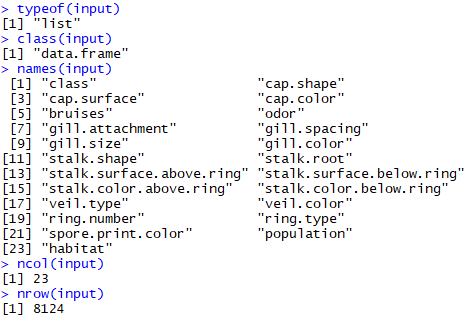
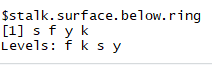
**CS358 Homework 1**

นาย สุรพล กุลเสวตร์ 5709680085 (https://github.com/kokore/CS358HW1\_-DecisionTree)

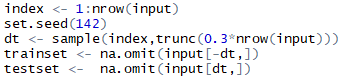
1. การเตรียมชุดข้อมูล (Data acquisition)
   1. โหลด data mushrooms จาก Kaggle https://www.kaggle.com/uciml/mushroom-classification
   2. เปลื่ยน path เป็นโพลเดอร์ที่เก็บไฟล์
   3. ใช้คำสั่งอ่านไฟล์ read.csv ไปที่ตัวแปร data  
      
   4. ใช้คำสั่ง typeof เพื่อดู type ของข้อมูล คำสั่ง class นิยาม type ของข้อมูล คำสั่ง names เพื่อดูชื่อทุกๆ attribute คำสั่ง ncol เพื่อดูจำนวนหลัก คำสั่ง nrow เพื่อดูแถว  
        
      การใช้คำสั่ง input$ตามด้วยชื่อ attribute และ คำสั่ง unique()   
        
      จะเห็นค่าที่เป็นไปได้ทั้งหมดของ attribute นี้โดยค่า attribute จะไม่ซ้ำกัน ทำให้สรุปข้อมูลได้ดังนี้ ข้อมูลเห็ดเป็นชนิด data.frame ที่มี attribute เป็น list type มีทั้งหมด 23 attribute มีแถวทั้งหมด 8124 แถว และ หลักทั้งหมด 23 หลัก ซึ่ง attribute ในแต่ละ attribute จะให้จะมีค่าได้แก่  
        
      attribute class ซึ่งให้ค่า e และ p   
        
        
      attribute cap-shape ให้ค่า b c x f k และ s   
        
        
        
      attribute cap-surface ให้ค่า f g y และ s  
        
      attribute cap-color ให้ค่า n b c g r p u e w และ y  
        
        
      attribute bruises ให้ค่า t และ f  
        
        
      attribute odor ให้ค่า a l c y f m n p และ s  
        
        
      attribute gill-attachment ให้ค่า a และ f   
        
        
      attribute gill-spacing ให้ค่า c และ w   
        
        
      attribute gill-size ให้ค่า b และ n  
        
        
      attribute gill-color ให้ค่า k n e o p r u w y b h และ g  
        
        
      attribute stalk-shape ให้ค่า e และ t  
        
      attribute stalk-root ให้ค่า b c e r และ ?  
        
        
      attribute stalk-surface-above-ring ให้ค่า f y k และ s  
        
        
      attribute stalk-surface-below-ring ให้ค่า f y k และ s  
        
        
      attribute stalk-color-above-ring ให้ค่า n b c g o p e w และ y  
        
        
      attribute stalk-color-above-ring ให้ค่า n b c g o p e w และ y  
        
        
      attribute veil-type ให้ค่า p   
        
        
      attribute veil-color ให้ค่า n o w และ y  
        
        
        
      attribute ring-number ให้ค่า n o และ t  
        
        
      attribute ring-type ให้ค่า c e f l n และ p   
        
        
      attribute spore-print-colorให้ค่า k n b h r o u w และ y  
        
        
      attribute population ให้ค่า a c n s v และ y  
        
        
      attribute habitat ให้ค่า g l m p u w และ d
2. การแบ่งข้อมูลเพื่อ Train และ Test แบบจำลอง (Data partitioning)
   1. สร้างตำแหน่ง 1 ถึง แถวสุดท้ายของข้อมูลเห็ด

Set.seed เพื่อให้ข้อมูลกระจายตัว

Dt แบ่งข้อมูลโดยใช้คำสั่ง sample โดยใช้ค่าพารามิเตอร์ 2 ตัวได้แก่ index คือ ตำแหน่งของทุก ๆ ตัวในข้อมูลเห็น trunc คือ จะส่งจำนวนเต็มที่ตัดออกจากแถวของข้อมูลเห็ดคือ 30 %

Trainset คือข้อมูลเห็ด 70 % โดยna.omit คือฟังก์ชั่นที่จะนำmissing values

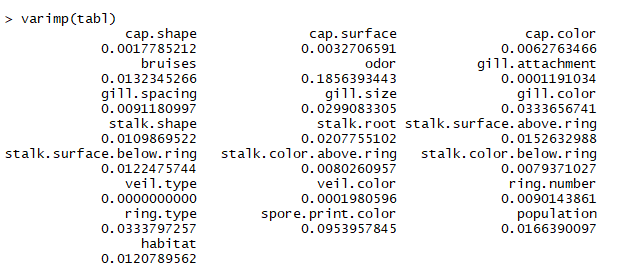
Testset คือข้อมูลเห็ด 30 % โดยna.omit คือฟังก์ชั่นที่จะนำmissing values



1. การเลือก Attribute เพื่อสร้างแบบจำลอง (Attribute selection)

ผมได้เลือก Random Forest Metod เป็นตัวเลือก Attribute  
ใช้หลักการ Decision Tree ในการเลือก Attribute  
การสุ่มเลือก attribute ของข้อมูลเพื่อเลือก features  ขึ้นมา split tree และสร้าง tree ต้นใหม่ไปเรื่อยๆเป็นร้อยเป็นพันต้น แล้วค่อยเฉลี่ยผล prediction

1. การแสดงภาพเกี่ยวกับ Attribute ที่เลือก (Attribute visualization)  
   เริ่มจากการ install package party  
     
   เรียกใช้ library  
     
   เรียกใช้คำสั่ง cforest ไปที่ตัวแปล tabl โดยใช้พารามิเตอร์ 2 ตัวคือ พารามิเตอร์คือ class และ attribute อื่น ๆ ที่ไม่ใช่class พารามิเตอร์อีกตัวคือข้อมูลเห็ดทั้งหมด

Varimp ทำการสุ่ม attribute ไปคำนวณ decision tree เรื่อย ๆ และค่อยเฉลี่ยนผล prediction  
  


1. Classification with Decision tree

ทำการลง package data.tree

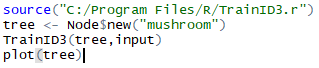


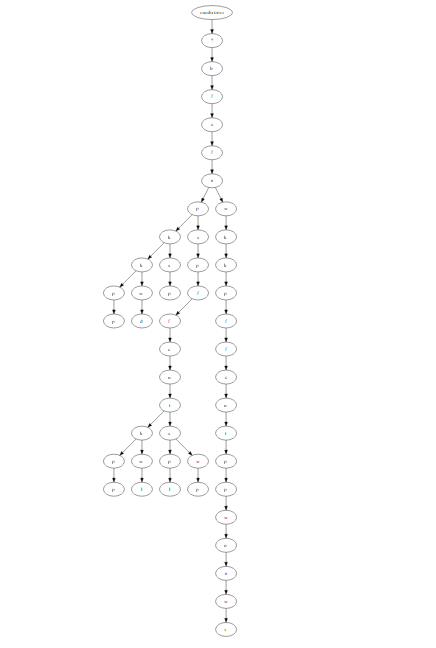
เรียกใช้ library



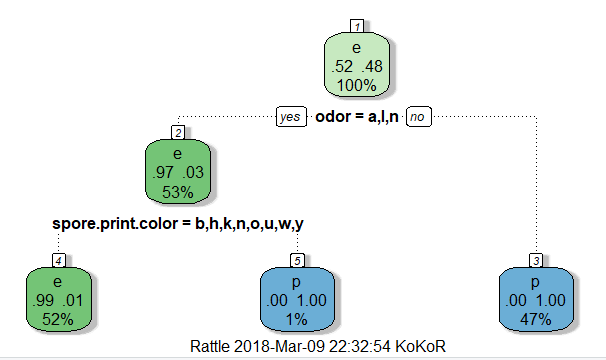
สร้าง tree โดยให้ root node ชื่อ mushroom

เรียกใช้ TrainID3.r เพื่อหาค่า informationgain ในการหาค่า attribute ที่จะนำมาสร้าง Child





ทำการลง package rpart  
  
เรียกใช้ library  
  
เรียกใช้ฟังก์ชั่น rpart โดยเรียกใช้พารามิเตอร์ 2 ตัวคือ พารามิเตอร์คือ class และ attribute อื่น ๆ ที่ไม่ใช่class พารามิเตอร์ method คือ พารามิเตอร์อีกตัวคือข้อมูลเห็ดทั้งหมด

1. สรุป  
   จากการจำลองสร้าง Decision Tree Model สำหรับข้อมูล Mushroom ทำให้ผมได้เรียนรู้การดึงข้อมูลจากไฟล์ csv มายังตัวแปรและยังช่วยทำความเข้าใจข้อมูลนั้น ๆ โดยใช้คำสั่งต่าง ๆ จากภาษาR การแบ่งข้อมูลออกเป็น Train และ Test เพื่อนำไปวิเคราะห์เปลี่ยบเทียบข้อมูลความถูกต้อง การหาAttributeที่มีความสำคัญโดยใช้ informationgain และ random forst การสร้างDecision Treeโดยใช้package ต่าง ๆ ทั้ง data.tree และ rpart