# Machine Learning and applications in AI
# Lecture III

Martin Palazzo

Universite de Technologie de Troyes
Universidad Tecnologica Nacional Buenos Aires

*mpalazzo@frba.utn.edu.ar*

November 29, 2022

# Overview

# Data Normalization

# Data pre-processing

In many cases the features of a data set does not belong to the same domain. imagine the features of a car like its weight, the number of kilometers recorded and the mili liters of it liquid break system. The features are measured in Kilometers, Kilograms and mL. This means that the scale of the original features can be really different. Learning algorithm can be sensitive to this situation. For this reason we can make a transformation of each feature along the samples known as auto-scaling.

## Feature standard scaling

$$z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j} \quad \forall j$$

# Data pre-processing

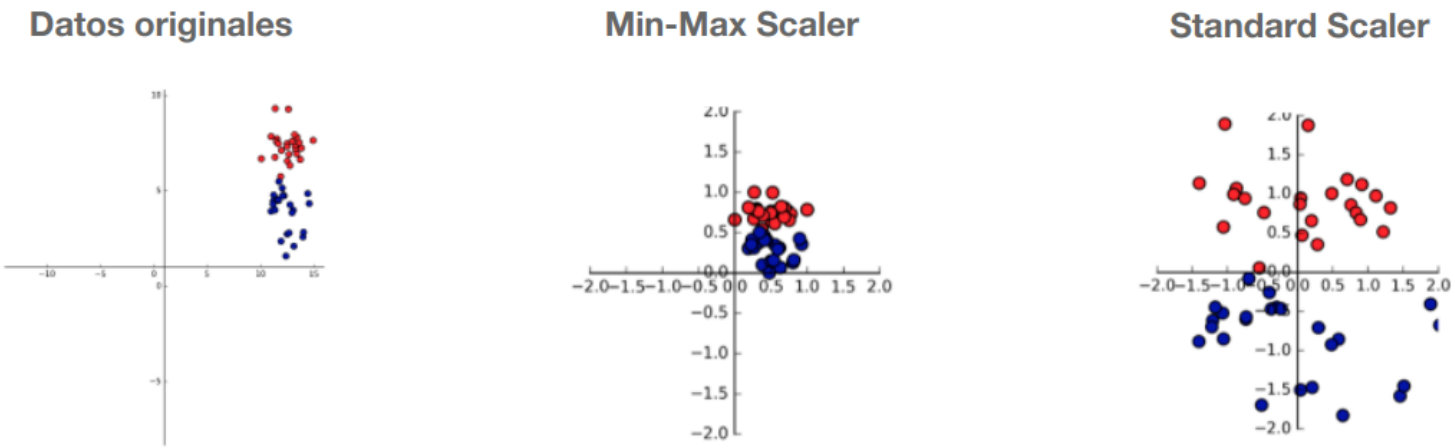



Figure: Data scaling

Source:
https://python-data-science.readthedocs.io/en/latest/normalisation.html

# Data pre-processing

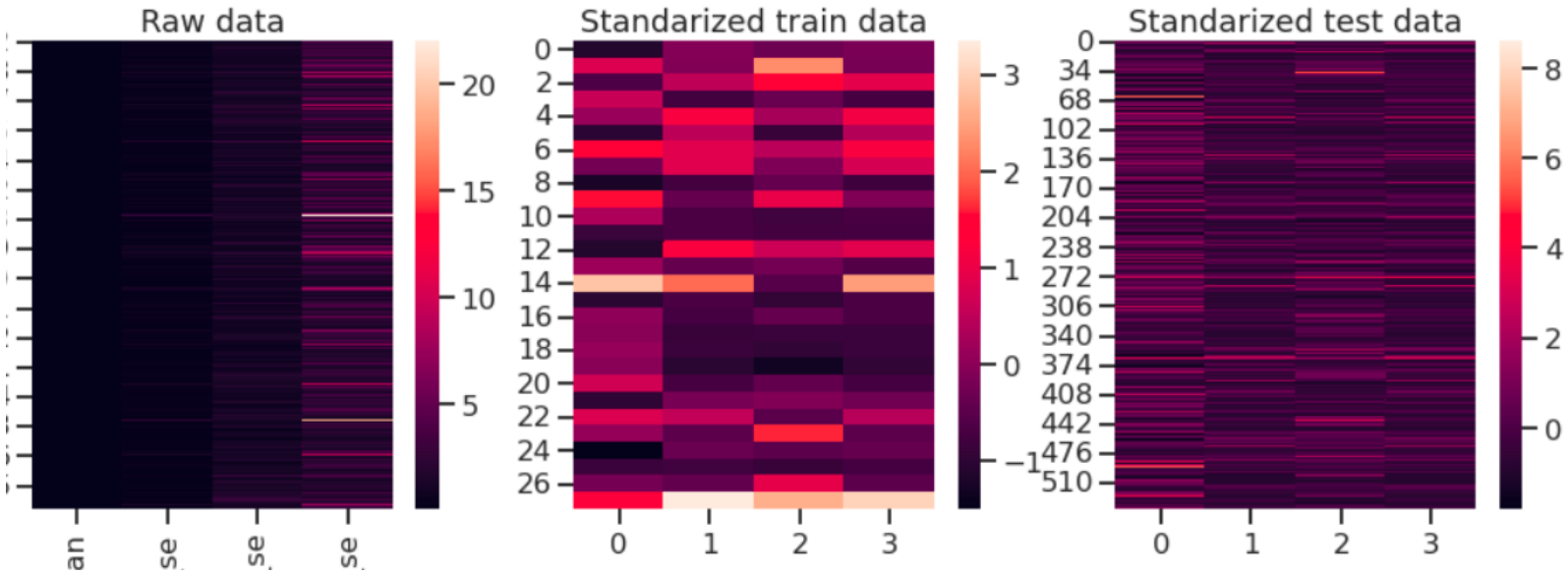



Figure: Data scaling

Source: https://github.com/clusterai

# Confusion Matrix

# Confusion Matrix

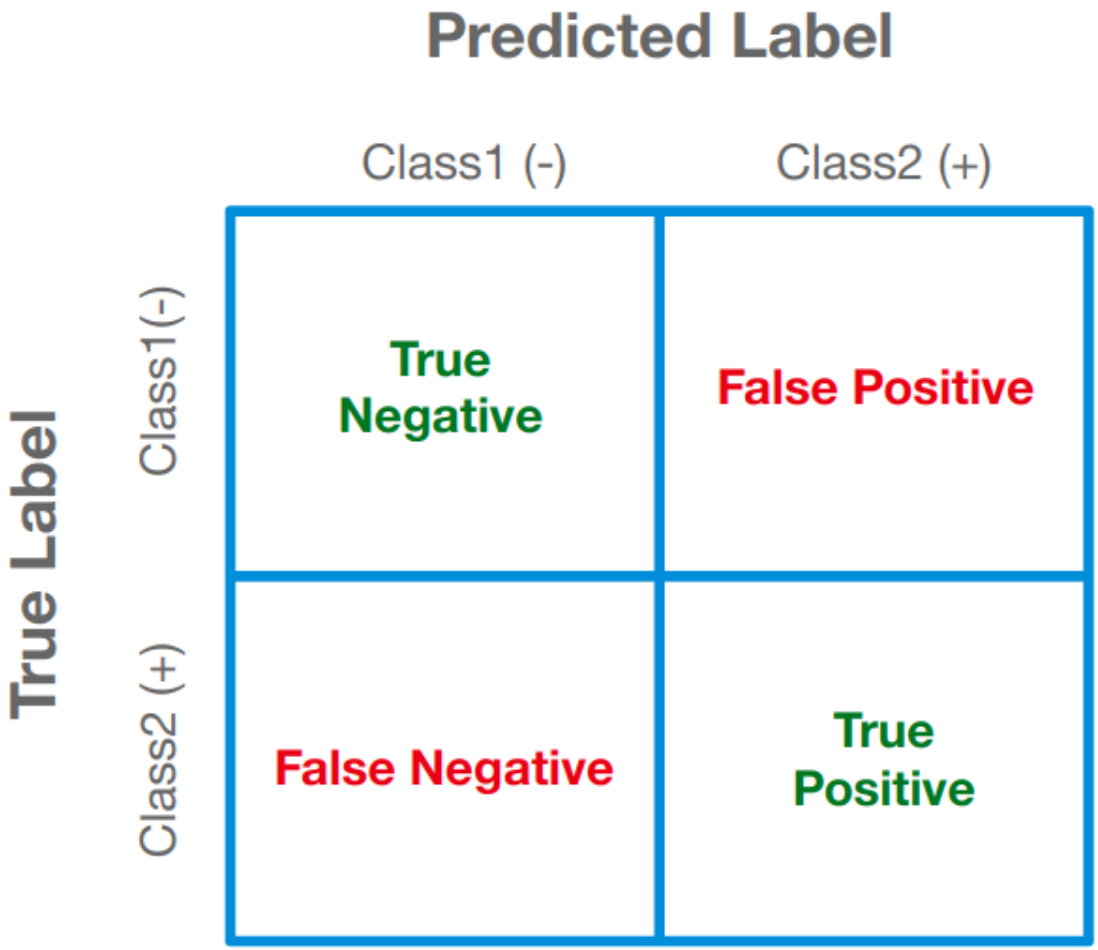



Figure: Confusion Matrix

# Confusion Matrix

From the confusion matrix it is possible to obtain the following classification performance metrics

$$\text{TP} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

# Area under the ROC curve

# The Area under the ROC curve

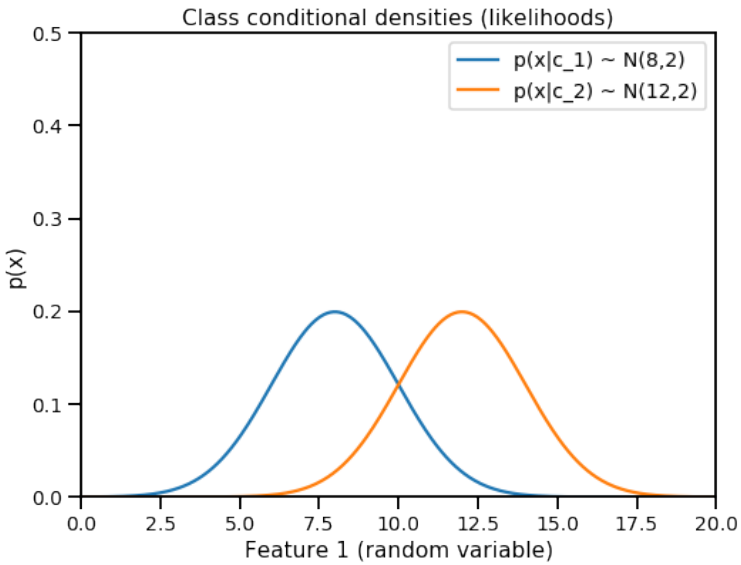
## ROC curve

A Receiver Operating Characteristic curve (ROC) curve displays how well a model can classify binary outcomes. An ROC curve is made by plotting a false positive (FP) rate against a true positive (TP) rate for each possible cutoff value.

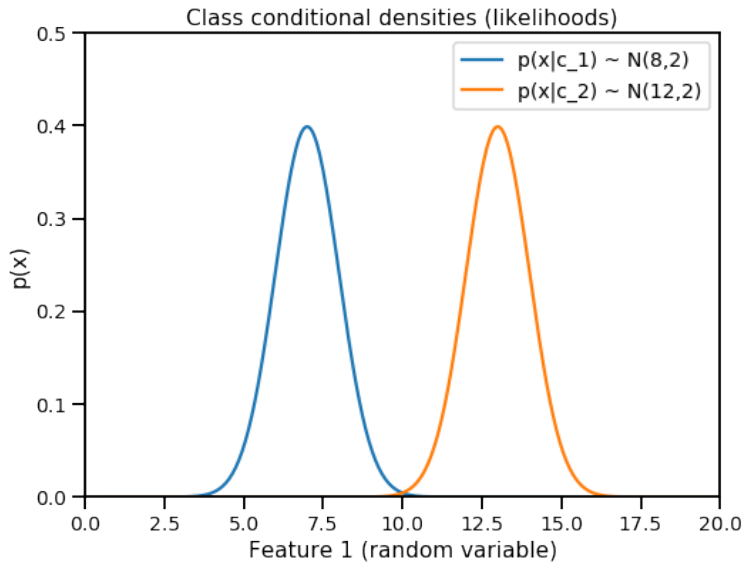
If we take the case of medical diagnosis, the ROC curve shows how well a model assigns correctly a healthy status to healthy people and a positive diagnosis to patients with disease. Considering the positive value to the Diagnosis status, by using different thresholds between the two classes it is analyzed:

- the fraction of positive diagnosis on patients with disease (TP rate).
- and the fraction of healthy patients that were incorrectly classified as positive diagnosis (FP rate).

# The Area under the ROC curve




(a) Overlapped



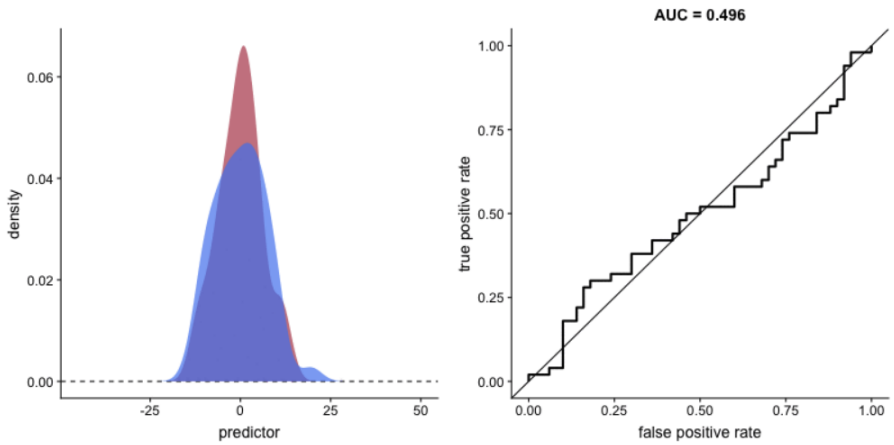

(b) Well separated

Figure: Class distributions overlapped and separated.

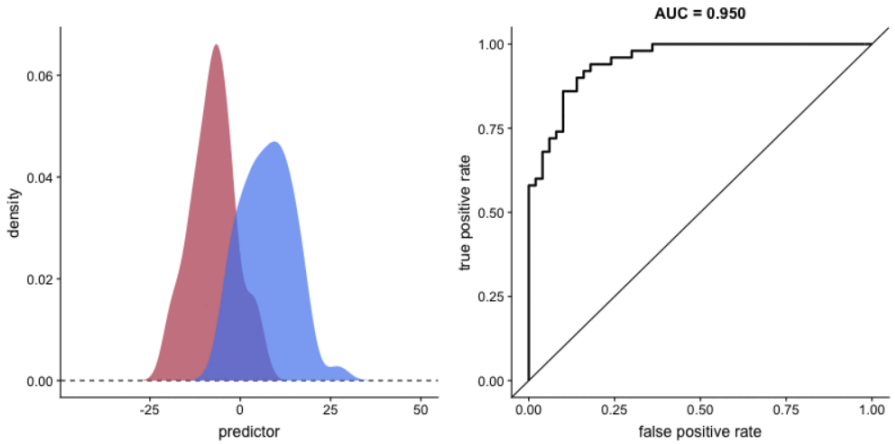
## Question

If we build a classifier using the bayes rule, which case will present a better area under the curve (AUC) ROC?

# The Area under the ROC curve




(a) Overlapped




(b) Well separated

## Remark

The AUC ROC depends on the density functions of each class and the classifier used.

# The Area under the ROC curve

## Exercise 01

Generate 100 univariate samples of class 0 $C_0$ with $mu_0 = 6$, $\sigma_0 = 2$. Then generate 100 univariate samples of class 1 $C_1$ with $mu_0 = 13$, $\sigma = 2$. Generate 10 different thresholds (different priors) and compute the AUC ROC.

## Exercise 02

Generate 100 univariate samples of class 2 $C_2$ with $mu_2 = 8$, $\sigma_2 = 2$. Then generate 100 univariate samples of class 3 $C_3$ with $mu_3 = 12$, $\sigma_3 = 2$. Generate 10 different thresholds (different priors) and compute the AUC ROC.

# Train, Validation and Test sets

# Train, Validation and Test sets

Given a data set $S = (x_1, y_1)...(x_n, y_n)$ we never will fit a classifier using the full set.

## Train set

The set of samples $X_{tr} \in S$ used to build the model (classifier, regression function).

## Validation set

The set of samples $X_{val} \in S$ used to measure the prediction performance of the classifier. Once it is validated, train and validation sets are merged to define a final model.

## Test set

The independent set of samples $X_{te} \in S$ used to measure the prediction performance of the final classifier. These samples are never used during the training and fit process.

# Cross Validation

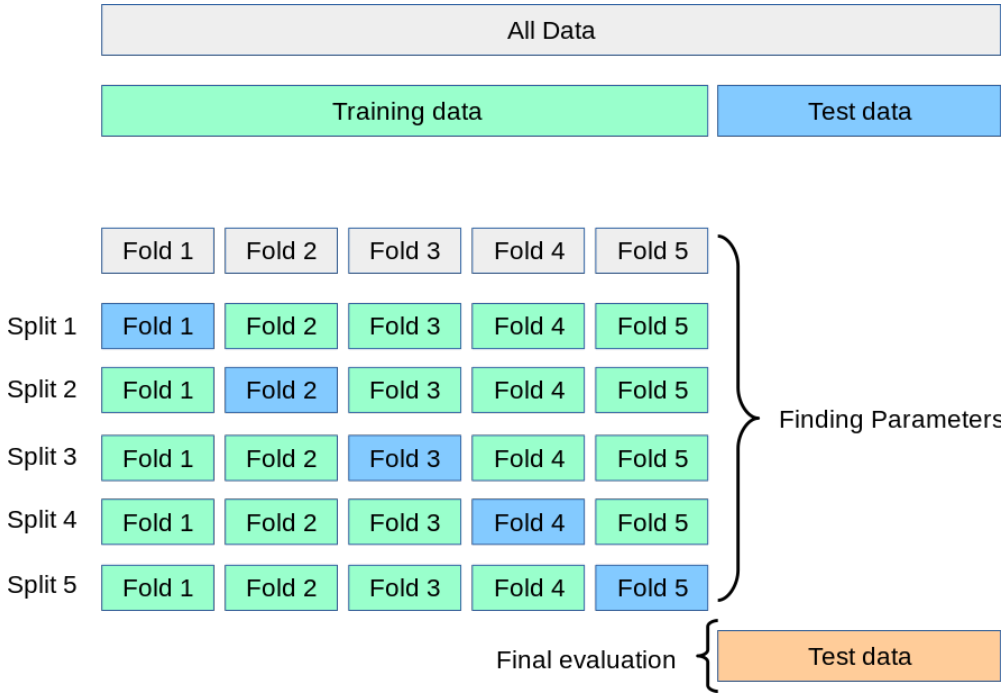



Figure: Cross validation process

## 5 fold cross validation

For cross validation first we can split the full dataset in 80% for train and 20% for test. Then make CV on training set.

# Train, Validation and Test sets

## Exercise 03

Given the breast cancer wisconsin dataset divide the data into train and test sets. Then using the training set train a KNN model by 5 fold cross validation. Compute the accuracy and empirical error of each fold. Select the best parameter and re-train using the full training set. Finally classify the rest of the test labels using the trained model.


Chollet, F. (2018). Keras: The python deep learning library. Astrophysics Source Code Library.


Bishop, C. M. (2006). Pattern recognition and machine learning. springer.


Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.