WOULD YOU HAVE SURVIVED IF YOU WERE ON TITANIC?

KAVEH **BAKHTIYARI**

Sr. Data Scientist

OBJECTIVE

Today, we are trying the achieve the following objectives:

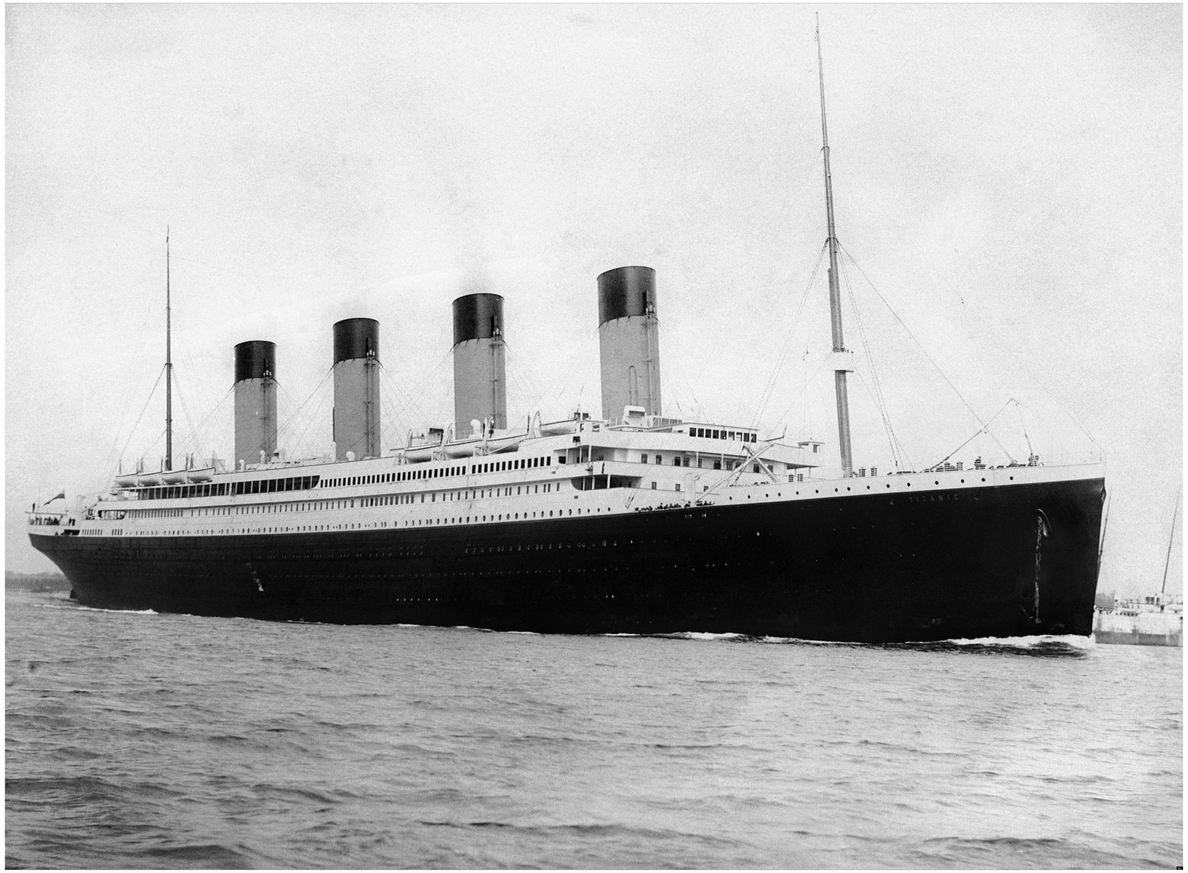
- To understand simple classification machine learning
- To train an ML model
- To test an ML model

And finally:

- To check if we would have survived if were on Titanic?

TITANIC

TITANIC



Titanic was a British passenger liner that sank in the North Atlantic Ocean on 15 April 1912 after striking an iceberg during her maiden voyage from Southampton, England, to New York City, United States.

Of the estimated **2,224 passengers and crew aboard**, **more than 1,500 died**, making it the deadliest sinking of a single ship up to that time.

AGENDA

1
TITANIC?

2
CLASSIFICATIO
N

3
DECISION
TREE

4
DATASET

5
RapidMiner

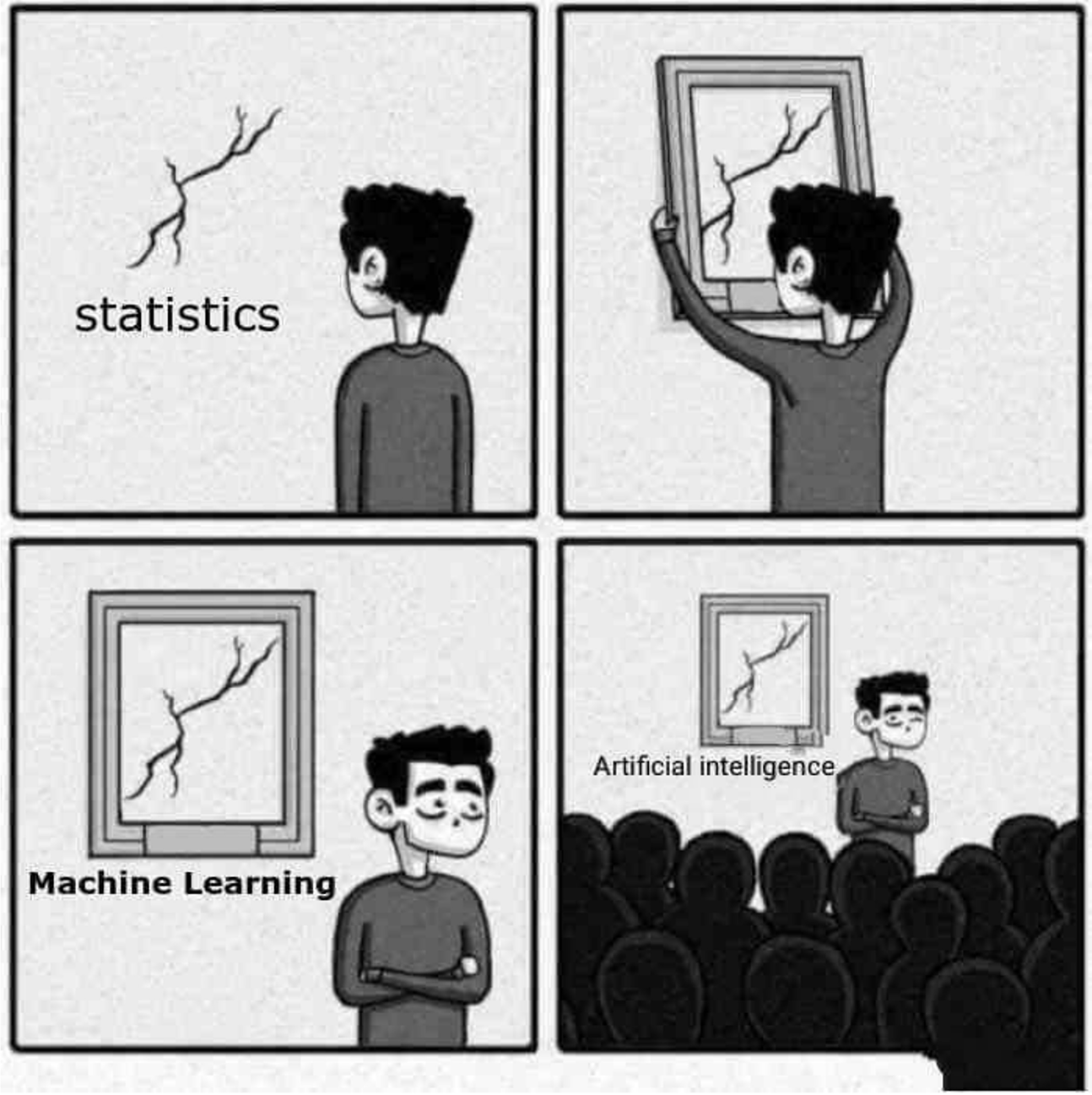
6
SURVIVAL?

7
Q&A

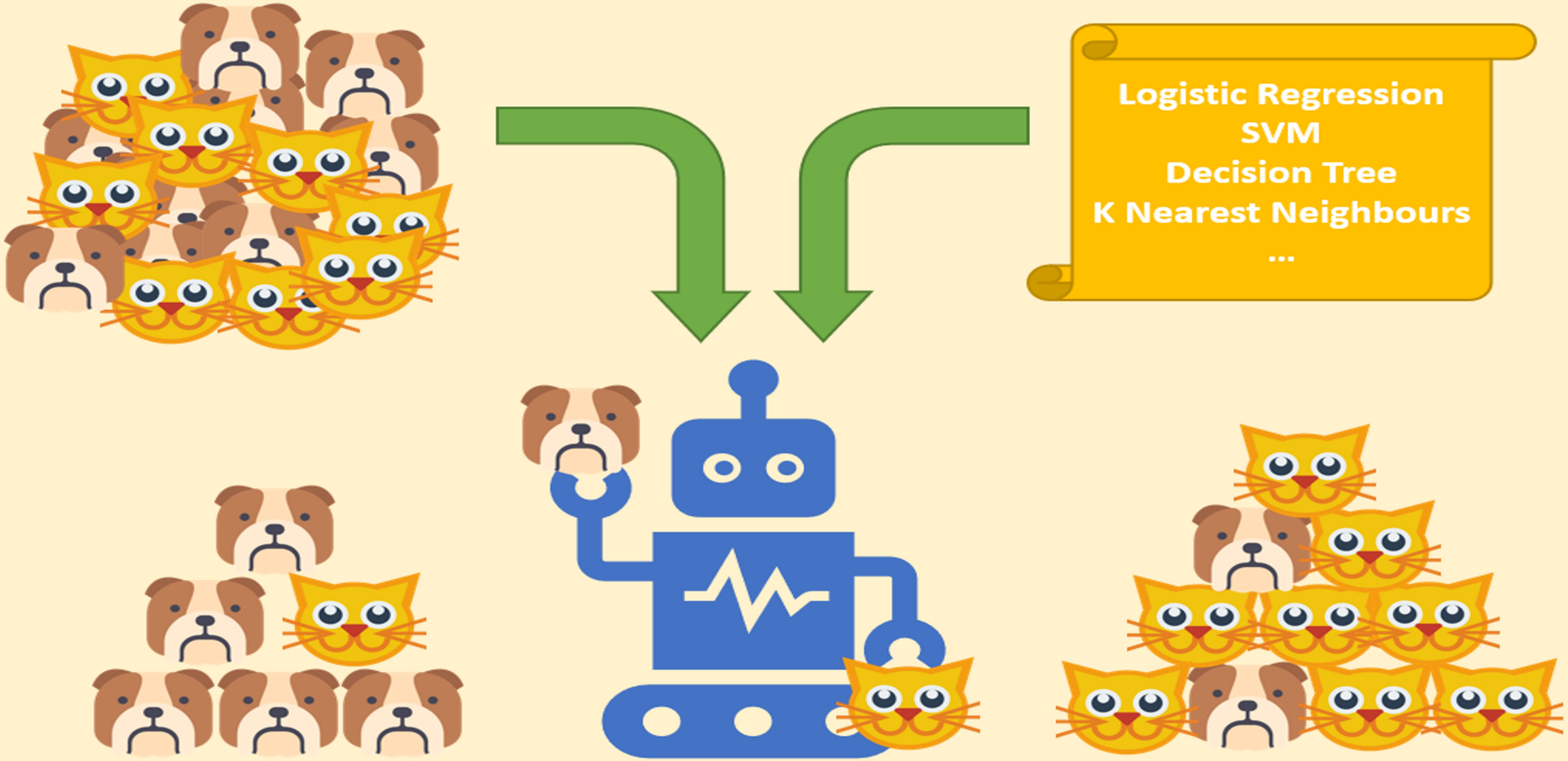
STATS, ML & AI

TITANIC

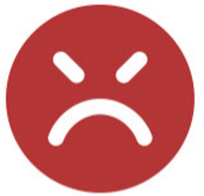
STATISTICS vs ML vs AI



CLASSIFICATION


Logistic Regression
SVM
Decision Tree
K Nearest Neighbours
...

CLASSIFICATION



Negative

I'm dissatisfied with your customer service. No one was able to help me with the problems I had with using your product.

Neutral

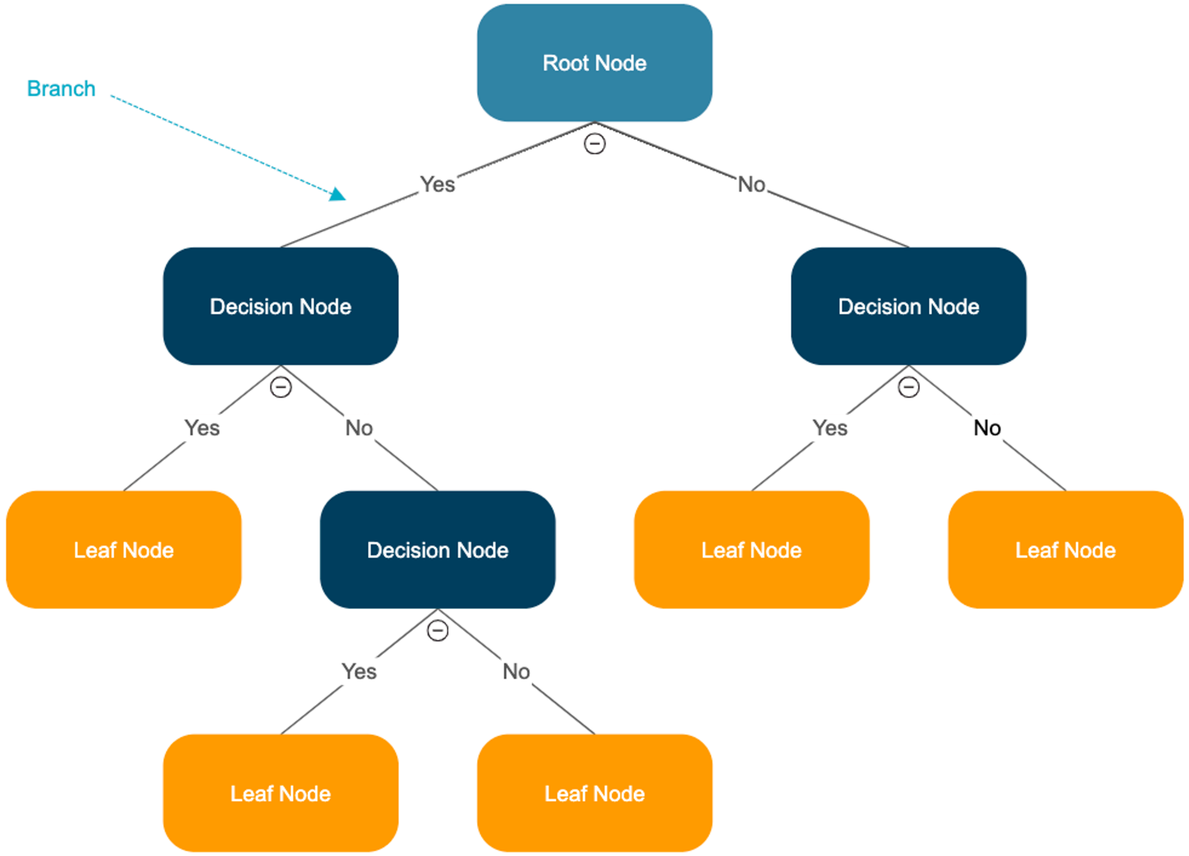
The product has multiple features that are suitable for users with different levels of experience.

Positive

I really enjoy how easy this product is to use and how it successfully helps my team complete their day-to-day tasks.

DECISION TREE

TITANIC



DECISION TREE

A decision tree is a decision support hierarchical model that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

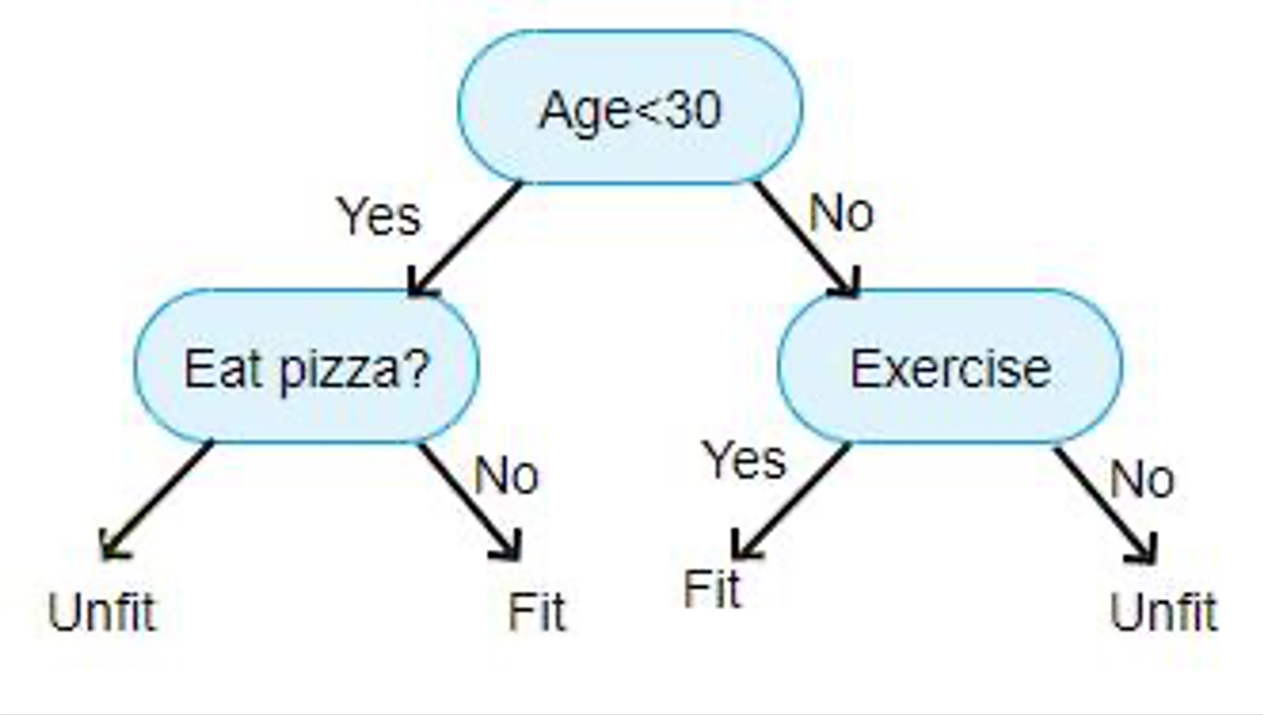
The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

PROS:

- Simple to understand and to interpret. Trees can be visualized.
- Requires little data preparation.
- Able to handle both numerical and categorical data.
- Able to handle multi-output problems.

CONS:

- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.



Example:

Are you fit?

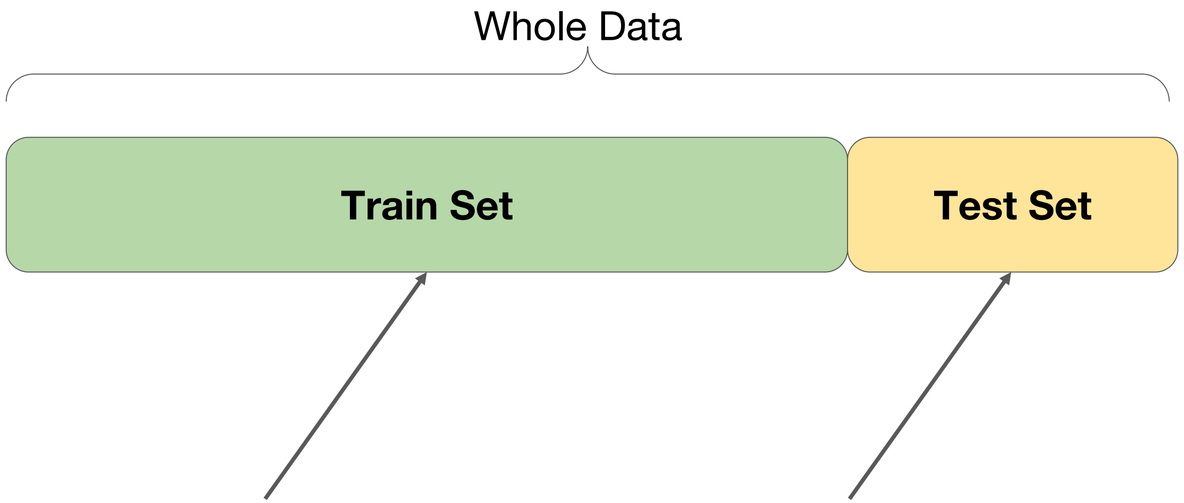
Features:

- Age
- Eat Pizza
- Exercise

Label (to be predicted):

- Fit

EVALUATION



- To train and build the ML model
- To test and evaluate the ML model

EVALUATION

Once the model is trained, it is time to evaluate the model and measure how accurate the model is.

The common metric for classification problems is “Accuracy” or the percentage of correct predictions out of all predictions.

METRIC:

- **ACCURACY** = True Predictions / Total Predictions

TRAIN/TEST Split:

Data should be split into two sets of Train and Test:

- **Train:**
  - Train set is being used to generate the decision tree
- **Test:**
  - Test set is being used to evaluate the decision tree

DATASET

TITANIC

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

DECISION TREE

Records: 1,309 records

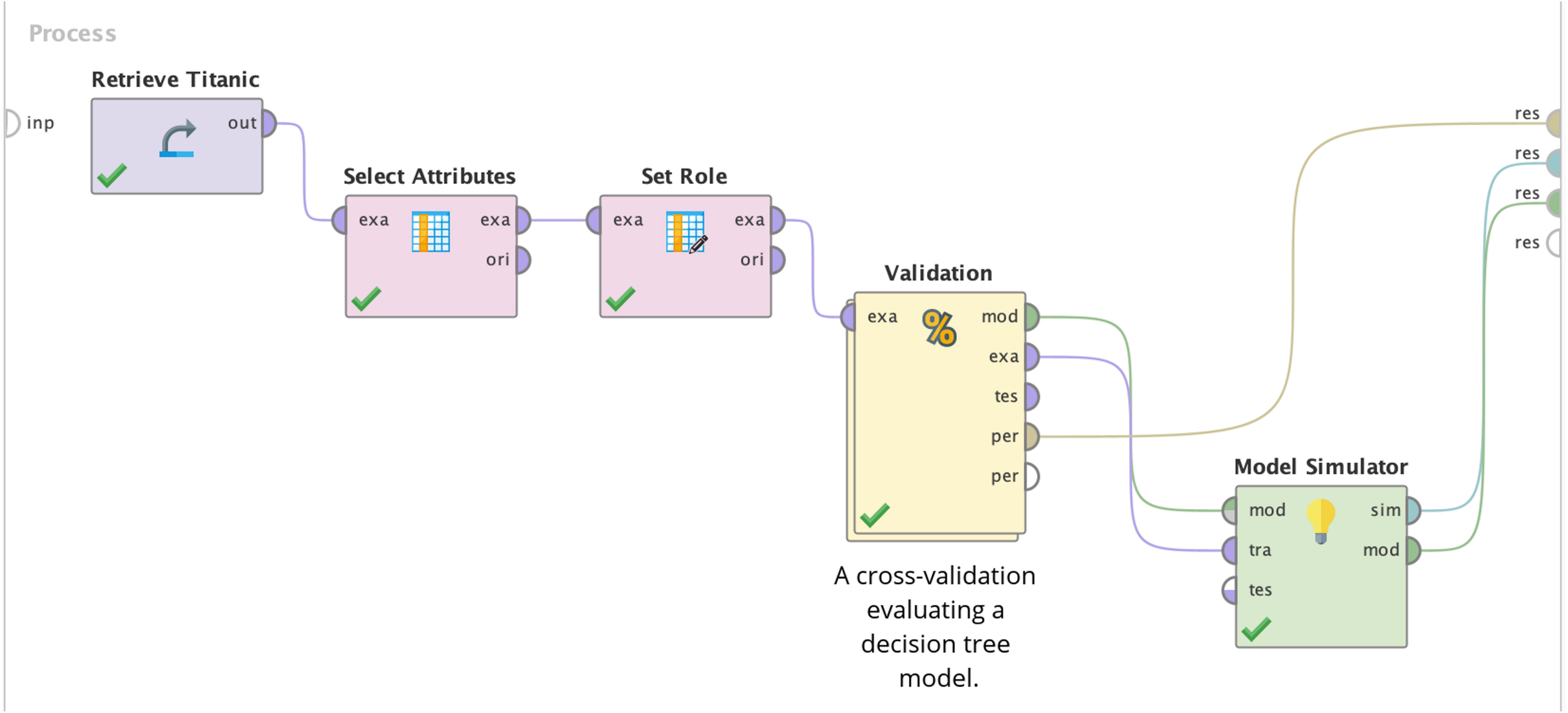
Columns:
- Passenger Name
- Passenger Class
- Sex
- Age
- Number of Siblings
- Number of Parents / Children Aboard
- Ticket Number
- Fare
- Cabin
- Port of Embark

And finally:
- Survived?

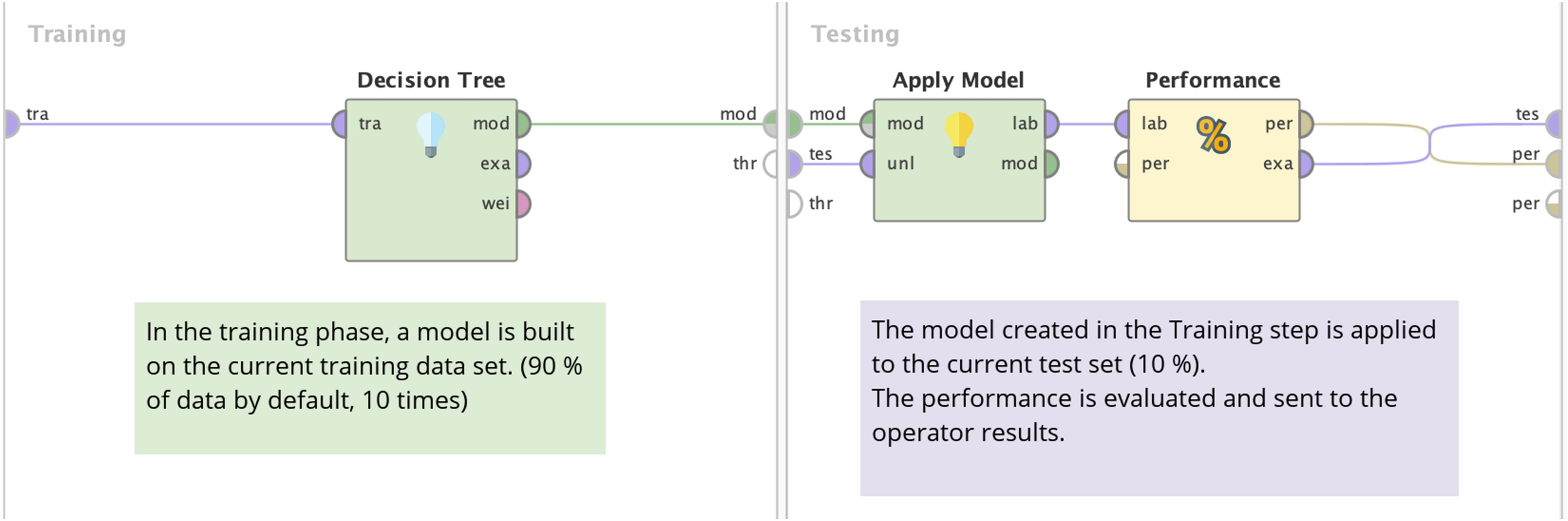
RapidMiner

Please download and install ***RapidMiner Studio***…

https://rapidminer.com/


Process
Retrieve Titanic
inp
out
Select Attributes
exa
exa
ori
Set Role
exa
exa
ori
Validation
exa
mod
exa
tes
per
per
A cross-validation evaluating a decision tree model.
Model Simulator
mod
tra
tes
sim
mod
res
res
res
res

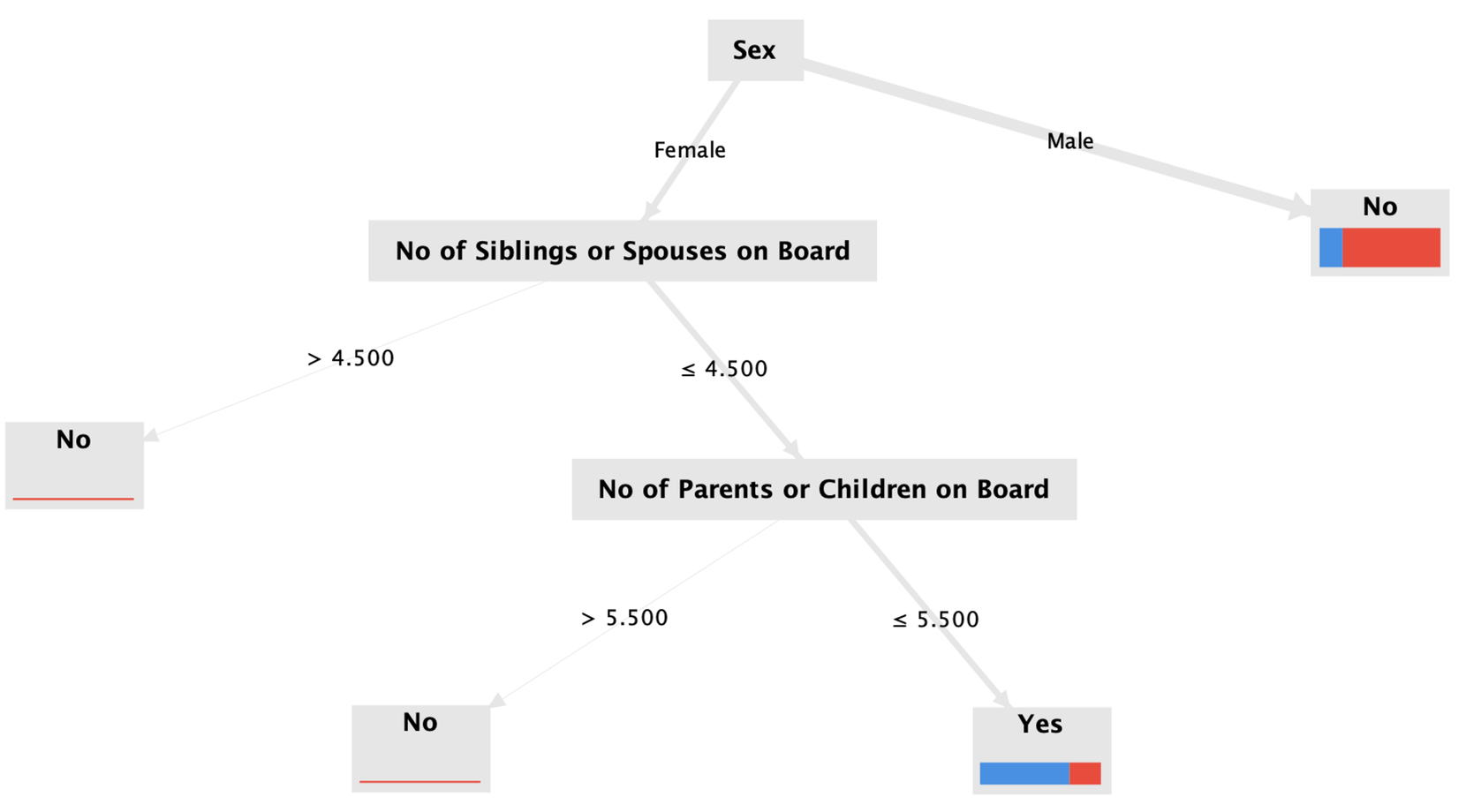
DEMO - TRAIN & TEST



In the training phase, a model is built on the current training data set. (90 % of data by default, 10 times)

The model created in the Training step is applied to the current test set (10 %).
The performance is evaluated and sent to the operator results.

DEMO - DECISION TREE

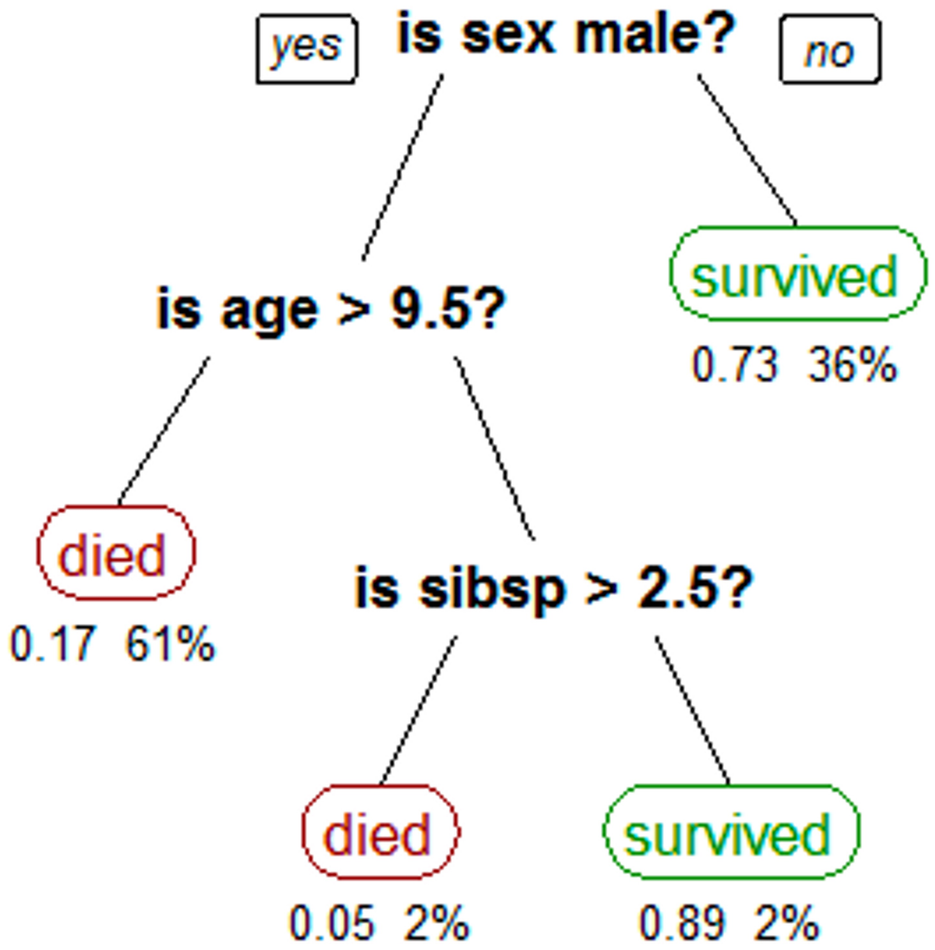


accuracy: 77.08% +/- 4.11% (micro average: 77.08%)

	true Yes	true No	class precision
pred. Yes	304	104	74.51%
pred. No	196	705	78.25%
class recall	60.80%	87.14%	

SURVIVAL?

TITANIC


yes
is sex male?
no
is age > 9.5?
survived
0.73 36%
died
0.17 61%
is sibsp > 2.5?
died
0.05 2%
survived
0.89 2%

CLOSING

- Machine Learning (ML) classification can be used to classify objects based on their given features (parameters)

- Decision Tree is one of the classification algorithms which is easy to interpret and understand.

- Accuracy of the decision tree is dependent on data quality and its relevance to the problem.

- Decision Tree finds hidden patterns in the data

- Decision Tree can help us recognize the importance of features in order to get the correct answer

Q&A