Intro to Machine Learning

About me

Ricardo Neves

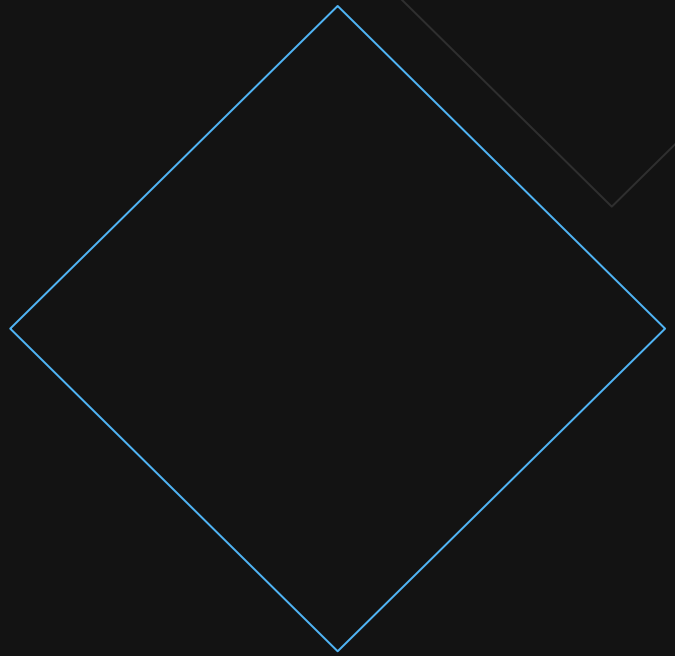
- Sr. Data Scientist @ DareData Engineering
- 6 years working with data, 4 as DS
- Bsc Physics
- Aveiro, Portugal
- Email: ricardo.asm.neves@gmail.com

Definitions

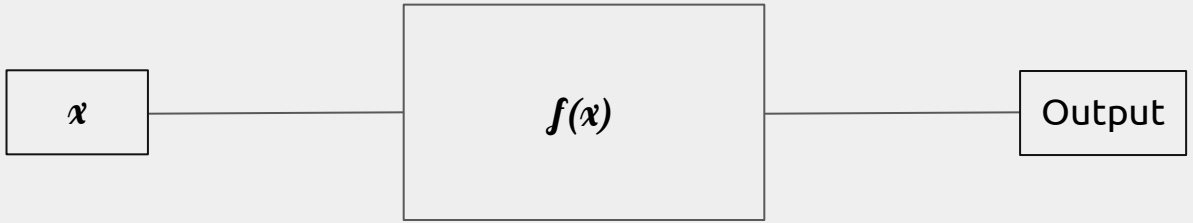
Artificial Intelligence (AI): it's a branch of computer science that aims to create systems capable of performing tasks that would normally require human intelligence.

Machine learning: Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

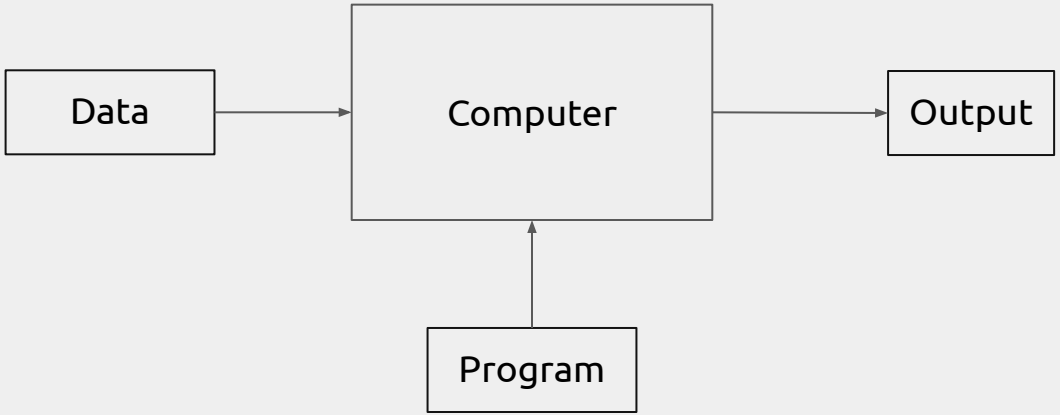
ML models



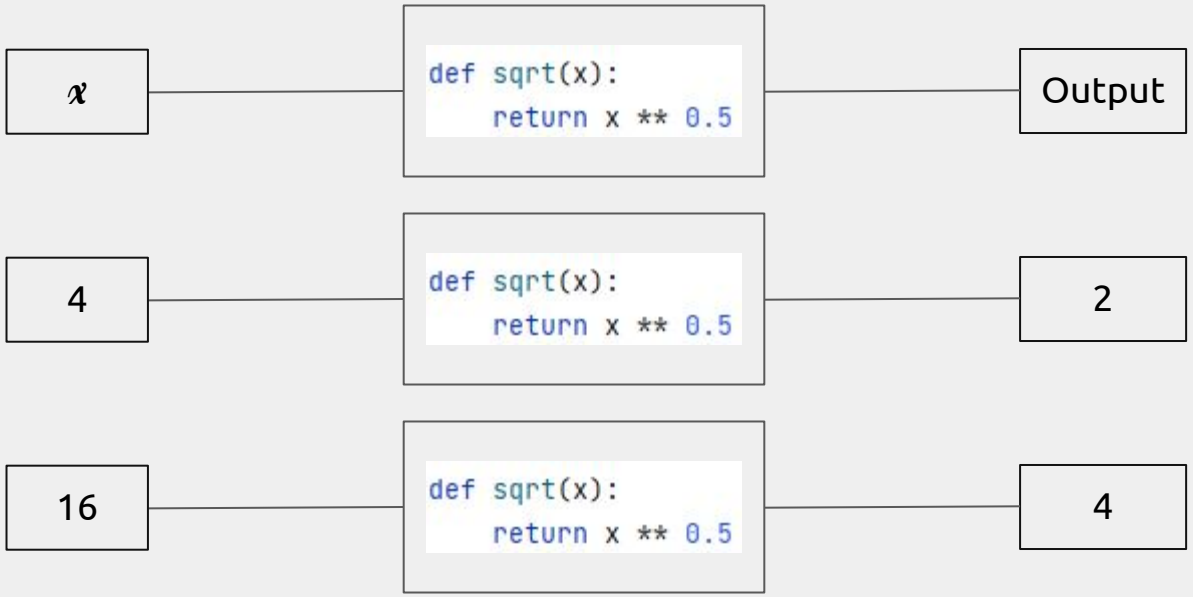
Function



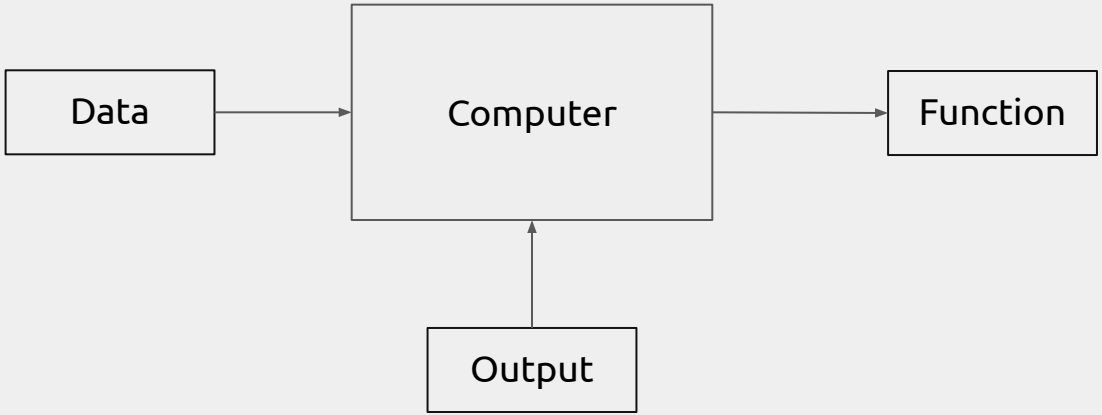
Traditional Programming



Traditional Programming



Machine Learning



Predicting stock price

Predicting stock price

Image Classification

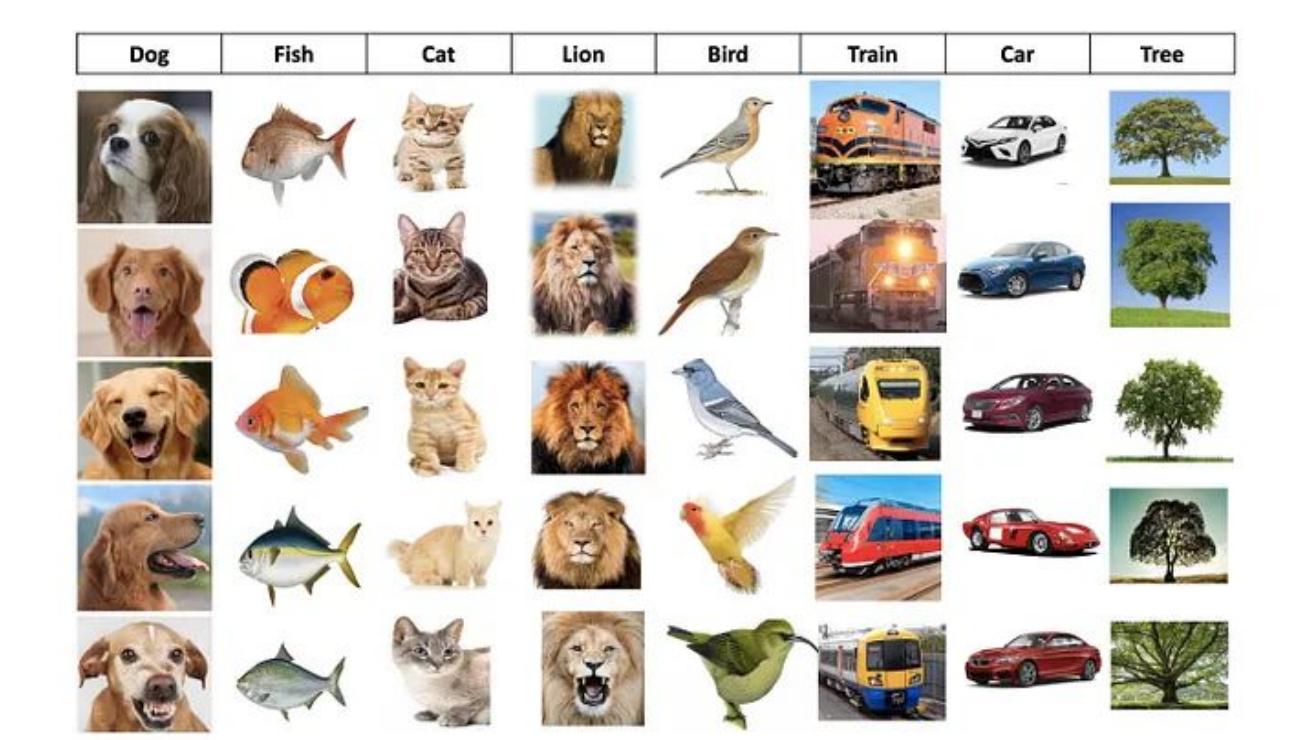
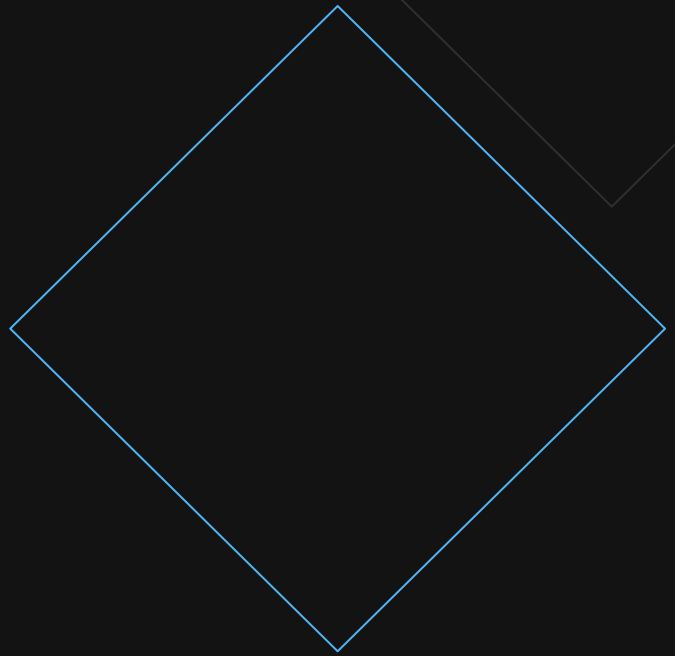


Image Classification

Types of learning



Supervised

Training data + Final result

Unsupervised

Training data

Semi-Supervised

Training data + Some results

Reinforcement learning

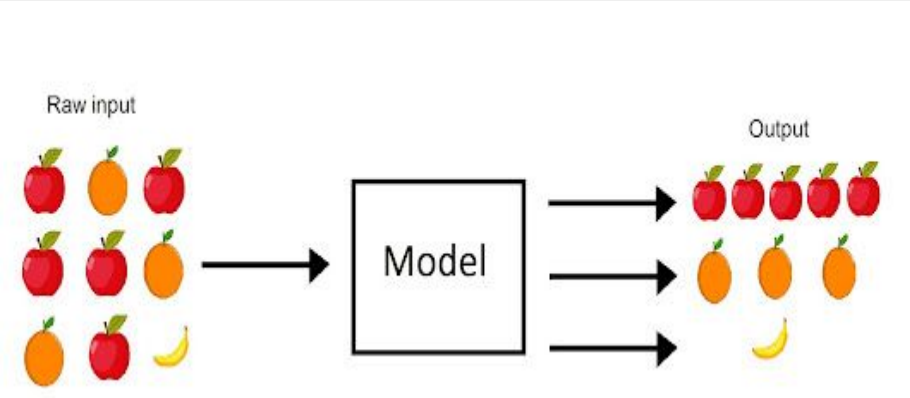
Training data + Reward/penalty

Supervised Learning

Classification

Models to predict discrete outcomes

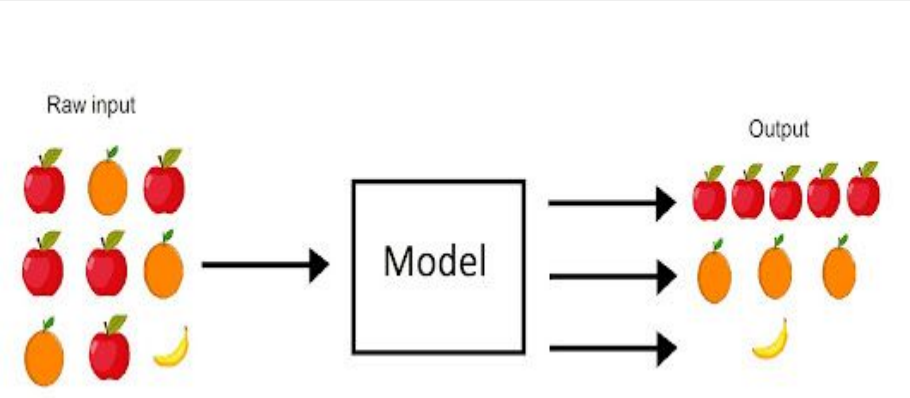
- Image classification
- Fraud detection
- Spam detection



Classification

Algorithms

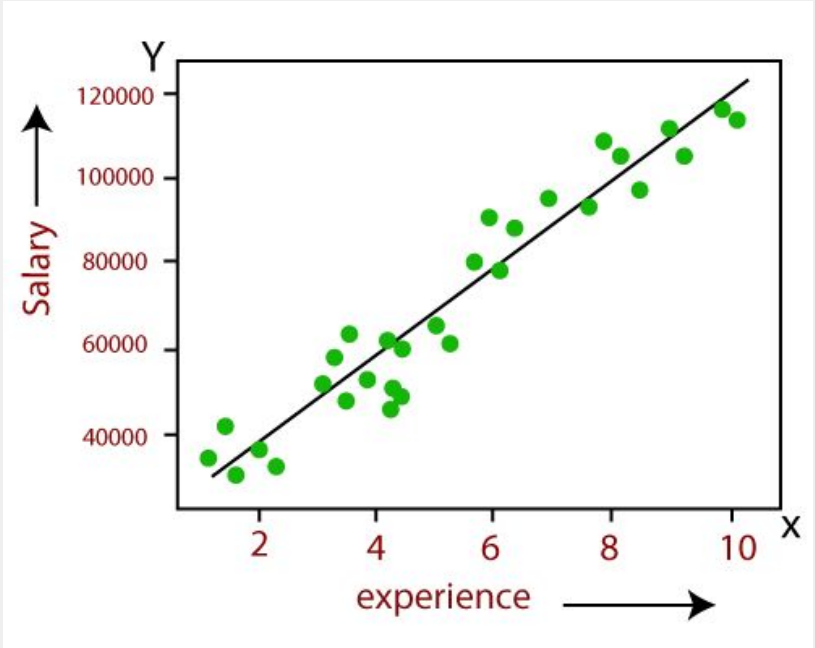
- Decision Tree
- Random Forest
- Logistic Regression
- Gradient Boosting (XGBoost, LightGBM, etc.)



Regression

Models to predict continuous outcomes

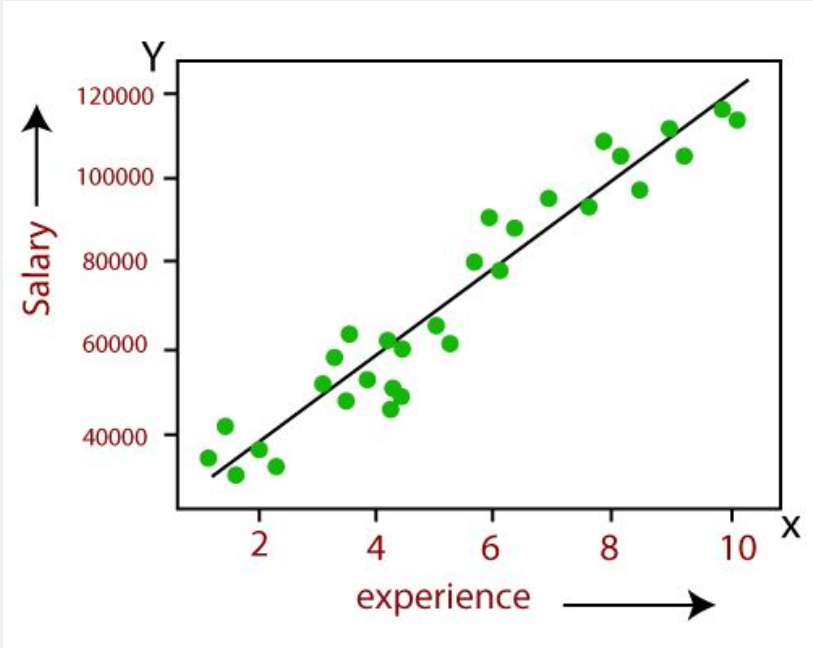
- Stock price
- Sales forecast
- House prices



Regression

Algorithms

- Linear Regression
- Polynomial Regression
- Decision Tree Regression
- Gradient Boosting Regression

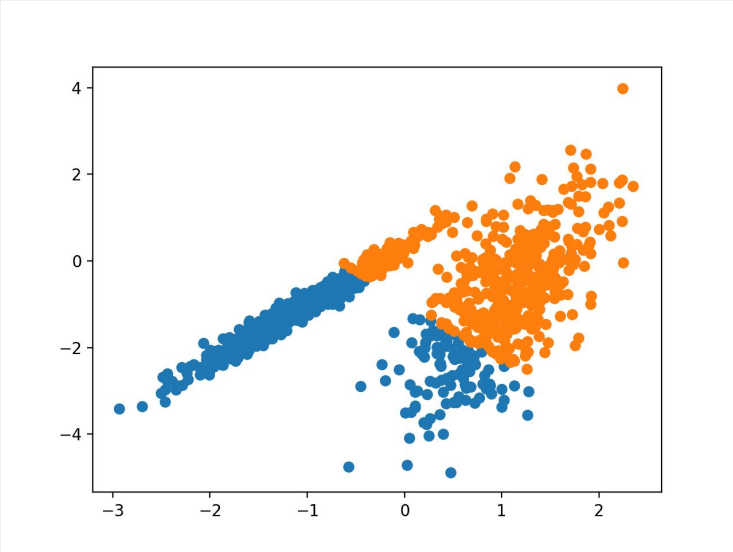


Unsupervised Learning

Clustering

Models that group data points based on how similar they are

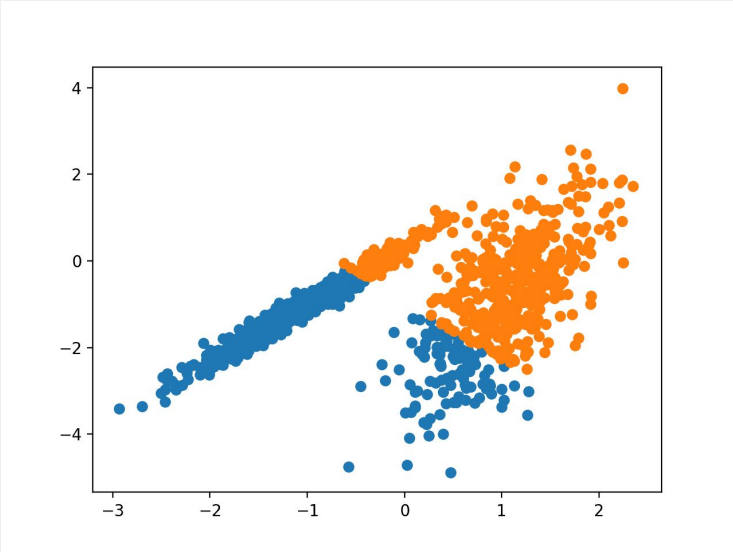
- Customer segmentation
- Image Segmentation
- Document Clustering



Clustering

Models that group data points based on how similar they are

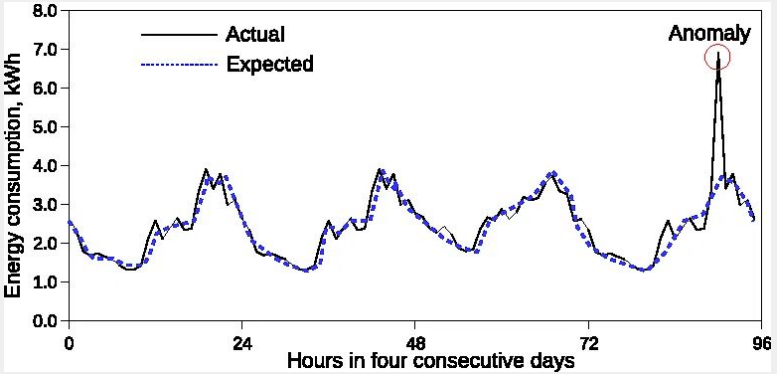
- K-means Clustering
- Hierarchical Clustering



Anomaly detection

Models find data points that are outside expected behaviour

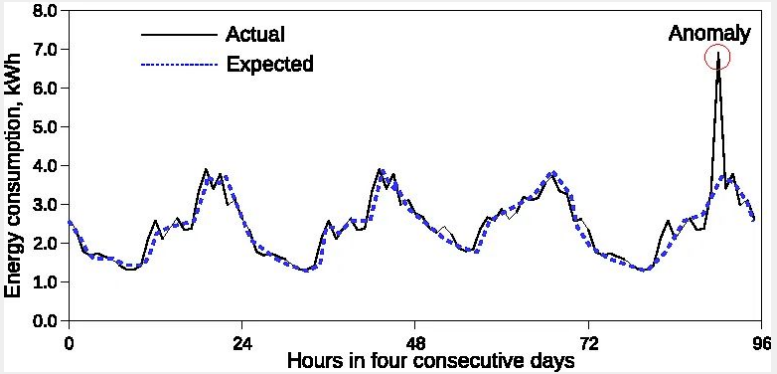
- Health Monitoring
- Intrusion detection
- Image Recognition
- Network Traffic Analysis



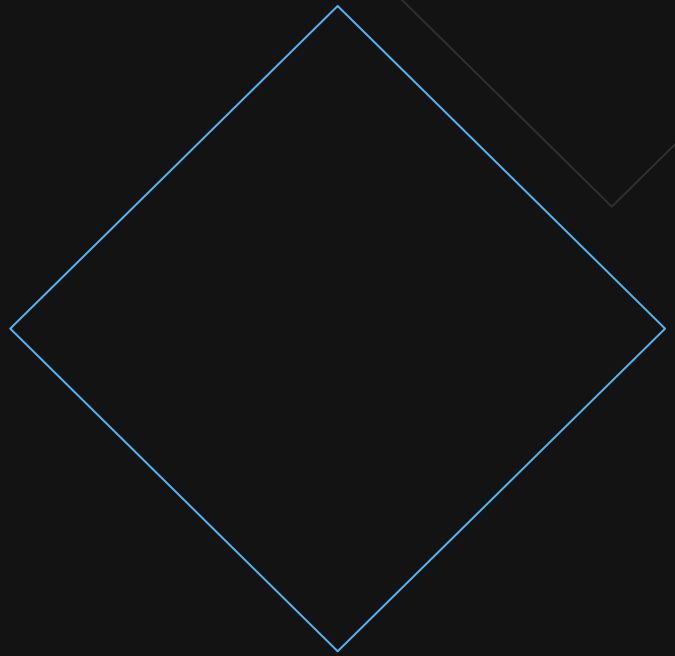
Anomaly detection

Algorithms

- Isolation Forest
- Local Outlier Factor
- ARIMA



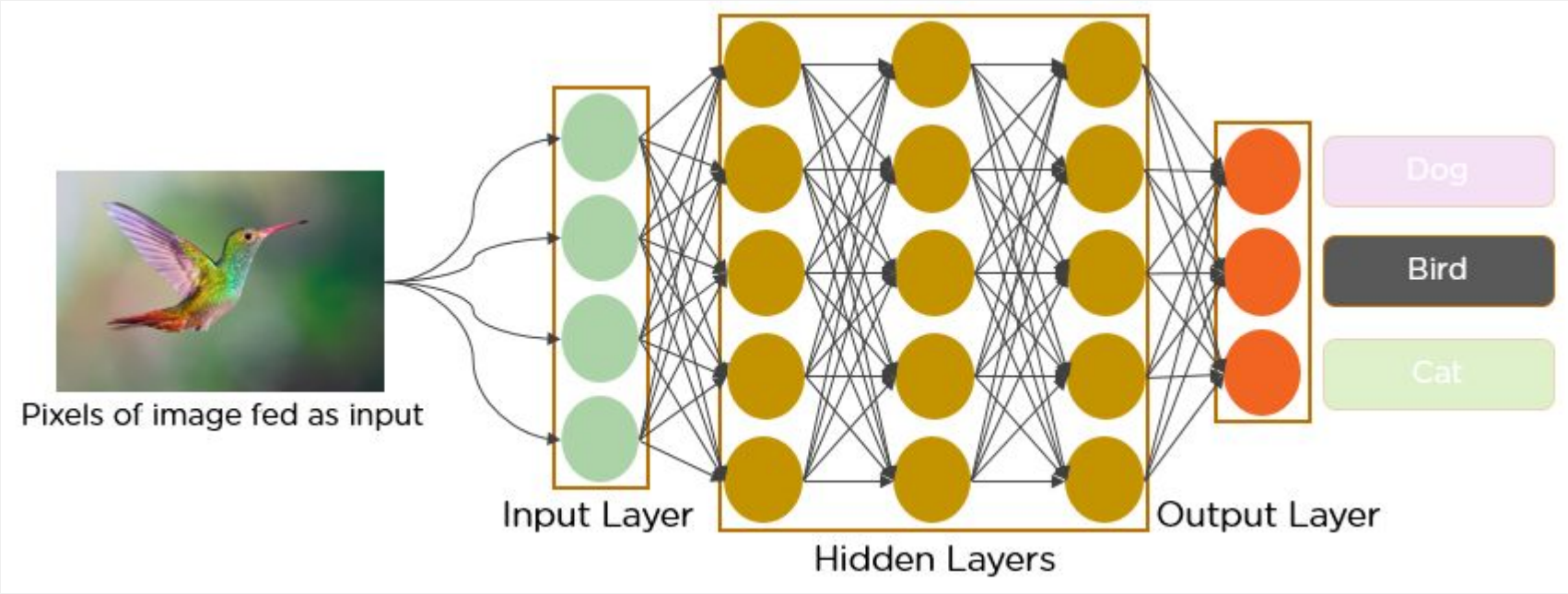
Deep Learning



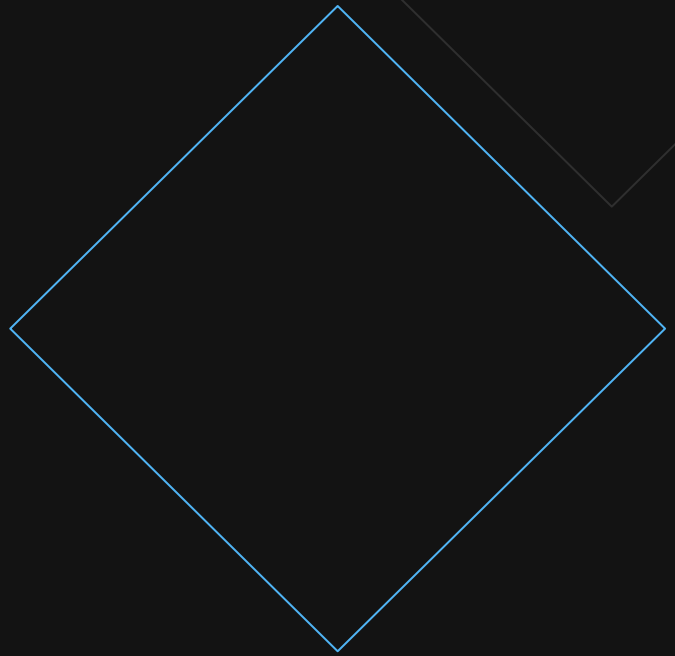
What is Deep Learning

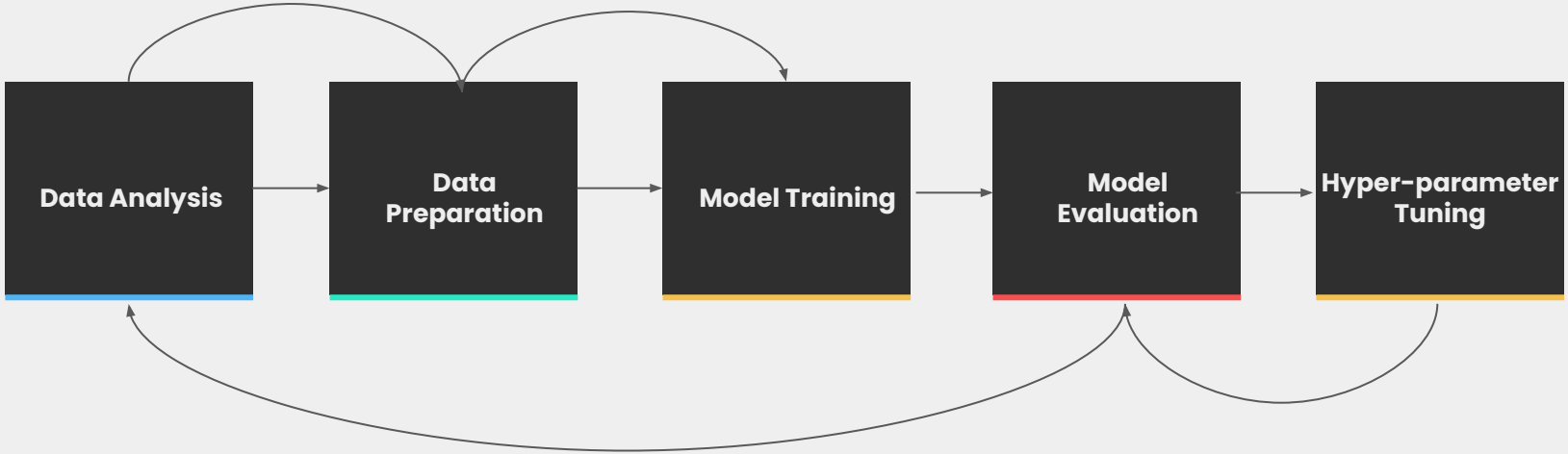
Deep Learning is a subfield of machine learning that focuses on algorithms inspired by the structure and function of the brain, called artificial neural networks. It's "deep" because the neural networks have many layers, which allows for more complex and abstract representations of data.

Neural Network



Training models




Data Analysis
Data Preparation
Model Training
Model Evaluation
Hyper-parameter Tuning

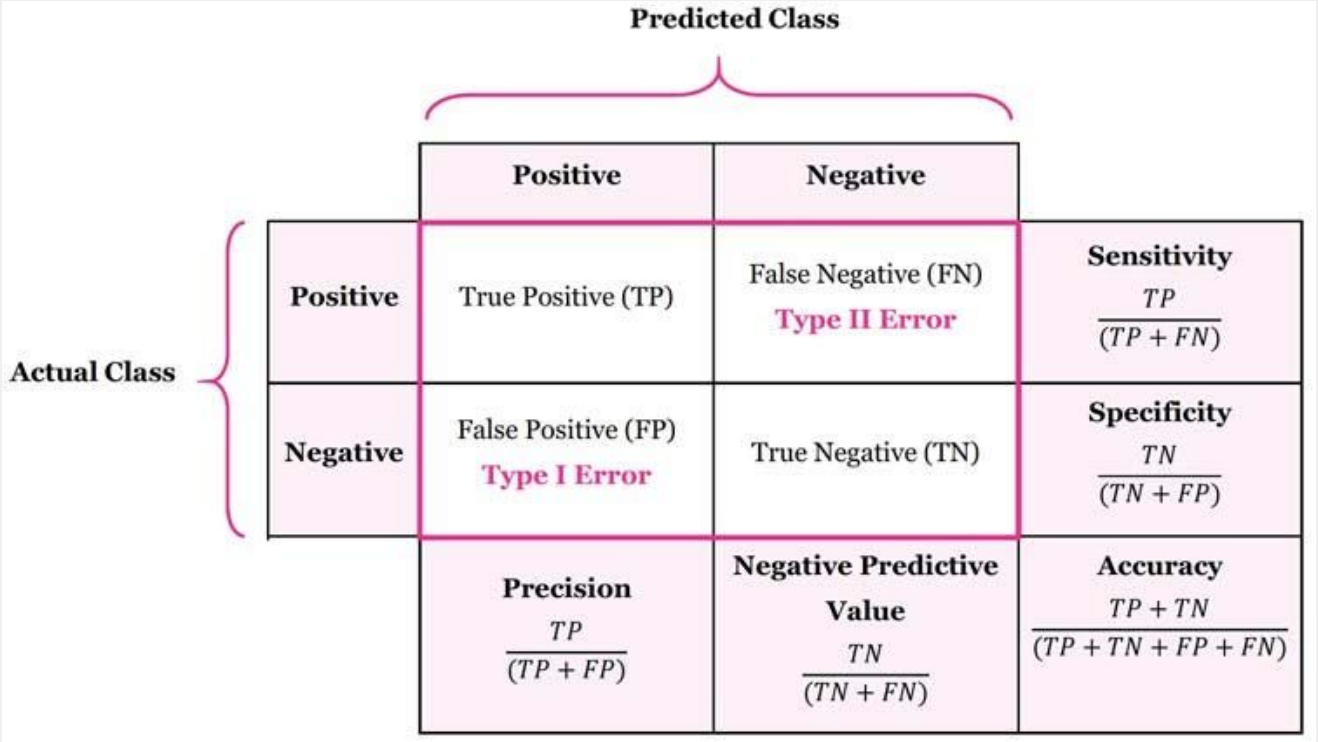
Data Preparation

- Cleaning
  - Remove null values
  - Treat outliers
  - Remove unacceptable values
- Normalization
- Encoding
- Feature engineering

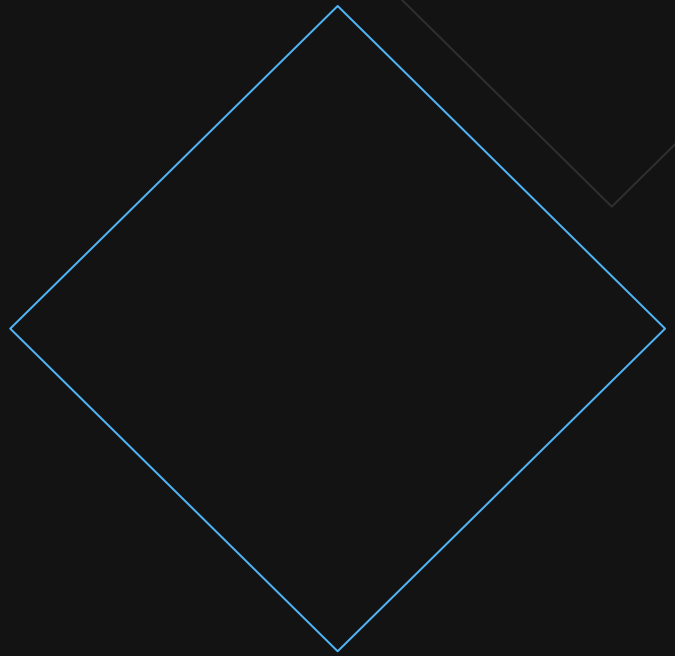
Model Training

- Define the problem
  - What is the output
- Split train and test data
  - Test data is never seen by the training algorithm
  - If model is time-dependent split time based
- Test multiple algorithms

Model Evaluation Metrics - Classification



ChatGPT



GPT4 vs ChatGPT

GPT4

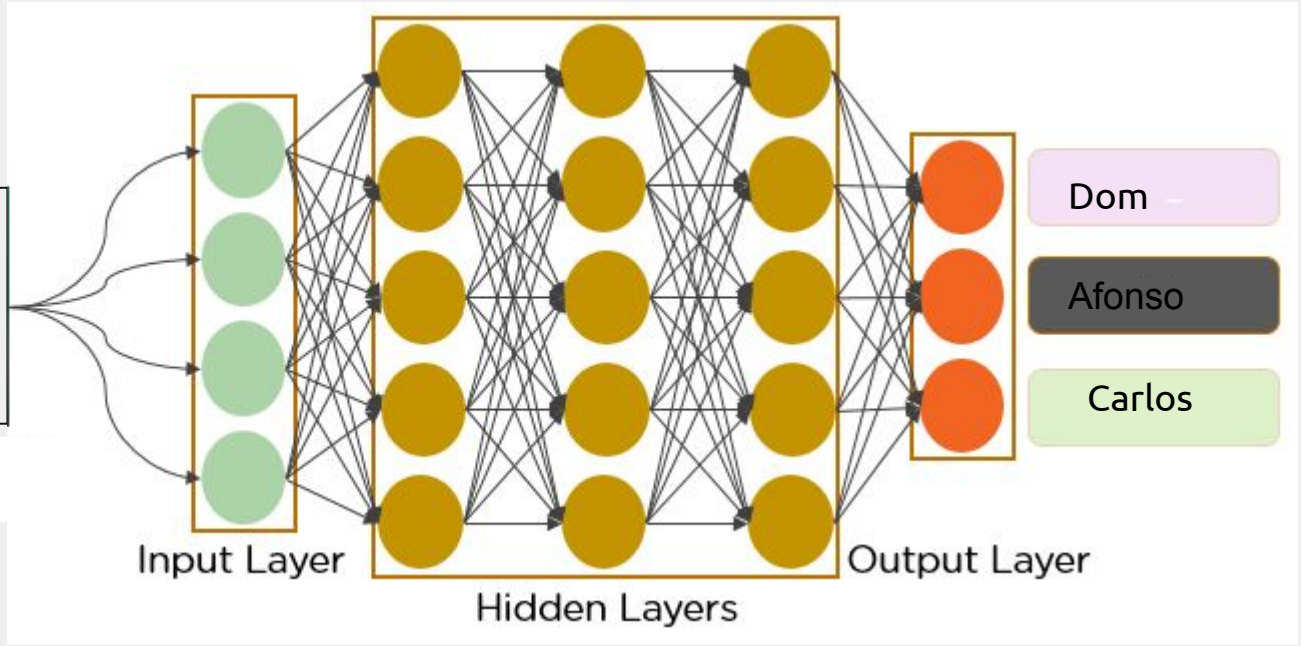
- Neural Network
  - Large Language Model (LLM)
  - Predict next word
- Supervised learning
- Trained on **Millions** of text data points

ChatGPT

- Application that uses an LLM
- GPT4 + fine tuning
  - Reinforcement learning from human feedback (RLHF)

GPT

Quem foi o primeiro rei de Portugal?



Q&A

Thank you

Find out more about us
daredata.engineering

Or read one of our Articles at
blog.daredata.engineering

Email
contact@daredata.engineering

Phone
(PT) +351 932174951

Address
Rua da Prata, n.º 80. 1700-051, Lisbon, PT

DAREDATA
ENGINEERING