CO 544 Machine Learning and Data Mining

PreLab

Create a new directory called weka. To run Weka, change into that directory and type

java -jar weka.jar

Data Preprocessing in Weka

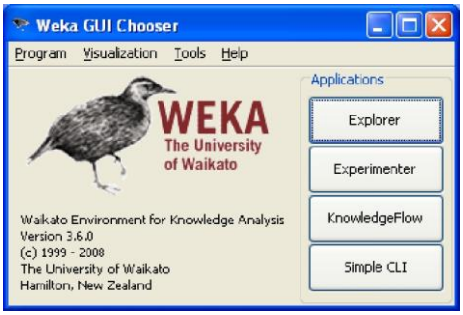
1. Download the zooData.txt dataset.

This Data set contains details of animal name, hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, fins, legs, tail, domestic, catsize and type for each of seven different types of animals (mammal, bird, reptile, fish, amphibian, insect, invertebrate). It contains 100 samples. Each example of an animal is termed a *sample*, or *instance*.

2. Converting dataset

   Use the given zooData.txt file to complete the zoo.arff file **manually**.

3. Start up Weka, you will see the ***Weka GUI Chooser*** screen.

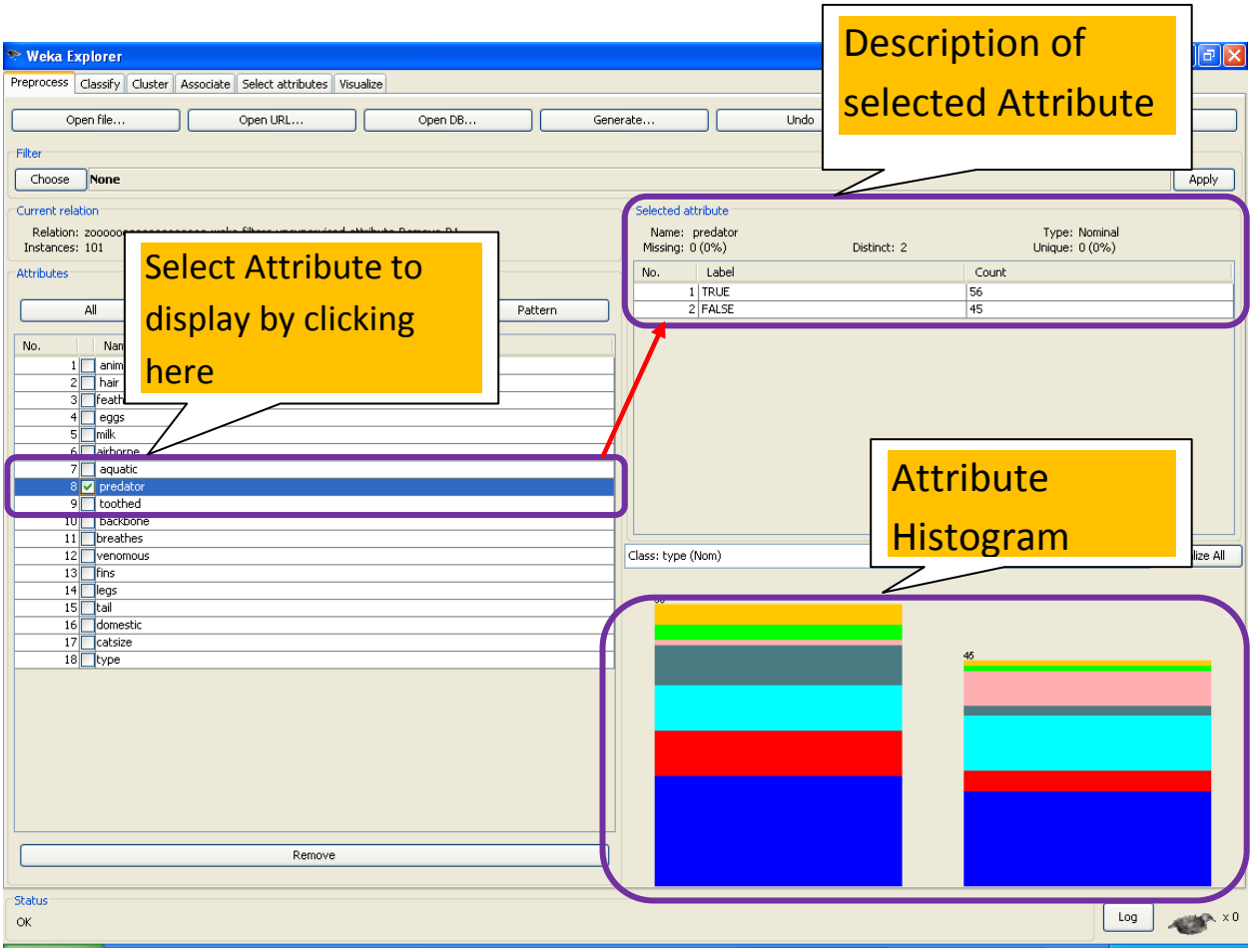


4. Explorer Screen

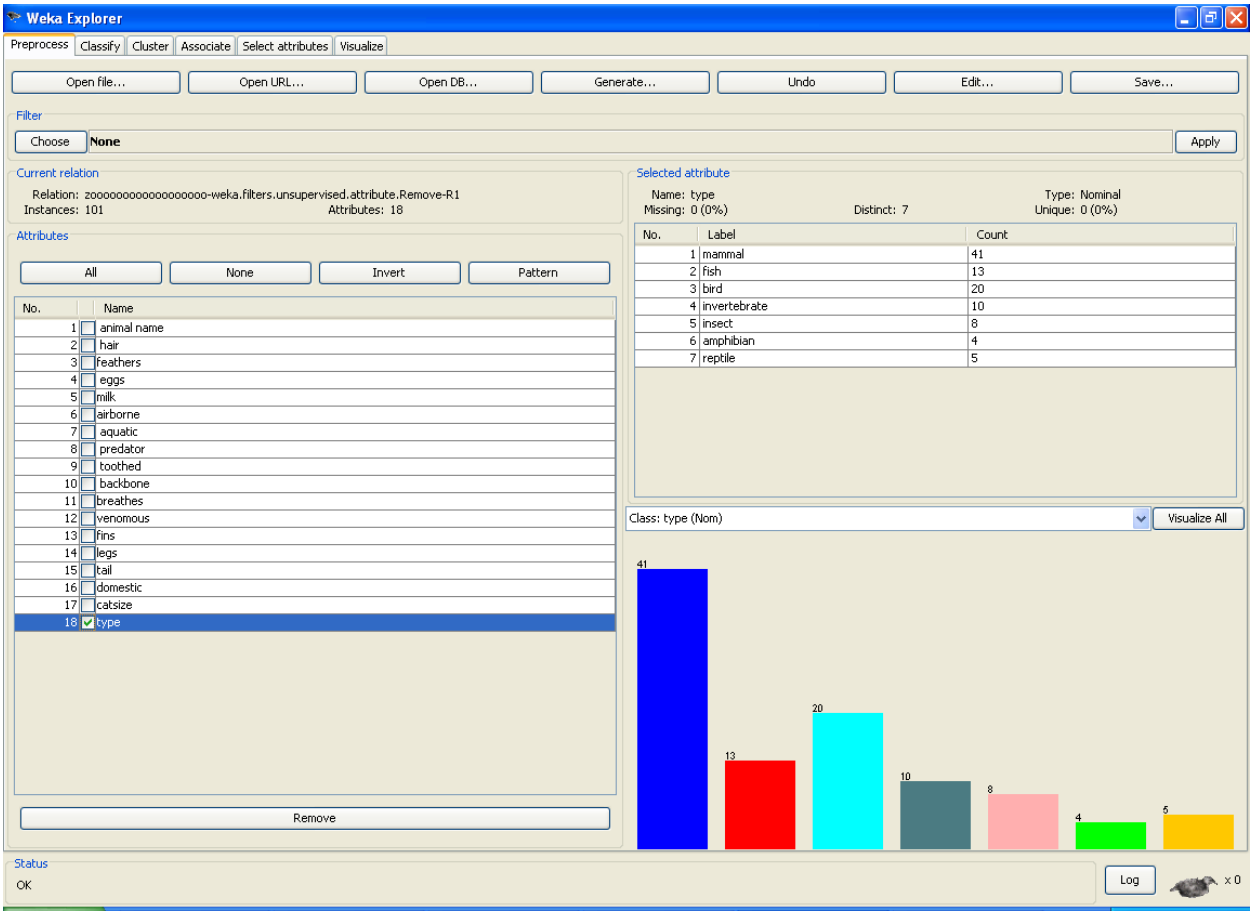
There are three elements of the screen.

1. Attribute statistics
2. Class designator
3. Attribute histogram

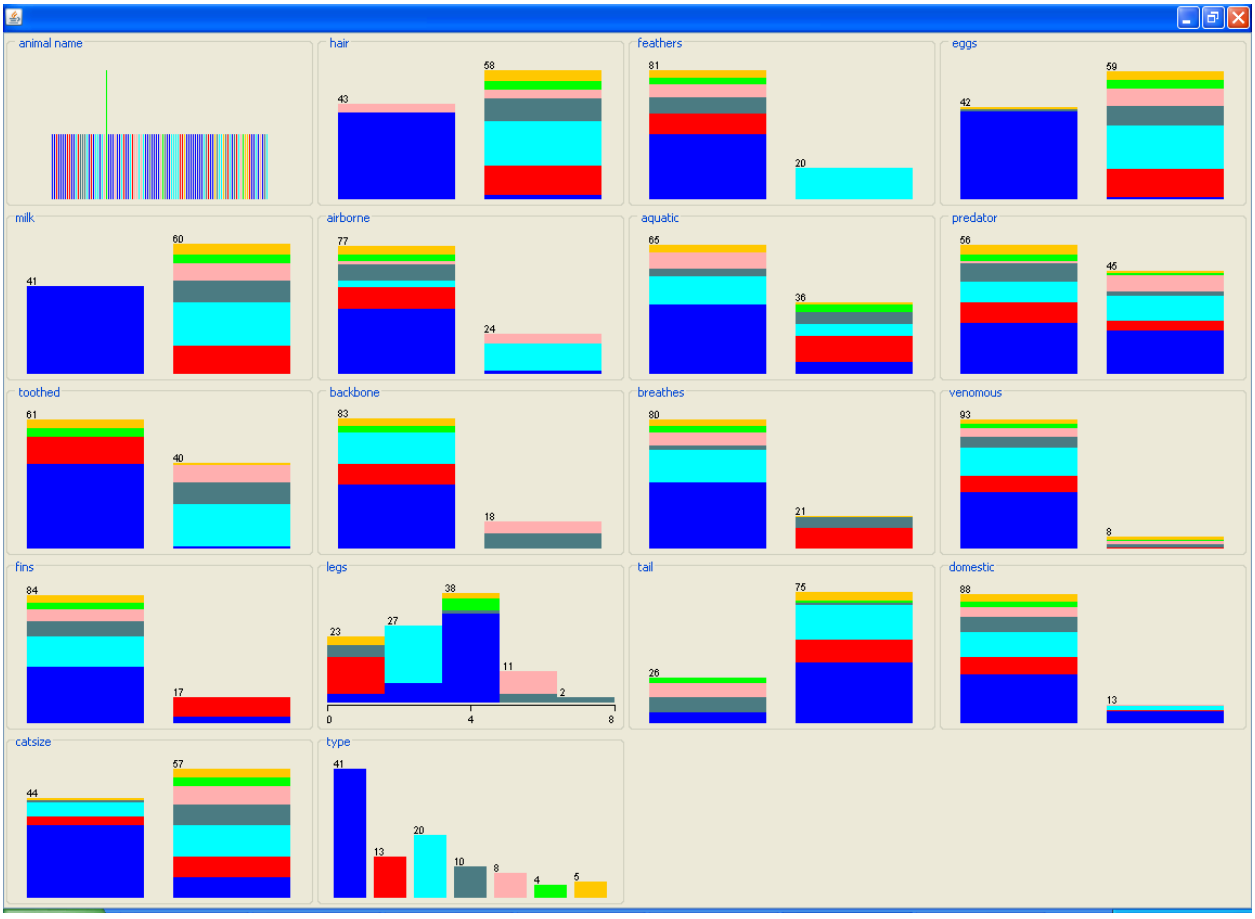
5. To see a histogram for any attribute, select it in the ***Attributes*** section of the ***Preprocess*** tab .Here we see the histogram for the predator attribute. The histogram shows us the distribution of predator for all the seven types of animals.

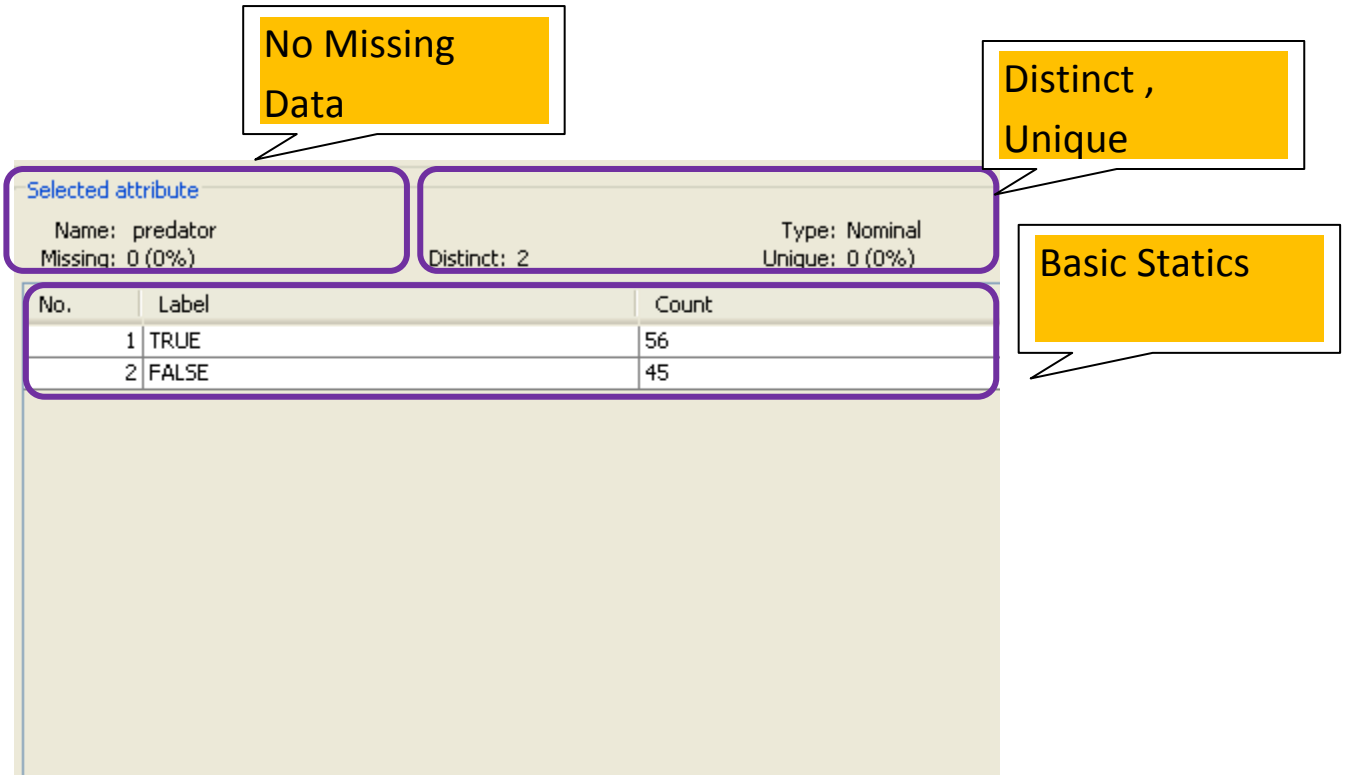


When you select the Class attribute the Histogram will look like the bellow diagram.



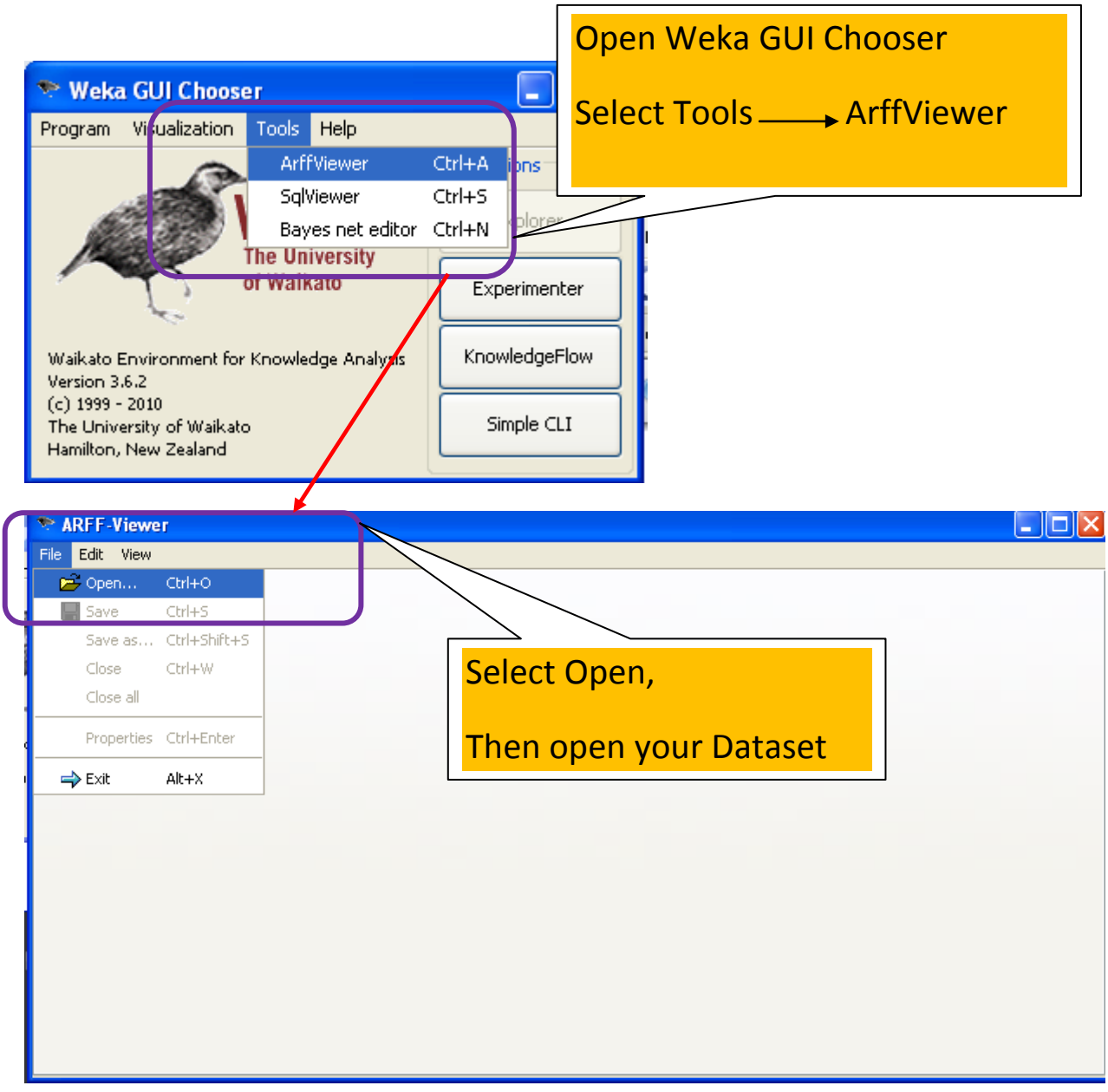
Now, click on the ***Visualize All*** button to see the histograms of all the attributes together.

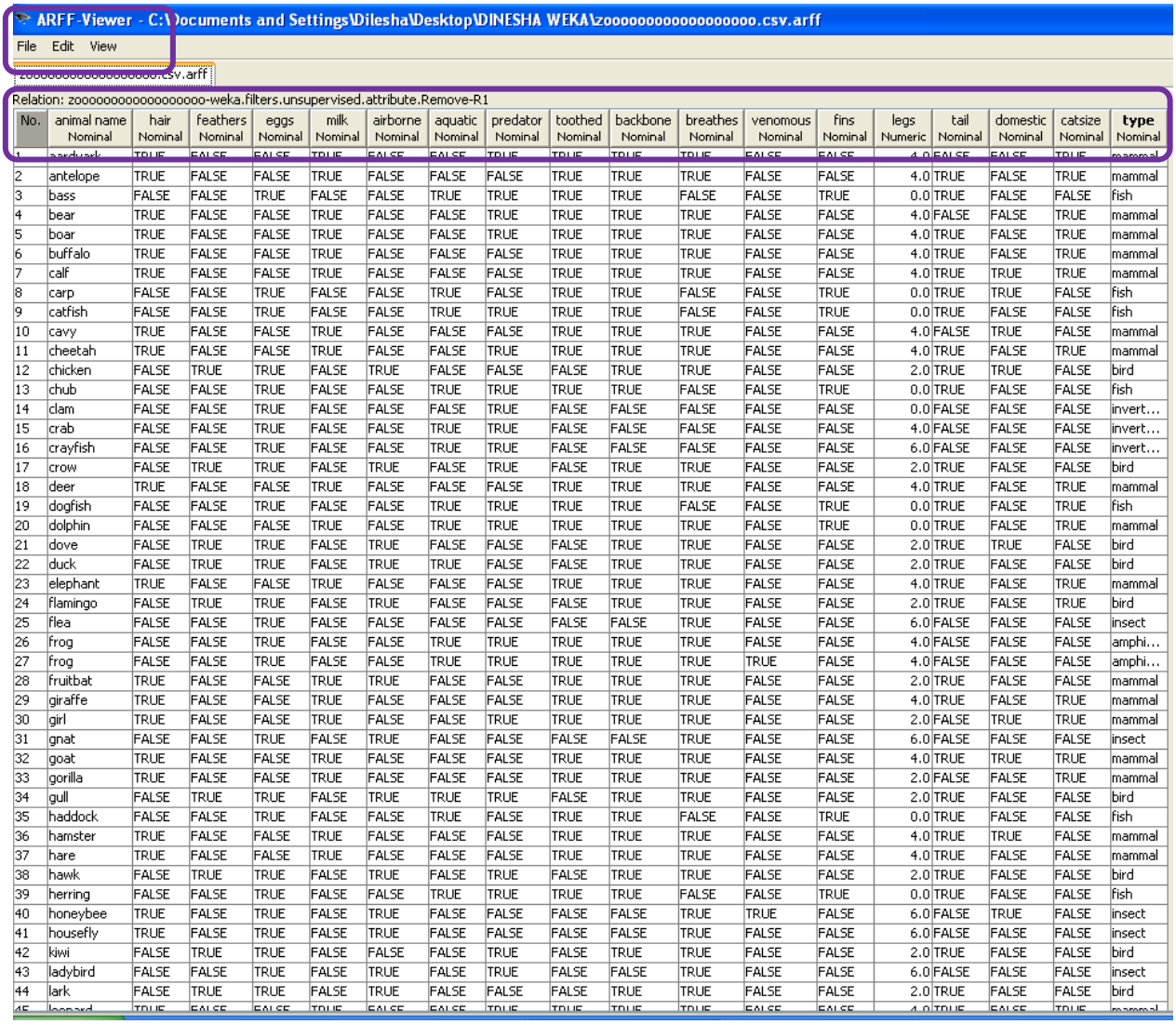




Here, for the predator length attribute, we can see that we have no missing data, in other words, there are no instances in the dataset which have no prdator measurement. We also see basic statistics for this attribute. This screen also provides values for two characteristics termed ***Distinct*** and ***Unique***.

10. Additional Useful Weka Capabilities
    - ARRF file viewer – View and modify datasets.
        i. Open Weka to view the ***Weka GUI Chooser***.
        ii. On the ***Weka GUI Chooser***, select ***Tools***, then ***ARFFViewer***.
        iii. When the ***ARFF Viewer*** opens up, select ***Open***.
        iv. Then find and open the dataset file.


Open Weka GUI Chooser
Select Tools ⟶ ArffViewer
Weka GUI Chooser
Program
Visualization
Tools
Help
ArffViewer Ctrl+A
SqlViewer Ctrl+S
Bayes net editor Ctrl+N
The University of Waikato
Experimenter
KnowledgeFlow
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand
ARFF-Viewer
File
Edit
View
Open... Ctrl+O
Save Ctrl+S
Save as... Ctrl+Shift+S
Close Ctrl+W
Close all
Properties Ctrl+Enter
Exit Alt+X
Select Open,
Then open your Dataset



No.	animal name Nominal	hair Nominal	feathers Nominal	eggs Nominal	milk Nominal	airborne Nominal	aquatic Nominal	predator Nominal	toothed Nominal	backbone Nominal	breathes Nominal	venomous Nominal	fins Nominal	legs Numeric	tail Nominal	domestic Nominal	catsize Nominal	**type** Nominal
1	aardvark	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	FALSE	FALSE	TRUE	mammal
2	antelope	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
3	bass	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
4	bear	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	FALSE	FALSE	TRUE	mammal
5	boar	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
6	buffalo	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
7	calf	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	TRUE	TRUE	mammal
8	carp	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	TRUE	FALSE	fish
9	catfish	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
10	cavy	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	FALSE	TRUE	FALSE	mammal
11	cheetah	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
12	chicken	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	TRUE	FALSE	bird
13	chub	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
14	clam	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	0.0	FALSE	FALSE	FALSE	invert...
15	crab	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	4.0	FALSE	FALSE	FALSE	invert...
16	crayfish	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	6.0	FALSE	FALSE	FALSE	invert...
17	crow	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	bird
18	deer	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
19	dogfish	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	TRUE	fish
20	dolphin	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	0.0	TRUE	FALSE	TRUE	mammal
21	dove	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	TRUE	FALSE	bird
22	duck	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	bird
23	elephant	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
24	flamingo	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	TRUE	bird
25	flea	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	6.0	FALSE	FALSE	FALSE	insect
26	frog	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	FALSE	FALSE	FALSE	amphi...
27	frog	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	4.0	FALSE	FALSE	FALSE	amphi...
28	fruitbat	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	mammal
29	giraffe	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal
30	girl	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	2.0	FALSE	TRUE	TRUE	mammal
31	gnat	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	6.0	FALSE	FALSE	FALSE	insect
32	goat	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	TRUE	TRUE	mammal
33	gorilla	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	2.0	FALSE	FALSE	TRUE	mammal
34	gull	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	bird
35	haddock	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
36	hamster	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	TRUE	FALSE	mammal
37	hare	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	FALSE	mammal
38	hawk	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	bird
39	herring	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	0.0	TRUE	FALSE	FALSE	fish
40	honeybee	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	6.0	FALSE	TRUE	FALSE	insect
41	housefly	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	6.0	FALSE	FALSE	FALSE	insect
42	kiwi	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	bird
43	ladybird	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	6.0	FALSE	FALSE	FALSE	insect
44	lark	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	2.0	TRUE	FALSE	FALSE	bird
45	leopard	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	4.0	TRUE	FALSE	TRUE	mammal

Building Decision Trees in Weka

1. Move from the ***Preprocess*** tab to the ***Classify*** tab.To select the Weka J4.8 algorithm, click on ***Choose*** and Follow the path

weka ⟶ classifiers ⟶ trees ⟶ J48

2. As we can see, the J4.8 decision tree has been loaded.Now, left-click on ***J48*** to open up the Weka ***Generic Object Editor.***For our first try at building a decision tree, we will accept all of the defaults here, except that we will change ***saveInstanceData*** to ***true***.This will allow us to find out how each sample is classified after we build the tree.

3. Click the "***Use Training set***" radio button in the ***Test options*** box.
.Then, Weka will build the tree using the training set.

4. Click ***Start*** on the ***Classify*** tab. The ***Classifier output*** box shows the results of classification.

5. To see the tree, right-click on the highlighted ***Result list*** entry for the tree we just built, and then click ***Visualize tree***.

You will have a tree like this