CME 4403 Introduction to Machine Learning

Term Project Report

Berkay ÇELİK 2013510085

May, 2018

DATA SET INFORMATION

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope - a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

**Number of Instances**: 4177

**Number of Attributes**: 8

**Attribute information:**

Given is the attribute name, attribute type, the measurement unit and a

brief description. The number of rings is the value to predict: either

as a continuous value or as a classification problem.

**Name Data Type Meas. Description**

---- --------- ----- -----------

Sex nominal M, F, and I (infant)

Length continuous mm Longest shell measurement

Diameter continuous mm perpendicular to length

Height continuous mm with meat in shell

Whole weight continuous grams whole abalone

Shucked weight continuous grams weight of meat

Viscera weight continuous grams gut weight (after bleeding)

Shell weight continuous grams after being dried

Rings integer +1.5 gives the age in years

**Statistics for numeric domains:**

**Length Diam Height Whole Shucked Viscera Shell Rings**

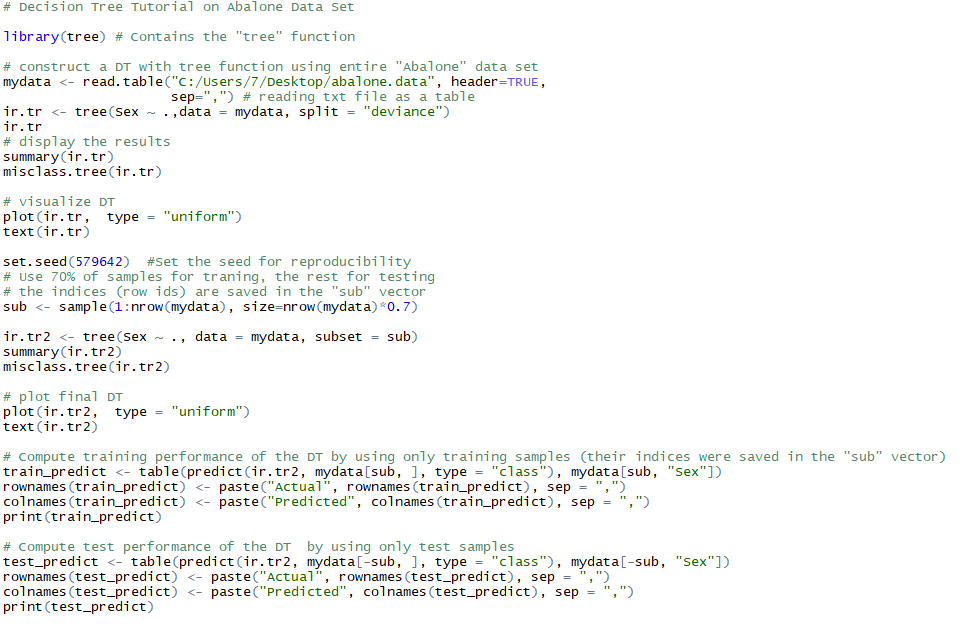
Min 0.075 0.055 0.000 0.002 0.001 0.001 0.002 1

Max 0.815 0.650 1.130 2.826 1.488 0.760 1.005 29

Mean 0.524 0.408 0.140 0.829 0.359 0.181 0.239 9.934

SD 0.120 0.099 0.042 0.490 0.222 0.110 0.139 3.224

Correl 0.557 0.575 0.557 0.540 0.421 0.504 0.628 1.0



Firstly, data set is read as a table onto variable that named mydata. The mydata is integrated onto a tree structure. The target feature is “sex” column. Other columns are Length Diameter, Height, WholeWeight, ShuckedWeight, VisceraWeight, ShellWeight, Rings.

DECISION TREE

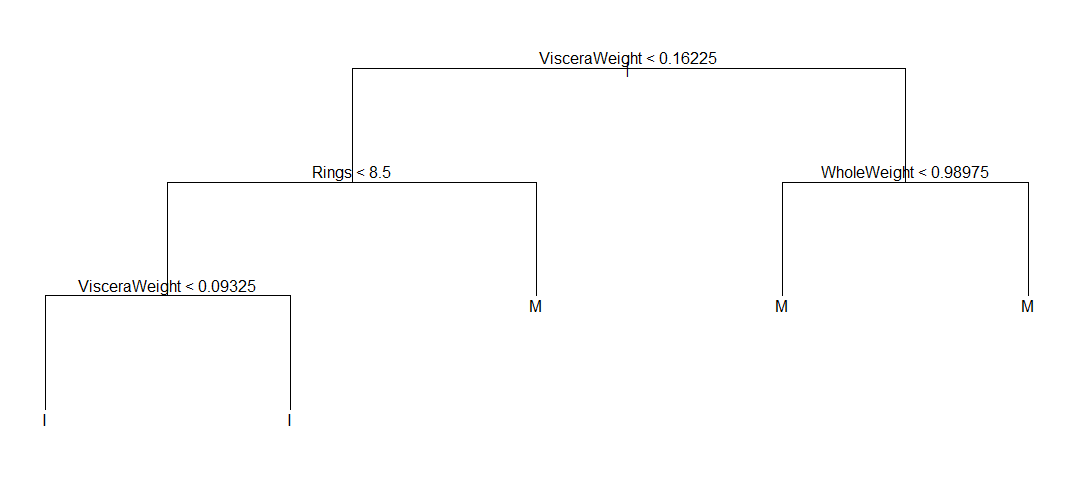


Figure : Decision tree made from all data

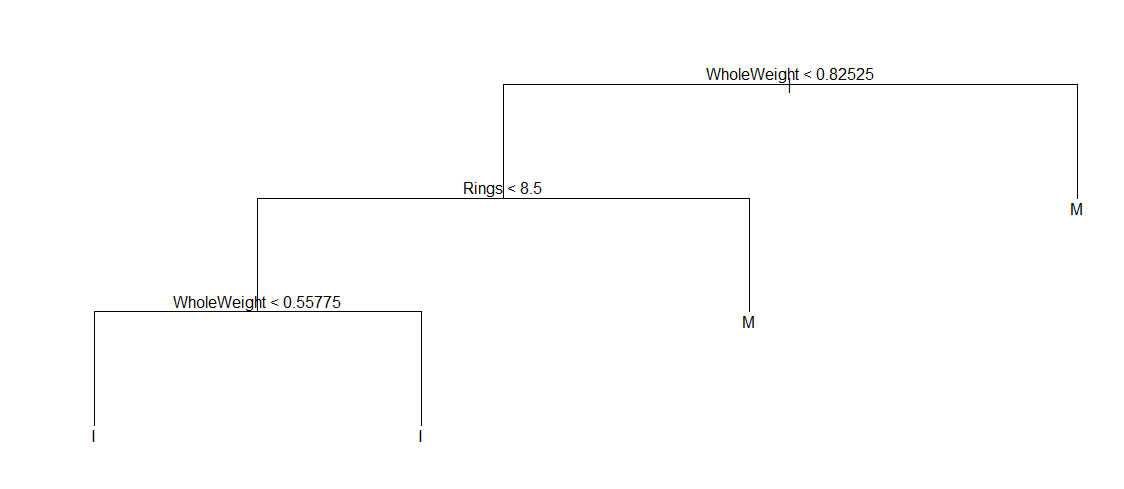
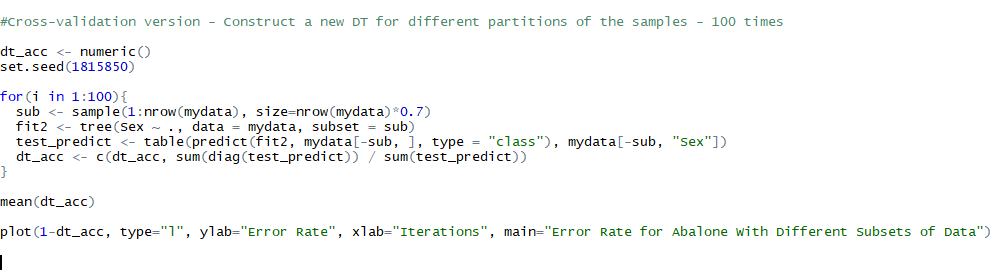


Figure : Decision tree use 70% of samples for traning, the rest for testing

It is possible to test the correctness and consistency of the machine learning model. Although it gives the correct results for the data in our hand, it may not work consistently for all of the data. For this reason, validation methods are needed. In our project, we used the cross validation method with 100 repetitions.



K-Nearest Neighbor

The k-nearest neighbors algorithm is a non-parametric method used for classification and regression. The input consist of the k closest training examples in feature space. The output is a class membership. An object is classified by a majority vote of its neighbors, The object is assigned to class most common neighbor among its k nearest neighbor.

