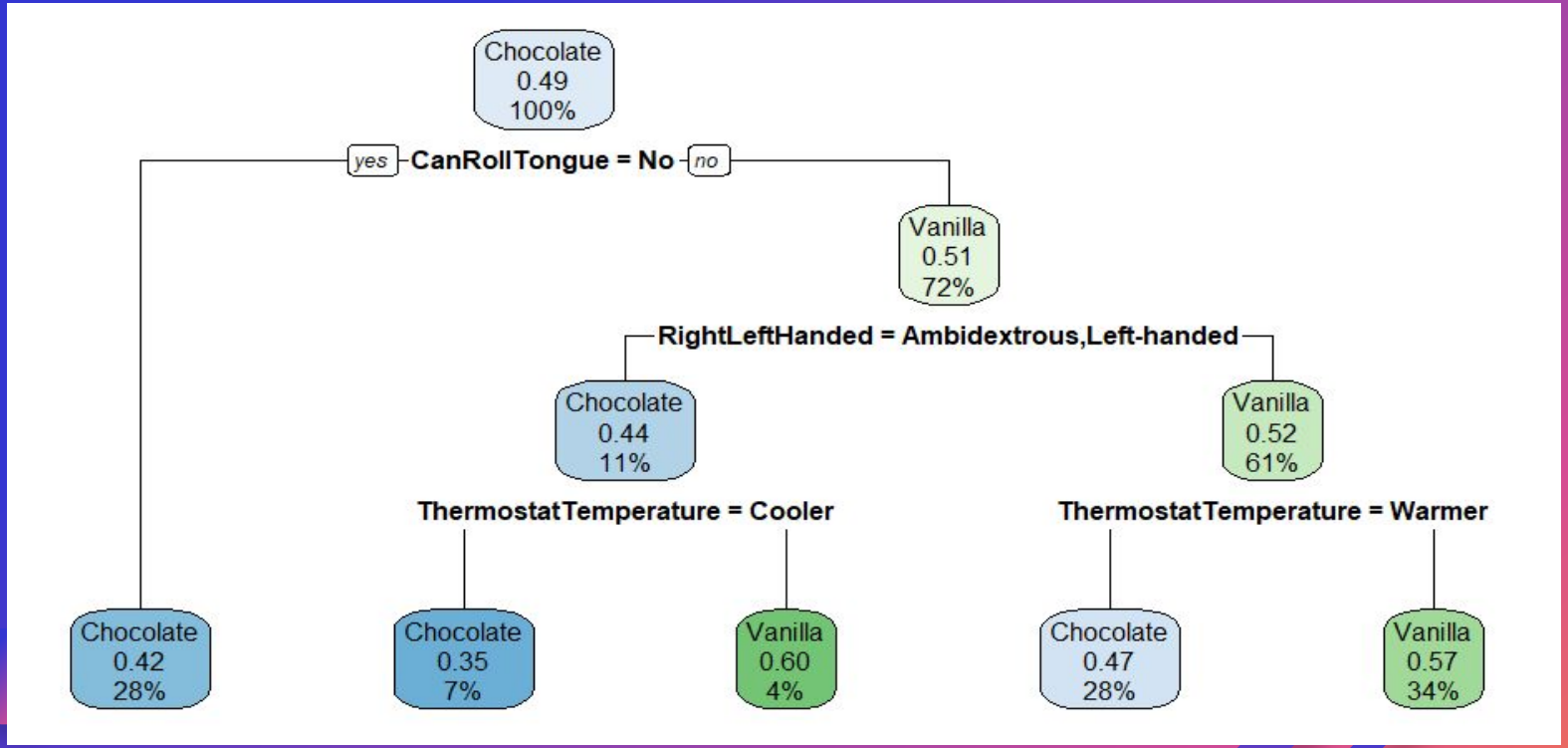
ASSIGNMENT 10
DATA TREE

~DHRUVAL PATEL

Tree 1

tree1=rpart(ChocolateOrVanilla~RightLeftHanded+CanRollTongue+ThermostatTemperature,Expanded)



Identify the best relative error,Test 1

```
> printcp(tree1)

Classification tree:
rpart(formula = ChocolateOrVanilla ~ RightLeftHanded + CanRollTongue +
    ThermostatTemperature, data = Expanded)

Variables actually used in tree construction:
[1] CanRollTongue         RightLeftHanded       ThermostatTemperature

Root node error: 120/247 = 0.48583

n= 247

        CP nsplit rel error xerror     xstd
1 0.033333      0   1.00000 1.0833 0.065394
2 0.029167      1   0.96667 1.2167 0.064388
3 0.016667      3   0.90833 1.2167 0.064388
4 0.010000      4   0.89167 1.0667 0.065441
```

Whe we use printcp() without any interpretation of the minbucket or minspilt. It gives an error of 0.89 which is <0.95.
But we will test the graph with minsplit and minbucket to get the the lowest error

Use of minsplit in tree 1,test 2

```
> printcp(rpart(ChocolateOrVanilla~RightLeftHanded+CanRollTongue+ThermostatTemperature,data=Expanded,method="class", minsplit=35))

Classification tree:
rpart(formula = ChocolateOrVanilla ~ RightLeftHanded + CanRollTongue +
    ThermostatTemperature, data = Expanded, method = "class",
    minsplit = 35)

Variables actually used in tree construction:
[1] CanRollTongue         RightLeftHanded       ThermostatTemperature

Root node error: 120/247 = 0.48583

n= 247

        CP nsplit rel error xerror     xstd
1 0.033333      0   1.00000 1.0000 0.065458
2 0.029167      1   0.96667 1.2167 0.064388
3 0.010000      3   0.90833 1.1333 0.065148
```

When we use the mini split function at 35, we get an error which is greater that the original . Thus we reject this method.

Use of minbucket in tree 1, test 3

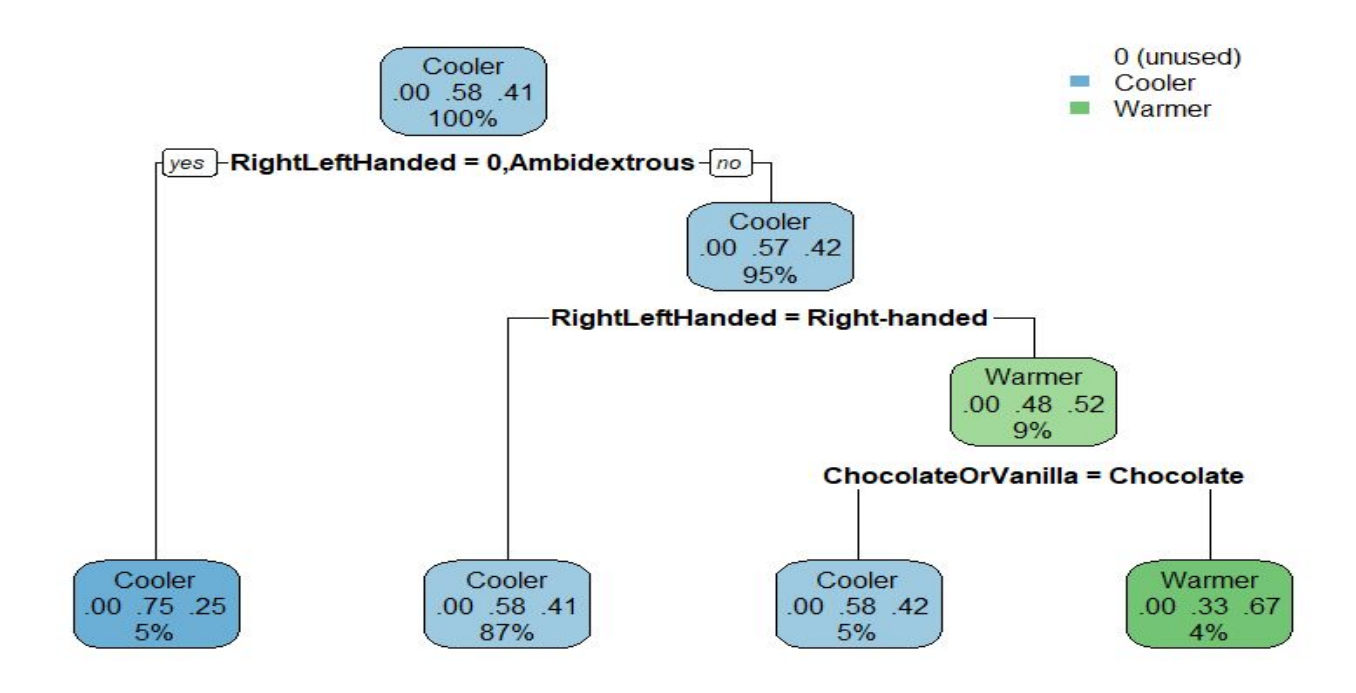
```
> printcp(rpart(ChocolateOrVanilla~RightLeftHanded+CanRollTongue+ThermostatTemperature,data=Expanded,method="class", minbucket=35))

Classification tree:
rpart(formula = ChocolateOrVanilla ~ RightLeftHanded + CanRollTongue +
    ThermostatTemperature, data = Expanded, method = "class",
    minbucket = 35)

Variables actually used in tree construction:
[1] CanRollTongue         ThermostatTemperature

Root node error: 120/247 = 0.48583

n= 247

        CP nsplit rel error xerror     xstd
1 0.033333      0   1.00000 1.0750 0.065419
2 0.016667      1   0.96667 1.0917 0.065363
3 0.010000      2   0.95000 1.0667 0.065441
```

The use of minbucket products an error which is much greater than the previous two test and is also greater than the normal error acceptance range that is 0.95.

Thus as low errors are better,we reject both minsplit and minbucket and rather use fare directly which were not split at the best predictive values . Thus using test 1 gives best result.

Tree 2 -(rejected graph)

tree2=rpart(ThermostatTemperature~OddEvenSection+RightLeftHanded+ChocolateOrVanilla,Expanded,cp=0.005) noice how information about oddEven section is not displayed



Test 1

```
> printcp(tree2)

Classification tree:
rpart(formula = ThermostatTemperature ~ OddEvenSection + RightLeftHanded +
    ChocolateOrVanilla, data = Expanded, cp = 0.005)

Variables actually used in tree construction:
[1] ChocolateOrVanilla RightLeftHanded

Root node error: 103/247 = 0.417

n= 247

        CP nsplit rel error xerror     xstd
1 0.0097087      0   1.00000 1.0000 0.075234
2 0.0050000      3   0.97087 1.0097 0.075335
```

When we use this to test our error rates of the tree model , we find that without any fare buckets and and splits we have an error of 0.97 which is greater then the normal error range of below 0.95.

Thus we can not accept this and we need to use either fare bucket or split to find the data which has the minimum error.

Use of minsplit in tree 2,test 2

```
> printcp(rpart(ThermostatTemperature~OddEvenSection+RightLeftHanded+ChocolateOrVanilla,data=Expanded,cp=0.005, minsplit=15))

Classification tree:
rpart(formula = ThermostatTemperature ~ OddEvenSection + RightLeftHanded +
    ChocolateOrVanilla, data = Expanded, cp = 0.005, minsplit = 15)

Variables actually used in tree construction:
[1] ChocolateOrVanilla OddEvenSection     RightLeftHanded

Root node error: 103/247 = 0.417

n= 247

        CP nsplit rel error xerror     xstd
1 0.0194175      0   1.00000 1.0000 0.075234
2 0.0097087      1   0.98058 1.0388 0.075608
3 0.0050000      4   0.95146 1.0388 0.075608
```

When we use the fare split, we get an error which is less than the previous one and

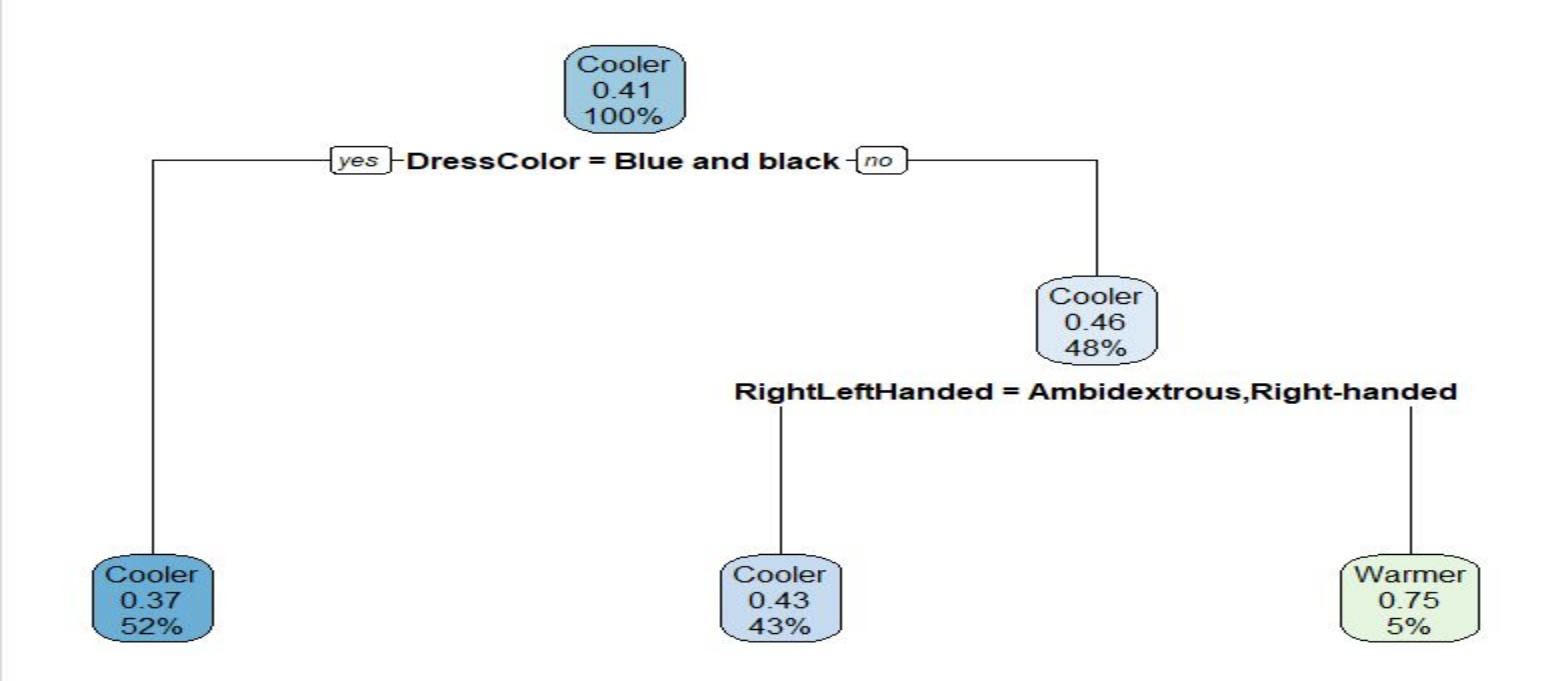
Use of minbucket in tree 2, test 3

```
> printcp(rpart(ThermostatTemperature~OddEvenSection+RightLeftHanded+ChocolateOrVanilla,data=Expanded,cp=0.005,minbucket=15))

Classification tree:
rpart(formula = ThermostatTemperature ~ OddEvenSection + RightLeftHanded +
    ChocolateOrVanilla, data = Expanded, cp = 0.005, minbucket = 15)

Variables actually used in tree construction:
[1] RightLeftHanded

Root node error: 103/247 = 0.417

n= 247

        CP nsplit rel error xerror     xstd
1 0.0097087      0   1.00000 1.0000 0.075234
2 0.0050000      1   0.99029 1.0194 0.075432
```

When we used minbucket, we get a higher that is greater than the previous and also above the range.
Therefore, we can say that mon of the result was appropriate and we need to test with new independent variables

Restes with different variables

tree2=rpart(ThermostatTemperature~RightLeftHanded+DressColor,Expanded,cp=0.005)

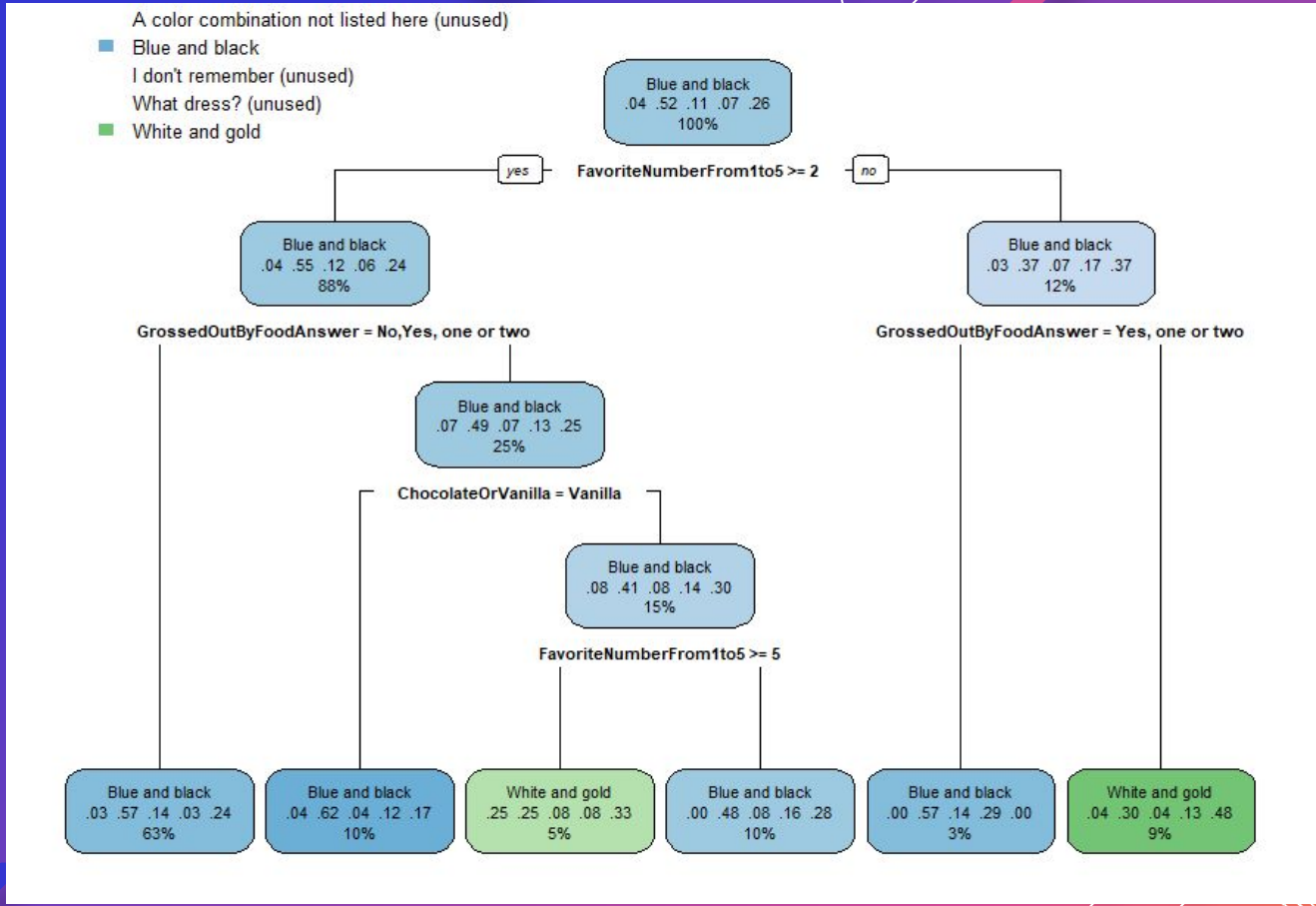


Test with minsplit and minbucket

We tested min split, minbucket and each further had the same value of 0.94 of relative error which is in the range of acceptable errors

```
Classification tree:
rpart(formula = ThermostatTemperature ~ RightLeftHanded + DressColor,
    data = Expanded, cp = 0.005, minsplit = 30)

Variables actually used in tree construction:
[1] DressColor      RightLeftHanded

Root node error: 102/246 = 0.41463

n=246 (1 observation deleted due to missingness)

        CP nsplit rel error xerror     xstd
1 0.029412      0   1.00000 1.0000 0.075755
2 0.005000      2   0.94118 1.0882 0.076517
```

TREE3


A color combination not listed here (unused)
Blue and black
I don't remember (unused)
What dress? (unused)
White and gold
Blue and black
.04 .52 .11 .07 .26
100%
yes
FavoriteNumberFromOnetoFive >= 2
no
Blue and black
.04 .55 .12 .06 .24
88%
Blue and black
.03 .37 .07 .17 .37
12%
GrossedOutByFoodAnswer = No,Yes, one or two
GrossedOutByFoodAnswer = Yes, one or two
Blue and black
.07 .49 .07 .13 .25
25%
ChocolateOrVanilla = Vanilla
Blue and black
.08 .41 .08 .14 .30
15%
FavoriteNumberFromOnetoFive >= 5
Blue and black
.03 .57 .14 .03 .24
63%
Blue and black
.04 .62 .04 .12 .17
10%
White and gold
.25 .25 .08 .08 .33
5%
Blue and black
.00 .48 .08 .16 .28
10%
Blue and black
.00 .57 .14 .29 .00
3%
White and gold
.04 .30 .04 .13 .48
9%

```
tree3=rpart(DressColor~FavoriteNumberFrom1to5+ChocolateOrVanilla+GrossedOutByFoodAnswer,Expanded,cp=0.00005)
```

Identify the best relative error,Test 1

```
> printcp(tree3)

Classification tree:
rpart(formula = DressColor ~ FavoriteNumberFrom1to5 + ChocolateOrVanilla +
    GrossedOutByFoodAnswer, data = Expanded, cp = 5e-05)

Variables actually used in tree construction:
[1] ChocolateOrVanilla     FavoriteNumberFrom1to5 GrossedOutByFoodAnswer

Root node error: 117/246 = 0.47561

n=246 (1 observation deleted due to missingness)

        CP nsplit rel error xerror     xstd
1 0.017094      0   1.00000 1.0000 0.066948
2 0.002849      2   0.96581 1.0855 0.066992
3 0.000050      5   0.95726 1.1624 0.066652
```

Without using any fare buckest and split we get an relative error of 0.96 which is outside the acceptable range

Use of minsplit in tree 3,test 2

```
> printcp(rpart(DressColor~FavoriteNumberFrom1to5+ChocolateOrVanilla+GrossedOutByFoodAnswer,data=Expanded,cp=0.0005,minsplit=4))

Classification tree:
rpart(formula = DressColor ~ FavoriteNumberFrom1to5 + ChocolateOrVanilla +
    GrossedOutByFoodAnswer, data = Expanded, cp = 5e-04, minsplit = 4)

Variables actually used in tree construction:
[1] ChocolateOrVanilla     FavoriteNumberFrom1to5 GrossedOutByFoodAnswer

Root node error: 117/246 = 0.47561

n=246 (1 observation deleted due to missingness)

         CP nsplit rel error xerror     xstd
1 0.0170940      0   1.00000 1.0000 0.066948
2 0.0064103      2   0.96581 1.0342 0.067018
3 0.0034188      6   0.94017 1.1026 0.066948
4 0.0005000     11   0.92308 1.1538 0.066708
```

Here after using the fare split, we get an error 0.92 which is relatively low and in the range of the acceptable relative error.

Use of minbucket in tree 1, test 3

```
> printcp(rpart(DressColor~FavoriteNumberFrom1to5+ChocolateOrVanilla+GrossedOutByFoodAnswer,data=Expanded,cp=0.0005,minbucket=10))

Classification tree:
rpart(formula = DressColor ~ FavoriteNumberFrom1to5 + ChocolateOrVanilla +
    GrossedOutByFoodAnswer, data = Expanded, cp = 5e-04, minbucket = 10)

Variables actually used in tree construction:
[1] ChocolateOrVanilla     FavoriteNumberFrom1to5 GrossedOutByFoodAnswer

Root node error: 117/246 = 0.47561

n=246 (1 observation deleted due to missingness)

        CP nsplit rel error xerror     xstd
1 0.008547      0   1.00000 1.0000 0.066948
2 0.002849      2   0.98291 1.0769 0.067007
3 0.000500      5   0.97436 1.1026 0.066948
```

After using minbucket we got a relative error of 097 which is much greater than acceptable range and thus we reject this.

We accept the use of minsplit which is gives us significantly low error that is acceptable in the range.

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**