Machine learning: prediction, classification and clustering

UBB Faculty of Sociology

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 Networks

#7 Help Final Project

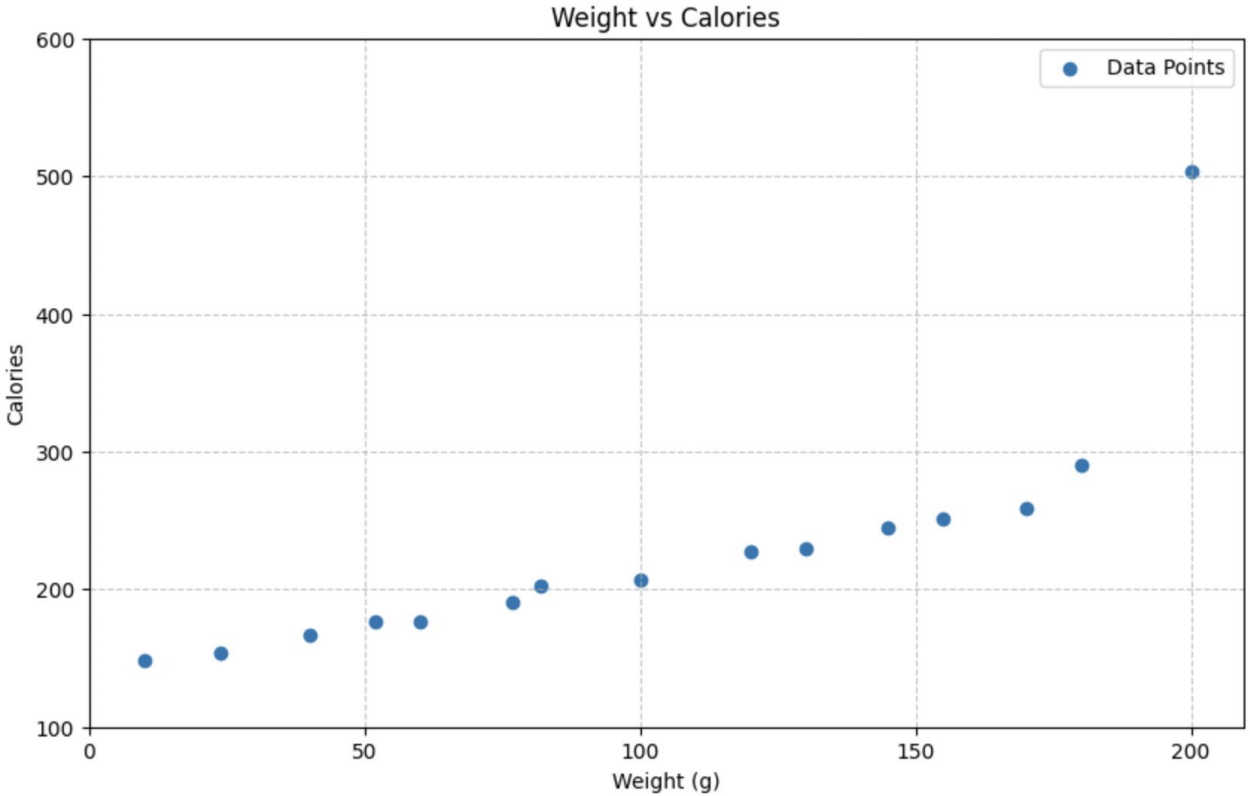
#8 Help Final Project

2. Linear Regression, Decision Trees

#1.1 Linear Regression recap
#1.2 Decision Trees Regression
#1.3 Regression implementation
#1.4 Final Project Task 2 - Census Data EDA
#1.5 Questions & Further reading

1.1 Linear Regression recap

Predict calories



Eq of the line:
$y = ax + b$

Where:
a: slope
b: intercept
x: data

How lines work:
Calories =
Slope * Weight +
Intercept

Predict the calories

Calories for 15 apple pie slices:

148, 154, 167, 176, 176,

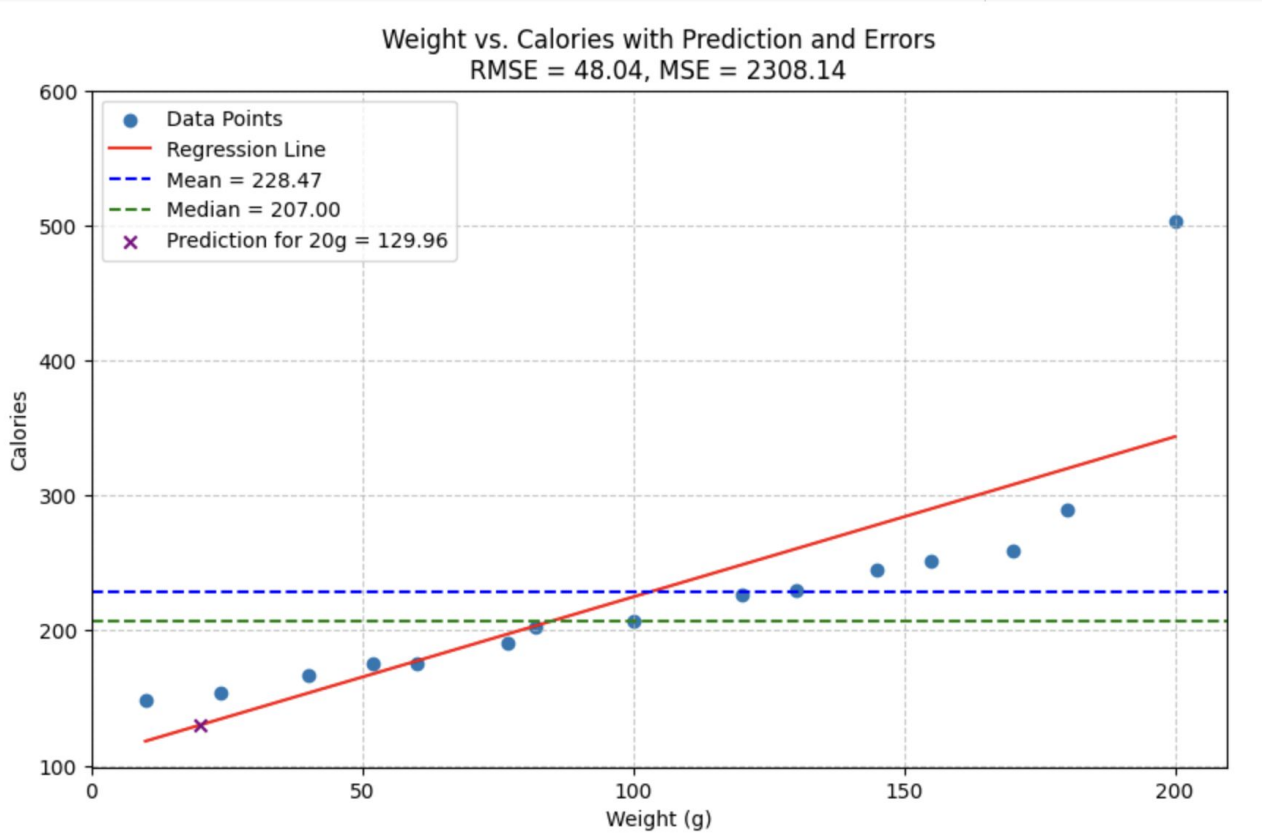
191, 203, **207**, 227, 230,

245, 251, 259, 290, 503

Mean: 228,47

Median: 207

Predict calories



Eq of the line:
y = ax + b

Where:
a: slope
b: intercept
x: data

How lines work:
Calories =
1.19*20 + 106.2=131

Code on Colab

{MSE formula}

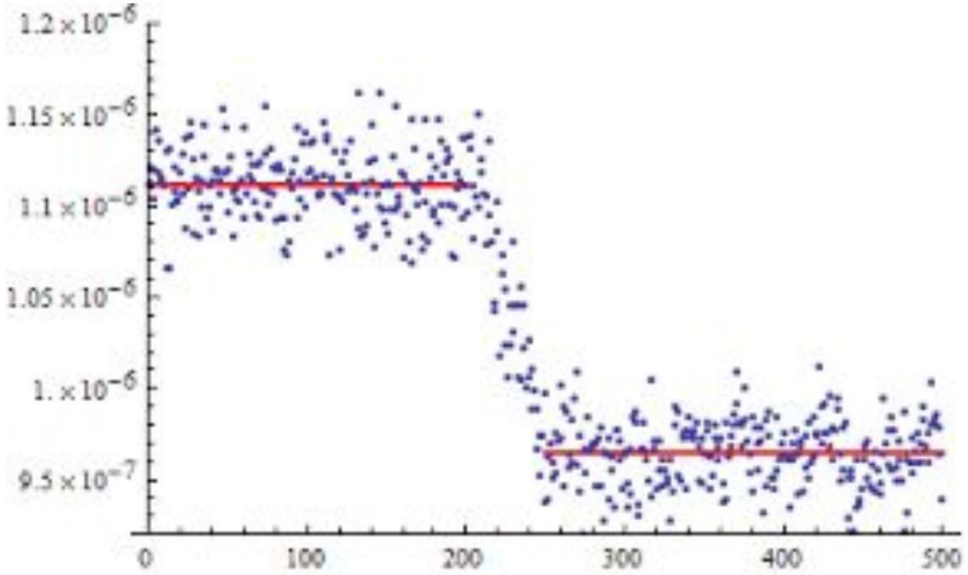
{LASSO MSE formula}

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



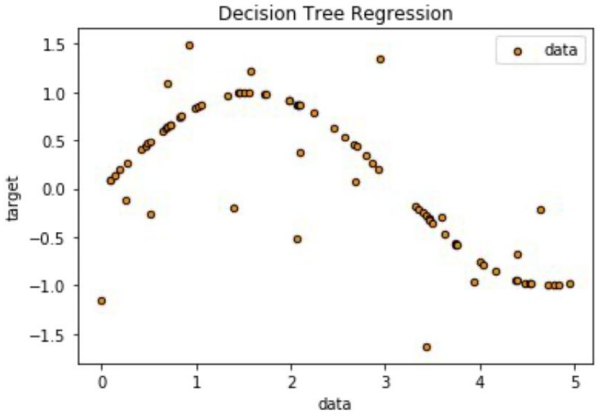
1.2 Decision Trees Regression

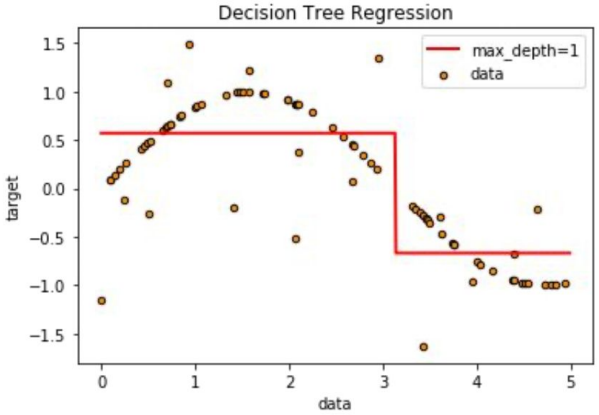
Decision Tree Regression

Decision Trees Regression is a very fast algorithm How it works: A two step repeating process

1. Take some random samples in the subtree split the dataset in two
2. Save the random split which separates the data by value the best

Then, repeat for the newly created subtrees

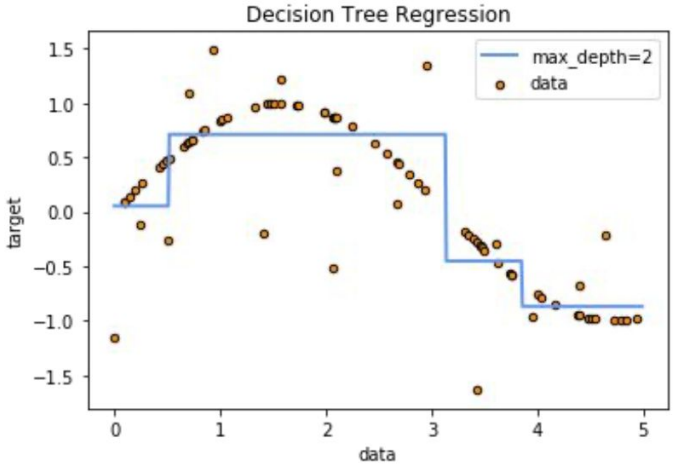


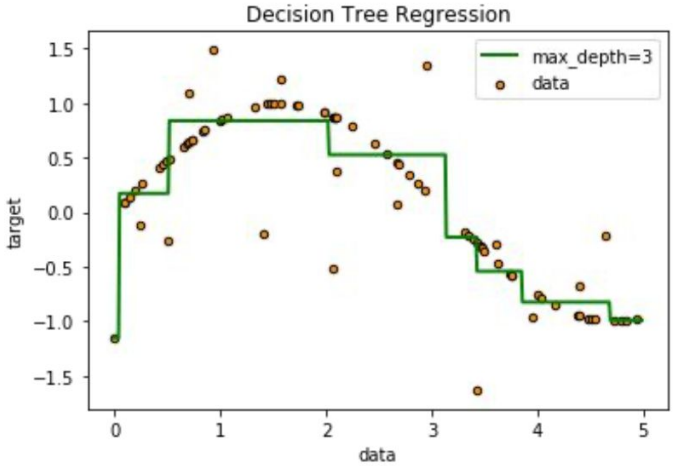


Decision Tree Regression

Decision Trees Regression is a very fast algorithm How it works: A two step repeating process

1. Step 1
    1. Take some random samples in the subtree split the dataset in two
    2. Save the random split which separates the data by value the best
2. Step 2
    1. Take some random samples in the subtree split the dataset in two
    2. Save the random split which separates the data by value the best





Decision Tree Regression

Decision Trees Regression is a very fast algorithm How it works: A two step repeating process

3. Step 3
    1. Take some random samples in the subtree split the dataset in two
    2. Save the random split which separates the data by value the best

    ……..

n. Step n

    n.1 Take some random samples in the subtree split the dataset in two

    n.2 Save the random split which separates the data by value the best

Decision Tree Regression

Advantages:

- We get a nice, explainable model, easy to understand by humans
- Very fast prediction for new samples

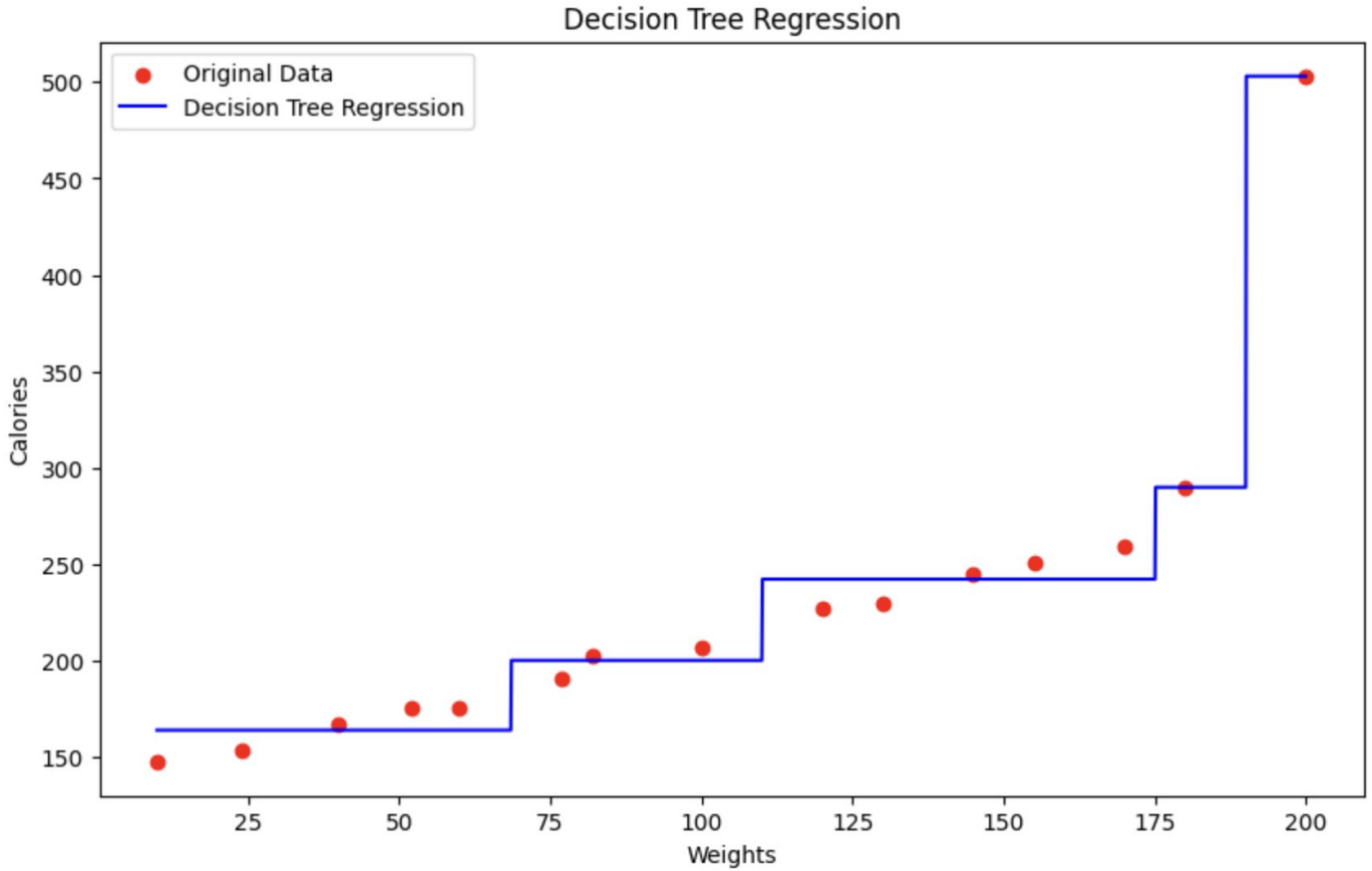
Decision Tree Regression

Disadvantages:

- Leaves grow exponentially by depth
- Very prone to overfitting

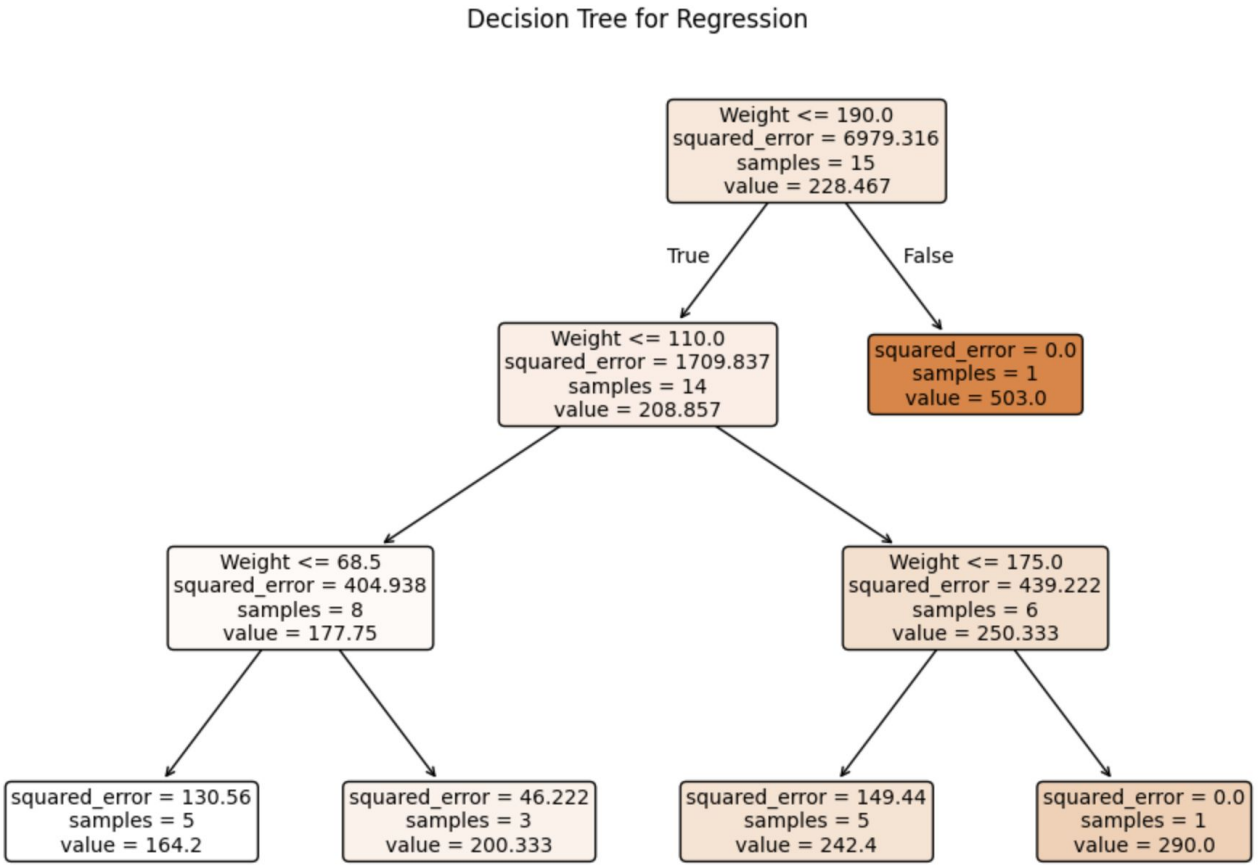
Decision Tree Regression Calories

[Code on Colab]



Decision Tree Regression Calories

[Code on Colab]



1.3 Regression implementation

- Penguins size
- California Housing

1.4 Final Project Task 2 - Census Data EDA

https://github.com/zahariesergiu/ubb-sociology-ml/blob/main/final_project/Final_Project_Task_2_Census_Data_EDA.ipynb

1.5 Further Reading & Questions

#1 Regression Trees explained: https://www.youtube.com/watch?v=g9c66TUylZ4

#2 Decision Tees Split Criteria: https://scientistcafe.com/ids/splitting-criteria

#3 Self-supervised learning: https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

#4 Feature scaling @ Kaggle: https://www.kaggle.com/code/mysarahmadbhat/all-about-feature-scaling/notebook

#5 Nature article: statistics and machine learning:
https://www.nature.com/articles/nmeth.4642?source=post_page-----64b49f07ea3----------

#6 If correlation doesn't imply causation, then what does?:
https://michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/

#7 Requirements for talking about cause and effect:
https://towardsdatascience.com/are-you-guilty-of-using-the-word-experiment-incorrectly-9068baeab7a4

Thank you !!

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