# SNGULAR

## A financial fraud predictive AI model integrated with Spring Boot.

### Predictive Model for Financial Industry

SNGULAR

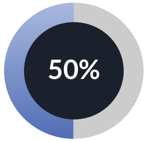
# Contents:

- What is Machine Learning
- Key Concepts
- Supervised Learning Process
- Anomaly Detection
- Example:
  - Data Analysis
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier

SNGULAR

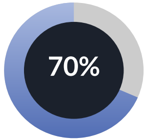
# Fraud in Bank Industry in 2024

More than half banks reported an increase in business fraud.



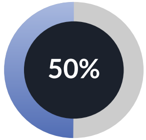

Increase of frauds

Check fraud losses in the Americas reached nearly $21 billion, and 70% of U.S. financial institutions reported check fraud.




Check Fraud

Over half report increasing investment in third-party fraud prevention.



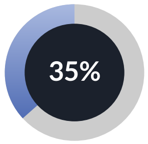

Increase of Investment


- *Thompson Reuters, Deloitte, and the Federal Trade Commission (FTC) https://www.alloy.com/state-of-fraud-benchmark-report-2024#component-marketo-embed*
- *SEC. (n.d.). https://www.sec.gov/files/fy24-oiad-sar-objectives-report.pdf*

# Bank Industry AI Adoption/Maturity in 2024

| COMPANY | SCORE ▼ |
|---|---|
| JPMorganChase | 1 |
| Capital One | 2 |
| Royal Bank of Canada | 3 |
| Wells Fargo | 4 |
| CommBank | 5 |
| UBS | 6 |
| HSBC | 7 |
| Citigroup | 8 |
| TD Bank | 9 |
| PNC Financial | 35 |

The **banking sector** emerged as a leader in AI adoption, with **35% of banks classified as AI leaders**.



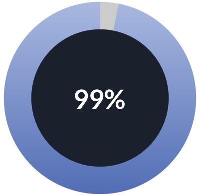

Banks AI Líderes

*Evident AI (EAI) https://evidentinsights.com/*

SNGULAR

# Preview Results

The predictive model obtained 99% of precision, detecting more than 13,500 frauds within one month transaction dataset.




Detection of Frauds

SNGULAR

# What is the best approach?

| # step | type | # amount | # isFraud |
|---|---|---|---|
| 430 | TRANSFER | 2828068.73 | 1 |
| 430 | CASH_OUT | 2828068.73 | 1 |
| 431 | PAYMENT | 4506.4 | 0 |
| 431 | PAYMENT | 20711.86 | 0 |
| 431 | PAYMENT | 14014.89 | 0 |
| 431 | PAYMENT | 3501.32 | 0 |
| 431 | PAYMENT | 13936.67 | 0 |
| 431 | PAYMENT | 1366.84 | 0 |
| 431 | PAYMENT | 5959.65 | 0 |

| # step | type | # amount |
|---|---|---|
| 431 | PAYMENT | 7536.7 |

SNGULAR

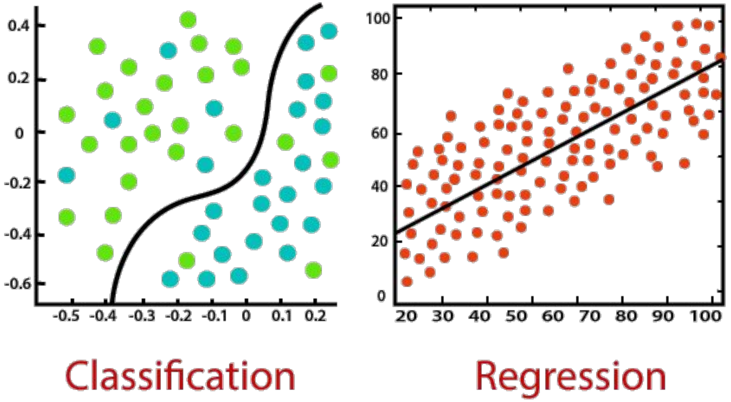
# What is the best approach?

# What is the best approach?

Machine learning

SNGULAR

# What Machine Learning(ML) is



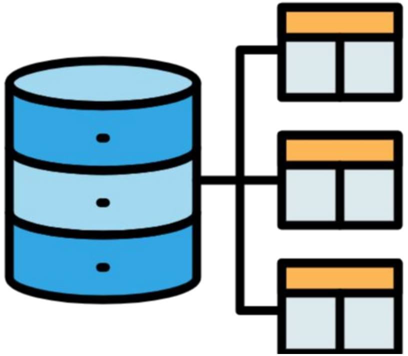

It's a sub-topic of Artificial Intelligence field, which is focused on predicting data, through information patterns.

It learns from data to enhance automation of decision making.

ML doesn't implement hard-coded rules. Instead, it predicts data from unseen information.

SNGULAR

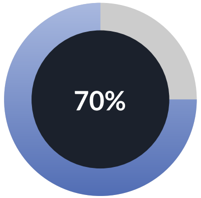
# Data



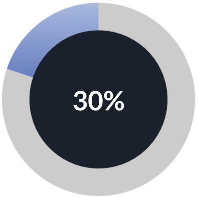
**01** The more data, the better.

**02** Data format can be numbers, text, images, sounds.

**03** It is divided into training (used to teach the model) and test (used to evaluate the model).




Train




Test

SNGULAR

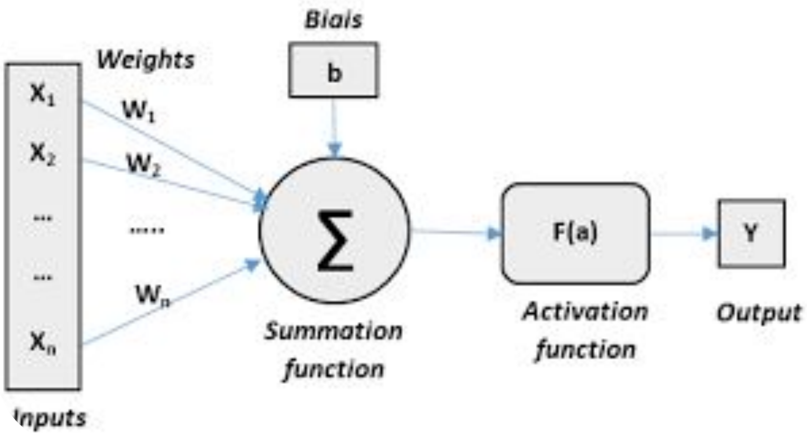
# Model

**01** Mathematical representation that learns from data.

**02** Models capture data patterns and relationships to make predictions and decisions.

**03** Typical models are:

- Logistic Regression
- Support Vector Machine
- Neural Networks, etc.



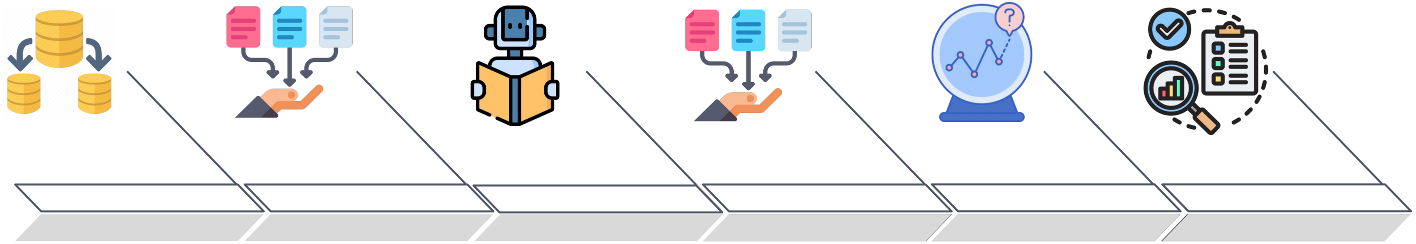
SNGULAR

# Training & Prediction

**01** Training is teaching mathematical model to recognize patterns in data.

**02** Data can be labeled(supervised learning) or unlabeled(unsupervised learning).

# Supervised Learning Process



**Split Data**

Training (70%)

Testing (30%)

**Provide Training Data**

Training data which is labeled

**Train Model**

Train with labeled data

**Provide Testing Data**

Testing data which is unlabeled

**Predict Labels**

Prediction of labels based on training data

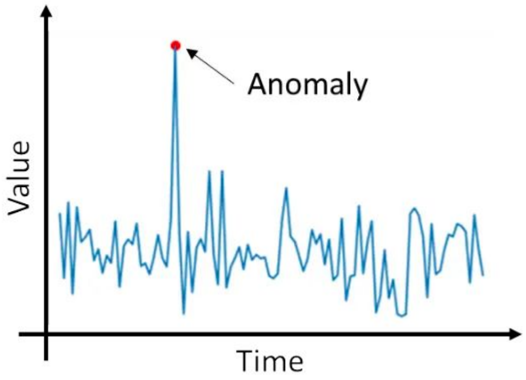
**Validate model**

Compare predicted labels with actual labels

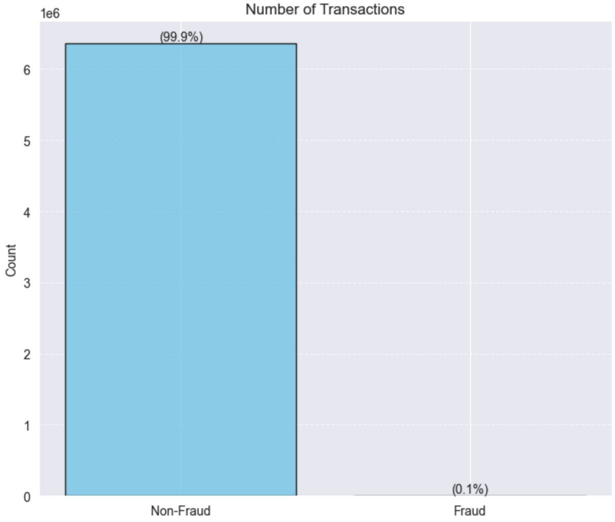
SNGULAR

# Anomaly Detection

Outliers




Machine learning model used to detect outliers in extreme unbalanced datasets, usually big enough to consider high-demand computational processing.




Unbalanced dataset over 24 million

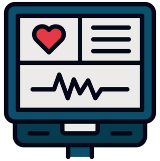
SNGULAR

# Anomaly Detection Use Cases

**01**

Financial Fraud:

Identifying unusual spending patterns on credit cards to detect fraudulent transactions.



**02**

Healthcare Monitoring:

Identifying unusual vital signs in patient data that could indicate a medical emergency.

**03**

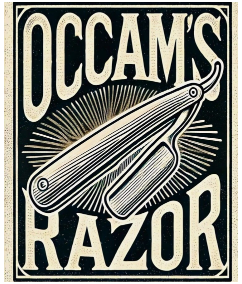
Retail Sales Analysis:

Identifying sudden spikes or drops in sales for specific products that could indicate issues with pricing or demand.

# Theoretical Explanation of Data Preprocessing

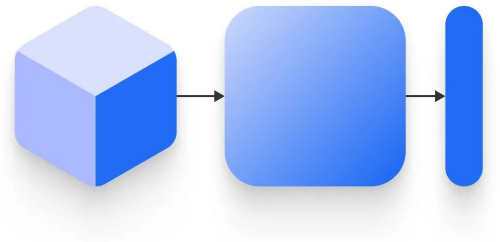
## Descriptive and Inferential analysis

Need to know what the behind data.




## Occam's Razor

Problem-solving principle that states the simplest explanation is usually the best.



## Dimensional Reduction

Reduce feature or variables of the dataset.
Retaining its most important properties.

SNGULAR

# Data Description

“PaySim is a financial simulator that simulates mobile money transactions based on an original dataset.”

The dataset has 24 million financial records.

*E. A. Lopez-Rojas, A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016*

SNGULAR

# Dataset Fields

- **step:** Represents a unit of time in the real world, with 1 step equating to 1 hour.
- **type:** Transaction types include CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER.
- **amount:** The transaction amount in the local currency.
- **nameOrig:** The customer initiating the transaction.
- **oldbalanceOrg:** The initial balance before the transaction.
- **newbalanceOrig:** The new balance after the transaction.
- **nameDest:** Transaction's recipient customer.
- **oldbalanceDest:** The initial recipient's balance before the transaction.
- **newbalanceDest:** The new recipient's balance after the transaction.
- **isFraud:** Identifies transactions conducted by fraudulent agents aiming to deplete customer accounts through transfers and cash-outs.

# Dimensional Reduction

Feature Selection:

**Shapiro-Wilk** test to determine null hypothesis - > not normal distribution. > 0.05 confidence level:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

$$(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{\|V^{-1} m\| = (m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}}$$

*V*: Variance Covariance Matrix
*m: x*

SNGULAR

# Dimensional Reduction

Feature Selection:

**Pearson Correlation**: Correlation between two normal distributed variables.

Coefficient correlation: -1 to 1

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Theta : Standard Deviation

Cov : Variance Covariance Matrix

SNGULAR

# Dimensional Reduction

Feature Selection:

**Spearman Correlation:** Correlation between two non-normal distributed variables.

Coefficient correlation: -1 to 1

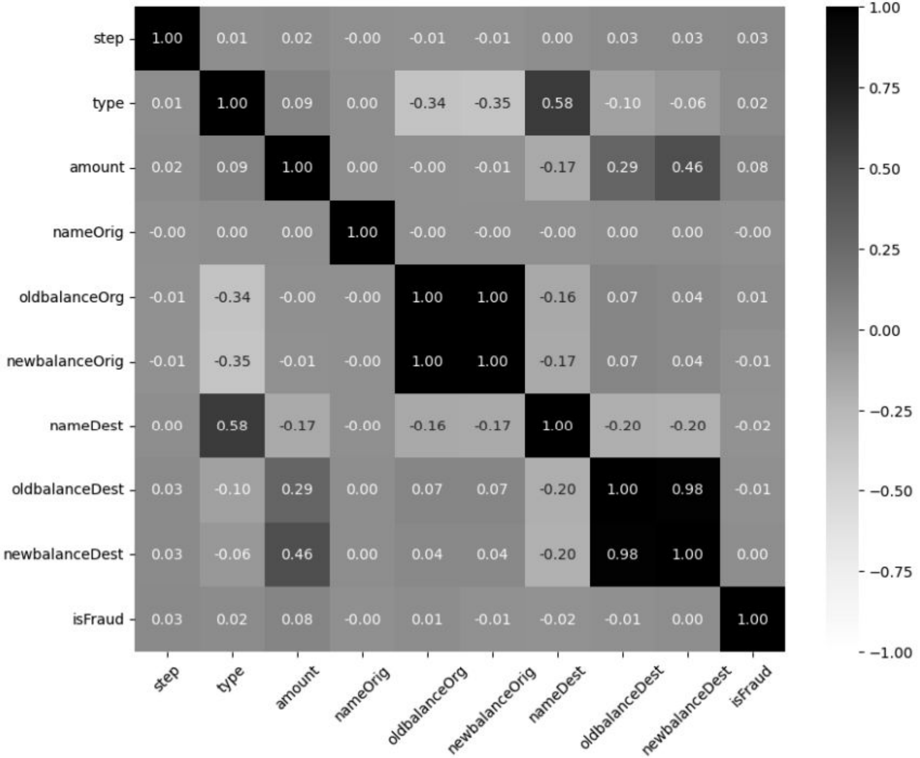
$$r_s = \rho\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right] = \frac{\operatorname{cov}\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right]}{\sigma_{\mathrm{R}[X]}\,\sigma_{\mathrm{R}[Y]}},$$

$\operatorname{cov}\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right]$: Covariance of ranked variables

$\sigma_{\mathrm{R}[X]}\,\sigma_{\mathrm{R}[Y]}$: Standard Deviation

# Dimensional Reduction

Pearson Correlation Matrix




Spearman Correlation Matrix

# Dimensional Reduction

Test Hypothesis

**Logistic Regression:** Relationship between two variables. Only applicable when dependent variable (y) is categorical, and independent variable (x) is continuous.

p-Value: 0 - 0.05 (confidence threshold)

$$p(x) = \frac{1}{1 + e^{-(\beta_0 - \beta_1 x)}}$$

Beta: Weight optimization parameter

# Dimensional Reduction

Test Hypothesis

**Chi-Square:** Relationship between two variables. Only applicable when dependent variable (y) is categorical, and independent variable (x) is also categorical.

p-Value: 0 - 0.05 (confidence threshold)

$$\chi^2 = \sum_{i=1}^{n} \frac{O_i^2}{E_i} - N.$$

$O_i$: Number of observations of variable i

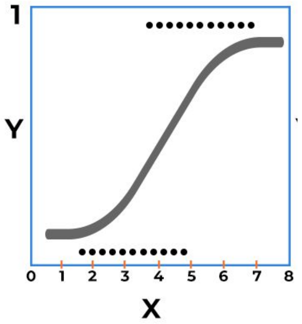
$N$: Total number of observations

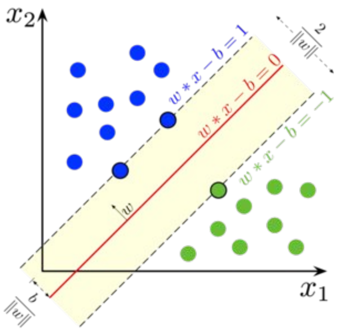
$E_i$: Expected number of type i

SNGULAR

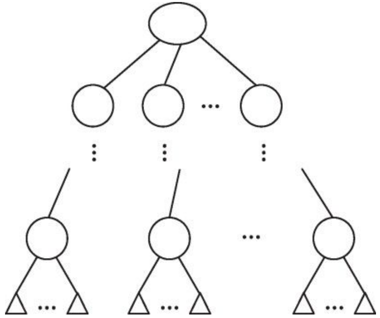
# Theoretical Explanation of Predictive Models

Logistic Regression



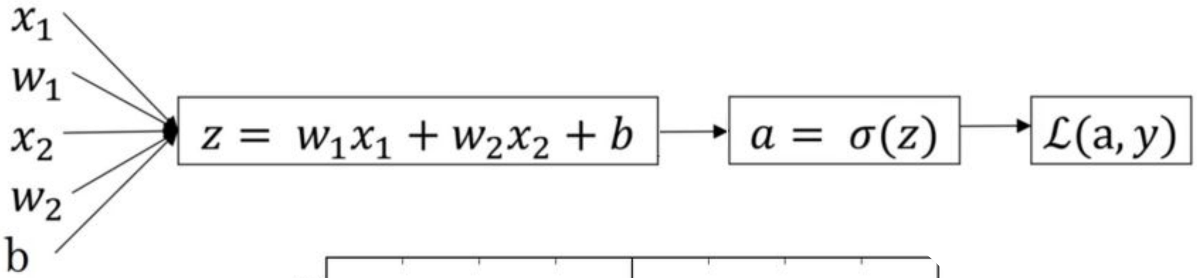

Support Vector Machine

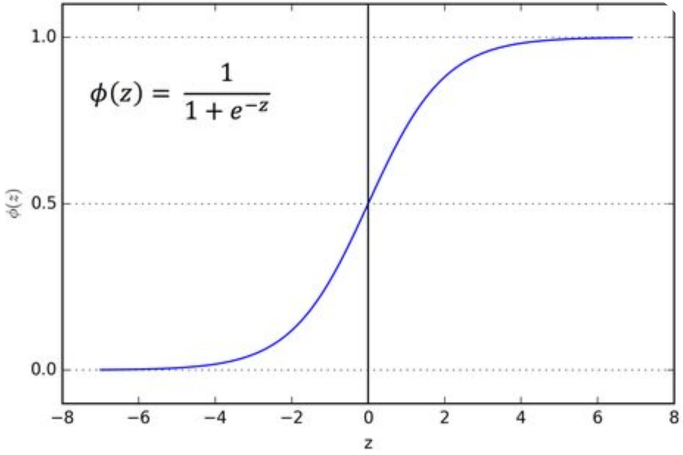


Decision Tree Classifier

# Logistic Regression

Classification Function



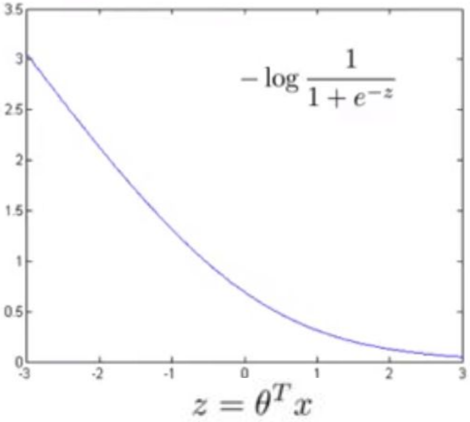



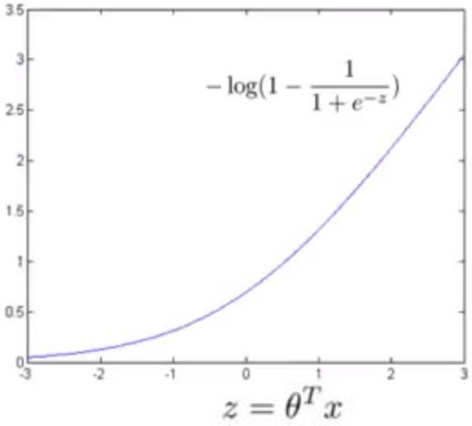

Sigmoid Function

# Logistic Regression

## Cost Optimization Function

$$\mathrm{Cost}(h_\theta(x), y) = -y \log \frac{1}{1 + e^{-\theta^T x}} - (1 - y) \log(1 - \frac{1}{1 + e^{-\theta^T x}})$$



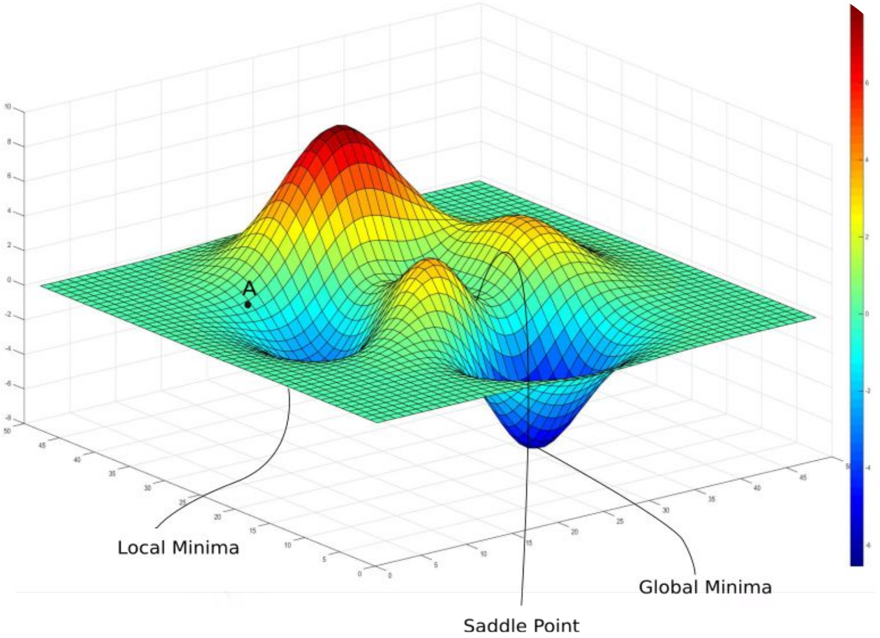
# Logistic Regression

## Cost Optimization Function

Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

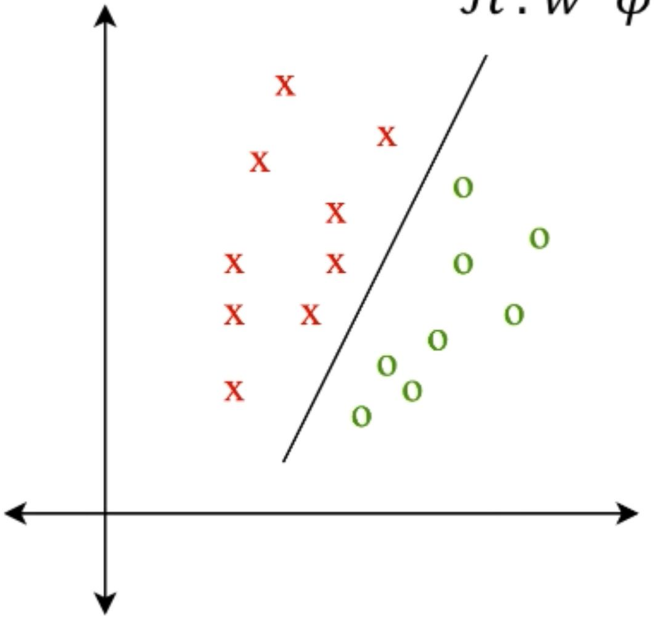


SNGULAR

# Support Vector Machine

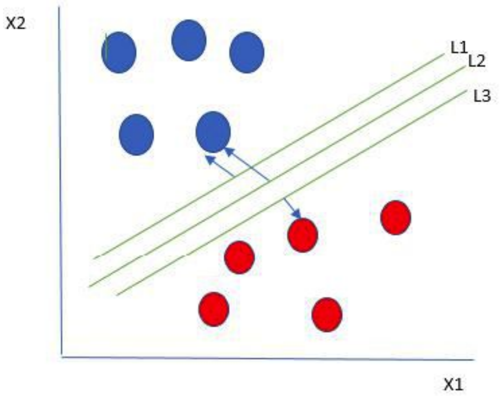
Classification Function Hyperplane Clasificator

$$\mathcal{H} : w^T \phi(x) + b = 0$$



# Support Vector Machine

Optimization Function




$$d_{\mathcal{H}}(\phi(x_0)) = \frac{|w^T \phi(x_0) + b|}{\|w\|_2}$$

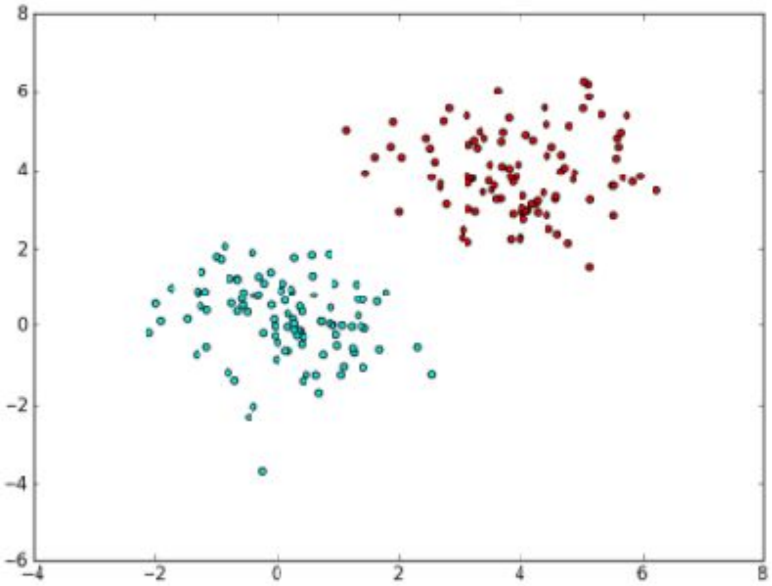
Distance of points from hyperplane

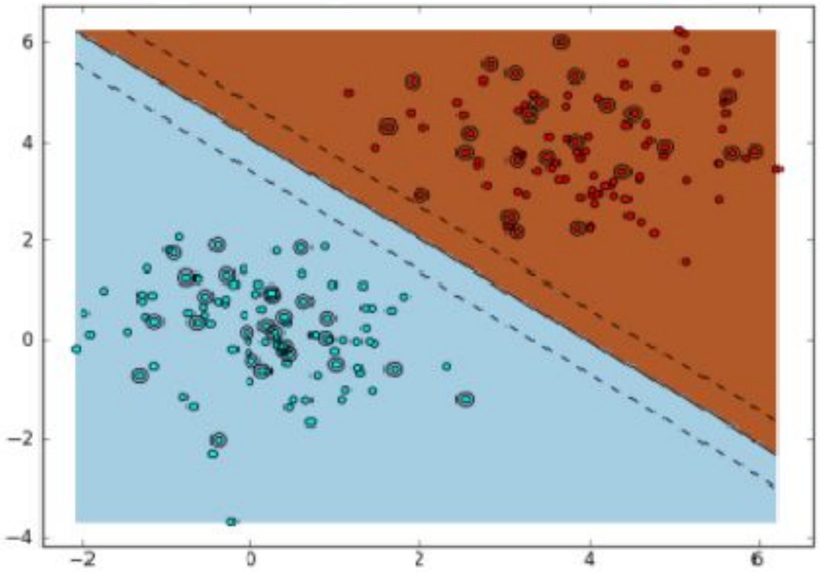
$$w^* = \arg\max_w \left[\min_n d_{\mathcal{H}}(\phi(x_n))\right]$$

**Optimization function:** Max distance of the hyperplane of the nearest point from the hyperplane

# Support Vector Machine

Kernel

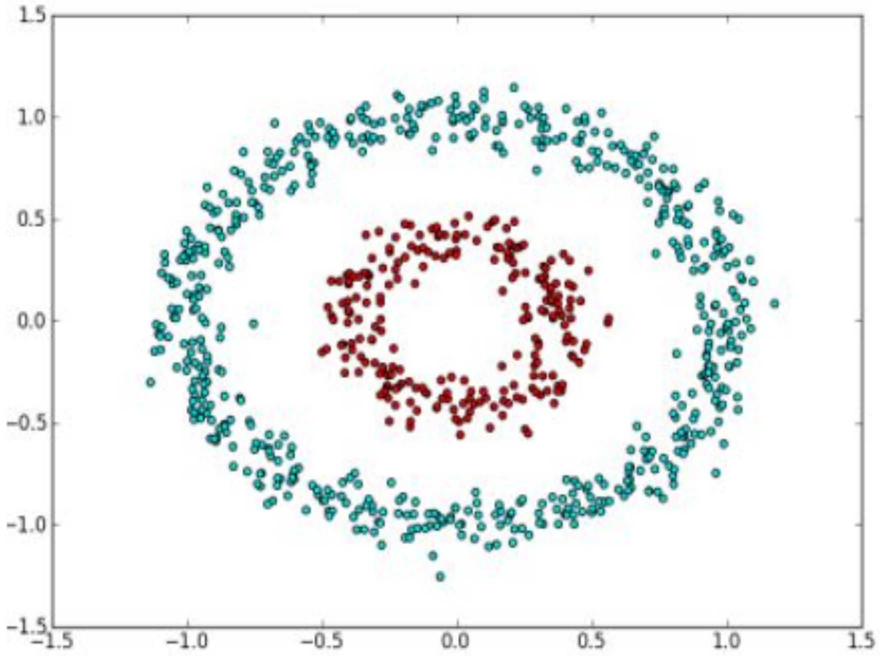


SNGULAR

# Support Vector Machine

Kernel

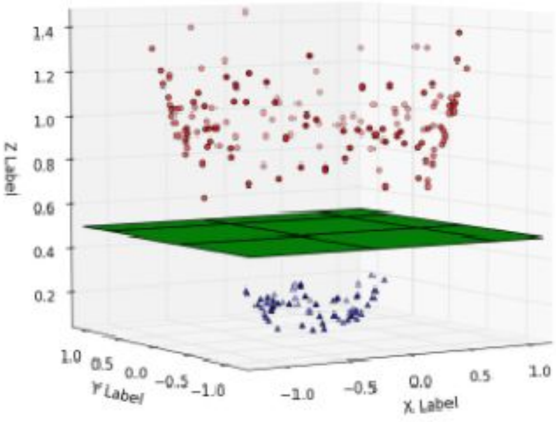


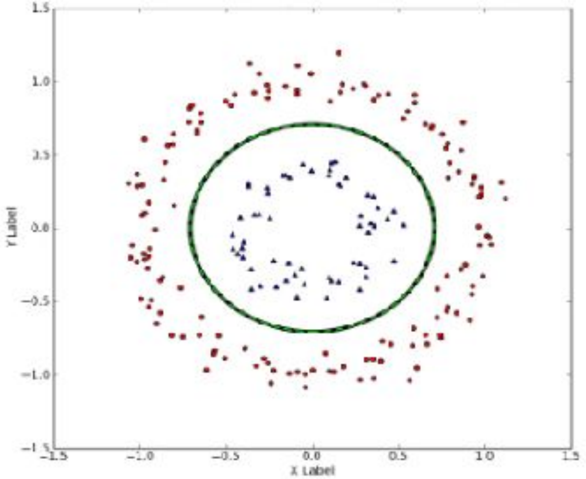
# Support Vector Machine

Kernel

Lagrange Multipliers

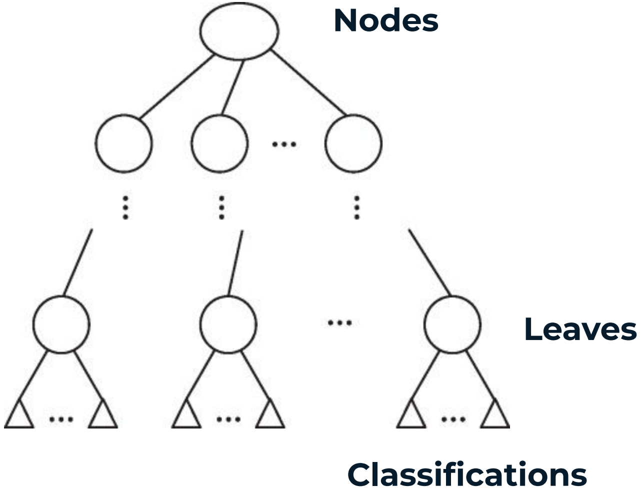
$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3 \,.$$

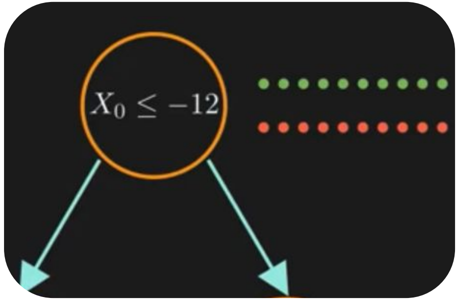



Hilbert Space

# Decision Tree Classifier - Tree Creation




**Leaves and nodes randomly generated** through seeds mechanism:

$$R_{i+1} = (a \cdot R_i + c) \mod m,$$

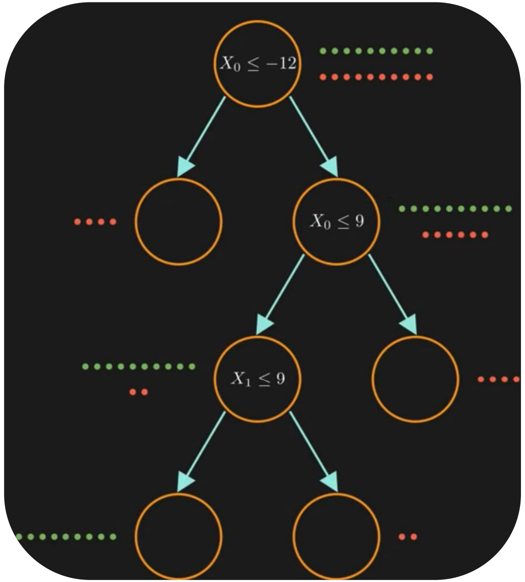
**Where**

***Ri*** *is current state*

***a, c , m*** *are random numbers generated*

SNGULAR

# Decision Tree Classifier - Evaluation Per Node




## Gini Index

$$I(D_{\text{node}}) = 1 - \sum_{c=1}^{C} p_c^2,$$

## Entropy

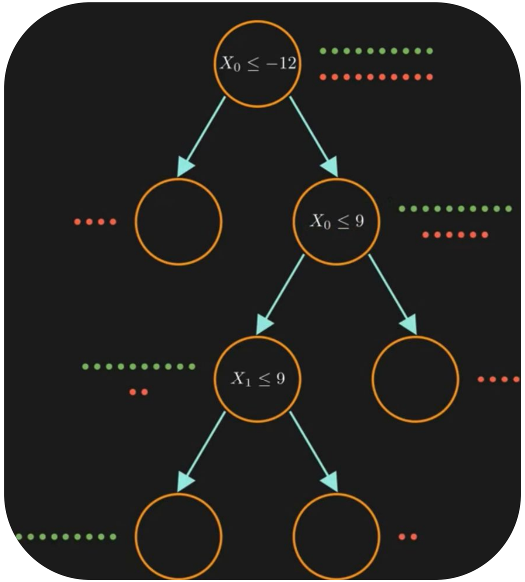
$$I(D_{\text{node}}) = - \sum_{c=1}^{C} p_c \log(p_c),$$

## Where

***Pc*** *is the proportion of samples c, which belongs to Dnode*

SNGULAR

# Decision Tree Classifier - Evaluation Horizontally




## Gini Index

$$\text{Information Gain} = I(D_{\text{node}}) - \left( \frac{|D_{\text{left}}|}{|D_{\text{node}}|} I(D_{\text{left}}) + \frac{|D_{\text{right}}|}{|D_{\text{node}}|} I(D_{\text{right}}) \right),$$

**Where**

***Dleft and Dright*** *are the data subsets resulting from the split.*

*|**D**| denotes the number of samples in DDD*

SNGULAR

# Implementation

01 Inferential Analysis.

02 Data Preparation.

03 Implementation of Logistic Regression.

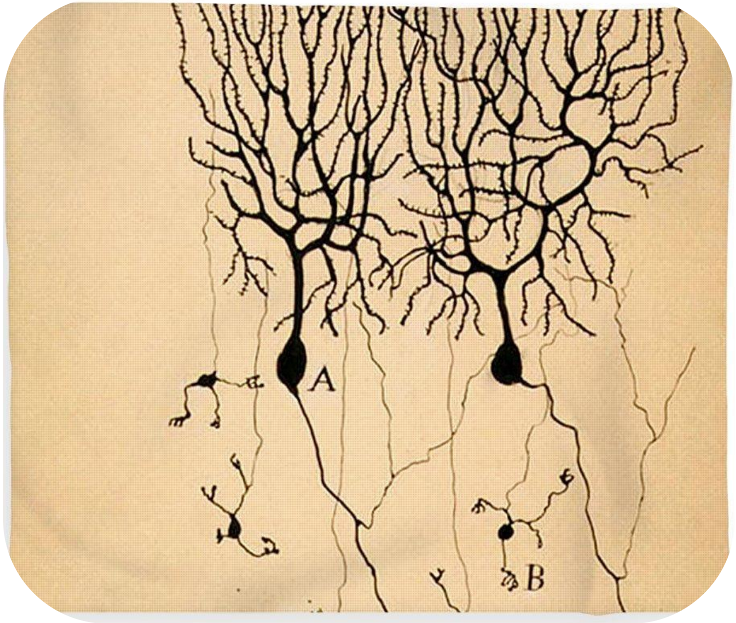
04 Implementation of Support Vector Machine.

05 Decision Tree Classifier

06 Spring Boot Integration.

SNGULAR

# Bonus

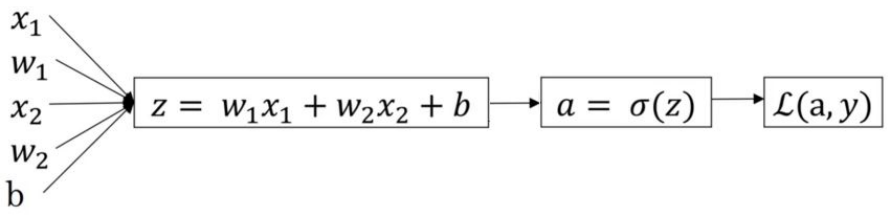



*Modelo Neuronal - Santiago Ramon y Cajal (1906)*

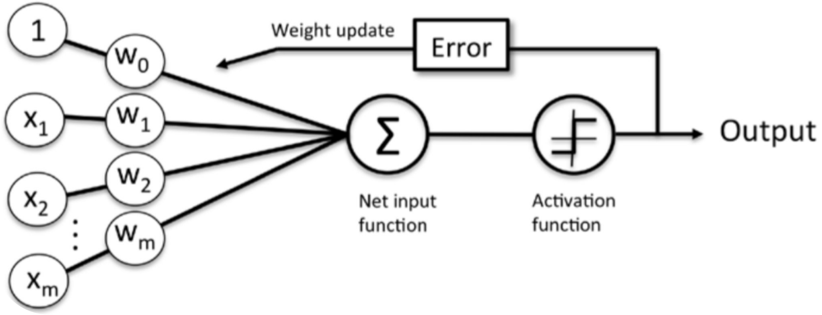
SNGULAR

# Bonus

## Logistic Regression Function

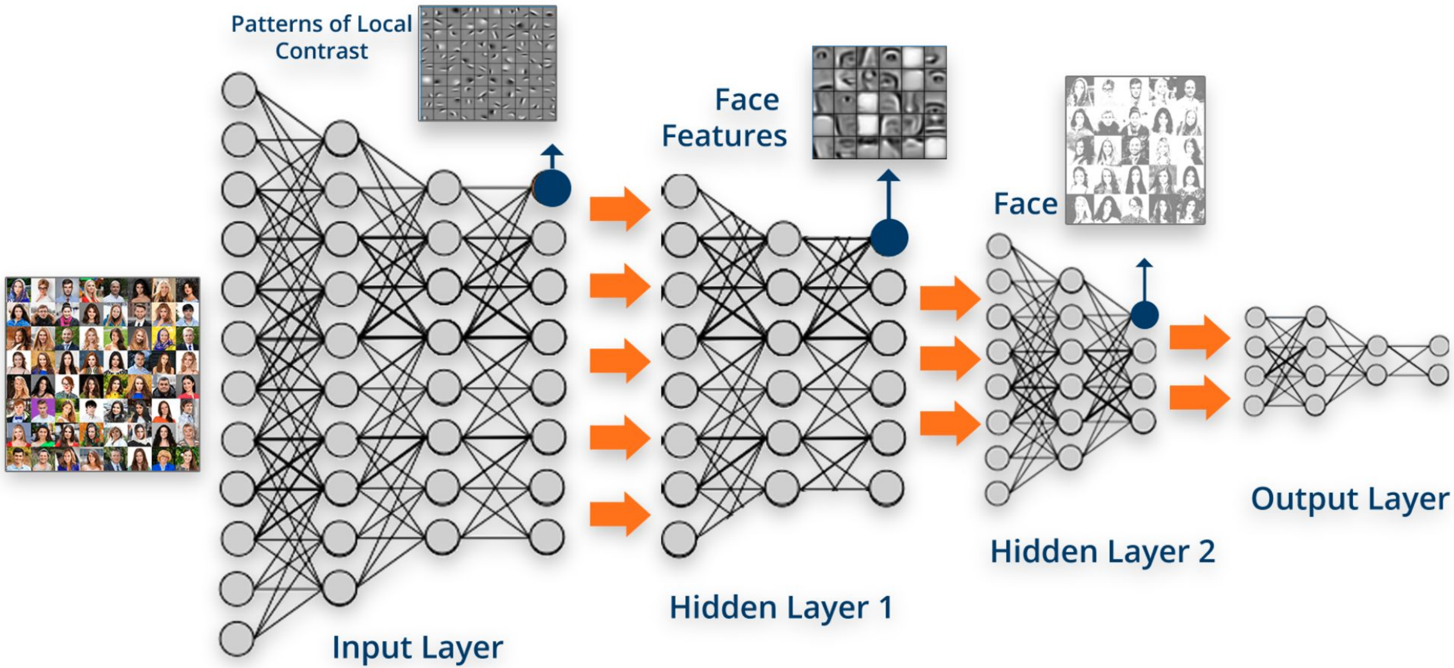



## Neuron Model



SNGULAR

# Bonus



SNGULAR

## Mindset of innovation

*"Machine learning is not just a tool. it is a mindset shift that enables us to transform data into actionable insights, driving smarter decisions and paving the way for innovation across every corner of the industry"*

# Bibliography


- *Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In 28th European Modeling and Simulation Symposium, EMSS, Larnaca (pp. 249-255). Dime University of Genoa.*

- *Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.*

- *Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. Expert systems With applications, 193, 116429.*

- *Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–272. doi:10.1038/s41592-019-0686-2*

- *Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.*

- *Komer, B., Bergstra, J., & Eliasmith, C. (2019). Hyperopt-sklearn. Automated Machine Learning: Methods, Systems, Challenges, 97-111.*

- *Sarkar, T. (2022). GPU-Based Data Science for High Productivity. In Productive and Efficient Data Science with Python: With Modularizing, Memory profiles, and Parallel/GPU Processing (pp. 299-326). Berkeley, CA: Apress*

- *SEC. (n.d.). https://www.sec.gov/files/fy24-oiad-sar-objectives-report.pdf*

SNGULAR

## Questions

Thanks for your attention!!!