Intro to Machine Learning

Supervised Learning in R

Gemma Dawson

13 September 2018

R Ladies

What is Machine Learning?

What is Machine Learning?

"The field of study that gives computers the ability to learn **without being explicitly programmed**."

What is Machine Learning?

"The field of study that gives computers the ability to learn **without being explicitly programmed**."

What is Machine Learning?

What is Machine Learning?

Supervised Learning

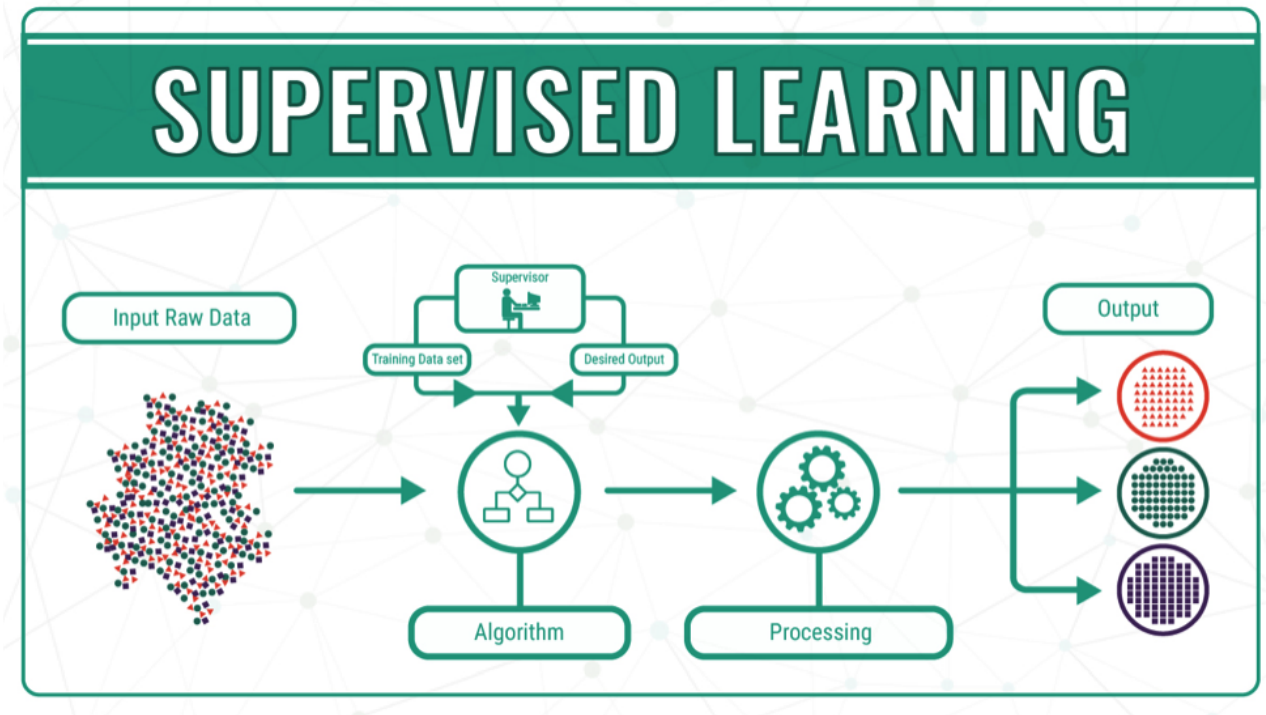
What is Machine Learning?

Supervised Learning

Unsupervised Learning

What is Supervised Learning?

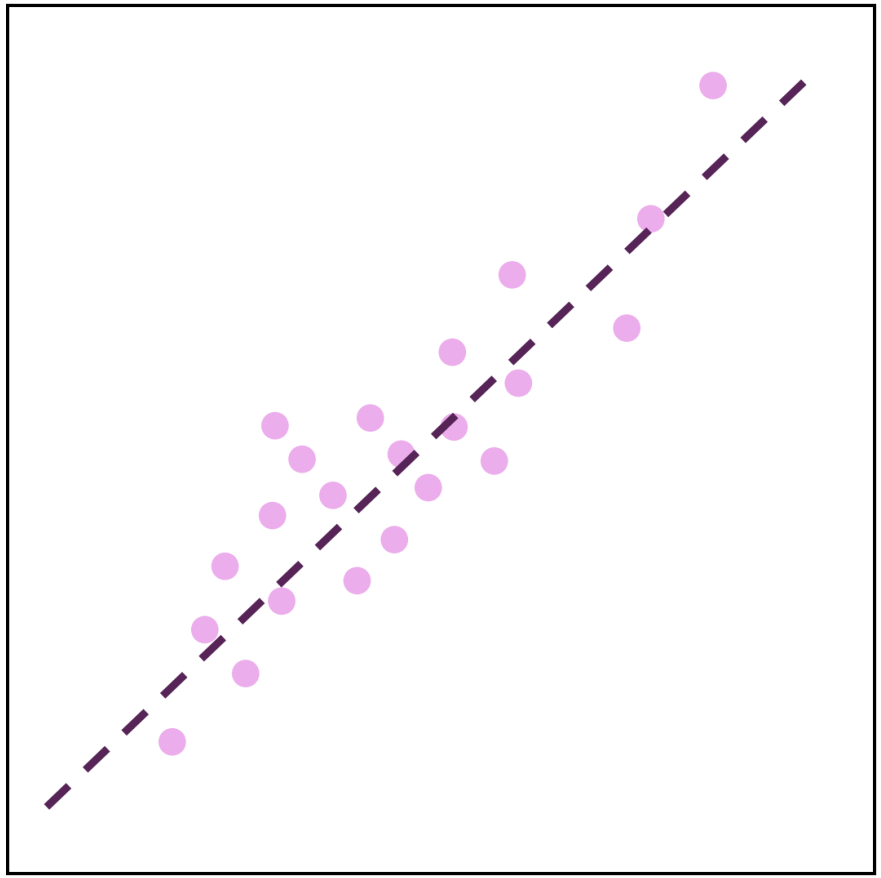
GOAL: create a predictive model



What is Supervised Learning?

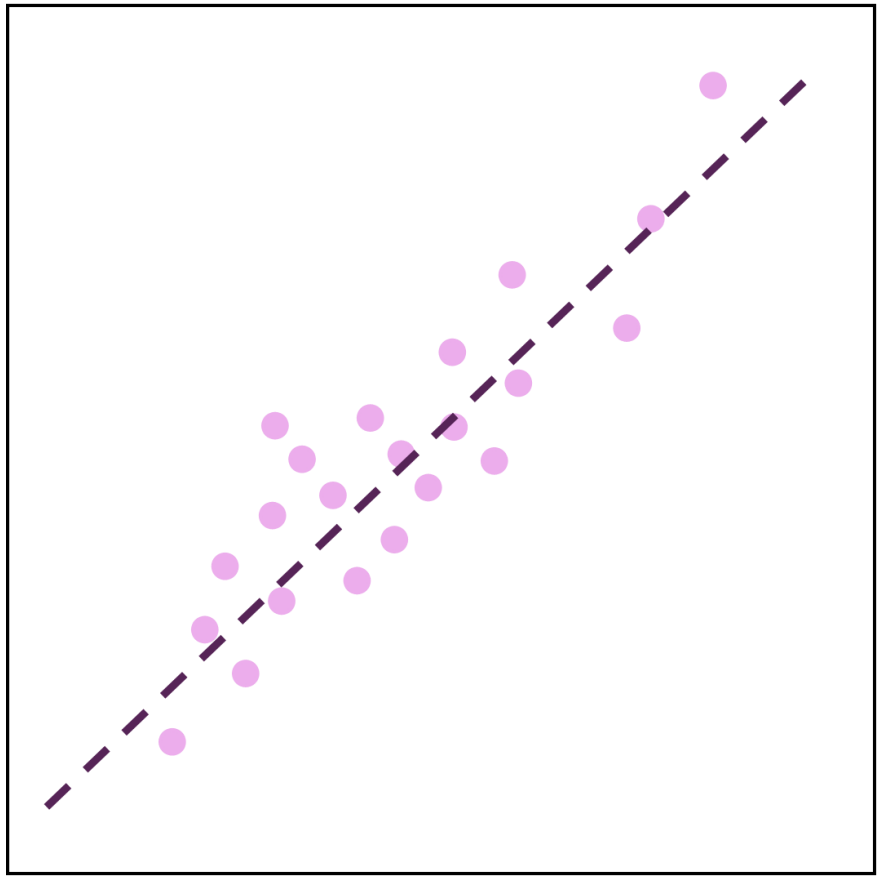
What is Supervised Learning?

Regression

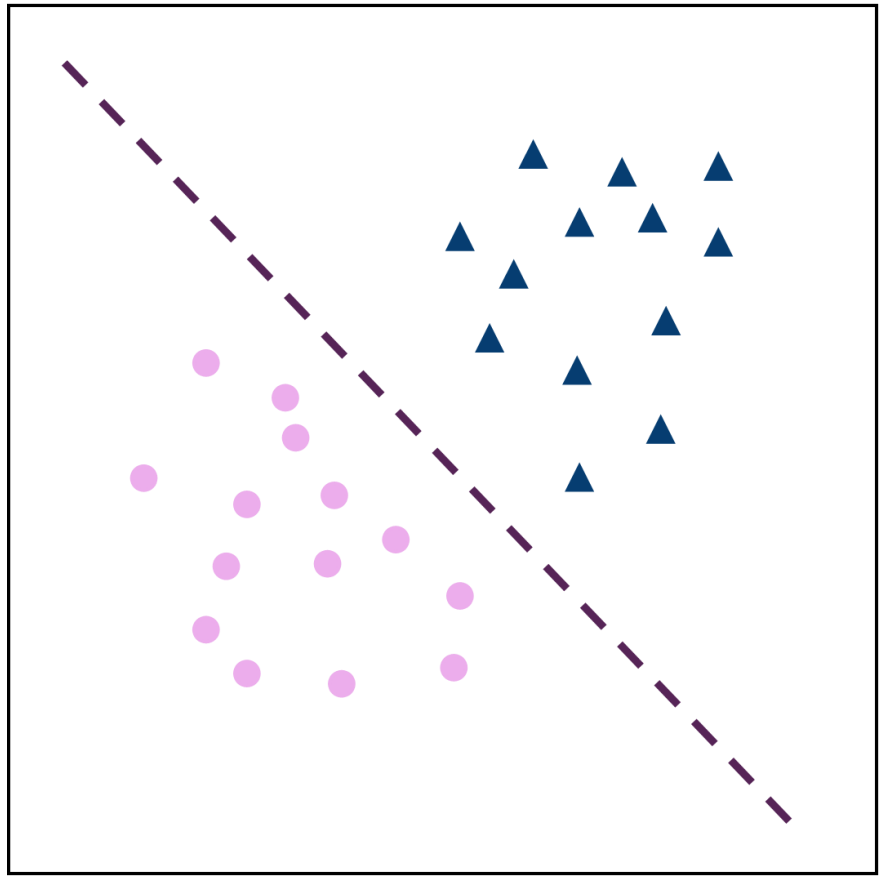


What is Supervised Learning?

Regression

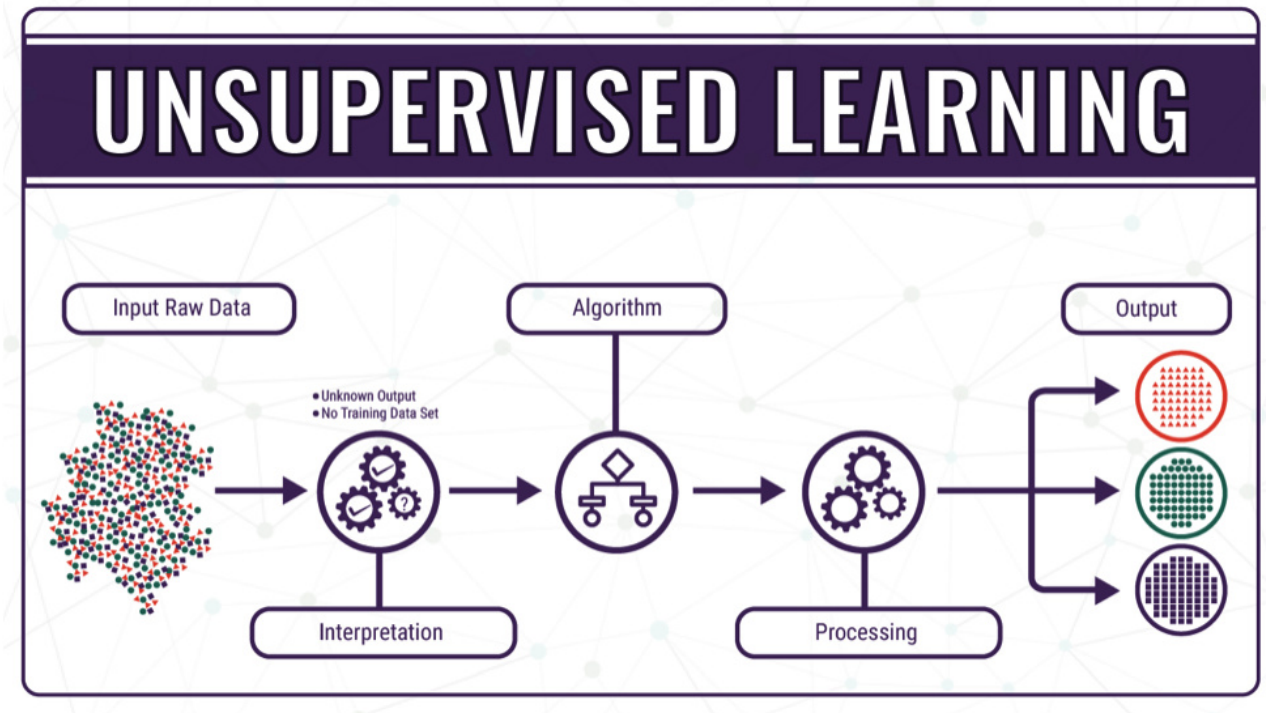


Classification



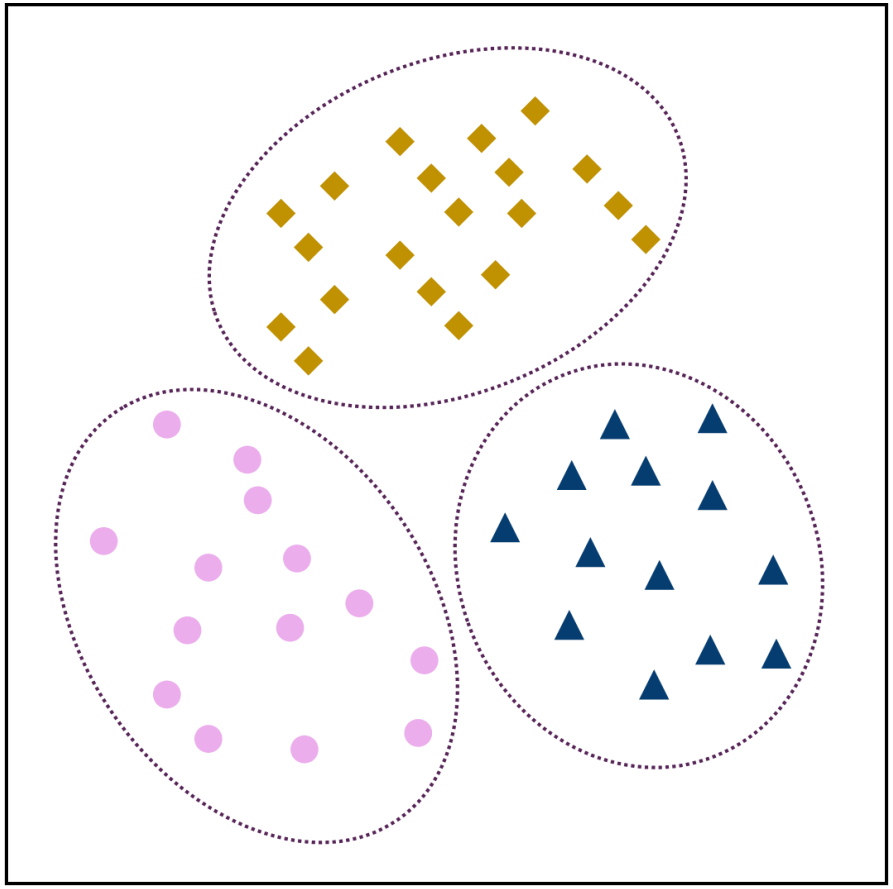
What is Unsupervised Learning?

GOAL: create a descriptive model

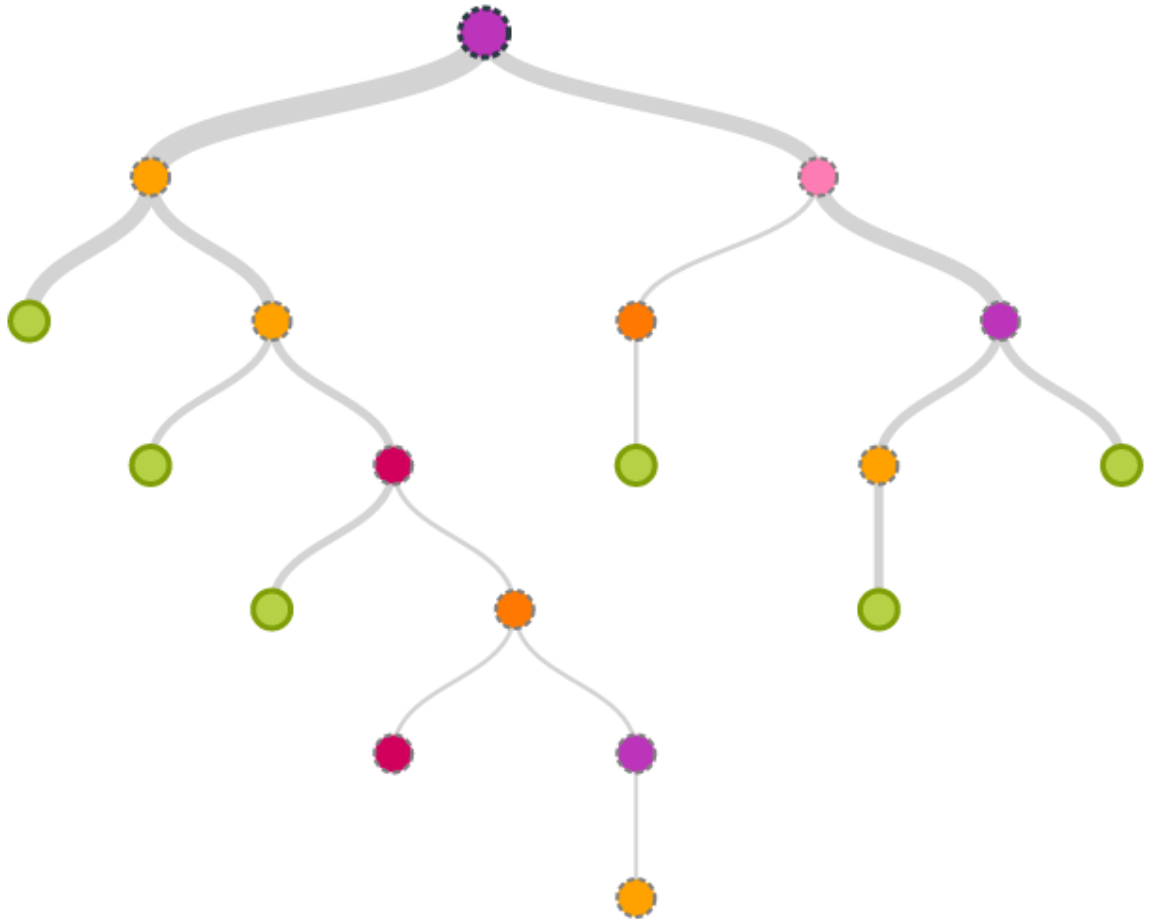


What is Unsupervised Learning?

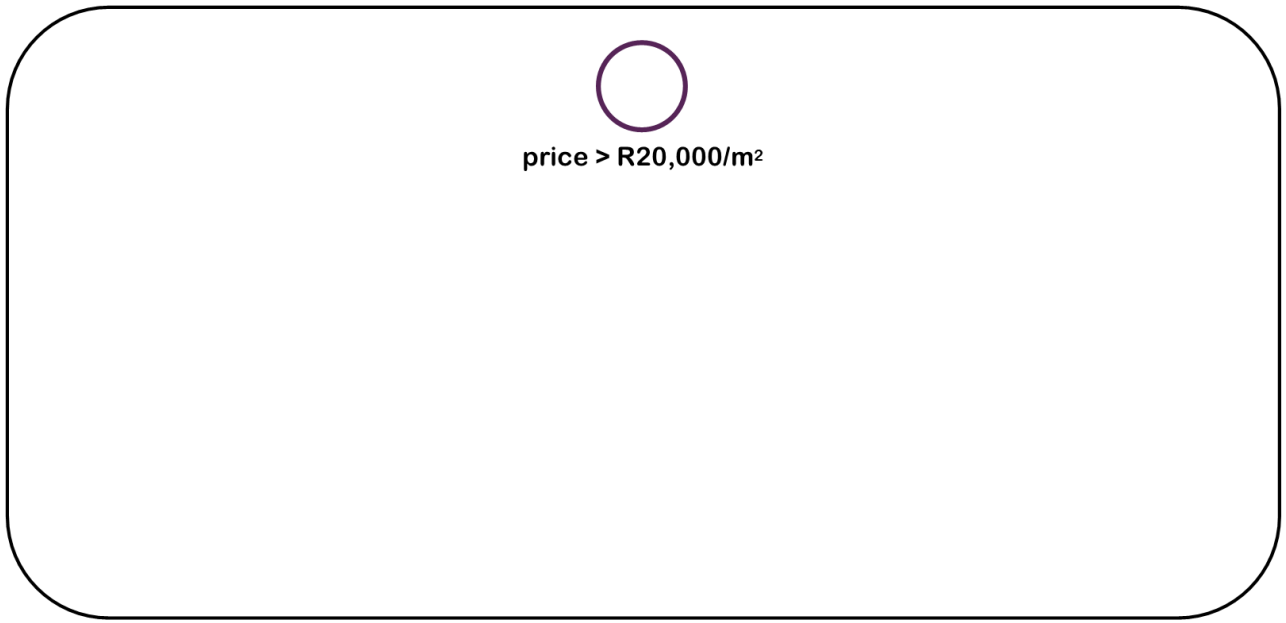
Clustering



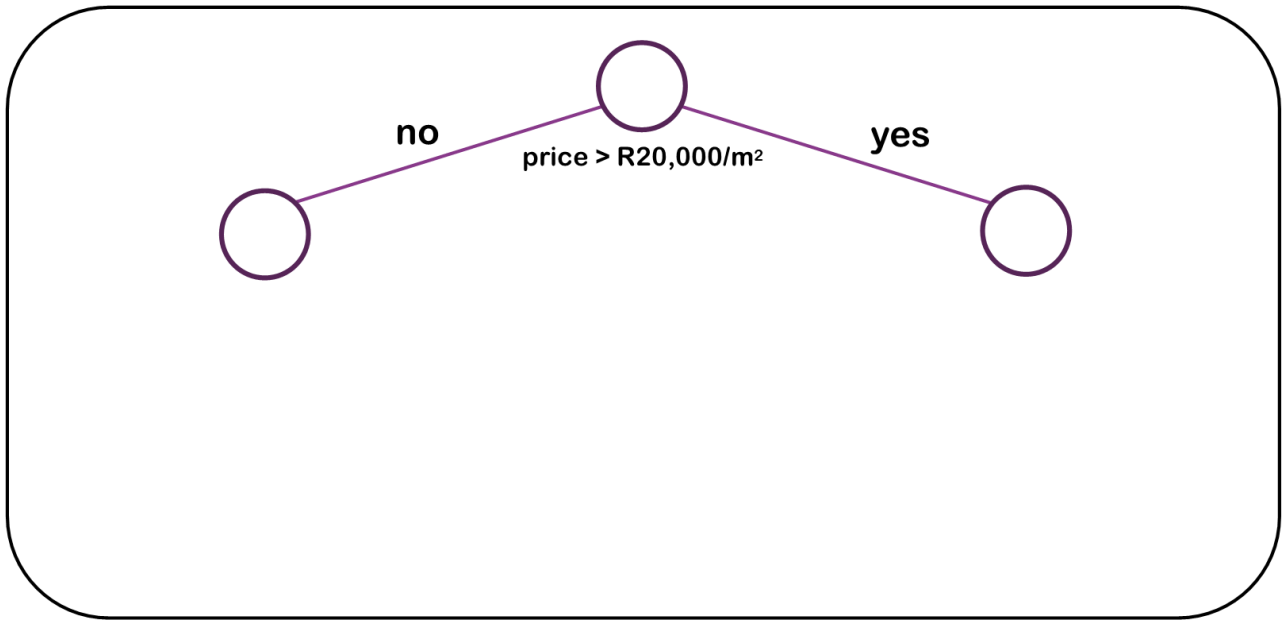
Supervised Learning: Decision Trees



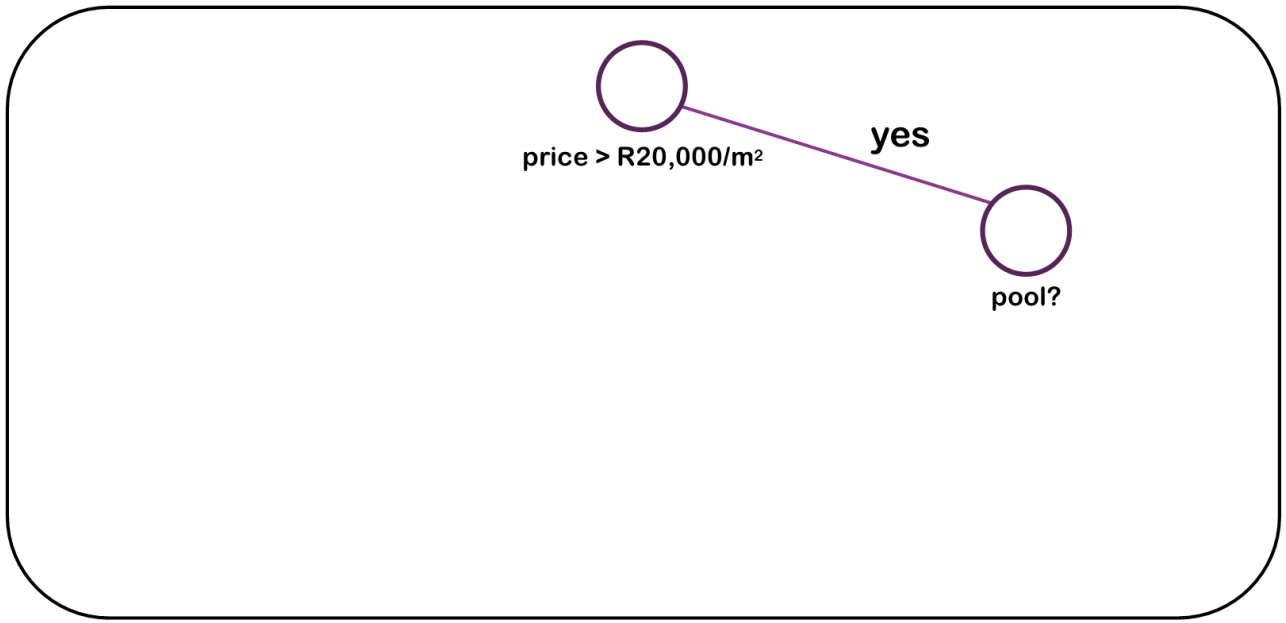
Supervised Learning: Decision Trees



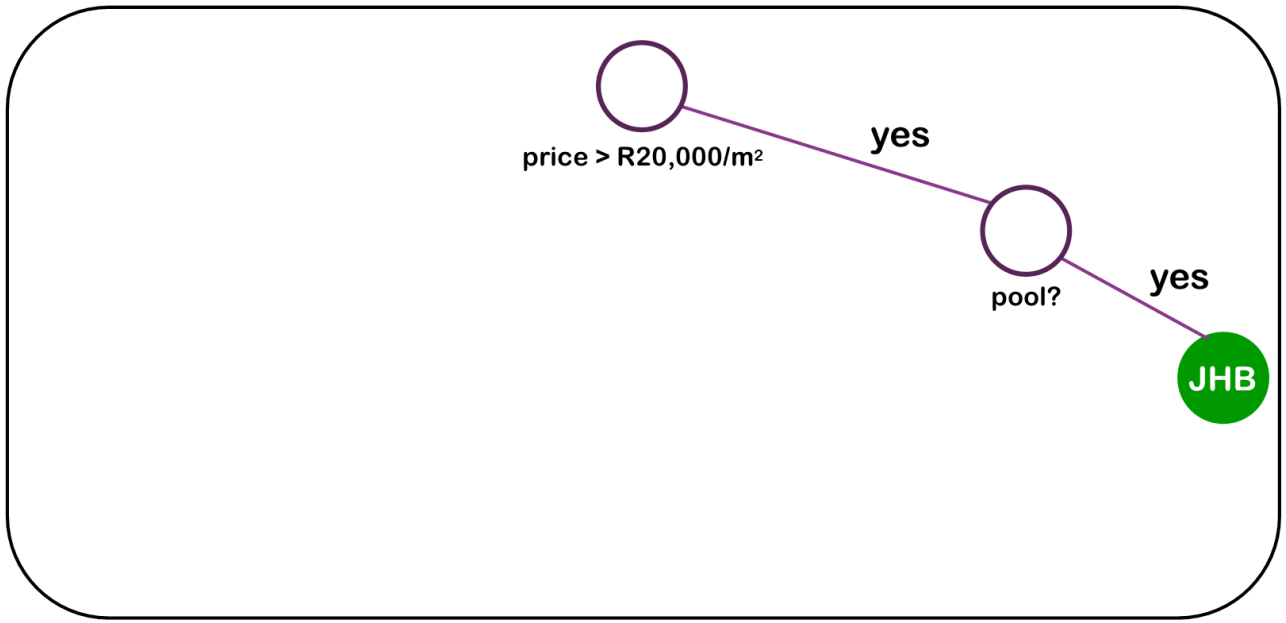
Supervised Learning: Decision Trees



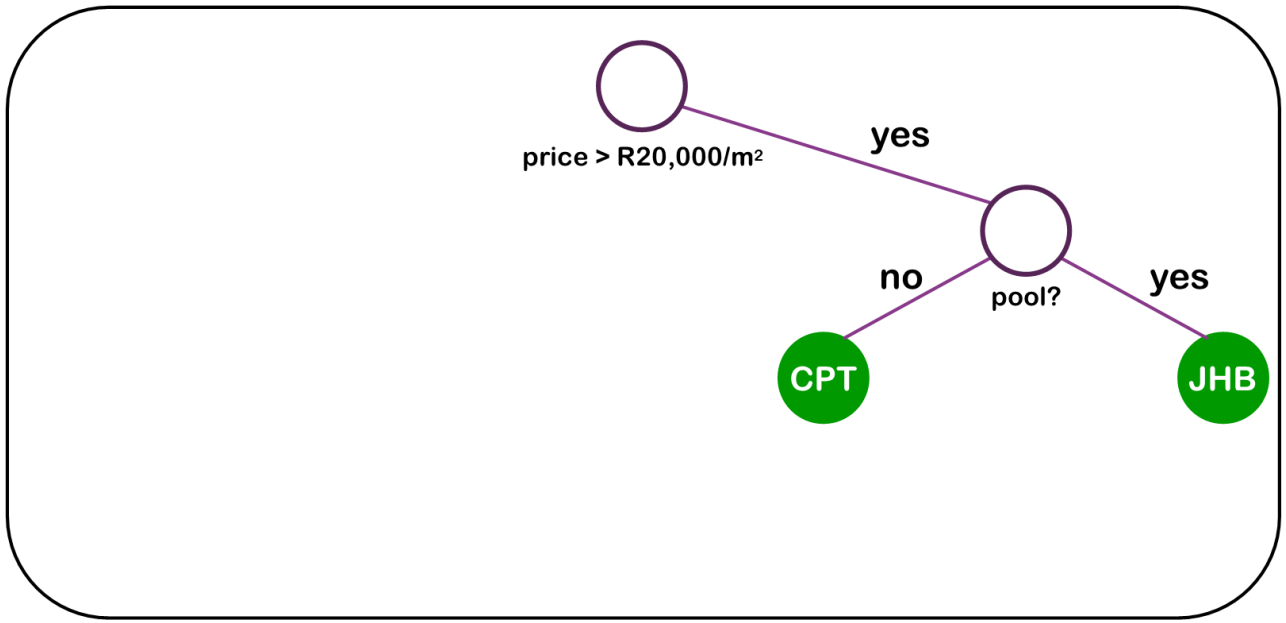
Supervised Learning: Decision Trees



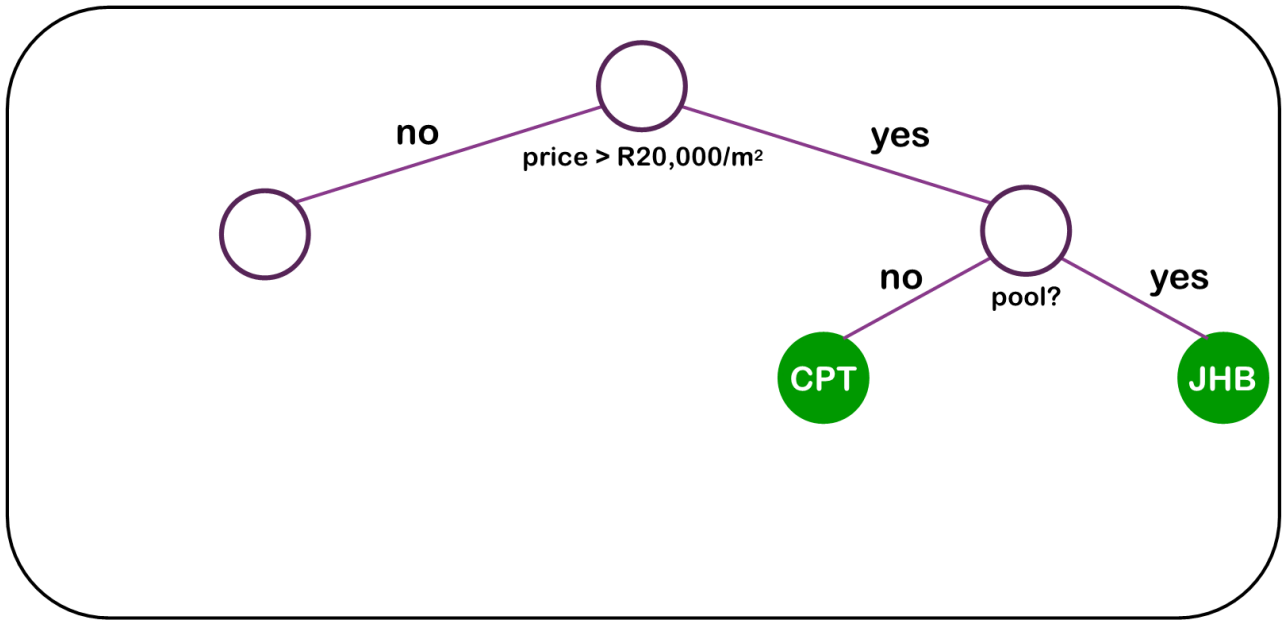
Supervised Learning: Decision Trees



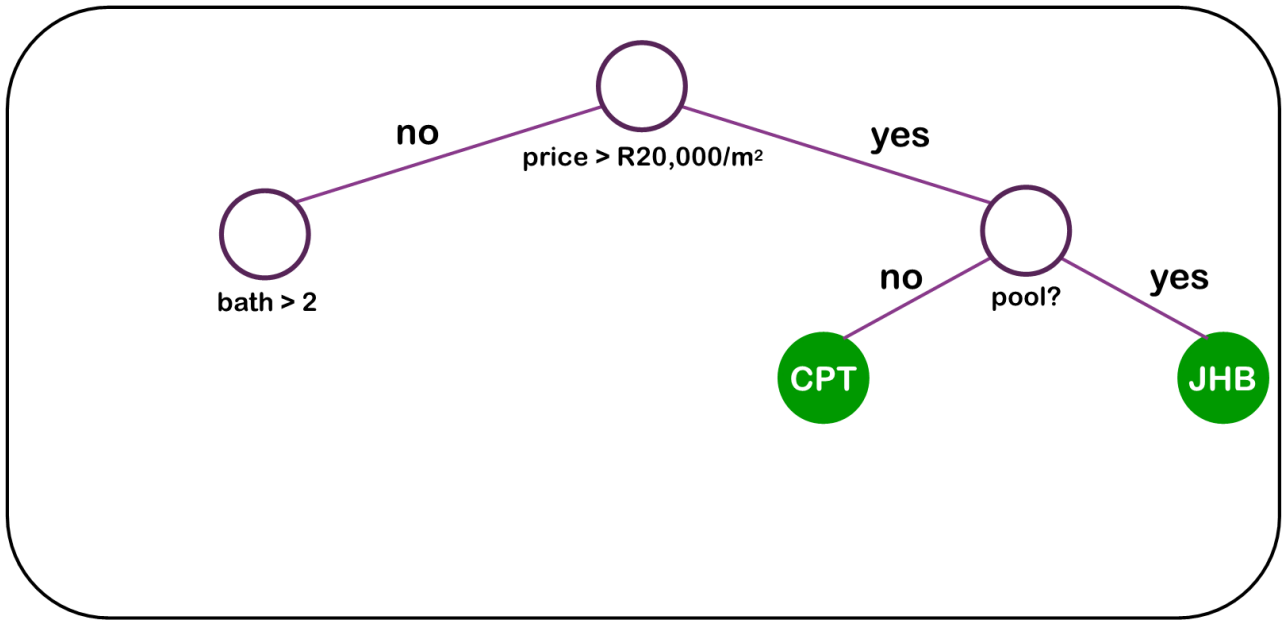
Supervised Learning: Decision Trees



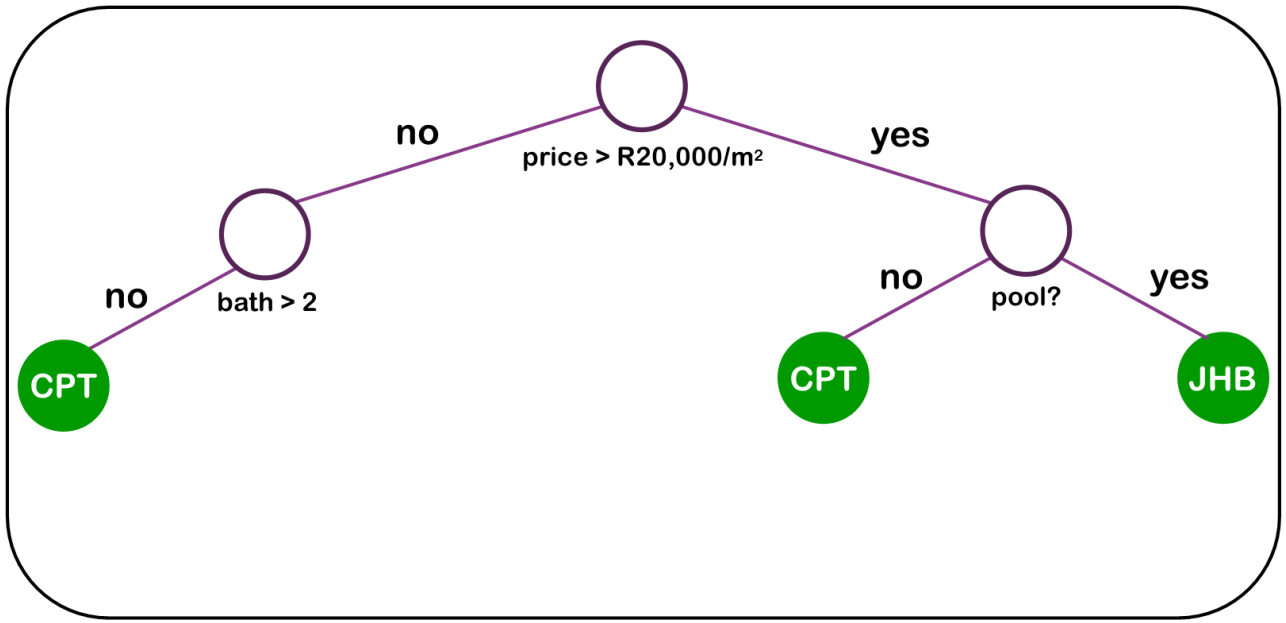
Supervised Learning: Decision Trees



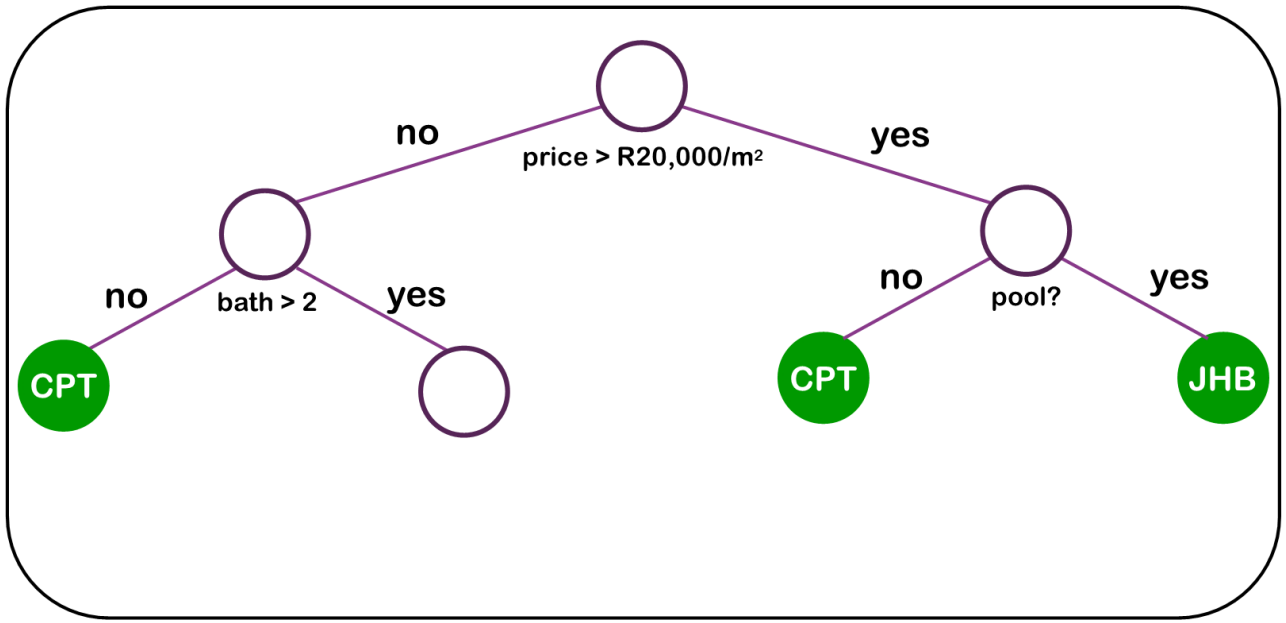
Supervised Learning: Decision Trees



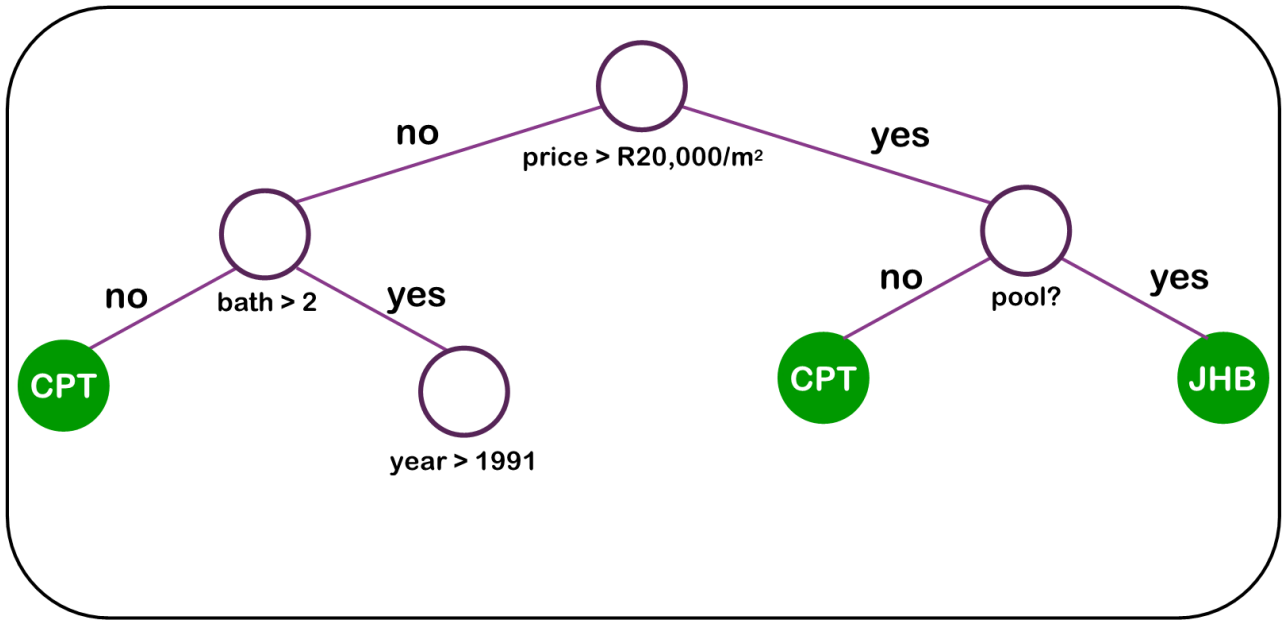
Supervised Learning: Decision Trees



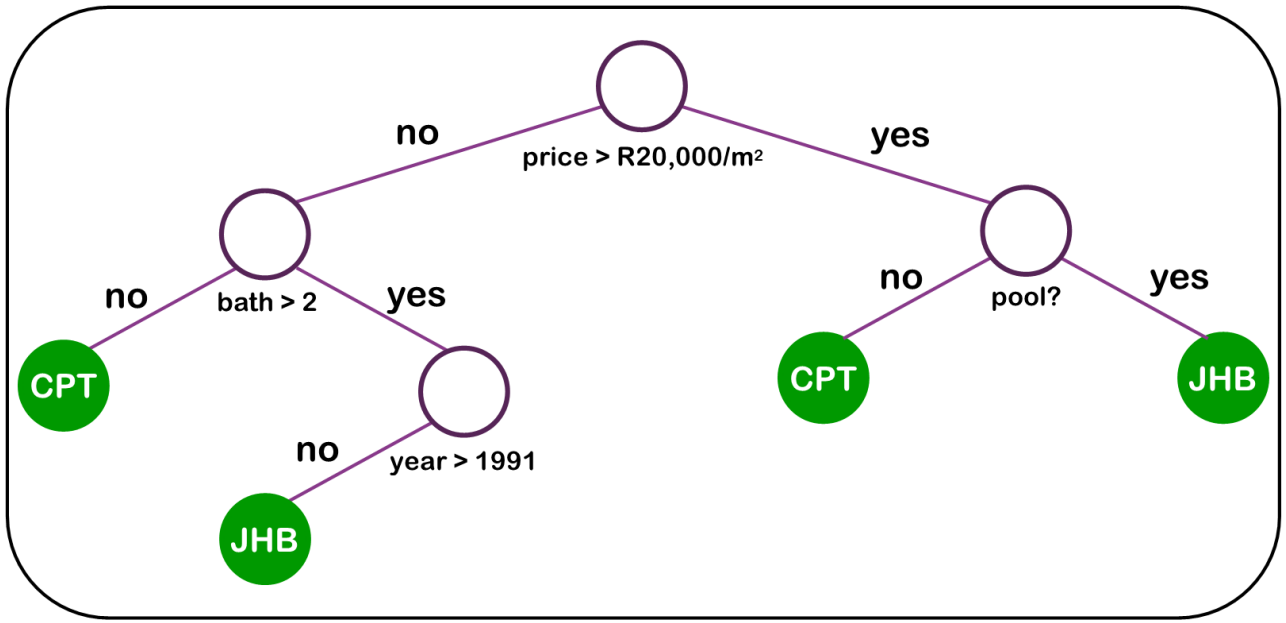
Supervised Learning: Decision Trees



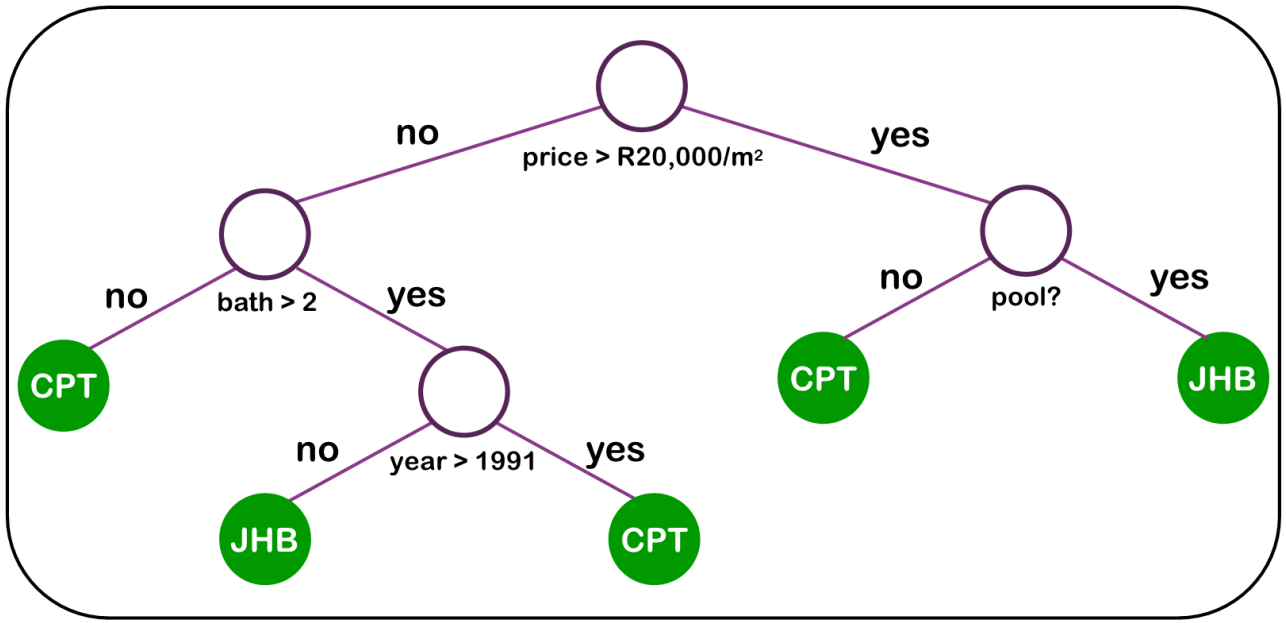
Supervised Learning: Decision Trees



Supervised Learning: Decision Trees



Supervised Learning: Decision Trees



R and RStudio

RLadies & R Users Groups

Titanic

Machine Learning from Disaster

Machine Learning from Disaster

Machine Learning from Disaster

kaggle™

training.csv

891 entries
11 variables

test.csv

418 entries
10 variables

Machine Learning from Disaster

kaggle™

GOAL: predict tof the Titanic's passengers survived

training.csv

891 entries
11 variables

test.csv

418 entries
10 variables

Load libraries

Load libraries

```
library(tidyverse)
library(rpart)
library(rpart.plot)
```

Load libraries

```
library(tidyverse)
library(rpart)
library(rpart.plot)
```

Import the data

Load libraries

```
library(tidyverse)
library(rpart)
library(rpart.plot)
```

Import the data

```
titanic <- read_csv(file = "data/titanic.csv", col_names = TRUE)
```

Let's look at the data

```
head(titanic)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare
##         <int>    <int>  <int> <chr>   <chr> <dbl> <int> <int> <chr>  <dbl>
## 1           1        0      3 Braund… male     22     1     0 A/5 2…  7.25
## 2           2        1      1 Cuming… fema…    38     1     0 PC 17… 71.3
## 3           3        1      3 Heikki… fema…    26     0     0 STON/…  7.92
## 4           4        1      1 Futrel… fema…    35     1     0 113803 53.1
## 5           5        0      3 Allen,… male     35     0     0 373450  8.05
## 6           6        0      3 Moran,… male     NA     0     0 330877  8.46
## # ... with 2 more variables: Cabin <chr>, Embarked <chr>
```

Tidy & Transform

```
summary(titanic$Pclass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.309   3.000   3.000
```

```
titanic$Pclass <- as.factor(titanic$Pclass)

summary(titanic$Pclass)
```

```
##   1   2   3
## 216 184 491
```

Tidy & Transform

```
summary(titanic$Survived)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3838  1.0000  1.0000
```

```
titanic$Survived <- if_else(titanic$Survived == 1,
                            "yes",
                            "no")

summary(titanic$Survived)
```

```
##    Length     Class      Mode
##       891 character character
```

Tidy & Transform

```
summary(titanic$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

```
titanic$Age <- if_else(is.na(titanic$Age),
                       mean(titanic$Age, na.rm = T),
                       titanic$Age)

summary(titanic$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   22.00   29.70   29.70   35.00   80.00
```

Tidy & Transform

```
titanic$Family <- titanic$SibSp + titanic$Parch

summary(titanic$Family)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.9046  1.0000 10.0000
```

Tidy & Transform

```
titanic <- titanic %>%
              select(Pclass, Sex, Age, Fare, Family, Survived)

glimpse(titanic)
```

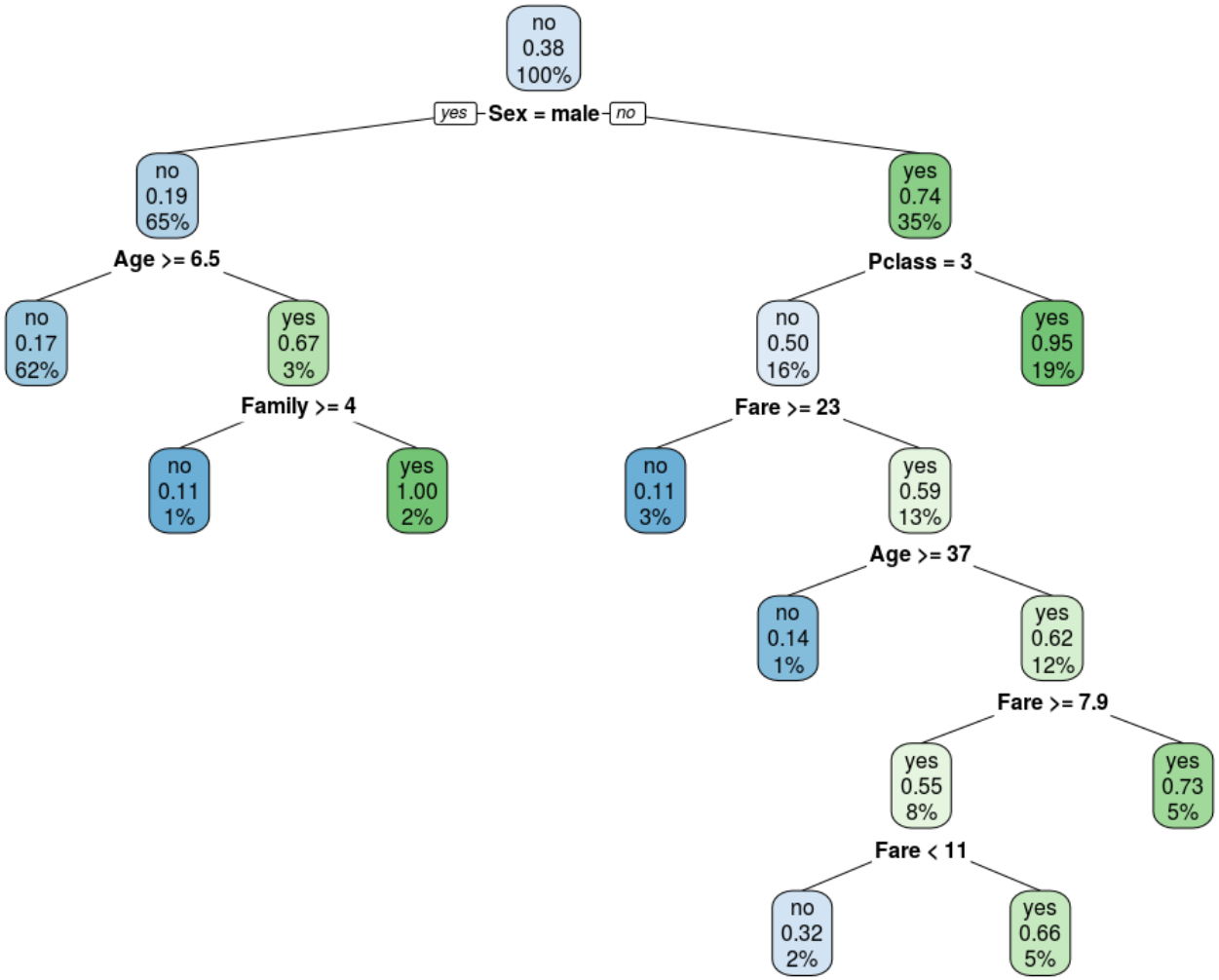
```
## Observations: 891
## Variables: 6
## $ Pclass   <fct> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2,...
## $ Sex      <chr> "male", "female", "female", "female", "male", "male",...
## $ Age      <dbl> 22.00000, 38.00000, 26.00000, 35.00000, 35.00000, 29....
## $ Fare     <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51....
## $ Family   <int> 1, 1, 0, 1, 0, 0, 0, 4, 2, 1, 2, 0, 0, 6, 0, 0, 5, 0,...
## $ Survived <chr> "no", "yes", "yes", "yes", "no", "no", "no", "no", "y...
```

Model

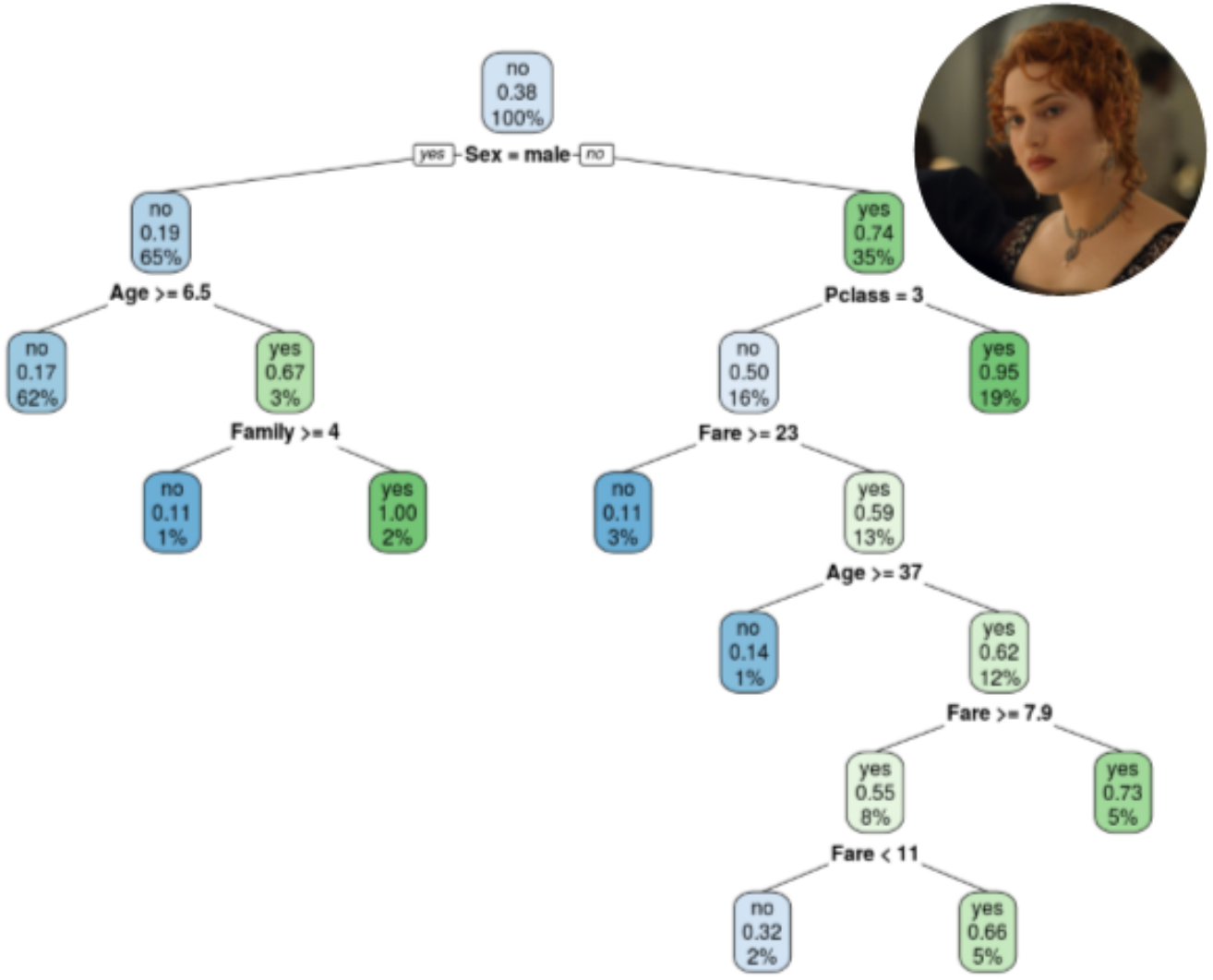
```
tree.titanic <- rpart(formula = Survived ~ ., data = titanic)

rpart.plot(tree.titanic, fallen.leaves = F)
```

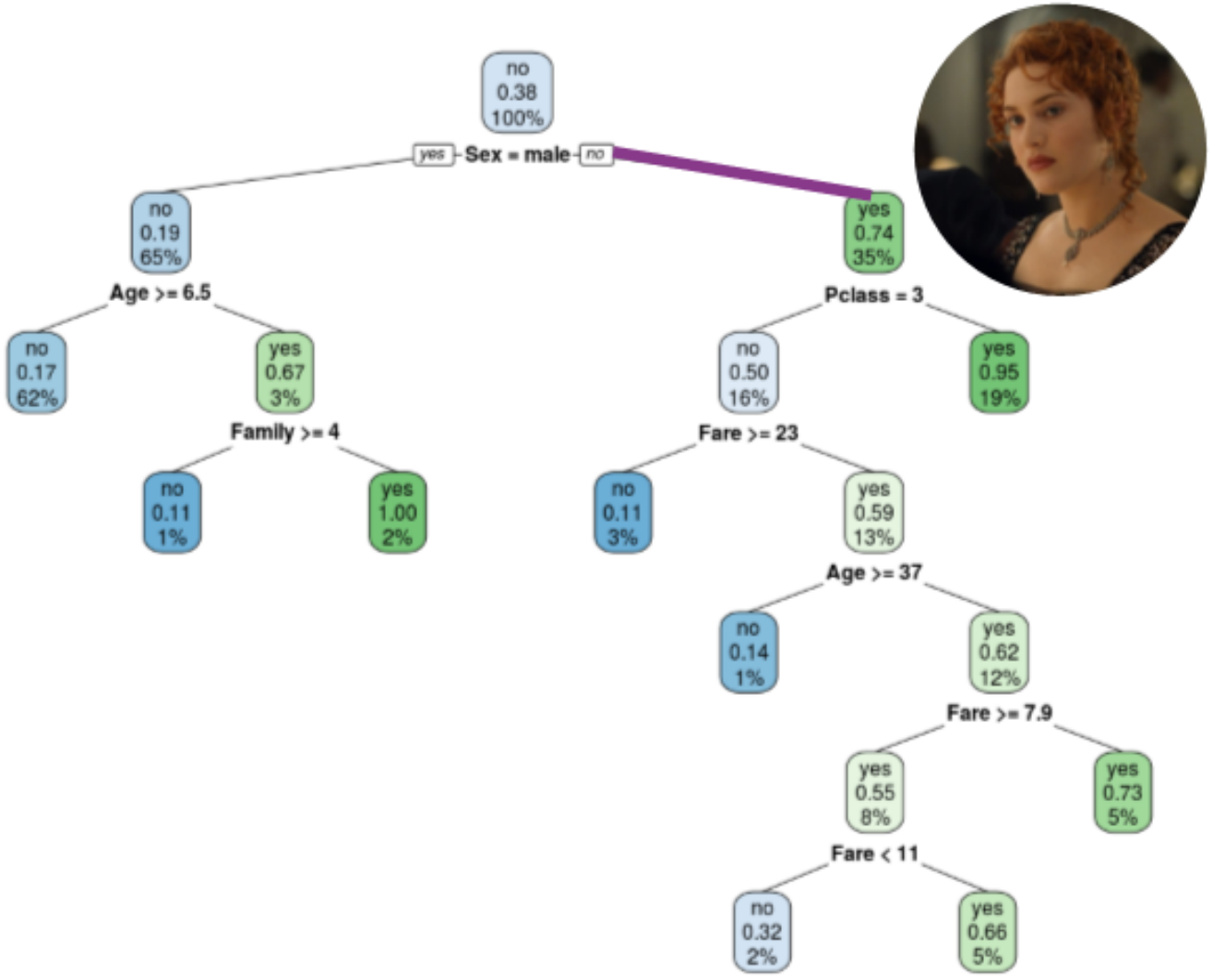
Visualise & Communicate



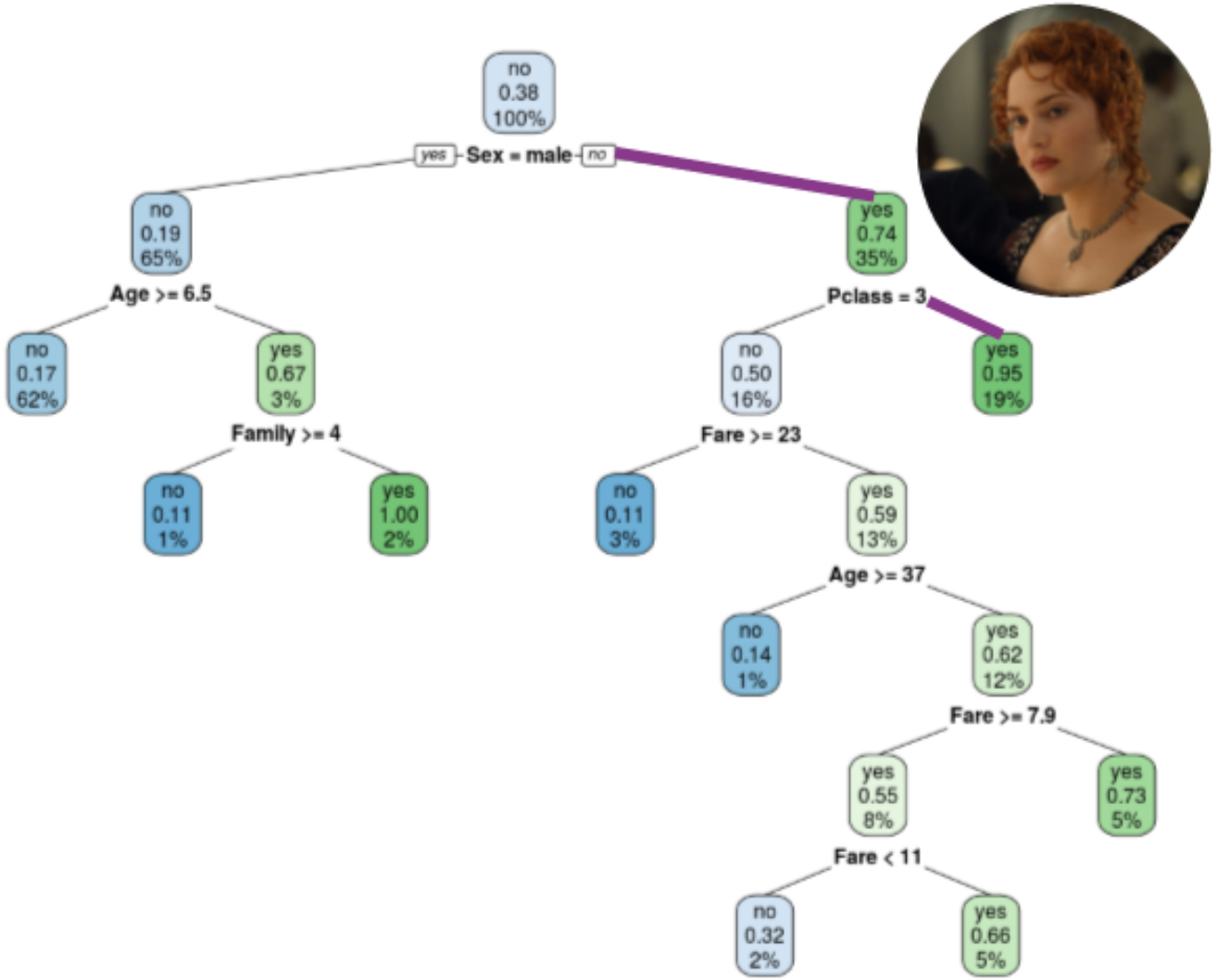
Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

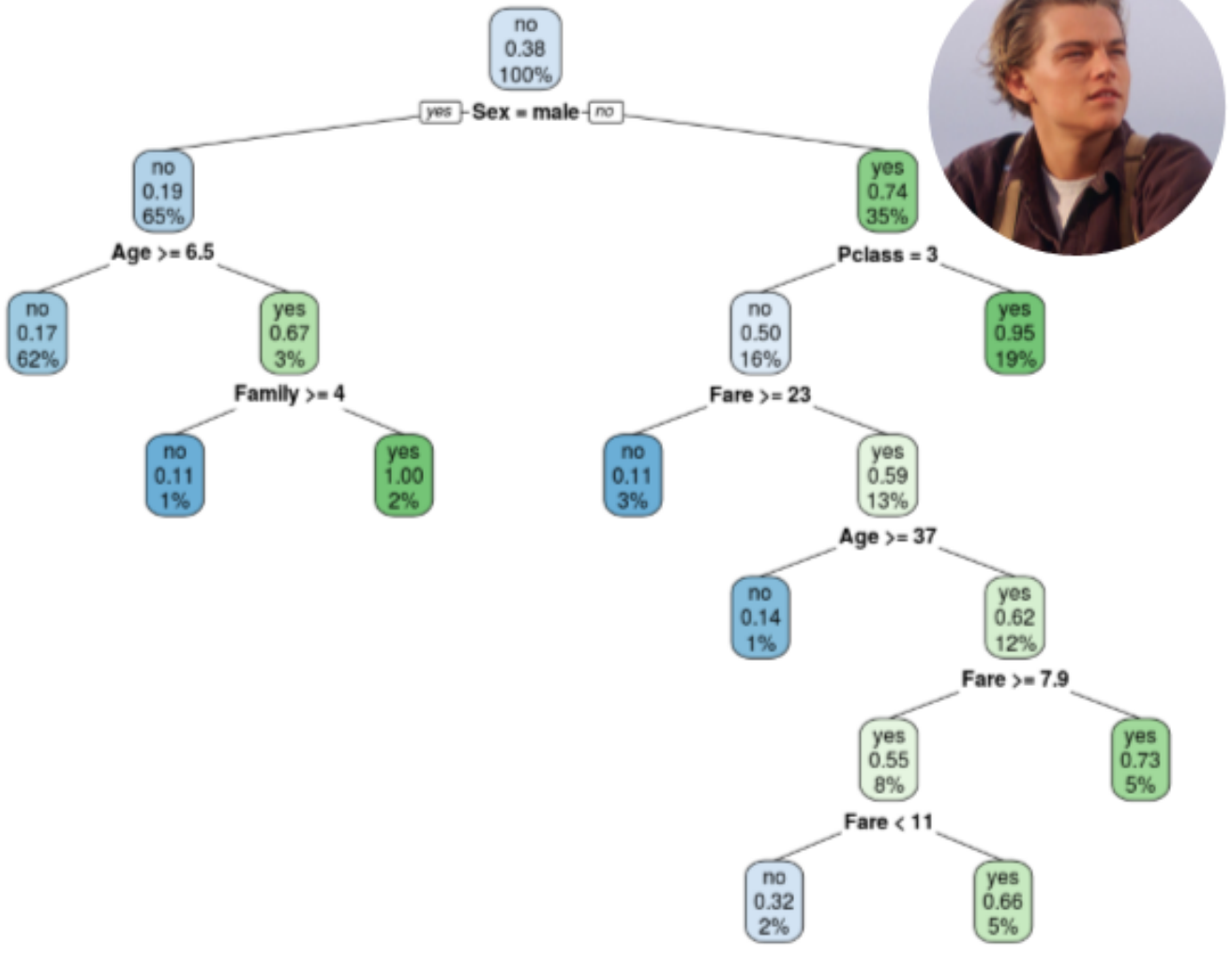
Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

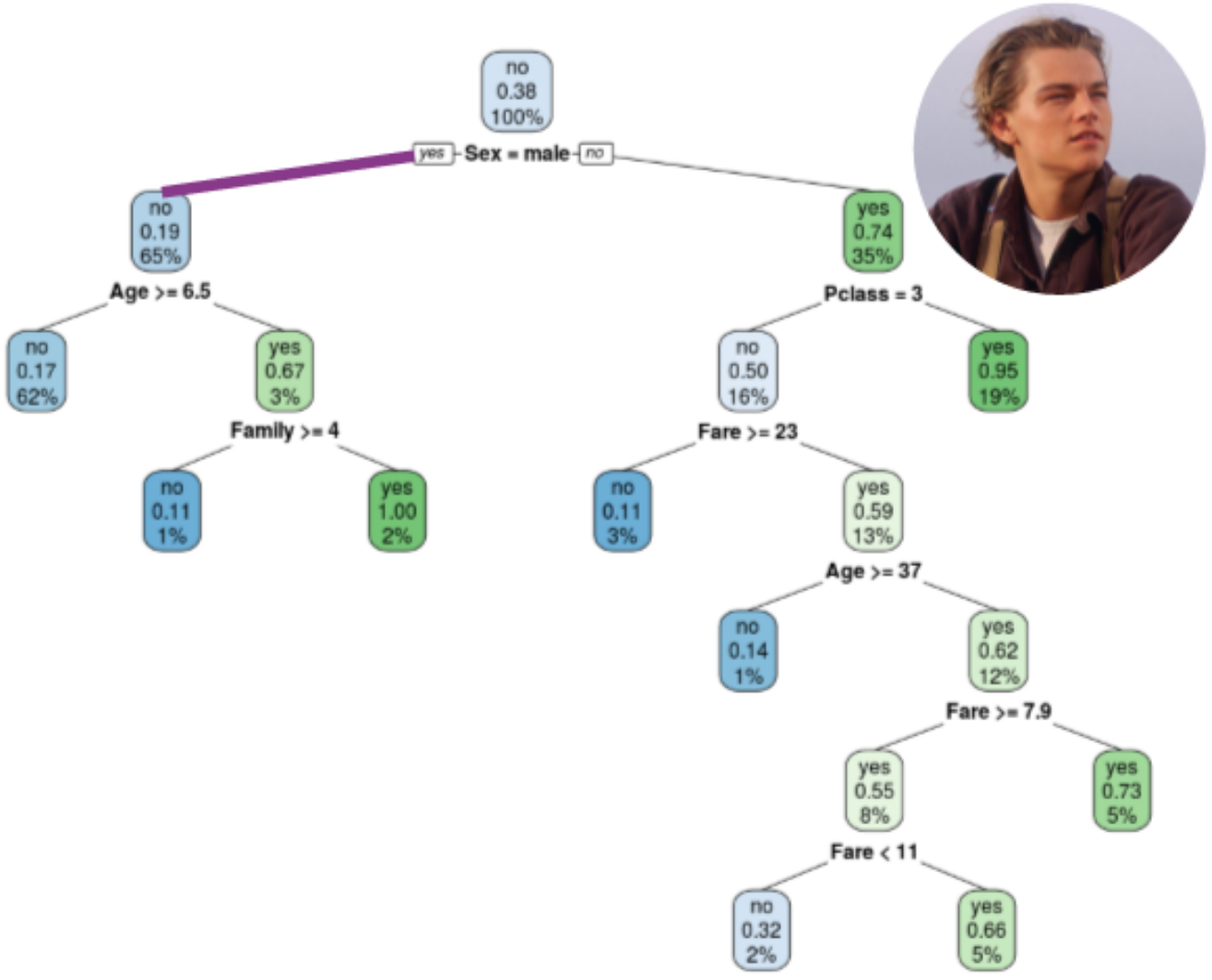
Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

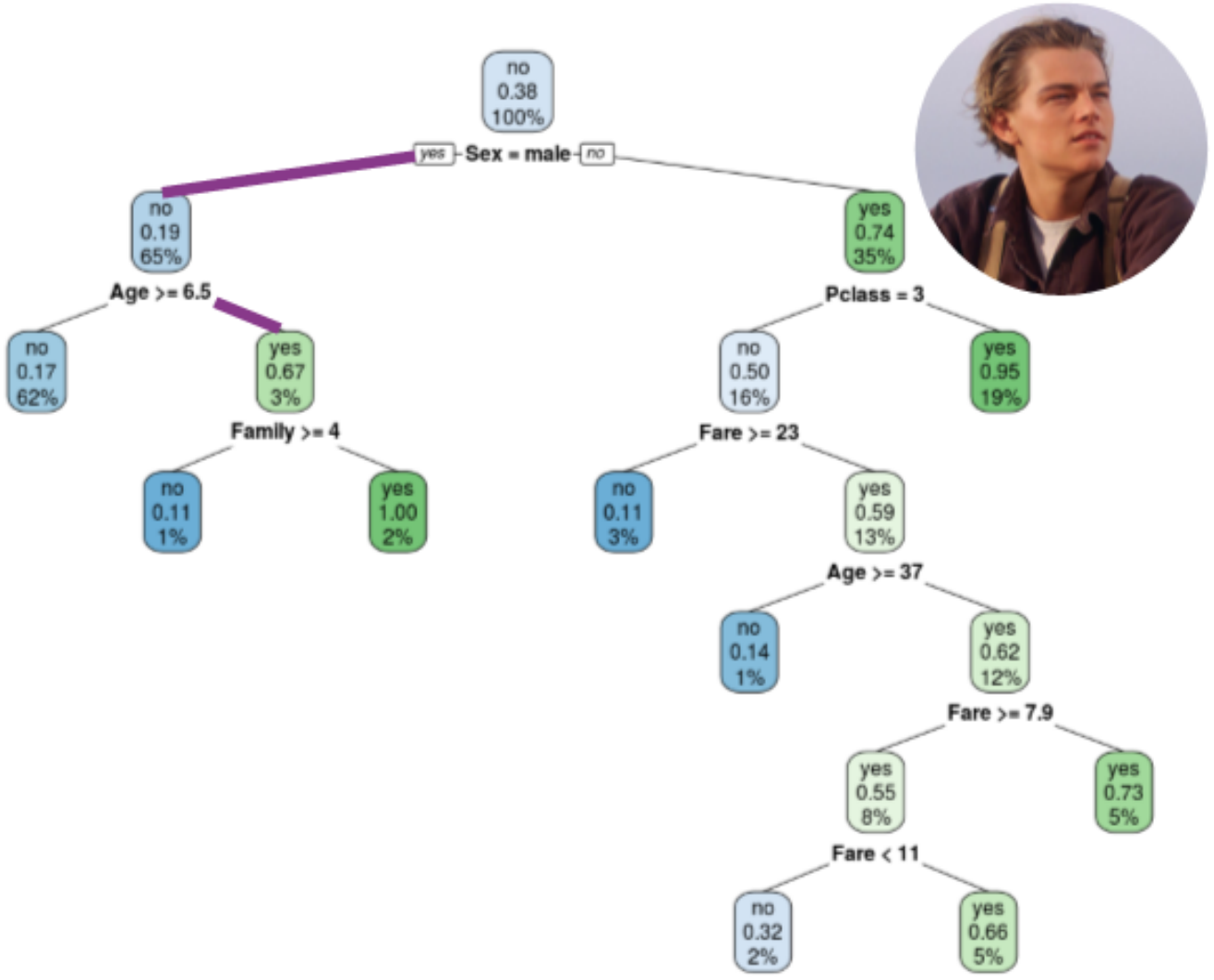
Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

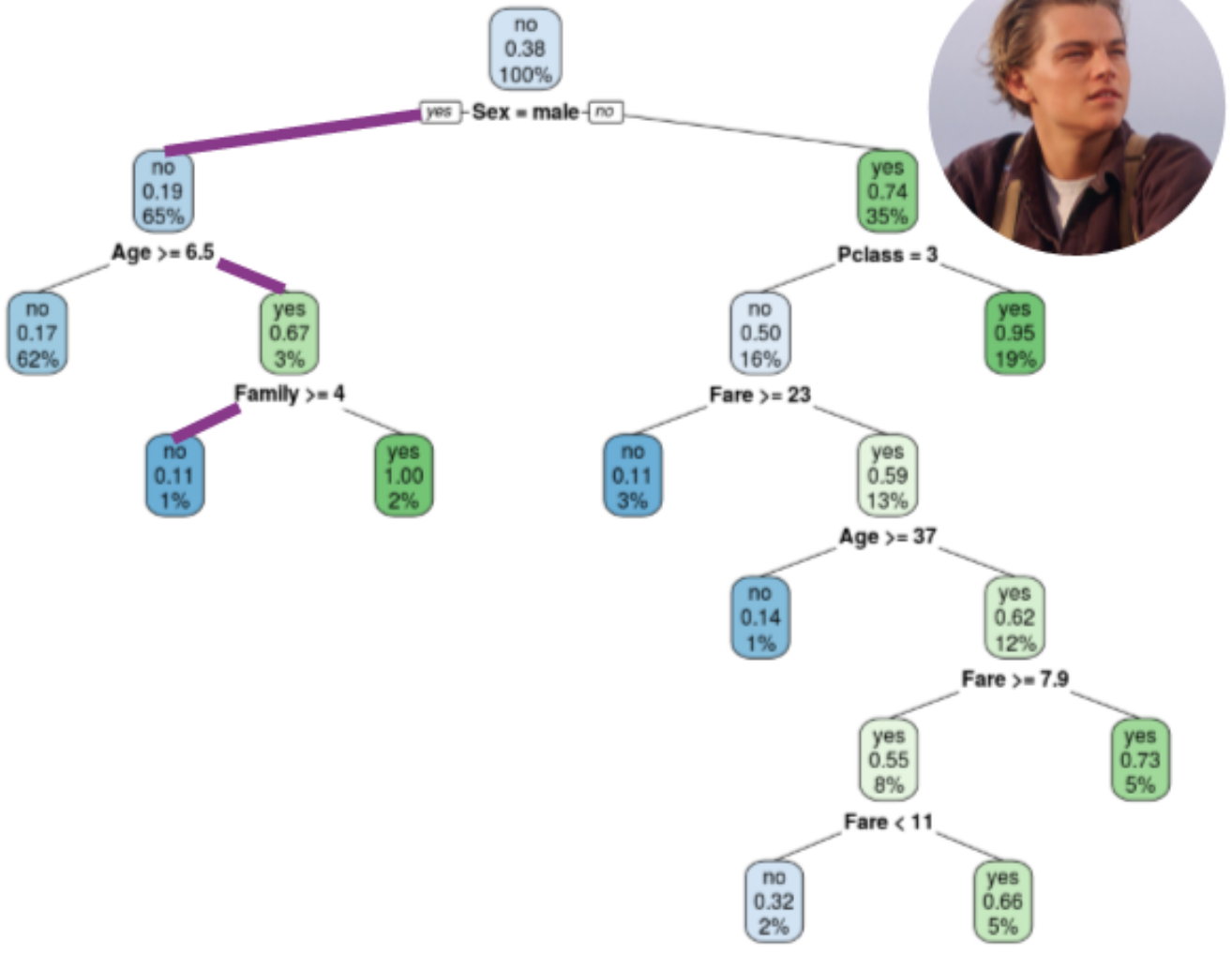
Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

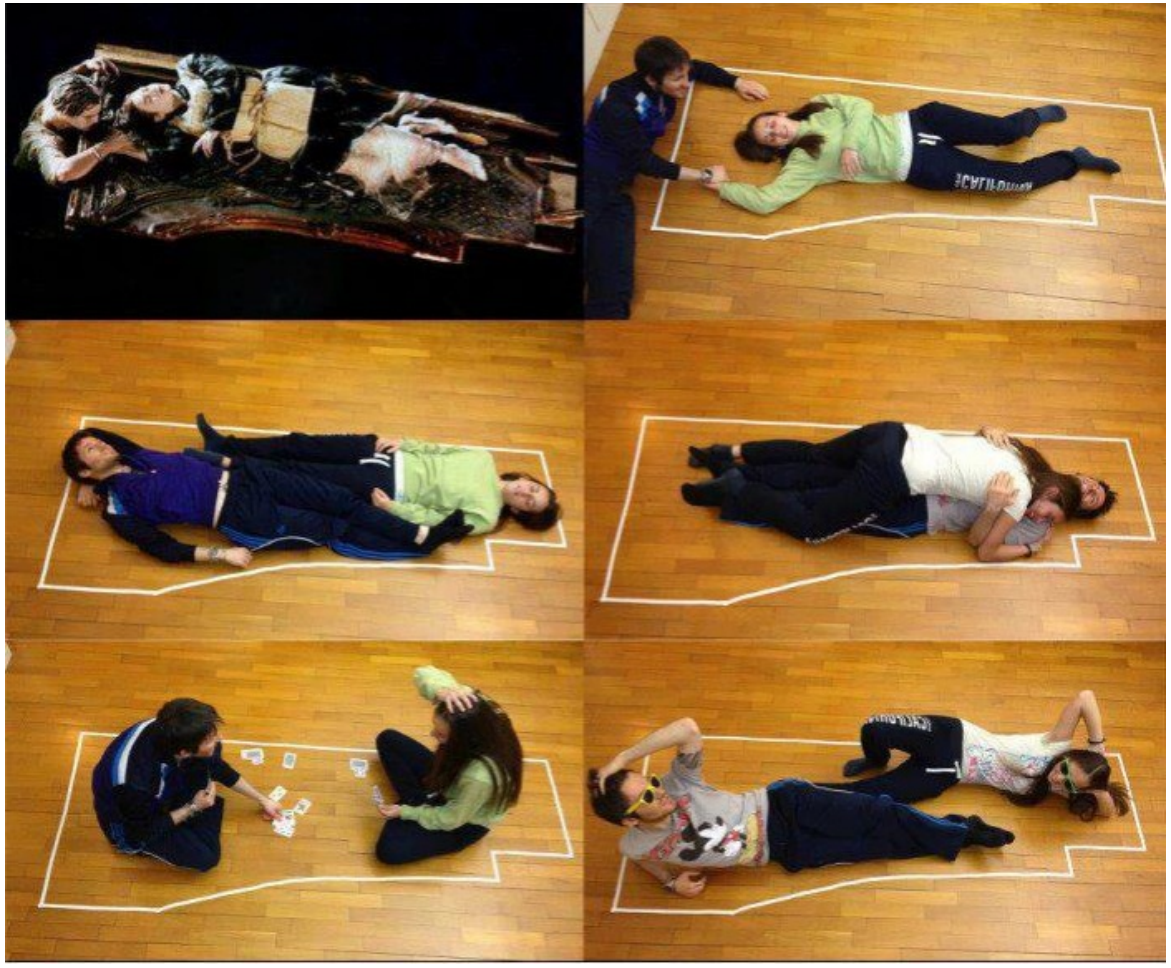
Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

Predict


no
0.38
100%
yes
Sex = male
no
no
0.19
65%
yes
0.74
35%
Age >= 6.5
Pclass = 3
no
0.17
62%
yes
0.67
3%
no
0.50
16%
yes
0.95
19%
Family >= 4
Fare >= 23
no
0.11
1%
yes
1.00
2%
no
0.11
3%
yes
0.59
13%
Age >= 37
no
0.14
1%
yes
0.62
12%
Fare >= 7.9
yes
0.55
8%
yes
0.73
5%
Fare < 11
no
0.32
2%
yes
0.66
5%

But we all know that Jack could have survived



Import, Tidy, & Transform the Test Data

```
# Import the test.csv file
titanic.test <- read_csv(file = "data/test.csv", col_names = TRUE)

# Convert the Passenger Class to a factor
titanic.test$Pclass <- as.factor(titanic.test$Pclass)

# Tidy the missing Age observations
titanic.test$Age <- if_else(is.na(titanic.test$Age),
                            mean(titanic$Age, na.rm = T),
                            titanic.test$Age)

# Create the Family variable
titanic.test$Family <- titanic.test$SibSp + titanic.test$Parch

# Grab the variables we originally used
titanic.test <- titanic.test %>%
                select(PassengerId, Pclass, Sex, Age, Fare, Family)
```

Import, Tidy, & Transform the Training Data

```
titanic.test$Predict <- predict(tree.titanic, titanic.test, type = "class")

summary(titanic.test$Predict)
```

```
##  no yes
## 284 134
```

```
titanic.test$Survived <- if_else(titanic.test$Predict == "yes", 1, 0)

titanic.submit <- titanic.test %>%
                    select(PassengerId, Survived)


write.csv(titanic.submit, file = "data/titanic_answer.csv", row.names = FALSE)
```

How did we do?

kaggle™

4931	new	zenchannel	0.77990	4	16h
4932	new	fredericohoffmann	0.77990	1	1d
4933	new	Dmitry Soroka	0.77990	3	1d
4934	▲ 3139	Cherma Ramalho	0.77990	20	12m
4935	new	Yarbreezy	0.77990	3	20h
4936	new	ryanrhall	0.77990	1	21h
4937	new	Gajendra Saraswat	0.77990	2	2h
4938	new	Gemma Dawson	0.77990	2	2h

Your Best Entry ↑

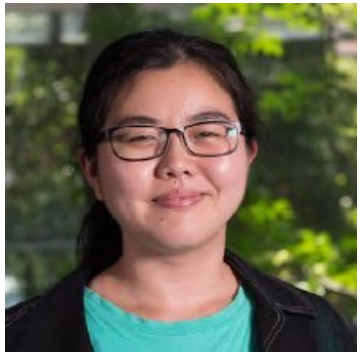
Your submission scored 0.77990, which is not an improvement of your best score. Keep trying!

4939	new	Potatoes Legend	0.77990	3	18h
4940	new	In Woo	0.77990	5	11h
4941	new	MMA Student	0.77990	3	10h
4942	new	Windowsill	0.77990	3	9h
4943	new	Maxime Godfroid	0.77990	8	9h
4944	new	Ayush Anshul	0.77990	15	8h
4945	new	kangyu1979	0.77990	1	6h

ma Dawson

www.icepack.ai
@GemmaDawson
@gemmadawsonza

Thank you



Emi Tanaka

@emitanaka

@statsgen