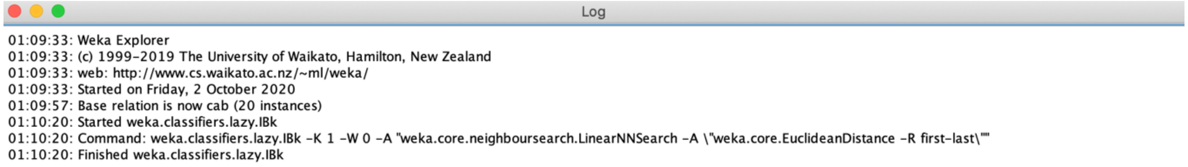
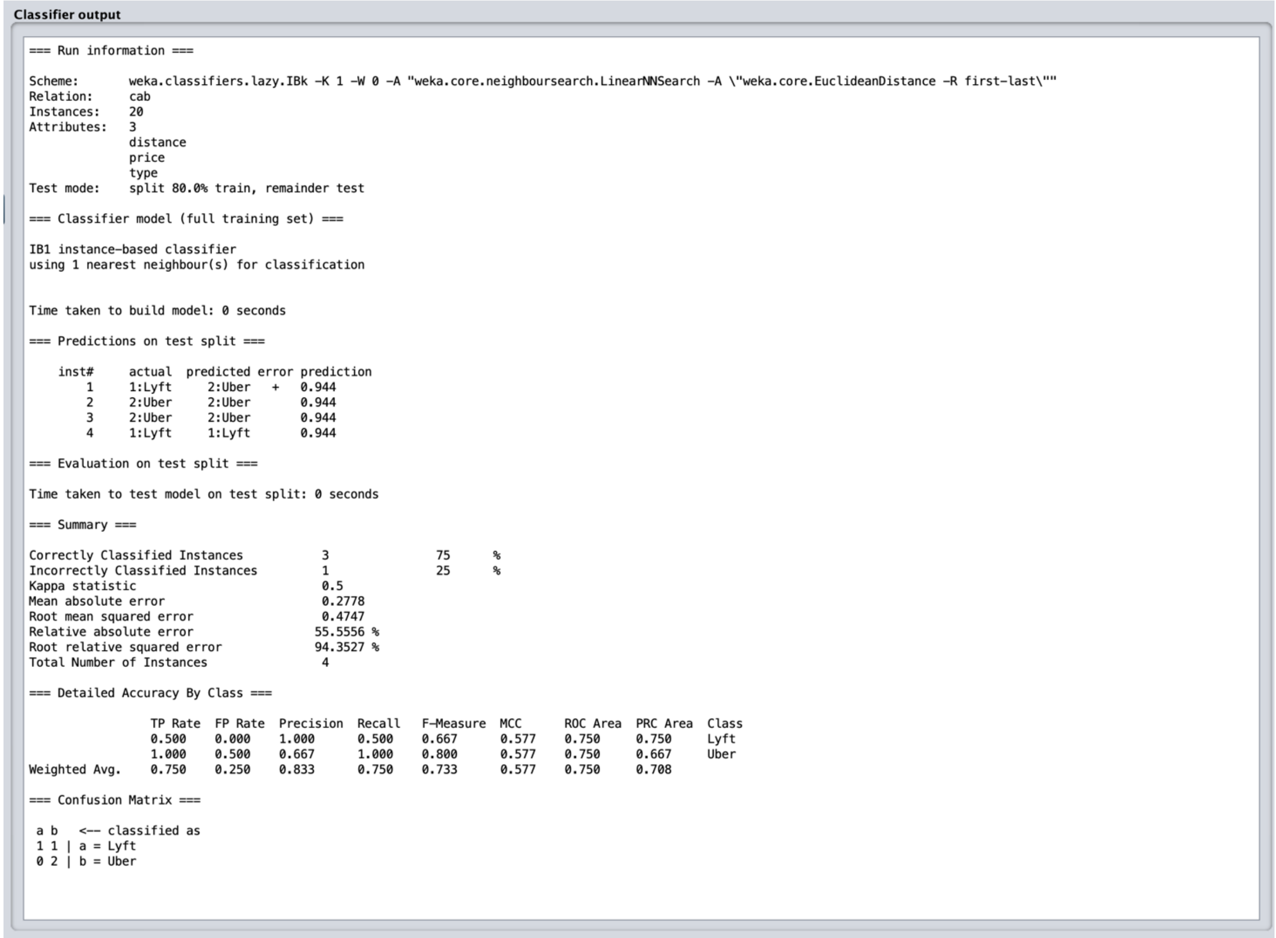
Man Tik Li
INFO7 – Data Mining Applications
Assignment #1
September 30, 2020

Weka Log



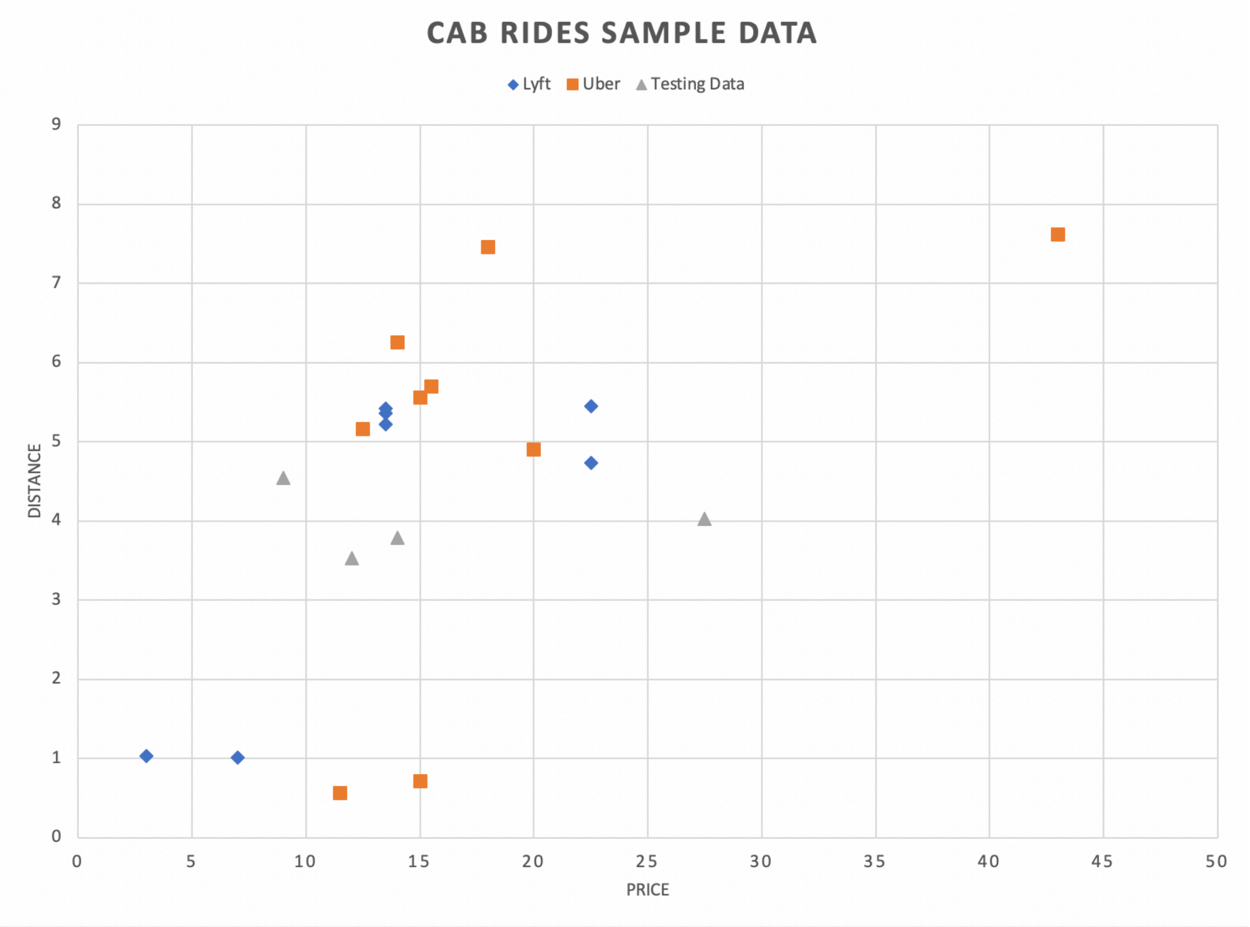
```
01:09:33: Weka Explorer
01:09:33: (c) 1999-2019 The University of Waikato, Hamilton, New Zealand
01:09:33: web: http://www.cs.waikato.ac.nz/~ml/weka/
01:09:33: Started on Friday, 2 October 2020
01:09:57: Base relation is now cab (20 instances)
01:10:20: Started weka.classifiers.lazy.IBk
01:10:20: Command: weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
01:10:20: Finished weka.classifiers.lazy.IBk
```

Weka Output



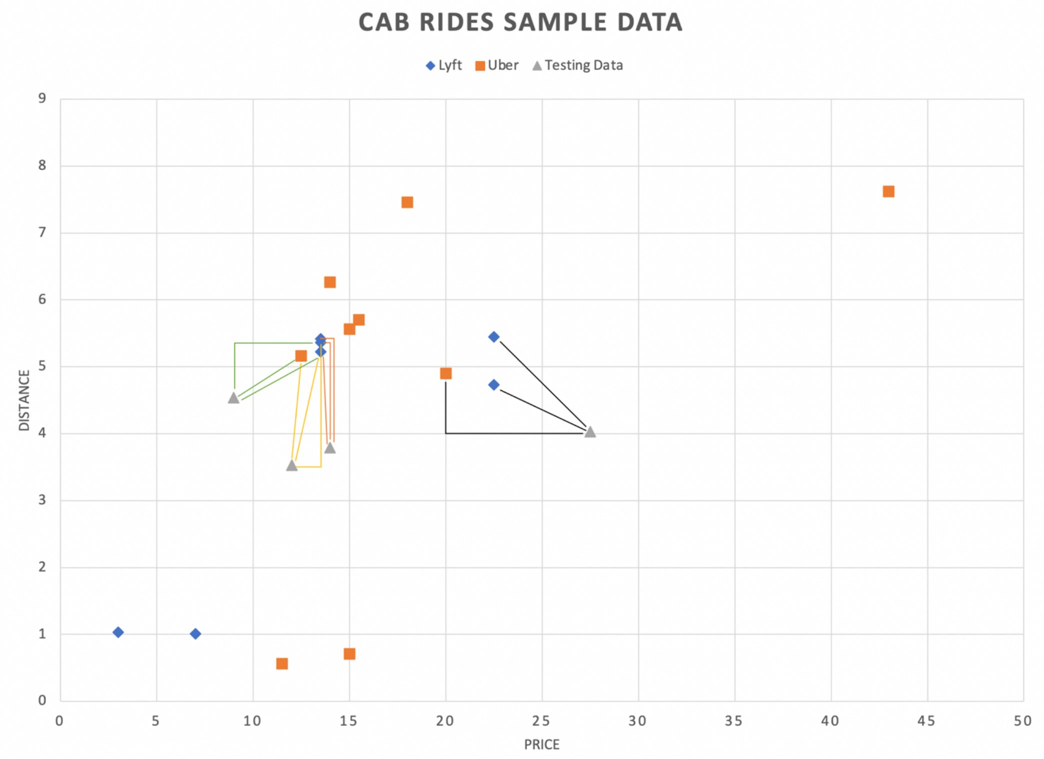
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     cab
Instances:    20
Attributes:   3
              distance
              price
              type
Test mode:    split 80.0% train, remainder test

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Predictions on test split ===

    inst#     actual  predicted error prediction
        1     1:Lyft     2:Uber   +   0.944
        2     2:Uber     2:Uber       0.944
        3     2:Uber     2:Uber       0.944
        4     1:Lyft     1:Lyft       0.944

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances           3               75      %
Incorrectly Classified Instances         1               25      %
Kappa statistic                          0.5
Mean absolute error                      0.2778
Root mean squared error                  0.4747
Relative absolute error                 55.5556 %
Root relative squared error             94.3527 %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.000    1.000      0.500    0.667      0.577    0.750     0.750     Lyft
                 1.000    0.500    0.667      1.000    0.800      0.577    0.750     0.667     Uber
Weighted Avg.    0.750    0.250    0.833      0.750    0.733      0.577    0.750     0.708

=== Confusion Matrix ===

 a b   <-- classified as
 1 1 | a = Lyft
 0 2 | b = Uber
```

Data Plot



Plotted in excel

Manually Examining 3 Nearest Neighbors



Point 1

- Closest Neighbor: Blue
- 3 Closest Neighbors: 2 Blue and 1 Orange
- No discrepancy between k=1 and k=3
- Actual Data: Blue (Lyft)

Point 2

- Closest Neighbor: Blue
- 3 Closest Neighbors: 2 Blue and 1 Orange
- No discrepancy between k=1 and k=3
- Actual Data: Orange (Uber)

Point 3

- Closest Neighbor: Blue
- 3 Closest Neighbors: 3 Blue
- No discrepancy between k=1 and k=3
- Actual Data: Orange (Uber)

Point 4

- Closest Neighbor: Blue
- 3 Closest Neighbors: 2 Blue and 1 Orange
- No discrepancy between k=1 and k=3
- Actual Data: Blue (Lyft)

Conclusion

The outcome of the point 1 and point 4 are match with their actual data with the data of closet neighbors. For point 2 and point 3, I occurred a different result. The possible errors, including insufficient data or the size of the data is small.