
# Machine Learning 101

Jeanne Goossens
The Space Leiden

21-01-2021

# Programma

Wat is Machine Learning?

Leren van data

Aan de slag met de Titanic

# Wat is machine learning?

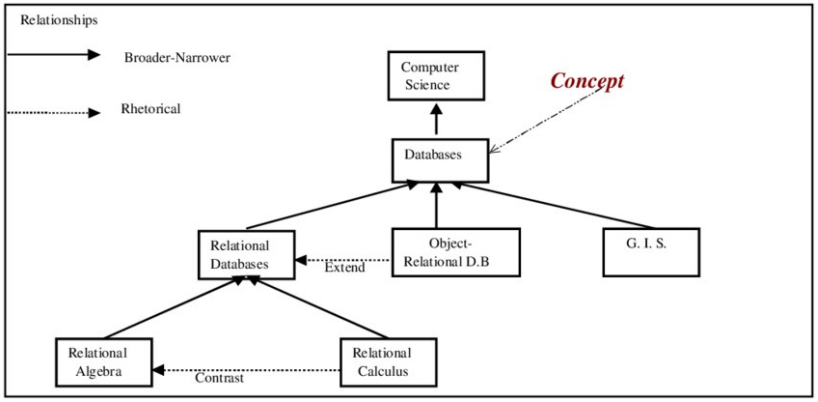
“Learning can be understood as a learning way to automatically find patterns and structure in data by optimizing the parameters of the model” (source).

# Wat is machine learning?

“Learning can be understood as a learning way to automatically find patterns and structure in data by optimizing the parameters of the model” (source).

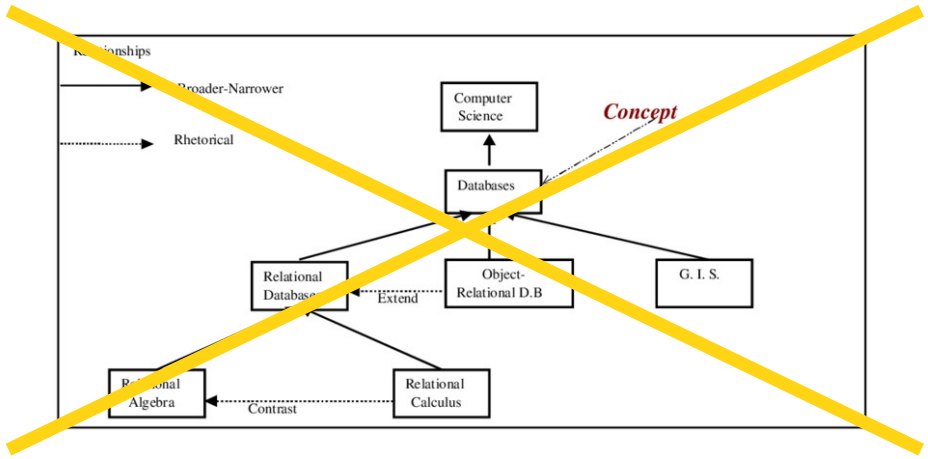
# Wat is machine learning?

"Learning can be understood as a learning way to automatically find patterns and structure in data by optimizing the parameters of the model" (source).



# Wat is machine learning?

"Learning can be understood as a learning way to automatically find patterns and structure in data by optimizing the parameters of the model" (source).



# Wat is machine learning?

“Learning can be understood as a learning way to automatically find patterns and structure in data by optimizing the parameters of the model” (source).

80% cleaning, 20% modelling

# Leren van data

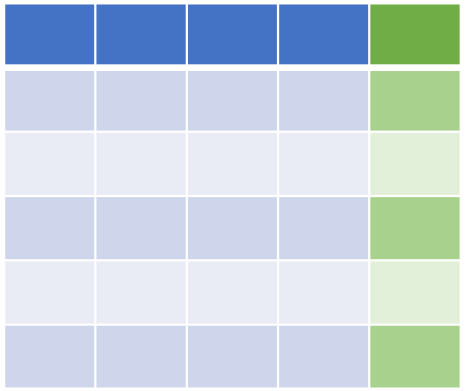
Onafhankelijke variabelen

Afhankelijke variabele(n)

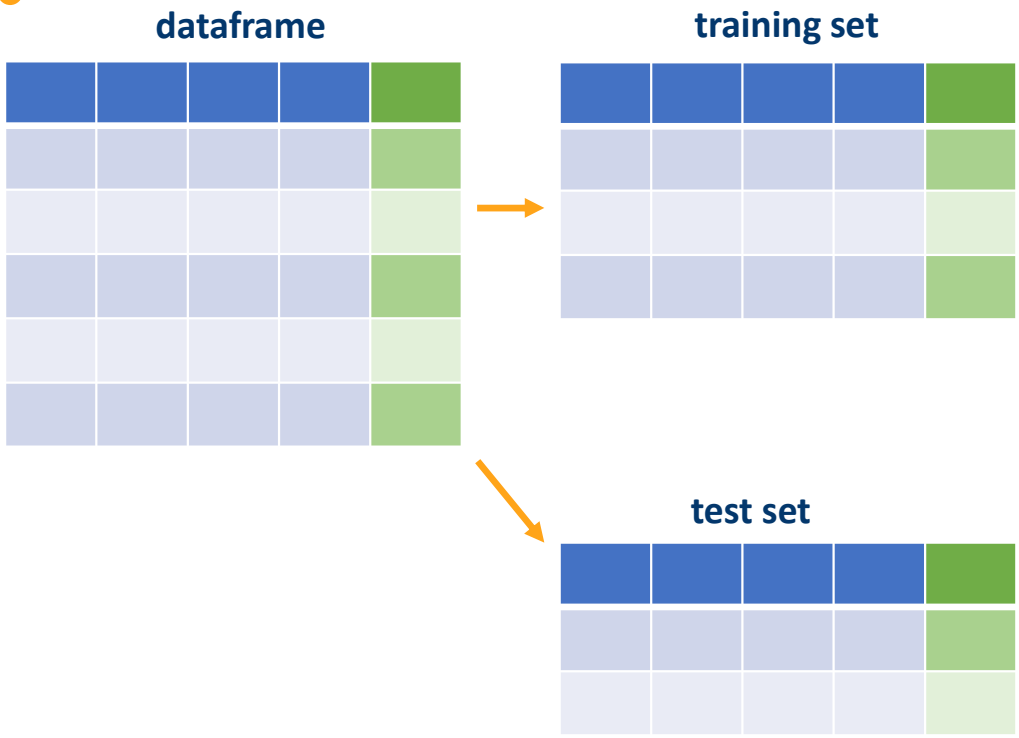
| Passenger id | Class | Age | Gender | Survival |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |

# Leren van data

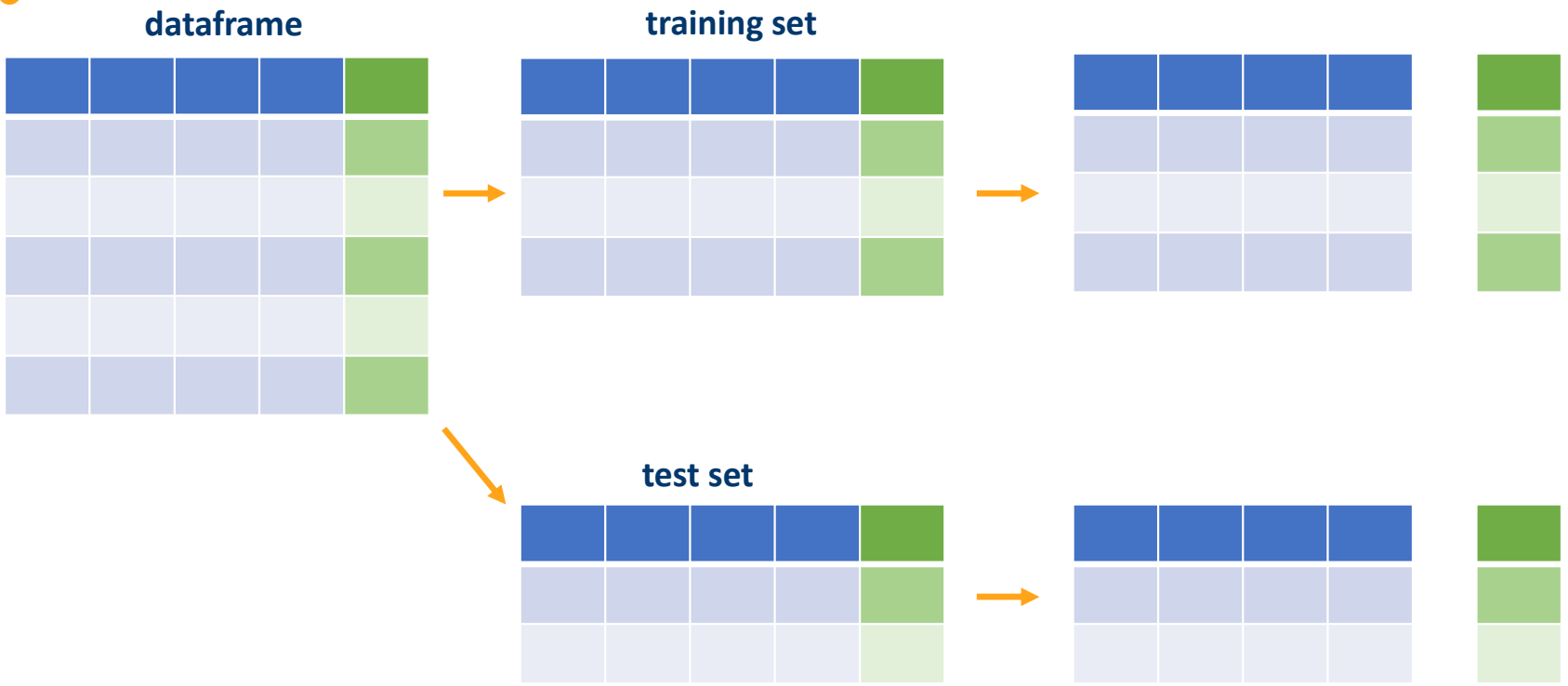
dataframe



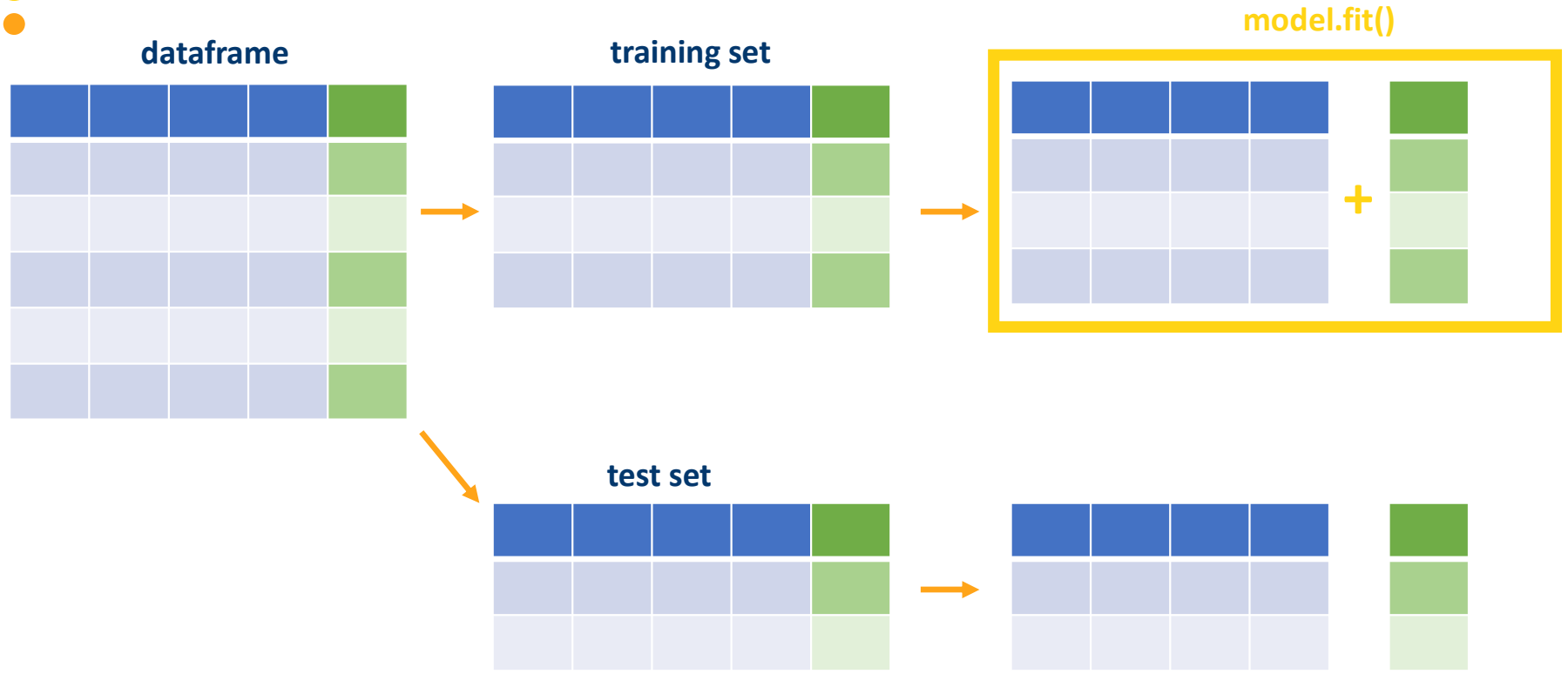
# Leren van data



# Leren van data

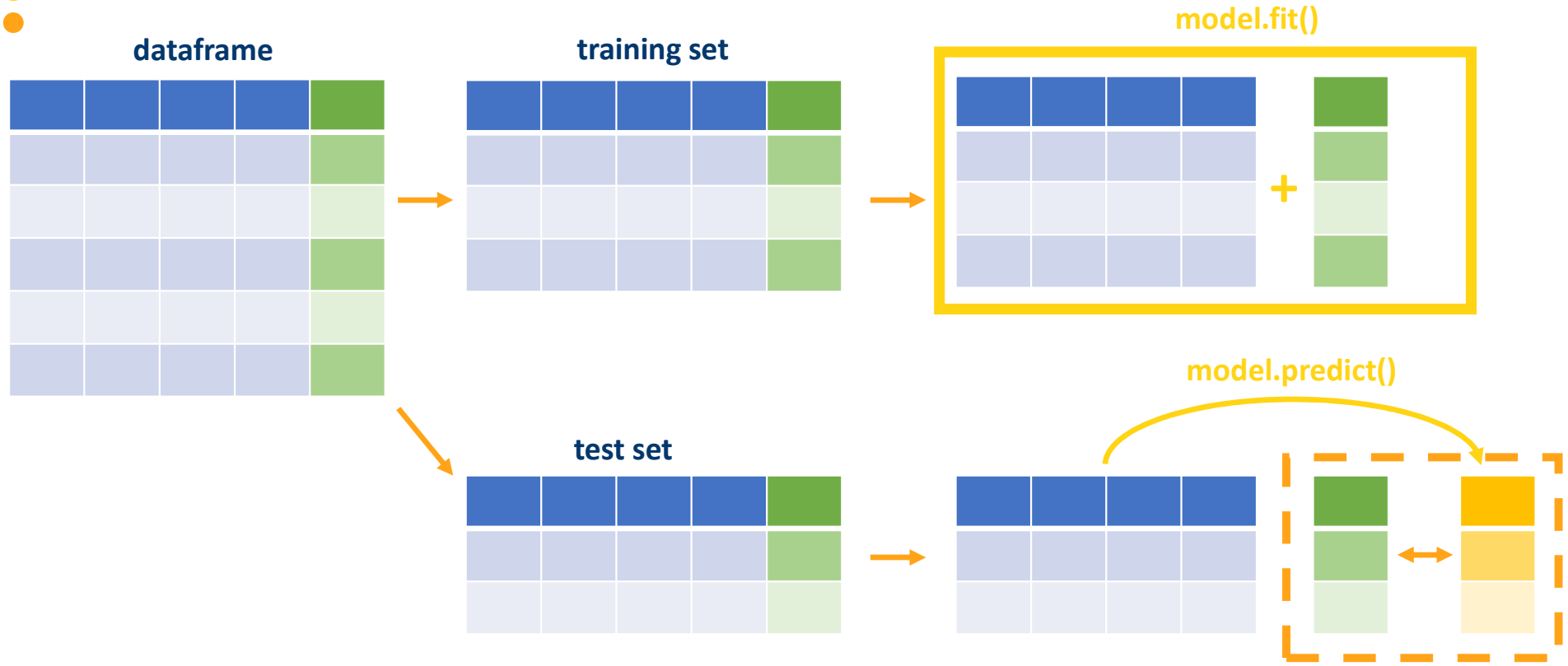


# Leren van data



# Leren van data

# Leren van data



# Aan de slag!

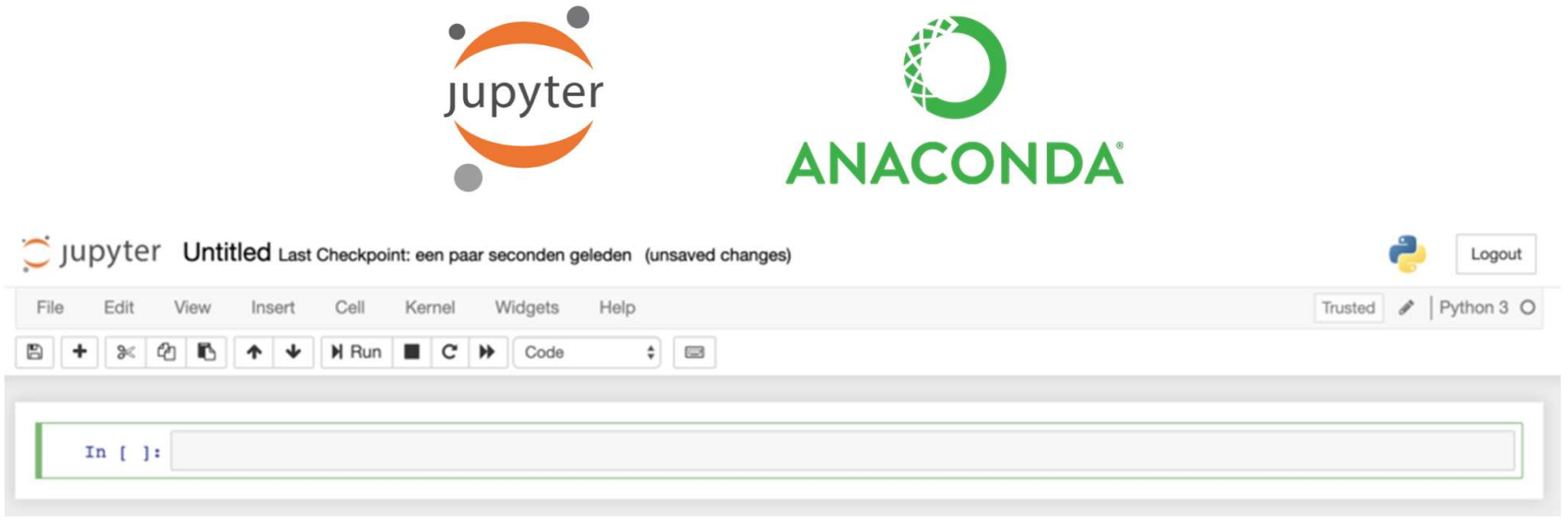
Als je lokaal wil draaien:

www.github.com/jeannegoossens/machine-learning-101

Als je in een virtual notebook wilt werken:

https://colab.research.google.com/drive/1_de4HeQVsBSuXKEt5c-gPTafY0k1b7R7?usp=sharing

# Jupyter notebooks




```
In [1]:  1 print('Hello world')
executed in 26ms, finished 18:07:47 2021-01-16
Hello world
```

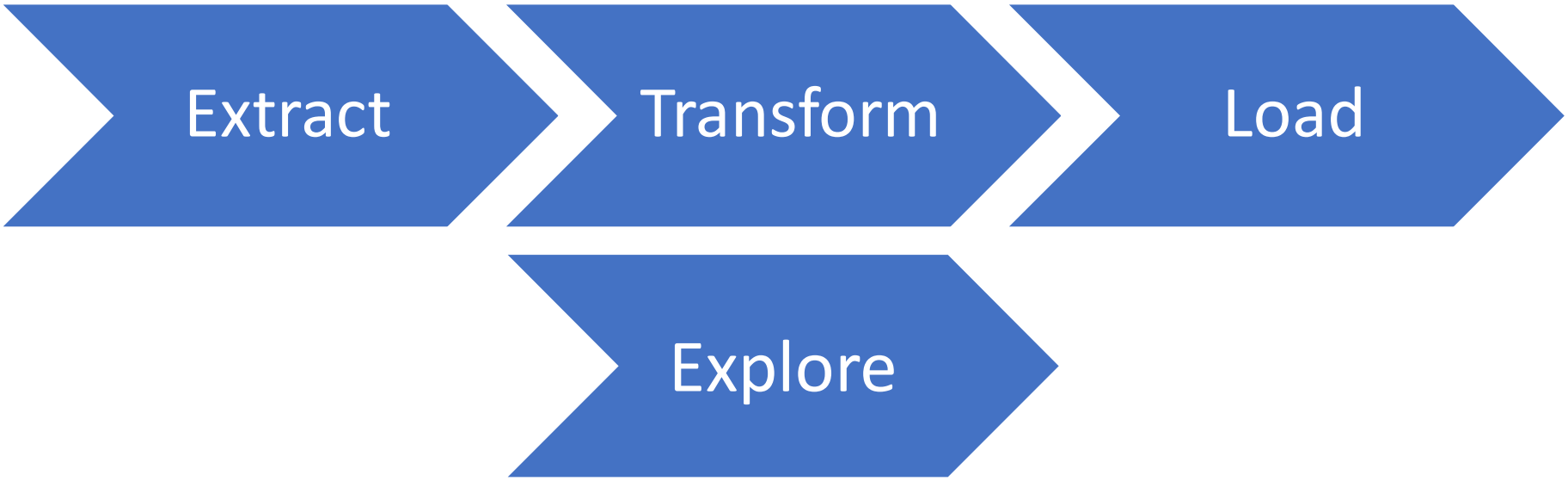
# Titanic.csv

**dataframe**

| class | name | sex | age | sibsp | parch | ticket | fare | cabin | embarked | boat | body | home.dest | survived |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Allen, Miss. Elisabeth Walton | female | 29 | 0 | 0 | 24160 | 211.3375 | B5 | S | 2 | ? | St Louis, MO | 1 |
| 1 | Allison, Master. Hudson Trevor | male | 0.9167 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | 11 | ? | Montreal, PQ / Chesterville, ON | 1 |
| 1 | Allison, Miss. Helen Loraine | female | 2 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | ? | ? | Montreal, PQ / Chesterville, ON | 0 |
| 1 | Allison, Mr. Hudson Joshua Creighton | male | 30 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | ? | 135 | Montreal, PQ / Chesterville, ON | 0 |
| 1 | Allison, Mrs. Hudson J C (Bessie Waldo Daniels) | female | 25 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | ? | ? | Montreal, PQ / Chesterville, ON | 0 |

# ETL

# ETL



# Extract

Out[3]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 431 | Yes | 1 | Bjornstrom-Steffansson, Mr. Mauritz Hakan | male | 28.0 | 0 | 0 | 110564 | 26.5500 | C52 | S |
| **1** | 664 | No | 3 | Coleff, Mr. Peju | male | 36.0 | 0 | 0 | 349210 | 7.4958 | NaN | S |
| **2** | 44 | Yes | 2 | Laroche, Miss. Simonne Marie Anne Andree | female | 3.0 | 1 | 2 | SC/Paris 2123 | 41.5792 | NaN | C |
| **3** | 347 | Yes | 2 | Smith, Miss. Marion Elsie | female | 40.0 | 0 | 0 | 31418 | 13.0000 | NaN | S |
| **4** | 891 | No | 3 | Dooley, Mr. Patrick | male | 32.0 | 0 | 0 | 370376 | 7.7500 | NaN | Q |

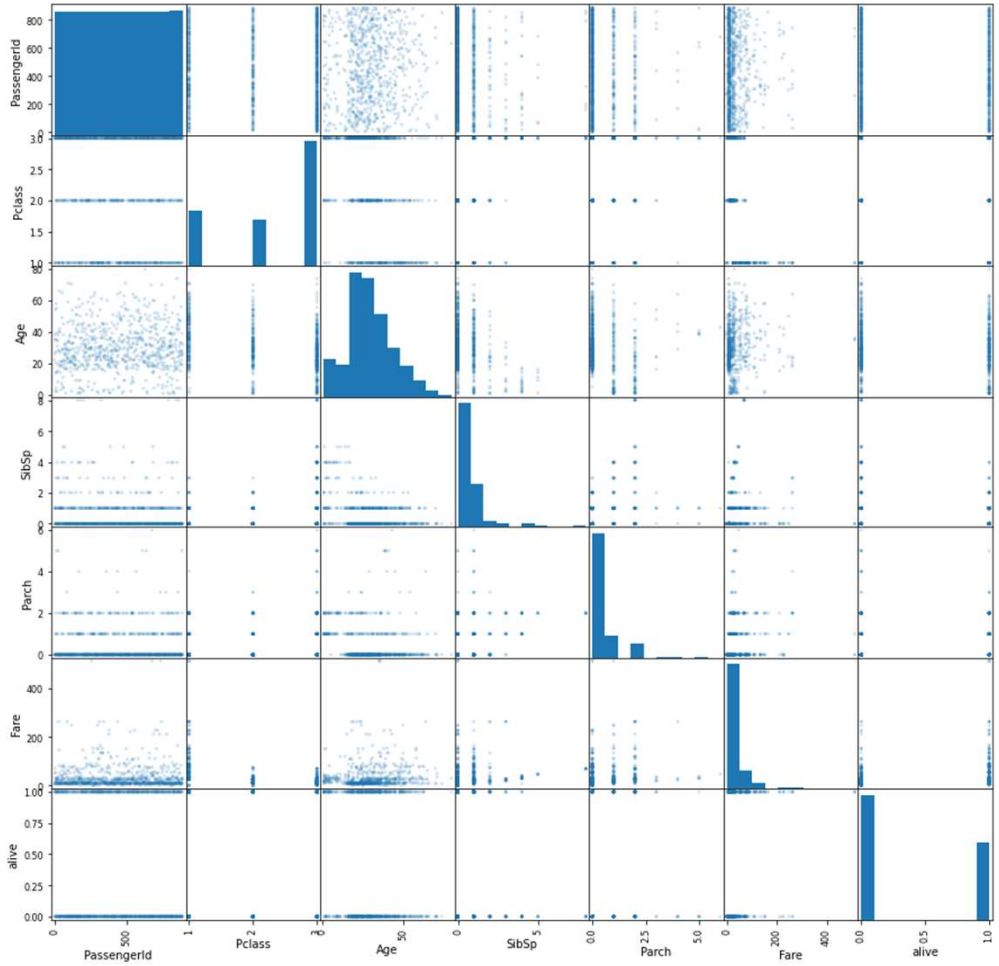
# Hypotheses?

Wat verwachten we dat de grootste voorspellende karakteristieken zijn?

- Leeftijd?
- Geslacht?
- Klasse?

Dit kunnen we testen op basis van onze data. Daarnaast kunnen we straks zien of het model deze kenmerken ook als relevante parameters ziet!

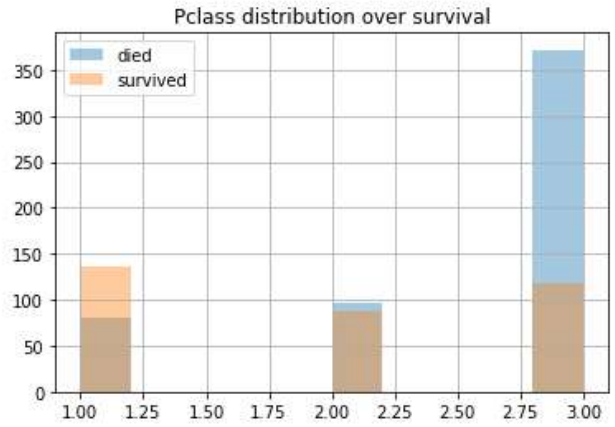
# Explore



# Hypotheses?

Klasse?

| Pclass | alive |
|---|---|
| 1 | 0.629630 |
| 2 | 0.472826 |
| 3 | 0.242363 |

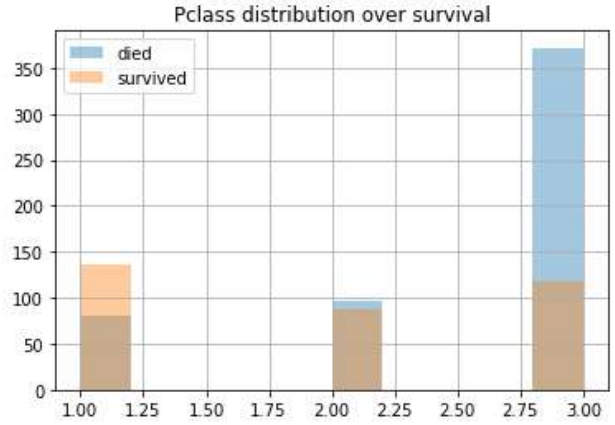


# Hypotheses?

Klasse?

Geslacht?

| Pclass | alive |
|---|---|
| 1 | 0.629630 |
| 2 | 0.472826 |
| 3 | 0.242363 |

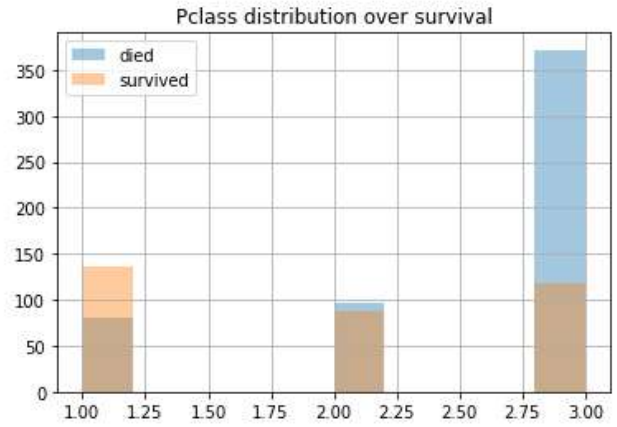



| Sex | alive |
|---|---|
| female | 0.742038 |
| male | 0.188908 |

# Hypotheses?

Klasse?

| Pclass | alive |
|---|---|
| 1 | 0.629630 |
| 2 | 0.472826 |
| 3 | 0.242363 |

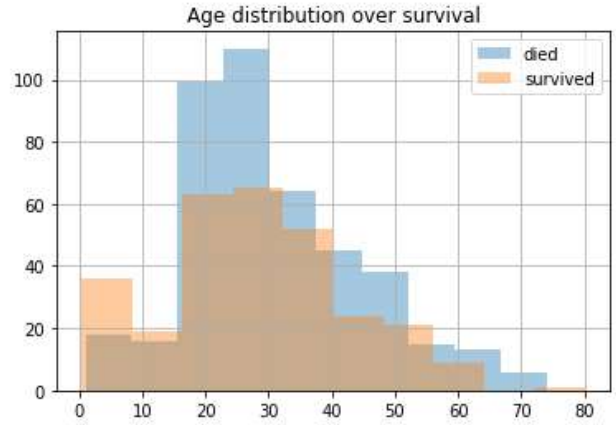



Geslacht?

| Sex | alive |
|---|---|
| female | 0.742038 |
| male | 0.188908 |

Leeftijd?

| adult | alive count | alive mean |
|---|---|---|
| No | 113 | 0.539823 |
| Yes | 601 | 0.381032 |



# Transform

Missende data opvullen?

- Alsnog achterhalen bij bron of Internet
- Achterhalen uit andere kolommen
- Opvullen met gemiddelde, modus, mediaan?
- Rij verwijderen of kolom verwijderen

Extra data toevoegen?

# Data voorbereiden

| Embarked |
| --- |
| S |
| C |
| Q |
| C |

# Data voorbereiden

| Embarked | Embarked_S | Embarked_C | Embarked_Q |
|---|---|---|---|
| S | 1 | 0 | 0 |
| C | 0 | 1 | 0 |
| Q | 0 | 0 | 1 |
| C | 0 | 1 | 0 |

# Data voorbereiden

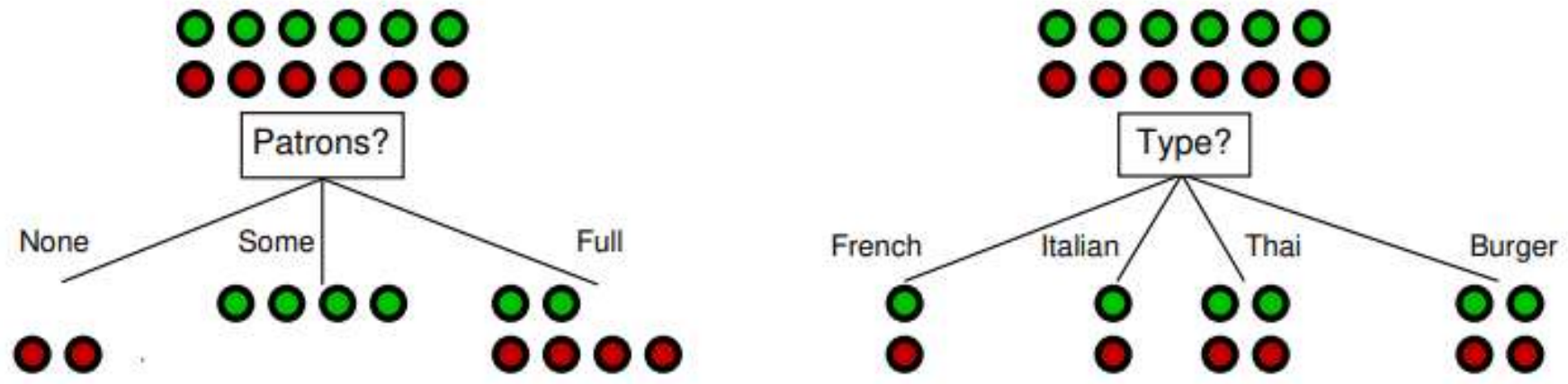
| Embarked |
| --- |
| S |
| C |
| Q |
| C |

| Embarked_S | Embarked_C | Embarked_Q |
| --- | --- | --- |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

# Data voorbereiden

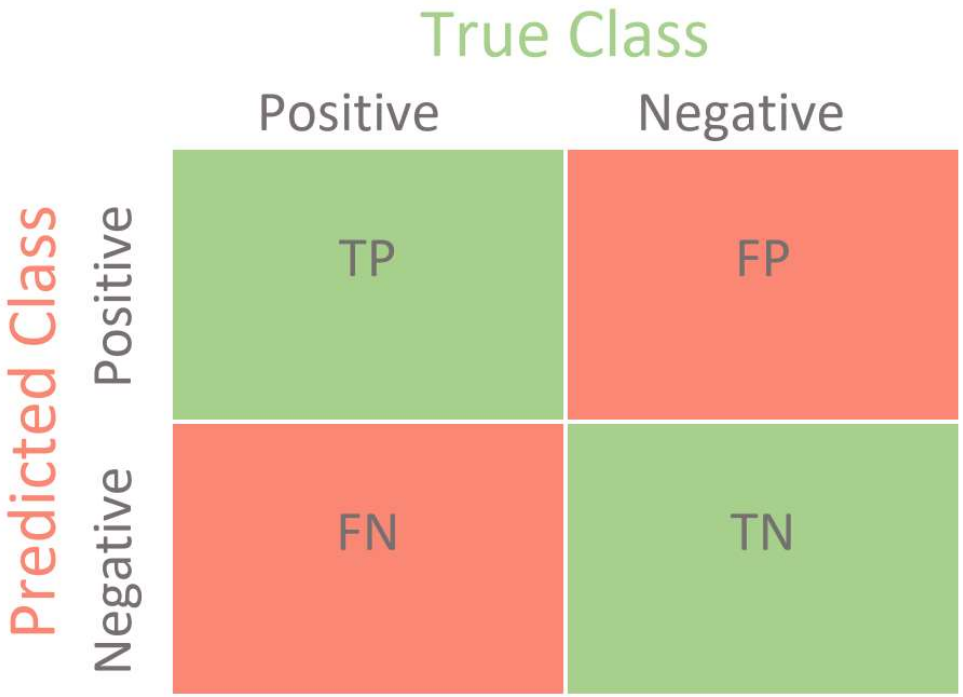
| Embarked | Embarked_S | Embarked_C | Embarked_Q |
|---|---|---|---|
| S | 1 | 0 | 0 |
| C | 0 | 1 | 0 |
| Q | 0 | 0 | 1 |
| C | 0 | 1 | 0 |

# Model trainen: decision tree



# Model trainen: decision tree

# Model evalueren



# Model tweaken

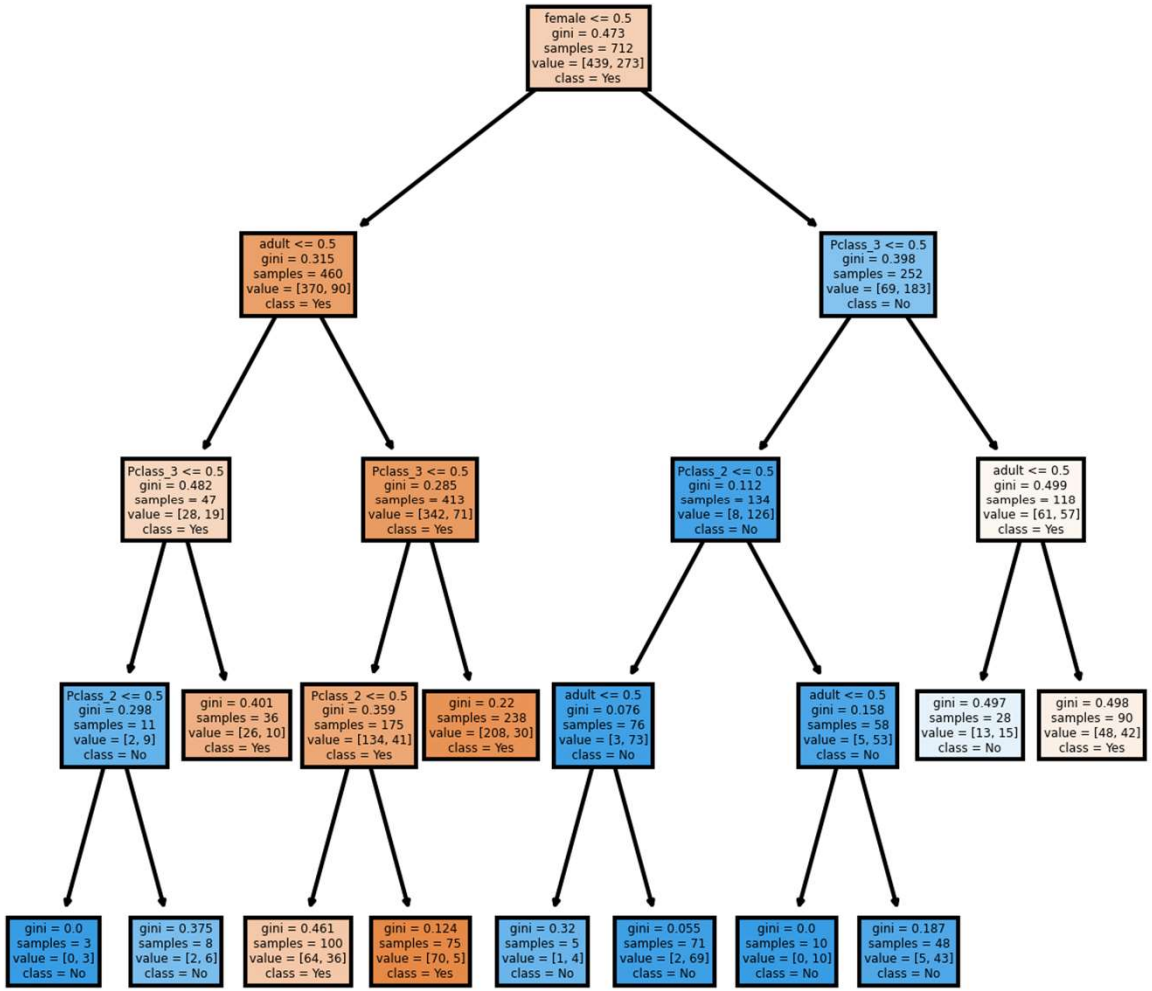
- Extra databewerking
  - Leeftijd categoriseren?
  - Persoonlijke titels gebruiken?
  - Fare categoriseren?

Probeer dingen uit, en test het effect op het model!

# Onze eigen hypotheses?

Wat zien we in het model terug van onze hypotheses? Wat als we een heel simple model draaien, met alleen die kolommen?

Het model wordt er beter van! Waarom is dat?


female <= 0.5
gini = 0.473
samples = 712
value = [439, 273]
class = Yes
adult <= 0.5
gini = 0.315
samples = 460
value = [370, 90]
class = Yes
Pclass_3 <= 0.5
gini = 0.398
samples = 252
value = [69, 183]
class = No
Pclass_3 <= 0.5
gini = 0.482
samples = 47
value = [28, 19]
class = Yes
Pclass_3 <= 0.5
gini = 0.285
samples = 413
value = [342, 71]
class = Yes
Pclass_2 <= 0.5
gini = 0.112
samples = 134
value = [8, 126]
class = No
adult <= 0.5
gini = 0.499
samples = 118
value = [61, 57]
class = Yes
Pclass_2 <= 0.5
gini = 0.298
samples = 11
value = [2, 9]
class = No
gini = 0.401
samples = 36
value = [26, 10]
class = Yes
Pclass_2 <= 0.5
gini = 0.359
samples = 175
value = [134, 41]
class = Yes
gini = 0.22
samples = 238
value = [208, 30]
class = Yes
adult <= 0.5
gini = 0.076
samples = 76
value = [3, 73]
class = No
adult <= 0.5
gini = 0.158
samples = 58
value = [5, 53]
class = No
gini = 0.497
samples = 28
value = [13, 15]
class = No
gini = 0.498
samples = 90
value = [48, 42]
class = Yes
gini = 0.0
samples = 3
value = [0, 3]
class = No
gini = 0.375
samples = 8
value = [2, 6]
class = No
gini = 0.461
samples = 100
value = [64, 36]
class = Yes
gini = 0.124
samples = 75
value = [70, 5]
class = Yes
gini = 0.32
samples = 5
value = [1, 4]
class = No
gini = 0.055
samples = 71
value = [2, 69]
class = No
gini = 0.0
samples = 10
value = [0, 10]
class = No
gini = 0.187
samples = 48
value = [5, 43]
class = No

# Zou jij overleven?

Nu we een getraind model hebben, kunnen we ook kijken of wij zelf overleefd zouden hebben. Welke klasse zou jij minimal moeten reizen om de ramp overleefd te hebben?

# The Space planning

Volgende week: The Space - Capture The Flag door Niels

Daarna: Github CI door Wesley