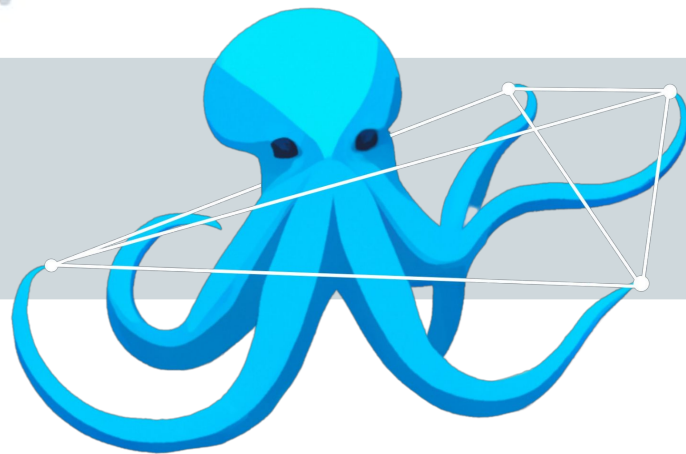


Machine Learning Month

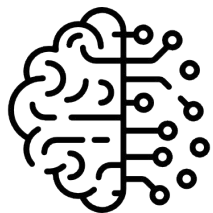
Introduction to ML with Sklearn



Who are we?

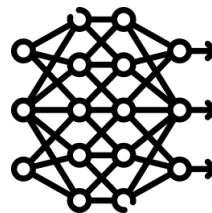
Fully Connected Graph

Overview of the Month



Intro to ML
Week 1

Natural Language
Processing
Week 2



Neural Networks
Week 3

Award Ceremony



Welcome to the First Lecture

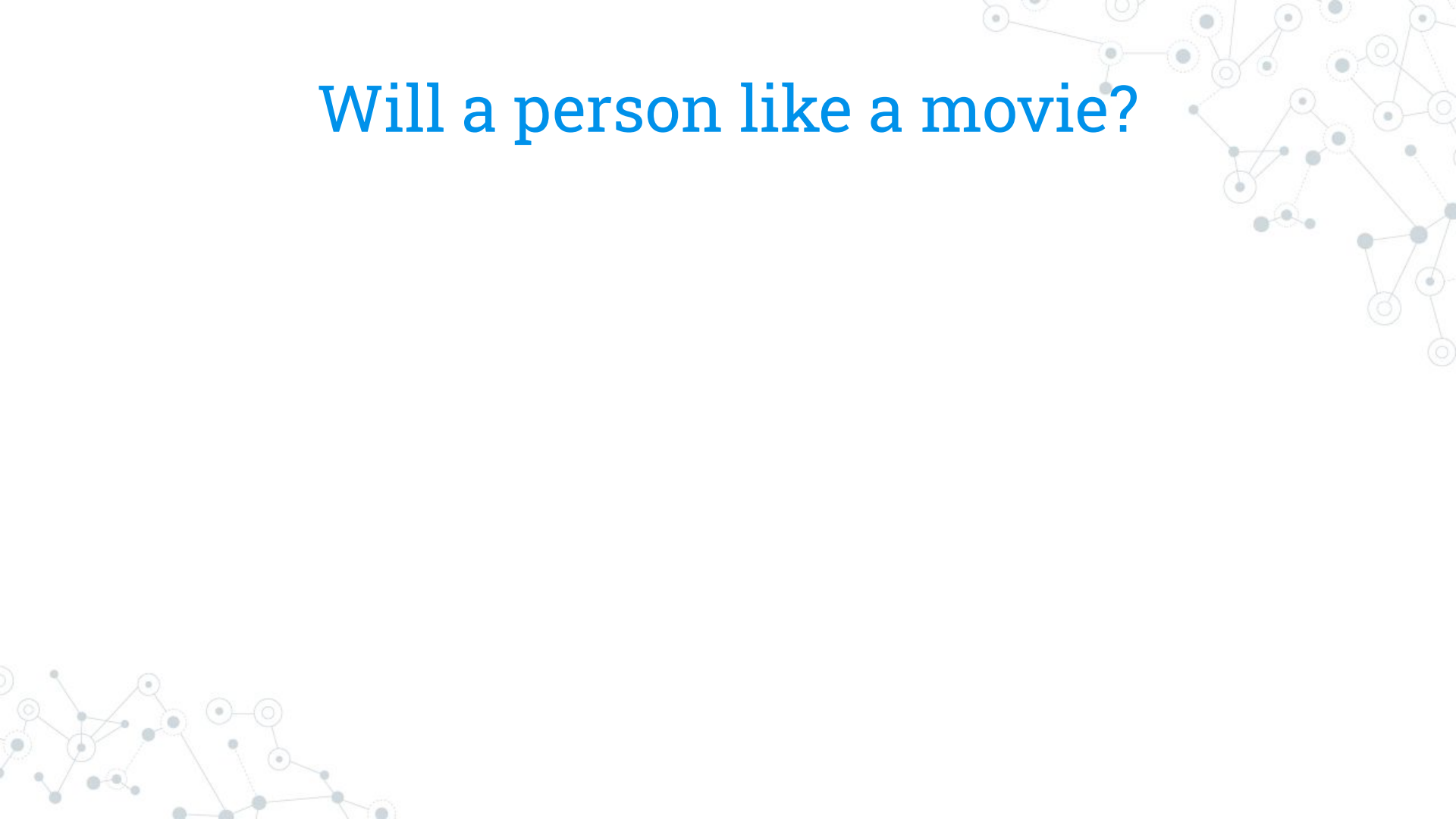
Our Objectives:

- ① Why ML? Types of problems we can solve
- ① Identify types of data and how to prep them
- ① Multiple machine learning algorithms
- ① Practice implementation in Colab
- ① Prepare for the Kaggle competition



Overview of ML

Will a person like a movie?



Will a person like a movie?



Yes.

Will a person like a movie?



Yes.



Will a person like a movie?



Yes.





Definitions

Some Definitions

- ◎ x – the object, it's features
- ◎ X – the space of objects
- ◎ $y(x)$ – the answer for an object, target feature
- ◎ Y – the space of answers

Types of Machine Learning

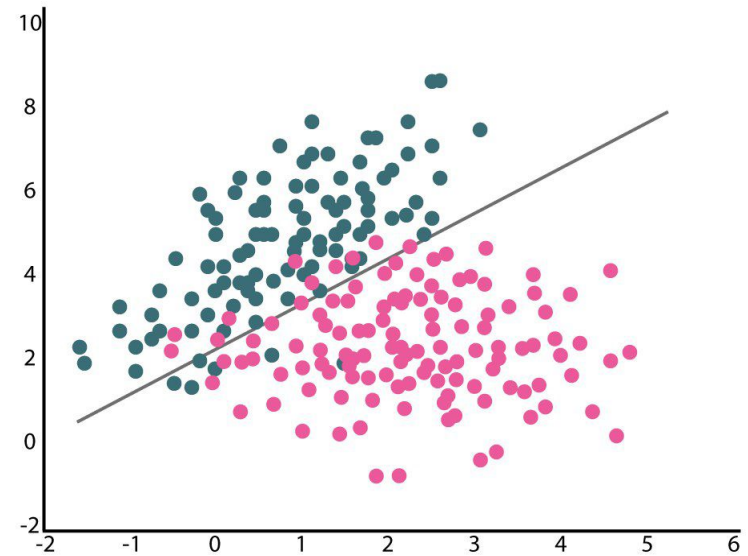
Supervised Learning:

- ◎ Data provided is complete
- ◎ Useful for prediction and classification

Example of Machine Learning Tasks

- ⊙ Will a user like a film?
- ⊙ Will a person return a loan?

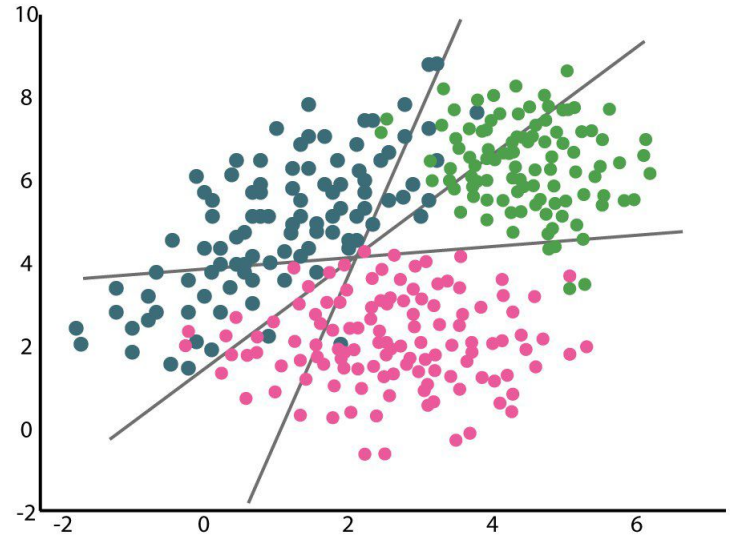
$$Y = \{0, 1\}$$



Binary Classification

Example of Machine Learning Tasks

- ⊙ What is the topic of an article?
- ⊙ What sort of an apple is this?
- ⊙ What type of vehicle is in the image: motorcycle, car or a van?

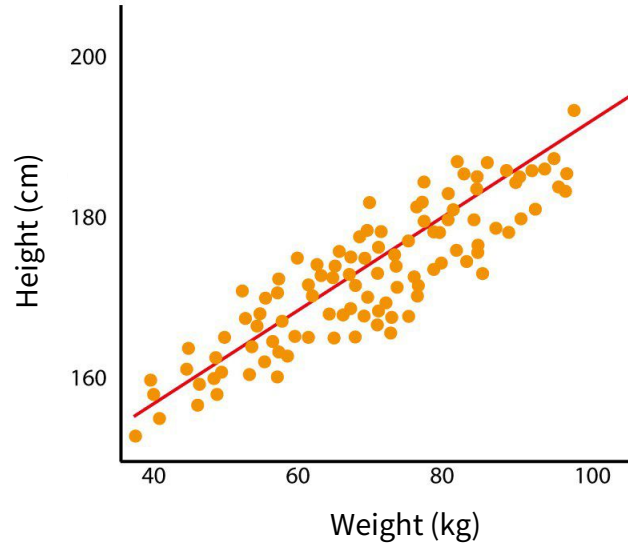


Multi-Class Classification

Example of Machine Learning Tasks

- ⊙ Weather forecast for tomorrow
- ⊙ Revenue prediction
- ⊙ Determining age by photograph

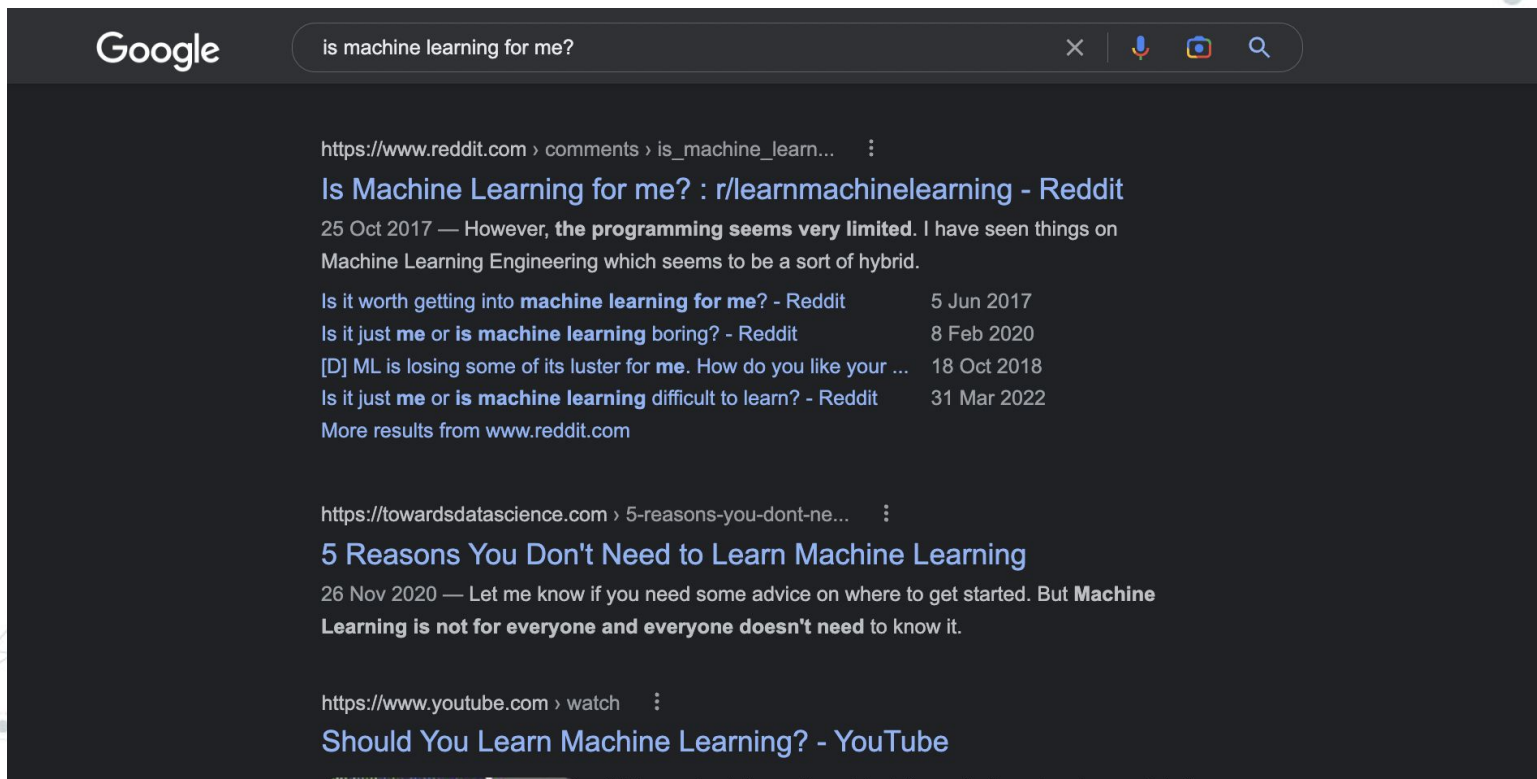
y is real valued



Regression

Example of Machine Learning Tasks

Ranking



The image shows a screenshot of a Google search results page. The search query is "is machine learning for me?". The results are ranked and include:

- [https://www.reddit.com > comments > is_machine_learn...](https://www.reddit.com/comments/is_machine_learn...)
 - Is Machine Learning for me? : r/learnmachinelearning - Reddit**

25 Oct 2017 — However, **the programming seems very limited**. I have seen things on Machine Learning Engineering which seems to be a sort of hybrid.
 - Is it worth getting into **machine learning for me?** - Reddit 5 Jun 2017
 - Is it just **me** or **is machine learning** boring? - Reddit 8 Feb 2020
 - [D] ML is losing some of its luster for **me**. How do you like your ... 18 Oct 2018
 - Is it just **me** or **is machine learning** difficult to learn? - Reddit 31 Mar 2022
 - More results from www.reddit.com
- [https://towardsdatascience.com > 5-reasons-you-dont-ne...](https://towardsdatascience.com/5-reasons-you-dont-ne...)
 - 5 Reasons You Don't Need to Learn Machine Learning**

26 Nov 2020 — Let me know if you need some advice on where to get started. But **Machine Learning is not for everyone and everyone doesn't need** to know it.
- [https://www.youtube.com > watch](https://www.youtube.com/watch)
 - Should You Learn Machine Learning? - YouTube**


Types of Machine Learning

A decorative network diagram in the top right corner, consisting of various sized circles (nodes) connected by thin lines (edges). Some nodes are solid grey, while others are hollow white with a grey outline. The connections form a complex, interconnected web.

Supervised Learning:

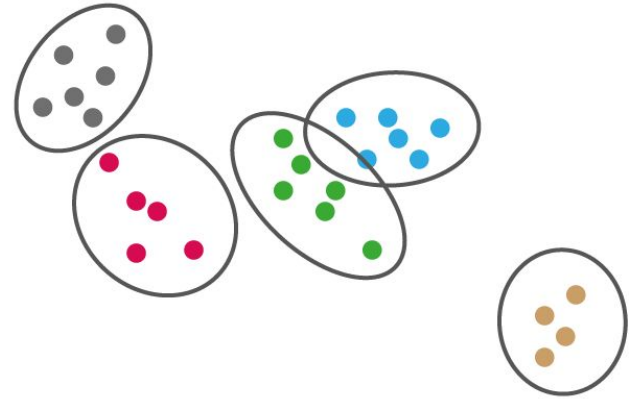
- ◎ Data provided is complete
- ◎ Useful for prediction and classification

Unsupervised Learning:

- ◎ Data is missing the goal value
 - ◎ Useful for uncovering hidden patterns
- 
- A decorative network diagram in the bottom left corner, similar to the one in the top right. It features a cluster of nodes connected by lines, with some nodes highlighted in solid grey and others in hollow white.

Example of Machine Learning Tasks

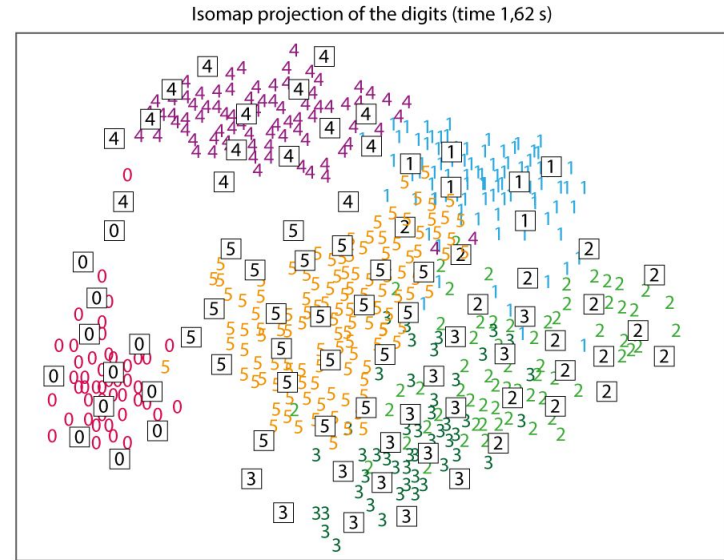
- ◎ User segmentation
- ◎ Search for similar users in social media
- ◎ Search for genes with similar representations



Clustering

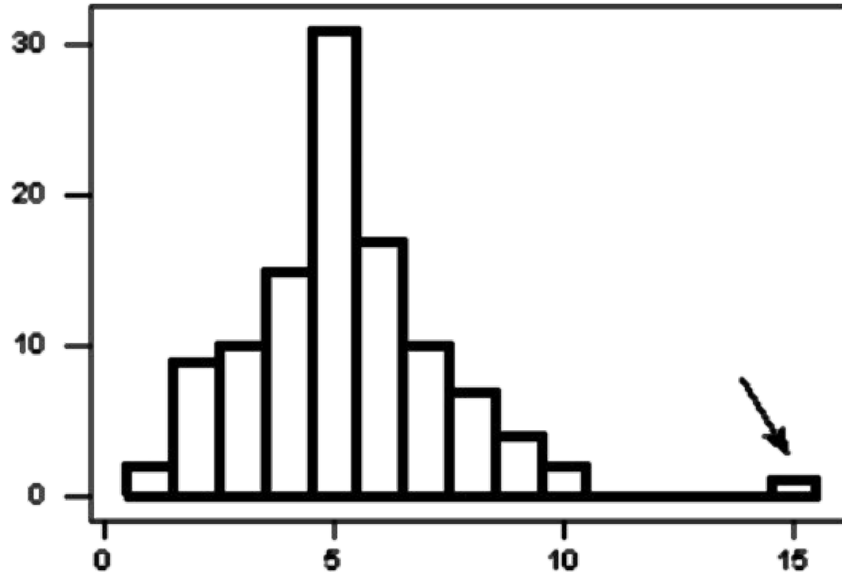
Example of Machine Learning Tasks

- Visualising d-dimensional data in a way that visually shows the structure of data



Visualisation

Example of Machine Learning Tasks



Anomaly detection

The background is a dark blue gradient. It is decorated with several light blue elements: solid dots and hollow circles of varying sizes, scattered across the frame. Some are clustered on the left side, while others are more sparsely distributed on the right and top.

Types of Data

Binary features

Can take up one of two values $D_j = \{0, 1\}$

- ⊙ Is this a cat?
- ⊙ Did the revenue increase?

Answer yes corresponds to 1, the answer no – 0

Continuous features

Can take up values $D_j = \mathbb{R}$

- ⊙ Age
- ⊙ Area of an apartment
- ⊙ Number of products bought

The last feature corresponds to the set of natural numbers, but it's treated as continuous

Categorical features

They take up values from D_j – an unordered set

- ◎ Color of eyes
- ◎ City
- ◎ Education Level

Categorical features

They take up values from D_j – an unordered set

- ◎ Color of eyes
- ◎ City
- ◎ Education Level (sometimes can have an order)

They are hard to deal with. There are new methods made of how to account for them in machine learning models.

Ranking features

Type of categorical features D_j that can be ordered

- ◎ Role in the movie (main, secondary, background)
- ◎ Type of populated area (city, town, village)
- ◎ Education Level (PhD, Master, Bachelor, Undergrad)

Ranking features

Type of categorical features D_j that can be ordered

- ◎ Role in the movie (main, secondary, background)
- ◎ Type of populated area (city, town, village)
- ◎ Education Level (PhD, Master, Bachelor, Undergrad)

The *distance* between two features doesn't make sense

Features of a dataset

| | A | B | C | D | E | F | G |
|----|----|-----------------|----------------|------------|-----------|-----------|---------|
| 1 | id | title | city | postalCode | latitude | longitude | areaSqm |
| 2 | 0 | West-Varkenoor | Rotterdam | 3074HN | 51.896601 | 4.514993 | 14 |
| 3 | 3 | Ruiterakker | Assen | 9407BG | 53.013494 | 6.561012 | 16 |
| 4 | 8 | Brusselseweg | Maastricht | 6217GX | 50.860841 | 5.671673 | 16 |
| 5 | 10 | Donkerslootstra | Rotterdam | 3074WL | 51.893195 | 4.516478 | 25 |
| 6 | 12 | Vorselenburgstr | Alphen aan den | 2405XJ | 52.122335 | 4.661434 | 10 |
| 7 | 17 | Groenhoven | Amsterdam | 1103LW | 52.326211 | 4.976048 | 19 |
| 8 | 18 | Noorderhagen | Enschede | 7511EL | 52.221643 | 6.894667 | 21 |
| 9 | 19 | Jaersveltstraat | Rotterdam | 3082SJ | 51.890481 | 4.466388 | 16 |
| 10 | 20 | Tongerseweg | Maastricht | 6213GB | 50.841744 | 5.670447 | 17 |
| 11 | 21 | Lange Marktstra | Leeuwarden | 8911AD | 53.197261 | 5.790455 | 19 |
| 12 | 22 | Guido Gezellest | Eindhoven | 5615HL | 51.431324 | 5.475464 | 16 |
| 13 | 23 | Ank van der Mo | Amsterdam | 1065LH | 52.352244 | 4.824007 | 12 |
| 14 | 25 | Beatrixstraat | Enschede | 7511KL | 52.221827 | 6.902143 | 20 |
| 15 | 28 | Tesselschadestr | Leeuwarden | 8913HA | 53.198638 | 5.782587 | 51 |

Target feature

Targets refer to the values that we are trying to predict

In the competition, the only target is the rent of a property

| rent |
|------|
| 500 |
| 290 |
| 425 |
| 600 |
| 425 |
| 750 |
| 240 |
| 500 |
| 660 |
| 412 |

The background is a dark blue gradient. Scattered across the top and bottom edges are several light blue decorative elements: some are solid circles of varying sizes, and others are hollow circles. The text "Break Time" is centered in a white, bold, serif font.

Break Time


The background is a dark blue gradient. It is decorated with several light blue elements: solid dots of varying sizes and hollow circles of varying diameters, scattered across the frame. The text is centered in a bold, white, sans-serif font.

Data Preprocessing

What is Data Preprocessing?

A decorative network diagram in the top right corner, consisting of various sized grey circles (nodes) connected by thin grey lines (edges). Some nodes are highlighted with a darker grey or blue color.

Before passing data to a ML model, it has to be prepared.

- ⦿ Data normalization
 - ⦿ Missing values
 - ⦿ Handling categorical features
- 
- A decorative network diagram in the bottom left corner, consisting of various sized grey circles (nodes) connected by thin grey lines (edges). Some nodes are highlighted with a darker grey or blue color.


Handling Categorical Data



We can use one hot encoding to represent categorical data.

In this technique, we represent the item as multiple binary values for each possible outcome.

Ex: Suppose we had a category ‘color’ which consisted of “red”, “green” and “blue”



Handling Categorical Data

In one-hot encoding, the data would be represented as:

| ID | Red | Green | Blue |
|----|-----|-------|------|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |

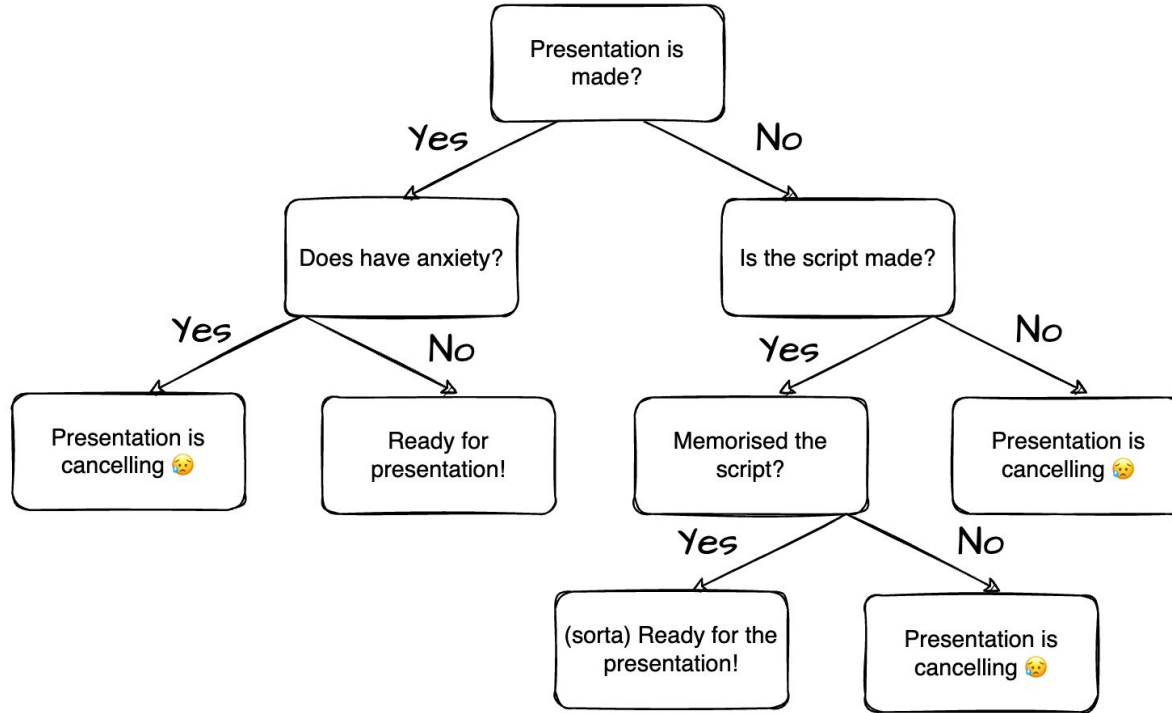
The background is a dark blue gradient. Scattered across the page are several light blue decorative elements: solid circles of various sizes and hollow circles. Some are clustered on the left side, while others are more sparsely distributed on the right and top.

ML Model

A few ML techniques

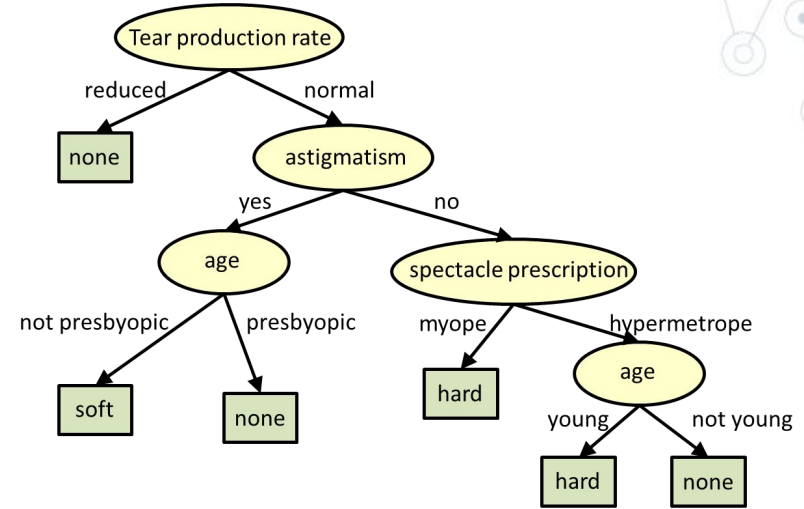
- ⊙ Decision Trees
- ⊙ Random Forests
- ⊙ Linear Regression
- ⊙ Gradient Boosting
- ⊙ Neural Networks (Discussed in Lecture 3)

Decision Trees



Decision Trees

- Predicts the dependent variable using inference rules from the given data
- Groups samples with similar values together
- Arrives at a prediction by answering a series of if ... else statements



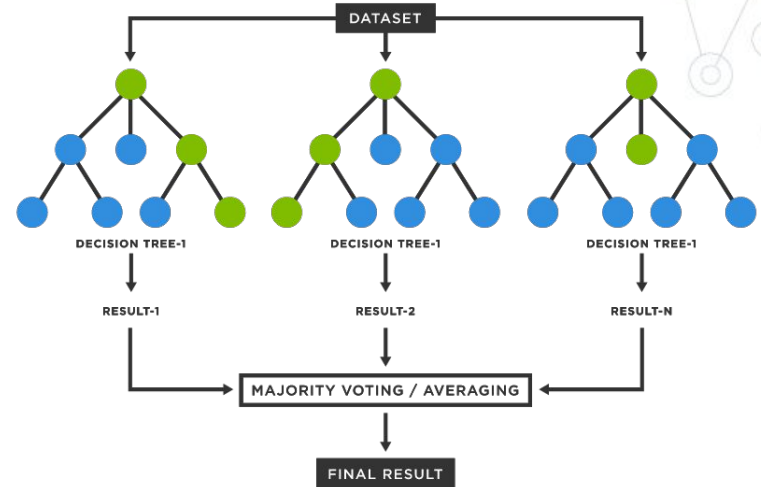
Decision Trees

| Strengths | Weaknesses |
|--|--|
| Can be visualized and explained easily | Prone to overfitting |
| Can handle both numerical and categorical features | Unstable since even small alterations can change results |



Random Forests

- Derivative of Decision Trees
- The dataset is randomly split into several components and a decision tree is created from each of these
- The output value is the value majority voted for



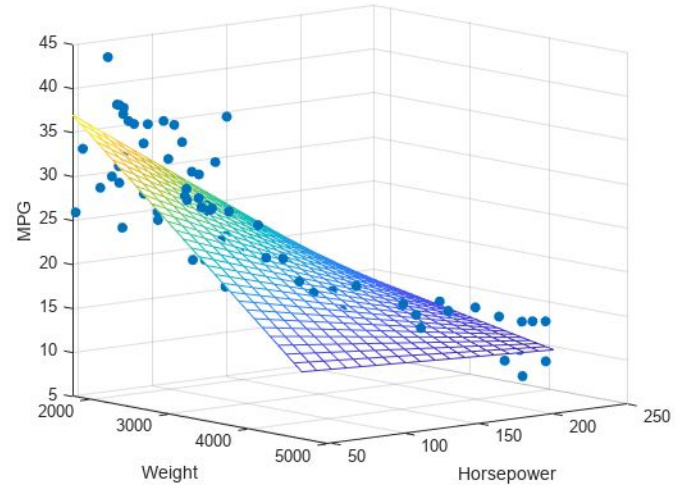
Random Forests

| Strengths | Weaknesses |
|--------------------------|--------------------------|
| Resistant to overfitting | Loss of Interpretability |
| Higher accuracy than DTs | |



Linear Regression

- Simple model that linearly predicts the goal given features.
- The objective of linear regression is to generate a line of best fit which minimizes the Residual Sum of squares.



$$y(x_1, x_2, \dots, x_n) = c_0 + c_1x_1 + c_2x_2 + \dots + c_nx_n$$

$$y(\mathbf{x}) = c_0 + \mathbf{c}\mathbf{x}$$

Linear Regression

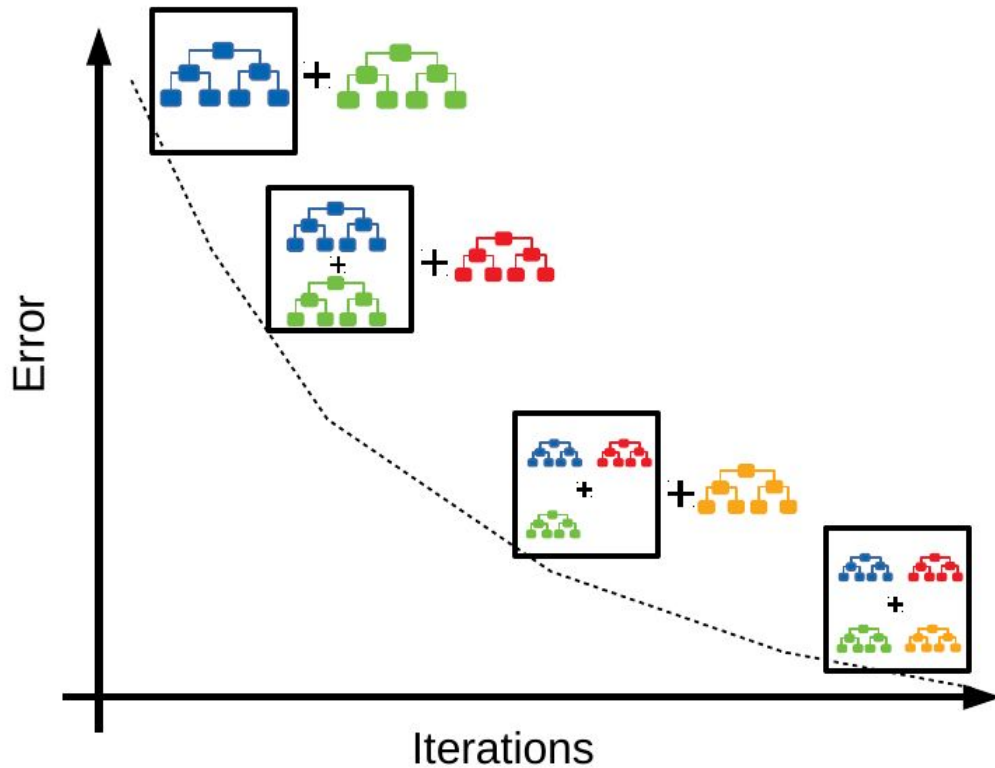
| Strengths | Weaknesses |
|--|----------------------------------|
| Performs well for linearly separable data | Sensitive to Outliers |
| Resistant to overfitting due to generalization | Cannot handle categorical values |



Gradient Boosting



Gradient Boosting



Gradient Boosting

- ① Derivatives from decision trees
- ① Further trains the trees on sections of the data that it has problems with
- ① Over time, these weaknesses are covered, leading to high accuracy



Gradient Boosting

| Strengths | Weaknesses |
|----------------------------------|--------------------|
| High Accuracy | Expensive to train |
| Highly flexible and customizable | Prone to overfit |



The background is a dark blue gradient. It features several decorative elements: a cluster of blue dots and hollow circles in the top-left corner, and another cluster of blue dots and hollow circles in the bottom-right corner. The text 'Practice Session' is centered in a bold, white, sans-serif font.

Practice Session

Installation

The logo for Google Colab, featuring the word "colab" in a bold, lowercase, orange sans-serif font. The letters are slightly shadowed, giving it a 3D appearance.

colab

Machine Learning is built using an extensive set of libraries.

It can be difficult to get code working locally


For convenience, we can use [Google Colab](#) to save time

Using Colab

A decorative network diagram in the top right corner, consisting of various sized circles (nodes) connected by thin lines (edges). Some nodes are solid grey, while others are hollow white with a grey border. The connections form a complex, interconnected web.

Jupyter notebooks allow us to execute code while also using markdown to provide comments

Open the link and go to the lecture materials
mlmonth.svcover.nl

A decorative network diagram in the bottom left corner, similar to the one in the top right. It features a cluster of nodes connected by lines, with some nodes highlighted in solid grey and others in hollow white with grey borders.

The Problem

We have acquired a dataset of the grades of students in Portugal throughout a year. The dataset also contains their demographic data as well.

Using the acquired data, we wish to predict the final grade (G3) of a student.

Tools

The image features a dark blue background with the word "Tools" centered in a large, white, bold, sans-serif font. Scattered around the text are several decorative elements: small solid blue dots and larger hollow blue circles. These elements are primarily located in the upper-left and lower-right quadrants, creating a sense of depth and movement. The overall aesthetic is clean and modern.

Model training with sklearn

- ⦿ Free ML Library in Python
- ⦿ Contains useful features such as pre-processing
- ⦿ Contains implementations of several ML models





Getting Started is Simple

```
pip install sklearn
```



Data Handling with Pandas

Along with sklearn, pandas is helpful in handling data

The pandas dataframe is powerful and allows for convenient data access.

Installation: `pip install pandas`



Training

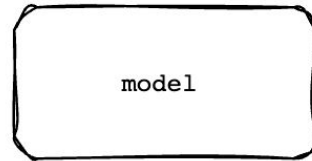
Train data



Learning algorithm

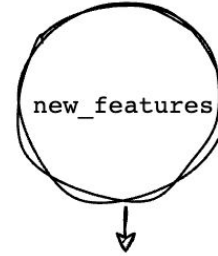


Trained algorithm

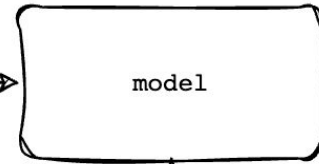


Application

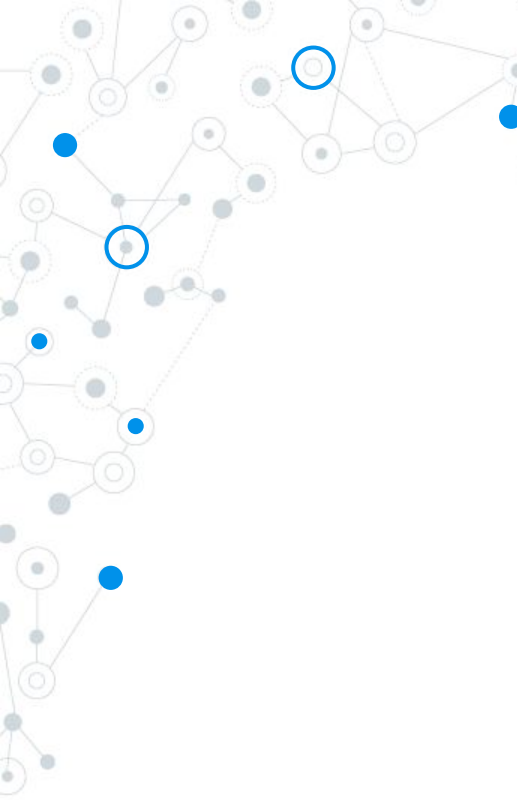
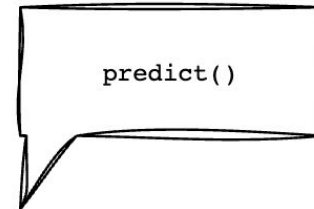
new data



Trained algorithm



Predictions





Kaggle Preparation

Preparation for the Competition

Important things to know:

- ① [Link to Competition](#)
- ② Read the New to Kaggle section
- ③ Build your model
- ④ Test your model
- ⑤ Submit your results
- ⑥ Improve!



November 24

Introduction to
Natural Language
Processing



Thank you for your attention!

Do you have any
more questions?
Join our [Discord](#)
server

