the Greater Wellington region over the years

As can be seen from the figure the number of fish caught in the Greater Wellington region has increased dramatically over the 90 year period this data was collected. This shows an expected trend as the number of fish caught has increased along with commercialised fishing in New Zealand. The figure also helps prove the problem statement, "Whatever happens, over-fishing today will lead to collapses tomorrow", as can be seen at the most recent years 2009-2012 in the dataset where the number of fish caught is significantly smaller than previous years like 70's to the early 2000's where there seemed to be an abundance fish caught. The cause of this may be increased exposure to over fishing resulting in fishermen limiting what they catch or it could simply imply that fish stocks in these areas have been so depleted there are not enough fish to catch.

Another interesting plot that can be shown from the dataset is the number of fish caught per month of the year as can be seen in **figure 2**.

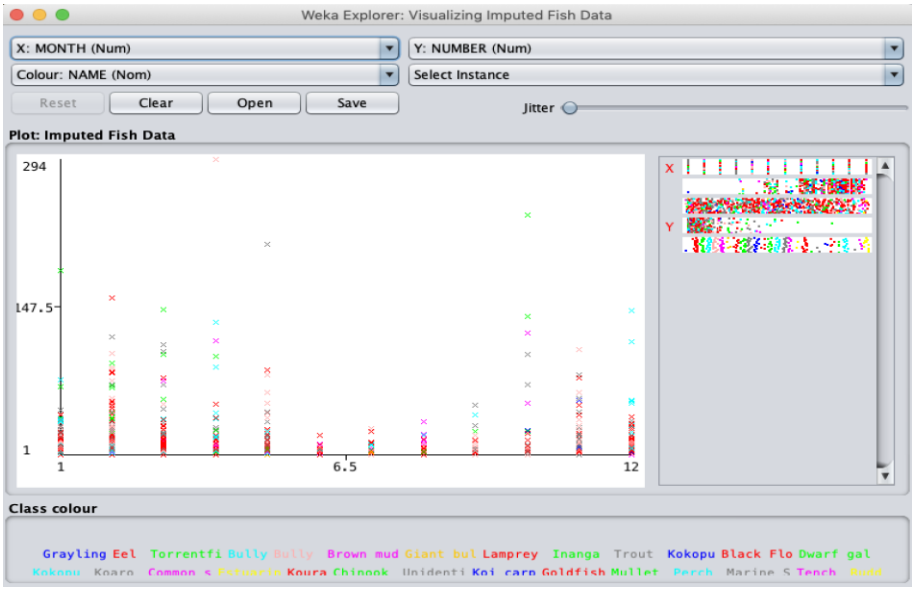

Figure 2: Number of fish caught in each month of the year

This plot shows the greater number of fish caught in the warm summer months compared to the cold winter months which is an expected trend when it comes to the sport of fishing and the New Zealand climate. While this does not help toward the problem statement it is still interesting to note and it validates the realism of the dataset.

A pipeline was created to apply machine learning algorithms to classify the data. The pipeline is shown in **figure 3**.

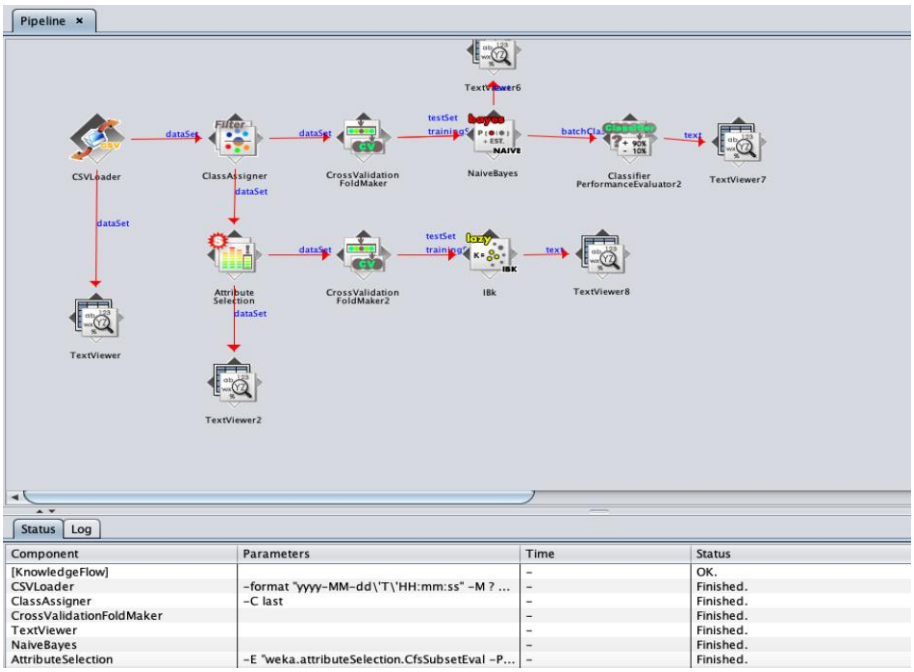

Figure 3: Pipeline to analyse dataset

The pipeline begins with a CSV loader to import the dataset as a CSV file exported from excel where all the data imputation took place. The text viewer is there to view the dataset and enable the user to ensure the data in the pipeline is what they expect. The CSV loader leads to a class assigner which assigns the class of the dataset to be the last attribute which is the name of the fish. This pipeline is then split with one path leading straight to a cross validation fold maker which creates stratified cross-validation folds from the incoming data and the other path leads to an attribute selector and then to a cross validation fold maker. The first path leads to a BayesNet classifier and the second to an IBK classifier which is a WEKA algorithm that performs a K – Nearest Neighbours type analysis. Both of these paths end with text viewers to view the classifying results of the machine learning algorithms.

The results of the BayesNet analysis are shown in **figure 4**.

```
=== Evaluation result ===

Scheme: BayesNet
Options: -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Relation: Imputed Fish Data 2-weka.filters.unsupervised.attribute.ClassAssigner-Clast


Correctly Classified Instances         489               30.9102 %
Incorrectly Classified Instances      1093               69.0898 %
Kappa statistic                          0
Mean absolute error                      0.0634
Root mean squared error                  0.1779
Relative absolute error                 99.8836 %
Root relative squared error             99.9984 %
Total Number of Instances             1582

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    0.000    ?          0.000    ?          ?        0.199     0.002     Giant bully
               0.000    0.000    ?          0.000    ?          ?        0.463     0.013     Lamprey
               1.000    1.000    0.309      1.000    0.472      ?        0.498     0.308     Eel
               0.000    0.000    ?          0.000    ?          ?        0.477     0.028     Inanga
               0.000    0.000    ?          0.000    ?          ?        0.497     0.101     Trout
               0.000    0.000    ?          0.000    ?          ?        0.499     0.133     Bully
               0.000    0.000    ?          0.000    ?          ?        0.499     0.019     Kokopu
               0.000    0.000    ?          0.000    ?          ?        0.490     0.083     Kokopu
               0.000    0.000    ?          0.000    ?          ?        0.487     0.065     Koaro
               0.000    0.000    ?          0.000    ?          ?        0.476     0.011     Common smelt
               0.000    0.000    ?          0.000    ?          ?        0.468     0.020     Torrentfish
               0.000    0.000    ?          0.000    ?          ?        0.150     0.002     Estuarine triplefin
               0.000    0.000    ?          0.000    ?          ?        0.491     0.063     Koura
               0.000    0.000    ?          0.000    ?          ?        0.299     0.003     Black Flounder
               0.000    0.000    ?          0.000    ?          ?        0.050     0.001     Chinook salmon
               0.000    0.000    ?          0.000    ?          ?        0.489     0.048     Bully
               0.000    0.000    ?          0.000    ?          ?        0.099     0.001     Koi carp
               0.000    0.000    ?          0.000    ?          ?        0.199     0.003     Mullet
               0.000    0.000    ?          0.000    ?          ?        0.473     0.027     Brown mudfish
               0.000    0.000    ?          0.000    ?          ?        0.453     0.013     Dwarf galaxias
               0.000    0.000    ?          0.000    ?          ?        0.499     0.013     Perch
               0.000    0.000    ?          0.000    ?          ?        0.399     0.004     Tench
               0.000    0.000    ?          0.000    ?          ?        0.499     0.006     Goldfish
               0.000    0.000    ?          0.000    ?          ?        0.424     0.009     Rudd
               0.000    0.000    ?          0.000    ?          ?        0.050     0.001     Grayling
               0.000    0.000    ?          0.000    ?          ?        0.299     0.003     Unidentified galaxiid
               0.000    0.000    ?          0.000    ?          ?        0.050     0.001     Marine Species
Weighted Avg.  0.309    0.309    ?          0.309    ?          ?        0.486     0.144
```

Figure 4: BayesNet analysis on dataset

A bayesian network is part of the bayesian tribe of AI and is a probabilistic graphical model which shows the probability that a certain known cause was the reason for a specific event (Soni, 2018). This was chosen to classify the dataset as it does well when multiple attributes can be linked to one another. In this dataset the number of fish, the type of fish, the location of the fish and the month the fish was caught are all attributes that can be related to each other so using a bayesian network can help with regards to accurately classifying a dataset. The BayesNet classifier correctly classified 30.9102% of test cases. This is not a high percent of correctly classified test cases however it is significantly higher than it would have been if some unnecessary attributes like location or object ID remained in the dataset or if the types of fish had not been reduced. The reason for such low classification was due to the high number of classes and small number of instances in relation to those classes. For

example, some classes such as Greyling and Chanook Salmon only had one instance in the whole dataset where this type of fish was caught.

The results of the IBK analysis are shown in **figure 5**.

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch
Relation:     Imputed Fish Data 2
Instances:    1582
Attributes:   4
               MONTH
              YEAR
              NUMBER
              NAME
Test mode:    5-fold cross-validation

=== 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         437               27.6233 %
Incorrectly Classified Instances      1145               72.3767 %
Kappa statistic                          0.1127
Mean absolute error                      0.0572
Root mean squared error                  0.2104
Relative absolute error                 90.1702 %
Root relative squared error            118.2195 %
Total Number of Instances             1582
```

Figure 5: IBK analysis on dataset

The IBK algorithm is one of WEKA's versions of a K – Nearest Neighbours algorithm. This works by taking every new piece of data in a dataset and comparing it to all other prior pieces of data and based on this comparison the algorithm determines what class this piece of data is. The IBK analysis resulted in 27.6233% of test cases being correctly classified. Once again this is quite a small number of correctly classified instances, once again due to previously mentioned problems with the dataset.

Based on this dataset some interesting questions can be asked about the current state and the future state of fishing in New Zealand. Is the reason for the decrease in fish caught in recent years due to public understanding and cooperation with regards to overfishing or is it the case that the fish stocks have been so depleted over the years that there is no coming back?

Bibliography

Council, G. W. (2016, May 2). *NIWA Freshwater Fish Sites*. Retrieved from data.govt.nz: https://catalogue.data.govt.nz/dataset/niwa-freshwater-fish-sites1

Soni, D. (2018, June 8). *Introduction to Bayesian Networks* . Retrieved from towardsdatascience : https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e