MACHINE LEARNING TUTORIAL 1 REPORT

Enrolment Number: 15535029
Name: Prakhar Dhama
Class: M.Tech CSE 1st Yr.

PROBLEM 1

Objective

Construct a decision tree for abalone dataset. Also,

1. Compare accuracy of the model for all the 29 classes with the accuracy provided with dataset for previous experiments.
2. Compare accuracy of the model by treating data set as a 3-category classification problem (grouping ring classes 1-8, 9 and 10, and 11 on) with accuracy provided with dataset.

Total Classes in Dataset

There are 29 ring classes in the given dataset.

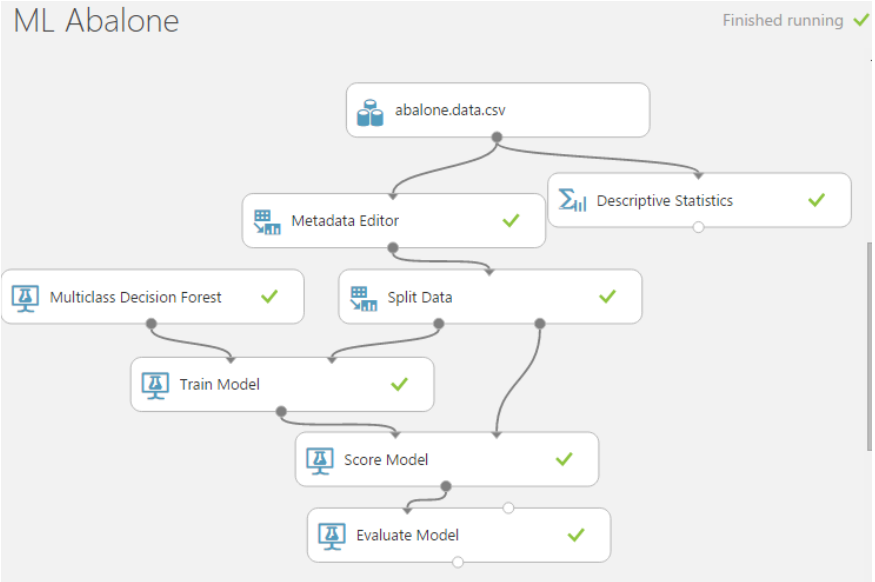
Methodology

1. Tools Used: Microsoft Azure Machine Learning cloud computing platform and R programming language.
2. Features/Preprocessing:
    - Apart from sex field all other are numerical so it is made categorical using meta data editor module.
    - R script is used for grouping the ring classes for second problem.

Results

Problem I: With 29 Classes

A. Model

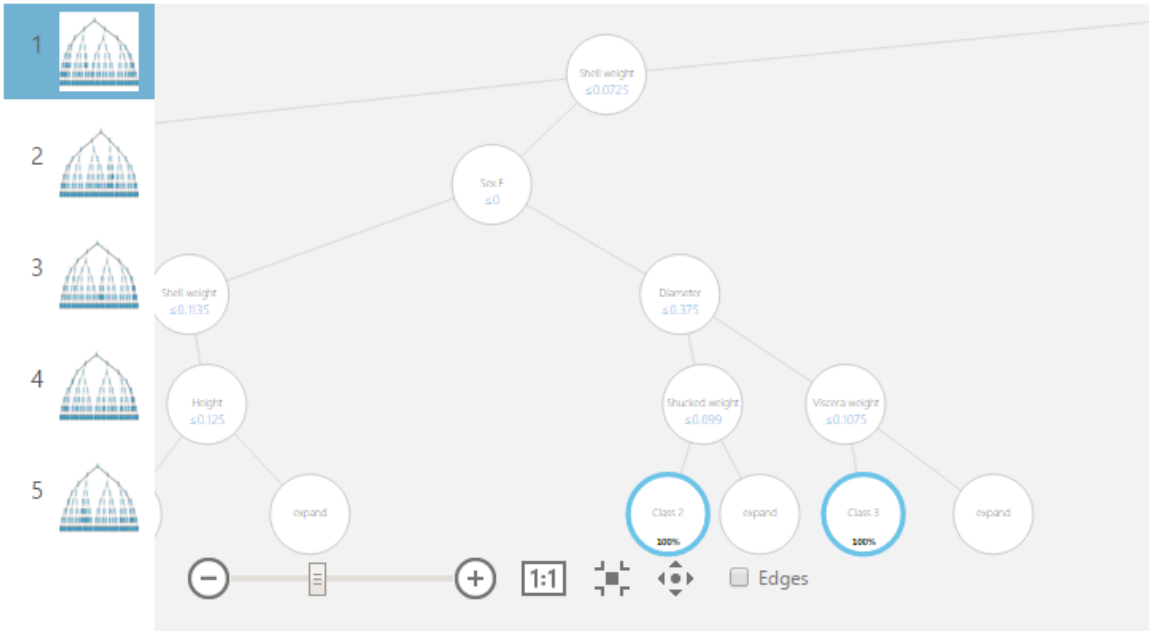


B. Sample Tree

ML Abalone › Train Model › Trained model

trees constructed

5



C. Accuracy

Our Accuracy: **27.5%**

ML Abalone › Evaluate Model › Evaluation results

Metrics

Overall accuracy	0.275862
Average accuracy	0.948276
Micro-averaged precision	0.275862
Macro-averaged precision	NaN
Micro-averaged recall	0.275862
Macro-averaged recall	NaN

Accuracy Mentioned with dataset for previous experiments: 21~26%

```
Sam Waugh (1995) "Extending and benchmarking Cascade-Correlation", PhD
thesis, Computer Science Department, University of Tasmania.

-- Test set performance (final 1044 examples, first 3133 used for training):
 24.86% Cascade-Correlation (no hidden nodes)
 26.25% Cascade-Correlation (5 hidden nodes)
 21.5%  C4.5
  0.0%  Linear Discriminate Analysis
  3.57% k=5 Nearest Neighbour
   (Problem encoded as a classification task)
```

D. Confusion Matrix

ML Abalone › Evaluate Model › Evaluation results

Confusion Matrix

Predicted Class

Actual Class	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	2	24	25	26	27	29
1																							100.0%					
3			50.0%																				50.0%					
4			58.3%	33.3%	8.3%																							
5			19.2%	46.2%	19.2%	11.5%	3.8%																					
6			3.1%	12.5%	20.3%	39.1%	17.2%	4.7%	1.6%			1.6%																
7				4.2%	12.6%	42.0%	26.9%	10.9%	2.5%	0.8%																		
8				0.7%	4.6%	12.4%	32.0%	35.3%	11.1%	1.3%		2.0%	0.7%															
9					1.1%	6.1%	23.3%	37.2%	26.7%	4.4%	0.6%	0.6%																
10						3.2%	7.6%	33.1%	38.2%	15.9%		1.3%	0.6%															
11				0.9%	0.9%	2.7%	5.5%	29.1%	30.9%	22.7%	1.8%	2.7%	1.8%						0.9%									
12						1.6%	9.7%	17.7%	46.8%	11.3%	1.6%	6.5%		1.6%	3.2%													
13							14.6%	20.8%	20.8%	22.9%	2.1%	12.5%	2.1%	2.1%	2.1%													
14						3.8%	3.8%		46.2%	26.9%	3.8%	15.4%																
15						3.7%	7.4%	11.1%	37.0%	14.8%	11.1%	7.4%	3.7%					3.7%										

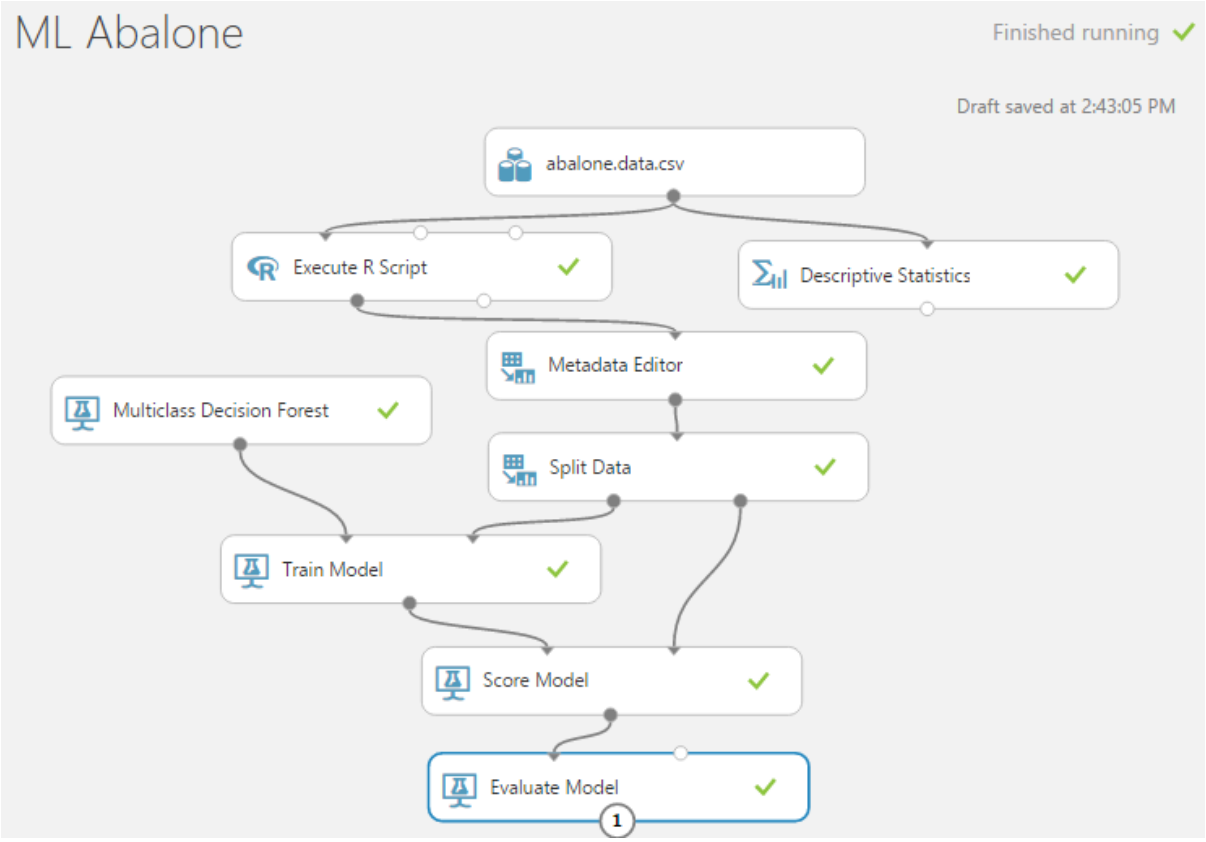
Problem II: With 3 Groups

A. R Script

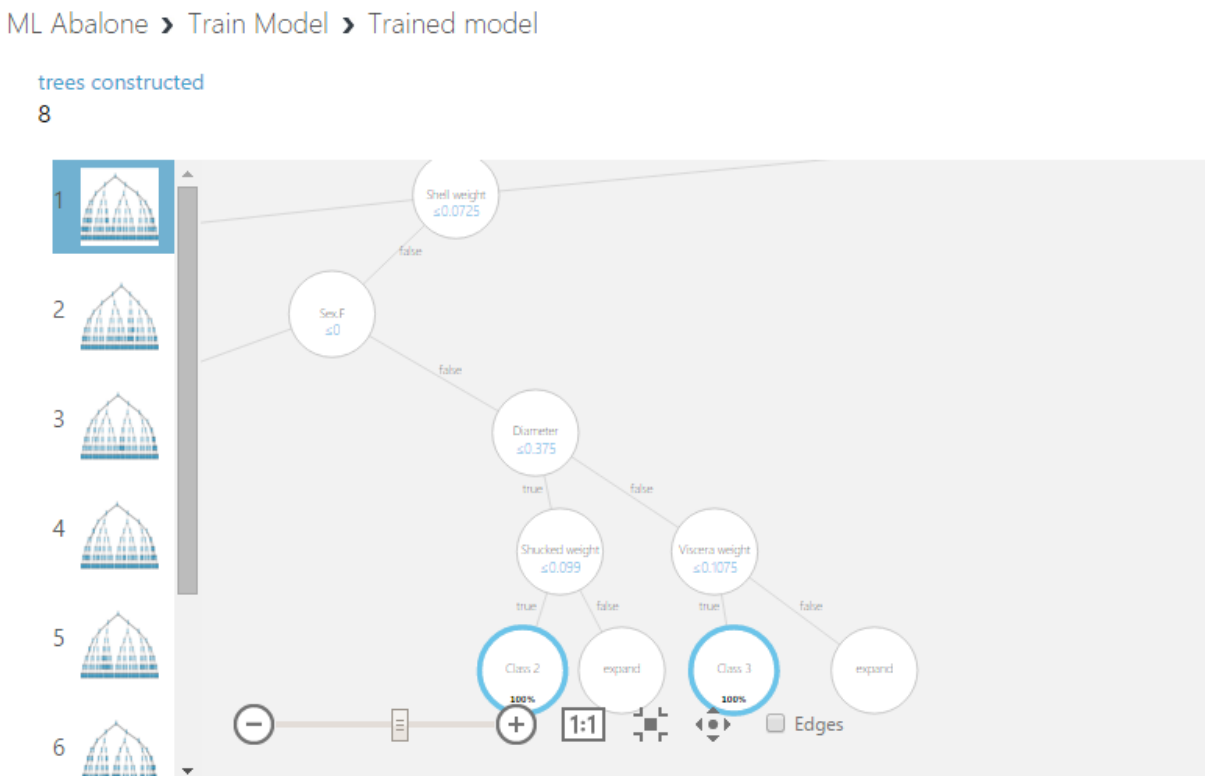
R Script

```
# Map 1-based optional input ports to variables
abalone <- maml.mapInputPort(1) # class: data.frame
RingsGroup <- abalone$Rings
RingsGroup[RingsGroup<=8] <- 1
RingsGroup[RingsGroup==9] <- 2
RingsGroup[RingsGroup==10] <- 2
RingsGroup[RingsGroup>=11] <- 3
RingsGroup[RingsGroup==1] <- "Class 1"
RingsGroup[RingsGroup==2] <- "Class 2"
RingsGroup[RingsGroup==3] <- "Class 3"
abalone$Rings <- NULL
data.set <- cbind(abalone, RingsGroup)
# Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("data.set");
```

B. Model



C. Sample Tree



D. Accuracy

Our accuracy: **65.8%**

ML Abalone › Evaluate Model › Evaluation results

Metrics

Overall accuracy	0.658046
Average accuracy	0.772031
Micro-averaged precision	0.658046
Macro-averaged precision	0.651906
Micro-averaged recall	0.658046
Macro-averaged recall	0.65484

Accuracy Mentioned with dataset for previous experiments: 59~65%

```
David Clark, Zoltan Schreter, Anthony Adams "A Quantitative Comparison of
Dystal and Backpropagation", submitted to the Australian Conference on
Neural Networks (ACNN'96). Data set treated as a 3-category classification
problem (grouping ring classes 1-8, 9 and 10, and 11 on).

-- Test set performance (3133 training, 1044 testing as above):
 64%    Backprop
 55%    Dystal
-- Previous work (Waugh, 1995) on same data set:
 61.40% Cascade-Correlation (no hidden nodes)
 65.61% Cascade-Correlation (5 hidden nodes)
 59.2%  C4.5
 32.57% Linear Discriminate Analysis
 62.46% k=5 Nearest Neighbour
```

E. Confusion Matrix

ML Abalone › Evaluate Model › Evaluation results

Confusion Matrix

Predicted Class

Actual Class	Class 1	Class 2	Class 3
Class 1	74.8%	19.4%	5.8%
Class 2	19.0%	50.7%	30.3%
Class 3	8.5%	20.6%	70.9%

PROBLEM 2

Objective

Clustering the dataset using K Means on attribute Plant Name based on their locations.

Total Classes in Dataset

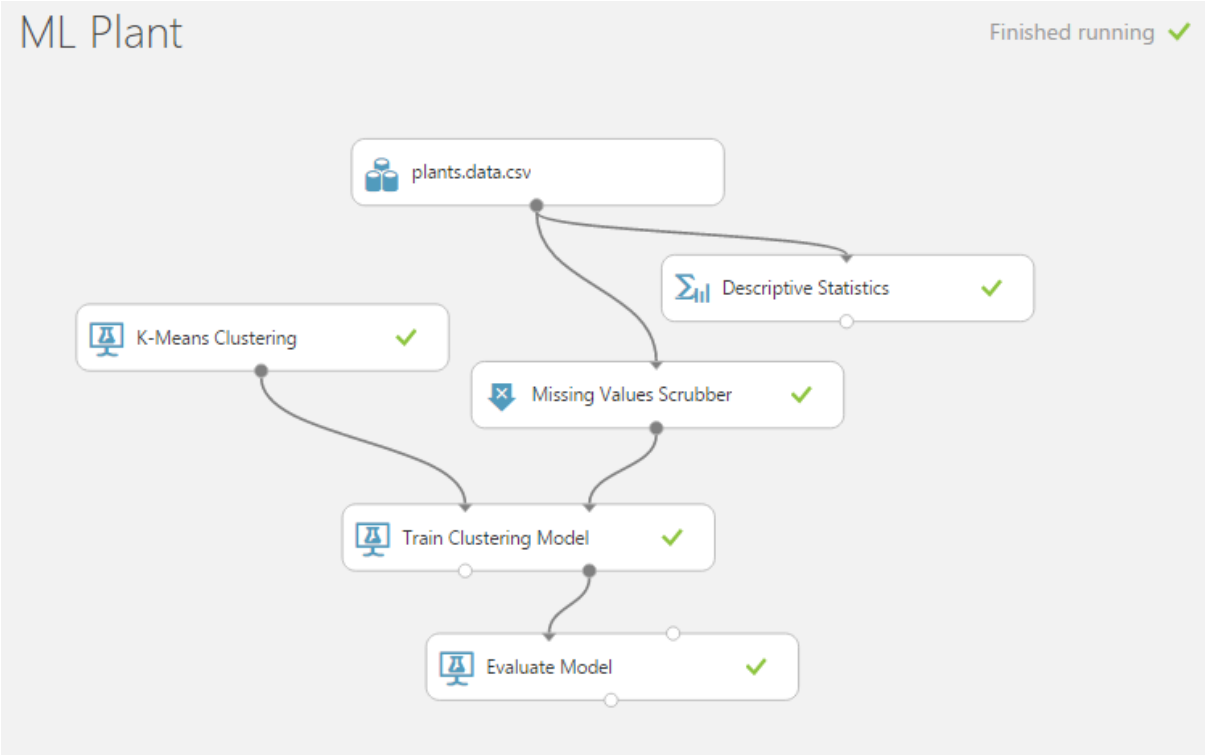
There are 2 classes in the given dataset, one US and other is Canada.

Methodology

1. Tools Used: Microsoft Azure Machine Learning cloud computing platform and R programming language.
2. Features/Preprocessing:
    - Fields with states as dengl (Denmark) and fraspm (France) are not included in clustering using Scrubber module.

Results

A. Model



B. Result

Plants clustered in cluster 0: **16.8%**

Plants clustered in cluster 1: **83.2%**

ML Plant › Evaluate Model › Evaluation results

rows 3 columns 5

view as

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	2.75244	5.680784	34781	9.640171
Evaluation For Cluster No.0	6.20971	7.727542	5837	9.640171
Evaluation For Cluster No.1	2.055228	5.268024	28944	5.173279

PROBLEM 3

Objective

Classification of the dataset using KNN on attribute Class Name.

Total Classes in Dataset

There are 3 classes in the given dataset i.e. (49 balanced, 288 left, 288 right).

Methodology

1. Tools Used: Microsoft Azure Machine Learning cloud computing platform and R programming language.
2. Features/Preprocessing:
    - Class name is made categorical using meta data editor.

Results

A. Accuracy

ML Balance-Scale › Evaluate Model › Evaluation results

Metrics

Overall accuracy	0.8
Average accuracy	0.866667
Micro-averaged precision	0.8
Macro-averaged precision	0.622222
Micro-averaged recall	0.8
Macro-averaged recall	0.606481

B. Confusion Matrix

ML Balance-Scale › Evaluate Model › Evaluation results

Confusion Matrix

Predicted Class

Actual Class	B	L	R
B	11.1%	44.4%	44.4%
L	10.7%	87.5%	1.8%
R	13.3%	3.3%	83.3%