ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

DATA ENGINEERING AND ANALYTICS

Professor: JESÚS GARCÍA SAN LUIS
E-mail: jgarcias@faculty.ie.edu

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Modeling

SESSION 4

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

Data Science Projects

Bussiness Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
• Business Objectives • Technical constraints • Project Planning	• Raw Data • Graphics • Quality Verification	• Cleaning: empty values, outlayers • Feature selection	• Model selection • Parameters selection	• Technical evaluation • Business utility	• Results presentation • Application Architecture

- Modeling is only 20% of the total effort
- As a consequence, Data Science is not about learning algorithms. Is about dealing with data in a comprehensive and systematic way to select the best algorithms and apply them correctly
- You must use a systematic approach to develop data science projects

What is Modeling

- Fit a function to parameters to obtain a result as close as possible to the related outputs

- Dataset (first five lines of csv sample):

```
"preg";"plas";"pres";"skin";"test";"mass";"pedi";"age";"class"
6;148;72;35;0;33.6;0.627;50;1
1;85;66;29;0;26.6;0.351;31;0
8;183;64;0;0;23.3;0.672;32;1
1;89;66;23;94;28.1;0.167;21;0
0;137;40;35;168;43.1;2.288;33;1
. . .
```

- Inputs (features or attributes): `preg, plas, skin, test, mass, pedi, age`
- Outputs: `class`

- Modeling: to find a function f such that f(inputs) is as close to outputs as possible.
  - The 'closeness' is measured by using a loss function, dependent of the kind of model and the problem.
  - We have available a list of 'models' in our machine library. Once one model is chosen, the library will fit the model parameters minimizing the loss function for the given inputs.

What is Modeling

- In the programs we will use the following naming conventions:
  - Inputs -> X
  - Output -> y (usually is a vector but not necessarily)

- In the example:
  - X =

```
6 148 72 35   0   33.6 0.627 50
1  85 66 29   0   26.6 0.351 31
8 183 64  0   0   23.3 0.672 32
1  89 66 23  94   28.1 0.167 21
0 137 40 35 168   43.1 2.288 33
```

  - y =

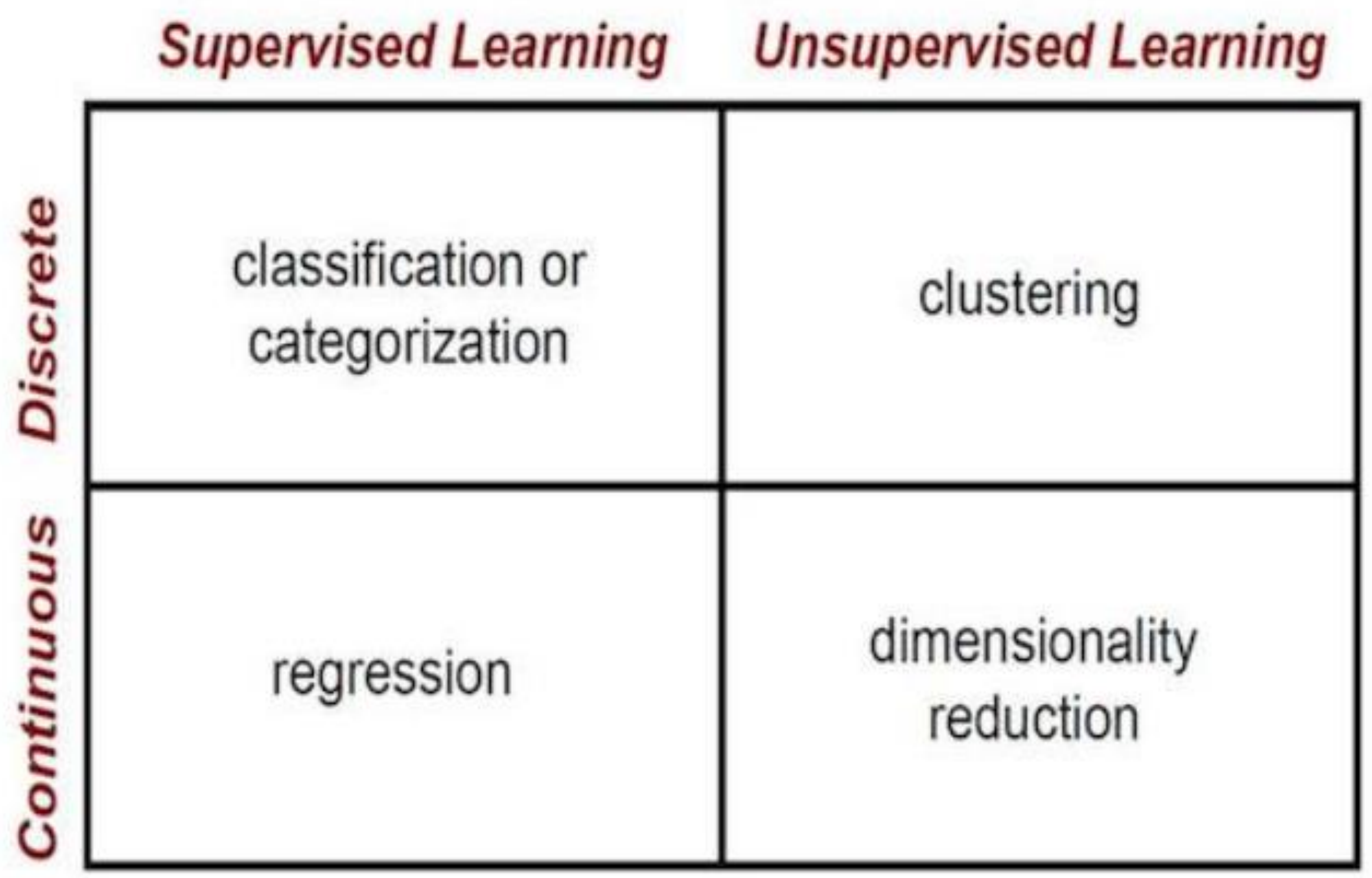
```
1
0
1
0
1
```

Models

Model types

- Regression / Classification
- Supervised / Unsupervised / Semi-supervised

- Unsupervised Classification = Clustering / Association (rule-based)
- Unsupervised Regression = ???



Source: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

Typical Models

The source: https://scikit-learn.org/stable/supervised_learning.html

- Classification
  - Logistic regression
  - Nearest Neighbors (KNeighbors, RadiusNeighbors)
  - Decision Trees
  - Support Vector Machines
  - Neural Networks

- Regression
  - Linear Regression
  - Regularizations: Ridge and Lasso
  - Elastic-Net (combination of Ridge and Lasso)
  - Support Vector Machines (SVR)
  - XGBoost
  - Neural Networks

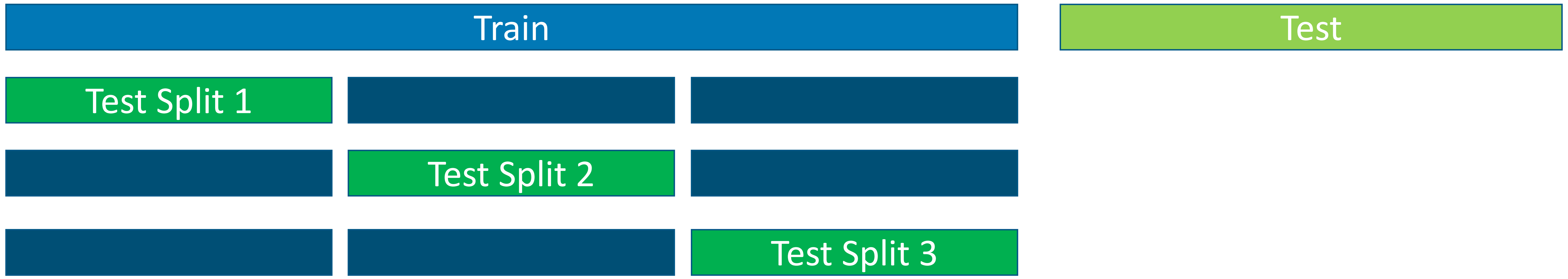
- Clustering
  - k-means (convex surfaces)
  - Autoencoders

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

Model Selection

Cross Validation

- Why do we need cross validation?
- What is cross validation?



- Scikit learn:

```
from sklearn import metrics
scores = cross_val_score(clf, X, y, cv=3, scoring='f1_macro’)
```

- Stratified cross validation

Model Selection

Hyper-parameter tuning

- Why hyper-parameter tuning?

- What do you need to tune parameters?
  - Parameter space (to calculate the parameters combinations)
  - Method for sampling candidates
  - Cross validation and score function

- Methods
  - Exhaustive grid search
  - Randomized optimization
  - Tournament or successive halving (saving resources). Can be used for the two methods above

https://scikit-learn.org/stable/modules/grid_search.html

Model Selection

Scores

- Regression Scores
  - Mean squared error / mean absolute error / max error…
  - Mean squared logarithmic
  - Explained variance

- Classification Scores
  - Accuracy
  - Confusion matrix
  - Receiver Operating Characteristic and Area Under the Curve

Model Selection

Scores – Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	123	2
	Positive	6	432

Accuracy: **(TN + TP) / samples**

Model Selection

Scores – Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	432	2
	Positive	1	11

Accuracy: (TN + TP) / samples = 434 / 446 = 0,97
Precision: TP / (TP + FP) = TP / Total Predicted Positive = 11/12 = 0,92

Model Selection

Scores – Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	432	2
	Positive	1	11

Accuracy: (TN + TP) / samples
Precision: TP / (TP + FP) = TP / Total Predicted Positive = 11/12 = 0,92
Recall or Sensivity: TP / (TP + FN) = TP / Total Actual Positive = 11/13 = 0,85
Specifity: TN / (TN + FP) = 432/434 = 0,99

Model Selection

Scores – Confusion Matrix

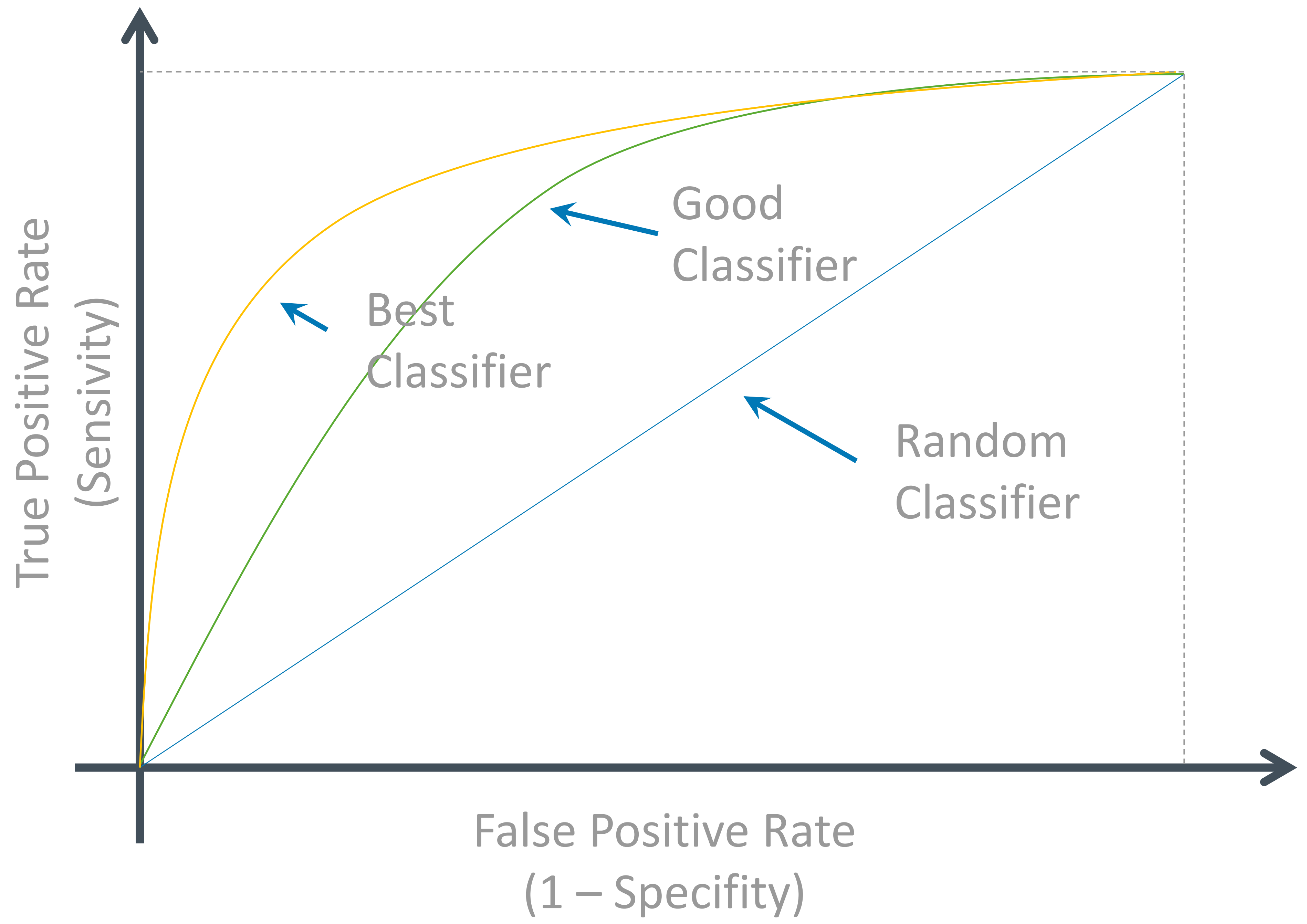
		Predicted	
		Negative	Positive
Actual	Negative	432	2
	Positive	1	11

Accuracy: (TN + TP) / samples
Precision: TP / (TP + FP) = TP / Total Predicted Positive = 11/12 = 0,92
Recall or Sensivity: TP / (TP + FN) = TP / Total Actual Positive = 11/13 = 0,85
Specifity: TN / (TN + FP) = 432/434 = 0,99
F1: 2 * (Precision * Recall) / (Precision + Recall) = 2 * 0,92 * 0,85 / (0,92 + 0,85) = 0,88

Model Selection

Scores – ROC - AUC

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



$$FPR = \frac{FP}{FP + TN}$$

The ROC curve measures how well the classifier separates the probabilities of positive and negative cases