Machine learning: prediction, classification and clustering

UBB Faculty of Sociology

Hello!
I’m Sergiu Zaharie

Machine Learning Engineer / Data Scientist
Senior Data Scientist @BetFair Romania Development

zahariesergiu@gmail.com
+40742087160
https://www.linkedin.com/in/zahariesergiu/

Course Agenda

#1 Intro, Simple Linear Regression

#2 Python recap, Git, Handling data, EDA

#3   Regression, Decision Trees

#4   Bias, Variance, Overfitting, Classification, Metrics

#5   Random Forest Classifier, Clustering

#6   Neural Newtorks

#7   Help Final Project

#8   Help Final Project

How it works

- Grade based on attendance, participation and final project.
- Collaborative course: Google Colab as our main support.

1. Intro to Machine Learning

#1.1 What is Machine Learning
#1.2 Visual example
#1.3 Classification Exercise
#1.4 What is Machine Learning?
#1.5 Types of Machine Learning
#1.6 App examples exercise
#1.7 Simple Linear Regression
#1.8 Questions & Further reading

1.1 What is Machine Learning?

How many people here have heard about machine learning?

How many people here have heard about machine learning?

How many people here have done some kind of machine learning course, training, or technical reading?

How many people here have heard about machine learning?

How many people here have done some kind of machine learning course, training, or technical reading?

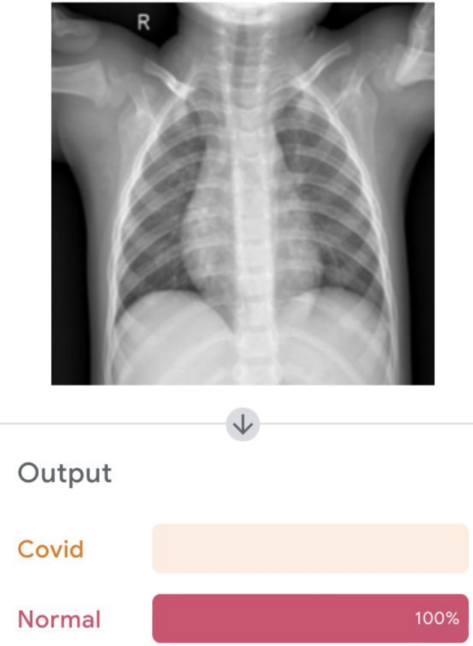
ML VS AI?

Machine Learning is a subset of Artificial Intelligence that enables systems to learn from data, while AI covers all technologies that mimic human intelligence.

Non ML: Chatboots, hard-coded AI game

1.2 Visual Example

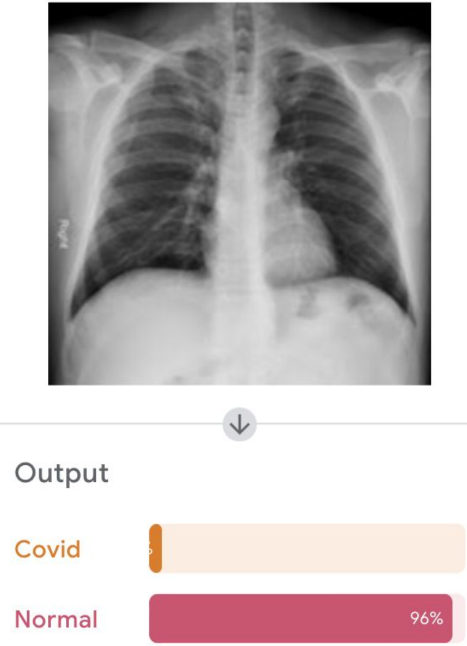
https://teachablemachine.withgoogle.com/



Output

Covid

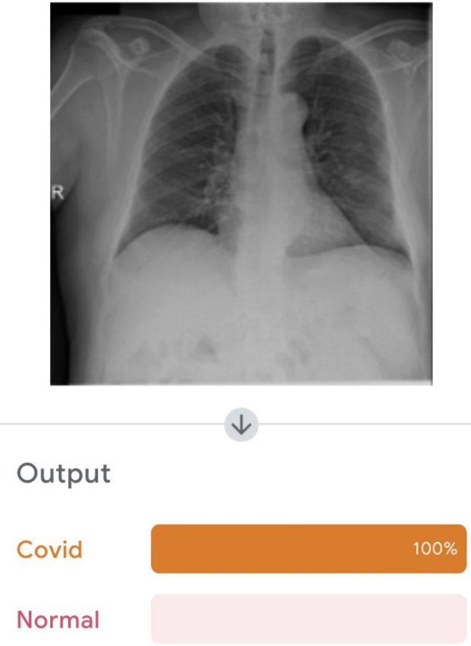
Normal 100%



Output

Covid

Normal 96%



Output

Covid 100%

Normal

Always about **single tasks**.

A particular machine learning system only ever “does one thing”

a **single function**.

1.3 Classification exercise

Suppose you want to build a dataset for Decathlon to classify runners and bikers by reading smartwatch data.

You’ll get to ask everyone in your city park questions, to collect data.

What questions do you ask?

Suppose you want to build a dataset for Decathlon to classify runners and bikers by reading smartwatch data.

You’ll get to ask everyone in your city park questions, to collect data.

What questions do you ask?

1. **What is your average speed during this activity?**
   - Purpose: Speed is a critical factor, as bikers typically move faster than runners.
2. **What is your average heart rate while performing this activity?**
   - Purpose: Runners and bikers may have distinct heart rate patterns due to the intensity and nature of the activity.
3. **Are you currently moving or stationary during your activity?**
   - Purpose: This helps capture whether the individual is in an active phase, like running or cycling, versus resting or pausing.

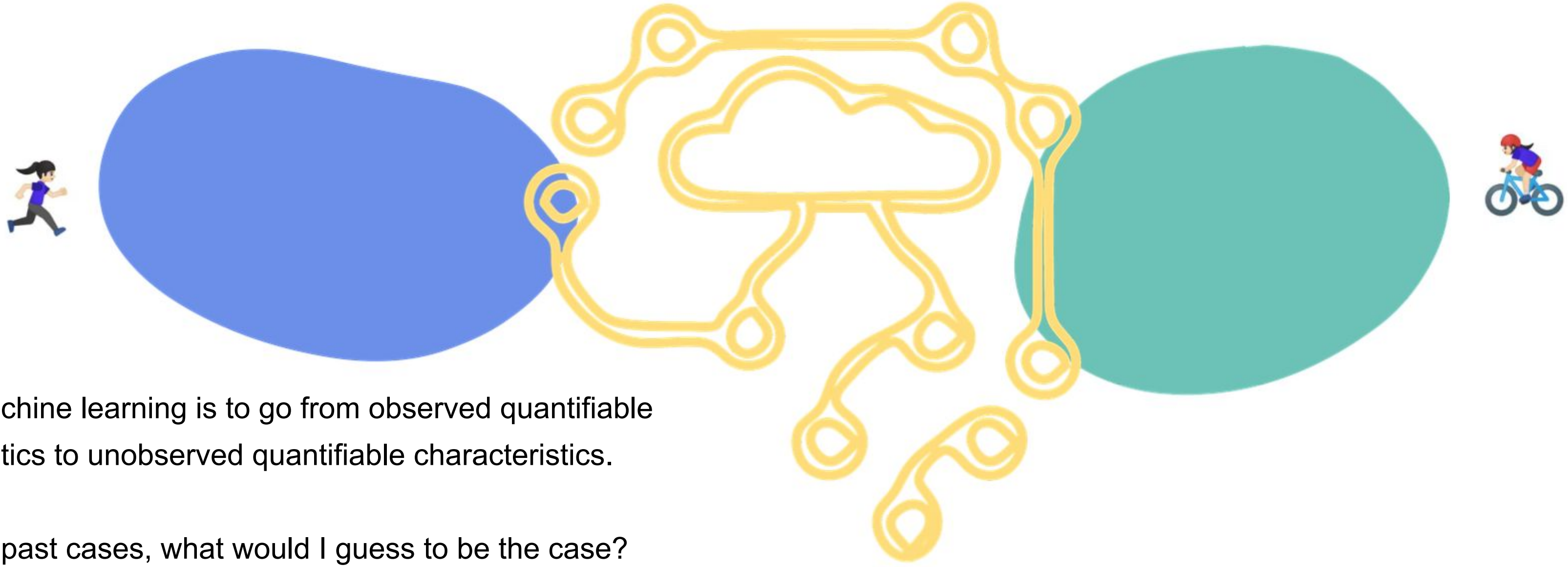
These will be features for your classification model

4. **What is your maximum speed during this activity?**
    - Purpose: Maximum speed can help differentiate between runners (lower max speeds) and bikers (higher max speeds).
5. **What is the typical cadence or step rate you maintain during this activity?**
    - Purpose: Runners have step counts (e.g., strides per minute), while bikers might have a pedal cadence.
6. **Do you take any rest breaks during this activity?**
    - Purpose: Identifying rest periods and how they reflect in the smartwatch data could help classification.
7. **Do you experience frequent stops during your activity, such as at traffic lights or intersections?**
    - Purpose: Helps differentiate bikers (who often stop more frequently) from runners (who may have more continuous movement).
8. **Do you notice significant elevation changes during your activity?**
    - Purpose: Bikers might cover longer elevation profiles, while runners might prefer flatter terrains. Or reversed i.e. trail running
9. **What is the duration of your typical session?**
    - Purpose: Bikers often engage in longer sessions compared to runners.

1.4 What is Machine Learning?

"Machine learning is the science of getting computers to act without being explicitly programmed."

Andrew Ng



Goal of machine learning is to go from observed quantifiable characteristics to unobserved quantifiable characteristics.

Looking at past cases, what would I guess to be the case?

Machine learning is basically fancy ways of asking this question, by associating “features”
“This feature was often associated with riding a biking.”

#1.1 What is Machine Learning?

"Machine learning is the science of getting computers to act without being explicitly programmed."

Andrew Ng

#1.1 What is Machine Learning?

"Machine learning is the science of getting computers to act without being explicitly programmed."

Andrew Ng

Machine learning is how we teach computers to learn from examples instead of giving them step-by-step instructions.

#1.1 What is Machine Learning?

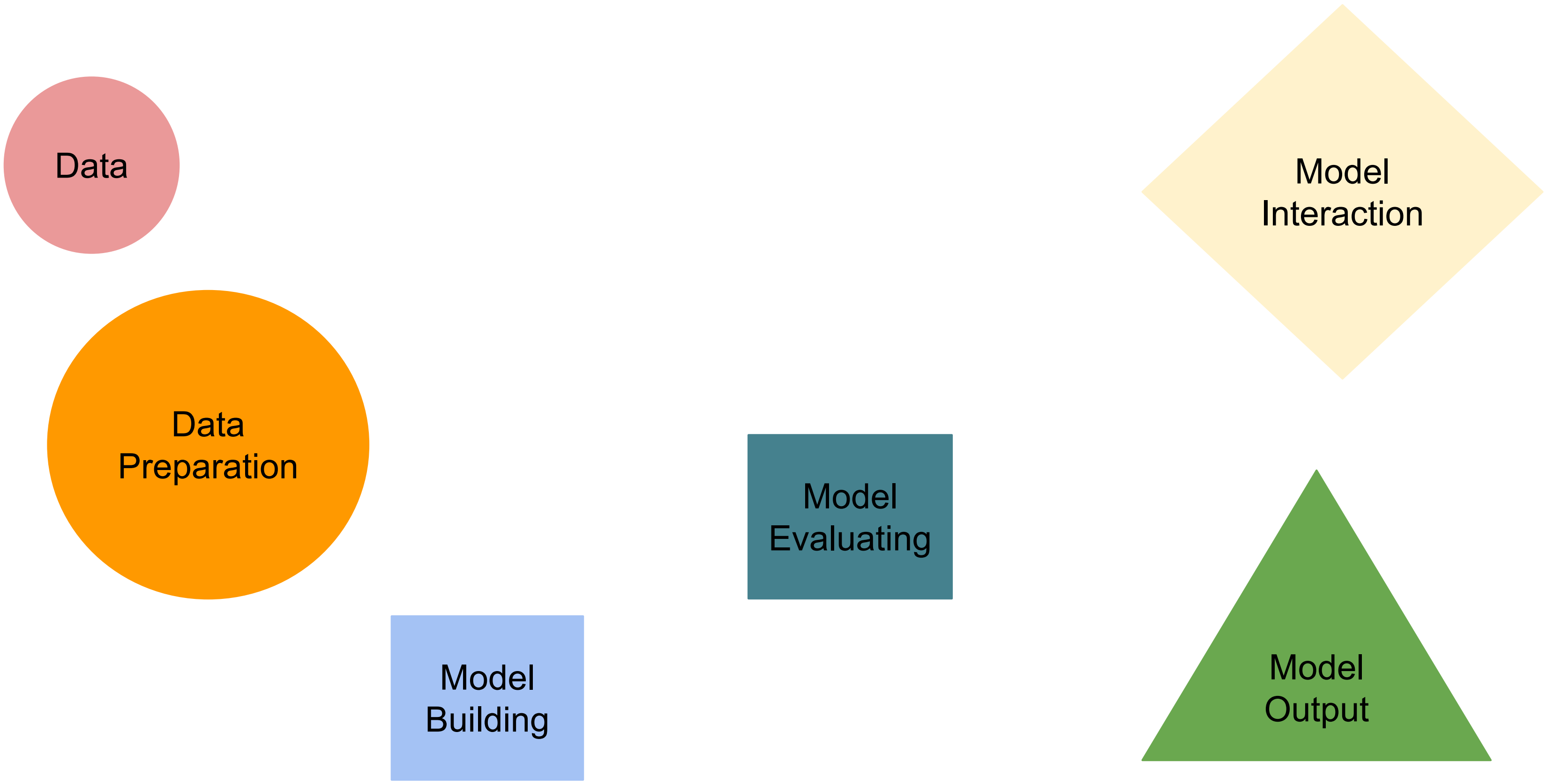
"Machine learning is the science of getting computers to act without being explicitly programmed."

Andrew Ng

Machine learning is how we teach computers to learn from examples instead of giving them step-by-step instructions.

Algorithms that are able to generalize from patterns

learned from data samples



Data

Data Preparation

Model Building

Model Evaluating

Model Interaction

Model Output

Data

- All machine learning models need data
- Where does your data come from?
- What is the type (string/integer) of each of the features (columns)

- Train data - data to train on
- Test data – Keep out of training, used only for testing the model.

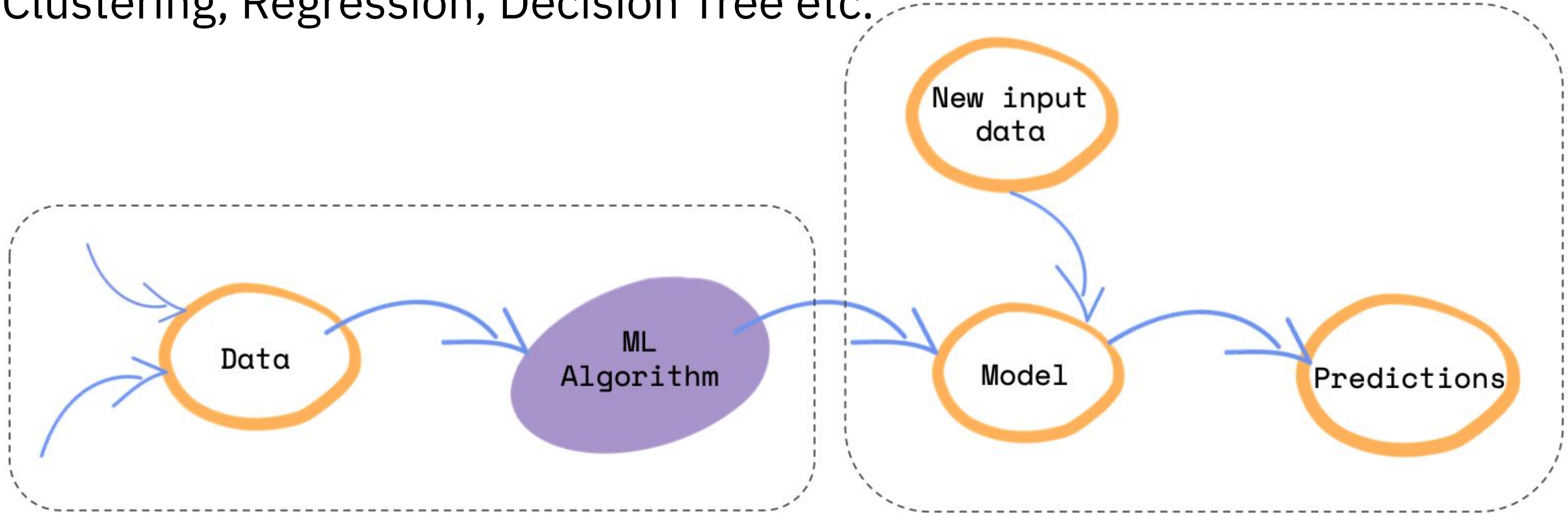
Data Preprocess

- Explore your data - look for trends that might inform you
- Remember - how was your data collected?
- How is it going to be used?

- Format the data in a way that the computer can read it
- Clean
- Normalize - adjusting your data to a common scale ●
- Might choose to exclude missing values
- Feature engineer etc

Model Building

- Ask yourself: What type of problem are you trying to solve?
- Data + Algorithm = **Model**
  - Algorithm:
    - Clustering, Regression, Decision Tree etc.



Model Evaluation

How well can your model [predict] unseen data?

- Train data - data to train on
- **Test data** – Keep out of training, used only for testing the model.
- Metrics:
    - Accuray fractions pf predictions model got right
    - Precision - proportions of positive predictions that are correct
    - Recall - proportions of actual positives that are correctly identified
    - MSE - mean squared difference between actual and predicted
    - RMSE - root mean squared difference between actual and predicted

Model Output

What will the output of your model look like?

Regression Models

Output: **Continuous numeric values** (e.g., predicted price, loan amount, temperature).

Classification / Decision Models

Should the system explain why a loan is denied (e.g. insufficient income) to comply with regulations ?

Output: **Approve / Deny**, or **class probabilities / scores**.

Recommendation / Ranking Models

Helps users discover relevant products or content; can improve engagement and retention. How many? Sorted or not?

Output: **Top-N recommendations** or **ranked items** based on user preferences, behavior, or relevance scores.

Sequence / Time Series Models

Predictions can be **deterministic** (single value) or **probabilistic** (distribution / prediction intervals) to reflect uncertainty in future outcomes.

Output: **Predicted sequences of values** (e.g., stock prices, demand forecasts, next words in text).

Model Interactions

- How do you give feedback to your model?
- How can you leverage the model capabilities to make it more impactful?
- What do you need to do to transform the model output to make it usable?

Examples:

- Collect user feedback
- Regulary add examples of user queries for diverse contexts
- Monitor outputs for biases or errors and correct the in the training data
- User experience - design to improve usability

1.5 Types of Machine Learning

Supervised Learning

Regression - Interpolating data: generate a point

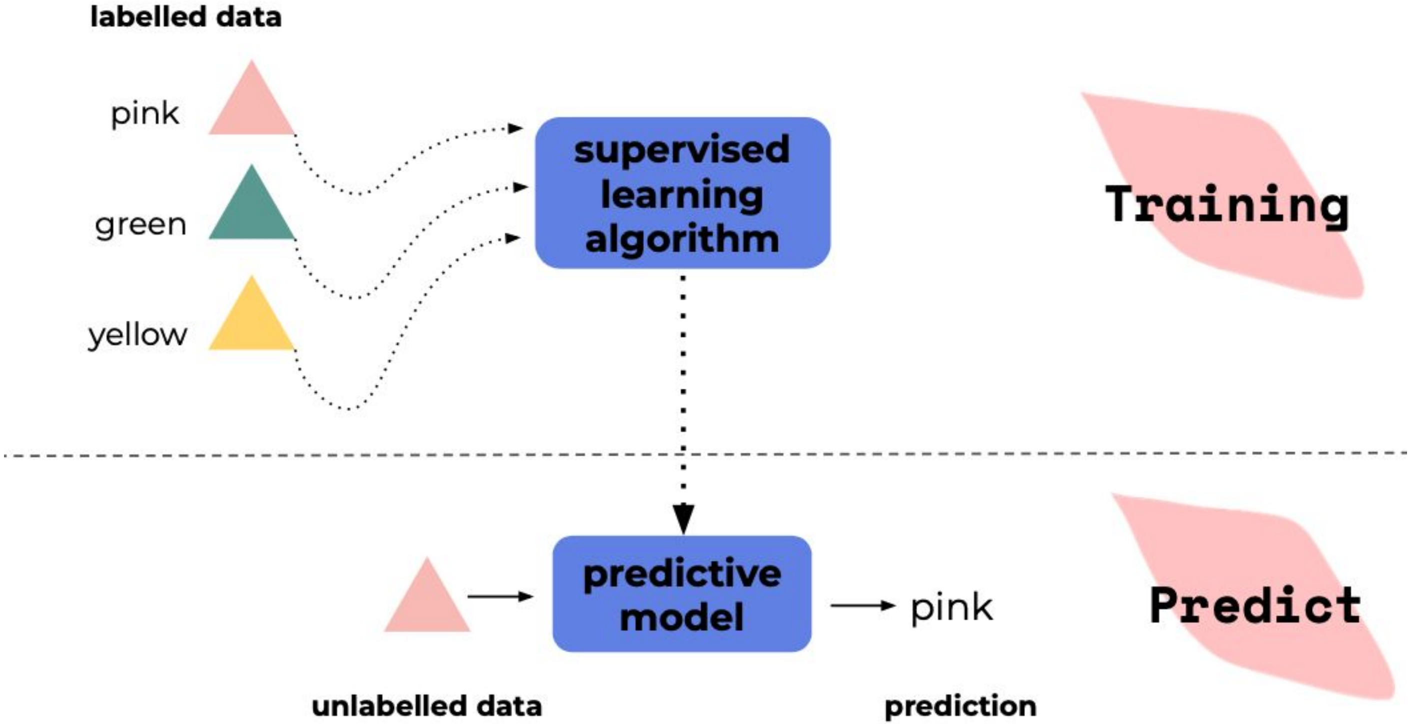
Classification - Label data: dog or cat

Unsupervised Learning

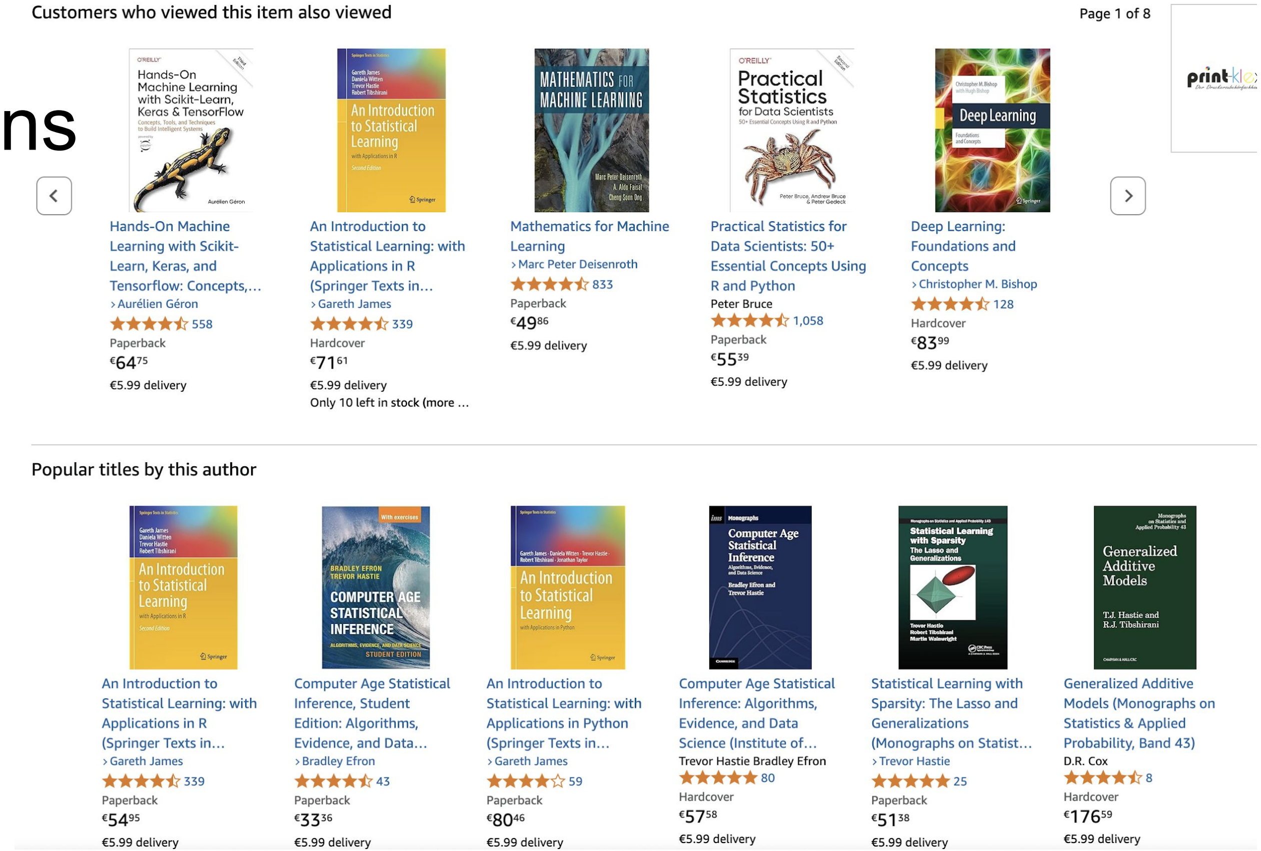
Clustering - Group data

Reinforcement learning

Supervised Learning



Amazon recommendations



Customers who viewed this item also viewed

Page 1 of 8

Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts,…
› Aurélien Géron
558
Paperback
€64⁷⁵
€5.99 delivery

An Introduction to Statistical Learning: with Applications in R (Springer Texts in…
› Gareth James
339
Hardcover
€71⁶¹
€5.99 delivery
Only 10 left in stock (more …

Mathematics for Machine Learning
› Marc Peter Deisenroth
833
Paperback
€49⁸⁶
€5.99 delivery

Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python
Peter Bruce
1,058
Paperback
€55³⁹
€5.99 delivery

Deep Learning: Foundations and Concepts
› Christopher M. Bishop
128
Hardcover
€83⁹⁹
€5.99 delivery

Popular titles by this author

An Introduction to Statistical Learning: with Applications in R (Springer Texts in…
› Gareth James
339
Paperback
€54⁹⁵
€5.99 delivery

Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data…
› Bradley Efron
43
Paperback
€33³⁶
€5.99 delivery

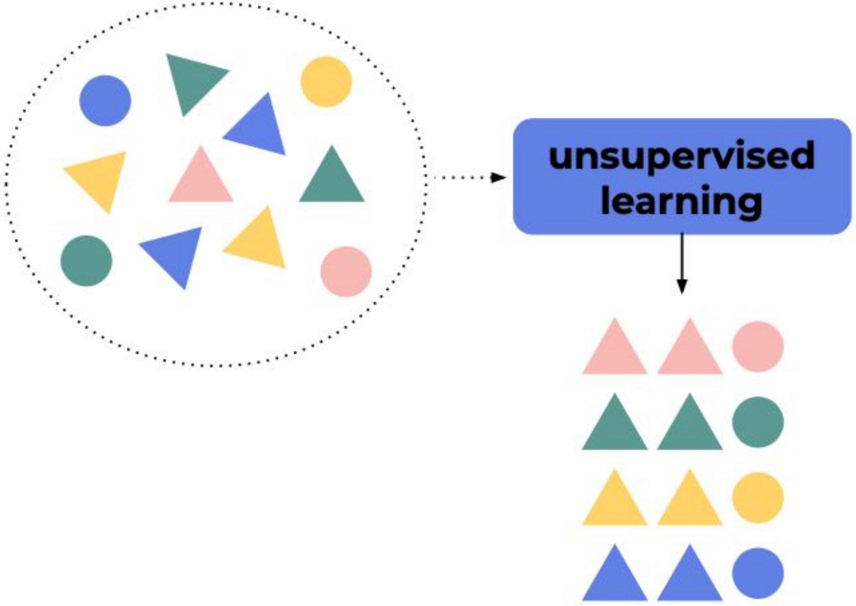
An Introduction to Statistical Learning: with Applications in Python (Springer Texts in…
› Gareth James
59
Paperback
€80⁴⁶
€5.99 delivery

Computer Age Statistical Inference: Algorithms, Evidence, and Data Science (Institute of…
Trevor Hastie Bradley Efron
80
Hardcover
€57⁵⁸
€5.99 delivery

Statistical Learning with Sparsity: The Lasso and Generalizations (Monographs on Statist…
› Trevor Hastie
25
Paperback
€51³⁸
€5.99 delivery

Generalized Additive Models (Monographs on Statistics & Applied Probability, Band 43)
D.R. Cox
8
Hardcover
€176⁵⁹
€5.99 delivery

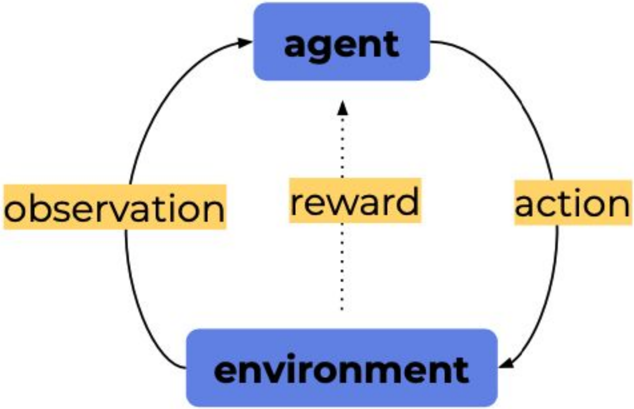
Unsupervised Learning



Reinforcement learning

Reinforcement learning is a type of machine learning where an agent interacts with an environment by taking actions, observing the resulting states, and receiving rewards.

Through this feedback loop, the agent learns to choose actions that **maximize cumulative rewards over time**.



AlphaGo

https://youtu.be/WXuK6gekU1Y

1.6 App Examples Exercises

Describe a few real applications, where classification can be useful.

Describe y variable (label, target) and the X variables (features, predictors, attribute),

Same for regression.

Classification Binary: $y \in \{0,1\}$

- **Email Spam Detection**: Predict whether an email is spam or not.
    - **Features** : Word Frequency (numerical continuous), Email Length (numerical continuous), Attachment Presence (categorical binary).
    - **Target**: Spam or Not Spam.
- **Credit Card Fraud Detection**: Detect whether a credit card transaction is fraudulent\
    - **Features**: Transaction Amount (numerical continuous), Time (numerical continuous or discrete), Merchant Category (categorical nominal), Device Type (categorical nominal).
    - **Target**: Fraud or Not Fraud.

Classification Multiclass: $y \in \{0, 1, 2, 3, \dots\}$

- **Image Recognition**: Classify images into categories like cat, dog, or bird.
    - **Features**: Pixel Intensity Values (numerical continuous).
    - **Target**: Cat or dog or bird
- **Predicting Weather**: Classify weather into categories: sunny, cloudy, rainy, or snowy.
    - **Features**: Temperature (numerical continuous), Humidity (numerical continuous), Wind Speed (numerical continuous), Pressure (numerical continuous).
    - **Target**: Sunny or cloudy or rainy or snowy.

Classification Multilabel $y \in \{0,1\}$ for each class

- **Multi-Genre Movie Classification**: Assign multiple genres (e.g., Action, Comedy, Drama, Family) to a movie.
    - **Features**: Movie Description (textual unstructured), Director (categorical nominal), Cast (categorical nominal), Keywords (categorical nominal).
    - **Target**: 0 or more: Action, Comedy, Drama, Family
- **Tagging News Articles**: Assign multiple tags (e.g., Politics, Sports, Technology) to a news article.
    - **Features**: Article Text (textual unstructured), Title (textual unstructured), Keywords (categorical nominal).
    - **Target**: 0 or more: Politics, Sports, Technology

Regression: $y \in R$, continuous value

- **Predicting Stock Prices**: Forecast the closing price of a stock.
    - **Features**: Opening Price (numerical continuous), Daily High/Low (numerical continuous), Trading Volume (numerical continuous), Market Sentiment (numerical continuous or categorical ordinal).
    - **Target**: Closing stock price as continuous value
- **Predicting Employee Salaries**: Estimate an employee's salary in dollars.
    - **Features**: Years of Experience (numerical continuous), Education Level (categorical ordinal), Job Role (categorical nominal), Location (categorical nominal)
    - **Target**: Salary as continuous value

1.7 Simple Linear Regression

Apple pie recipe

What's inside:

- 1 kg apples
- 1 secret ingredient
- random pie ingredients

Mix ingredients. Put in the oven.
Take a break. Enjoy.

Predict calories in a slice of pie

What’s inside:

- 1 kg apples
- 1 secret ingredient
- random pie ingredients

Predict calories in a slice of pie

227 calories

What’s inside:

- 1 kg apples
- 100 g of lard (untura)
- random pie ingredients

Predict calories in a slice of pie

203 calories

What’s inside:

- 1 kg apples
- 80 g of butter
- random pie ingredients

Predict the calories

Calories for 15 apple pie slices:

227, 203, 207, 148, 230,

154, 176, 290, 167, 245,

176, 259, 251, 503, 191

Predict the calories

Calories for 15 apple pie slices:

148, 154, 167, 176, 176,
191, 203, **207**, 227, 230,
245, 251, 259, 290, 503

Mean: 228,47

Median: 207

Predict calories in a slice of pie

130 calories

What’s inside:

- 1 kg apples
- 20 g of dates (curmale)
- random pie ingredients

Predict calories in a slice of pie

131 calories

What’s inside:

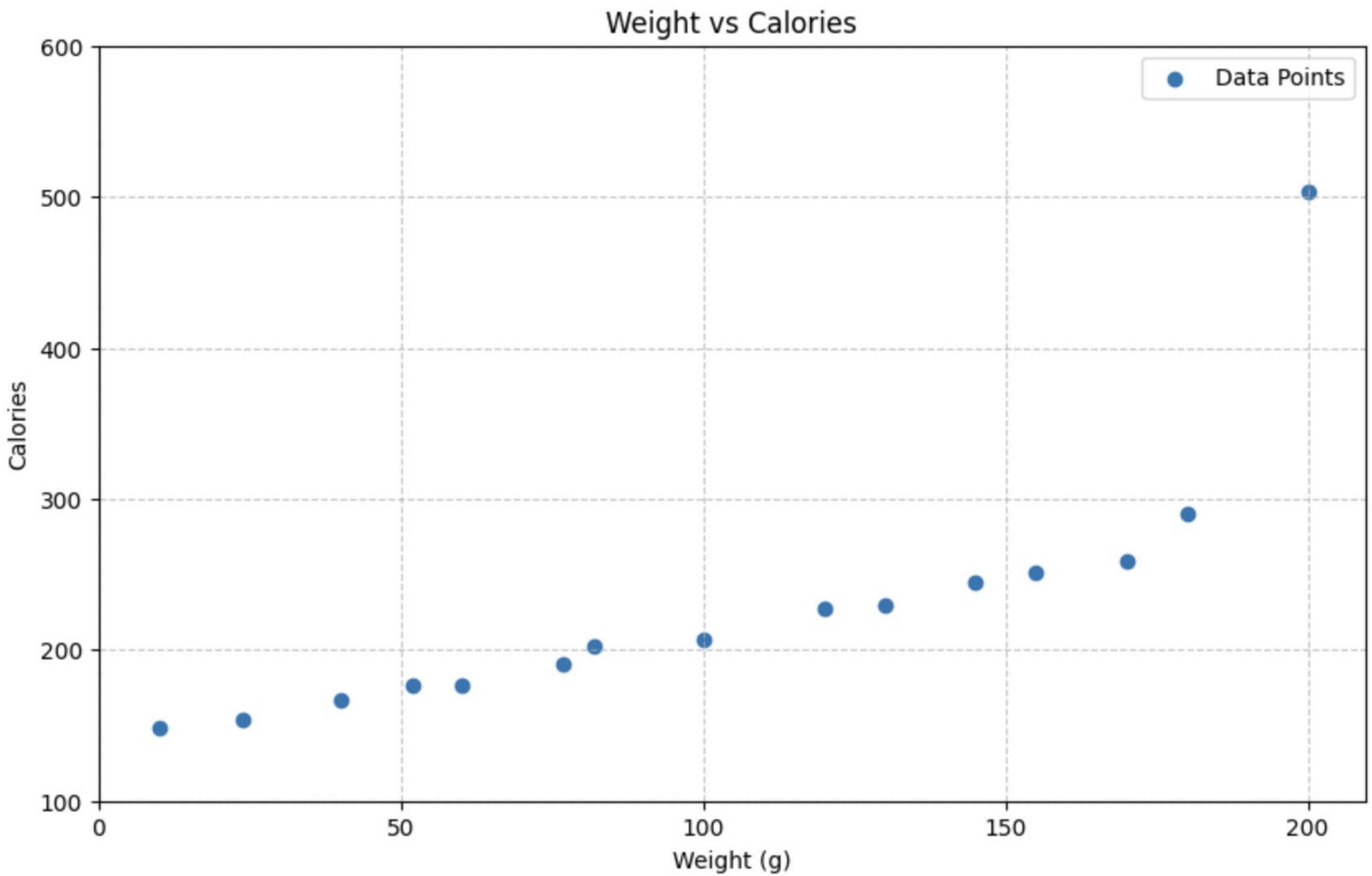
- 1 kg apples
- 20 g of dates (curmale)
- random pie ingredients

How did we do?

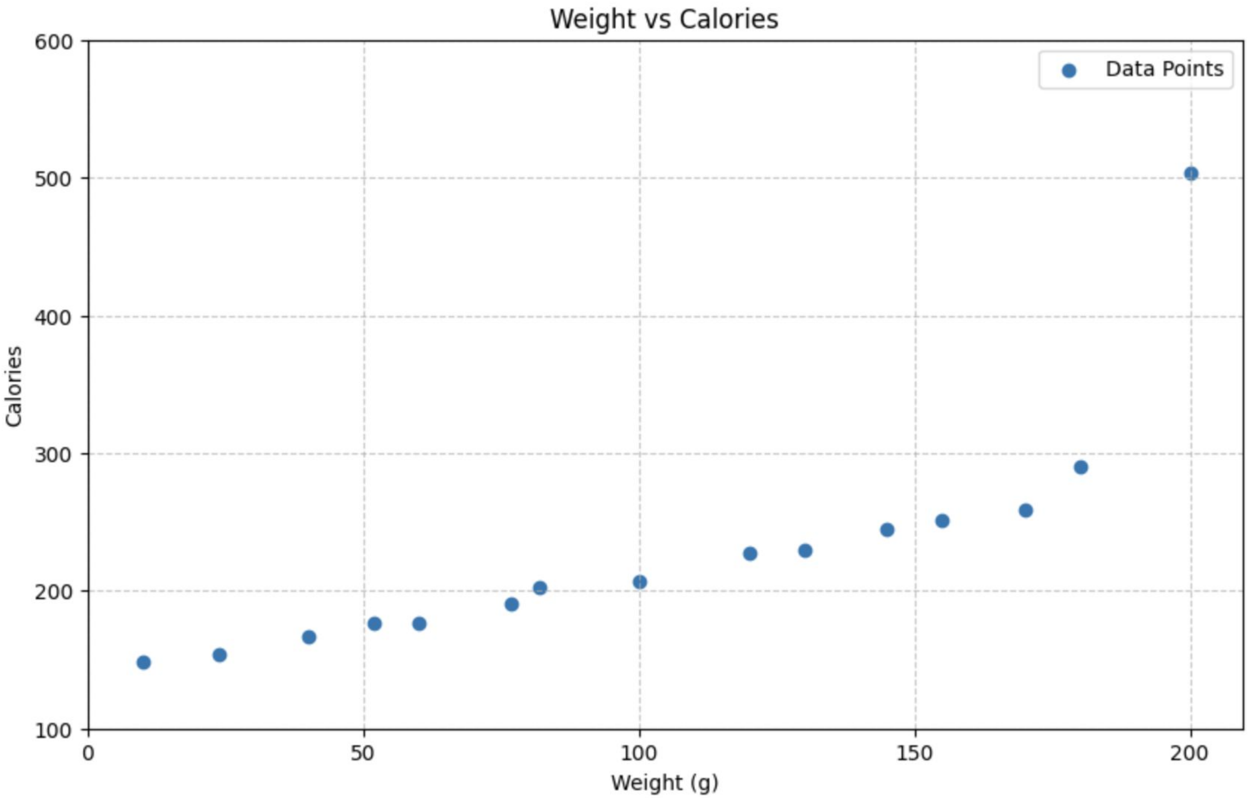
Mean = 228

Did we overestimate? Error = 228-131 = 97

Predict calories



Predict calories

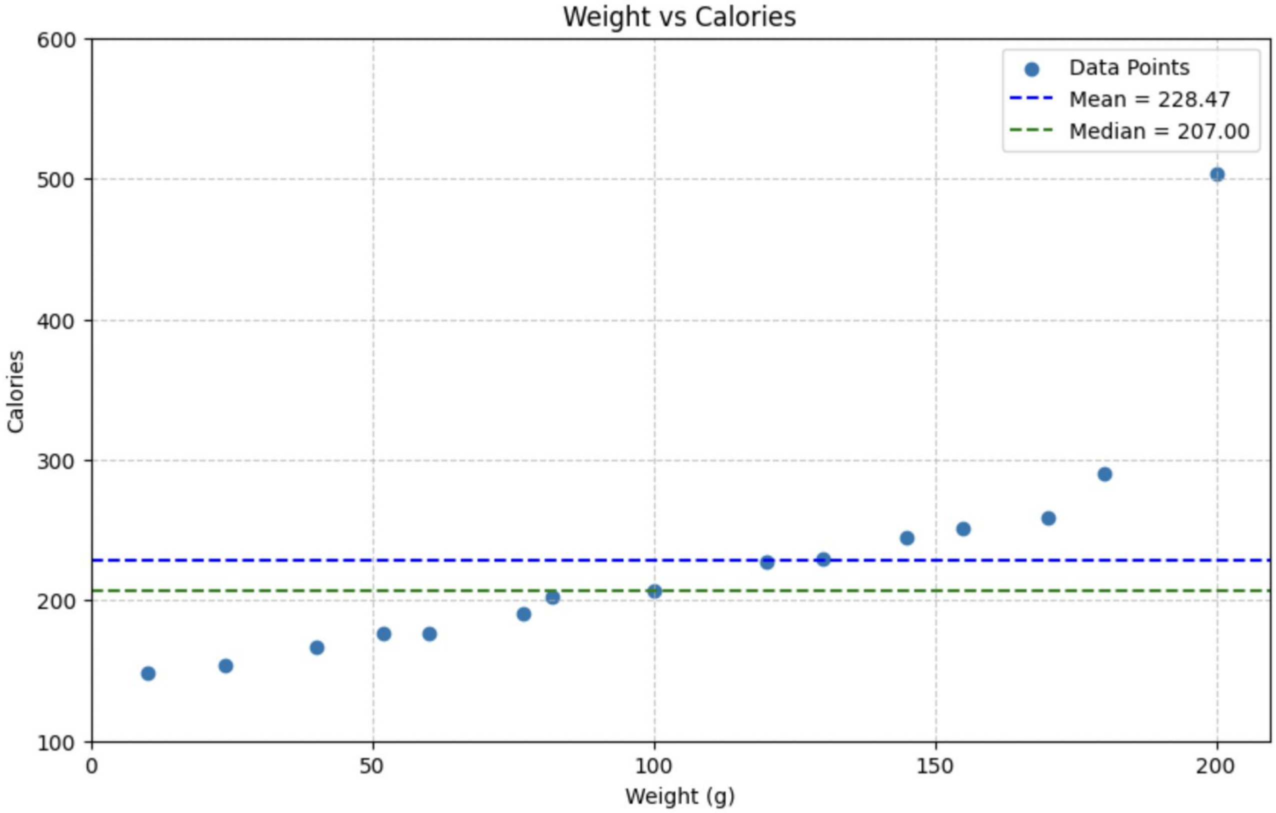


Eq of the line:
y = ax + b

Where:
a: slope
b: intercept
x: data

How lines work:
Calories =
Slope * Weight +
Intercept

Predict calories



Eq of the line:
y = ax + b

Where:
a: slope
b: intercept
x: data

How lines work:
Calories =
0 * Weight + 228

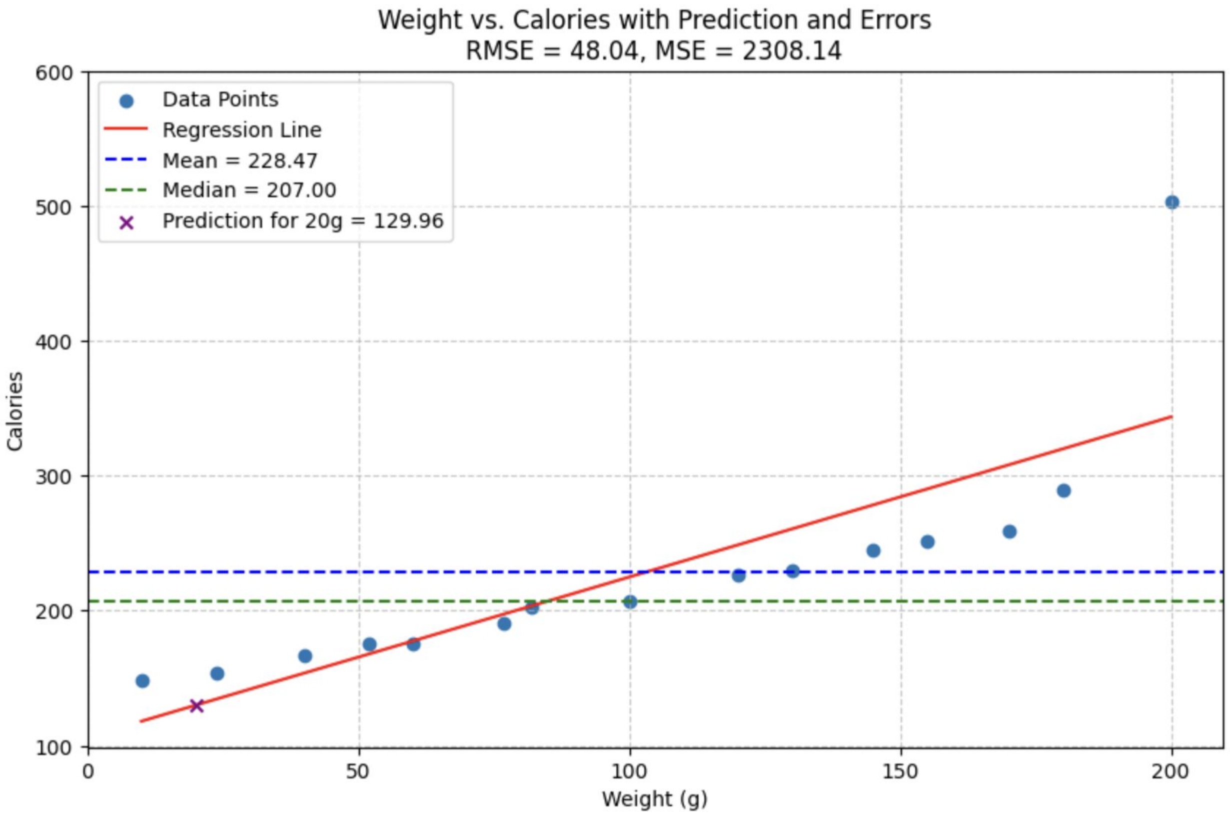
Let’s find a line that’s as close to the points as possible!

Define error as point - line.

If we want a line close to the points, we don’t like errors.

An easy function of the errors to minimize is root mean squared error (RMSE).

Predict calories



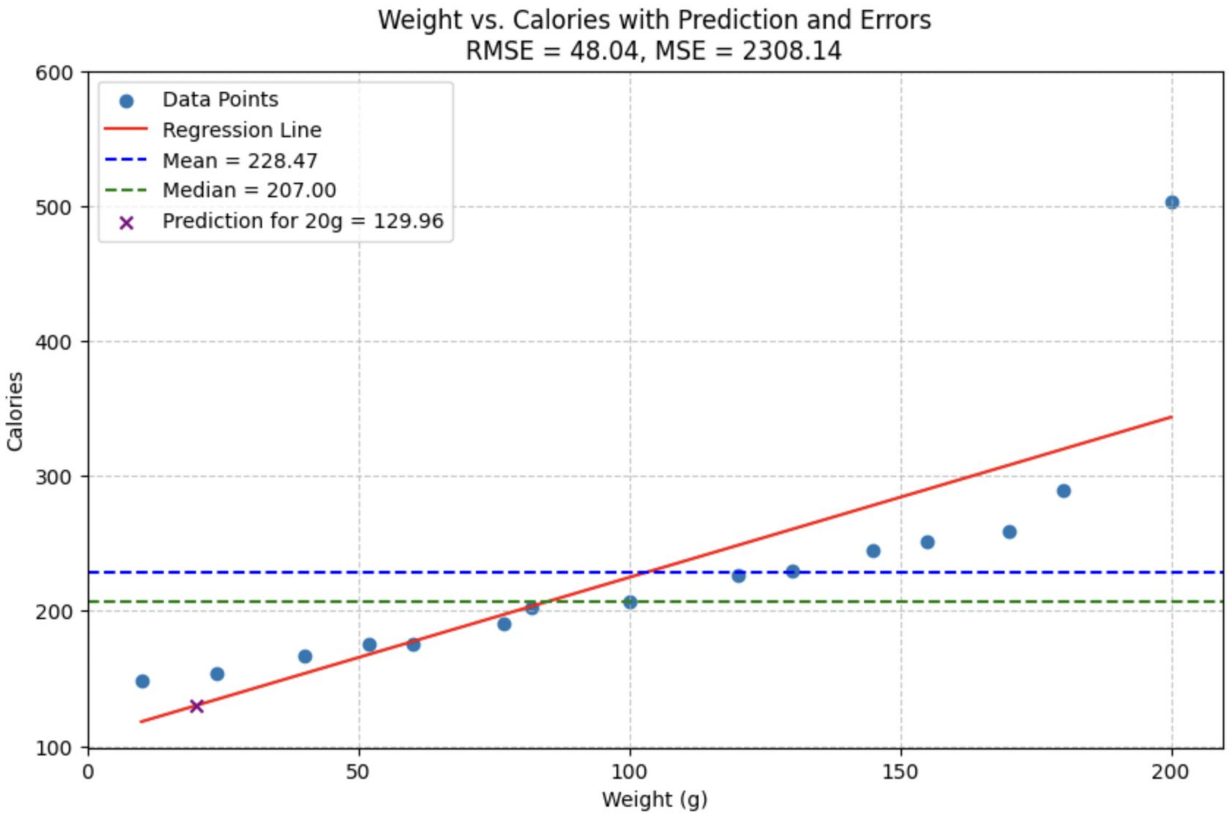
Eq of the line:
y = ax + b

Where:
a: slope
b: intercept
x: data

How lines work:
Calories =
1.187 * Weight + 106.2

Code on Colab

Predict calories



Eq of the line:
y = ax + b

Where:
a: slope
b: intercept
x: data

How lines work:
Calories =
1.187 * Weight + 106.2 = 129.94

Code on Colab

Data is not always linear

No problem!

There are non-linear regression models:

- Decision Tree Regression
- SVM with non-linear kernels
- Neural networks
- KNN
- Ensemble methods

1.8 Questions & Further reading

#1 Google machine learning glossary: https://developers.google.com/machine-learning/glossary

#2 What is machine learning: https://www.youtube.com/watch?v=iLu9XyZ55oI

#3 Steps of Machine Learning: https://www.youtube.com/watch?v=nKW8Ndu7Mjw&t=394s

#4 Understanding the 3 most common loss functions for Machine Learning Regression: https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression-23e0ef3e14d3

Thank you !!

Machine Learning Engineer / Data Scientist
zahariesergiu@gmail.com
https://www.linkedin.com/in/zahariesergiu/
https://github.com/zahariesergiu/ubb-sociology-ml