Assignment 2

Ahsan Nabi FA24-PCS-005

The aim of this assignment is to understand the complete ML pipeline.

Q1.1 Hand crafted features:

Extract as many input features as you can by manually observing the text, i.e., the names of people. Create ARFF file(s).

Answer: Names, Length, First (character), Last (character), Sum of Ascii values of the characters, Number of vowels, number of consonants, gender of name (manually added), Class (+ or -), and Class10 (0 for - and 1 for +)

Q1.2 Convert the output feature, i.e., + and – symbols to their numeric equivalent (1 and 0).

Answer: done on excel by the formula: =IF(I2="+", 1, IF(I2="-", 0, ""))

Q2.1 ML experiments in WEKA: View different characteristics of the data (WEKA's main window). If you notice anything interesting about the dataset, record it.

Answer: I run the InfoGainAttributeEvaluator for ranking the attributes by information gain and I found the following best attributes:

Ranked attributes:

1.0000000000000002	9 Class
1.0000000000000002	1 Names
0.7294602478794109	3 First
0.1319344939991074	4 Last
0	Rest of the attributes

The Class attribute had 1 Information gain because they mapped one on one with the label Class10. Therefore it was not a new or interesting feature. Names also mapped one on one to class and could not find any new relationships. However, First character created some new rules like names starting with `A` most likely were classified 0 or – and names ending with `a` most likely were classified as 1 or + (See the dataset visualization in Figure 1).

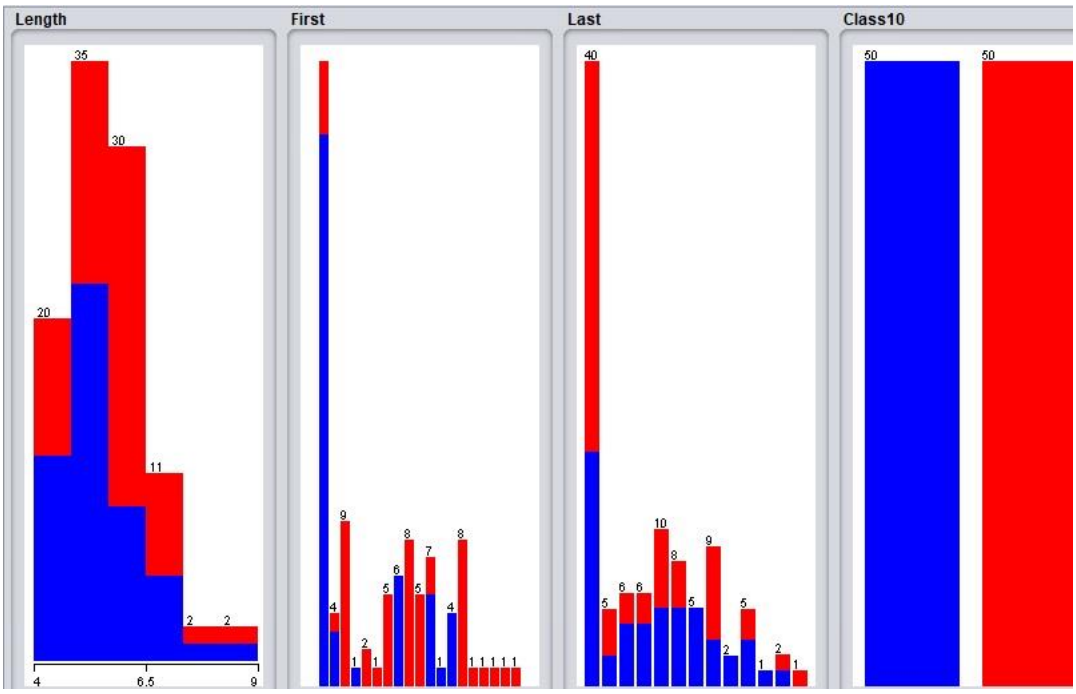

Figure 1: Attributes First and Last mostly determine the class for names starting with A (mostly class +) or ending with a (mostly class -)

Hence, we run our J48 algorithm on First and Last character attributes.

Q2.2 Run the j48 classification algorithm and observe/record the results.

Result 1:

Attributes: Names, Length, First, Last, Sum_ASCII, Vowels, Gender, Consonants, Class, and Class10

J48 pruned tree:

Class = -: 0 (50.0)

Class = +: 1 (50.0)

Accuracy: 100%

No new information from Result 1.

Result 2:

Attributes: Length, First, Last, Class10

J48 pruned tree:

First = A: 0 (34.0/4.0)

First = K: 0 (4.0/1.0)

First = F: 1 (9.0)

First = O: 0 (1.0)

First = L: 1 (2.0)

First = D: 1 (1.0)

First = B: 1 (5.0)

First = I: 0 (6.0)

First = H: 1 (8.0)

First = N: 1 (5.0)

First = S: 0 (7.0/2.0)

First = G: 0 (1.0)

First = U: 0 (4.0)

First = M: 1 (8.0)

First = E: 1 (1.0)

First = Z: 1 (1.0)

First = Q: 1 (1.0)

First = J: 1 (1.0)

First = R: 1 (1.0)

Accuracy: 88.23%

The accuracy can be increased by increasing the percentage split from 66% to 75%. In that case the accuracy becomes 92%. Note that 92% accuracy is determined by the First character attribute only.

Q3. Write a paragraph about your experience of working with the standard ML pipeline in your own words:

Answer: Weka 3.8 has Workbench, but Weka 3.6 does not. Still, you can use Explorer to get the same results. Data preprocessing in Weka is naïve, so to convert class from +/- to 1/0 required preprocessing in excel. Other attributes can also be calculated and added in excel using user-defined formulas. Excel can be converted to CSV and then imported in Weka. There it can be saved as ARFF. The Weka tutorial in Assignment 0 is basically a list of slides which do not show many new features like Data Mining Process, Workflow, etc. So I did not create ML pipeline, but used Workbench and Explorer tabs to do preprocessing, classification and attribute selection. I could save the model, visualize J48 tree and predict classes of instances. On the whole, Weka saved my time finding the interesting attributes/features and creating decision tree. However, we need data of better quality so that a combination of attributes could determine the class with more accuracy and minimally descriptive model.