Assignment #2

Real-World Data Handling, Modelling and Visualisation

COMP 309 Machine Learning Tools and Techniques

Qiangqiang Li(Aaron)

ID: 300422249

Part 1: Core: Evidence related to fish stocks in New Zealand

Business Objectives: In order to analyze fish-related information, we use machine learning tools to analyze open relation databases, and illustrate the most important aspects of NZ fisheries.

Data Mining Goal:Compare different techniques to analyze the open databases, and find a best method to , and find the main reason for the fishing over-fishing.

Produce Project Plan:Find some useful databases, through different techniques to analyze these databases, and obtain the important factor for the issue.

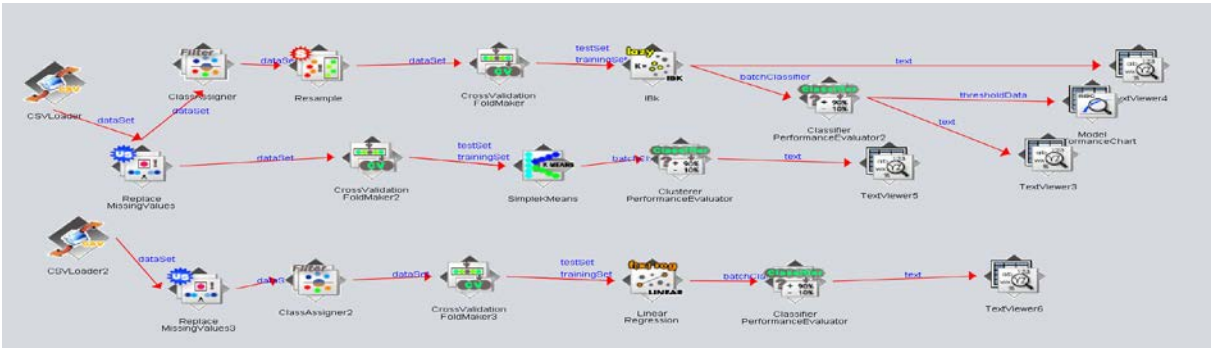
1. Reply :

Data Information: **NIWA_Freshwater_Fish_Sites.csv.** A range of Environmental datasets for the GW Region, including fish distribution, ecological sites, land cover, geology, catchments, and contours. ***Website:*** "https://catalogue.data.govt.nz/dataset/niwa-freshwater-fish-sites1".

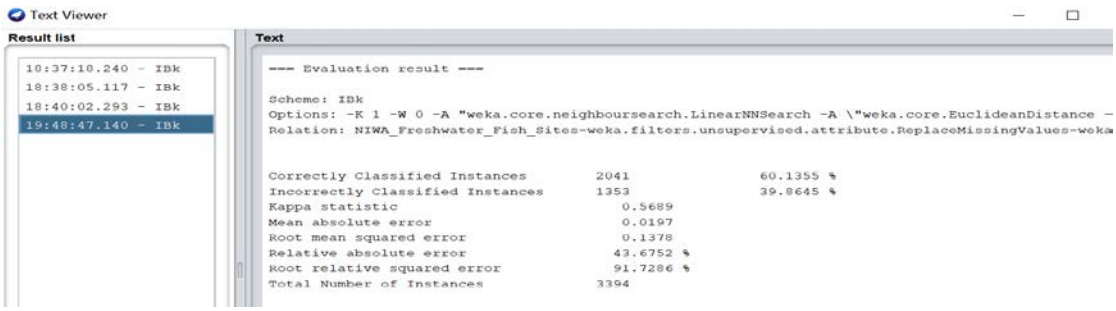
Classification: is a process related to categorization, the process in which ideas and objects are recognized, differentiated and understood. ***In this case, the fish data includes many*** fish distribution **LOCALITY** and fish **NAME**, the classifier could to use the fish name as **label**, it will find the different area have the number of the fish. This using **KNN(IBK)** technique to show the result.

Clustering: is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). This data includes many information about the fish data and not to know the label, So this using **K-means** technique to show the result, this can find the same place have the number of fish data.

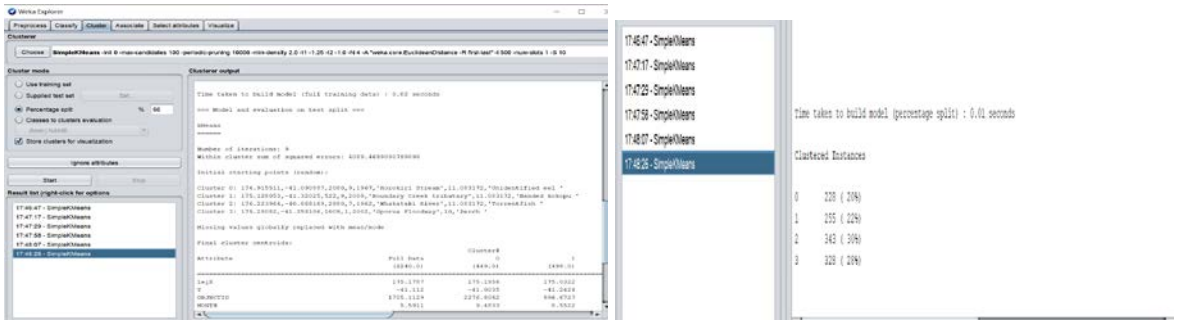
Regression: is a set of statistical processes for estimating the relationships among variables. This is a good way to find the relation between the different attributes. This using Linear Regression technique to show the result. this is a good method to find the relation between these attributes, and find the important attributes whether will affect the number of fish.



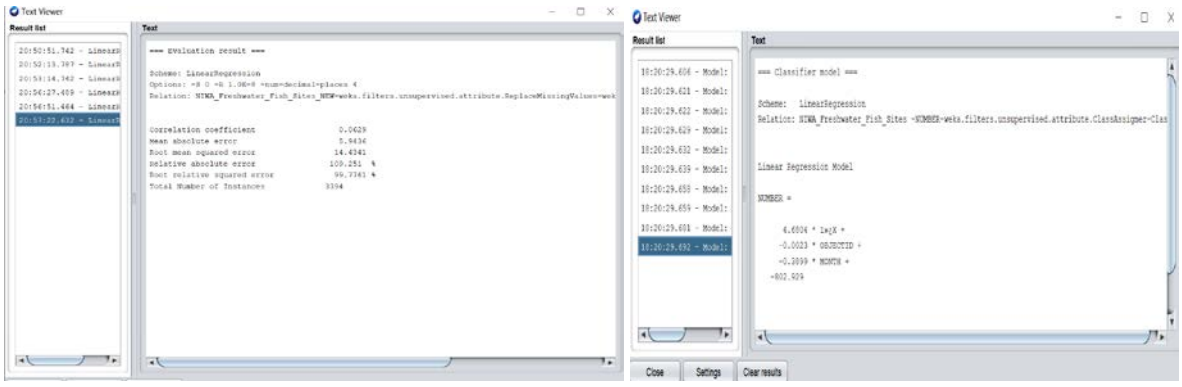
2. Describe the results of each technique used on the one dataset. Note the most appropriate form of results may differ between each technique. Exercise your skill and judgement to decide how the results should be communicated.



KNN: the Correctly Classified Instances is 60.1355%, it is not high to classifier. Because the data have too many missing data, and the data includes the location and number, the fish name. these data not to clear to classifier the number of the different fish in the same fish distribution.



K-means: not to need the label, and select K value will effect the result: Like k=4, the result was show 4 cluster, the data will show the same group. This result can not show the location relation to the number of fish. **So** it not suit for this data.



Linear Regression: These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. In this case, the correlation coefficient is too low, only at 6.295%. Because this data has many missing data and the real data may be not very well. So the data was presented the low correlation coefficient. But this method usually use to analyze the relationship in these variables.

- ***3. Identify how these aspects of the techniques are different, e.g. how do the results from clustering differ from classification techniques. Please refer to the dataset when describing the differences in the techniques.***

Reply: The ***differences*** is that classification is used in supervised learning technique where predefined labels are assigned to instances by properties, on the contrary, clustering is used in unsupervised learning where similar instances are grouped, based on their features or properties. Like this case, the data includes **LOCALITY information** and fish **NAME**, the classifier using the **fish name** as label can analyze how many fish in differences distribution. But using the ***Regression*** need to transform the data type, it is find the relation of varies attributes, in this case, the data is too small and independent, it is difficult to show a good result.

On the other hand, the similarity between two objects is measured by the **similarity function** where the distance between those two object is measured. Shorter the distance higher the similarity, conversely longer the distance higher the dissimilarity. In this case, through data analysis of the number of fish in the same group, it can not show the location relation to the number of fish.

- ***4.Please revisit the business understanding based on your exploration of the data. It is noted that a simple question to ask is "is there any evidence of fish stocks collapsing in NZ waters?". Please create and describe two other questions that could be asked of the data as well.***

Reply: **NIWA_Freshwater_Fish_Sites is** a range of Environmental datasets for the GW Region, including fish distribution, ecological sites, land cover, geology, catchments, and contours. This data might use the location site and the number of catchments will to represent the problem in fish stocks collapsing in NZ water. In these results, the change of number didn't relate with the location. According to the dataset, this fishing data has not enough data to show the real-world issue, the data was record from few fish distribution, the data includes many not useful data, and may be effect from other factors. So, the next two question is:

1. Is there the economic effect fish stocks collapsing in NZ?
2. Is there the environment effect fish stocks collapsing in NZ?

These questions will improve the more information to analyze the trend of the fishing in New Zealand.

Part 2: Completion: Feature importance to Fish stocks in New Zealand

- ***1. Reply:***

The one question is "what features can be considered as the underlying cause for any increase(or decrease) in fish numbers?". Fish stocks may be affected by the environment and the economy, so we can use more data to analyze which characteristics are related to this information.

My idea of the whole analysis is to find a way to relate "fish" with "commercial". In order to do this, not only the plain fishing commercial datasets are needed, there could potentially exist more factors that pose impact on the Environmental, which is why more datasets are needed. I selected other 2 datasets:

Fish-stocks-meeting: Our fish stocks are affected by commercial, customary, and recreational fishing, and environmental pressures (eg ocean temperature, acidity, and productivity). The Ministry for Primary Industries uses three performance measures to assess influences on fish stocks: **a soft limit** (below which a rebuilding plan is required), **a hard limit** (below which closing a fishery should be considered), and **an overfishing threshold** (where the rate of extraction is higher than the of

replenishment).**Website**:https://data.mfe.govt.nz/table/52514-landings-from-stocks-meeting-or-exceeding-performance-thresholds-200914/data/.

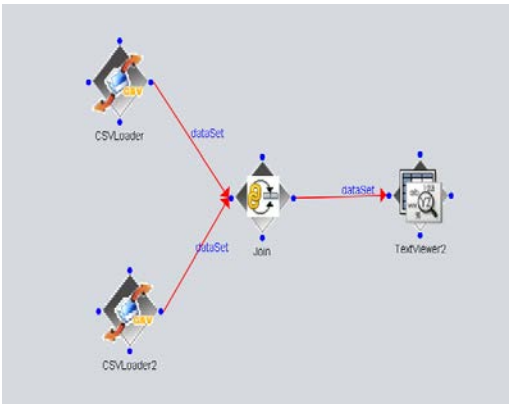
Fish-monetary-stock-account: Environment-economic accounts show how our environment contributes to **our economy**, **the impacts of economic activity on our environment**, and how we respond to environmental issues. **Website**: https://catalogue.data.govt.nz/dataset/environmental-economic-accounts-2019-tables/resource/9c0fdf15-1d92-4163-a32c-0e2f98e75b76.

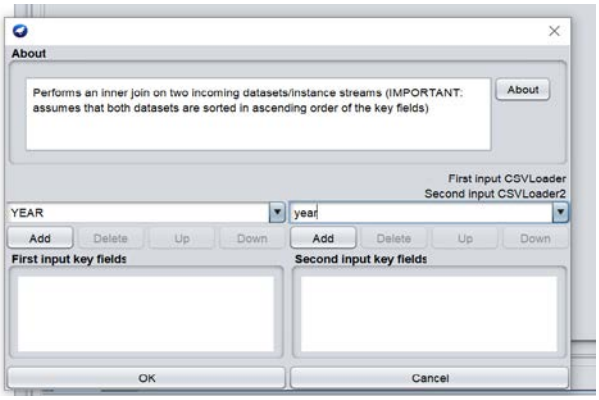
2. **Reply: this merging steps:**

   1. *Change the database to the Suitable database:* Use the **Excel's Transpose** function to change the databases and the year use the rang form 1996 to 2012. The three databases have the same year, and easy to *merging;*

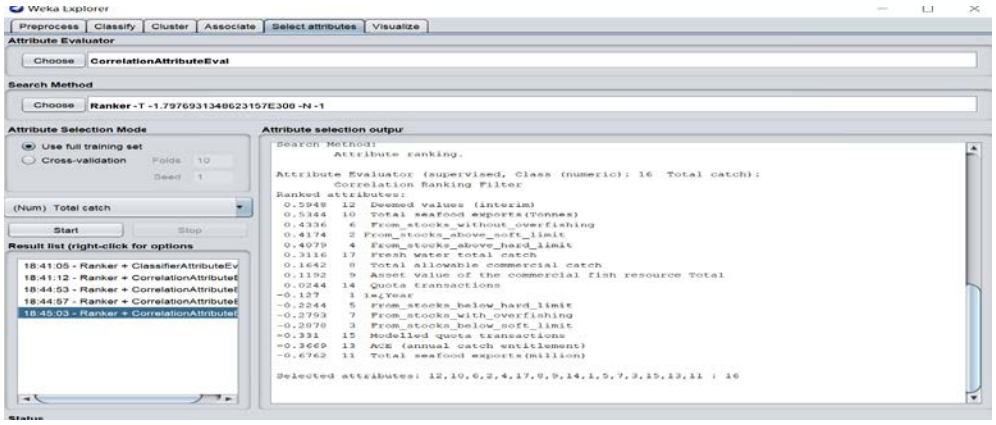
- **NIWA_Freshwater_Fish_Sites:** use **Excel** tool to sum **the number of catch fish** in each year.
- **Fish-monetary-stock-account:** find the different tables to merge by **each year. The Attributes** include TACC, Asset Value of CFRT, seafood exports(Tonnes), seafood exports(million), Deemed values, ACE, Quota transactions, Modelled quota transactions ,Total catch. Different Attributes from the different tables.
- **Fish-stocks-meeting:** This dataset relates to the "State of fish stocks" measure on the Environmental Indicators**. It Includes** a soft limit, a hard limit and an overfishing threshold.

   2. *Using the **Weka knowledge Flow** to merge the different table, using the **Join*** function to merge the two dataset.





3. **Reply: Using the Weka Select attributes, select the important attributes to show the data information.**



This can selected the attributes: Year, From_stocks_above_soft_limit, From_stocks_above_hard_limit, From_stocks_without_overfishing, Total seafood exports(Tonnes), Deemed values (interim), Total catch. Remove 10 ***attributes. Like the below data:***

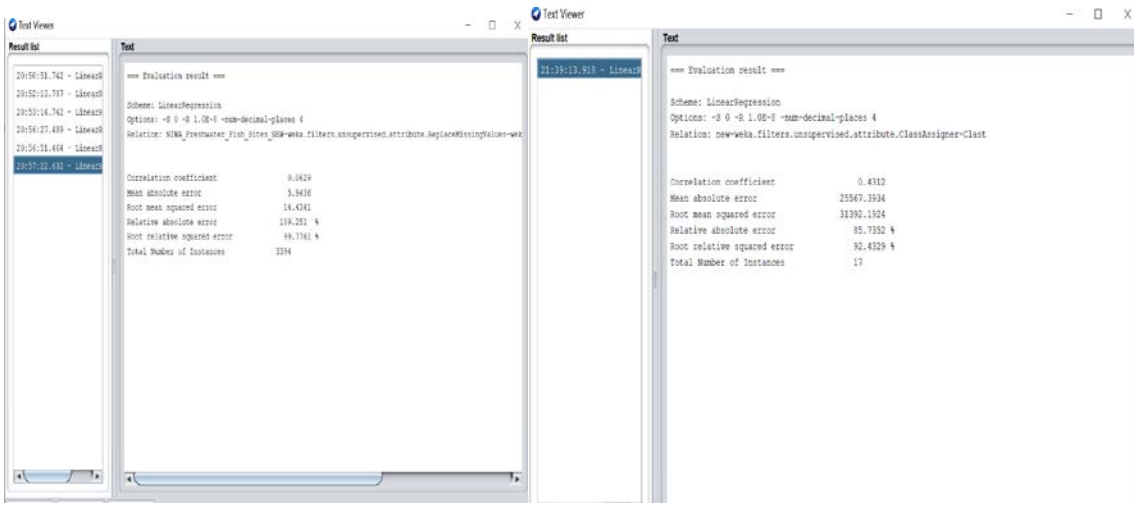
```
@relation 'fishing_stock-weka.filters.unsupervised.attribute.Remove-R3,5,7-9,11,13-15,17'

@attribute Year numeric
@attribute From_stocks_above_soft_limit numeric
@attribute ' From_stocks_above_hard_limit' numeric
@attribute ' From_stocks_without_overfishing' numeric
@attribute ' Total seafood exports(Tonnes)' numeric
@attribute ' Deemed values (interim)' numeric
@attribute ' Total catch' numeric

@data
1996,240700,266000,225800,441255.65,337.9,413326.8
1997,240700,266000,225800,441255.65,337.9,413326.8
1998,240700,266000,225800,441255.65,337.9,413326.8
1999,240700,266000,225800,441255.65,337.9,413326.8
2000,240700,266000,225800,441255.65,337.9,413326.8
2001,240700,266000,225800,441255.65,337.9,413326.8
2002,240700,266000,225800,441255.65,337.9,475146.3
2003,240700,266000,225800,443875,337.9,495284.3
2004,240700,266000,225800,466825.7,337.9,475901.5
2005,240700,266000,225800,470912.3,337.9,469490.8
```

Handle the missing data, using the average value add the table.

Year	From_stocl	From_stocl	From_stocl	Total seafc	Deemed v	Total catch
1996	240700	266000	225800	441255.7	337.9	413326.8
1997	240700	266000	225800	441255.7	337.9	413326.8
1998	240700	266000	225800	441255.7	337.9	413326.8
1999	240700	266000	225800	441255.7	337.9	413326.8
2000	240700	266000	225800	441255.7	337.9	413326.8
2001	240700	266000	225800	441255.7	337.9	413326.8
2002	240700	266000	225800	441255.7	337.9	475146.3
2003	240700	266000	225800	443875	337.9	495284.3
2004	240700	266000	225800	466825.7	337.9	475901.5
2005	240700	266000	225800	470912.3	337.9	469490.8
2006	240700	266000	225800	474175	337.9	452408.8
2007	240700	266000	225800	474163.7	337.9	441711.1
2008	240700	266000	225800	456644.2	202.3	401016.4
2009	222700	243700	159400	419309.4	190.7	384859.5
2010	228300	251200	162700	438636.3	231.7	398244.2
2011	234000	256400	222700	449684.7	258.8	403719.4
2012	247400	275600	228900	437368.5	254.8	413326.8

- 4. Now only a subset of the original features and instances should exist. Analyse the output of the ML tool on this processed dataset to identify which features are important to a selected tools' output, e.g. which features are near the start of the decision tree for classification, or which features have the highest weights in regression and so fourth.

  Reply: Linear Regression: the **original features:6.29%, the new features:43.12%. So the new** features showed the better trend of the fishing.



Total catch = 4022.411*Year +630.6706 * Deemed values (interim) +(-7824680.1868)

```
Linear Regression Model

Total catch =

    4022.411  *  Year +
     630.6706 * Deemed values (interim) +
-7824680.1868

Time taken to build model: 0 seconds

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

Correlation coefficient                  0.4312
Mean absolute error                  25567.3934
Root mean squared error              31392.1924
Relative absolute error                 85.7352 %
Root relative squared error             92.4329 %
Total Number of Instances               17
```

IN this Linear Regression Model, it is clear that the number of catch fish is relation year and Deemed Values. So Deemed Values is the highest weights in regression.