KSCHOOL

Modelling Methodologies

Data Science

Irene Torres Valle

KSCHOOL

¿Who am I?

Modelling Methodologies

KSCHOOL

Index

KSCHOOL

Index

Modelling Methodologies

Every day

2,5 trillions of bytes
data

Every minute

72 hours
video

277.000
tweets

200 millions
emails

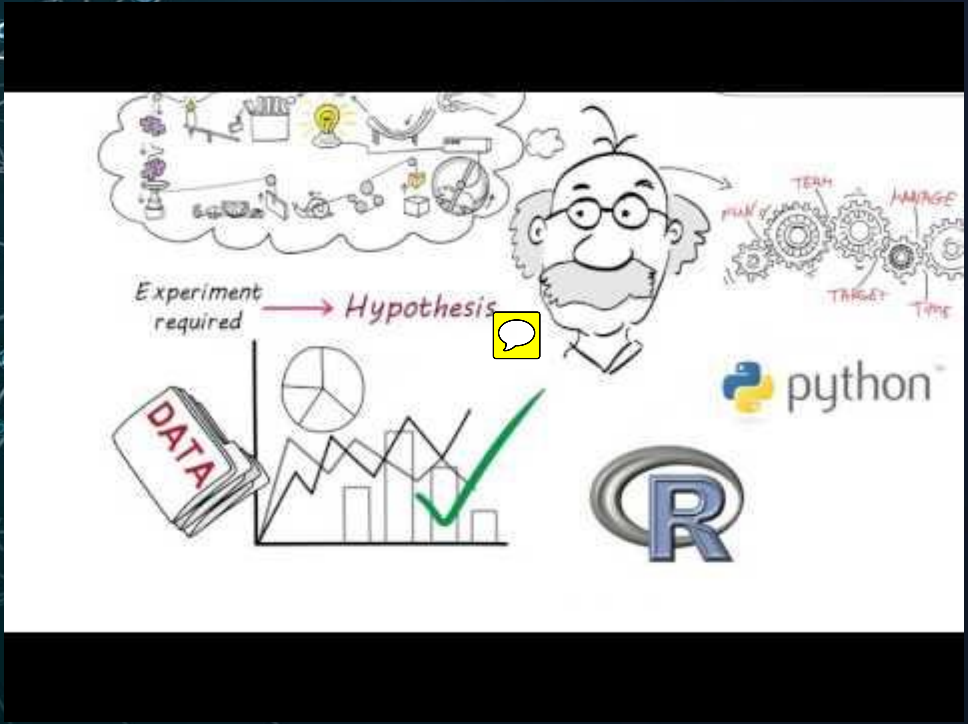
Data obtained from "IBM Acelera el BIG DATA":
https://www.ibm.com/developerworks/community/blogs/insider/entry/acelera_bigdata?lang=en

KSCHOOL

“Data Scientist:
The Sexiest Job of the 21st Century.”

Harvard Business Review

Modelling Methodologies



Types of data analysis

Modelling Methodologies

KSCHOOL

Descriptive

¿What did happen?

Diagnosis

¿Why does it happen?

Predictive

¿What will happen?

Prescriptive

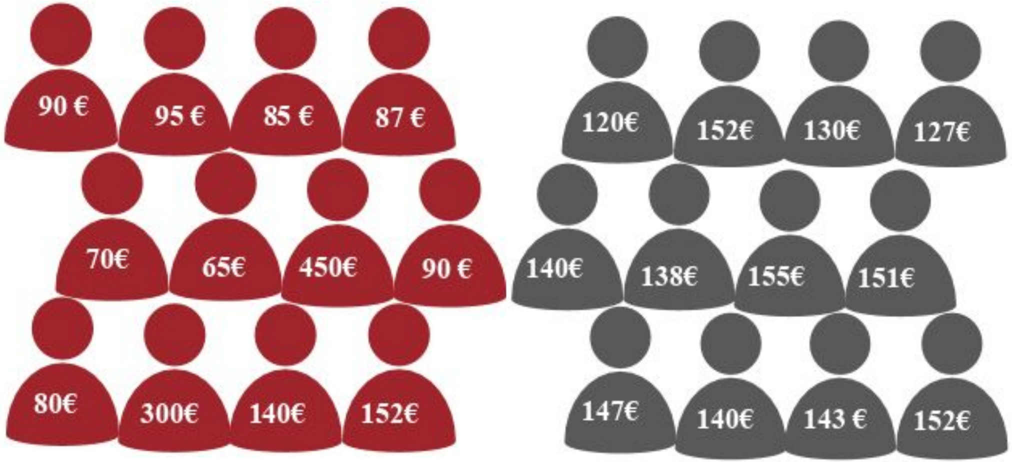
¿What do I need to do?

KSCHOOL

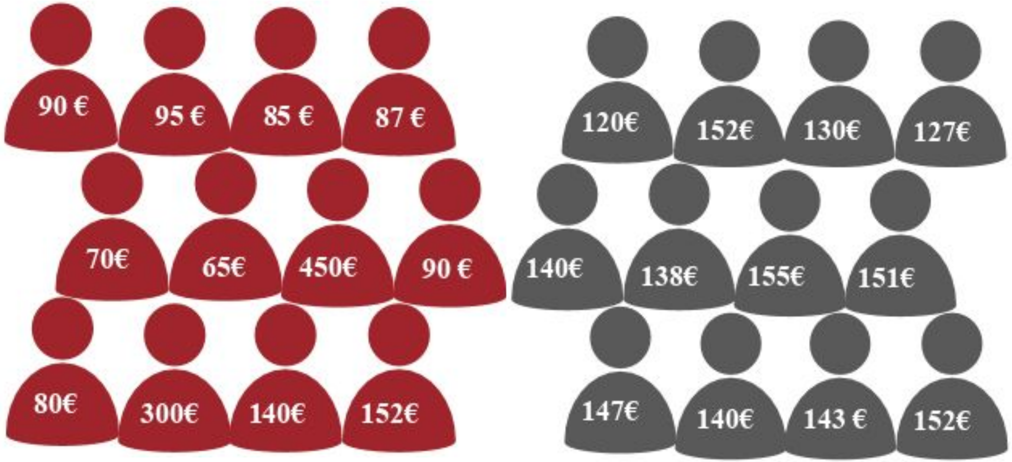
Some questionable conclusions

Be careful with the arithmetical mean in Descriptive Analysis

¿Which population is better?



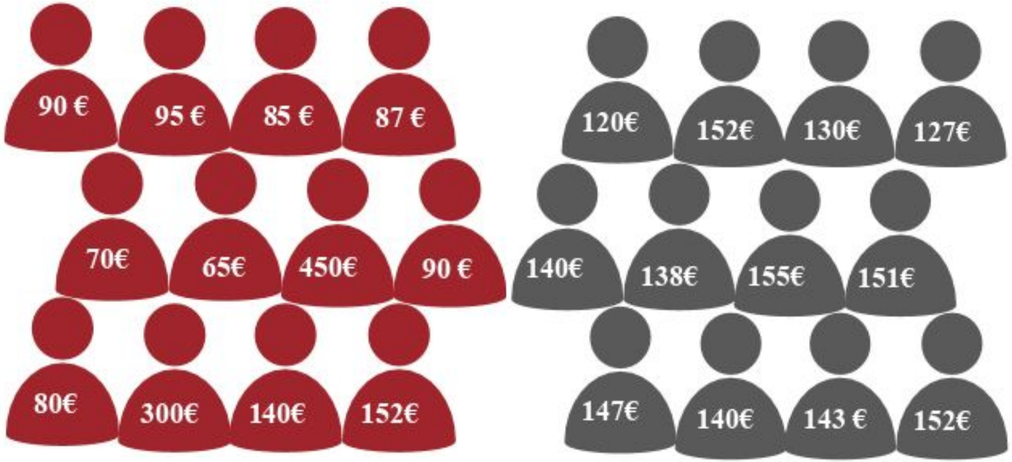
¿Which population is better?



Average = 142 €

Average = 141 €

¿Which population is better?



Average = 142 €	Average = 141 €
Median = 90 €	Median = 140 €

The importance of the statistical distribution of a variable

Modelling Methodologies

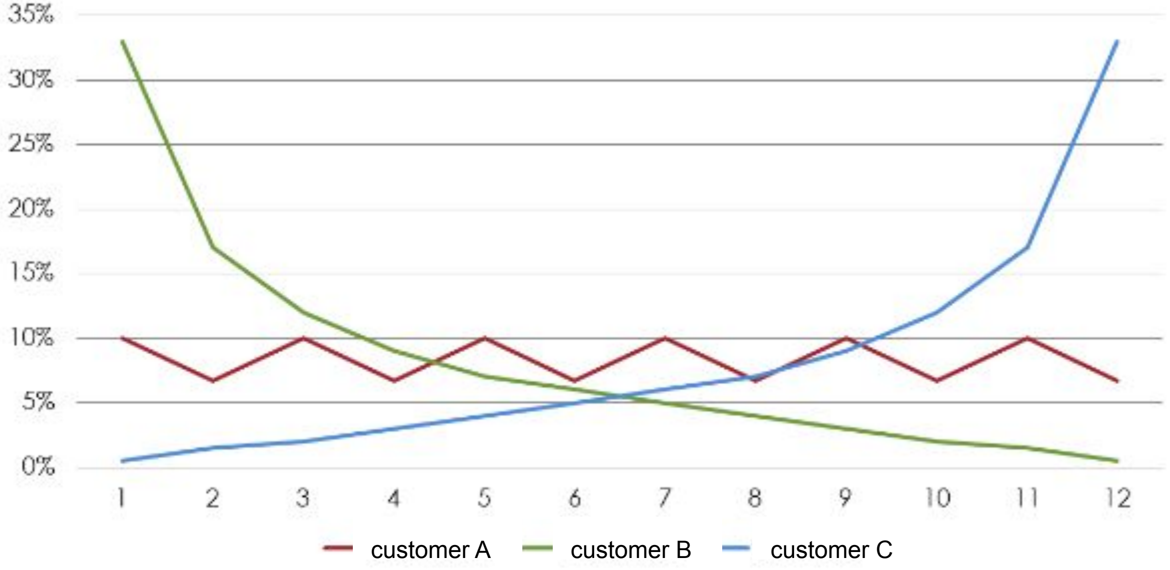
customer A customer B customer C

Spending $\mathbf{A} = \mathbf{1.000}$ €
Spending $\mathbf{B} = \mathbf{1.000}$ €
Spending $\mathbf{C} = \mathbf{1.000}$ €

Frequency $\mathbf{A} = \mathbf{12}$
Frequency $\mathbf{B} = \mathbf{12}$
Frequency $\mathbf{C} = \mathbf{12}$

Modelling Methodologies

Distribution of three customers' spending along 12 months



Spending A = 1.000 €
Spending B = 1.000 €
Spending C = 1.000 €

Frequency A = 12
Frequency B = 12
Frequency C = 12

KSCHOOL

Be careful with comparisons

Modelling Methodologies

Are trully comparable?

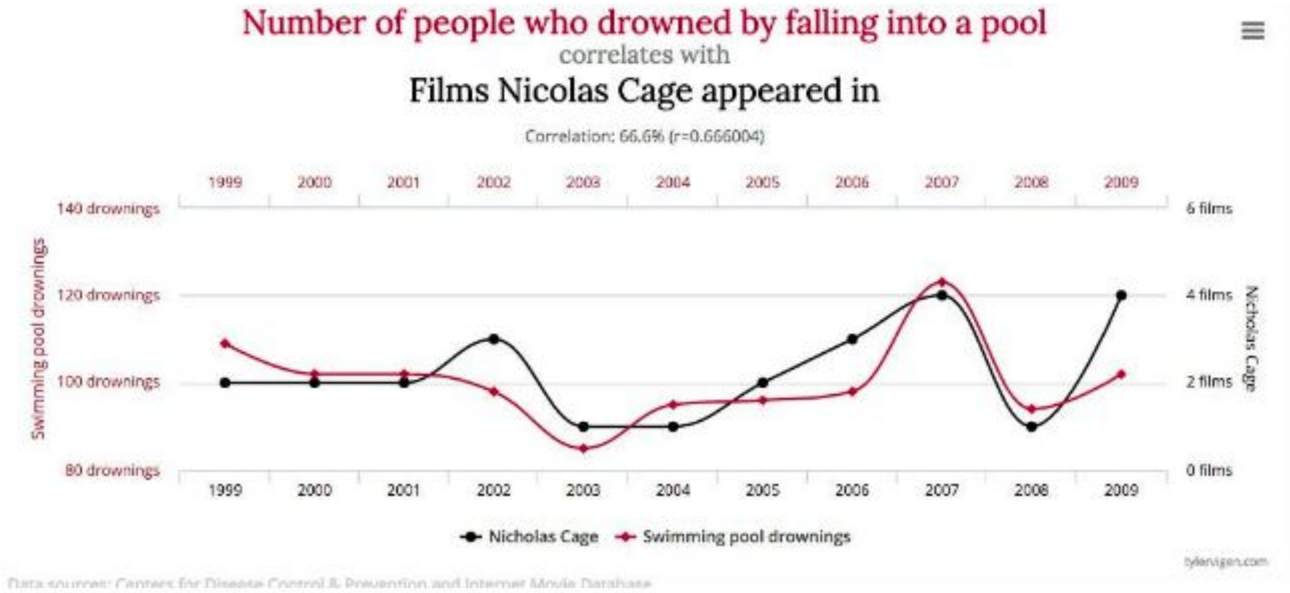
- Let’s imagine that one of them are women and the other one are men.
- Let’s imagine that one of them are customers and the other one are employes.

It is necessary the same starting point.
In another case, we will have the obvious, instead of learnings.

Relationship between variables does not imply causal relation

Modelling Methodologies

When Nicolas Cage acts, the number of people who drown by falling into a pool increase.



A mathematical model

Modelling Methodologies

KSCHOOL

A mathematical model is a description of a system using mathematical concepts and language. A model may help to explain a system and to study the effects of different components, and to make predictions about behaviour.

KSCHOOL

Example of a model:

Objetive

How could I measure the maximum heart rate?

KSCHOOL

Example of a model:

Data that I have

	Sexo	Edad	FCM
1	Mujer	37	185
2	Mujer	71	163
3	Hombre	25	210
4	Mujer	22	198
5	Hombre	48	179

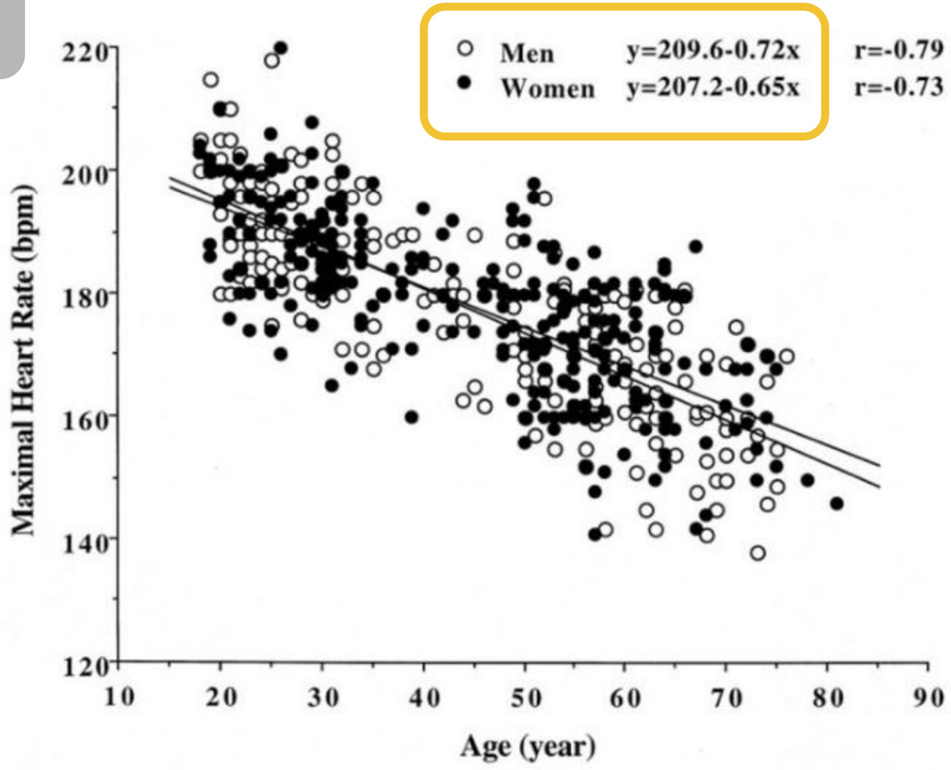
...

each line represents a person

Modelling Methodologies

Example of a model:

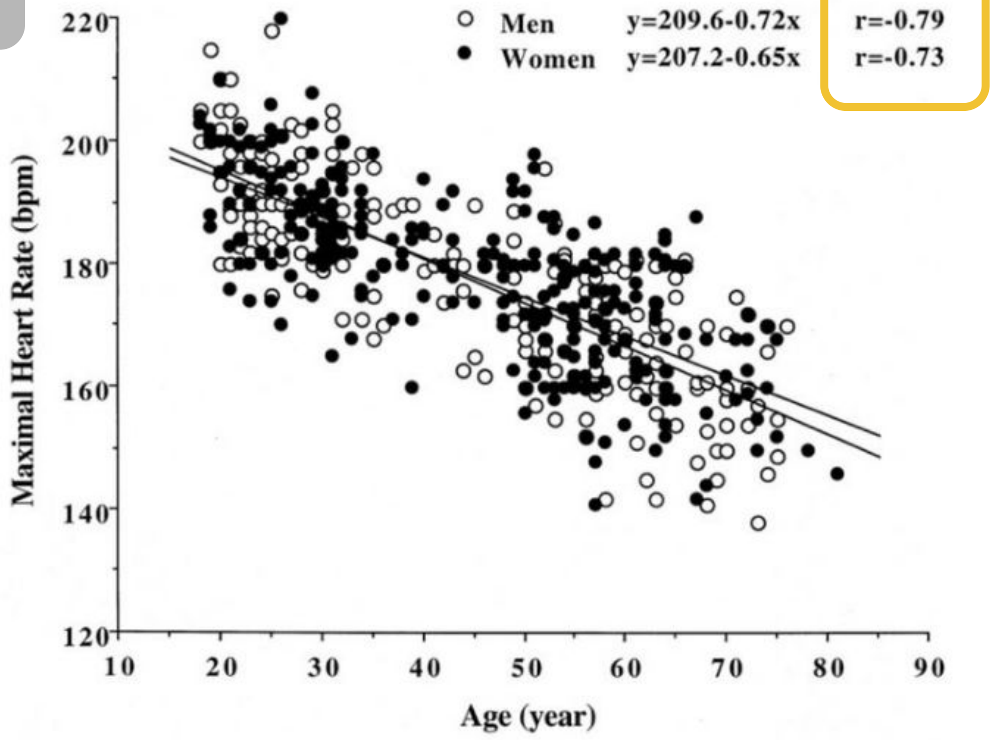
Results



Modelling Methodologies

Example of a model:

Results



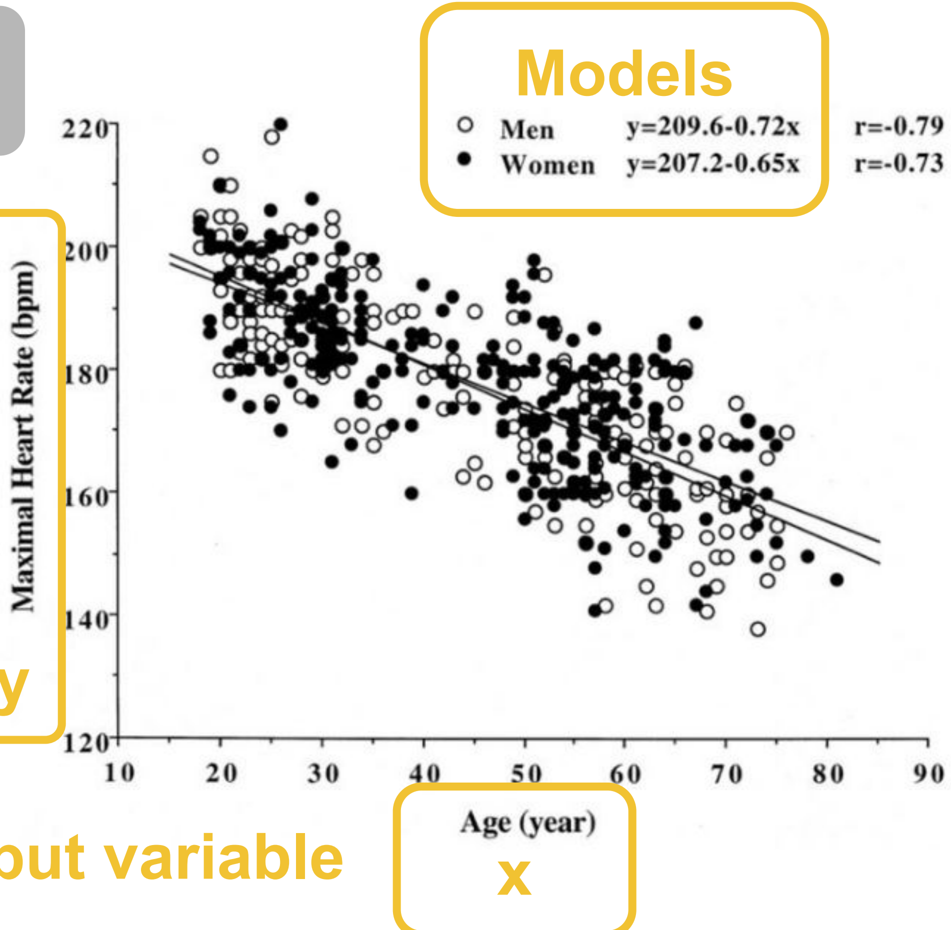
Modelling Methodologies

Example of a model:

Results

output variable

input variable



¿Machine Learning?

Modelling Methodologies

"Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with data, without being explicitly programmed."

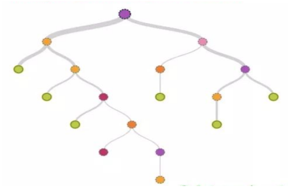
Wikipedia

Modelling Methodologies

General types of learnings

Supervised learning

Here is the data set where the right answers [labels] are given for each example. Please, produce more right answers.

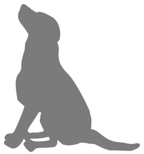


Unsupervised learning

Here is the unlabelled data. Please, find peculiarities, similarities or structures (e.g. clusters) in the data yourself.

Reinforcement learning

Learn to do something yourself purely by maximising your expected reward.



Modelling Methodologies

KSCHOOL

General types of learnings

Supervised learning

Here is the data set where the right answers [labels] are given for each example. Please, produce more right answers.

Unsupervised learning

Here is the unlabelled data. Please, find peculiarities, similarities or structures (e.g. clusters) in the data yourself.

Reinforcement learning

Learn to do something yourself purely by maximising your expected reward.

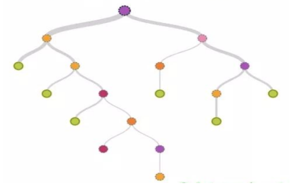
General types of problems

Regression

Predict a <u>continous</u> valued output.

Classification

Predict a <u>discrete</u> valued output (eg. a label or a class).



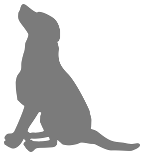
Clustering

Group similar examples into subsets called clusters.

Dimensionality reduction

Maybe you don't need all the data. What is the essence of the data?

This can be a goal on its own but is often used as a pre-processing step for other ML tasks



Modelling Methodologies

KSCHOOL

Supervised Learning

Learning from past elements

Regression

Predict a number:
eg PRICE

Classification

Predict a label:
eg PURCHASE / NON-PURCHASE

Modelling Methodologies

Unsupervised Learning

Learning by comparison

Clustering

Finding elements alike:
eg CUSTOMER SEGMENTATION according to their habits

Dimensionality Reduction

Explaining elements with less attributes

Example: Predicting House Prices

Problem Statement

We would like to predict the price of a house according to its characteristics.

Example: Predicting House Prices

Machine Learning requires explaining reality as a table of <u>features</u>.

Data

10010001110101010011010101010101010100101011101101
01010001001111101010101010101010010101001101010101
00100110101010101001010101100101100110111110101000
01010101010101010101010101010101010101100110010001
1101010100110101010101010101010010101110110101010001
0011111010101010101010100101010011010101010100110
10101010100101010110010110011011111010100001010101
01010101010101010101010101010110011010101010010101
01100101100110111110101000010101010101010101010101
01010101010101100110010001110101010011010101010101
01010010101110110101010001001111101010101010101010
01010100110101010100100110101010101001010101100101

Id	LotArea	GrLivArea	GarageArea	PoolArea	SalePrice
1	8450	196.0	1710	548	208500
2	9600	0.0	1262	460	181500
3	11250	162.0	1786	608	223500
4	9550	0.0	1717	642	140000
5	14260	350.0	2198	836	250000

KSCHOOL

Example: Predicting House Prices

Features Description

each line represents a house

Id	LotArea	GrLivArea	GarageArea	PoolArea	SalePrice
1	845	196.0	171	54	208500
2	960	0.0	126	46	181500
3	1125	162.0	178	60	223500
4	955	0.0	171	64	140000
5	1426	350.0	219	83	250000

target

KSCHOOL

Example: Predicting House Prices

Machine Learning uses <u>models</u> to explain relationship between features.

In this case, we look for a function that explains the SalePrice:

SalePrice = f(Features)

Example: Predicting House Prices

Linear Regression

A linear regression tries to find a <u>linear function</u>.

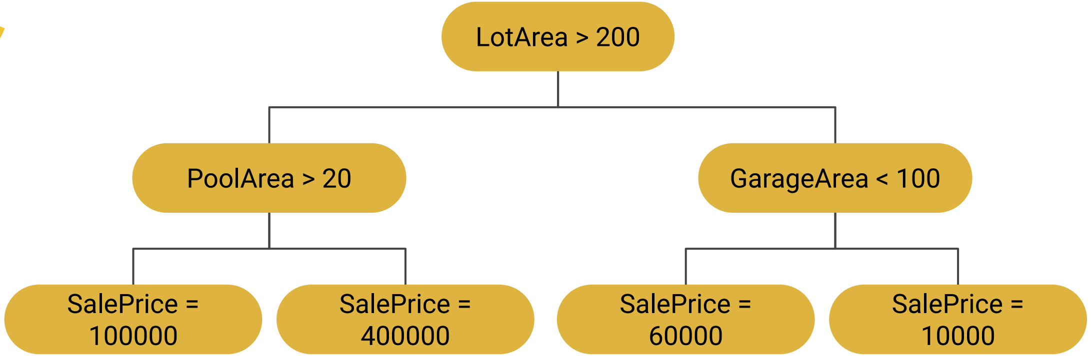
SalePrice = a * LotArea + b * GrLivArea + c * GarageArea + d * PoolArea

We need to find the best a,b,c,d to have the best relationship.

Example: Predicting House Prices

Decision Tree

A decision tree tries to find a path to explain the target:



We need to find the best splits for our predictions.

Example: Predicting House Prices

The Data Scientist works to find the best parameters for the model.

In this case:

- Linear Regression: a,b,c and d.
- Decision Tree: number of splits....

How do you find the best parameters?

Modelling Methodologies

KSCHOOL

Example: Predicting House Prices

There are several metrics that can be chosen for this task. For example:

- Bias: Average of the errors $MBE = \frac{1}{N}\sum_{i=1}^{N}(x_{f,i} - x_{o,i})$
- MAE: Average of the absolute values of errors $MAE = \frac{1}{N}\sum_{i=1}^{N}|x_{f,i} - x_{o,i}|$
- RMSE: Square root of Average of the square of errors $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{f,i} - x_{o,i})^2}$

We chose the model parameters that provide the best METRIC.

In this case, let's use MAE:

$$\frac{|RealPrice1 - PredictedPrice1| + |RealPrice2 - PredictedPrice2| + \cdots}{Number\ of\ houses}$$

Example: Predicting House Prices

Once we have the best parameters for a type of model, we can rank the models with the best metric they had

In this case: MAE:
- Linear Regression: 15
- Decision Tree: 10

And we ultimately choose the model with the best metric.

Example: Predicting chewing gum buyers

Problem Statement

We would like to predict the customers who buy chewing gum.

Example: Predicting chewing gum buyers

Features Description

each line represents a customer

Id	Sex	Age	Smoker	Sweets	GumBuyer
1	Woman	19	Yes	No	Yes
2	Woman	45	No	No	No
3	Man	53	Yes	No	Yes
4	Woman	21	No	Yes	Yes
5	Man	35	No	Yes	No

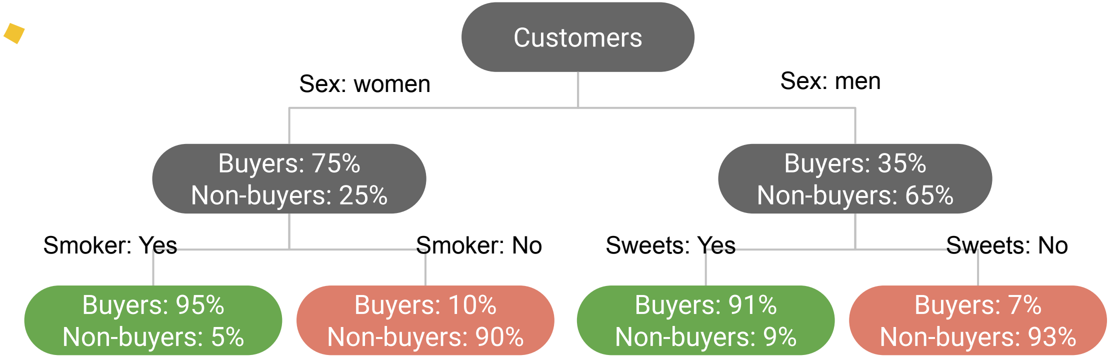
target

KSCHOOL

Example: Predicting chewing gum buyers

Decision Tree

A decision tree tries to find a path to explain the target:



We need to find the best parameters for the model: number of splits, minimum number of member per group...

Classification metrics: Confusion matrix

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)

KSCHOOL

Classification metrics: Confusion matrix

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

Example: Salmon and sea bass

Problem Statement

I would like to discern between salmon and sea bass.

Example: Salmon and sea bass

A fish-packing plant wants to automate the process of sorting incoming fish according to species.
As a pilot project, it is decided to try to separate sea bass from salmon using ML.

We know they are different but we don't know what features make them different.

Let's go.

Statistical Pattern Recognition. Ricardo Gutierrez-Osuna. Wright Sate University.

KSCHOOL

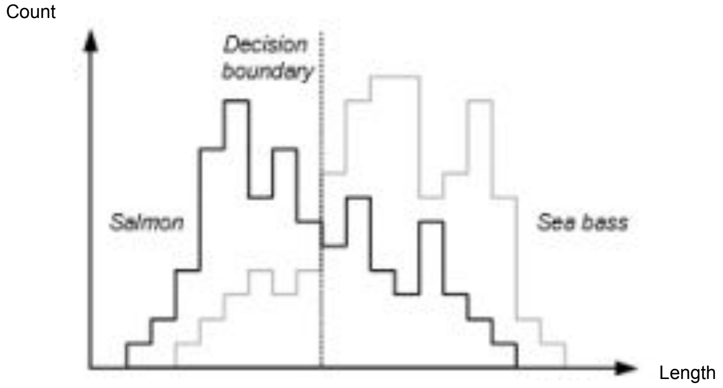
Example: Salmon and sea bass

Some features to explore:

- Length
- Scale density
- Width
- Position of the mouth
- ...

Example: Salmon and sea bass

Start by checking length of each fish:



They have different distributions. Notice that salmon tends to be shorter than sea bass.

Example: Salmon and sea bass

Find the best length L threshold:

fish length < L → classify as a salmon

fish length > L → classify as a sea bass

After searching through all possible thresholds L, the best L= 9, and still 30% of fish is misclassified.

We cannot reliably separate sea bass from salmon by length alone!

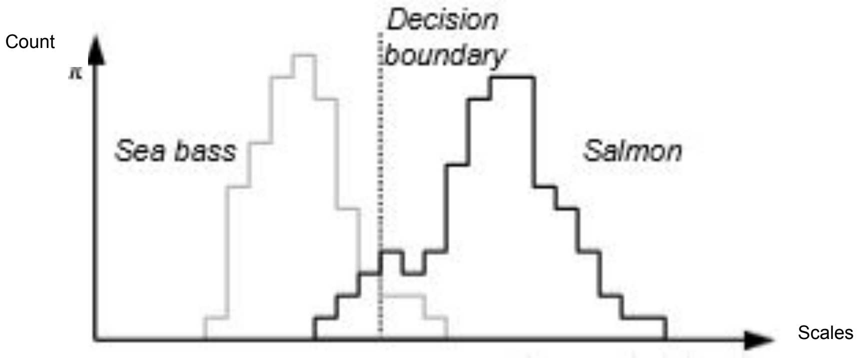
Example: Salmon and sea bass

I have to continue looking for...

Example: Salmon and sea bass

Next feature I can consider:

Scale density

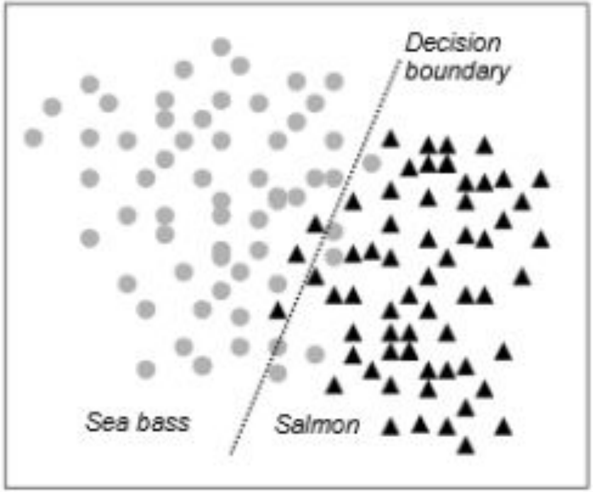


The two classes are much better separated!

Example: Salmon and sea bass

Lets to combine the two features:

length



scales

We have reduced the error rate below 5%.

Should we be satisfied with this result?

Example: Salmon and sea bass

And… what could we do now?

KSCHOOL

Example: Salmon and sea bass

Options we have:

- Consider additional features:
  - Which ones?
  - Some features may be redundant (e.g., if eye color perfectly correlates with width, then we gain no information by adding eye color as feature.)
  - It may be costly to attain more features
  - Too many features may hurt the performance
- Use a more complex model

KSCHOOL

Example: Salmon and sea bass

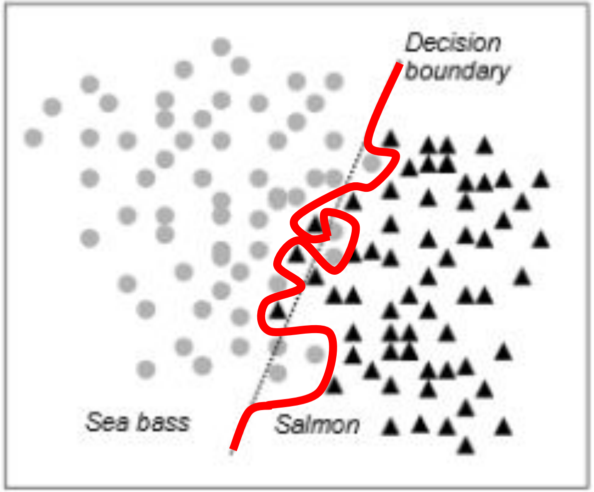
Options we have:

- Consider additional features:
  - – Which ones?
  - – Some features may be redundant (e.g., if eye color perfectly correlates with width, then we gain no information by adding eye color as feature.)
  - – It may be costly to attain more features
  - – Too many features may hurt the performance
- Use a more complex model

Example: Salmon and sea bass

With a more complex model all data are perfectly separated :

length



scales

Should we be satisfied now??

Example: Salmon and sea bass

We must consider:

Which decisions will the classifier take on novel patterns, i.e. fishes not yet seen?
Will the classifier suggest the correct actions?

This is the issue of GENERALIZATION

KSCHOOL

Measurement of the model

We take the available features table and we split in 2 sets:

1 set to calculate parameters: Training Set

1 set to calculate metrics: Test Set

	Length	DensityScale	Width	...
1	10000	37	1385	...
2	12350	71	1613	...
3	9871	25	2100	...
4	15431	22	1985	...
5	5200	48	1791	...
..	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Summary: ML is about features, models and metrics

- ML requires explaining reality as a table of features
- We then use models to explain relationships between features for the given ML task
- Each model is evaluated according to a metric

Modelling Methodologies

KSCHOOL

Supervised Learning
some of the main techniques

- Linear / Logistic Regresion
- Decision Tree
- Random Forest
- XGBoost
- K-neighbors
- Neural Networks
- Support Vector Machines - SVM
- ...

And many others... The list is growing thanks to the active research.

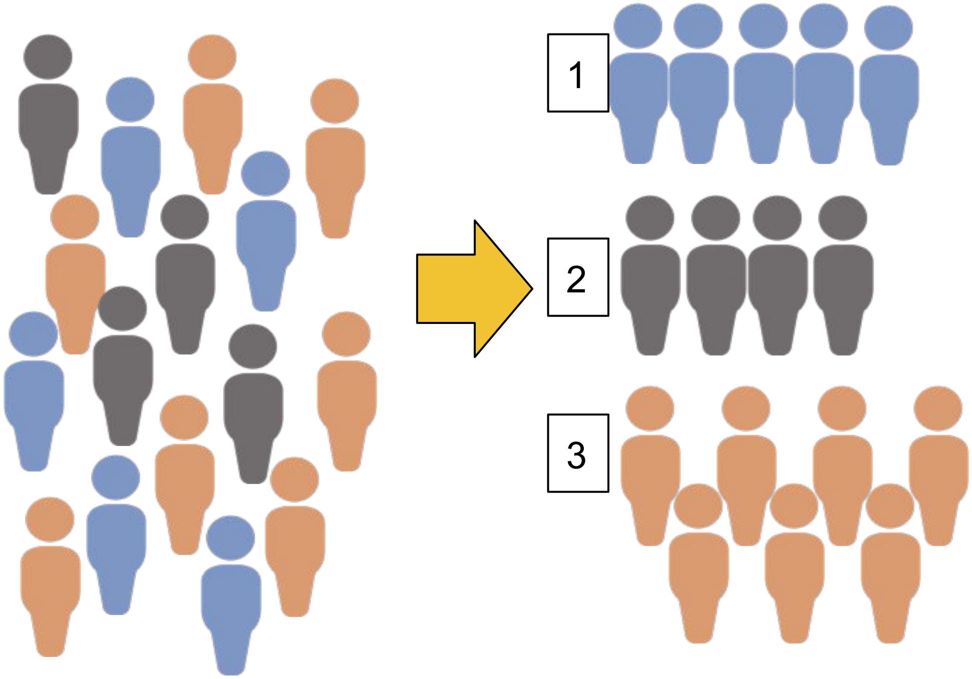
Hands on Excercises

Let’s try to do next two exercises:

1. Exercise_1: Compare models’ predictions and decide which is the best one with different metrics.
2. Exercise_2: Build the confusion matrix and calculate main metrics about it.
   a. Which is the best model in general?
   b. Which identify better the 1 class?

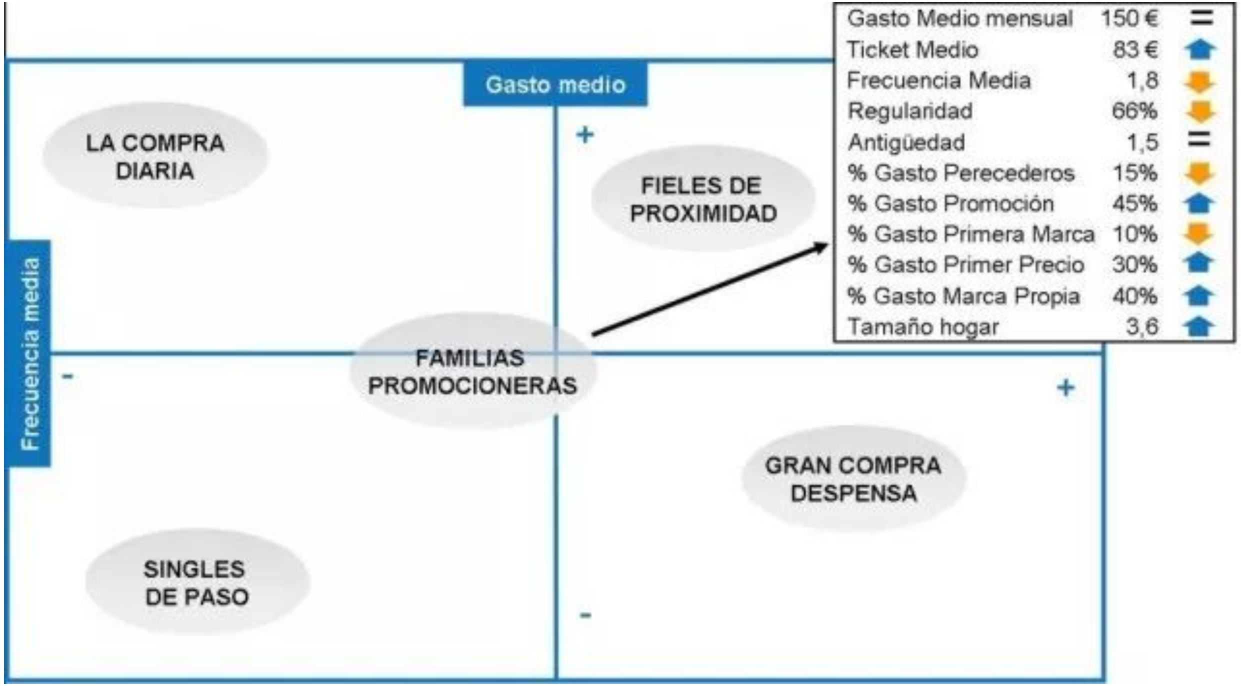
Clustering

Unsupervised Learning
the main type of problems



Modelling Methodologies

Example of clustering: department store



Picture from https://www.unica360.com/segmentacion-de-clientes-una-propuesta-de-clasificacion-i.

Modelling Methodologies

Machine
Learning
other concepts

Artificial Intelligence

AI may use Machine Learning to explain how humans think.

Deep Learning

Deep Learning is a type of ML models

Statistics

Statistics are a field of Mathematics. Some aspects of Statistics have been renamed as ML.

ML Canvas

KSCHOOL

Translating a Business Problem to ML

Any Business Problem can be analyzed through the following axes:

1. Business Value and Measure of Success.
2. Data Sources
3. ML task + metric
4. Features Extraction + Model Creation
5. Industrialization and Maintenance

Modelling Methodologies

KSCHOOL

Opportunity:

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

Machine Learning

ML task
Methods of evaluation

Data

Sources
Frequency
Legal
History

Acceptable Quality
Main Features

Ranking of Models

List of the models and their metric

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost

By:

Iteration:

Date:

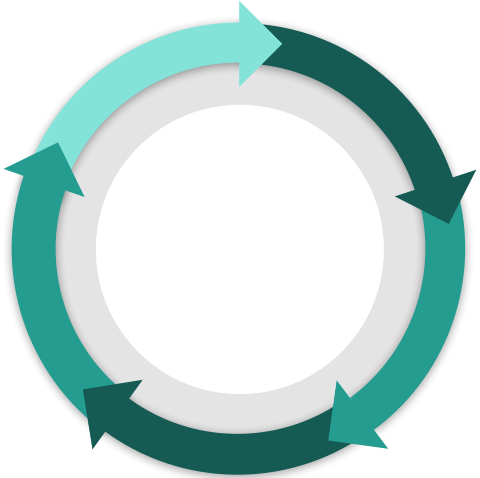
Data Science Process

Once the first version of the canvas is created, Data Scientist perform the following actions:

1. Extraction of features
2. Select of ML task and evaluation
3. Model Creation
4. Quantity of Data > Parameters
5. Loop until satisfied

KSCHOOL

The whole process is iterative



Product Managers and Data Scientists can revise their initial plans based on discoveries made during the different steps:

- Insufficient Data
- Metric not good enough for the task
- ...

Modelling Methodologies

KSCHOOL

Opportunity: House Price Estimation

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

Business Description:
Negotiate better house prices

Resulting Action:
Daily report of the new houses with offered price and estimated price.

Measure of Success (KPI):
Money saved

ML Initiative Cost:
1 month study

Machine Learning

ML task
Methods of evaluation

Regression
Method of evaluation: MAE

Data

Sources
Frequency
Legal
History

Data Sources:
Idealista, Kaggle Data, ...

Frequency:
Weekly

Legal:
Public Data

History:
1 year

Acceptable Quality
Main Features

Acceptable Quality:
Yes

Main Features:
Area of house, Bathrooms, ...

Ranking of Models

List of the models and their metric

Decision Tree
MAE:

Linear Regression:
MAE:

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost

Weekly job that retrains and produces new model.

Every morning, we run the predictions against a set of newly available houses.

Yearly evaluation to check changes

By:

Iteration:

Date:

Modelling Methodologies

Hands on Excercises

Let’s try to do that exercise for those 4 use cases:

1. Weather Forecast
2. Movie Recommendation
3. Email Spam Filters
4. Pollution Prediction

Some actual examples of Machine Learning

Modelling Methodologies

airbnb Enlace

UBER Enlace

NETFLIX Enlace

tinder Enlace

IBM SlamTracker Enlace

Spotify Enlace

KSCHOOL

Data science competitions:

synergic partners™

A Telefonica COMPANY

NETFLIX

BBVA

kaggle

KSCHOOL

“Cuando eres capaz de ver lo sutil, es fácil ganar”

Sun Tzu

Thank you!

irenetorresvalle@gmail.com