Homework 3

1. Run four models:
1. J48 with 10-fold CV (or some decision tree algorithm)

```
=== Run information ===

Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    HW3_1
Instances:   8416
Attributes:  23
             Class
             cap-shape
             cap-surface
             cap-color
             ruises
             odor
             gill-attachment
             gill-spacing
             gill-size
             gill-color
             stalk-shape
             stalk-root
             stalk-surface-above-ring
              stalk-surface-below-ring
             stalk-color-above-ring
             stalk-color-below-ring
             veil-type
             veil-color
              ring-number
             ring-type
              spore-print-color
             population
             habitat
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

odor = ALMOND: EDIBLE (400.0)
odor = ANISE: EDIBLE (400.0)
odor = NONE
|   spore-print-color = PURPLE: EDIBLE (0.0)
|   spore-print-color = BROWN: EDIBLE (1472.0)
|   spore-print-color = BLACK: EDIBLE (1424.0)
|   spore-print-color = CHOCOLATE: EDIBLE (48.0)
|   spore-print-color = GREEN: POISONOUS (72.0)
|   spore-print-color = WHITE
|   |   gill-size = NARROW
|   |   |   gill-spacing = CROWDED
|   |   |   |   population = SEVERAL: EDIBLE (72.0)
|   |   |   |   population = SCATTERED: EDIBLE (0.0)
|   |   |   |   population = NUMEROUS: EDIBLE (0.0)
|   |   |   |   population = SOLITARY: EDIBLE (0.0)
|   |   |   |   population = ABUNDANT: EDIBLE (0.0)
|   |   |   |   population = CLUSTERED: POISONOUS (16.0)
|   |   |   gill-spacing = CLOSE: POISONOUS (32.0)
|   |   gill-size = BROAD: EDIBLE (528.0)
|   spore-print-color = YELLOW: EDIBLE (48.0)
|   spore-print-color = ORANGE: EDIBLE (48.0)
|   spore-print-color = BUFF: EDIBLE (48.0)
odor = PUNGENT: POISONOUS (256.0)
odor = CREOSOTE: POISONOUS (192.0)
odor = FOUL: POISONOUS (2160.0)
odor = FISHY: POISONOUS (576.0)
odor = SPICY: POISONOUS (576.0)
odor = MUSTY: POISONOUS (48.0)

Number of Leaves  :     24

Size of the tree :      29


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8416              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             8416

=== Detailed Accuracy By Class ===
             TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
             1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     EDIBLE
             1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     POISONOUS
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 4488    0 |    a = EDIBLE
    0 3928 |    b = POISONOUS
```

2. RandomForests with 10-fold CV

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001
-S 1
Relation:     HW3_1
Instances:    8416
Attributes:   23
              Class
              cap-shape
              cap-surface
              cap-color
              ruises
              odor
              gill-attachment
              gill-spacing
              gill-size
              gill-color
              stalk-shape
              stalk-root
              stalk-surface-above-ring
               stalk-surface-below-ring
              stalk-color-above-ring
              stalk-color-below-ring
              veil-type
              veil-color
               ring-number
              ring-type
               spore-print-color
              population
```

```
            habitat
Test mode:    10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.21 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8416              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0004
Root mean squared error                  0.003
Relative absolute error                  0.0704 %
Root relative squared error              0.6078 %
Total Number of Instances             8416

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     EDIBLE
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     POISONOUS
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 4488    0 |    a = EDIBLE
    0 3928 |    b = POISONOUS
```

3. OneR with 10-fold CV (or some non-pruning rule algorithm)

```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
```

Homework 3

```
Relation:     HW3_1
Instances:    8416
Attributes:   23
              Class
              cap-shape
              cap-surface
              cap-color
              ruises
              odor
              gill-attachment
              gill-spacing
              gill-size
              gill-color
              stalk-shape
              stalk-root
              stalk-surface-above-ring
               stalk-surface-below-ring
              stalk-color-above-ring
              stalk-color-below-ring
              veil-type
              veil-color
               ring-number
              ring-type
               spore-print-color
              population
              habitat
Test mode:    10-fold cross-validation

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

odor:
                                                  ALMOND   -> EDIBLE
                                                  ANISE    -> EDIBLE
                                                  NONE     -> EDIBLE
                                                  PUNGENT  -> POISONOUS
                                                  CREOSOTE-> POISONOUS
                                                  FOUL     -> POISONOUS
                                                  FISHY    -> POISONOUS
                                                  SPICY    -> POISONOUS
                                                  MUSTY    -> POISONOUS
(8296/8416 instances correct)


Time taken to build model: 0.03 seconds
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8296              98.5741 %
Incorrectly Classified Instances       120               1.4259 %
Kappa statistic                          0.9713
Mean absolute error                      0.0143
Root mean squared error                  0.1194
Relative absolute error                  2.8644 %
Root relative squared error             23.9349 %
Total Number of Instances             8416

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.031    0.974      1.000    0.987      0.972    0.985     0.974     EDIBLE
                 0.969    0.000    1.000      0.969    0.984      0.972    0.985     0.984     POISONOUS
Weighted Avg.    0.986    0.016    0.986      0.986    0.986      0.972    0.985     0.979

=== Confusion Matrix ===

    a    b   <-- classified as
 4488    0 |    a = EDIBLE
  120 3808 |    b = POISONOUS
```

4. JRip with 10-fold CV (or some rule-pruning algorithm)

```
=== Run information ===

Scheme:       weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation:     HW3_1
Instances:    8416
Attributes:   23
              Class
              cap-shape
              cap-surface
              cap-color
              ruises
              odor
              gill-attachment
              gill-spacing
              gill-size
              gill-color
              stalk-shape
              stalk-root
              stalk-surface-above-ring
               stalk-surface-below-ring
              stalk-color-above-ring
              stalk-color-below-ring
              veil-type
              veil-color
               ring-number
              ring-type
               spore-print-color
              population
              habitat
Test mode:    10-fold cross-validation

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

JRIP rules:
===========

(odor = FOUL) => Class=POISONOUS (2160.0/0.0)
(gill-size = NARROW) and (gill-color = BUFF) => Class=POISONOUS (1152.0/0.0)
(gill-size = NARROW) and (odor = PUNGENT) => Class=POISONOUS (256.0/0.0)
(odor = CREOSOTE) => Class=POISONOUS (192.0/0.0)
( spore-print-color = GREEN) => Class=POISONOUS (72.0/0.0)
( stalk-surface-below-ring = SCALY) and (stalk-surface-above-ring = SILKY) => Class=POISONOUS
(80.0/0.0)
(stalk-color-above-ring = YELLOW) => Class=POISONOUS (8.0/0.0)
(population = CLUSTERED) and (cap-color   = WHITE) => Class=POISONOUS (8.0/0.0)
 => Class=EDIBLE (4488.0/0.0)

Number of Rules : 9


Time taken to build model: 0.18 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       8416             100      %
Incorrectly Classified Instances        0               0      %
Kappa statistic                         1
Mean absolute error                     0
Root mean squared error                 0
```

```
Relative absolute error              0     %
Root relative squared error            0     %
Total Number of Instances           8416

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     EDIBLE
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     POISONOUS
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 4488    0 |    a = EDIBLE
    0 3928 |    b = POISONOUS
```

2. Compare and contrast the models: time to build, size, accuracy/error measure, confusion matrix, etc. Discuss your results.

1. J48 with RandomForests

	J48	**RandomForests**
Time to build	0.02 seconds	0.21 seconds
Size	29	-
Accuracy/Error measure	100 %	100 %
Confusion matrix	a b <-- classified as 4488 0 | a = EDIBLE 0 3928 | b = POISONOUS	a b <-- classified as 4488 0 | a = EDIBLE 0 3928 | b = POISONOUS

J48 and RandomForest are both from the Trees model. J48 takes much less time than RandomForest to build with the cross validation of 10.

2. OneR with JRip

	OneR	**JRip**
Time to build	0.03 seconds	0.18 seconds
Size	-	-
Accuracy/Error measure	98.5741 %	100 %
Confusion matrix	a b <-- classified as 4488 0 | a = EDIBLE 120 3808 | b = POISONOUS	a b <-- classified as 4488 0 | a = EDIBLE 0 3928 | b = POISONOUS

OneR and JRip are both from the Rules model. In comparison of JRip, OneR takes much less time to build with the cross validation of 10.

3. Trees with rules

J48 and RandomForest are models of Trees. OneR and JRip are models of Rules. First comparison of J48, which is a model of Trees with OneR, which is a model of Rules. In this comparison, the time to build for J48 is less than the time taken for OneR, when cross-validation folds is 10. The accuracy for J48 is 100% while the accuracy for OneR is 98.5741%. Second comparison of RandomForest, which is a model of Trees with JRip, which is a model of Rules. In this comparison, the time to build for JRip is less than the time taken for RandomForest, when cross-validation folds is 10.

3. Examine the results of your models. Which is more understandable?
How would you present these results to the user? Which model do you prefer, and why?

There are parts of two models which I examined i.e., Trees and Rules. For the tree-based, it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. For the rule-based, it is sometimes problematic because those who want to comply with rules are not always sure of everything they need to look at. In the tree-based, the accuracy will be 100% but, in the rule-based, it is not sure the accuracy to be 100%. These are the reason I will prefer the Trees model. The graph is a way to present these results to the user. We can use matplotlib from which we can display scatter plot, bar plot, etc. In the Tree model, I would prefer a J48 model, as the time taken to build this model is 0.02 seconds which is less than all the examined models. Also, the accuracy of this model is 100%.