Co-funded by
the European Union

Alliance for Fostering
Business and Education Innovation
through Digital Supply Chains

Data Analytics

Familiarization with various machine learning techniques

Supervised Machine Learning

Co-funded by the European Union

- **Definition**: Learning from labeled data to predict outcomes for new data.
- **Common Methods**:
  - **Linear Regression**: Predicts continuous outcomes.
  - **Logistic Regression**: Used for binary classification tasks.
  - **Decision Trees**: Versatile for classification and regression.
  - **Random Forest**: An ensemble method that improves prediction accuracy and robustness.
  - **Support Vector Machines (SVM)**: Effective in high-dimensional spaces for classification and regression.

Linear Regression

Co-funded by the European Union

- **Objective**: Predict a continuous outcome variable based on one or more predictor variables.
- **Definition**: Assumes a linear relationship between the dependent and independent variables.
- **Equation**: $Y = \beta 0 + \beta 1X + \epsilon$, where Y is the dependent variable, X is the predictor, $\beta 0$ is the intercept, $\beta 1$ is the slope, and ϵ is the error term.
- **Application**: Useful in scenarios like sales forecasting from advertising data.

Linear Regression

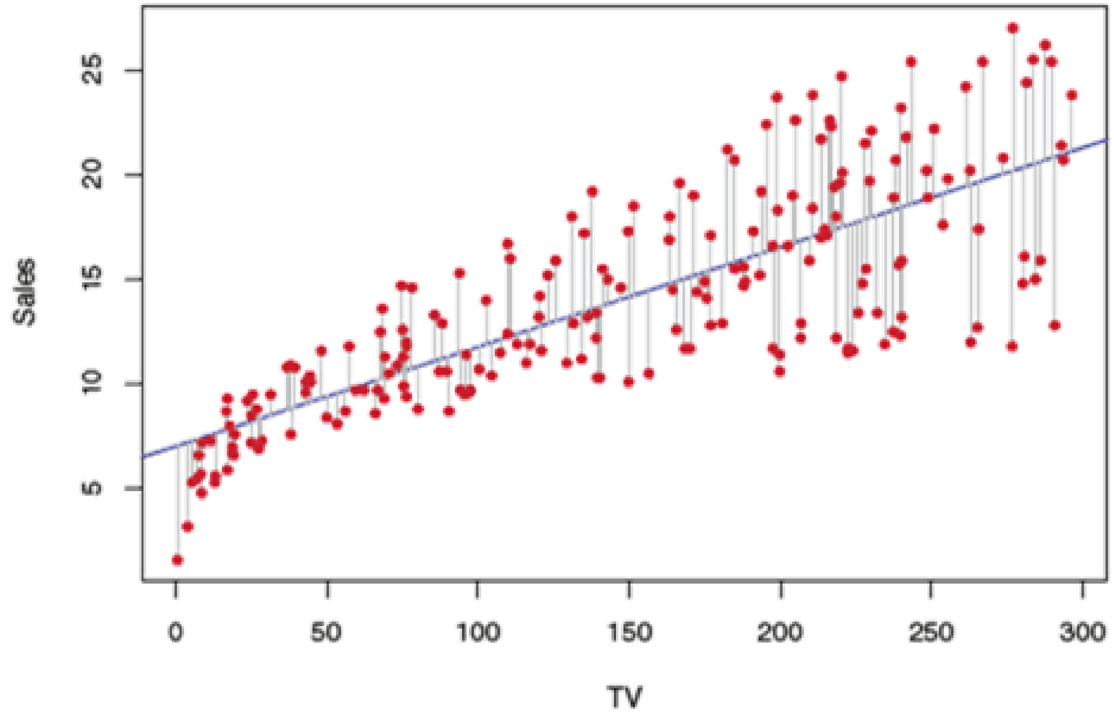
Co-funded by the European Union

intercept and *slope* terms

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Linear Regression

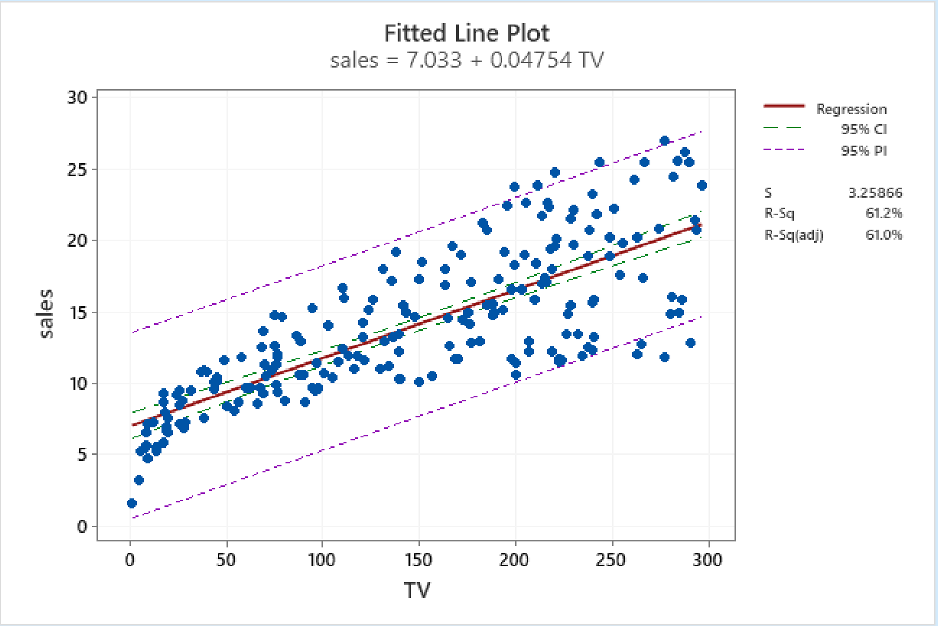
Co-funded by the European Union

confidence interval

$$\hat{y}_0 \pm t_{1-a/2,n-2}\sqrt{\frac{RSS}{n-2}\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

prediction interval

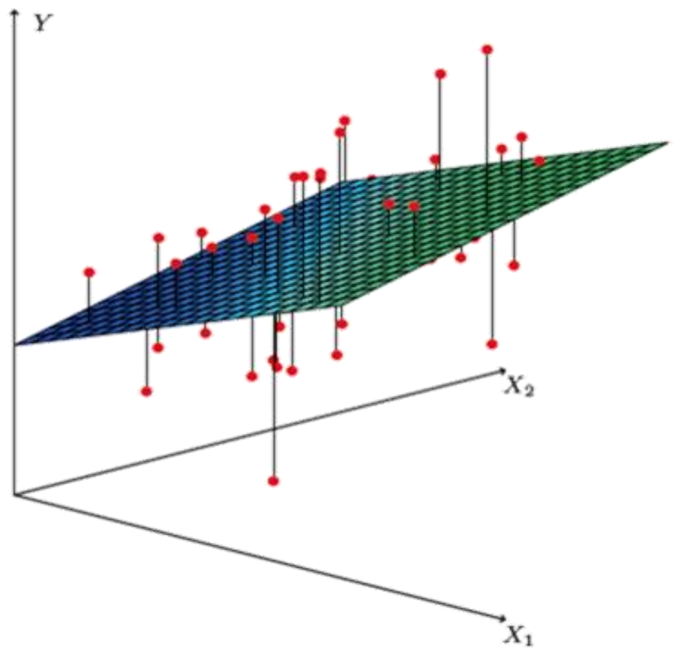
$$\hat{y}_0 \pm t_{1-a/2,n-2}\sqrt{\frac{RSS}{n-2}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$



Multiple Linear Regression

Co-funded by the European Union

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Logistic Regression

Co-funded by the European Union

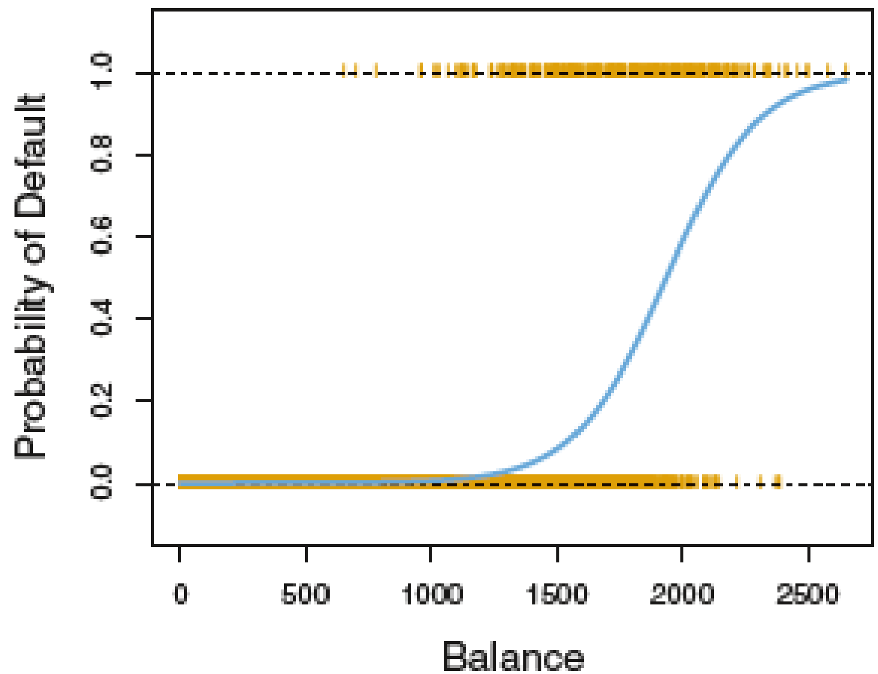
- **Objective**: Model the probability of a binary outcome from a set of predictor variables.
- **Definition**: Despite its name, used for binary classification, not regression.
- **Mechanism**: Uses the logistic function to map predicted values to probabilities.
- **Use Case**: Estimating the likelihood of a credit card default based on balance.

Logistic Regression

Co-funded by the European Union

The logistic regression model

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

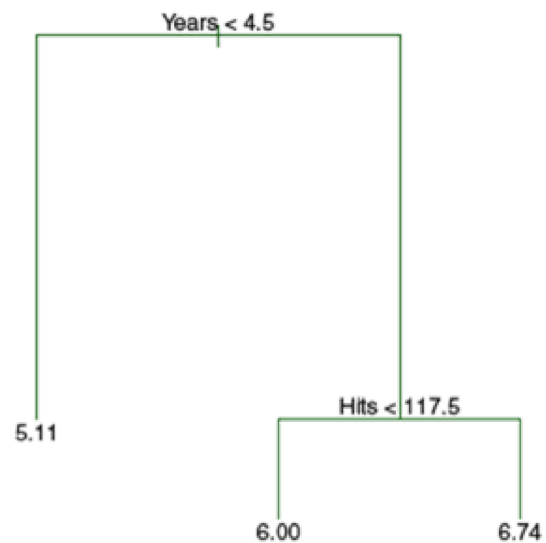
Decision Trees

- **Objective**: Predict an outcome based on decision rules inferred from the data features.
- **Structure**: Tree-like model of decisions and their possible consequences.
- **Types**: Used for both classification (categorical outcome) and regression (continuous outcome).
- **Advantage**: Easy to interpret and capable of handling both numerical and categorical data.

Regression Trees

Co-funded by the European Union

The aim of a regression tree is to predict a continuous output, with each leaf node representing a numerical value. Each internal node makes a decision depending on the value of a feature, and the tree is built by repeatedly dividing the data until a stopping condition is satisfied.

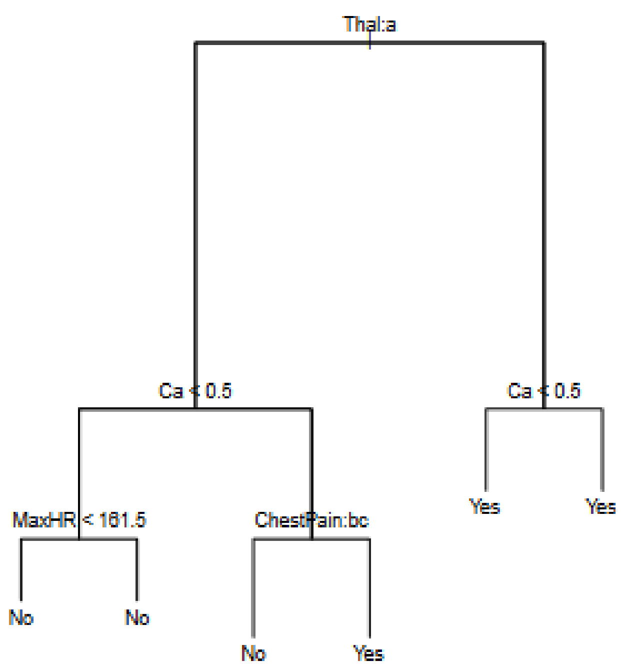


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Classification Trees

Co-funded by the European Union

The aim of a classification tree is to give a class label to every occurrence in the dataset, with each leaf node representing a class label. Each internal node makes a decision depending on the value of a feature, and the tree is built by repeatedly dividing the data into subsets until a halting condition is satisfied.



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

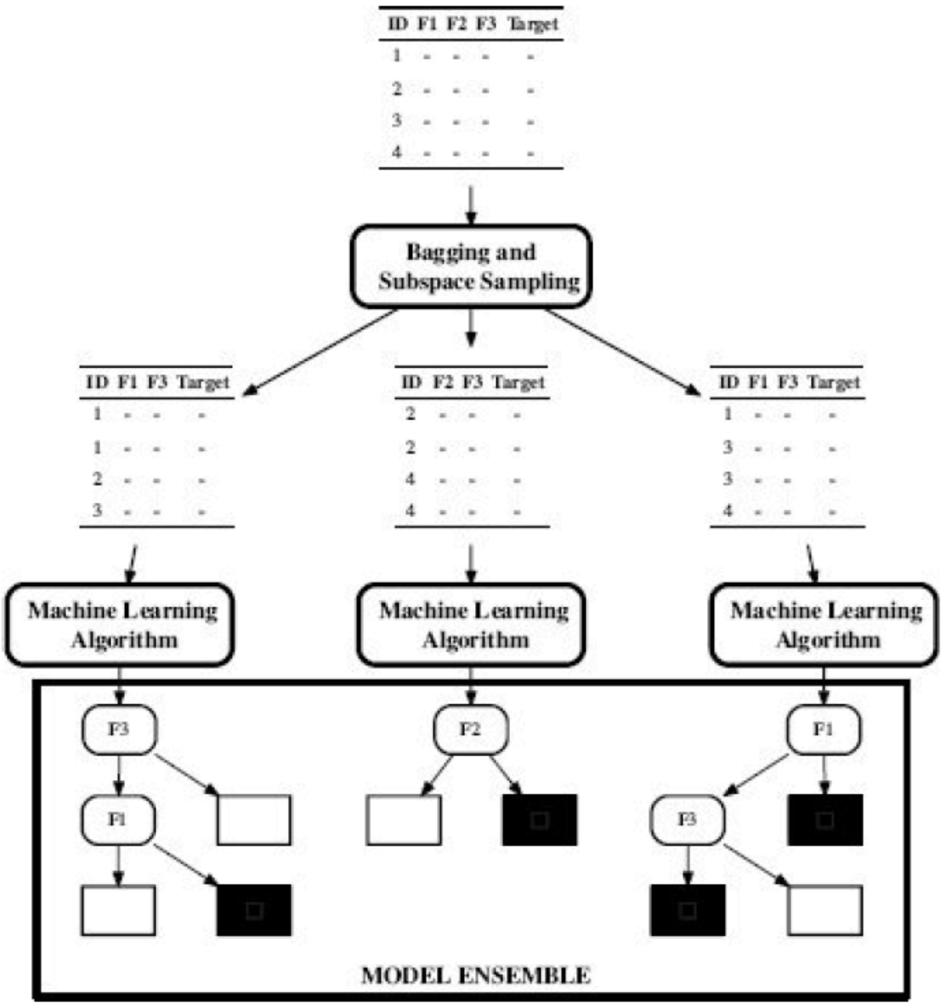
Random Forest

Co-funded by the European Union

- **Objective**: Improve predictive accuracy and control over-fitting compared to a single decision tree.
- **Definition**: An ensemble of decision trees, typically trained with the "bagging" method.
- **Functionality**: Each tree votes for a class, and the class with the most votes becomes the model’s prediction.
- **Benefit**: Robust against overfitting and highly effective across a wide range of classification and regression tasks.

Random Forest

Co-funded by the European Union



Source: Fundamentals of machine learning for predictive data analytics, Kumaresan Perumal, John D. Kelleher, Brian Mac Namee, Aoife D'Arcy, 2nd edition

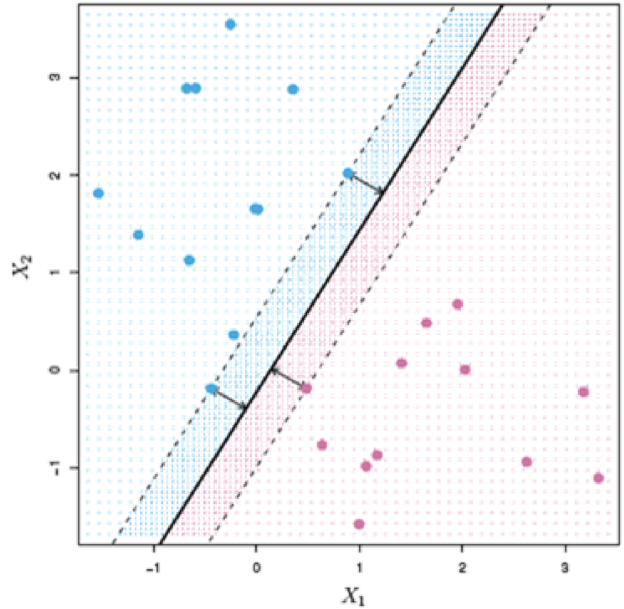
Support Vector Machines

Co-funded by the European Union

SVMs find a hyperplane that best separates classes in a high-dimensional space. They are effective for both linear and non-linear classification tasks.

Support vectors are the data points that lie closest to the decision boundary (hyperplane). These points influence the position and orientation of the hyperplane.

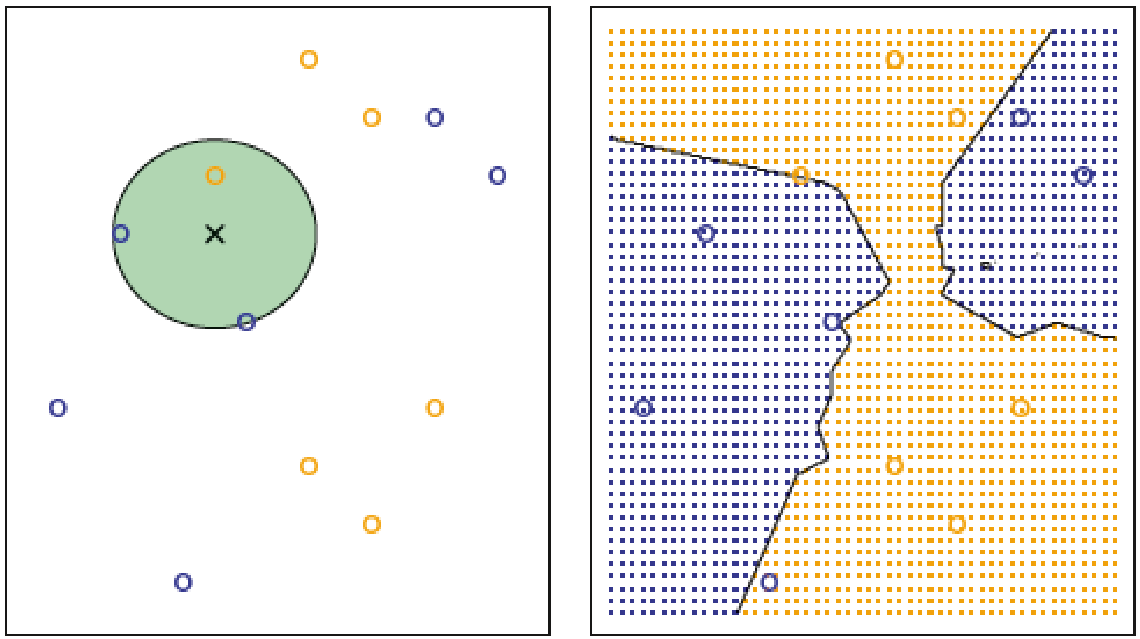
The ***margin*** is the distance between the hyperplane and the nearest data point from either class. SVM aims to maximize this margin.



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

K-Nearest Neighbors (KNN)

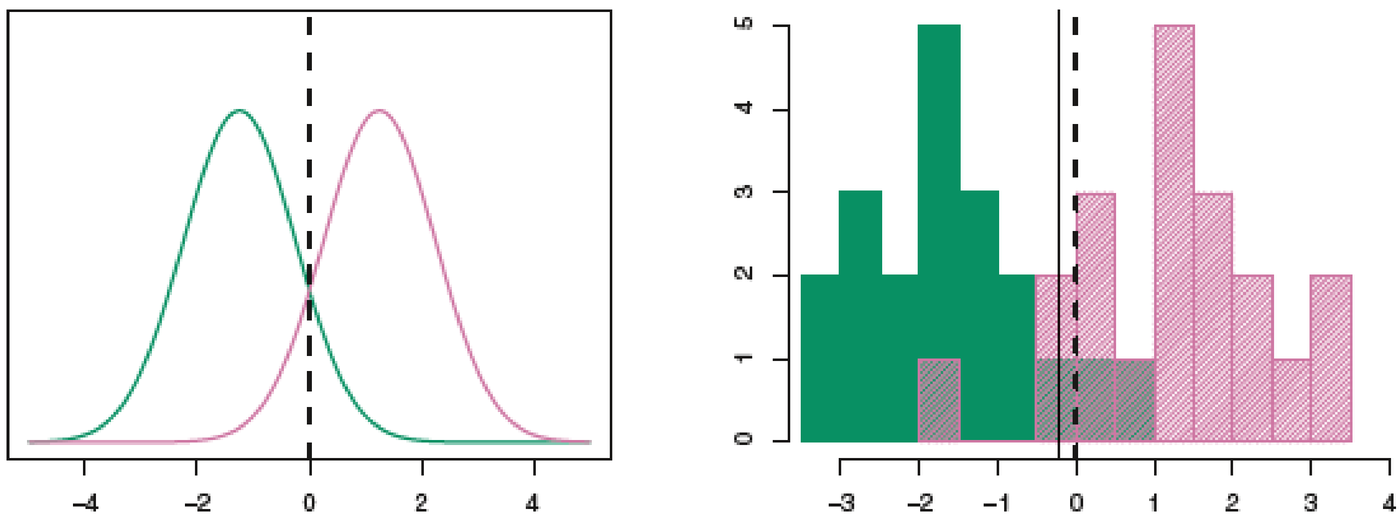
The algorithm predicts using the mean of the values of the closest neighbors (for regression) or the majority class (for classification)



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Linear Discriminant Analysis (LDA)

The goal is to find a linear combination of features that characterizes or separates two or more classes.

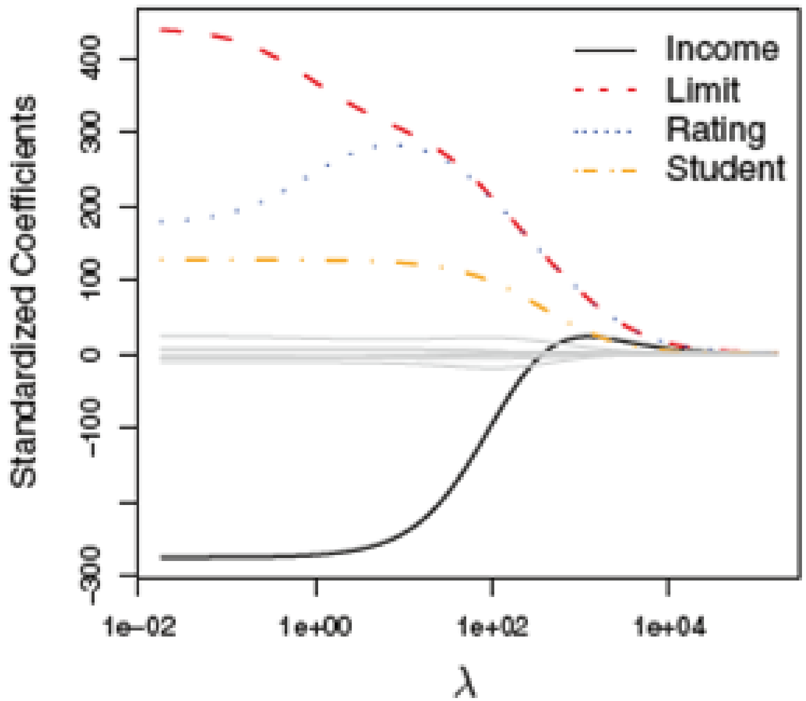


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Ridge Regression

Co-funded by the European Union

Ridge Regression is a linear regression technique that introduces a regularization term to the standard linear regression objective. The regularization term penalizes large coefficients, preventing them from becoming too influential in the model.

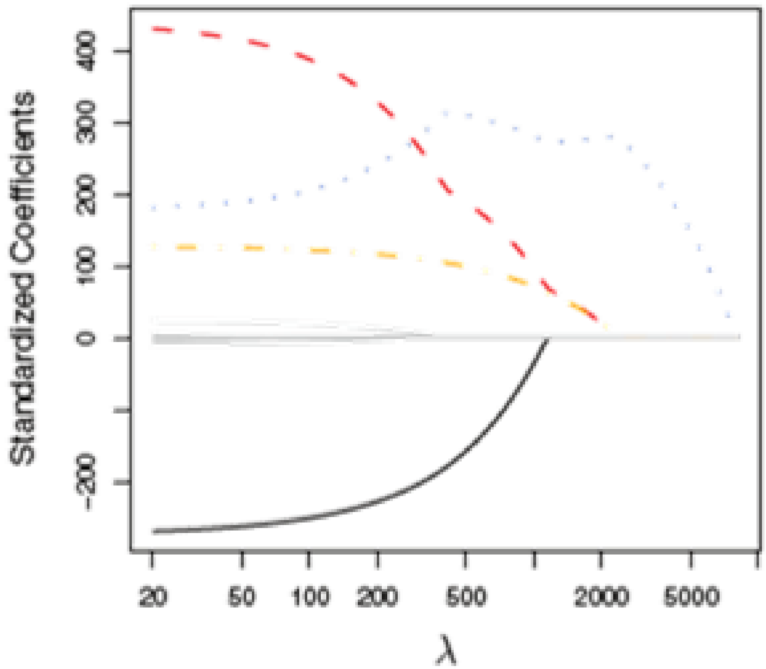


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Lasso Regression

Co-funded by the European Union

Lasso Regression is a linear regression technique that incorporates a regularization term to the standard linear regression objective. Similar to Ridge Regression, Lasso introduces a penalty term, but in this case, the penalty is proportional to the absolute values of the coefficients.



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Unsupervised Machine Learning

- **Definition**: Learning patterns from data without pre-assigned labels.
- **Common Methods**:
  - **Clustering (e.g., K-Means, Hierarchical)**: Identifies groups with similar characteristics.
  - **Principal Component Analysis (PCA)**: Reduces dimensionality while retaining variance.
  - **Autoencoders**: Neural networks for learning efficient codings.

Clustering

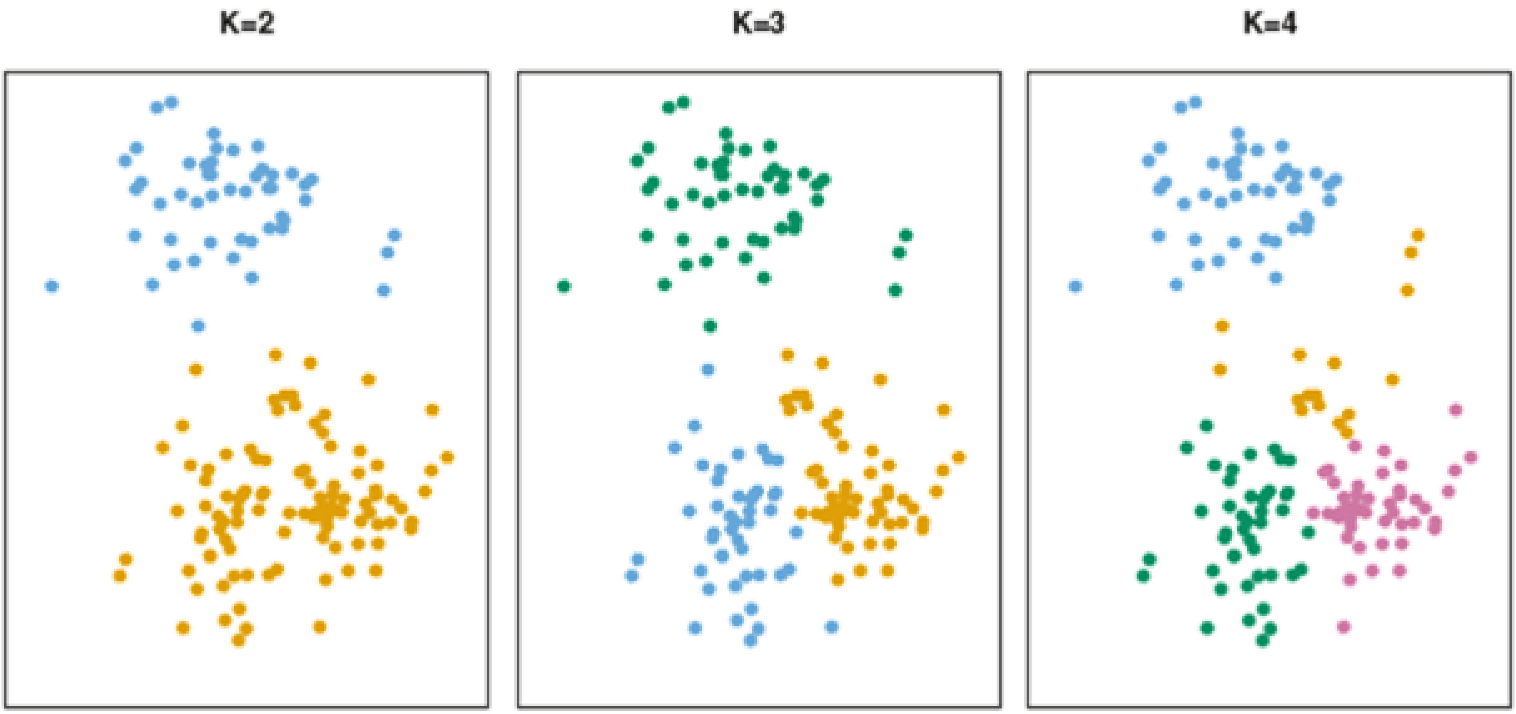
Co-funded by the European Union

- **Objective**: Partition the data into K distinct, non-overlapping clusters based on similarity.
- **Mechanism**: Assigns each data point to the nearest cluster while keeping the centroids as small as possible.
- **Application**: Market segmentation, identifying groups with similar characteristics within a dataset.

K-means Clustering

Co-funded by the European Union

$$\underset{C_1,\ldots,C_K}{\text{minimize}}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right\}$$

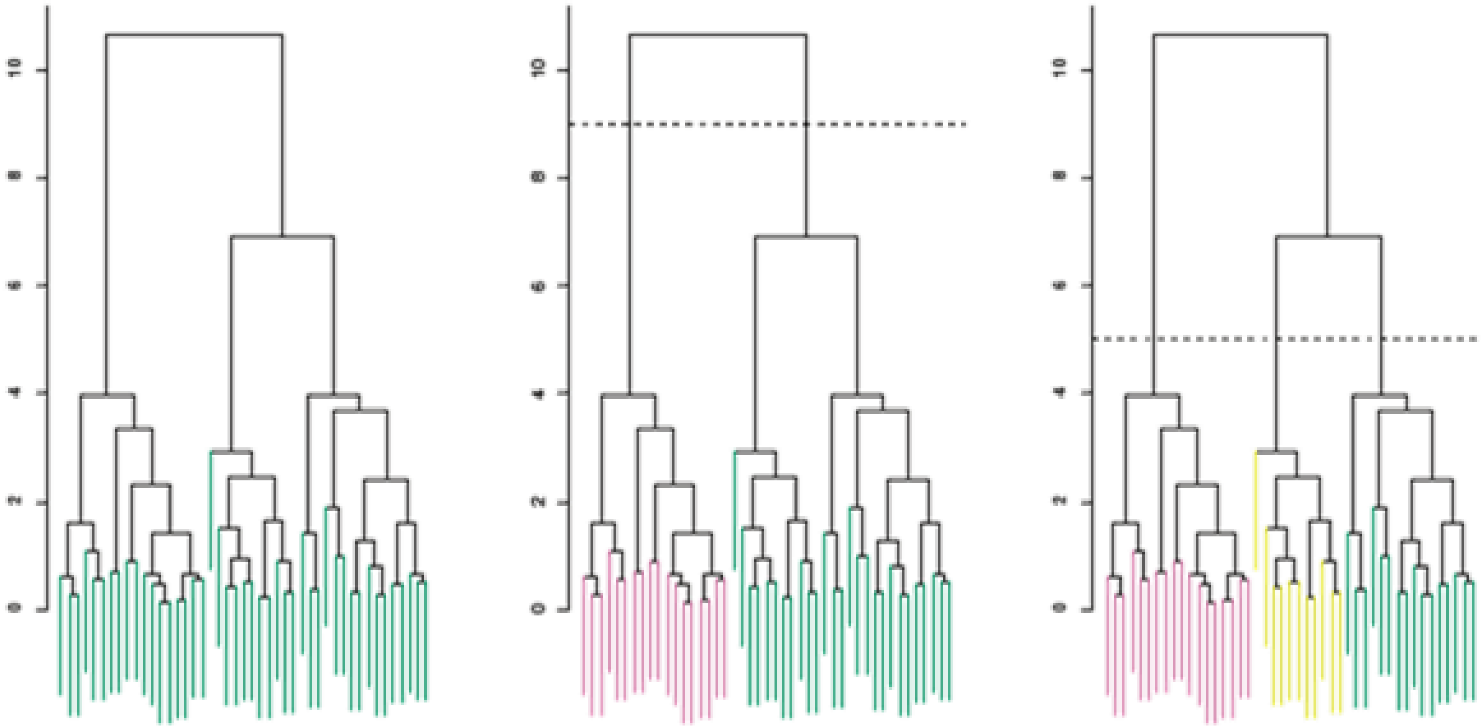


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Hierarchical Clustering

Co-funded by the European Union

Builds a tree of clusters, revealing hierarchical relationships.

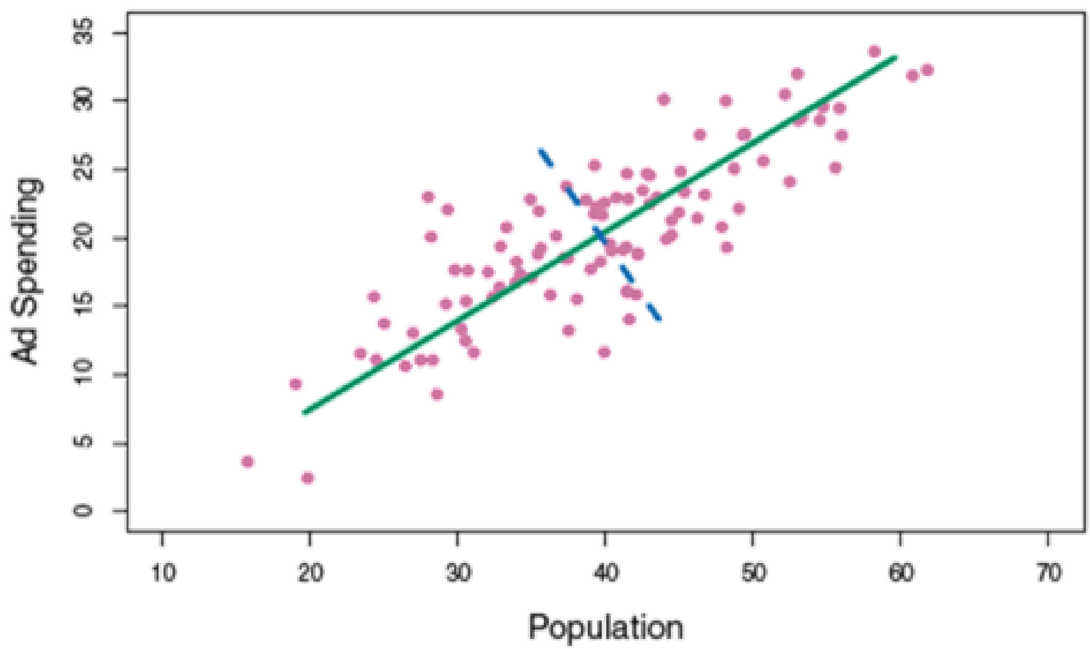


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Principal Component Analysis (PCA)

Co-funded by the European Union

- **Objective**: Reduce the dimensionality of a data set while retaining those characteristics that contribute most to its variance.
- **Process**: Transforms the original variables into a new set of variables, which are linear combinations of the original variables.
- **Utility**: Helpful in exploratory data analysis, visualization, and speeding up machine learning algorithms on large datasets.



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

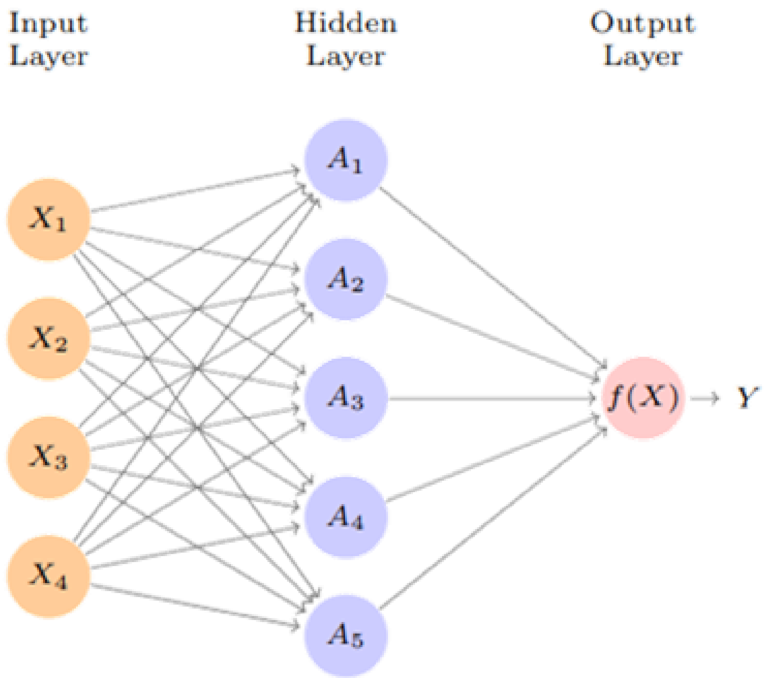
Deep Learning

- **Definition**: Advanced neural networks with multiple layers that learn from vast amounts of data.
- **Common Methods**:
  - **Convolutional Neural Networks (CNNs)**: Ideal for image and video recognition.
  - **Recurrent Neural Networks (RNNs)**: Effective for sequence prediction like text or speech.
  - **Autoencoders**: For data compression and denoising.

Neural Networks

Co-funded by the European Union

Neural networks are composed of multiple layers of interconnected nodes (neurons) that attempt to simulate the behavior of the human brain in order to "learn" from large amounts of data.

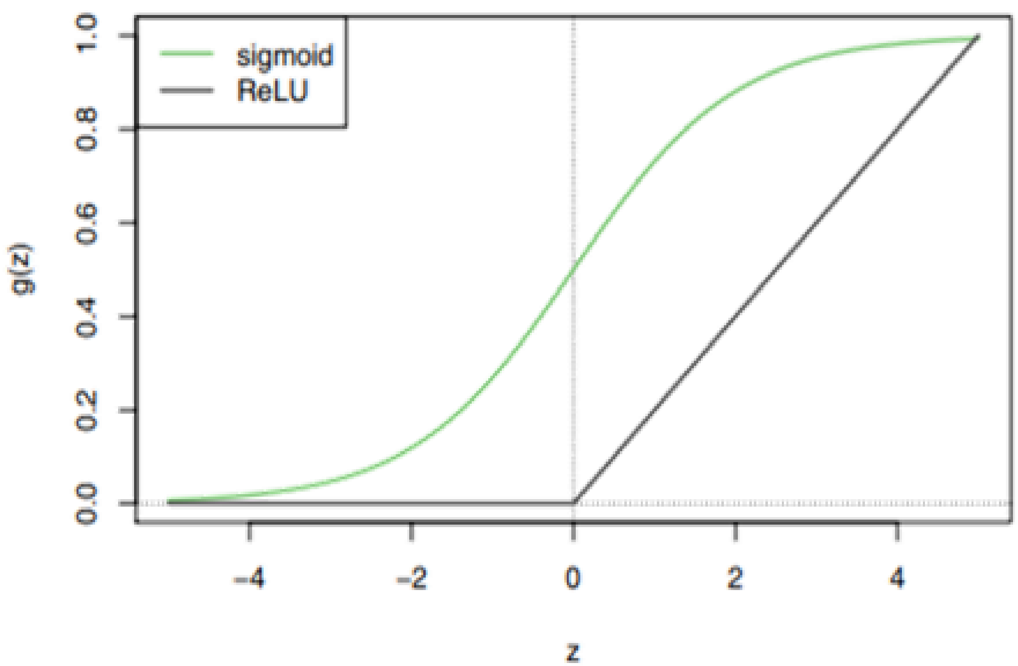


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Neural Networks – activation function

Co-funded by the European Union

The sigmoid and ReLU activation functions

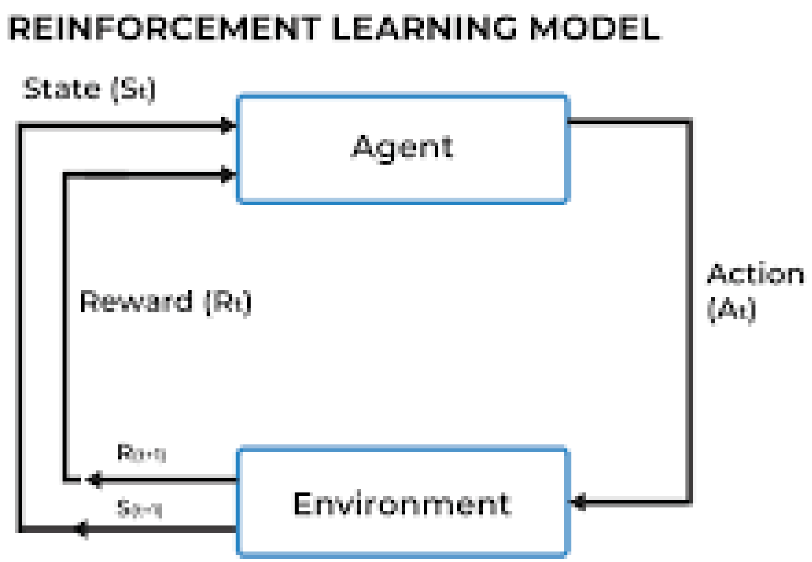


Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Reinforcement Learning

Co-funded by the European Union

- **Definition**: Learning optimal actions through trial and error to maximize a cumulative reward.
- **Objective**: Learn to make decisions by performing actions and receiving feedback.
- **Mechanism**: Involves an agent, a set of states, and actions leading to rewards or penalties.
- **Use Case**: Learning optimal strategies in games, robotic navigation, and real-time decision making.



Common Tools and Libraries for Machine Learning

Co-funded by the European Union

BEDigital

- **Python Ecosystem**:
  - **Scikit-learn**: General ML tasks.
  - **TensorFlow**: Large-scale deep learning.
  - **PyTorch**: Dynamic neural network training.
- **R Ecosystem**:
  - **Caret**: Streamlined model training.
  - **RandomForest**: For ensemble methods.
  - **Real-time decisions**: In environments like stock trading or autonomous vehicles.
- **Java/Scala**:
  - **Weka**: Data mining with easy access to algorithms.
  - **DL4J**: Deep learning in a JVM environment.

Choosing the Right ML Technique

Co-funded by the European Union

- **Factors to Consider**:
  - **Nature of the data**: Supervised vs. unsupervised.
  - **Problem complexity**: Need for simple models or deep learning.
  - **Computational resources**: Feasibility of training complex models.