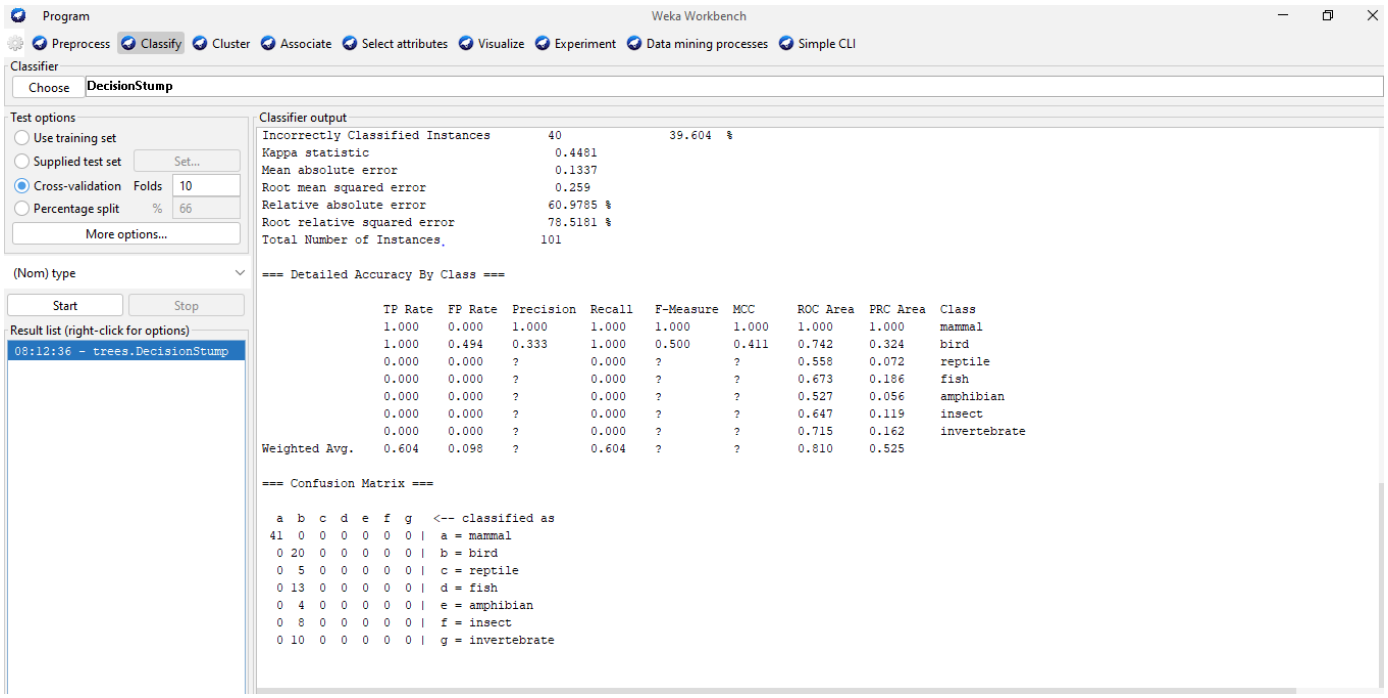
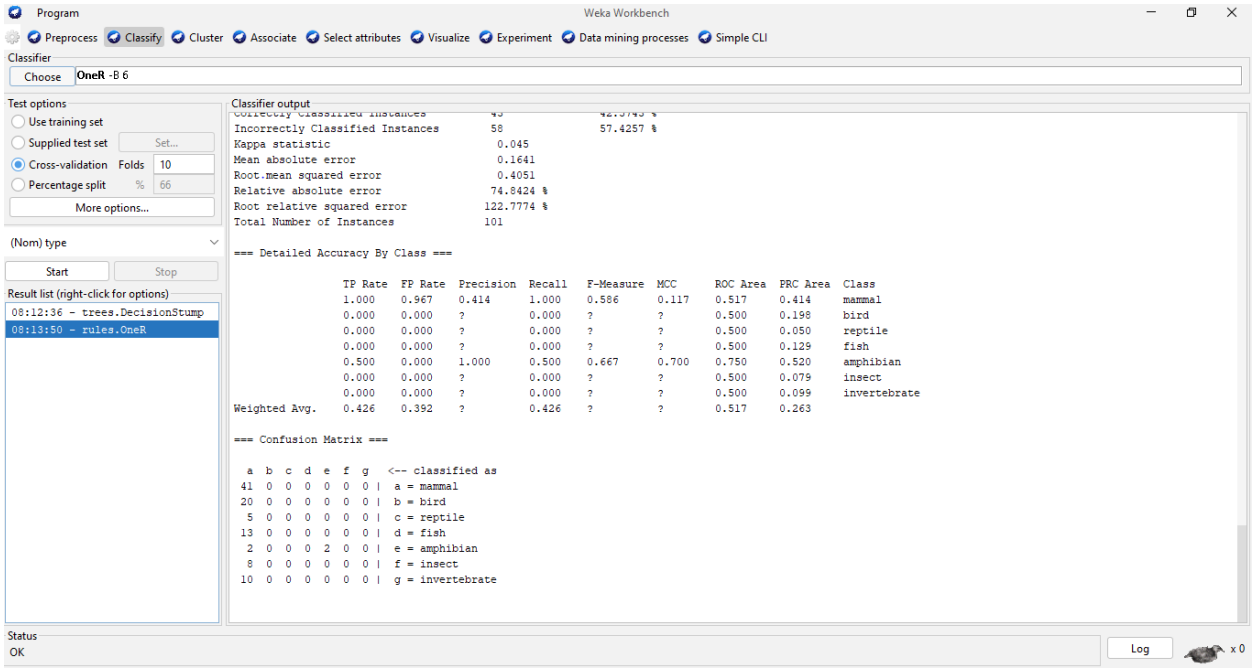
Introduction to WEKA

1. Use the following learning schemes to analyze the zoo data (in zoo.arff):

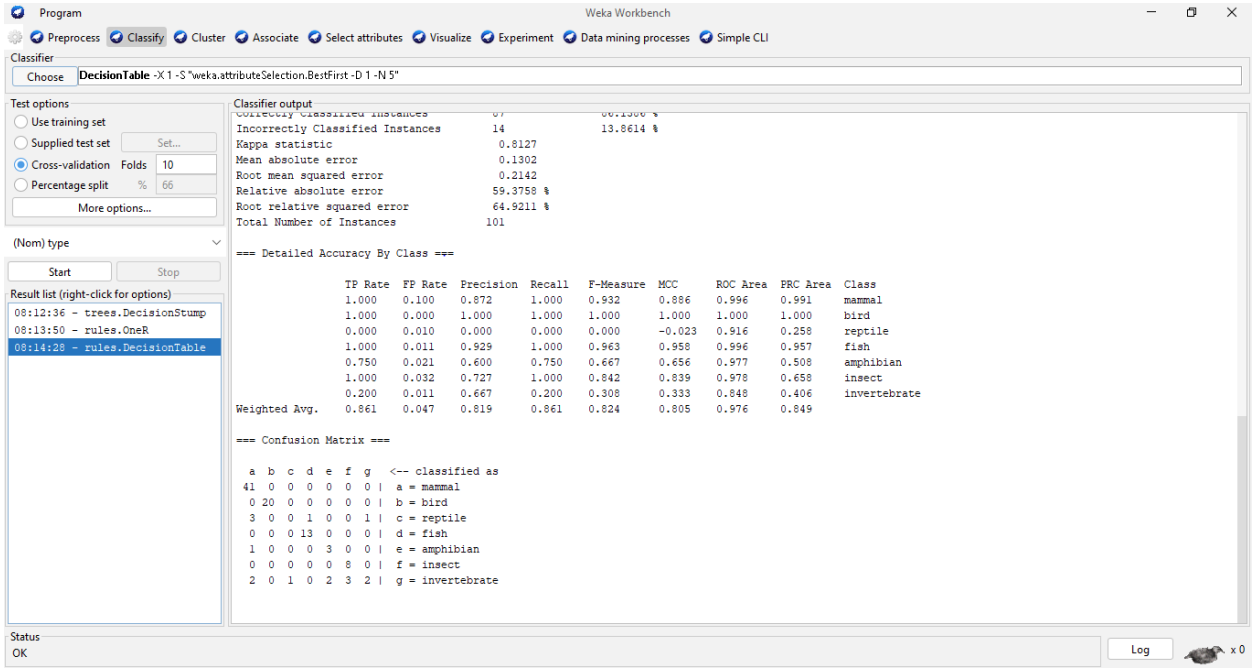
- Decision stump - weka.classifiers.DecisionStump



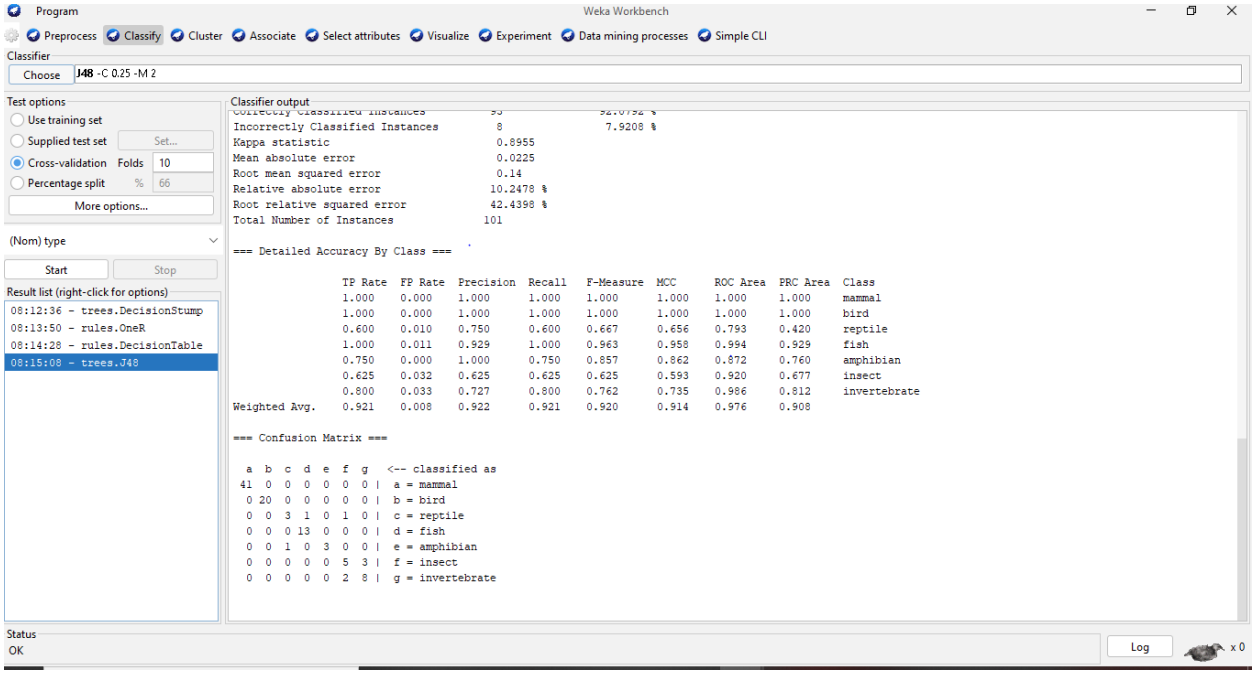
- OneR - weka.classifiers.OneR



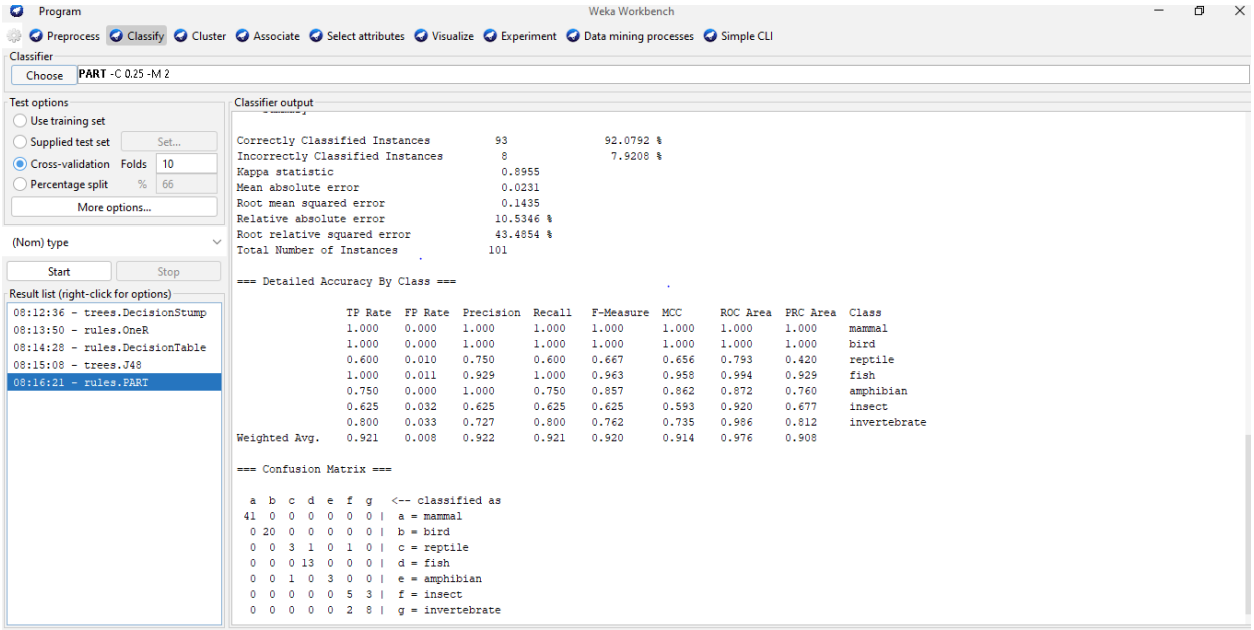
- Decision table - weka.classifiers.DecisionTable -R



- C4.5 - the J48 classifier



- PART - under "rules"



- How do the classifiers determine whether an animal is a mammal, bird, reptile, fish, amphibian, insect, or invertebrate?
    - Decision Stump: Nothing was identified correctly except the bird and mammal groups. There are plenty of groups that were identified as birds that shouldn't

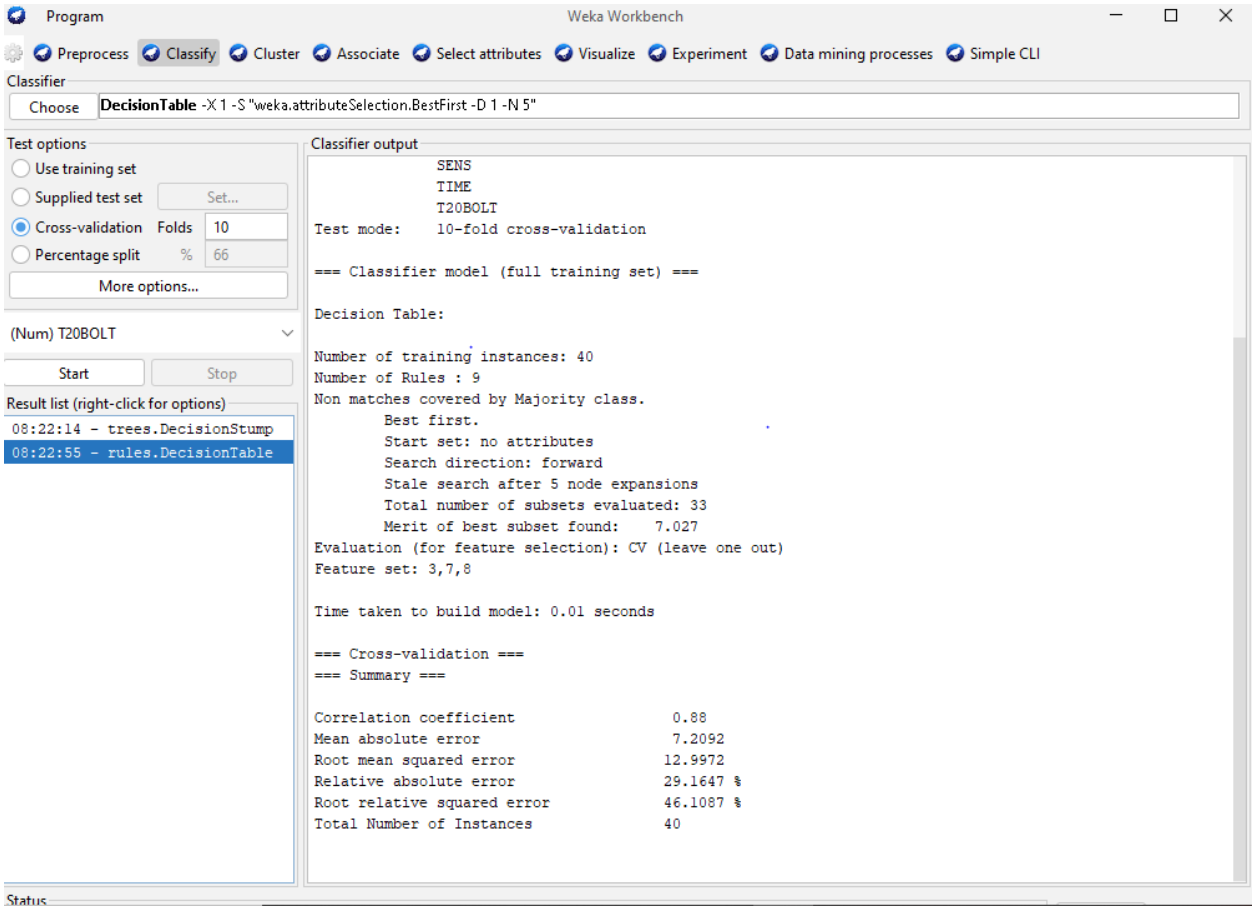
have been as well. Based on the viewer it appears as though it is not checking past the milk attribute so a lot of these variables are not being correctly identified.
    - Decision Table: It looks like amphibian, invertebrate, and reptile were misclassified. This is because they were only measured by a few attributes such as milk, legs, and tail.
    - OneR: Looks like only mammal was identified correctly, because only one variable was used to complete the matrix.
    - PART: Only fish, mammal, and bird were correctly identified, but since many decision rules were used it appears to be more accurate.
    - J48: Same as PART, even down to the CCI.
- Do the decisions made by the classifiers make sense to you?
    - It makes sense but I also do not understand why you would use OneR, because it only analyzes by one parameter compared to the others that are more accurate.
- What can you say about the accuracy of these classifiers when classifying an animal that has not been used for training?
    - J48 and the Decision Stump were the most accurate with a CCI over 90%.
- Why does OneR perform so badly?
    - It only analyzes by a single parameter since it is meant to be simple.

2. Use the following learning schemes to analyze the bolts data (bolts.arff without the TIME attribute):

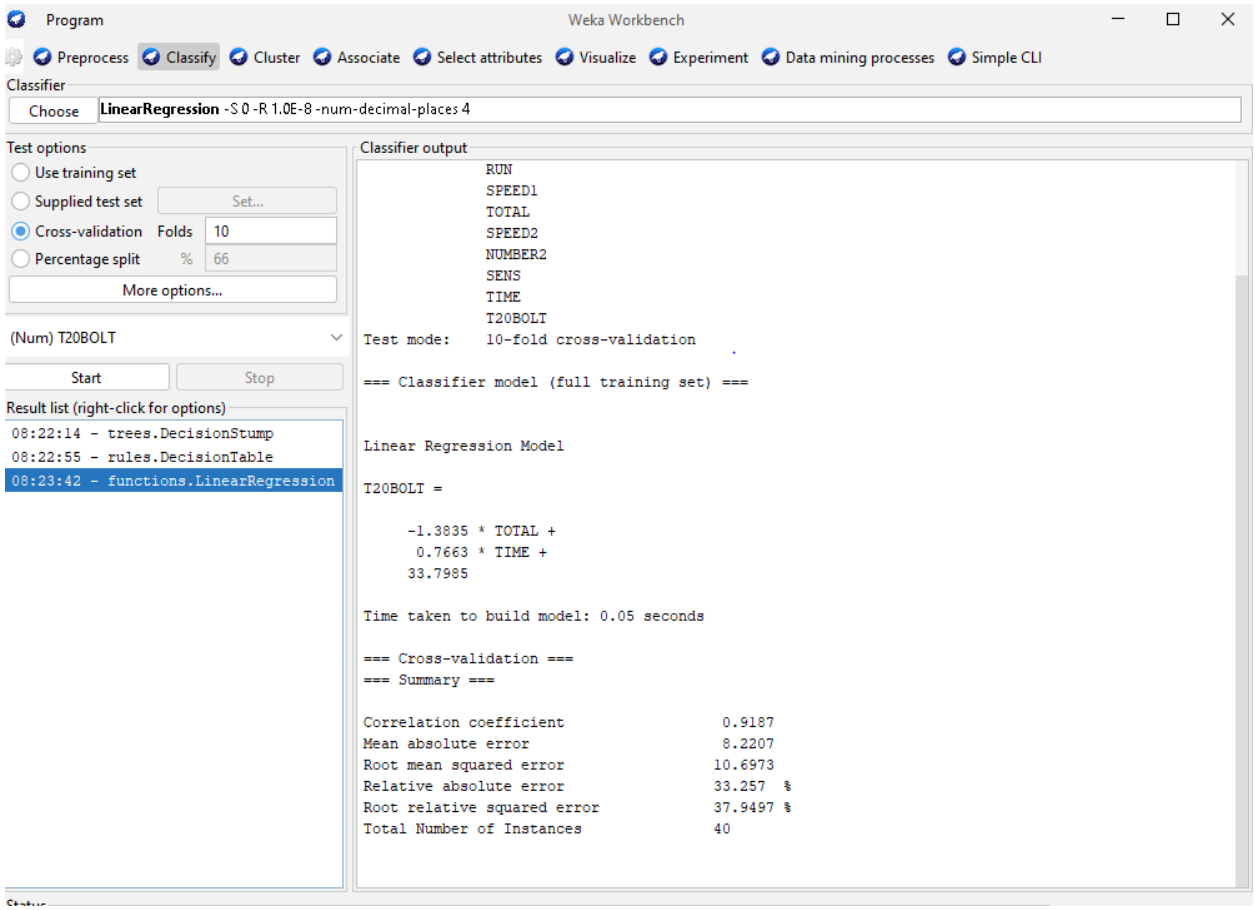
- Decision stump - weka.classifiers.DecisionStump

```
              TOTAL
              SPEED2
              NUMBER2
              SENS
              TIME
              T20BOLT
Test mode:    10-fold cross-validation

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

Decision Stump

Classifications

TIME <= 32.19 : 18.378275862068968
TIME > 32.19 : 74.94454545454549
TIME is missing : 33.93400000000001


Time taken to build model: 0 seconds

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

Correlation coefficient                  0.8677
Kendall's tau                            0.0657
Spearman's rho                           0.2955
Mean absolute error                      9.1809
Root mean squared error                 13.6626
Relative absolute error                 37.1415 %
Root relative squared error             48.4693 %
Total Number of Instances               40
```

- Decision table - weka.classifiers.DecisionTable -R



- Linear regression - weka.classifiers.LinearRegression



- M5' - weka.classifiers.M5'
- === Run information ===
-
- Scheme: weka.classifiers.rules.M5Rules -M 4.0 -num-decimal-places 4
- Relation: bolts
- Instances: 40
- Attributes: 8
- RUN
- SPEED1
- TOTAL
- SPEED2
- NUMBER2
- SENS
- TIME
- T20BOLT
- Test mode: 10-fold cross-validation
-
- === Classifier model (full training set) ===
-

- M5 pruned model rules
- (using smoothed linear models) :
- Number of Rules : 4
-
- Rule: 1
- IF
- TIME <= 32.19
- TOTAL > 15
- THEN
-
- T20BOLT =
- -0.8868 * TOTAL
- - 0.4179 * NUMBER2
- + 0.914 * TIME
- + 19.7865 [16/4.785%]
-
- Rule: 2
- IF
- TIME <= 23.225
- THEN
-
- T20BOLT =
- 1.162 * TOTAL
- + 0.8889 * TIME
- + 7.1663 [12/0%]
-
- Rule: 3
- IF
- RUN <= 28.5
- THEN
-
- T20BOLT =
- 0.3162 * RUN
- + 0.2518 * TOTAL
- - 0.0169 * TIME
- + 63.1843 [8/23.535%]
-
- Rule: 4
-
- T20BOLT =
- 1.2335 * TOTAL
- + 50.815 [4/30.472%]
-
-

-
- Time taken to build model: 0.03 seconds
-
- === Cross-validation ===
- === Summary ===
-
- Correlation coefficient 0.9104
- Mean absolute error 5.9185
- Root mean squared error 11.3655
- Relative absolute error 23.9433 %
- Root relative squared error 40.3201 %
- Total Number of Instances 40
-

- The dataset describes the time needed by a machine to produce and count 20 bolts. (More details can be found in the file containing the dataset.) Analyze the data.
- What adjustments have the greatest effect on the time to count 20 bolts?
    - I would say according to the above that the decision stump with 86.77 correctly classified instances and the linear regression with 91.87% correlation coefficient because they had the highest values of the analyzed set. I am not comparing PART to the LR because these are different measures.
- According to each classifier, how would you adjust the machine to get the shortest time to count 20 bolts?
    - Decision Stump: Adjust the speed 1 attribute to less than five.
    - Decision Table: Change Speed 1 to a value of 4 and the attribute total to 20.
    - Linear Regression: Make speed 1 the low value and adjust sens to be the high value
    - M5: Based on the two decision rules for speed I would increase the value of speed 1 and lower the values of both sens and total for both those greater and lower than five.

Rating: 9/10

- I believe this rating is fair because some of the questions were hard to grasp with only the reading and the videos, however, the videos were extremely helpful.