CS 422

Data Mining

Homework Assignment 1

Practicum Questions 1 and 4

Ronit Rudra
A20379221
rrudra@hawk.iit.edu

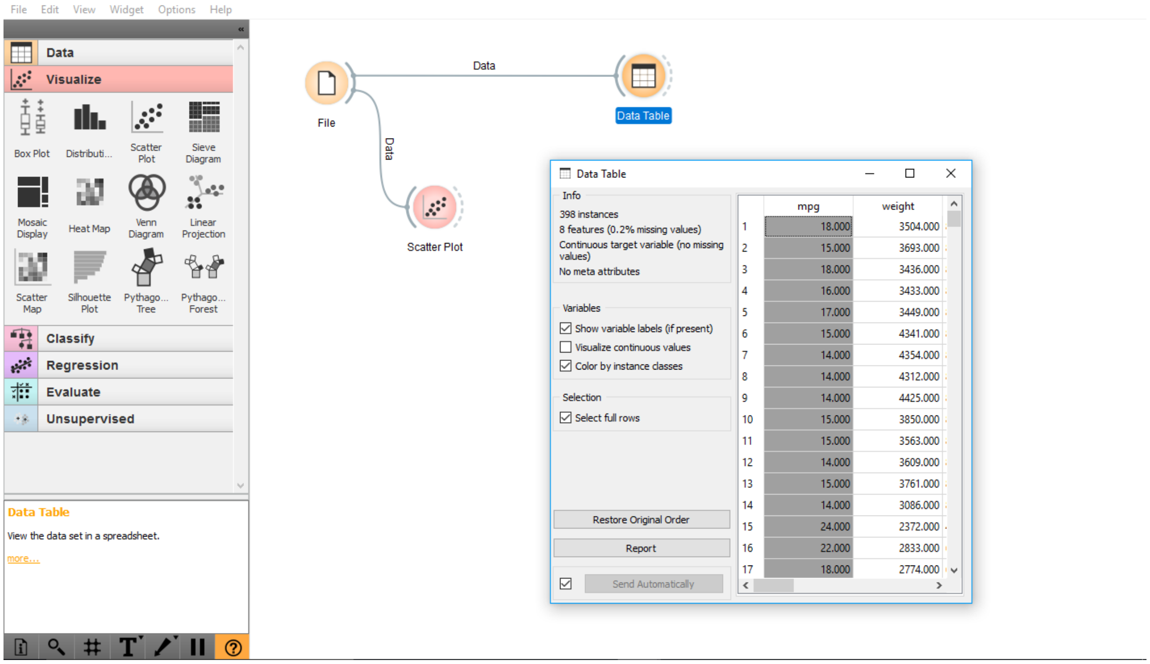
Course Instructor
Prof. Jawahar Panchal

1 Question 2:

Load the auto-mpg sample dataset into the Orange application, and visualize the dataset. Create a scatterplot between mpg and weight - what is the basic relationship between these variables using just visual inspection? Do the results make sense? Why?

- Creating appropriate workflow with *auto-mpg* dataset.

Figure 1: Workflow



Data Table

View the data set in a spreadsheet.

more...

Info

398 instances

8 features (0.2% missing values)

Continuous target variable (no missing values)

No meta attributes

Variables

☑ Show variable labels (if present)

☐ Visualize continuous values

☑ Color by instance classes

Selection

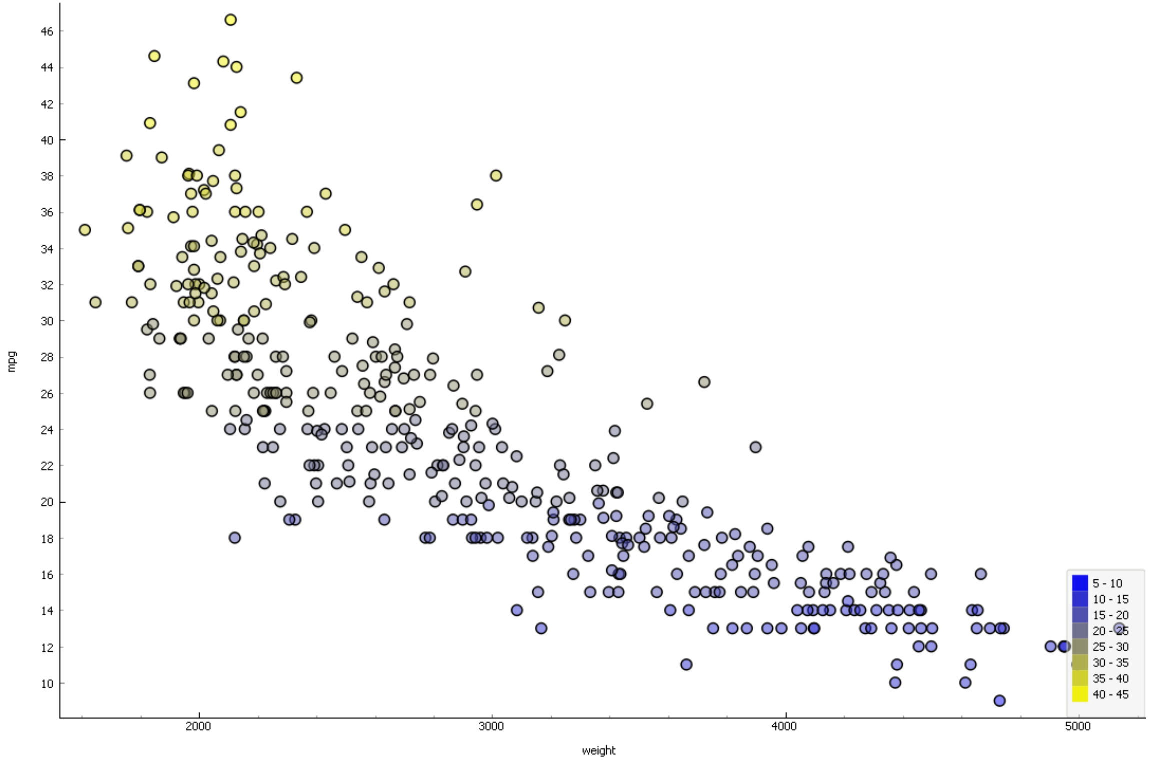
☑ Select full rows

☑ Send Automatically

	mpg	weight
1	18.000	3504.000
2	15.000	3693.000
3	18.000	3436.000
4	16.000	3433.000
5	17.000	3449.000
6	15.000	4341.000
7	14.000	4354.000
8	14.000	4312.000
9	14.000	4425.000
10	15.000	3850.000
11	15.000	3563.000
12	14.000	3609.000
13	15.000	3761.000
14	14.000	3086.000
15	24.000	2372.000
16	22.000	2833.000
17	18.000	2774.000

- Plotting a scatterplot between *weight* (x-axis) and *mpg* (y-axis)

Figure 2: Scatterplot of *mpg* vs *weight*



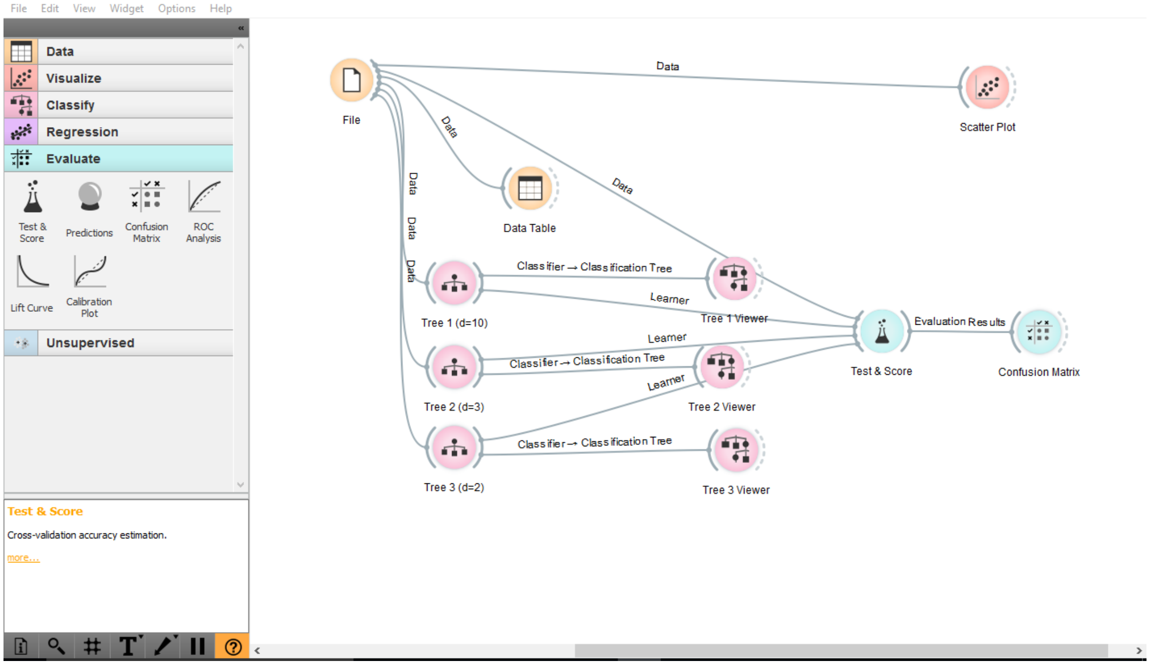
From Figure 2 it is observed that *mpg* and *weight* have an inverse non-linear relationship i.e. if *weight* increases then *mpg* decreases. This observation makes intuitive sense because more weight means a higher force required to change the speed of the vehicle and in turn consumes more gasoline. When the vehicle is lighter, the efficiency goes up as the inertia goes down.

2 Question 4:

Build two classification trees using the iris sample dataset within the Orange application. Keep all parameters for both classifiers the same (Feature Selection, Pruning), and modify the Limit Depth parameter to a smaller value than the default (e.g., from 10 to 2). How does this affect the Precision and Recall of the classifier? What types of flowers are misclassified? Why? What does Tan refer to as the border where these misclassifications occur?

- Creating appropriate workflow with *iris* dataset.

Figure 3: Workflow



Test & Score

Cross-validation accuracy estimation.

more...

- Creating 3 classification trees with varying depths, namely Tree 1 (depth= full), Tree 2 (depth=3) and Tree 3 (depth=2) while keeping all other parameters identical. The question asks for 2 trees but 3 were made to see just how classification accuracies change with varying depth.

Figure 4: Tree 1

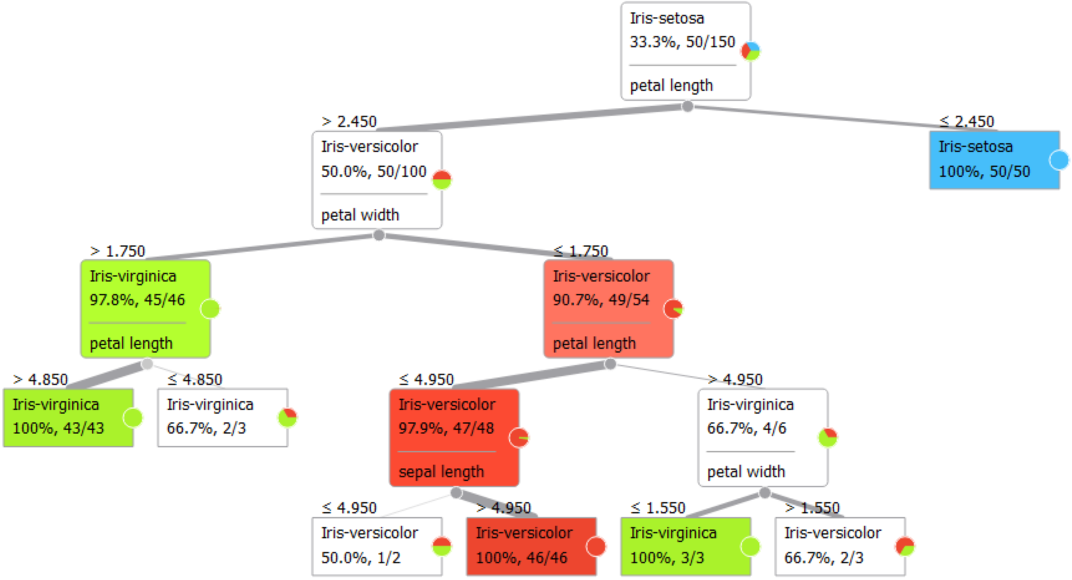


Figure 5: Tree 2

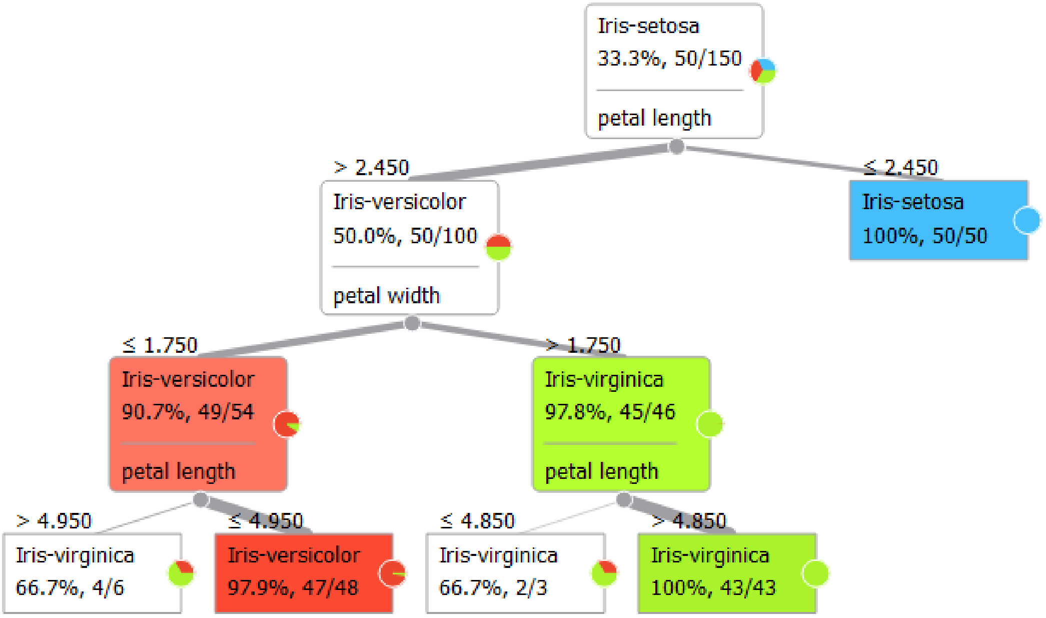
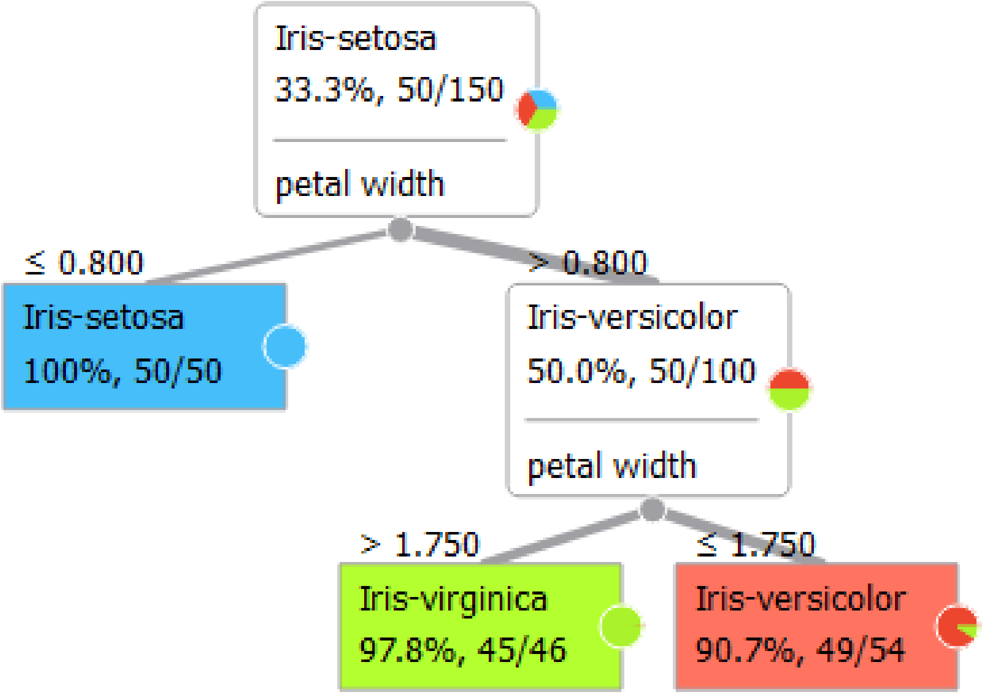
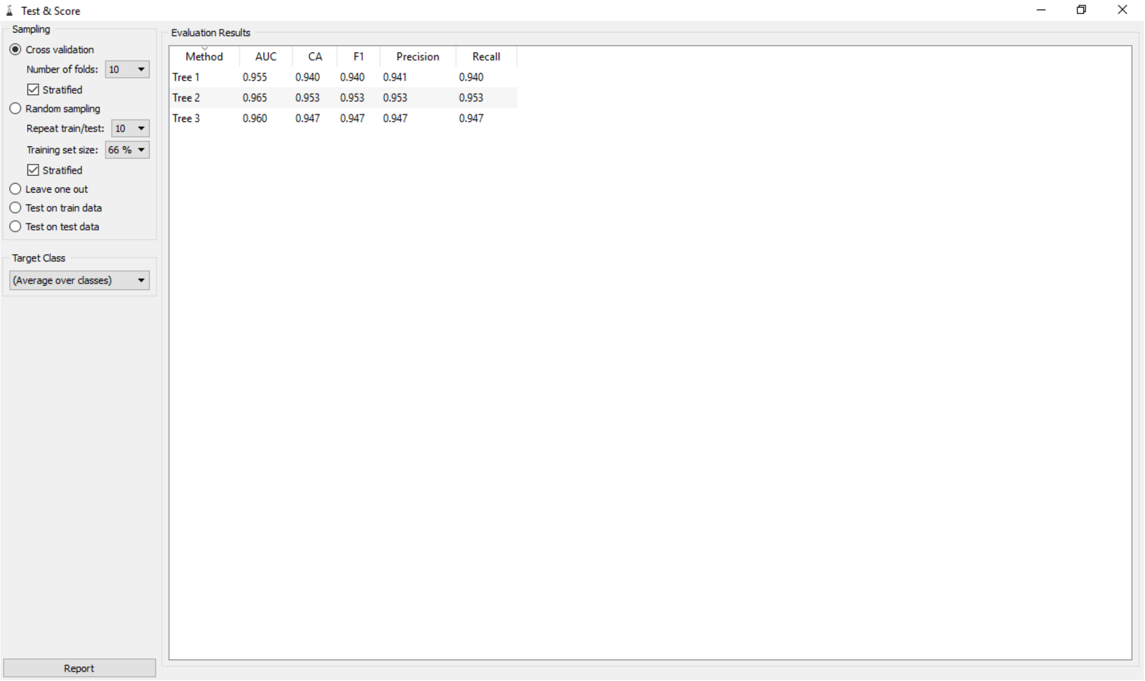


Figure 6: Tree 3



- The Results from the classification are:

Figure 7: Results



Test & Score

Sampling

● Cross validation

Number of folds: 10

☑ Stratified

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Tree 1	0.955	0.940	0.940	0.941	0.940
Tree 2	0.965	0.953	0.953	0.953	0.953
Tree 3	0.960	0.947	0.947	0.947	0.947

- Plotting the original data and predicted outcomes of the three classifiers. Scoring the plots resulted in *petal width* versus *petal length* being the most representative of the classification to be performed.

Figure 8: Original Data

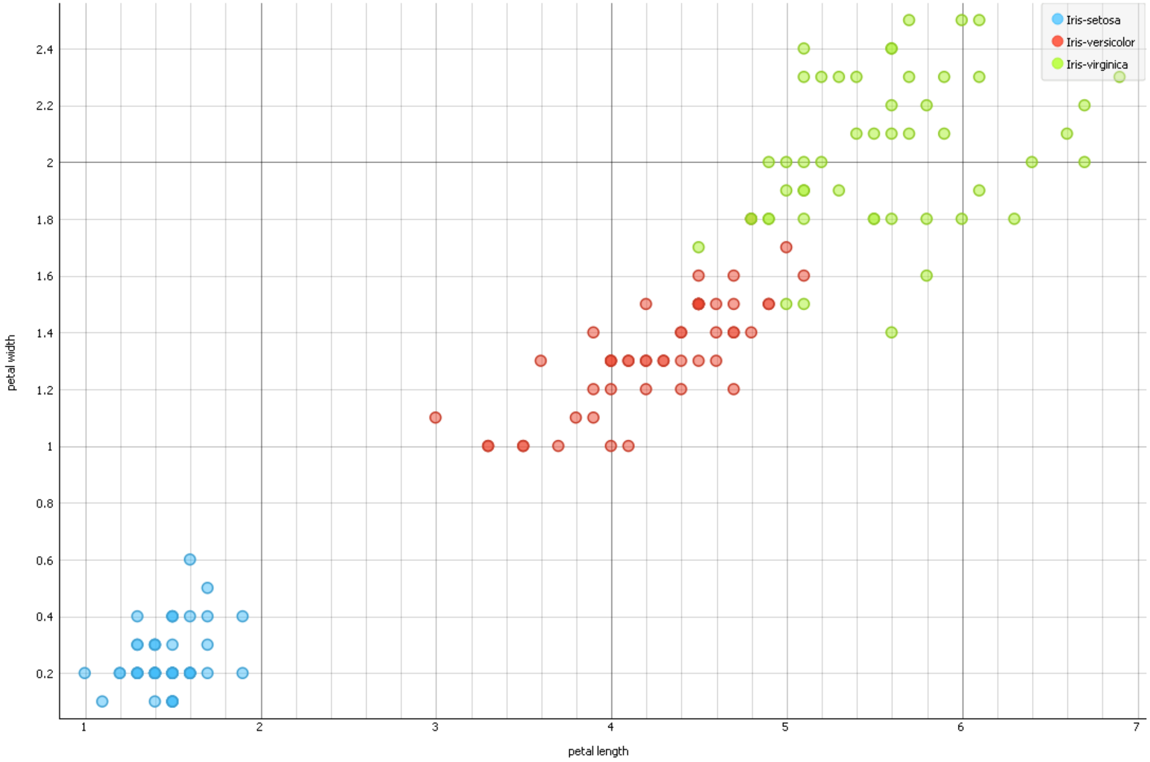


Figure 9: Tree 1 Predictions

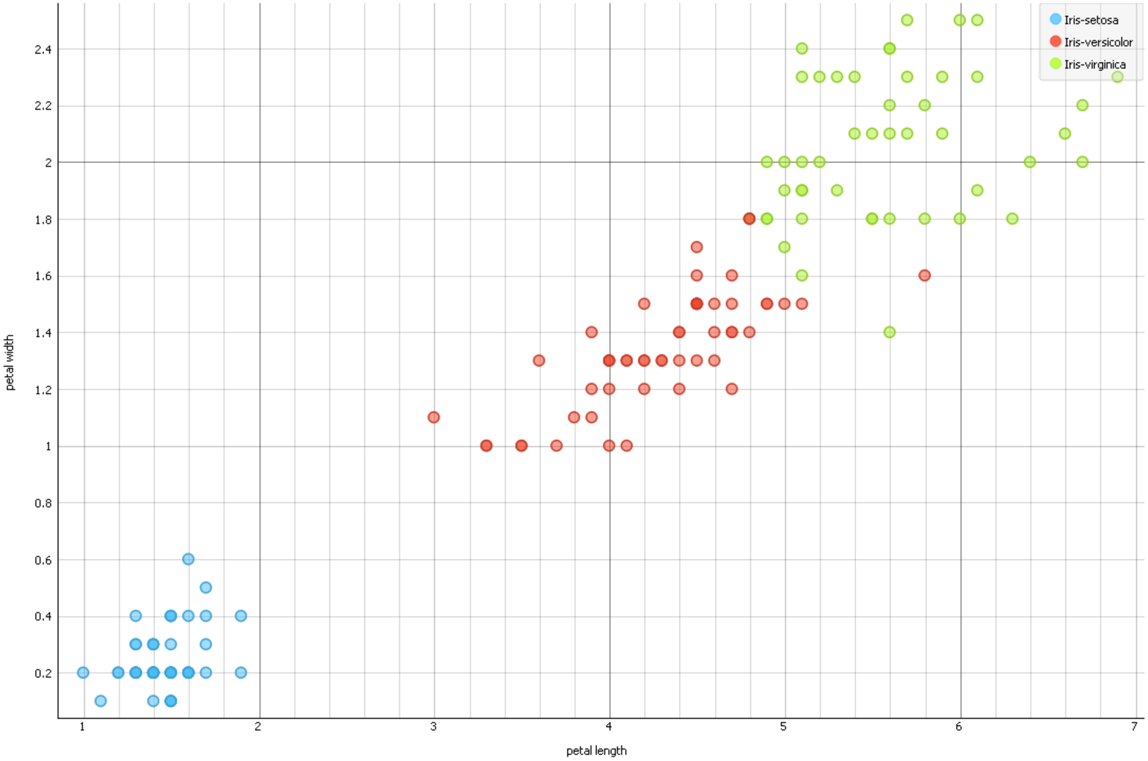


Figure 10: Tree 2 Predictions

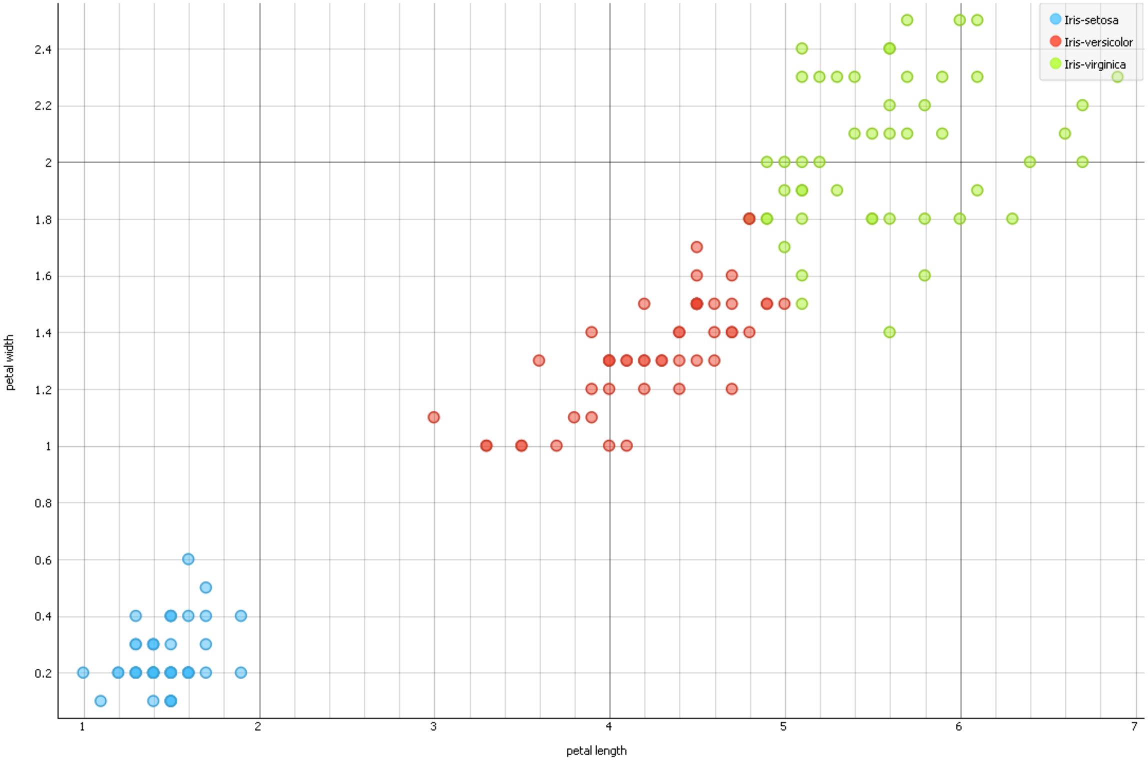
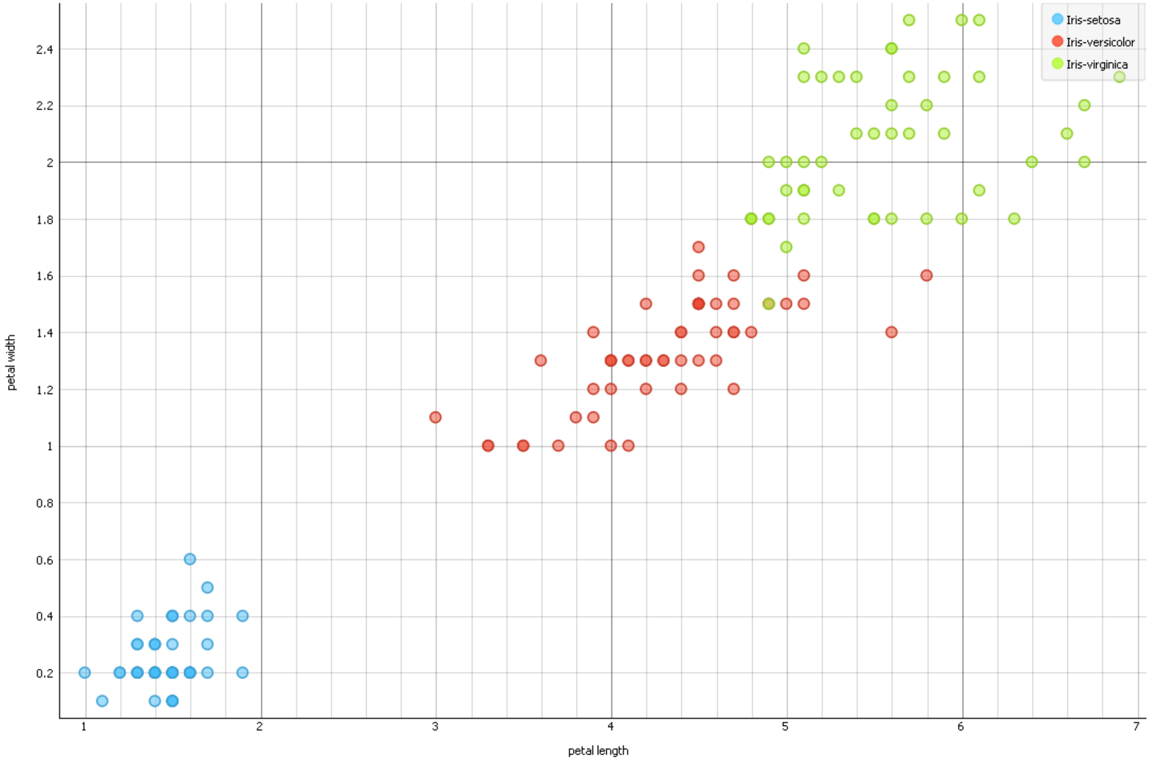


Figure 11: Tree 3 Predictions



All of the trees have somewhat similar classification accuracies with Tree 2 of height 3 being the highest. If only two trees were tested then one could have said that the precision and recall were higher in the tree with lower depth. But this is not the case if more than two depths are compared. The inverse relation does not hold. As the maximum depth a classifier reached was 4 for this dataset, it can be safely concluded that depth 3 has the highest accuracy, precision and recall (depth of 1 underfits). Speaking of misclassifications, all classifiers correctly classified Iris Setosa (it's cluster is clearly separable by a hyperplane in 2 dimensions which is a straight line). For Virginica and Versicolor, the misclassifications occured on the boundary between the two overlapping clusters as they are not easily separable (Figures 9 to 11 show how different classifiers misclassify at that particular region). Tan refers to this border of misclassification as the **Decision Boundary**.