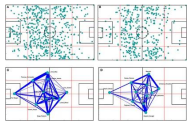
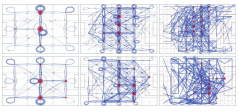
# Machine Learning for Sports

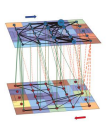
## Orange3

Escuela Universitaria Real Madrid
UNIVERSIDAD EUROPEA

JOHANN H. MARTÍNEZ

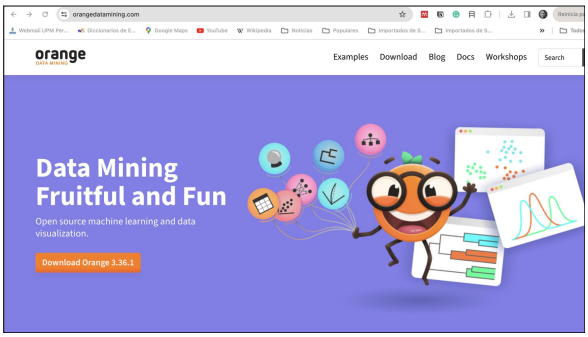




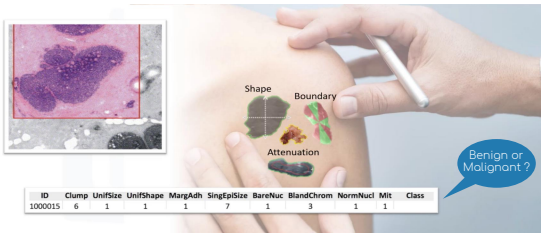



# Data, dataset and much more

GitHub link



## Skin Cancer




| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000015 | 6 | 1 | 1 | 1 | 7 | 1 | 3 | 1 | 1 | |

| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | malignant |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1017122 | 8 | 10 | 10 | 8 | 7 | 10 | | 7 | 1 | malignant |
| 1018099 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | benign |
| 1018561 | 2 | 1 | 2 | H | 2 | 1 | 3 | 1 | 1 | benign |
| 1033078 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | benign |
| 1033078 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | benign |

Modeling

| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000015 | 6 | 1 | 1 | 1 | 7 | 1 | 3 | 1 | 1 | Benign |

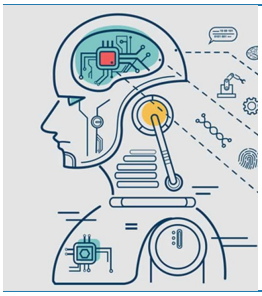
Prediction

Accuracy = 89%

## Examples


*"Machine Learning is the subfield of computer science that gives computers the ability to learn without being explicit programmed"*

Arthur Samuel, coin the term MachineLearning in 1959 at IBM

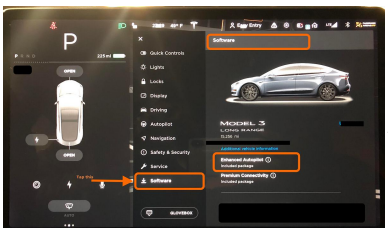


## Differences between

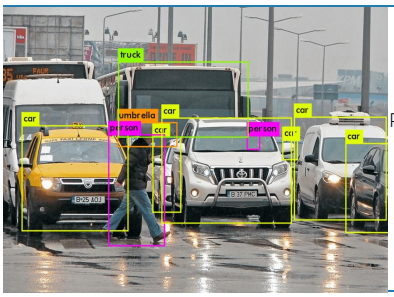
- Artificial Intelligence
- Machine Learning
- Deep Learning

## Artificial Inteligence(AI)

AI tries to make computers intelligent so that they mimic th cognitive functions of humans.




Autopilot in Tesla

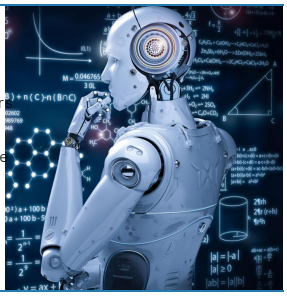



Pattern recognition

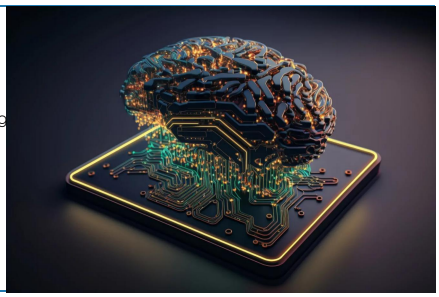
Using computer vision

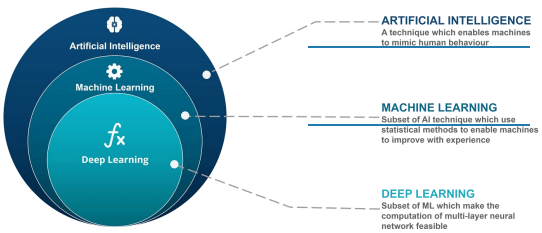

ML is the branch of AI that covers the **statistical** part of AI.

ML **teaches** the computers to solve problems by looking a **hundred** of examples.



Deep learning





# Classical Machine Learning

Task Driven

Data Driven

## Supervised Learning
( Pre Categorized Data )

- **Classification** ( Divide the socks by Color ) — Eg. Identity Fraud Detection
- **Regression** ( Divide the Ties by Length ) — Eg. Market Forecasting

## Unsupervised Learning
( Unlabelled Data )

- **Clustering** ( Divide by Similiarity ) — Eg. Targeted Marketing
- **Association** ( Identify Sequences ) — Eg. Customer Recommendation
- **Dimensionality Reduction** ( Wider Dependencies ) — Eg. Big Data Visualization

Obj: Predications & Predictive Models | Pattern/ Structure Recognition

| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | malignant |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1017122 | 8 | 10 | 10 | 8 | 7 | 10 | | 7 | 1 | malignant |
| 1018099 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | benign |
| 1018561 | 2 | 1 | 2 | H | 2 | 1 | 3 | 1 | 1 | benign |
| 1033078 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | benign |
| 1033078 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | benign |

## Supervised model

We **teach** the model by training it with some data from **labeled dataset**

Features

Class o label

Observation

Categorica

| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | malignant |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1017122 | 8 | 10 | 10 | 8 | 7 | 10 | | 7 | 1 | malignant |
| 1018099 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | benign |
| 1018561 | 2 | 1 | 2 | H | 2 | 1 | 3 | 1 | 1 | benign |
| 1033078 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | benign |
| 1033078 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | benign |

Numerical

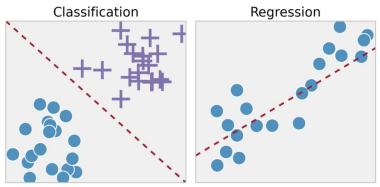
Labeled dataset

## I am going to teach you

1 —> 2
2 —> 4
5 —> 10
6 —> 12

10 —> ?

## Types of supervised techniques

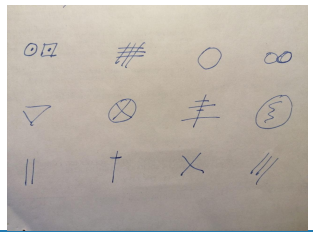



| Customer Id | Age | Edu | Years Employed | Income | Card Debt | Other Debt | Address | DebtIncomeRatio |
|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 2 | 6 | 19 | 0.124 | 1.073 | NBA001 | 6.3 |
| 2 | 47 | 1 | 26 | 100 | 4.582 | 8.218 | NBA021 | 12.8 |
| 3 | 33 | 2 | 10 | 57 | 6.111 | 5.802 | NBA013 | 20.9 |
| 4 | 29 | 2 | 4 | 19 | 0.681 | 0.516 | NBA009 | 6.3 |
| 5 | 47 | 1 | 31 | 253 | 9.308 | 8.908 | NBA008 | 7.2 |
| 6 | 40 | 1 | 23 | 81 | 0.998 | 7.831 | NBA016 | 10.9 |
| 7 | 38 | 2 | 4 | 56 | 0.442 | 0.454 | NBA013 | 1.6 |
| 8 | 42 | 3 | 0 | 64 | 0.279 | 3.945 | NBA009 | 6.6 |
| 9 | 26 | 1 | 5 | 18 | 0.575 | 2.215 | NBA006 | 15.5 |
| 10 | 47 | 3 | 23 | 115 | 0.653 | 3.947 | NBA011 | 4 |
| 11 | 44 | 3 | 8 | 88 | 0.285 | 5.083 | NBA010 | 6.1 |
| 12 | 34 | 2 | 9 | 40 | 0.374 | 0.266 | NBA003 | 1.6 |

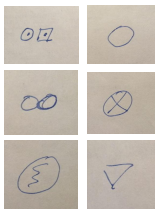
## Unsupervised model

The model **works on its own** to discover information may be not visible to the human eye.

## We also learn from unsupervised dataset

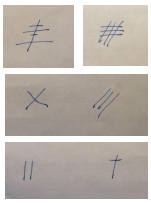


Polygonal language?



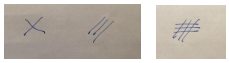
Two clusters

Linear language?



Fulled polygonal language?

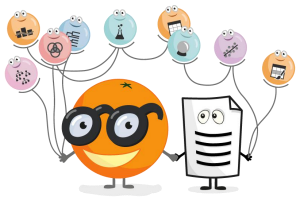
Diagonal Linear language?



Four clusters?

Empty polygonal language?

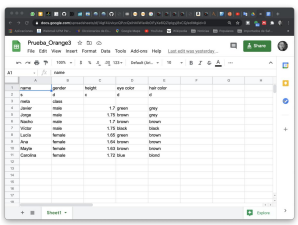
Straight Linear language?



ESCUELA DEL REAL MADRID

## First program

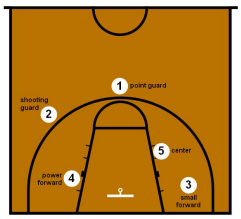
- File —> Iris
- Data Table
- Scatter plot
- Selected data table from scatter plot!



ESCUELA DEL REAL MADRID

## Own data & how to transform it to be read

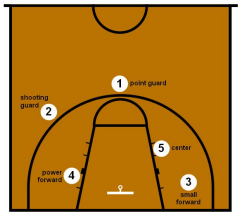
- Construir una base de datos en Google sheet.
- Importar los datos
- Adecuar los datos
  - s: string
  - d: discrete
  - c: continuous
  - meta: meta data
  - Class: class
- Save data




ESCUELA DEL REAL MADRID

## First visualizations

- Load innerNBA Players dataset
- Scatter plot
- Box plot

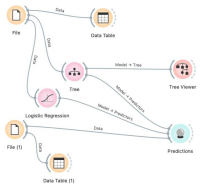



ESCUELA DEL REAL MADRID

## First visualizations

- DownloadNBA Players dataset
- From xls to xlsx to be well-read
- widget Transform to select Target
- Visualize it

**Bonus** Play with Titanic's survivals




ESCUELA DEL REAL MADRID

## First Predictions

- Download NBA Players dataset
- Tree —> Tree viewer
- Test file
- Prediction
- Logistic Regression

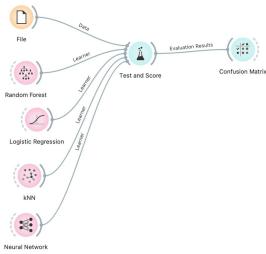
| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | malignant |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1017122 | 8 | 10 | 10 | 8 | 7 | 10 | | 7 | 1 | malignant |
| 1018099 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | benign |
| 1018561 | 2 | 1 | 2 | H | 2 | 1 | 3 | 1 | 1 | benign |
| 1033078 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | benign |
| 1033078 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | benign |

ESCUELA DEL REAL MADRID

## Classification

The goal is to predict a *class label*

- Binary Classification
- Multiclass Classification




ESCUELA DEL REAL MADRID

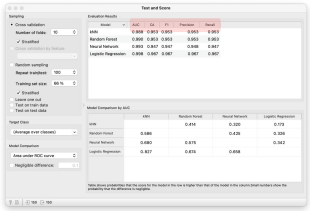
## Scoring

- K-cross validation
- Test & Score
- Confusion Matrix


ESCUELA DEL REAL MADRID

## Scoring

- K-cross validation




## Test and Score

- We are talking about some metrics:
  - CA. Classification Accuracy
  - Confusion matrix
    - To see where exactly was the classification wrong.
  - Jaccard index

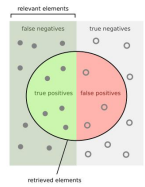
| Matriz de confusión | | Estimado por el modelo | |
|---|---|---|---|
| | | *Negativo (N)* | *Positivo (P)* |
| Real | *Negativo* | a: (TN) | b: (FP) |
| | *Positivo* | c: (FN) | d: (TP) |

| | |
|---|---|
| **Precisión (*"precision"*)** Porcentaje predicciones positivas correctas: | d/(b+d) |
| **Sensibilidad, exhaustividad (*"Recall"*)** Porcentaje casos positivos detectados | d/(d+c) |
| **Especificidad (*Specifity*)** Porcentaje casos negativos detectados | a/(a+b) |
| **Exactitud ("accuracy")** Porcentaje de predicciones correctas *(No sirve en datasets poco equilibrados)* | (a+d)/(a+b+c+d) |

Main diagonal: Correct estimated values (TP + TN)
Secondary diagonals:
- Upper means False Positives (FP)
- Bottom means False Negatives (FN)

## Confusion Matrix

TP = True Positives
FN = False Negatives
FP = False Positives
TN = True Negatives




$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$CA = \frac{TP+TN}{Total}$$

$$F1\text{-}score \in [0,1] = 2 * \frac{Precision * Recall}{Precision + Recall}$$

It thus symmetrically represents both precision and recall in one metric

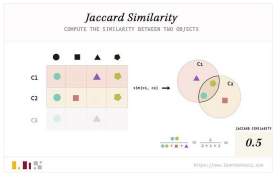
## Jaccard index
(The one of similarity)

$\mathbf{y}$ = True values (actual values)

$\hat{\mathbf{y}}$ =Predicted values of our model

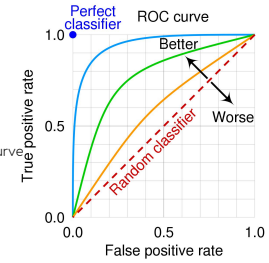
Jaccard index [0,1]

$$J(\hat{y}, y) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$




## AUC

Receiver Operating Characteristic Curve
(ROC) curve




## Matthews correlation coefficient (MCC)

MCC is a best single-value classification metric which helps to summarize the confusion matrix or an error matrix.

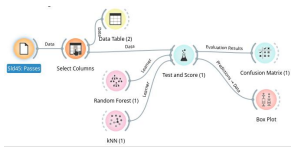
A confusion matrix has four entities:

**A**

| Actual \ Predicted | Control | Disease |
|---|---|---|
| Control | TN | FP |
| Disease | FN | TP |

**B**

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

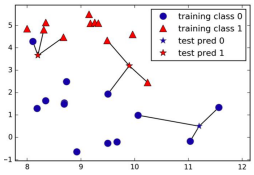



ESCUELA DEL REAL MADRID

## Scoring

- K-cross validation
- Test & Score
- Confusion Matrix

Classifiers

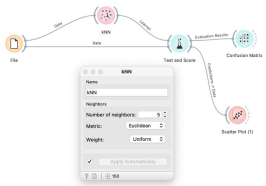

ESCUELA DEL REAL MADRID

## K-Nearest Neighbors (K-NN)

### Supervised

- A k-NN algorithm consists only of storing the training dataset.
- To make a prediction for a new data point, the algorithm finds the **closest data points** in the training dataset—its "nearest neighbors".

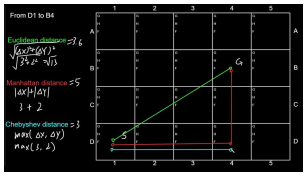

A lot of acronyms …




ESCUELA DEL REAL MADRID

## K-nearest Neighbors

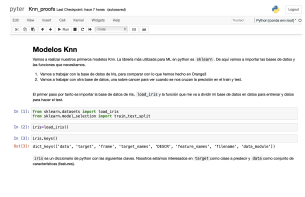
- Iris and Pases
- KNN —> Test Score —> Confusion Matrix
- Different values of K and distances.
- Scatter Plot (KNN) + Box plot




ESCUELA DEL REAL MADRID

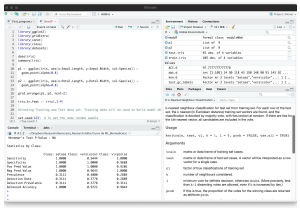
## K-nearest Neighbors

- Iris and Pases
- KNN —> Test Score —> Confusion Matrix
- Different values of K and distances.
- Scatter Plot (KNN) + Box plot



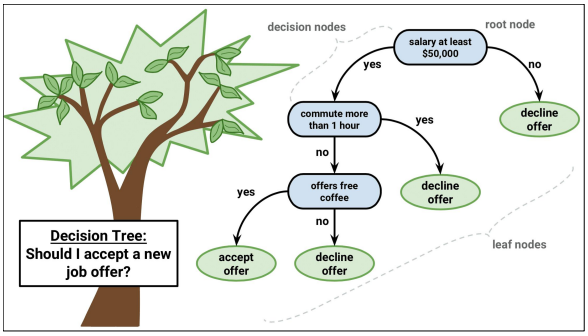

## K-nearest Neighbors in Python

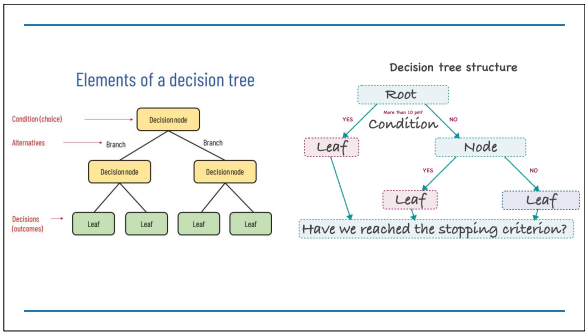
- Iris
- KNN —> Test Score —> Confusion Matrix



## K-nearest Neighbors in R

- Iris
- KNN —> Test Score —> Confusion Matrix




| Patient ID | Age | Sex | BP | Cholesterol | Drug |
|---|---|---|---|---|---|
| p1 | Young | F | High | Normal | Drug A |
| p2 | Young | F | High | High | Drug A |
| p3 | Middle-age | F | High | Normal | Drug B |
| p4 | Senior | F | Normal | Normal | Drug B |
| p5 | Senior | M | Low | Normal | Drug B |
| p6 | Senior | M | Low | High | Drug A |
| p7 | Middle-age | M | Low | High | Drug B |
| p8 | Young | F | Normal | Normal | Drug A |
| p9 | Young | M | Low | Normal | Drug B |
| p10 | Senior | M | Normal | Normal | Drug B |
| p11 | Young | M | Normal | High | Drug B |
| p12 | Middle-age | F | Normal | High | Drug B |
| p13 | Middle-age | M | High | Normal | Drug B |
| p14 | Senior | F | Normal | High | Drug A |
| p15 | Middle-age | F | Low | Normal | ? |

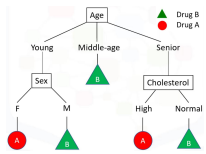
ESCUELA DEL REAL MADRID

## Decision Tree

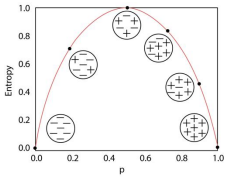
- What exactly is a decision tree?
- We'r doctors and we want to predict the best drug for a kind of patients

## How to build a decision tree?

1. Choose an atribute (Age):
2. Split it into its categories (Y, M-A, S)
   - It seems that M-A arrive to the solution to drug B
   - ... but Y, and S?
   - Split the remaining categories againg based on the other atributes:
     - For Y, now split it into Sex
       - Oh, look! It seems we arrive to solution A (for Y-female), and B (for Y-male)
     - For S, split it into Cholesterol
       - Oh, look! It seems we arrive to solution A (for S-Chol-H), and B (for S-Chol-N)



## How to compute the significance?




In every step, we want to dismiss the **impurity** of the nodes.

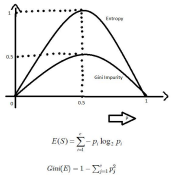
**Impurity** is computed by **Entropy** of data in the node

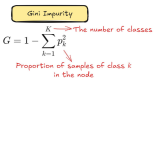
$$E = -\sum_i p_i \log_2(p_i)$$

**Entropy** is the amount of information disorder. If the samples are completely homogeneous the entropy is zero.

## How to compute the significance?

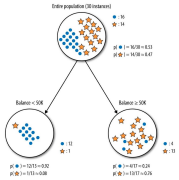
**Gini impurity** is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the branch.



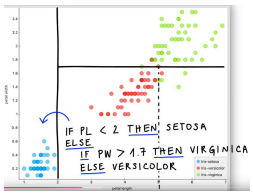



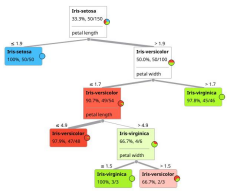

$$E(S) = \sum_{i=1}^{C} -p_i \log_2 p_i$$

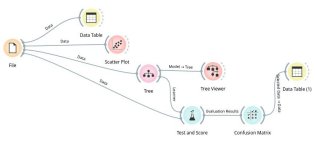
$$Gini(E) = 1 - \sum_{j=1}^{C} p_j^2$$




## Which attributes are the best?



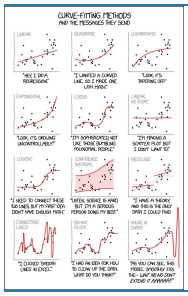




ESCUELA DEL REAL MADRID

## Decision Tree

- Iris dataset
- Test and Score —> confusion Matrix
- Tree —> Tree Viewer

# Linear Modelsfor Classification




Linear Modelsfor Classification

# Linear Modelsfor classification

**Regression**
What is the temperature going to be tomorrow?

PREDICTION
84°

Fahrenheit °F

**Classification**
Will it be Cold or Hot tomorrow?

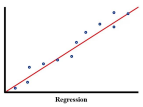
PREDICTION
COLD
HOT

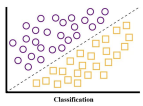
Fahrenheit °F

- They are a class of models that is **widely used** in practice.
- They make a prediction **using a linear function** of the input features

# Linear Modelsfor classification

| Regression | Classification |
|---|---|
| The task of predicting a continuous quantity is known as regression. | The classification process involves anticipating a discrete class label. |
| The task is to map the input value (x) to the continuous output variable (y). | The task is to map the input value(x) to the discrete output variable (y). |
| A regression model can predict a discrete value, but only in the form of an integer quantity. | A classification model can predict a continuous value, but it is in the form of a probability for a class label. |
| The output variable in regression must be of continuous value. | The output variable in Classification must be discrete. |



## Linear Binary classification

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent Variable · Population Y intercept · Population Slope Coefficient · Independent Variable · Random Error term

Linear component · Random Error component

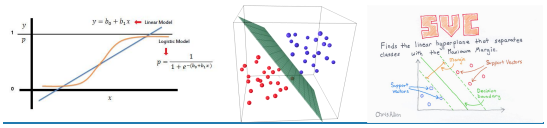
$X_i$ [0,..i] Features (columns)
$\beta$ [0,...i] Parameters of the model
$Y_i$ Prediction of the model

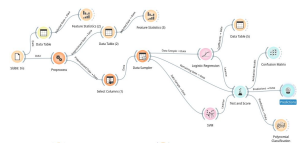
## Linear Models for classification

A linear classifier separates classes using a

- **line**
- **plane**
- **hyperplane**

**High-dimensional** problems: Where the number of features $p$ is much larger than the number of observations $N$, often written $p >> N$

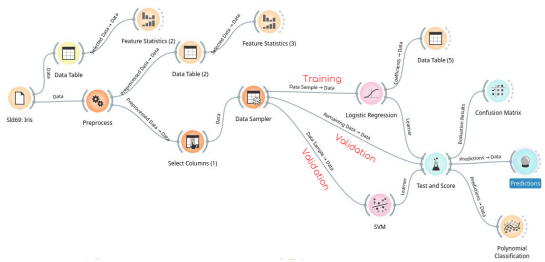


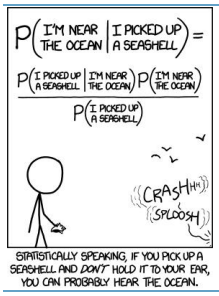



ESCUELA DEL REAL MADRID

## Linear Models

- Iris dataset
- LR -> Test and Score -> Confusion M.
- SVM -> Test and Score -> Confusion M.






# Naive Bayes Classifiers

**Bayes' theorem.**

Knowing the probability of **having a headache** given that one **has the flu** one could know *if one has any additional data*... the probability of **having the flu** if one **has a headache**

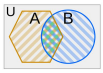
$P(A) =$ , $P(B|A) =$

$P(B) =$ , $P(A|B) =$

$P(A)\cdot P(B|A) =$ × =

$P(B)\cdot P(A|B) =$ × =

$= P(A)\cdot P(B|A)$ , i.e.

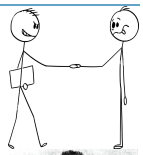



$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(B)}$$

$$P(B|A) = \frac{P(B)\cdot P(A|B)}{P(A)}$$

**Naive** means independency between features.

**Bayes** Based on Bayes' theorem.



**Naive Bayes classifier**

Assumes thatthe presence of a particular feature in a class is unrelated to the presence of any other feature

e.g., A fruit can be considered an apple if it is

- red
- round
- about 7 cm in diameter.

Even if these features depend on each other or upon the existence of the other features

BAYES THEOREM

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

BAE'S THEOREM

(1). Given a **dataset**

| | Outlook | Play |
|---|---|---|
| 0 | Rainy | Yes |
| 1 | Sunny | Yes |
| 2 | Overcast | Yes |
| 3 | Overcast | Yes |
| 4 | Sunny | No |
| 5 | Rainy | Yes |
| 6 | Sunny | Yes |
| 7 | Overcast | Yes |
| 8 | Rainy | No |
| 9 | Sunny | No |
| 10 | Sunny | Yes |
| 11 | Rainy | No |
| 12 | Overcast | Yes |
| 13 | Overcast | Yes |

(2) Convert it into **frequency** tables.

| Weather | No | Yes | |
|---|---|---|---|
| Overcast | 0 | 5 | 5 |
| Rainy | 2 | 2 | 4 |
| Sunny | 2 | 3 | 5 |
| All | 4 | 10 | 14 |

(3). Then, into a **probability** one.

| Weather | No | Yes | |
|---|---|---|---|
| Overcast | 0 | 5 | 5/14=0.35 |
| Rainy | 2 | 2 | 4/14=0.29 |
| Sunny | 2 | 3 | 5/14=0.35 |
| All | 4/14=0.29 | 10/14=0.71 | |

(3) Apply the **algorithm**

| Weather | No | Yes | |
|---|---|---|---|
| Overcast | 0 | 5 | 5/14=0.35 |
| Rainy | 2 | 2 | 4/14=0.29 |
| Sunny | 2 | 3 | 5/14=0.35 |
| All | 4/14=0.29 | 10/14=0.71 | |

$P(Sunny) = 0.35$

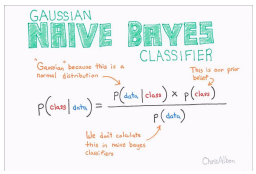
$P(No) = 0.29$ $P(Yes) = 0.71$

$P(Sunny|No) = 2/4 = 0.5$ $P(Sunny|Yes) = 3/10 = 0.3$

$P(No|Sunny) = P(Sunny|No)*P(No)/P(No)$
$P(Sunny|No) = 2/4 = 0.5$
$P(Sunny) = 0.35$
$P(No) = 0.29$
$P(No|Sunny) = 0.5*0.29/0.35 = 0.41$

$P(Yes|Sunny) = P(Sunny|Yes)*P(Yes)/P(Sunny)$
$P(Sunny|Yes) = 3/10 = 0.3$
$P(Sunny) = 0.35$
$P(Yes) = 0.71$
$P(Yes|Sunny) = 0.3*0.71/0.35 = 0.6$

Slang




Jargon

| The Posterior | The Evidence | The Prior |
|---|---|---|
| | The probability of getting this evidence if this hypothesis were true | The probability of H being true, before gathering evidence |

$$P(H|E) = \frac{P(H|E)\,P(H)}{P(E)}$$

The probability that the hypothesis (H) is true given the evidence (E)

The marginal probability of the evidence (Prob of E over all possibilities)


ESCUELA DEL REAL MADRID

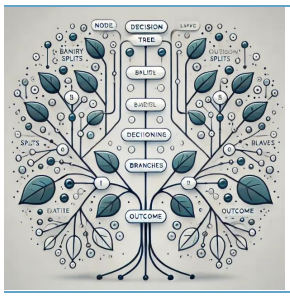
## Naive Bayes models

- Naive Bayes classifiers are **family** that are quite similar to the linear models.
- They tend to be **faster** in training.
- They often provide **generalizations slightly worse** than that of linear models.

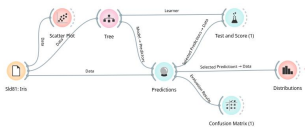

ESCUELA DEL REAL MADRID

## Naive Bayes models

- There are 3 kinds of Naive Bayes classifiers:
  - **GaussianNB**
    - Continuous data
  - **BernoulliNB**
    - Binary data
  - **MultinomialNB**
    - Count data
- They are great baseline models often used on **very large datasets** where training a linear one might take too long.




## Decision Trees



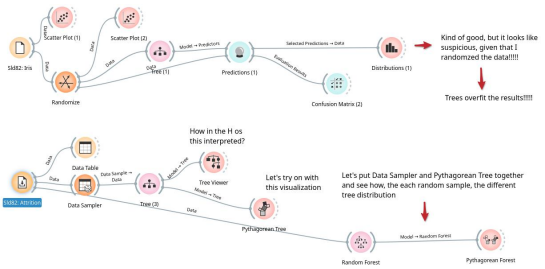

ESCUELA DEL REAL MADRID

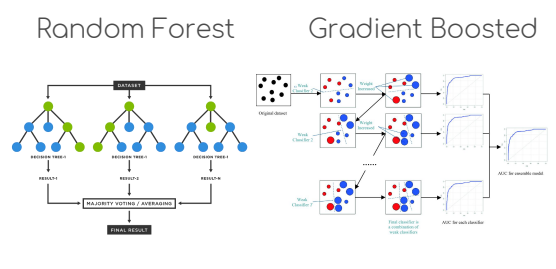
## Ensembles of Decision Trees

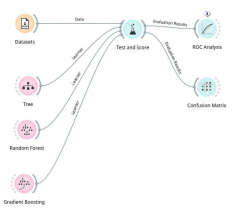
- **Ensembles** are combinations of multiple ML models to create a powerful one.
- There are two ensemble models that have proven to be effective, both of which use decision trees:
  - **Random forests (RF)**
  - **Gradient boosted decision trees.**

## Random Forest

- The main drawback of decision trees is that they **tend to overfit** the training data. RF are one way to address this problem.
- A RF is essentially **a collection of decision trees** where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data. If we build many trees, all of which work well and overfit in different ways, **we can reduce the amount of overfitting by averaging their results.**



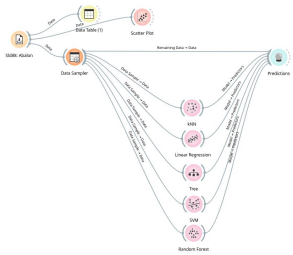





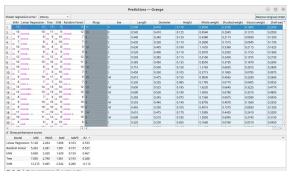
## Regression

the goal is to predict a *continuos value*




## Scoring

- Numeric Target
  - Abalone data
- Test & Score



## Evaluation metrics in classification

- We are talking about some metrics:
  - MSE: Mean Squared Error
  - RMSE: Root Mean Squared Error
  - MAE: Mean Absolute Error
  - R-Squared
  - MAPE

## Error coefficients

Mean Squared Error

Average squared difference between the estimated values and the actual value

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Root Mean Squared Error

-Quadratic mean of the differences between the observed values and predicted ones

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$$

## Error coefficients

Mean Absolute Percentage Error

Average magnitude of the absolute errors beetween the predicted and actual values.

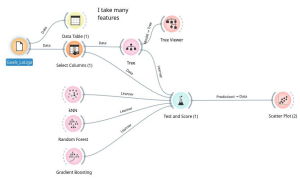
$$\mathrm{MAPE} = 100\frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|$$

$A_i$: $Actual value$
$F_i$: $Forecast value$

Coefficient of determination

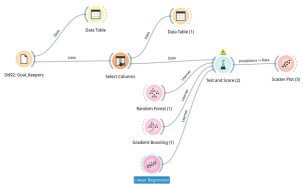
Proportion of the variation in the dependent variable th is predictable from the independent variable.

$$R^2 = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}} = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i (y_i - \bar{y})^2}$$



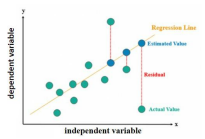

## Scoring

- Numeric Target
  - Number of goles (La Liga)
- Test & Score



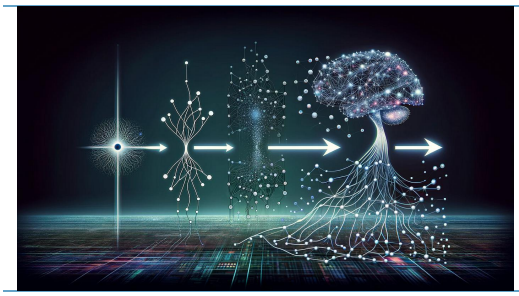

## Scoring

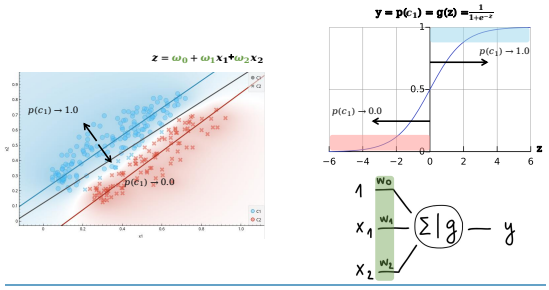
- Numeric Target
  - Goal Keepers
- Test & Score

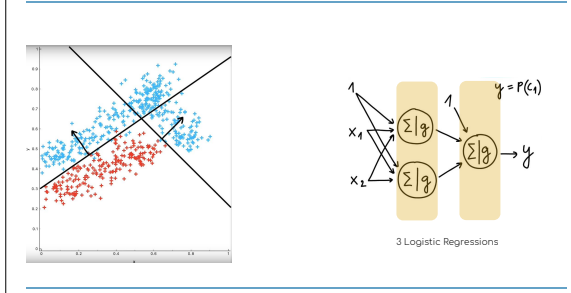


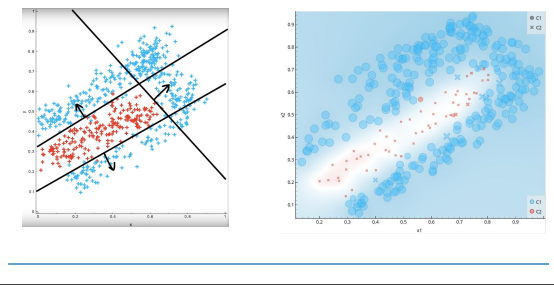

## Regression Algoritms

- Linear Regression
- Polynomial Regression
- Support Vector Regression (SVR)
- Decision Tree Regression
- Random Forest Regression








Unsupervised Learning

## Challenges in Unsupervised Learning

Usually applied tounlabeleddata.

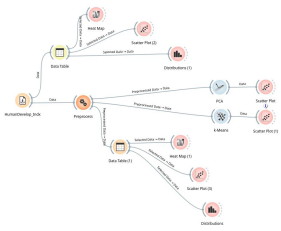
**Challenges**

- We don't know what the right output should be.
- It is tough to say whether a model"did well."
- It is tough to say is what modellearned "is useful"

UNSUPERVISED DATA

## Unsupervised learningalgorithms

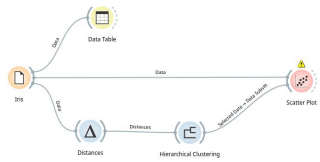
| Clustering | Dimensionality reduction | Association |
|---|---|---|
| K-means<br>Polynomial<br>Hierarchical<br>FuzzyC-means | PCA<br>Kernel-PA | Apriori Algorithm<br>Eclat Algorithm<br>FP-growth Algorithm |
| Identify clusters | Transform data to betetr understanding | Recommentation list |



ECOFLOR

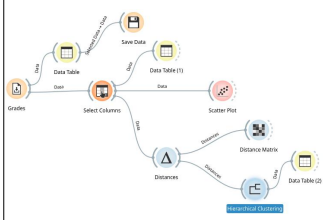
## Preprocessing and Scaling

- **We need first to standardize data.**
  - StandardScaler
  - RobustScaler
  - MinMaxScaler, etc



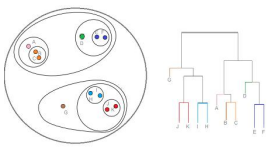

## Hierarchical clustering

- Iris —> Data Table
- Distances —>Hierarchical Clustering
- Scatter plot
  - Compare selected data with its visual representation.



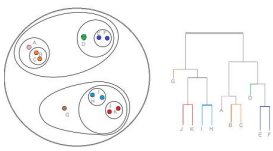

## Hierarchical clustering

- Grades—> Data Table& Save it
- Select columnsDistances —> HierarchicalClustering
  - Distance matrix
  - Dendogram
  - Look at the results



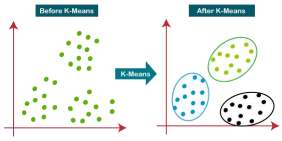
## HierarchicalClustering

- H.C. Agglomerative:
  - It starts with a data-point as a single cluster, then merges the closest pair into a large one, and so on.
- H.C. Divisive:
  - It starts with all points as a large cluster, then splits into smaller ones, and so on



## HierarchicalClustering

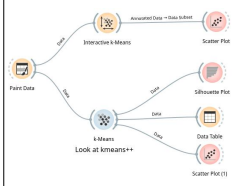
- H.C. handle non-spherical clusters, can produce clusters of different shapes and sizes.
- H.C. is:
  - computationally expensive for large datasets
  - sensitive to the choice of distance metric
  - produce suboptimal results when the clusters are highly non-linear or when there are outliers in the data.




## K-Means

Just distances among points and emergent centroids

- H.G. isn't the best for large datasets
  - Could you imagine a H.C. for genetic data?
    - Matriz distance of millions...
    - Almost 8Tb in memory...

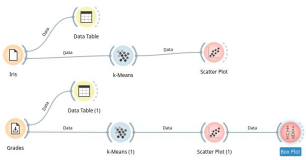



Playing with

## K-means

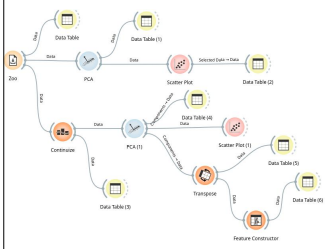
Real working

- Download Educational widget
- Paint 3 data clusters
- InteractiveK-M compared with Data
  - Fats Algorithm
    - Look! Each point is linked to a centroid
    - Then, it recompute centroids...
      - You can add centroids, and re-run.



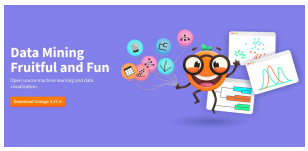

## K-means

- Iris
- Grades




## PCA

- Datasets->Zoo
- PCA
- Scatter Plot




## Bibliography


- @JohannH_M
- https://johemart.wixsite.com/neurocomplexity
- Orange Data Mining YouTube Chanel
- The Creativity Code, Marcus du Sautoy
- Introduction to Machine Learning with Python, A. Müller and S. Guido