Comp309 Assignment 2

Source of the dataset:

My dataset is Environmental-economic accounts.I get the dataset from this link in data.govt.nz.(<https://catalogue.data.govt.nz/dataset/environmental-economic-accounts-2019-tables/resource/9c0fdf15-1d92-4163-a32c-0e2f98e75b76>)

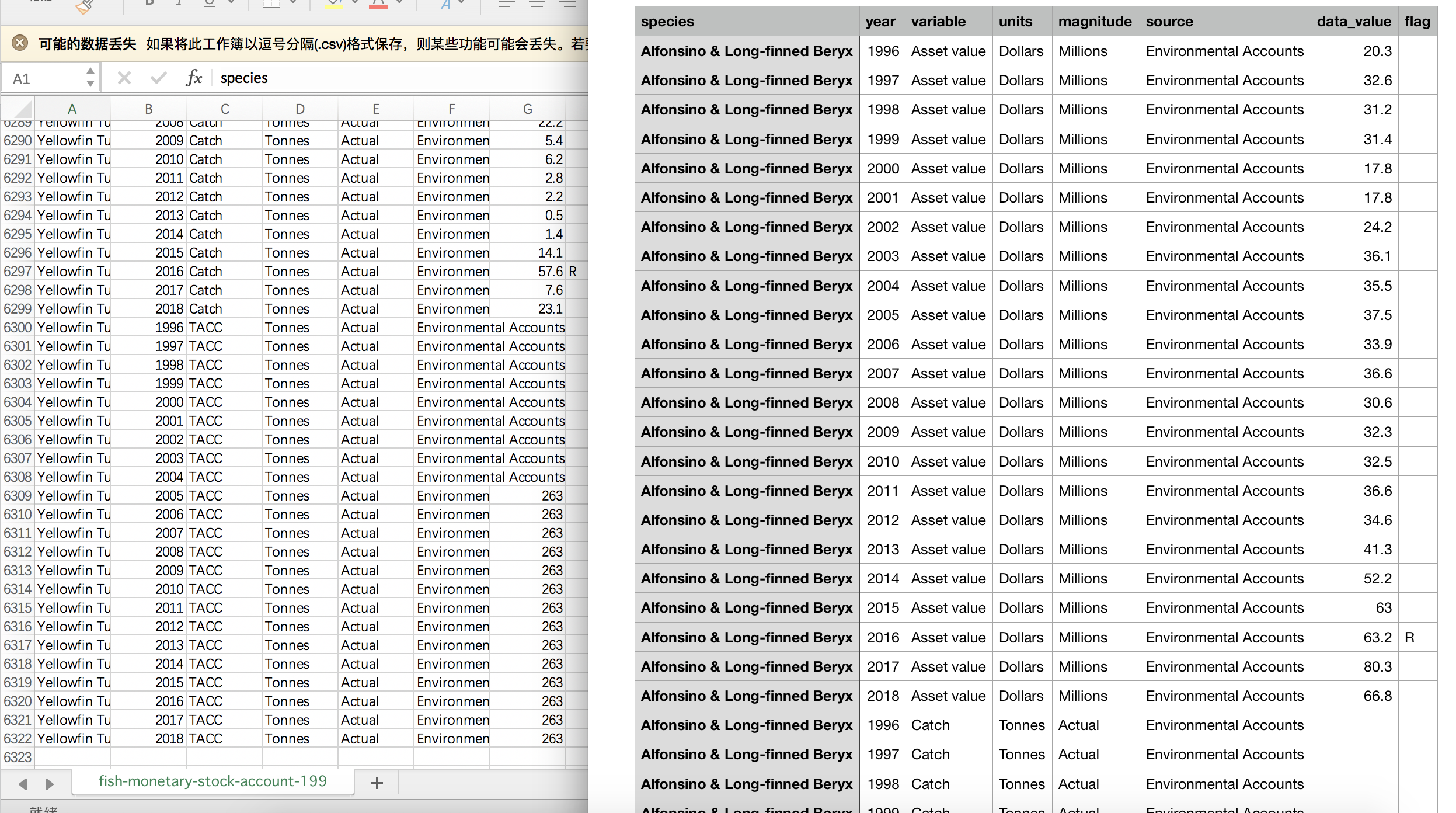
I chose Fish monetary stock account, 1996–2018–CSV in a series of datasets.

Description of the dataset :

This data set shows the catch and fishing trends of some fish from 1996 to 2018. And the annual government limits on the fishing of various fish species, as well as the benefits from these fish.

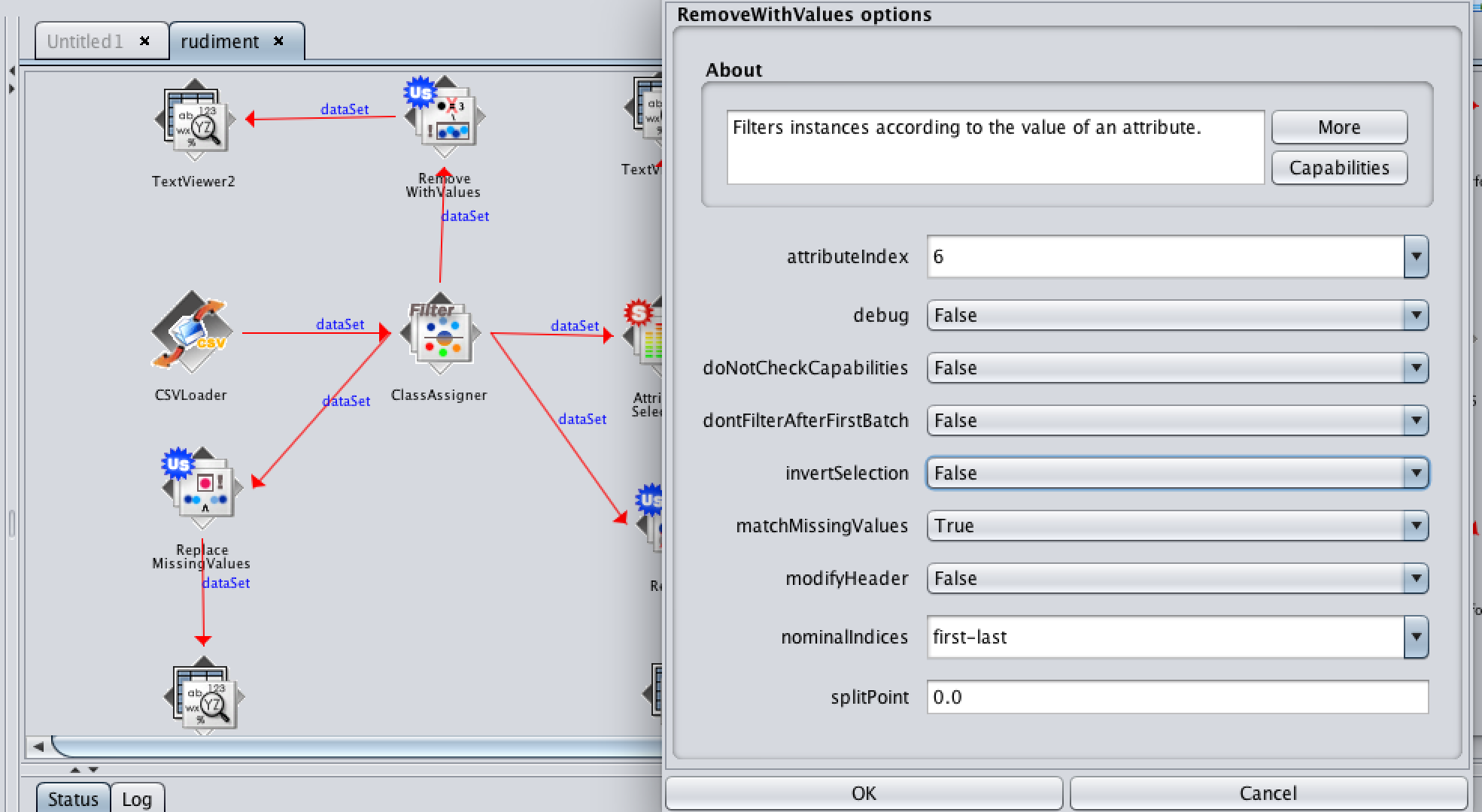
Environmental-economic accounts show how our environment contributes to our economy, the impacts of economic activity on our environment, and how we respond to environmental issues.

Part 1 : Core

When I open the dataset I find a lot of missing values and the dataset is too huge(6322 rows) for my analysis purposes. and complicated (too much species)which may mask the important features I want to figure out.So the first thing after loading the dataset is to filter my data.

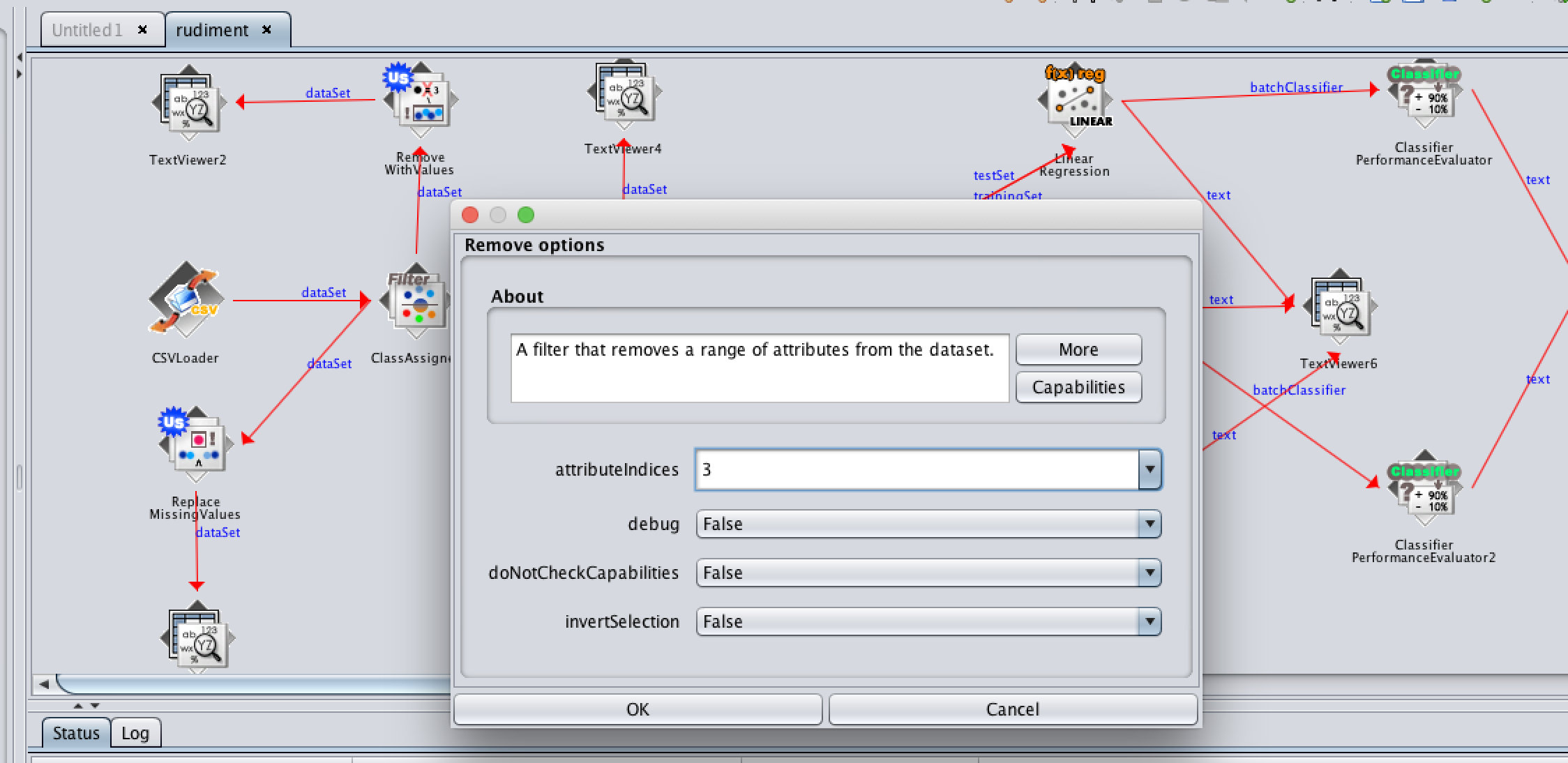
And this dataset has a lot of flaws：

1. For missing value ,I choose to remove rows with missing value( data\_value is the only attribute may have missing value ,its atrributeIndex is 6) because the number of data is enough.

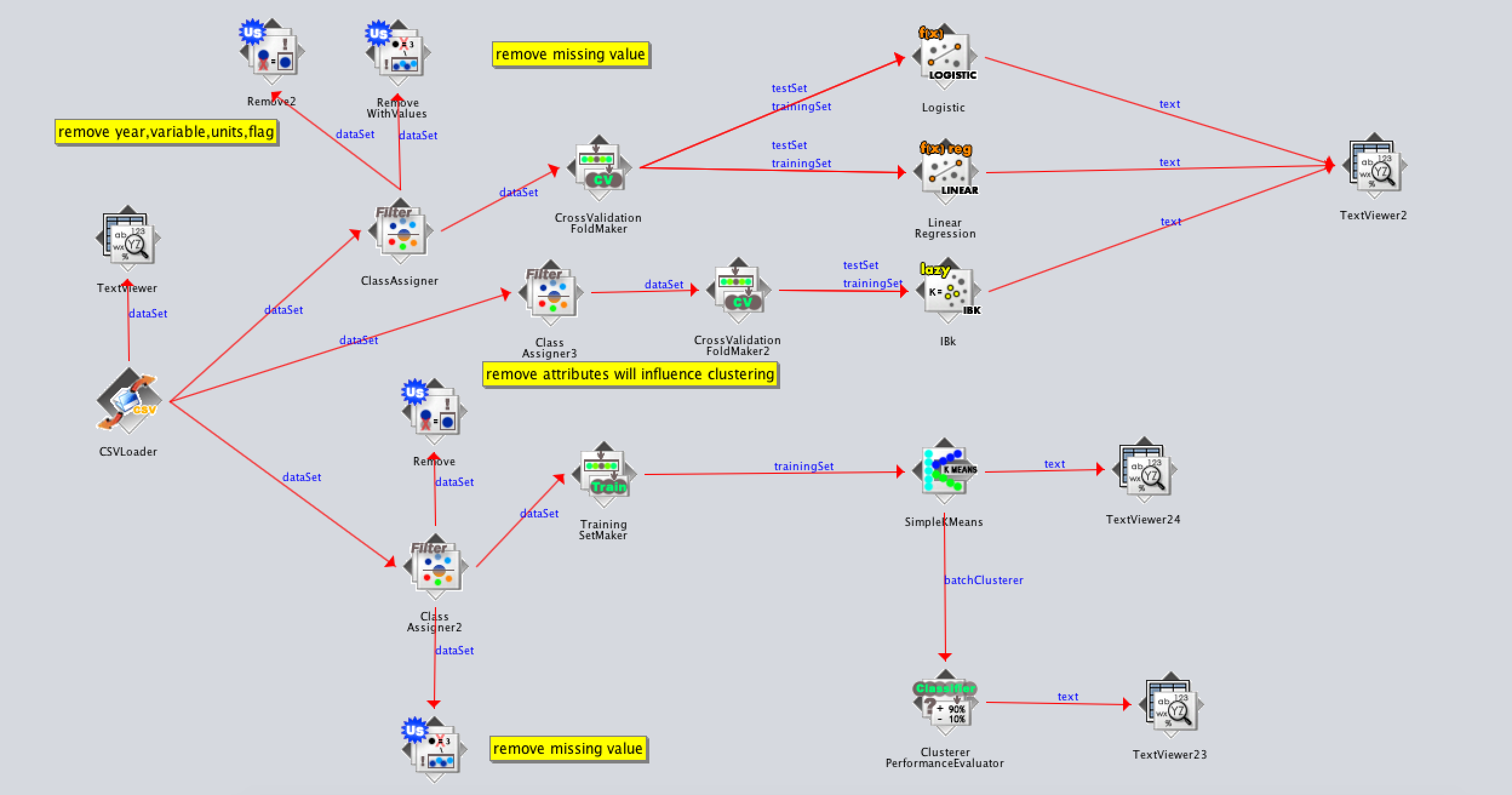
RemoveWithValues in Pipeline.

1. Remove the nominal attributes with only one variable when I do unsupervised learning. (When I did linear regression)This kind of data have no influence to the result so I remove them.

Remove.

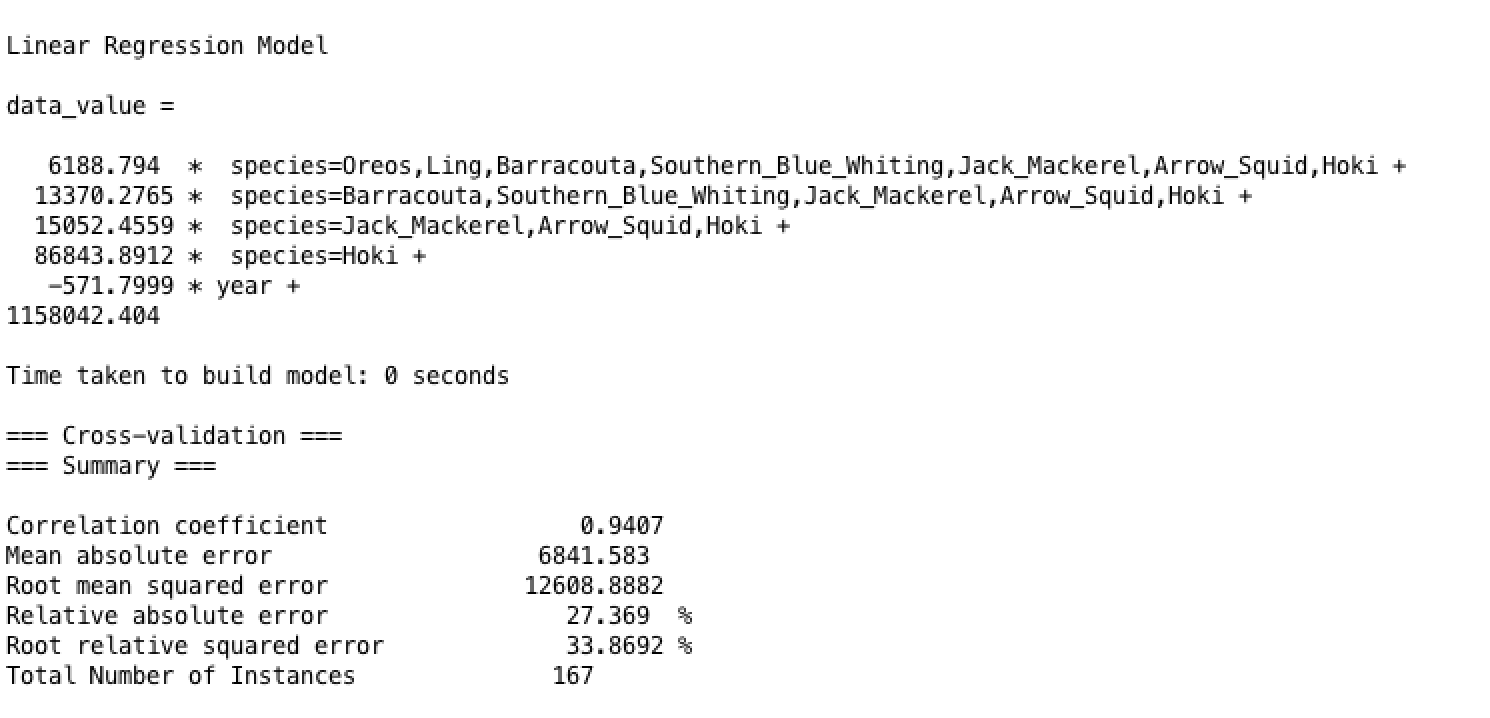


Pipeline:



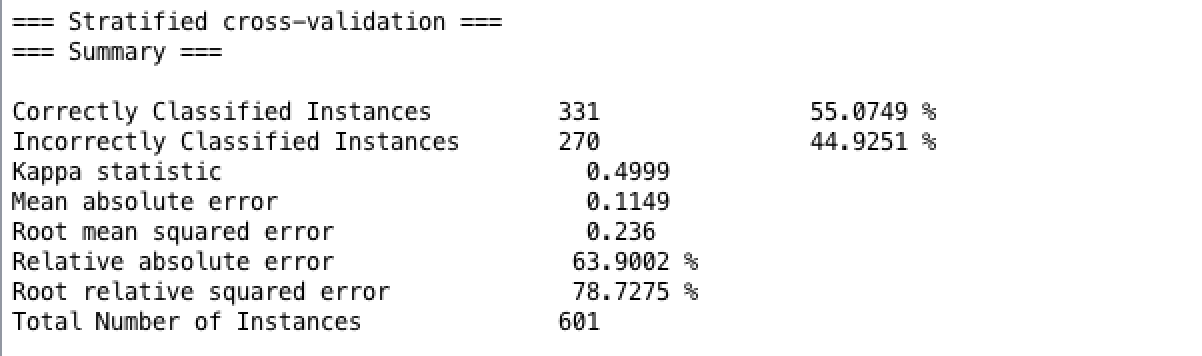
Reslut of Pipeline:

Linear regression:



由于我的pipeline过滤掉了无关数据，所以这现在进行线性回归算法的数据是只保留了每年的捕捞量（单位是吨）以及年份和鱼的种类。因为datavalue数据是连续的，所以可以使用线性回归，我们实验的目的也是为了通过已知数据去预测未来的趋势，这满足线性回归的目的，对于我的dataset来说也就是根据已有的每年的捕捞量去预测未来的捕捞量，最后通过相关系数不难看出结果很令人满意，所以线性回归算法是非常合适的。线性回归，就是能够用一个直线较为精确地描述数据之间的关系。这样当出现新的数据的时候，就能够预测出一个简单的值。

Logistic regression:

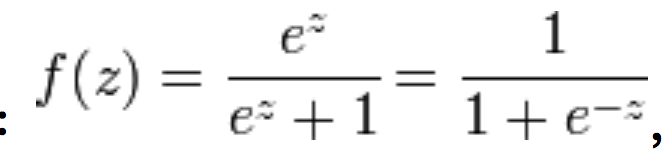


逻辑回归与线性回归不同，逻辑回归算法本质其实是一种分类算法，而并不是想要去预测某一个数值。所以结果的形式是与classification一样的。逻辑回归用于离散变量的**分类**，即它的输出y的取值范围是一个离散的集合，主要用于类的判别，而且其输出值y表示属于某一类的概

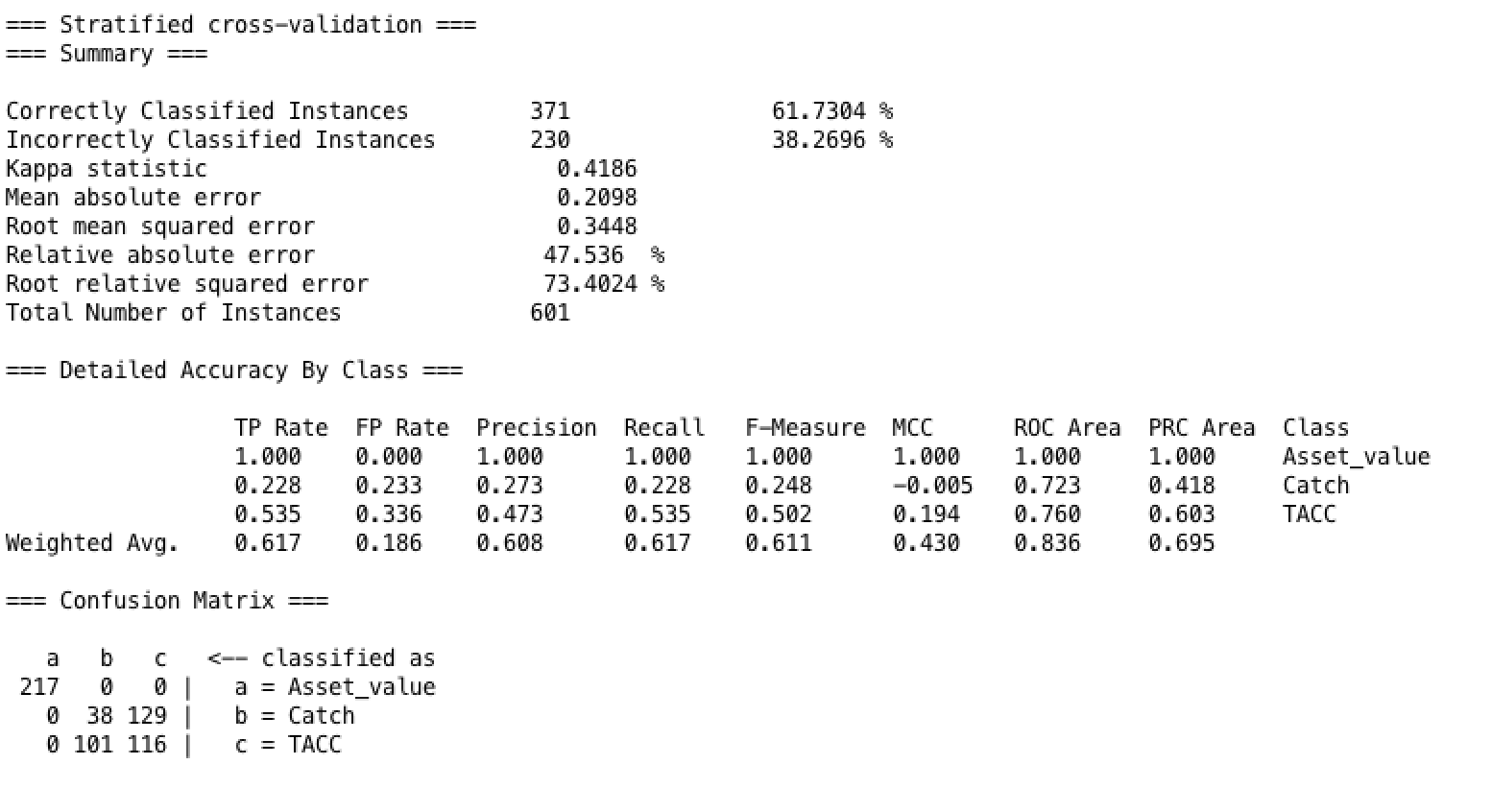
Logistic Regression逻辑回归主要用于分类问题，常用来预测概率，如知道一个人的年龄、体重、身高、血压等信息，预测其患心脏病的概率是多少。经典的LR用于二分类问题（只有0,1两类）。

## **Logistic函数**

对于任意的x值，对应的y值都在区间(0,1)内。

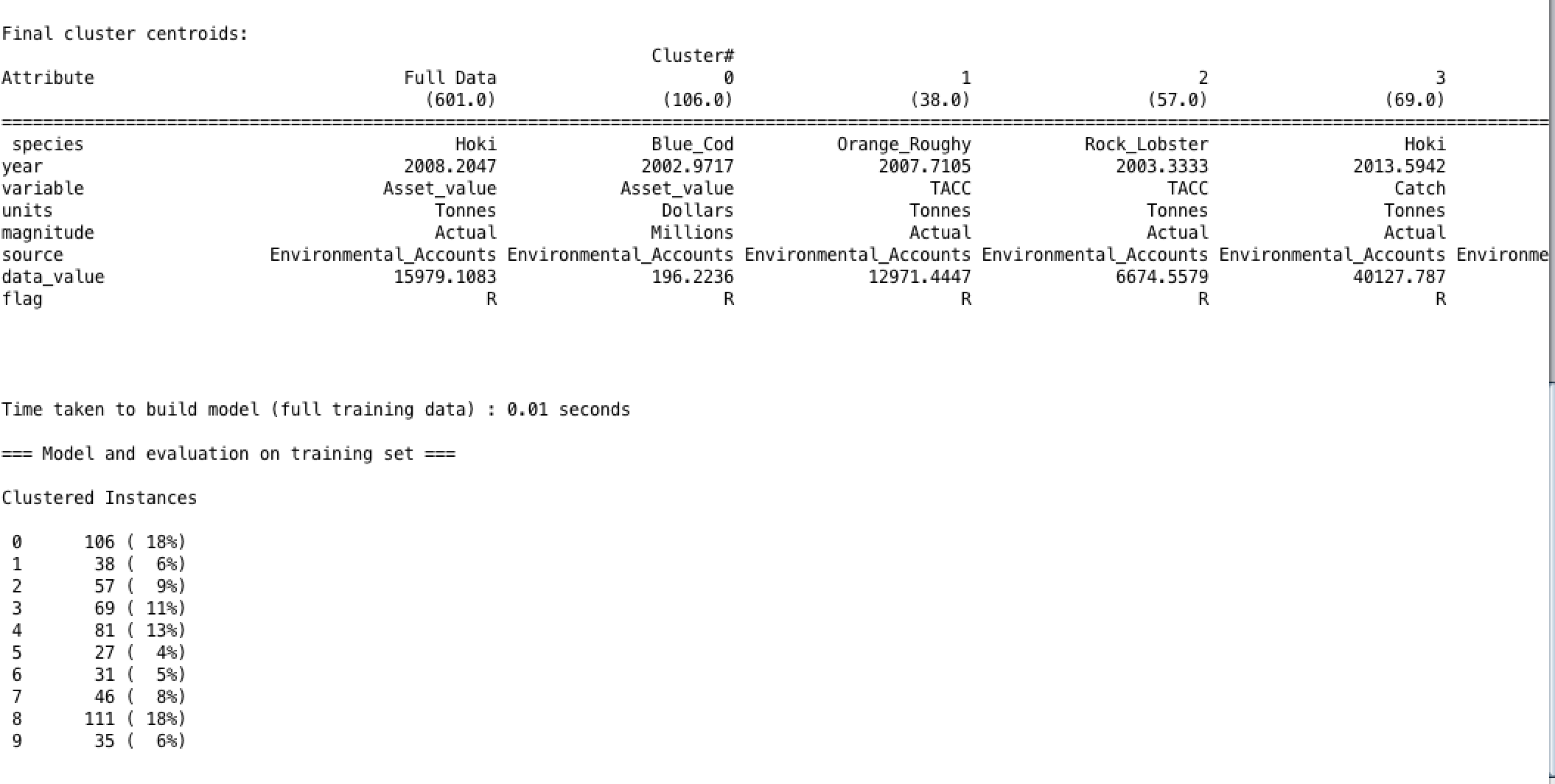
**函数公式为：**

IBK:



由于是分类算法所以结果是correctly classified instances，但其实实验目的并没有将dataset分类的需求。

SimplyKmean:



？？？

首先说上面两种不同的回归算法：线性回归和逻辑回归，普通线性回归主要用于连续变量的**预测**，即，线性回归的输出y的取值范围是整个实数区间（y∈R）

逻辑回归用于离散变量的**分类**，即它的输出y的取值范围是一个离散的集合，主要用于类的判别，而且其输出值y表示属于某一类的概率

单一的一个dataset是不足以证明过度捕捞对鱼类的的影响的，所以我们需要引入另一个dataset，这个dataset是hard和soft线of鱼类

<https://www.mpi.govt.nz/growing-and-harvesting/fisheries/fisheries-management/fish-stock-status/>

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

Question 2: Whether the fishing exceeds the limit？

Question 3: What is the relationship between catchment and asset value?

Part 2: completion

file:///Users/Patrick/Downloads/Environmental-economic-accounts-2018-data-to-2016-corrected.pdf

