Introduction To Machine Learning

Alex de Sá

a.desa@uq.edu.au
@alexgcsa

Plan

1st Part: Introduction to ML

2nd Part: Practical ML with Python

- Synthetic example (Classification)
- Real-world example (Classification and Regression)

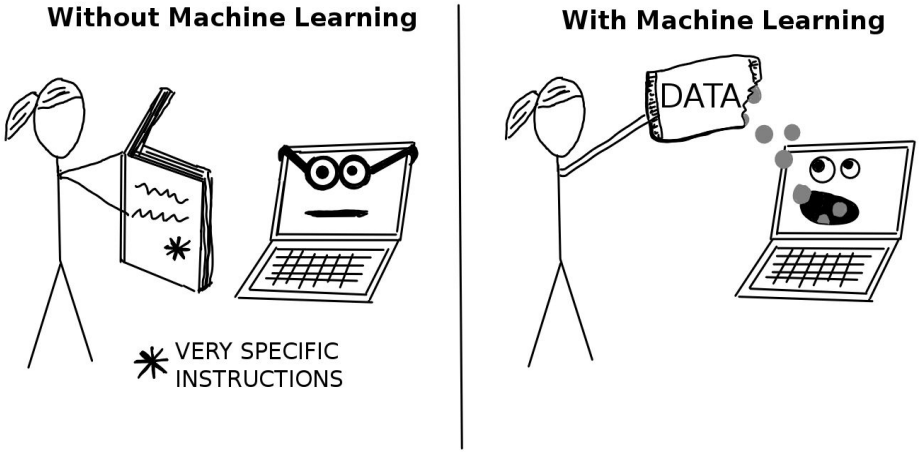
Highlight

You can interrupt me **at ANY time** if something is not clear

Machine Learning

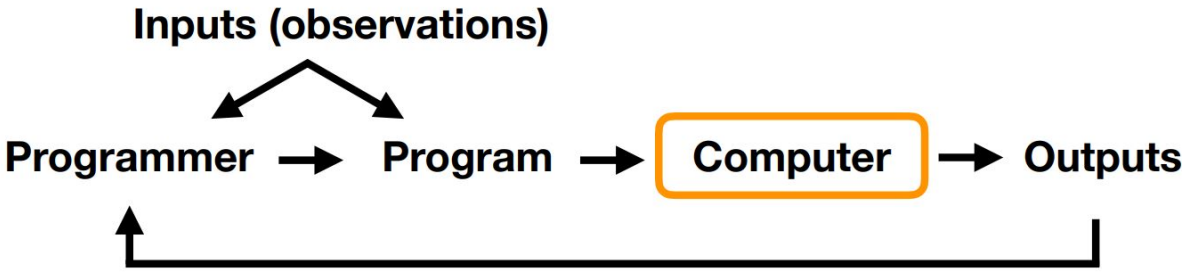
Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Arthur L. Samuel, AI pioneer, 1959



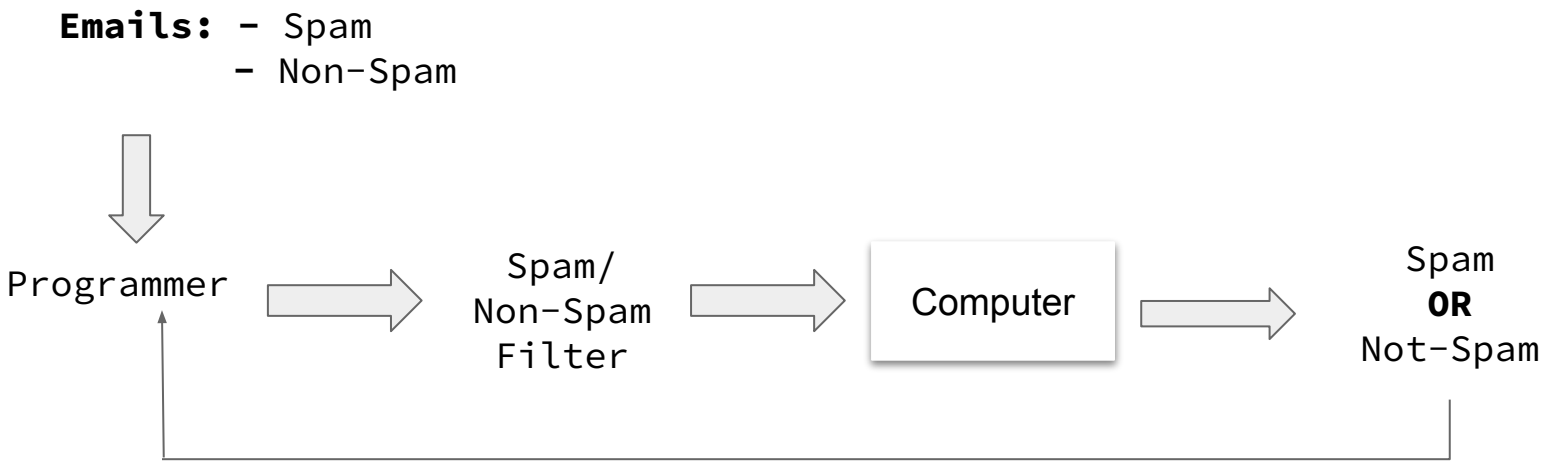
Source: Molnar, 2021

Traditional Programming versus Machine Learning



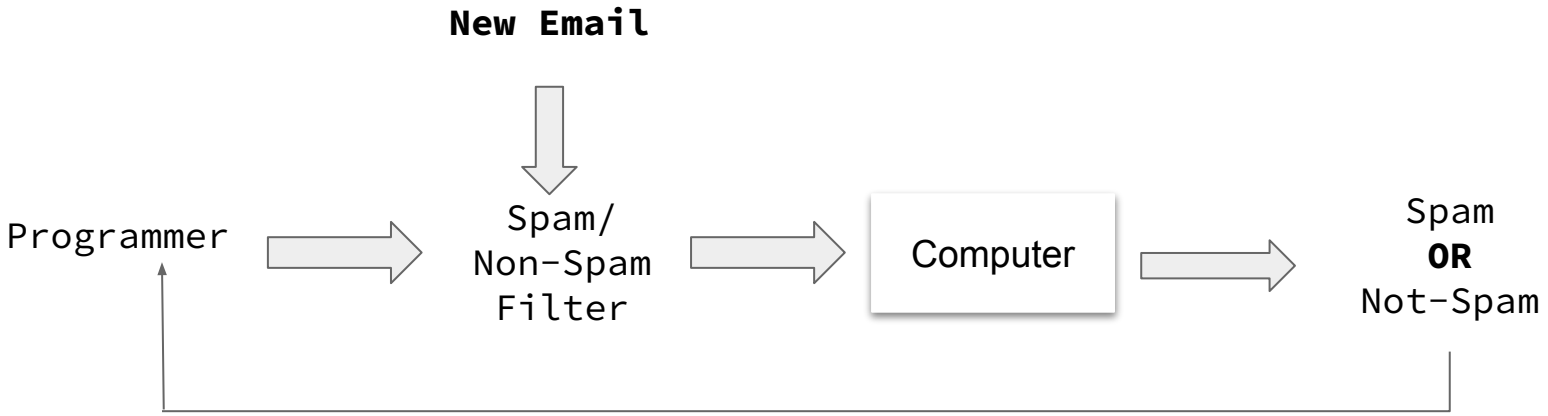
Raschka, 2021

Traditional Programming



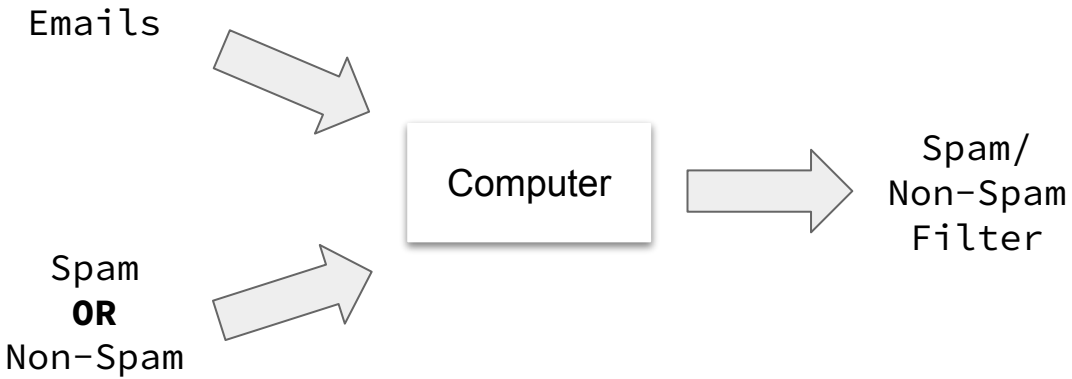
Raschka, 2021

Traditional Programming



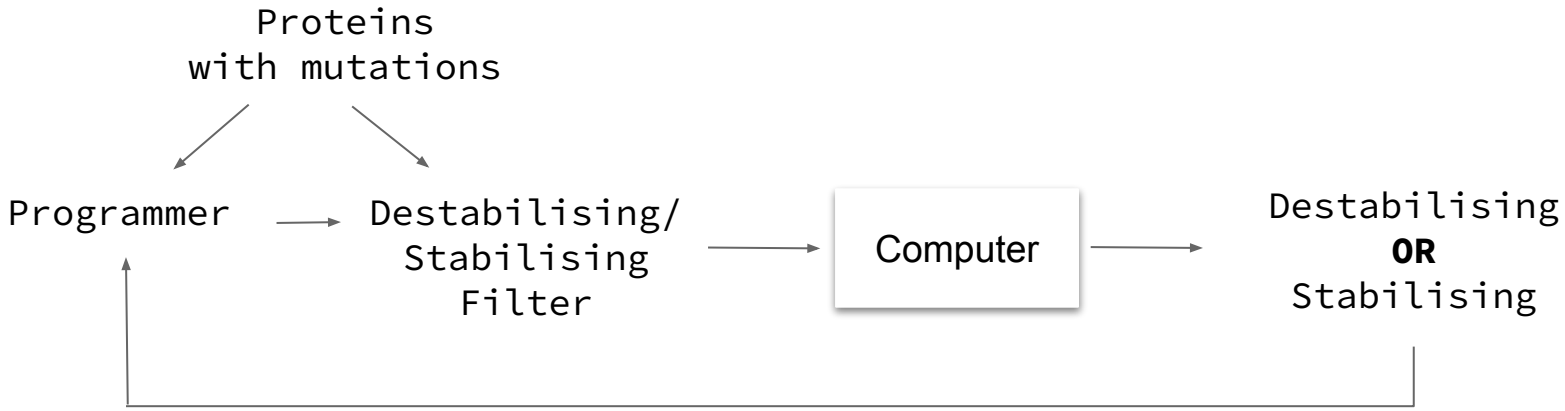
Raschka, 2021

Machine Learning



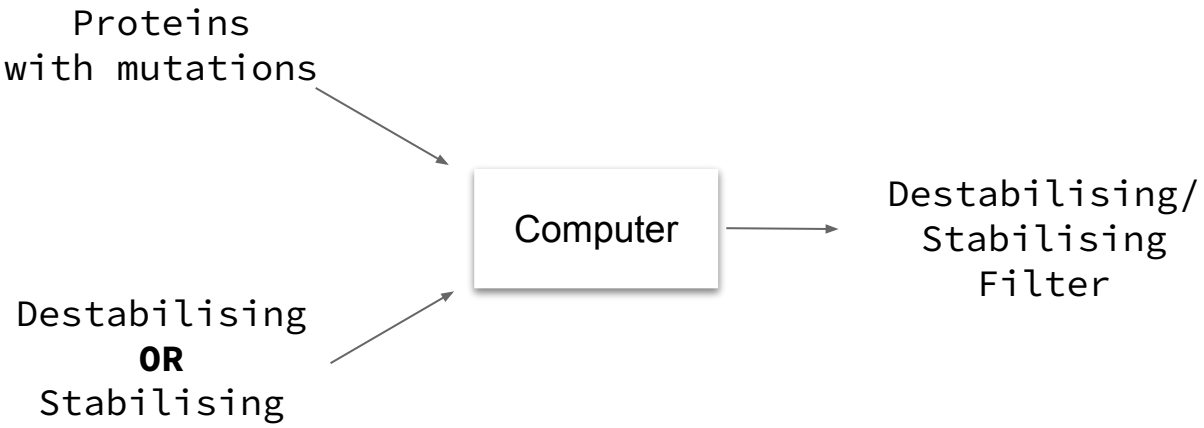
Raschka, 2021

Traditional Programming



Raschka, 2021

Machine Learning

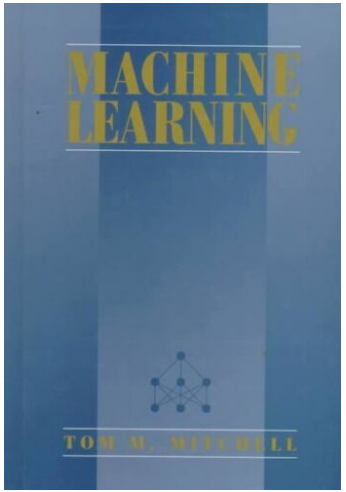


Raschka, 2021

Machine Learning Definition

A computer program is said to **learn** from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.”

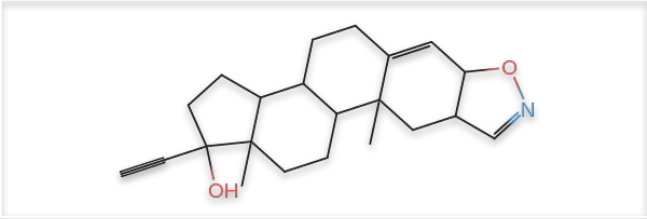
Tom Mitchell, Professor at Carnegie Mellon University

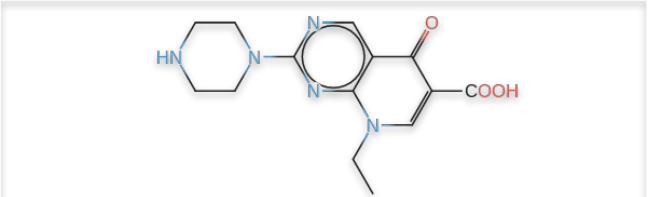


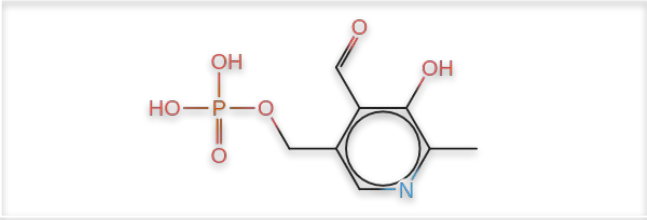
https://www.cs.cmu.edu/~tom/mlbook.html

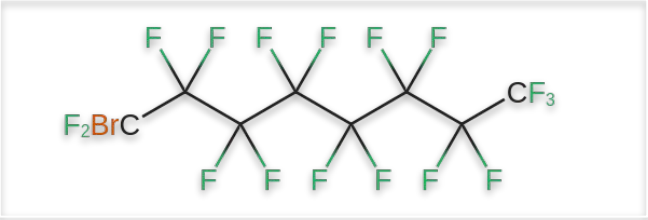
Example - Machine Learning Definition

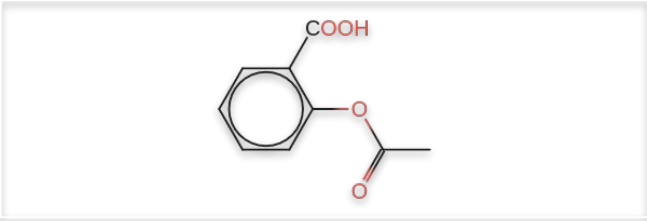
Identification of Hepatotoxicity in Small Molecules

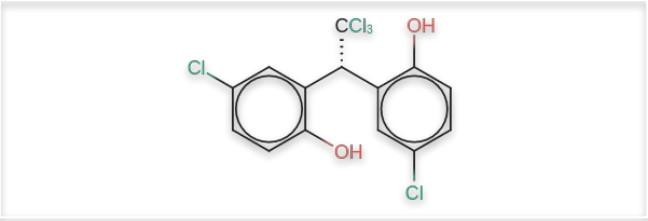




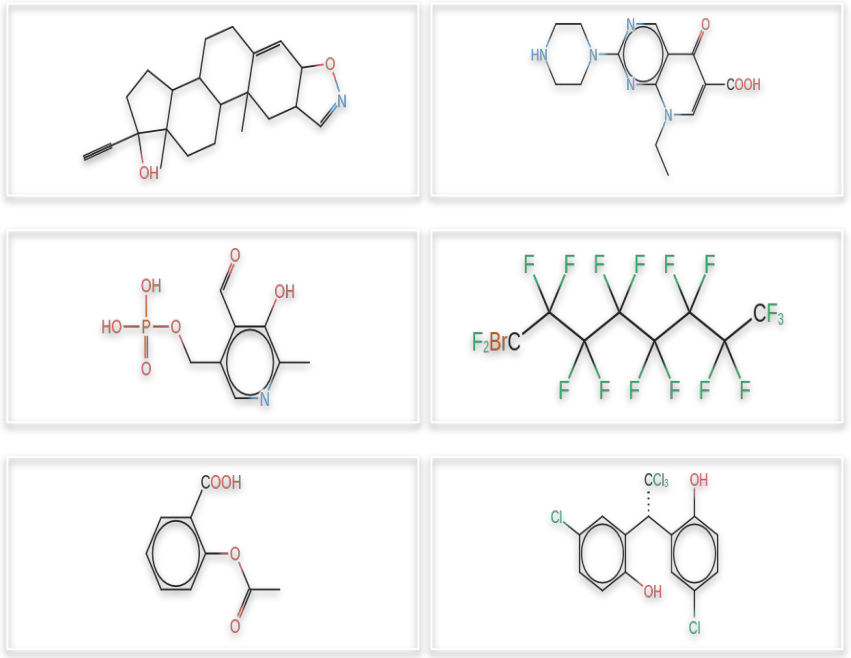






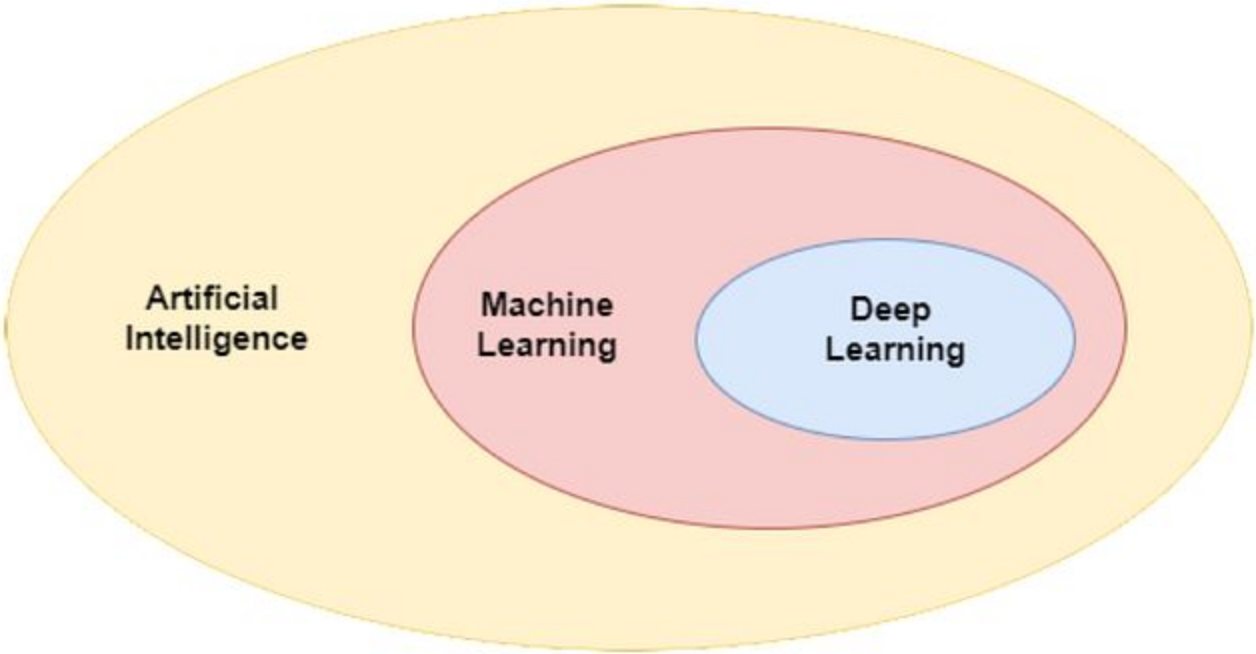


Identification of Hepatotoxicity in Small Molecules



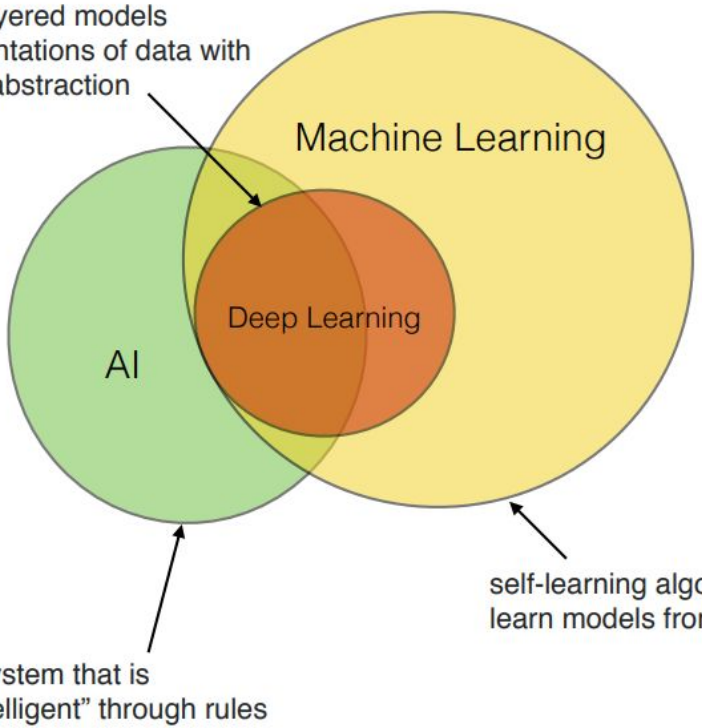
- **Task**: predicting hepatotoxicity from small molecules.
- **Performance**: percentage of molecules classified correctly as toxic.
- **Experience**: dataset of small molecules experimentally distinguishing them between toxic and non-toxic for the liver.

Taxonomy of Machine Learning



Towards Data Science, 2020

Taxonomy of Machine Learning



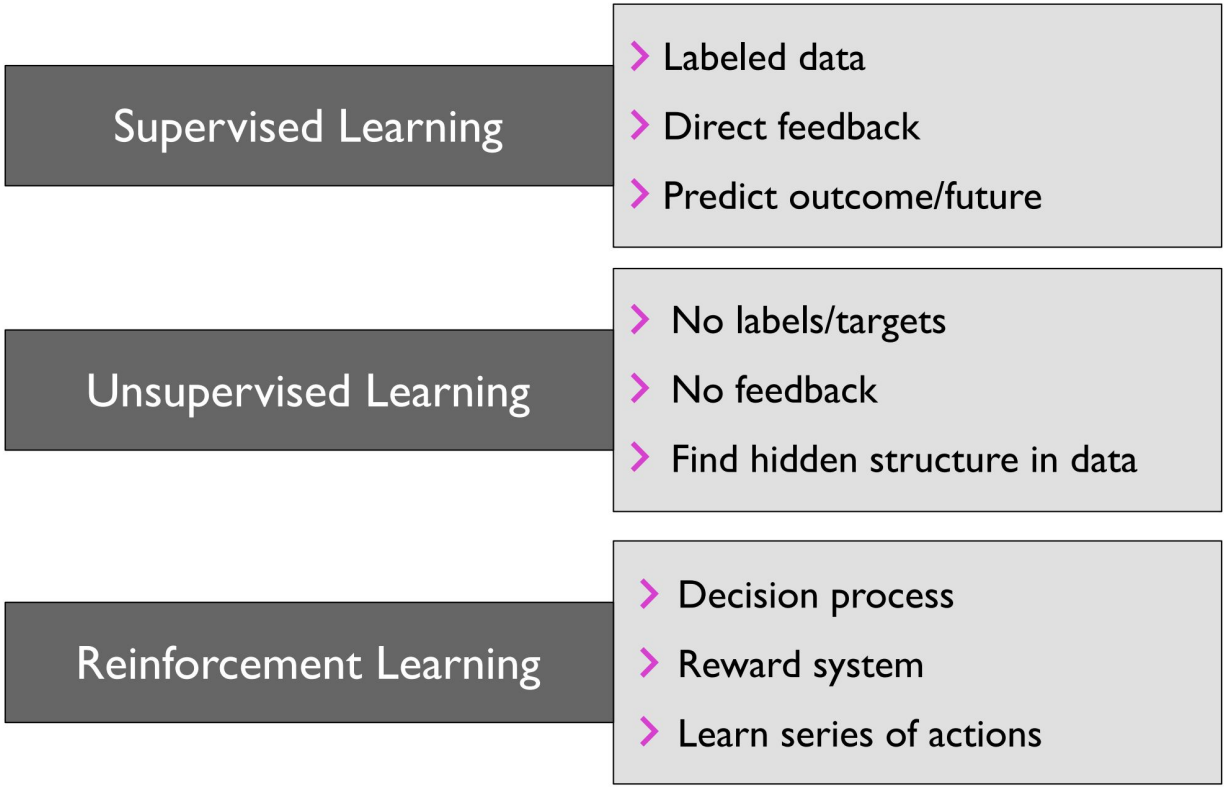
AI only: Using symbolic logic and knowledge representation for reasoning.

Human-based rules.

ML only: Predicting Flower Species

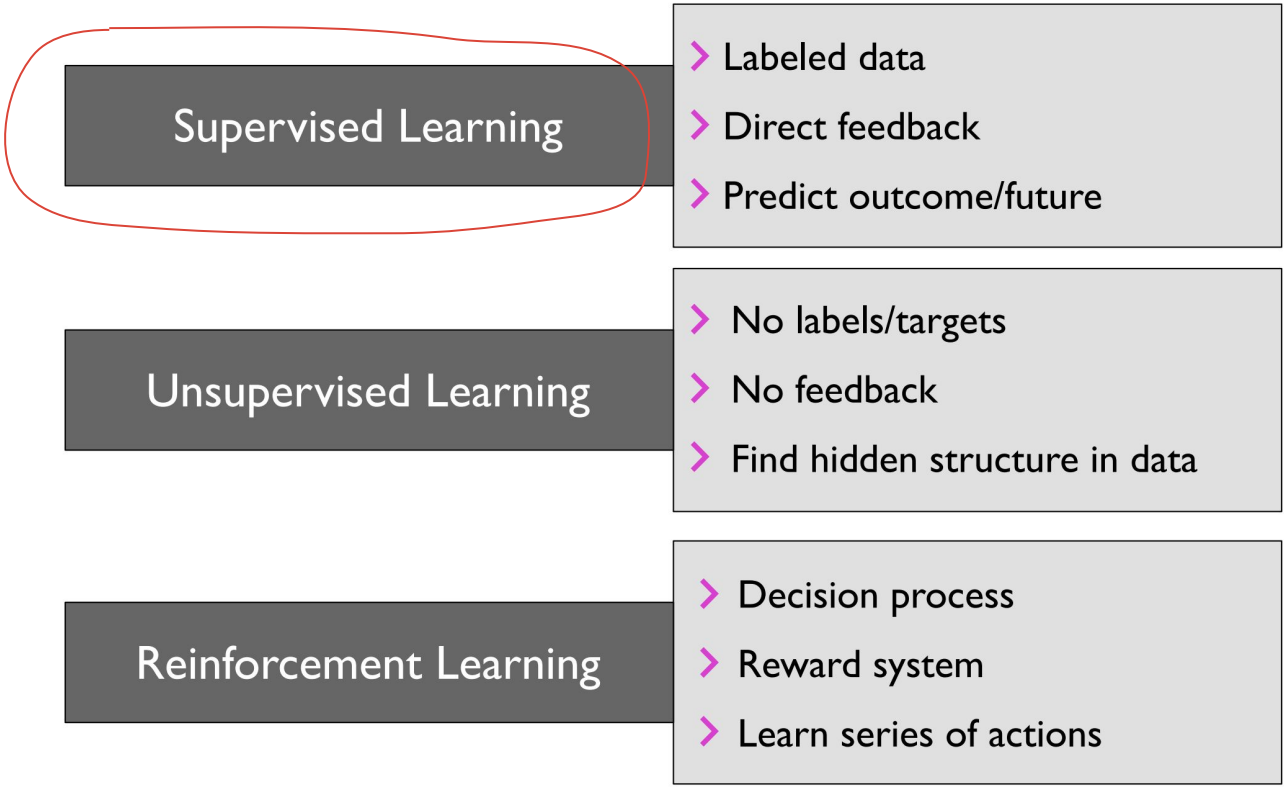
Raschka, 2021

Taxonomy of Machine Learning



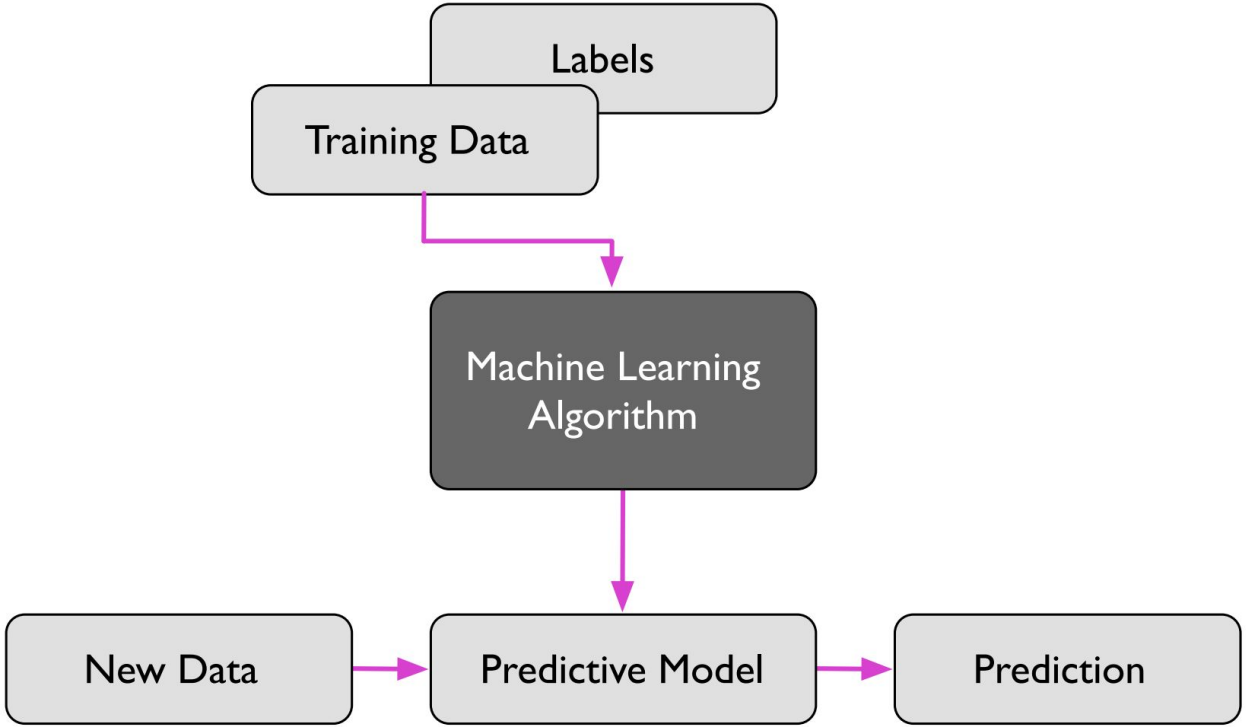
Raschka, 2021

Taxonomy of Machine Learning



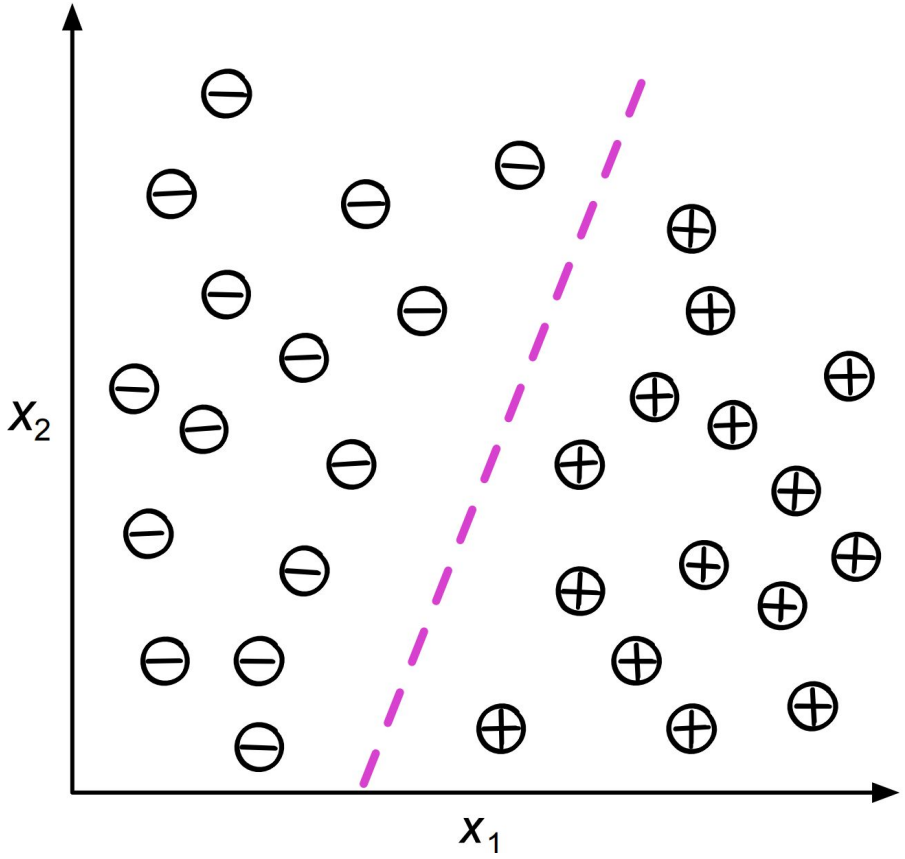
Raschka, 2021

Supervised Learning



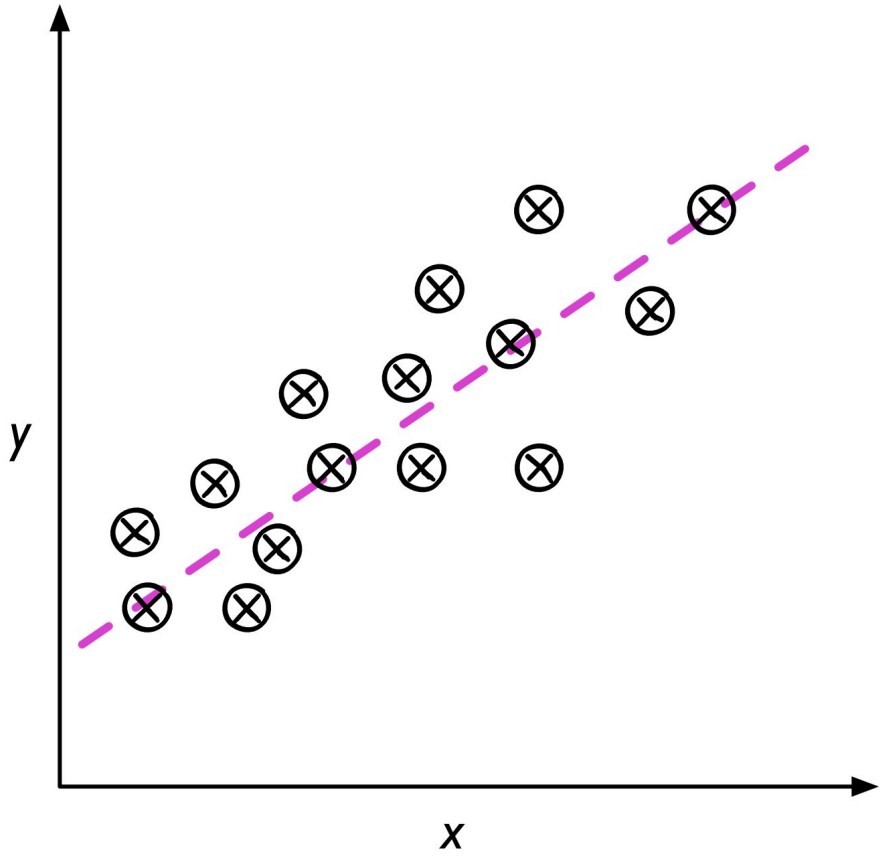
Raschka, 2021

Supervised Learning - Classification



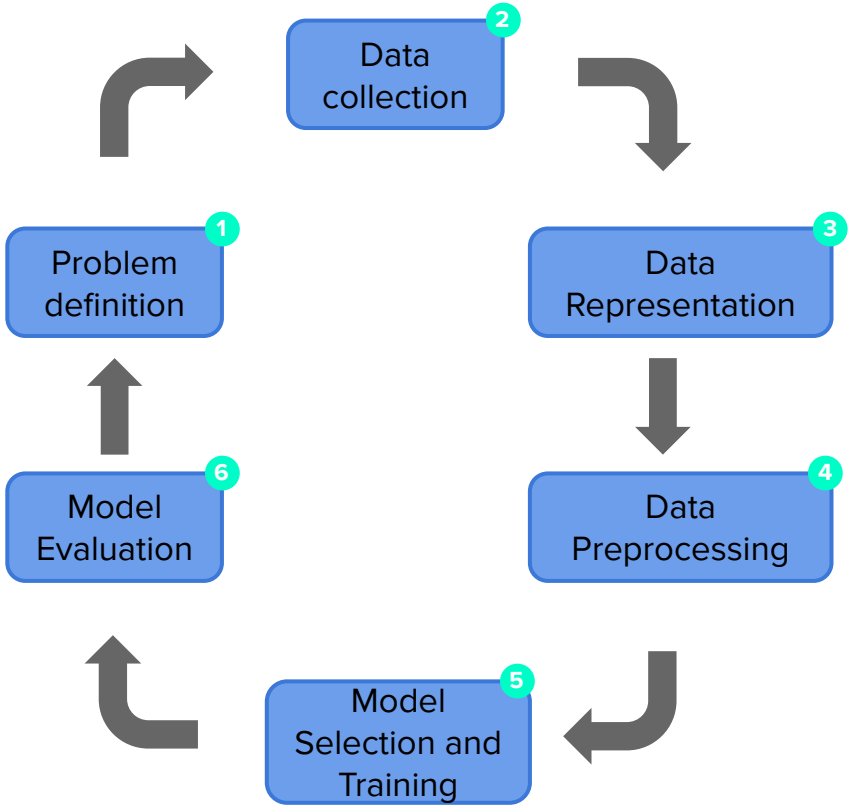
Raschka, 2021

Supervised Learning - Regression



Raschka, 2021

Machine Learning Workflow



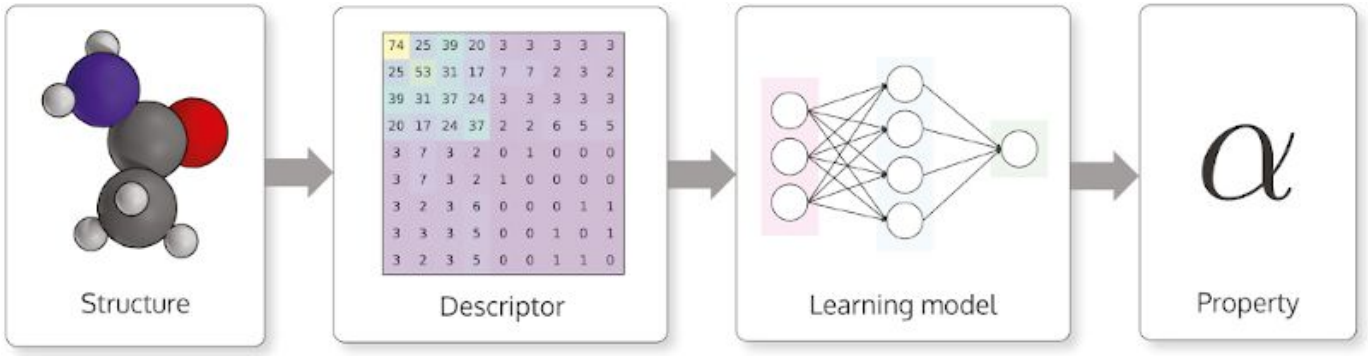
Data Representation

- Most of the recent Machine Learning tools (e.g., scikit-learn, etc) only accept numerical matrices (or dataframes) as inputs.

- We need to find ways to represent our biological, chemical, human data, etc in a numerical way.

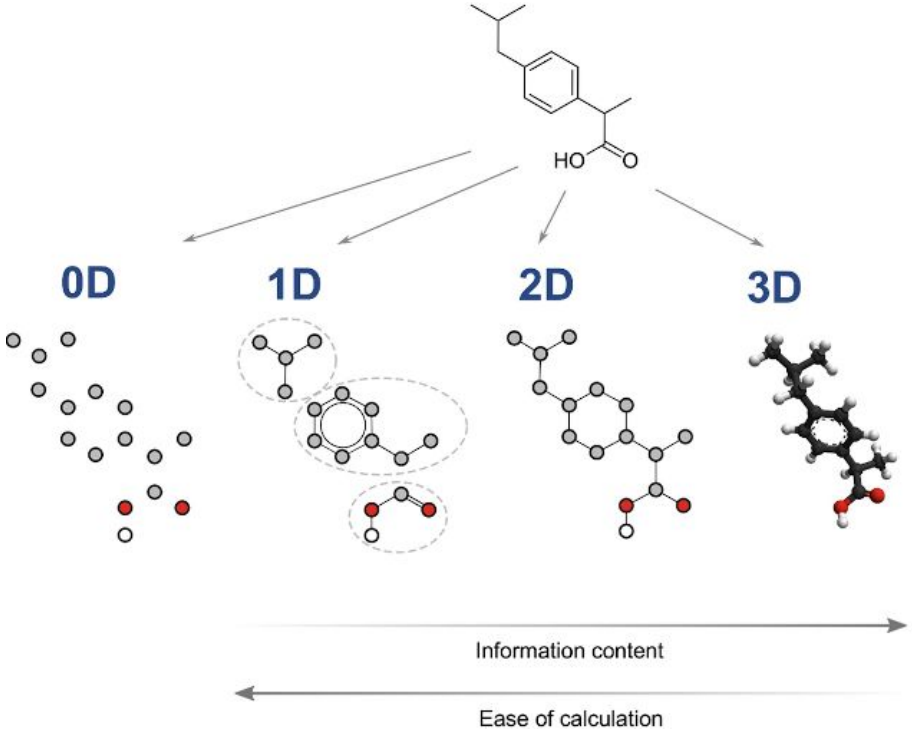
Data Representation - Small Molecules

Given the structure of the molecule, we are able to derive a list of descriptors aiming to characterise this input molecule to predict a given property



Chem Intelligence, 2021

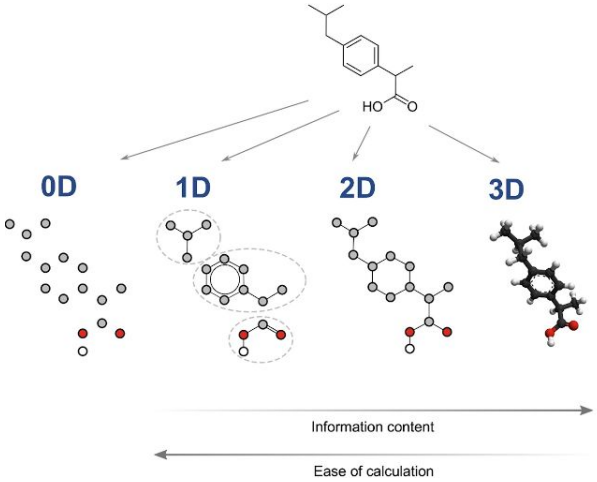
Data Representation - Small Molecules



Chem Intelligence, 2021

Data Representation - Small Molecules

- **0D Descriptors:** No information about structure and connectivity.
  - Atom counts, or molecular weights
- **1D Descriptors:** Partial information about the structure and connectivity.
  - Fingerprints.
- **2D Descriptors:** Information on molecular topology based on the graph representation
  - Atom distance matrix.
- **3D Descriptors:** Information about the spatial coordinates of atoms of a molecule
  - 3D fingerprints.



Chem Intelligence, 2021

RDKit
Open-Source Cheminformatics and Machine Learning

Sequence-based descriptors

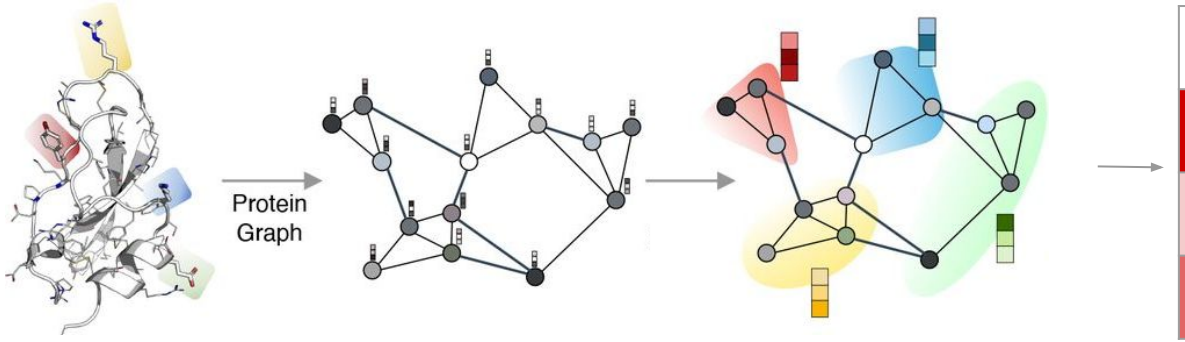
```
MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEEVWNIKQMIKLTQEH
IEALLDKFGGEHNPPSIYLEAYEEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTV
TSSSSSSLSVLPSSLSVFQNPTDVARSNPKSPQKPIVRVFLPNKQRTVVPARCGVTVRDS
LKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTGEELHVEVLENVPLTTHNFVRK
```

- Amino acid composition
- Physicochemical properties
- Disorder propensity scores
- ...

74	25	39	20	3	3	3	3	3
25	53	31	17	7	7	2	3	2
39	31	37	24	3	3	3	3	3
20	17	24	37	2	2	6	5	5
3	7	3	2	0	1	0	0	0
3	7	3	2	1	0	0	0	0
3	2	3	6	0	0	0	1	1
3	3	3	5	0	0	1	0	1
3	2	3	5	0	0	1	1	0

Data Representation - Proteins

- Structure-based descriptors:



D1	D2	…	D_{n-1}	D_n
1.0	100		1.5	1
0.1	1000		1.5	0
0.5	10		20.9	2

- residue depth
- solvent accessible surface area
- secondary structure distribution
- torsion angles
- Cumulative pair distances between pharmacophore groups

Data Representation - Proteins

- Structure-based descriptors:

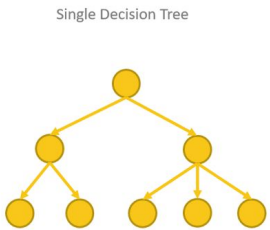
Protein ID	Feature 1	Feature 2	Feature 3	...	Feature M
Q5I0E9	0	5	0		0.150
O74717	10	3	1		0.164
P63033	5	2	1		0.457
...	7	4	1		0.973
Q8NIH1	0	12	0		0.667
D4APA9	14	3	1		0.011
O94864	2	9	1		0.199

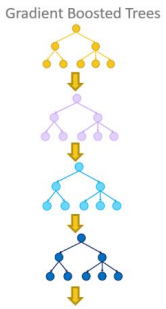
ML Algorithms

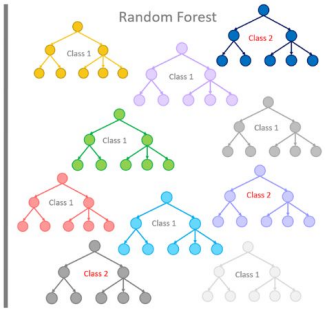
Several **Machine Learning (ML)** models with different characteristics:

- Deep learning models

- **Tree-based Models:**
  - Decision Trees
  - Random Forest
  - Extra Trees
  - eXtreme Gradient Boosting
  - Adaptive Boosting
  - Gradient Boosting

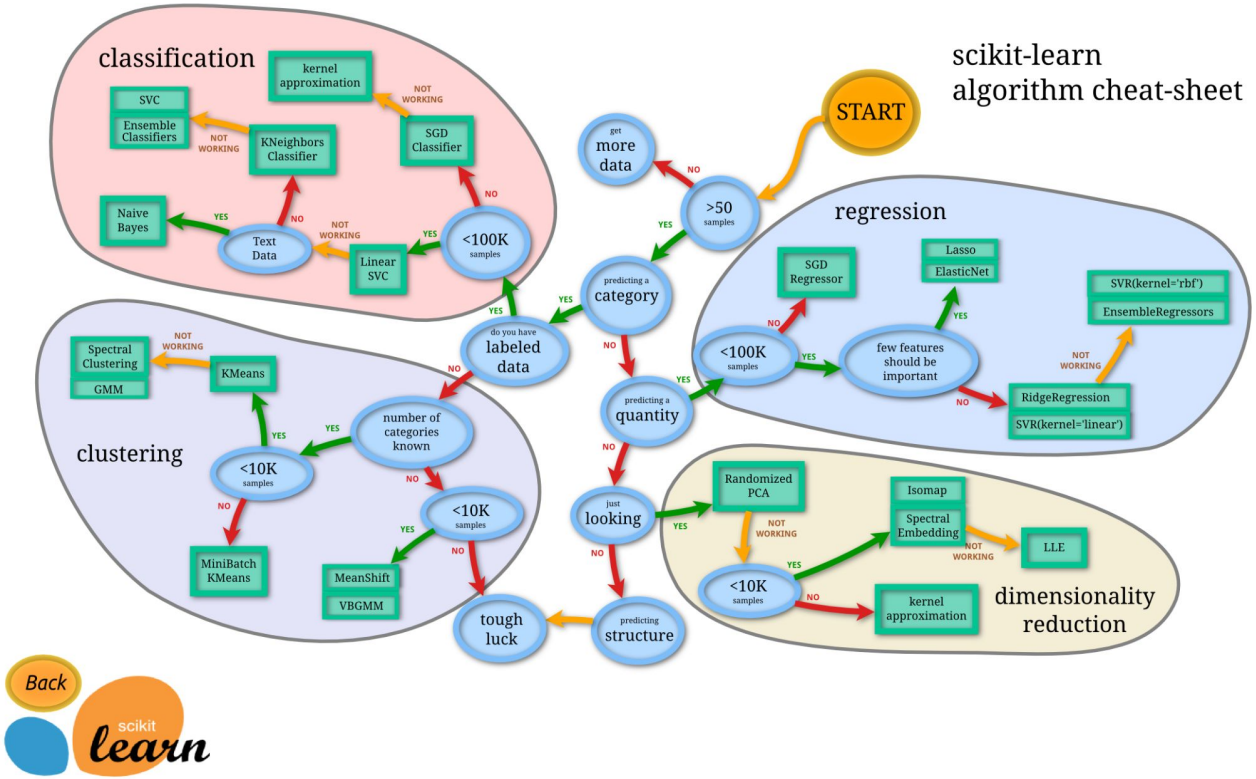






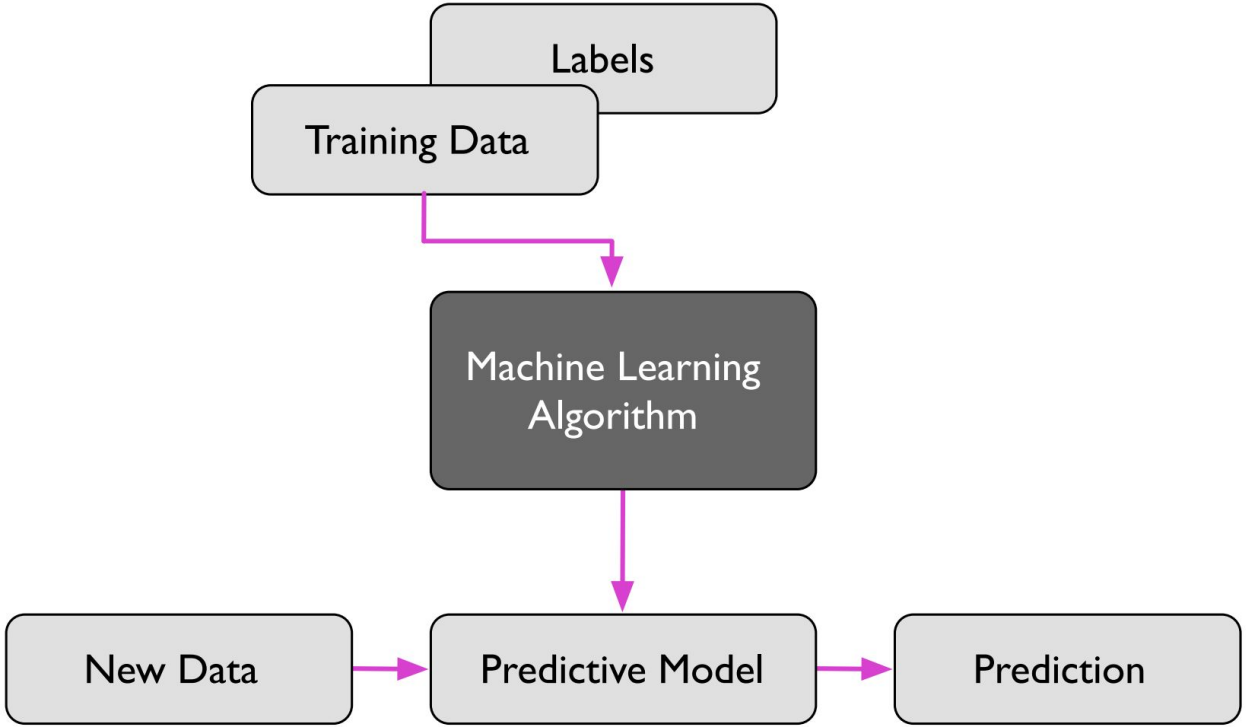
- **Support Vector Machines**
- Naïve Bayes
- ….

ML Algorithms



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Model Fitting

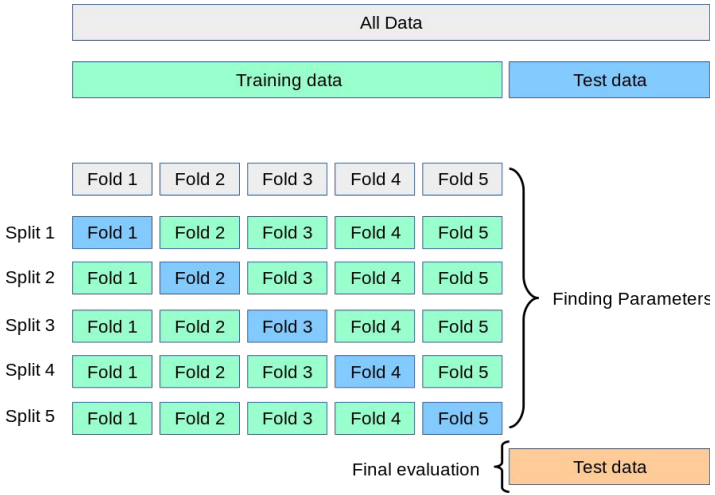


Raschka, 2021

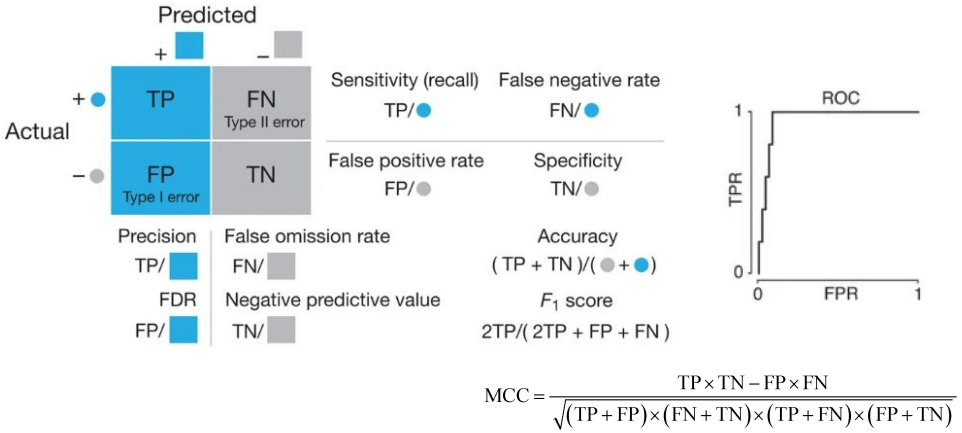
Selection and Evaluation of ML models

ML model's selection

Cross-Validation and Testing



Performance Evaluation for Classification Models

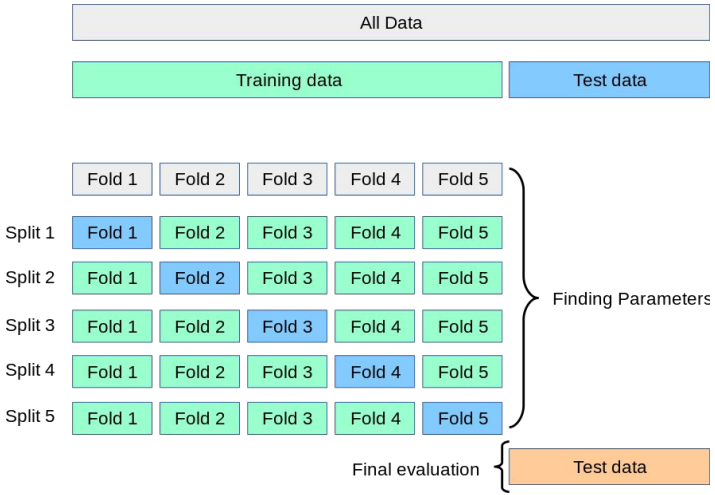


Lever et al., 2016

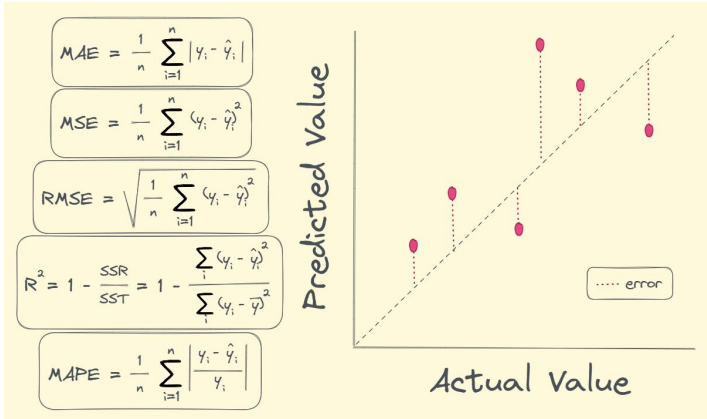
Selection and Evaluation of ML models

ML model's selection

Cross-Validation and Testing



Performance Evaluation for Regression Models

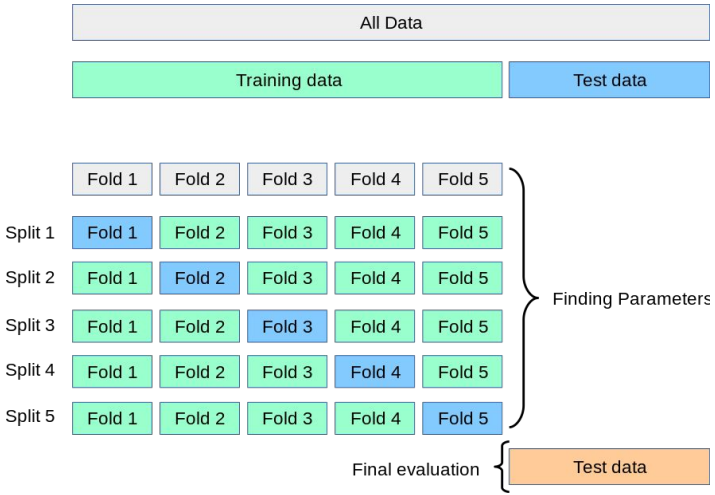


Dorfer, 2023

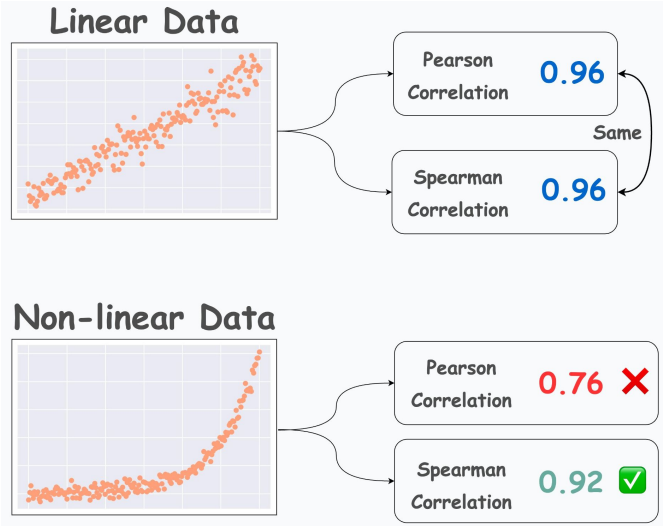
Selection and Evaluation of ML models

ML model's selection

Cross-Validation and Testing

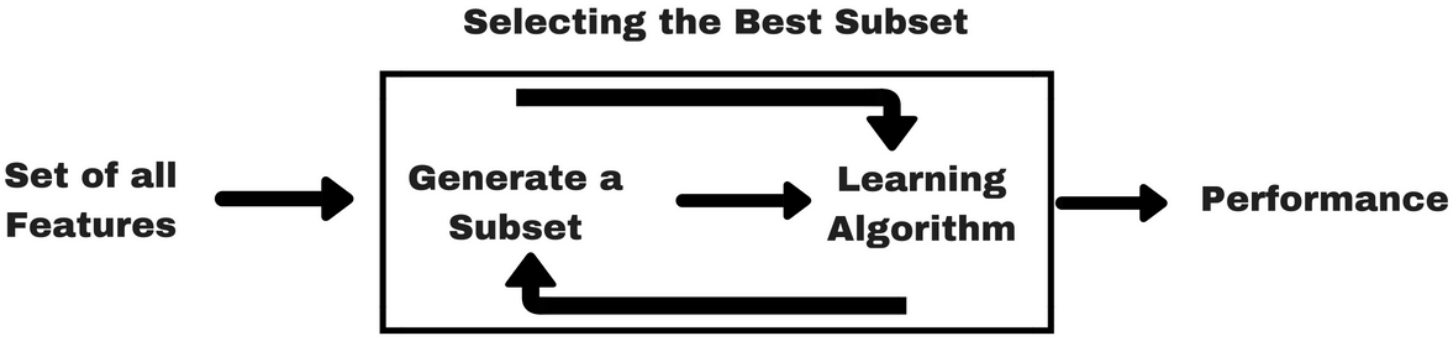


Performance Evaluation for Regression Models



Chawla, 2023

Machine Learning (ML) coupled with feature selection:

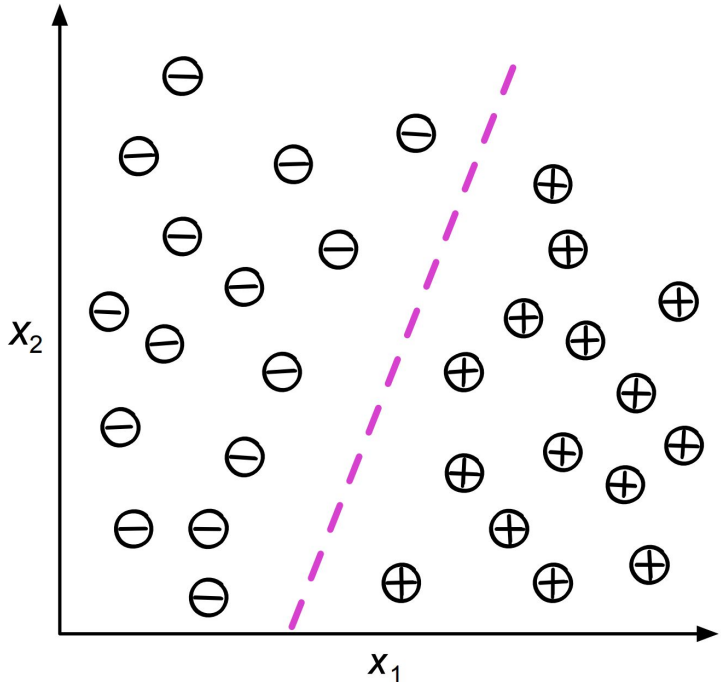


Analytics Vidhya

Focusing on Classification

Decision Boundary

Definition: A decision boundary, is a surface that separates data points belonging to different class labels. (Sahu, 2021)

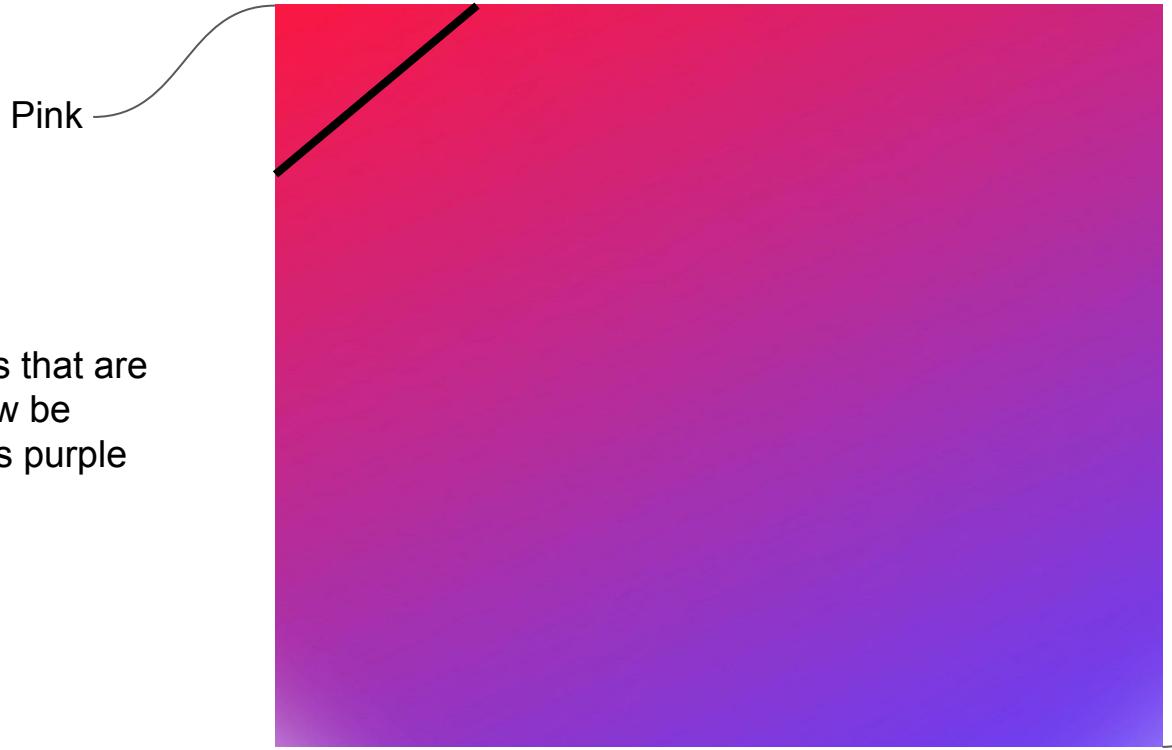


Decision Boundary

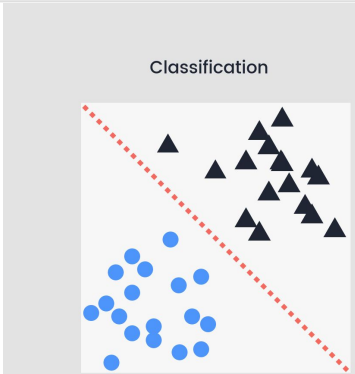
Binary
Classification

https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

Decision Boundary



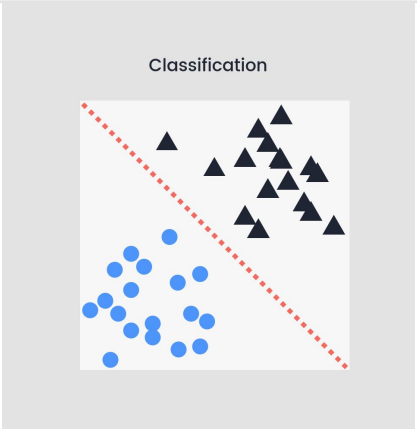
Many things that are pink will now be classified as purple



https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

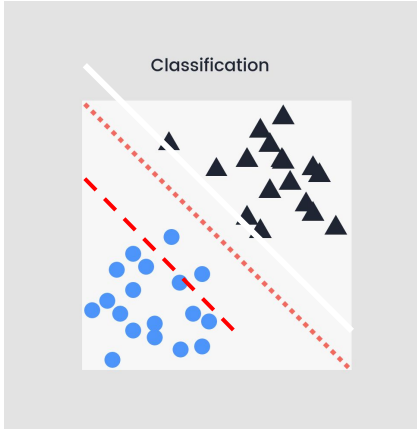
Decision Boundary

Many things that are pink will now be classified as purple



https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

Decision Boundary



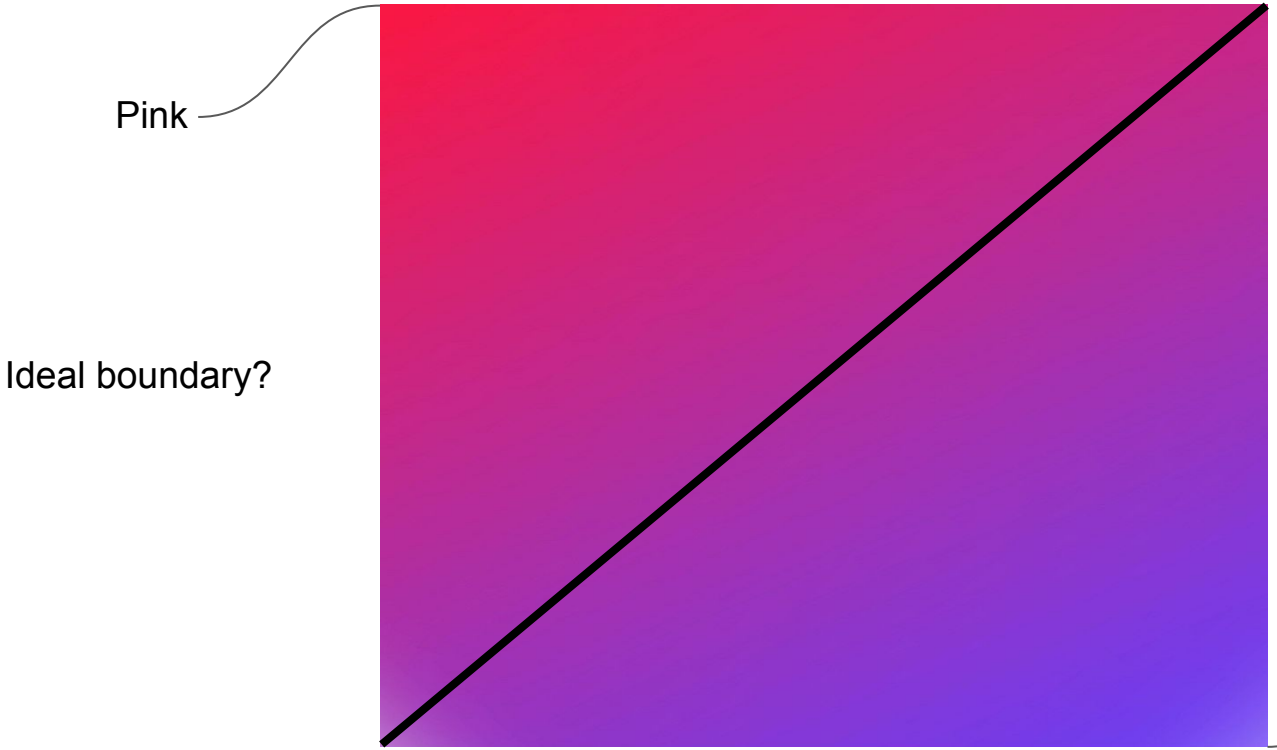
Pink

Purple

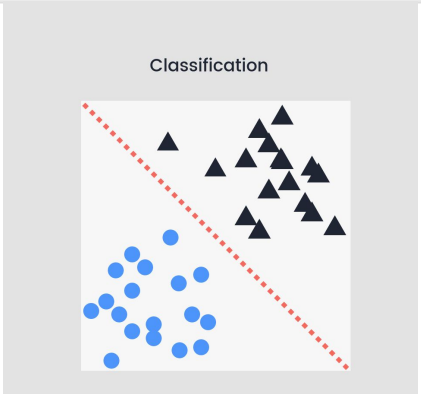
Many things that are purple will now be classified as pink

https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

Decision Boundary



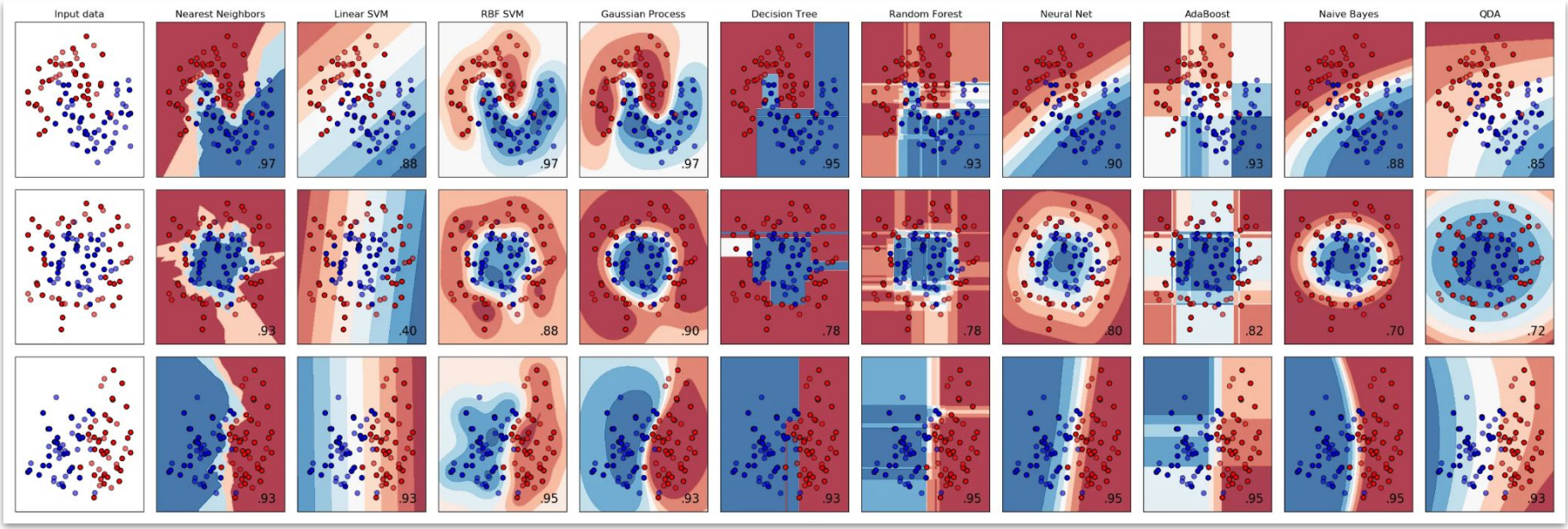
Ideal boundary?



https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

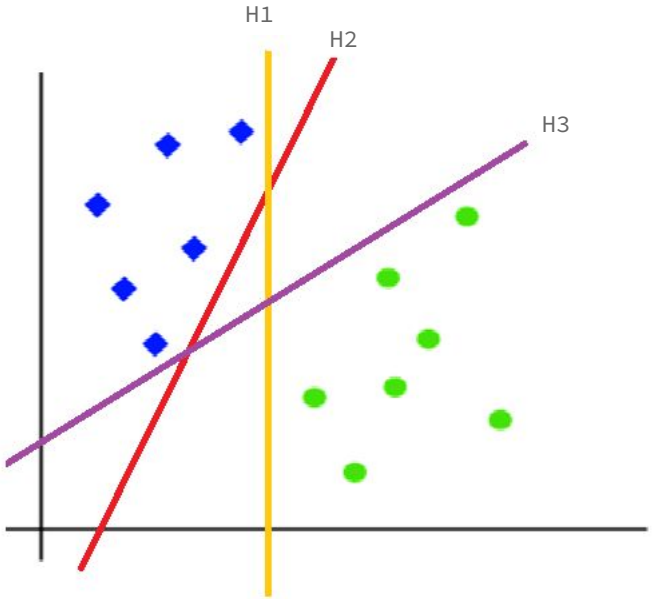
Decision Boundary

Comparison of the decision boundaries of 10 machine learning models:



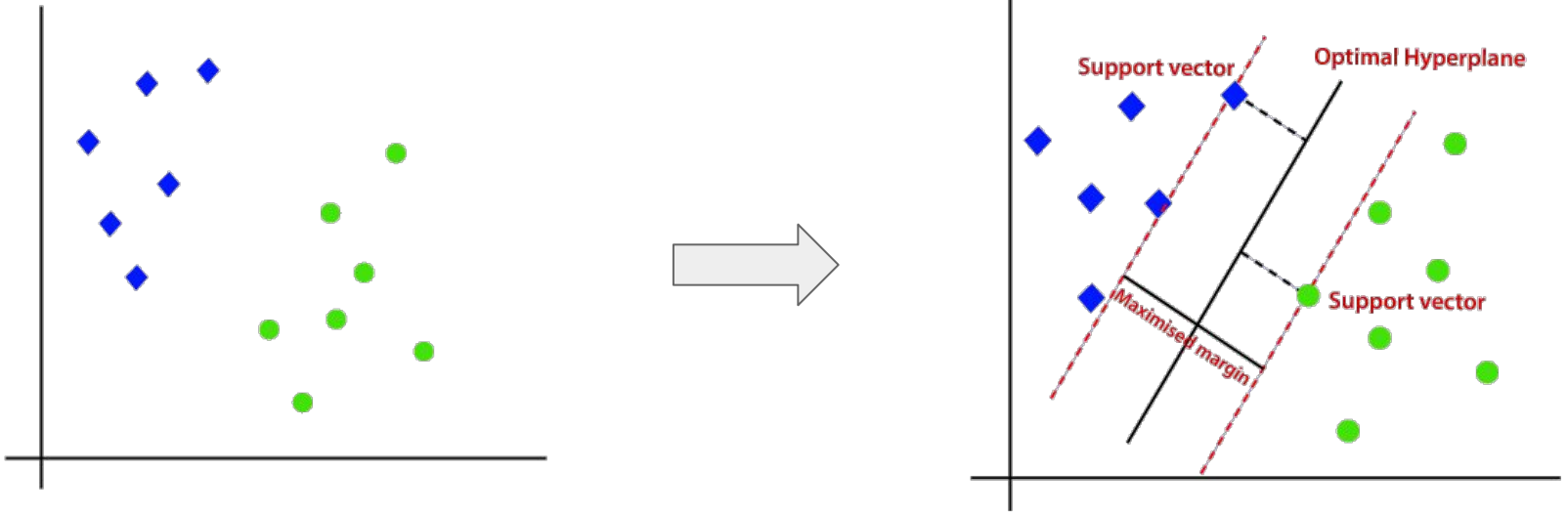
Varoquaux and Müller

Classical Support Vector Classifier (SVC)



Saini,2023

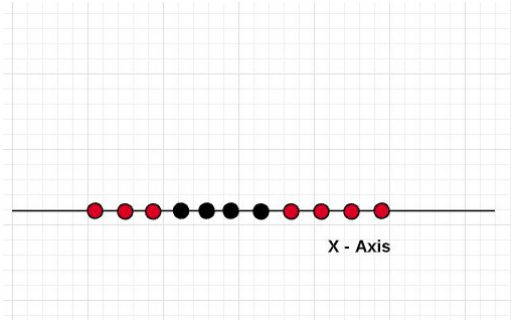
Classical Support Vector Classifier (SVC)



Saini,2023

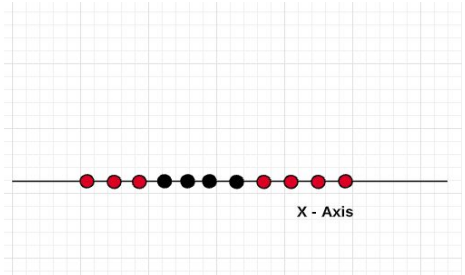
Classical Support Vector Classifier (SVC)

Support Vector Machines were developed to deal with linear data.
What happens when we take data like:



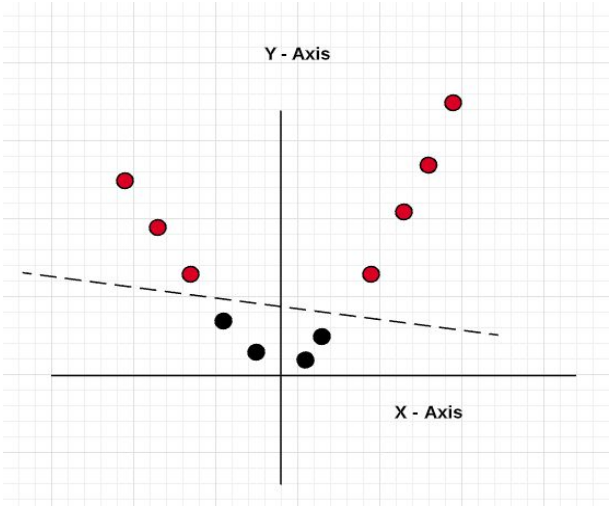
Non-linearly separable data

Classical Support Vector Classifier (SVC)



ϕ

$y = x^2$

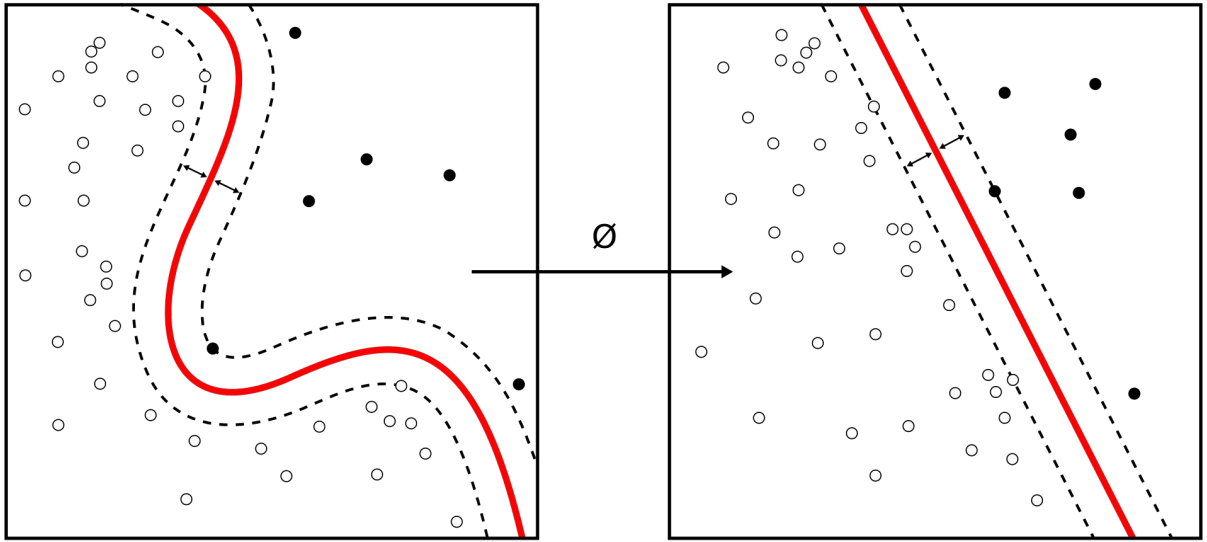


Non-linearly separable data

Support Vector Machines have a key component called **kernel machine.**

Classical Support Vector Classifier (SVC)

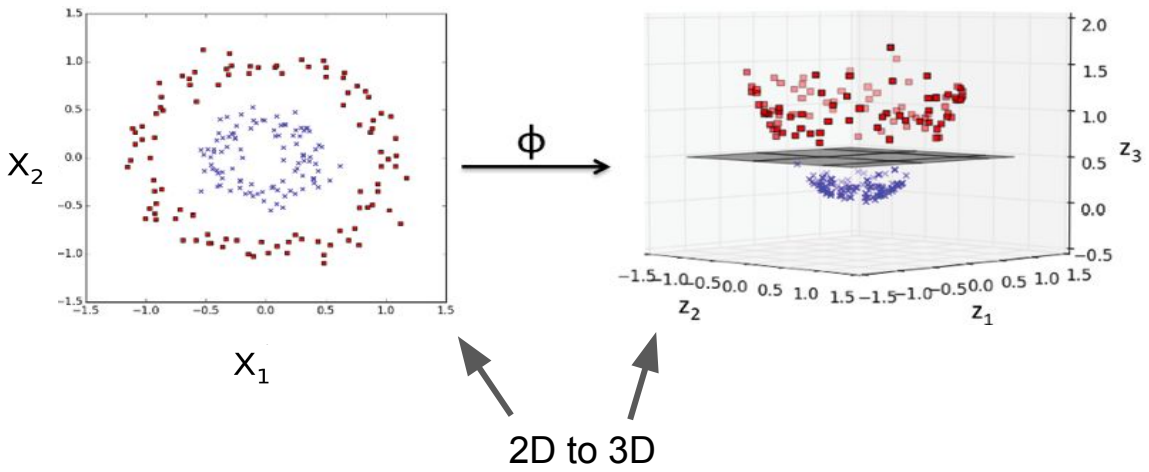
A **Kernel Function** manipulates the training data to transform a non-linear lower dimension space into a higher dimension space, which we can get a linear decision boundary



SVM - Wikipedia

Classical Support Vector Classifier (SVC)

A **Kernel Function** manipulates the training data to transform a non-linear lower dimension space into a higher dimension space, which we can get a linear decision boundary



SVM - Wikipedia

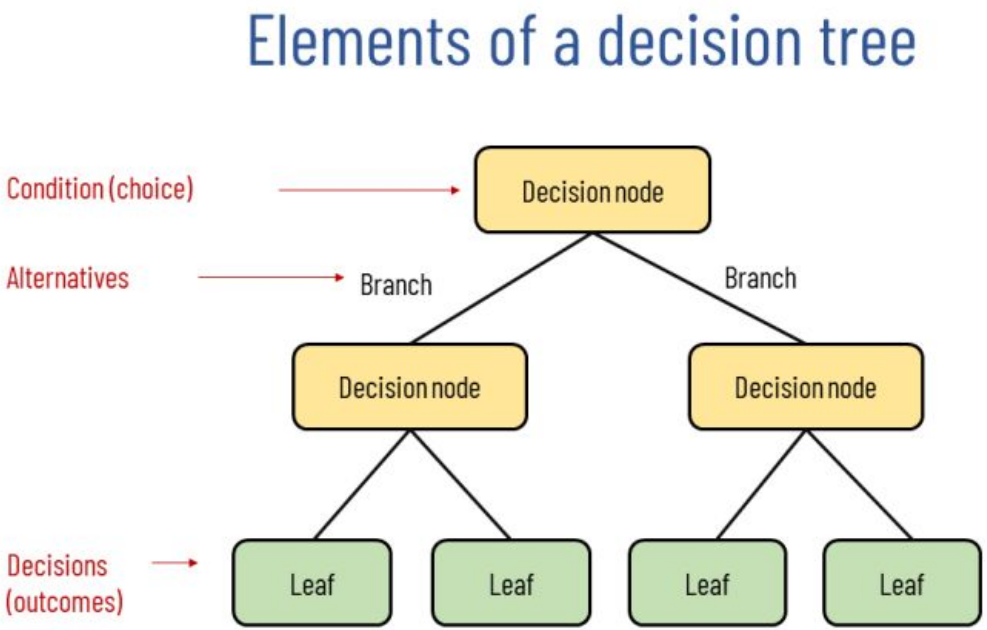
Part 1: Introduction to Support Vector Classifier (SVC)

https://github.com/alexgcsa/comp_life_sciences_2023

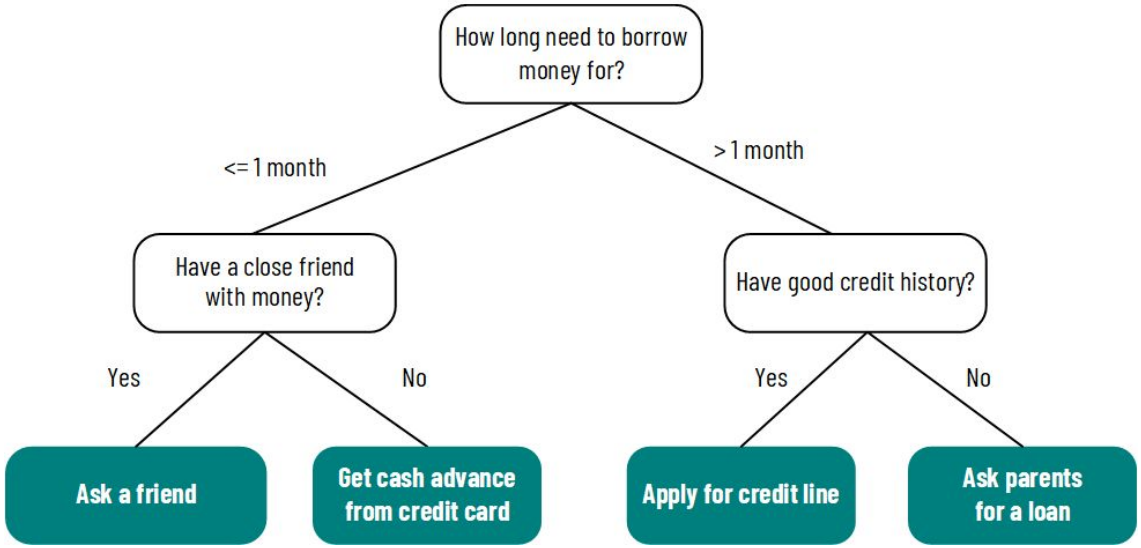
Part 2: Support Vector Classifier (SVC) for predicting small-molecule that inhibit protein-protein interactions

https://github.com/alexgcsa/comp_life_sciences_2023

Decision Trees



Kosarenko, 2021



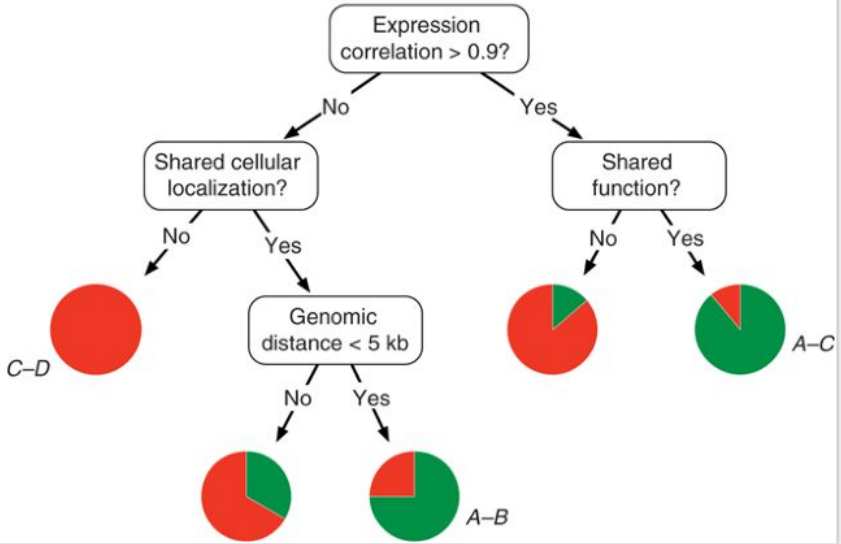
Kosarenko, 2021

Decision Trees

a

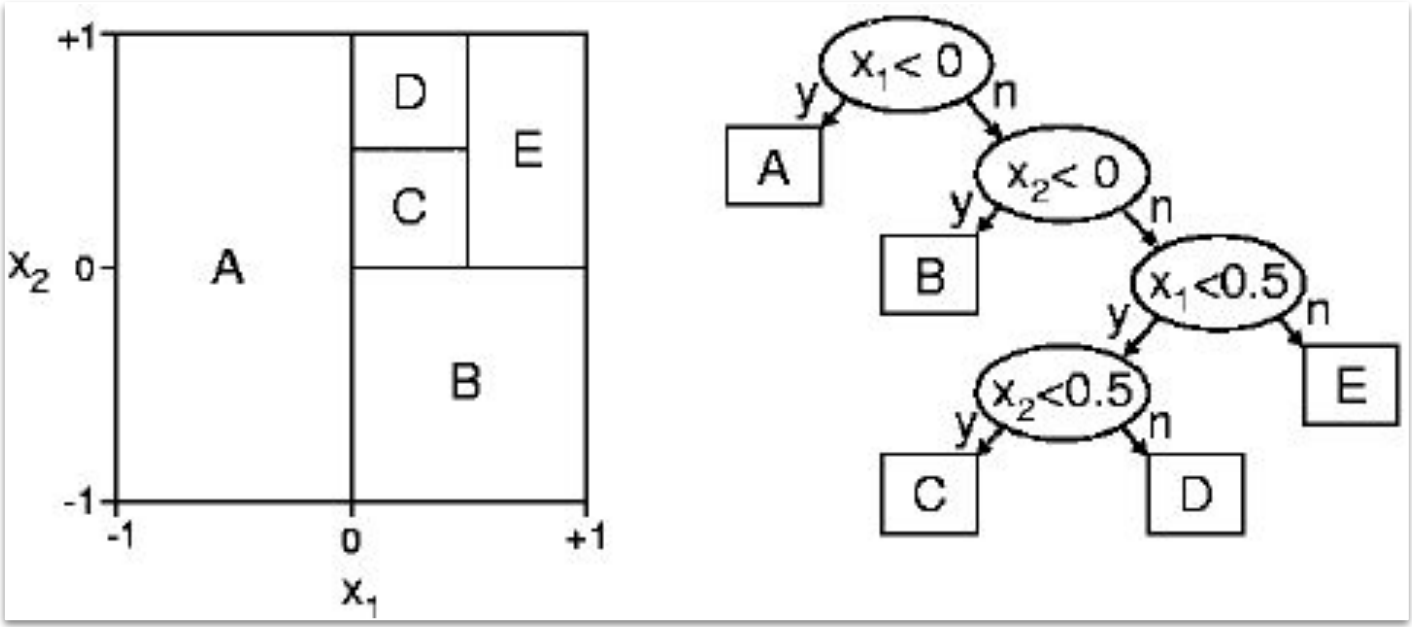
Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb
⋮					

b



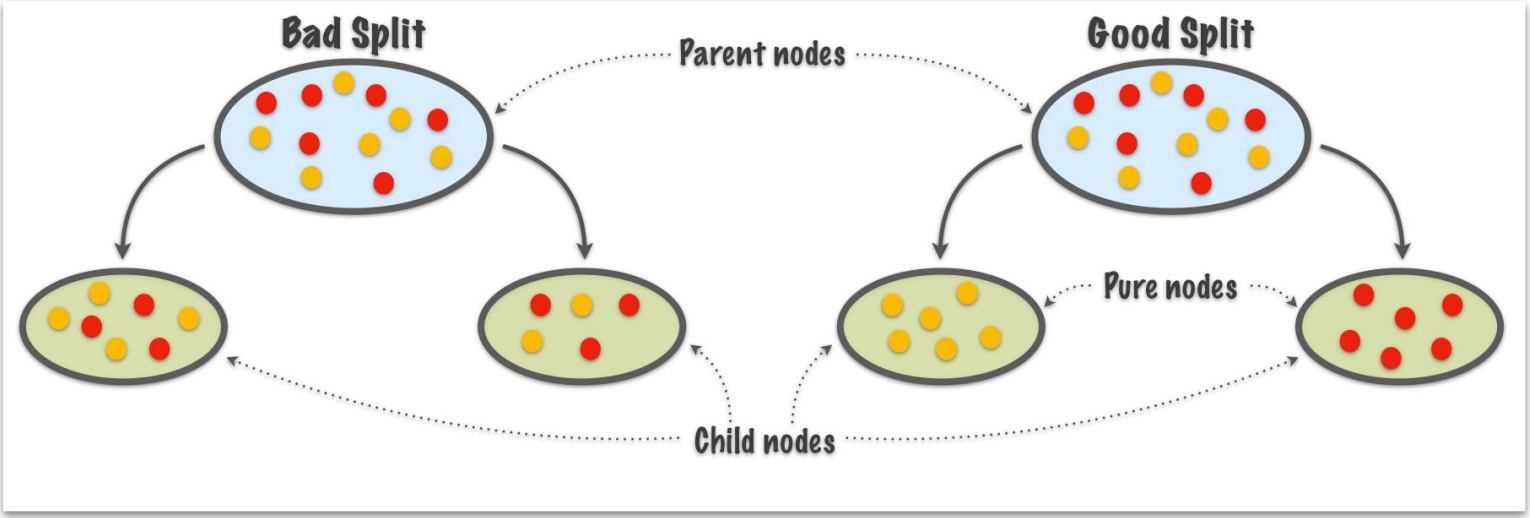
Kingsford & Salzberg, 2008

Decision Trees



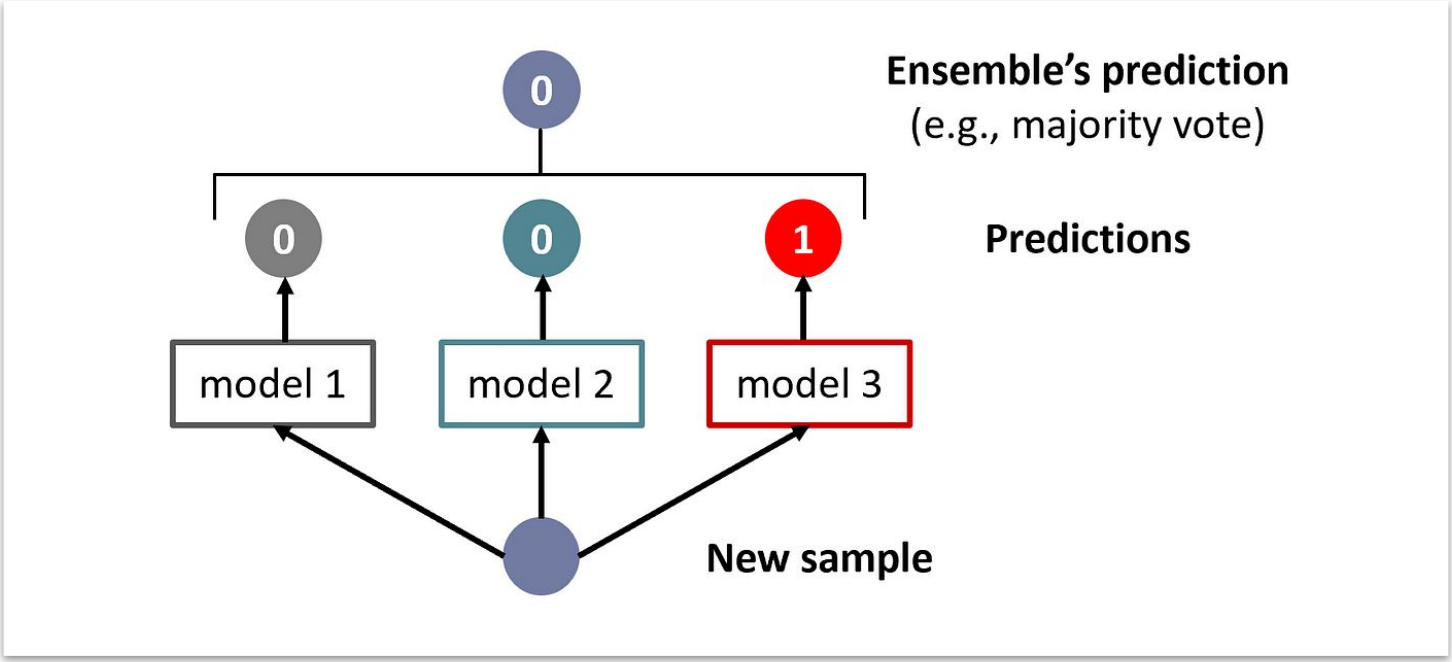
Potts and Sammut, 2000

Decision Trees



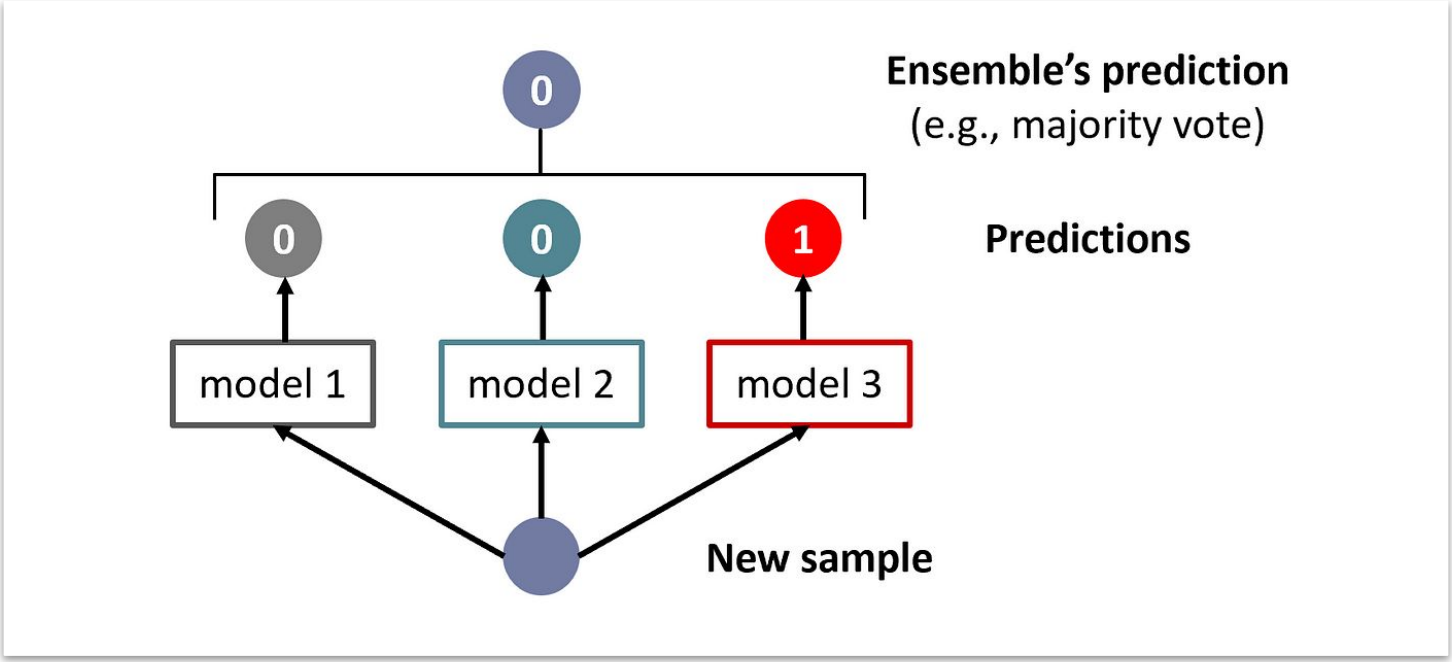
Jeffares, 2018

Ensembles



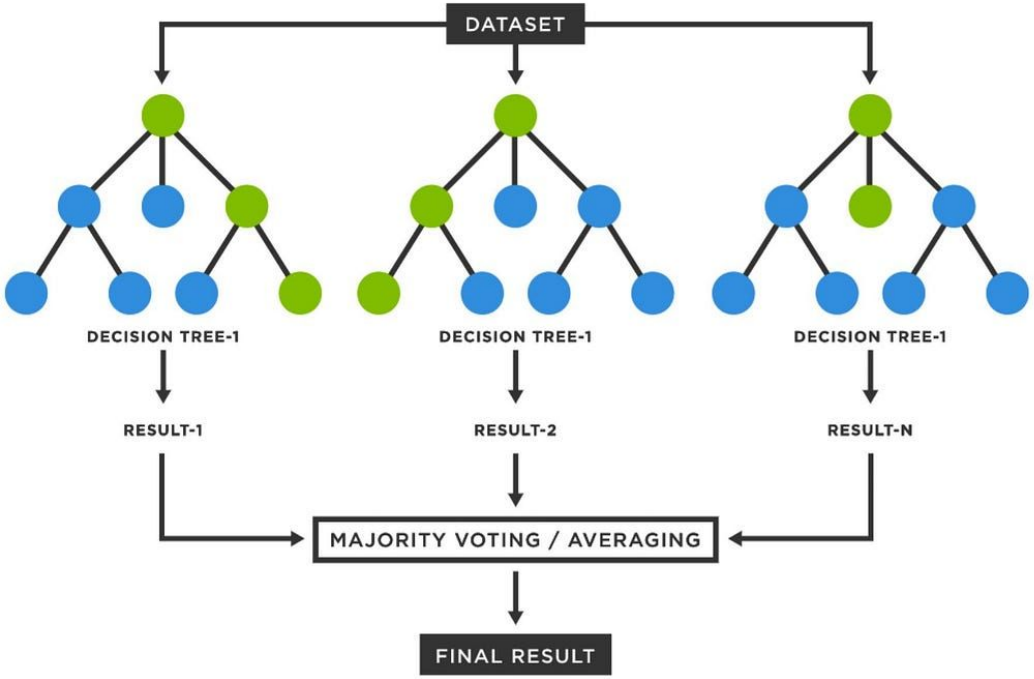
Yehoshua, 2023

Random Forest



Yehoshua, 2023

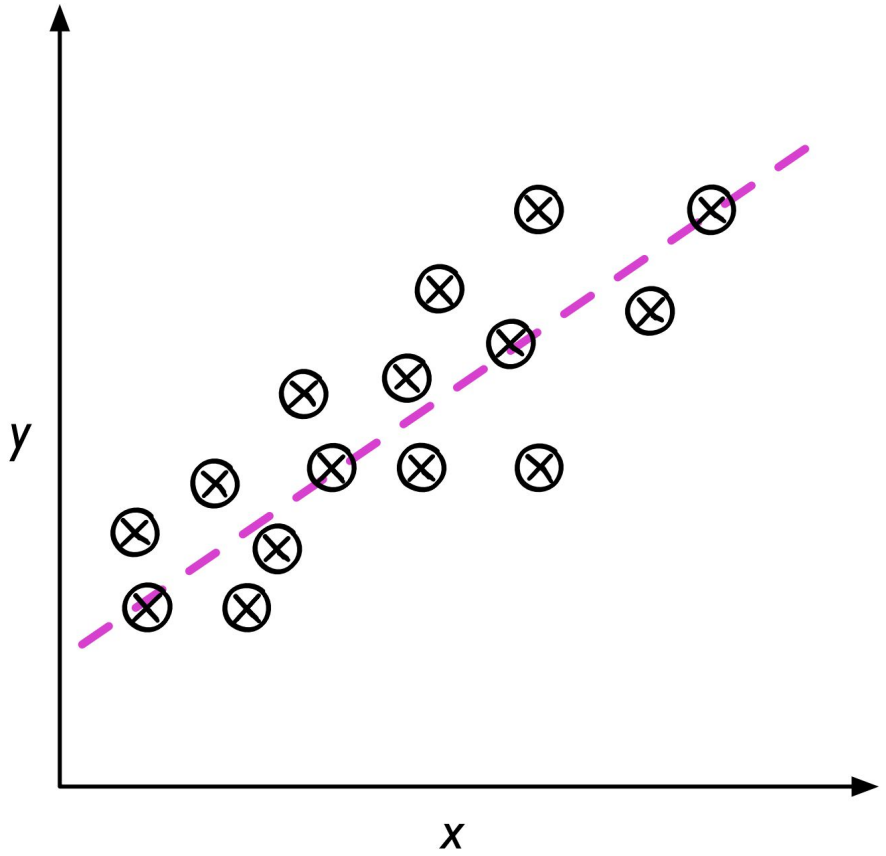
Random Forest



Kumar, 2021

Focusing on Regression

Regression Curve



Raschka, 2021

Part 3: Support Vector Regressor (SVR) for predicting small-molecule that inhibit protein-protein interactions

https://github.com/alexgcsa/comp_life_sciences_2023

ML Resources

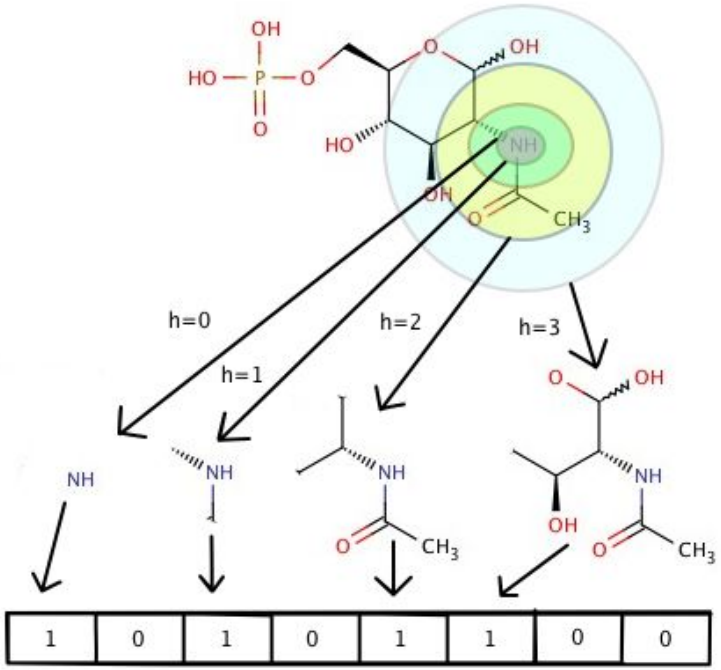
- Tom Mitchell's Book and Youtube Course:
  - https://www.cs.cmu.edu/~tom/mlbook.html
  - https://www.youtube.com/watch?v=m4NlfvrRCdg&list=PLl-BBnDxtUt1hLXmIwu27P22bTi6VwMkN
- Sebastian Raschka's Course:
  - https://sebastianraschka.com/blog/2021/ml-course.html
- Andrew Ng's Course:
  - https://www.coursera.org/specializations/machine-learning-introduction

Introduction To Machine Learning

Alex de Sá

a.desa@uq.edu.au
@alexgcsa

Morgan or Circular FingerPrint

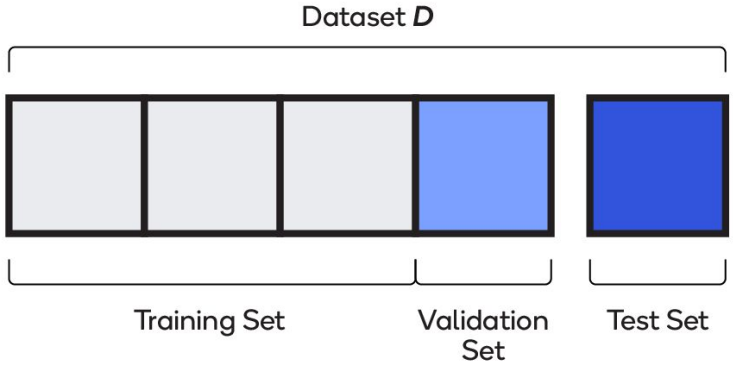


ChemBioInfo, 2011

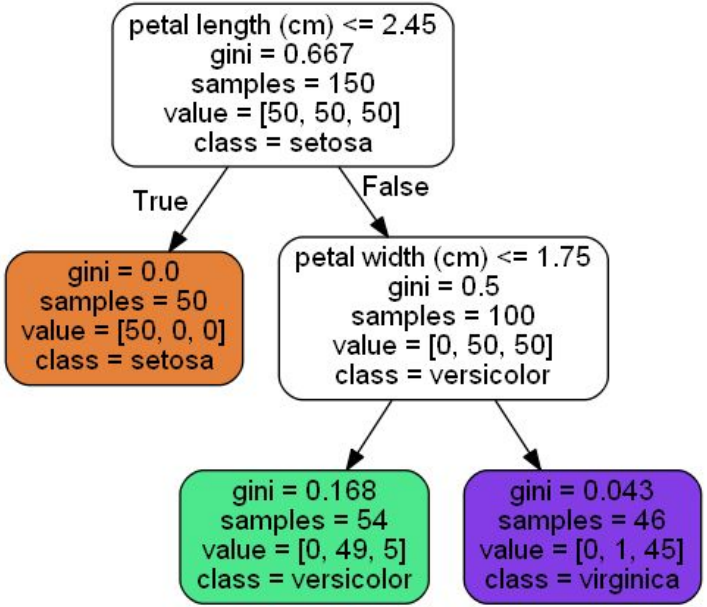
K-Fold Cross-Validation

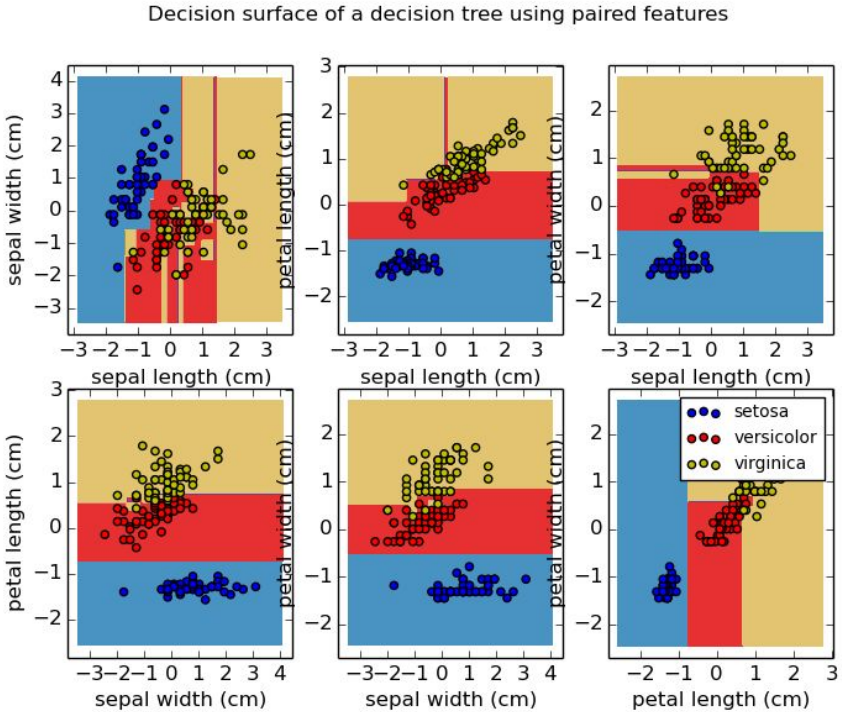
Train/Test

Test the model on new data, assessing its generalisation



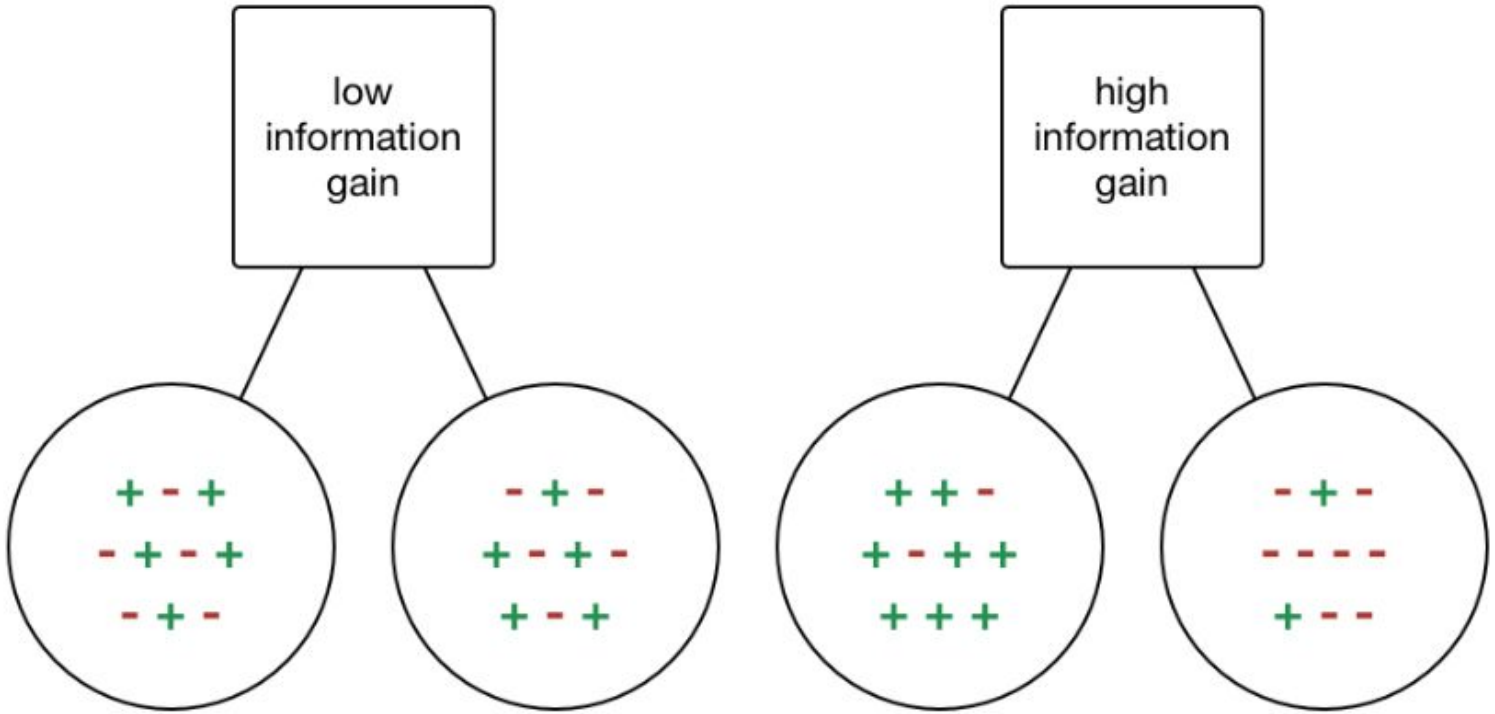
Decision Tree Classifier





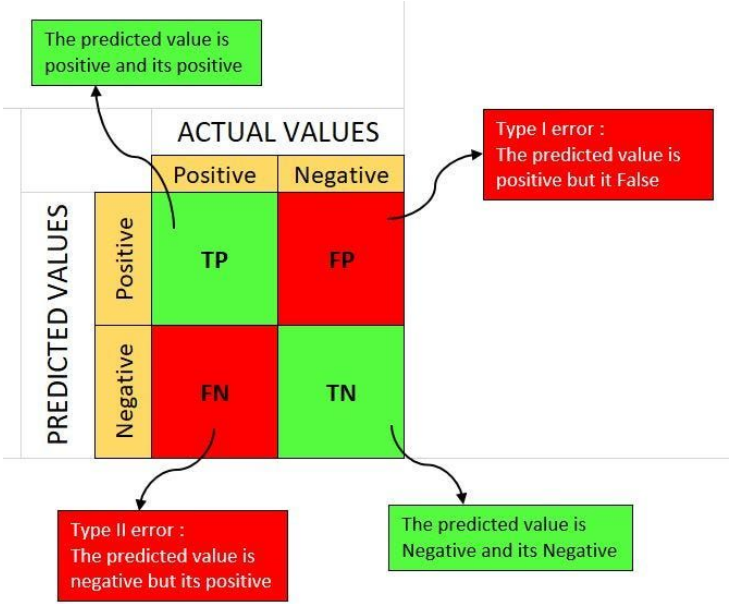
Towards Data Science, 2019

Decision Tree Classifier - Splitting



KDNuggets, 2022

Classification Metrics

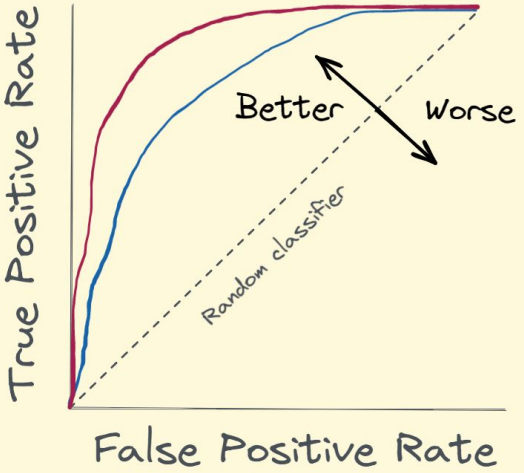


$$\text{Accuracy} = \frac{TP + TN}{\text{Total Samples}}$$

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

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

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



[Towards Data Science, 2023]

[Medium, 2020]