Predictive Modelling with Python

Jure Žabkar

April, 2023

Contents

Software installation

Introduction to scikit-learn

Artificial data sets, illustration of basic regression and classification techniques

Regression & Classification

Individual work: data preparation, Visualization, Modelling, Feature selection, Evaluation

Installation

The most elegant way to install the required software is by installing Conda.
You can either install:

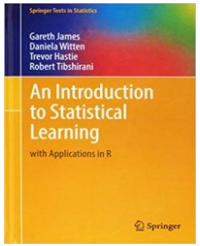
the entire set of packages in Anaconda, **or**

install Miniconda first, and manually add packages `scikit-learn`, `pandas` and `matplotlib`:

```
conda install -c intel scikit-learn
conda install pandas matplotlib
```

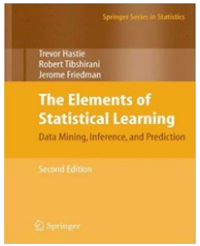
Github sources: jurezabkar/fri-ds-python-ml

Literatura



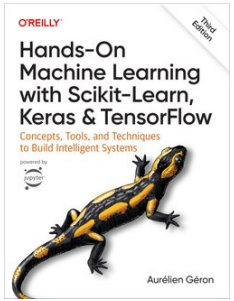
James G, Witten D, Hastie T, Tibshirani R (2013)

An introduction to statistical learning, Springer.



Hastie T, Tibshirani R, Friedman J (2017)

The Elements of Statistical Learning, 2nd Ed., Springer.



Geron, A (2022)

Hands-on machine learning with Scikit-Learn, Keras and TensorFlow, O'Reilly.

What will you learn?

- How to import the data
- Data preprocessing & visualization
- Computing basic data set statistics
- Basic regression and classification with sklearn
- How to tune the parameters of ML algorithms
- Proper evaluation

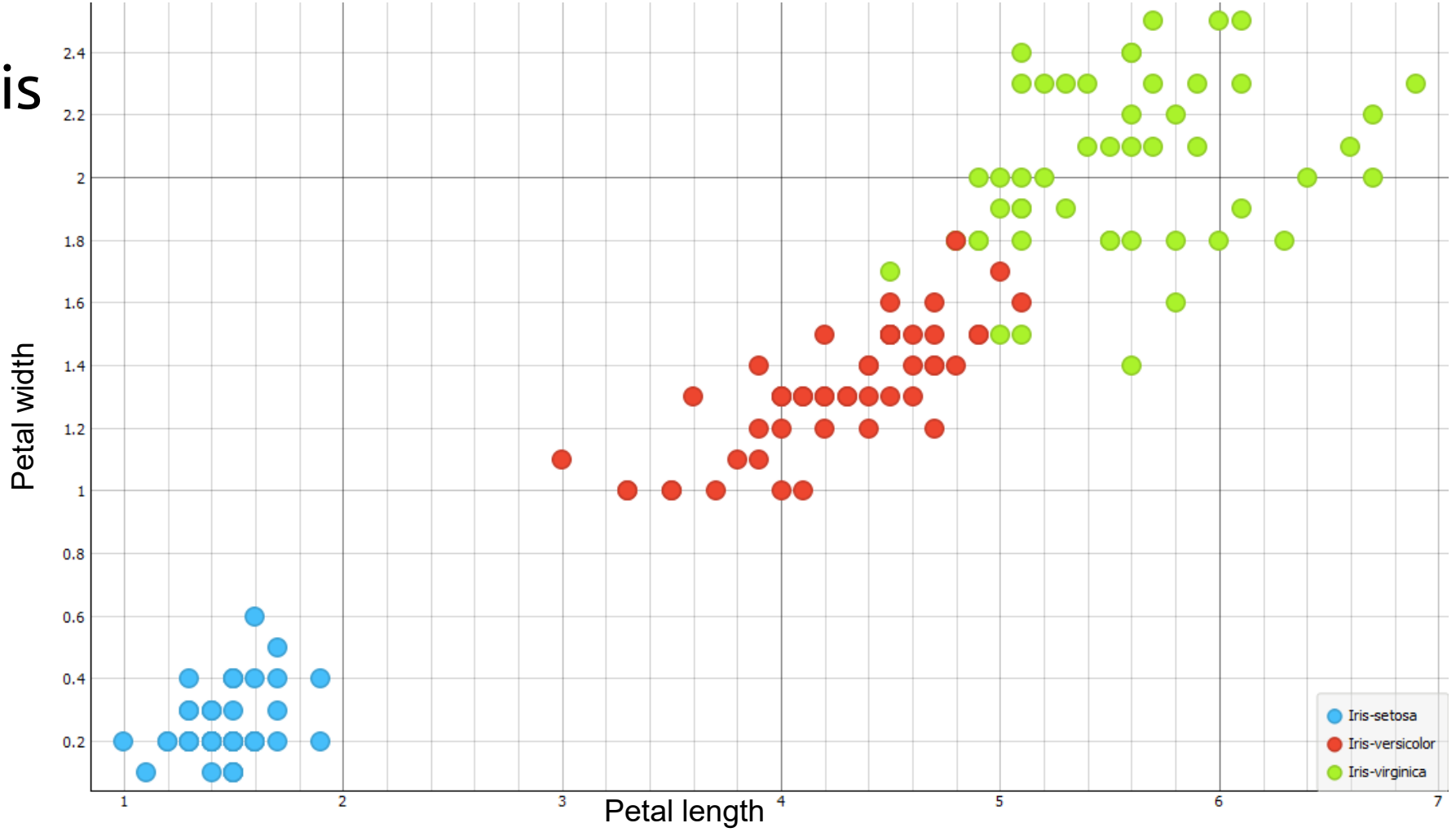
Classification: Iris dataset

3 types of Iris:

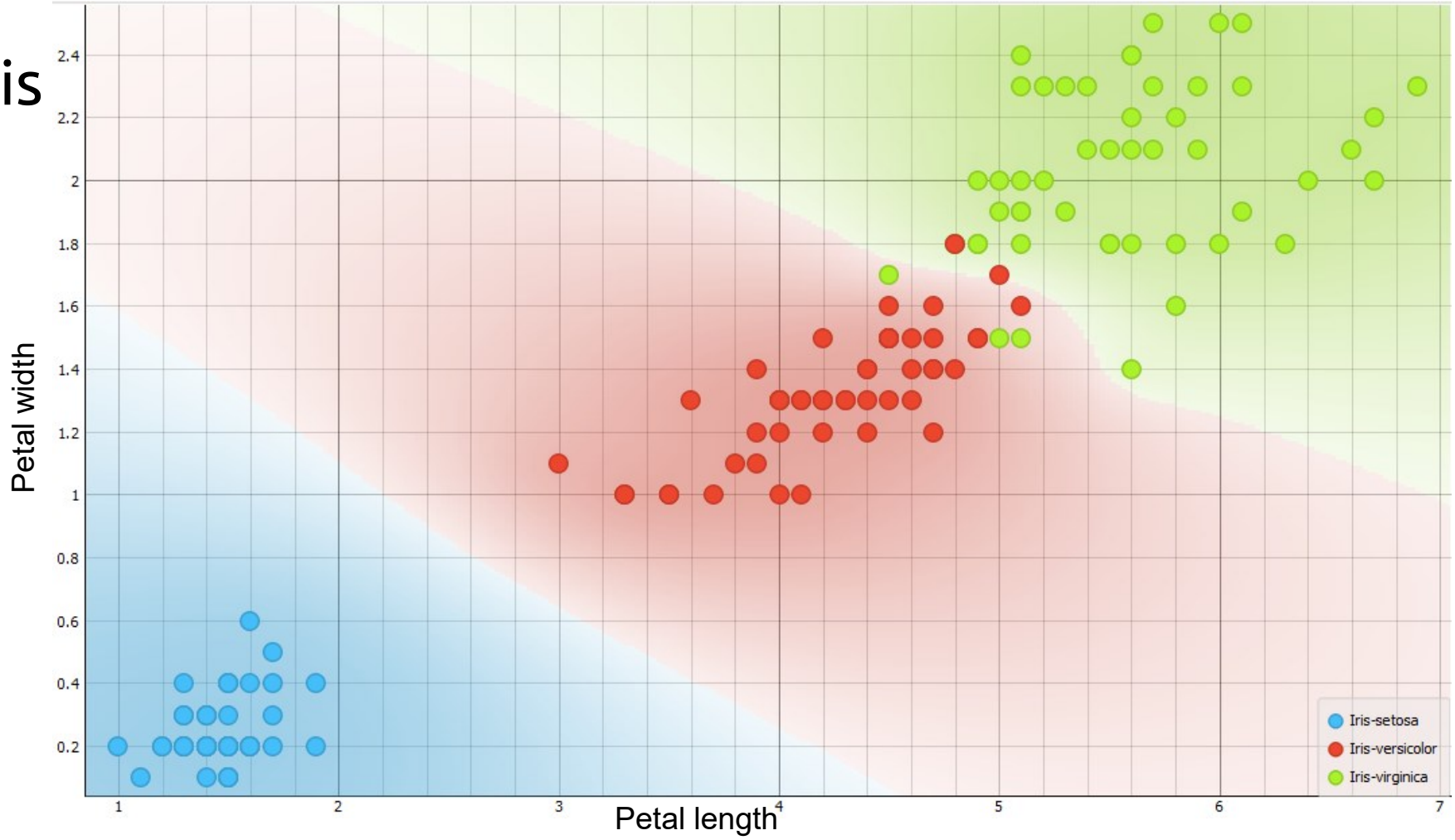
- Setosa
- Virginica
- Versicolor

iris	sepal length	sepal width	petal length	petal width
Iris-setosa	5.1	3.5	1.4	0.2
Iris-setosa	4.9	3.0	1.4	0.2
Iris-setosa	4.7	3.2	1.3	0.2
Iris-setosa	4.6	3.1	1.5	0.2
Iris-setosa	5.0	3.6	1.4	0.2
Iris-setosa	5.4	3.9	1.7	0.4
Iris-setosa	4.6	3.4	1.4	0.3

Iris

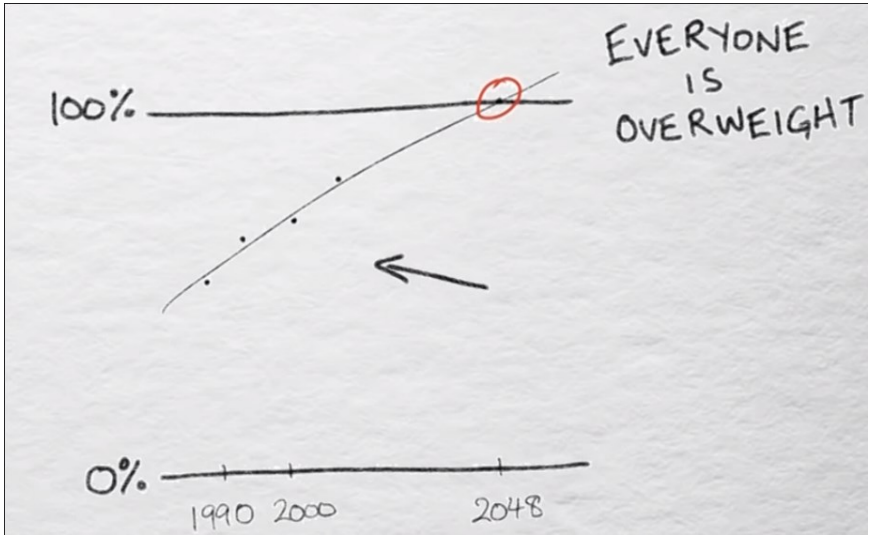

Petal width
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
Petal length
Iris-setosa
Iris-versicolor
Iris-virginica

Iris


Petal width
Petal length
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
Iris-setosa
Iris-versicolor
Iris-virginica

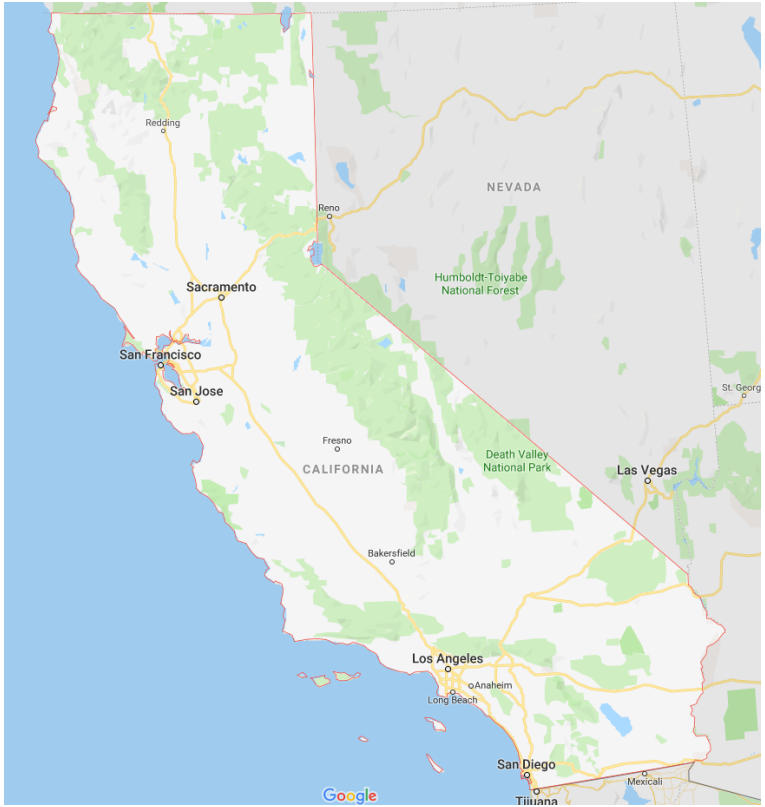
Regression: Obesity apocalypse

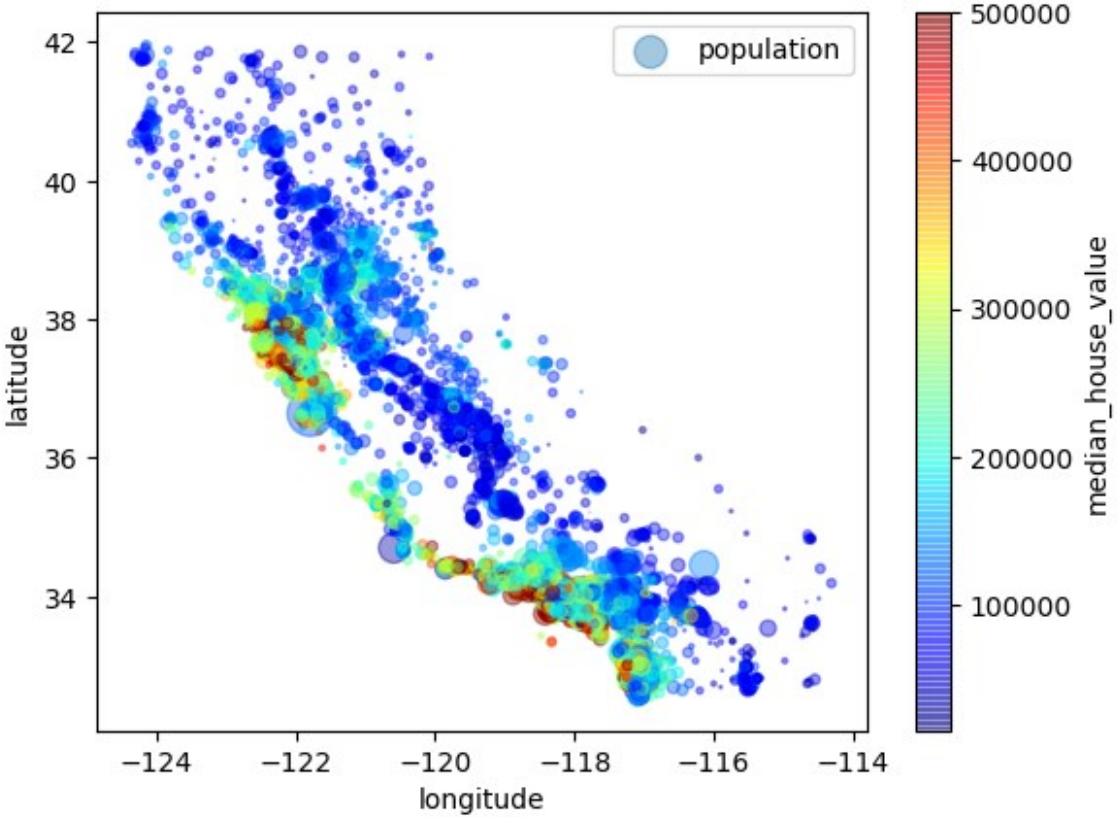
abcNEWS: "By 2048, all American adults would become overweight or obese."



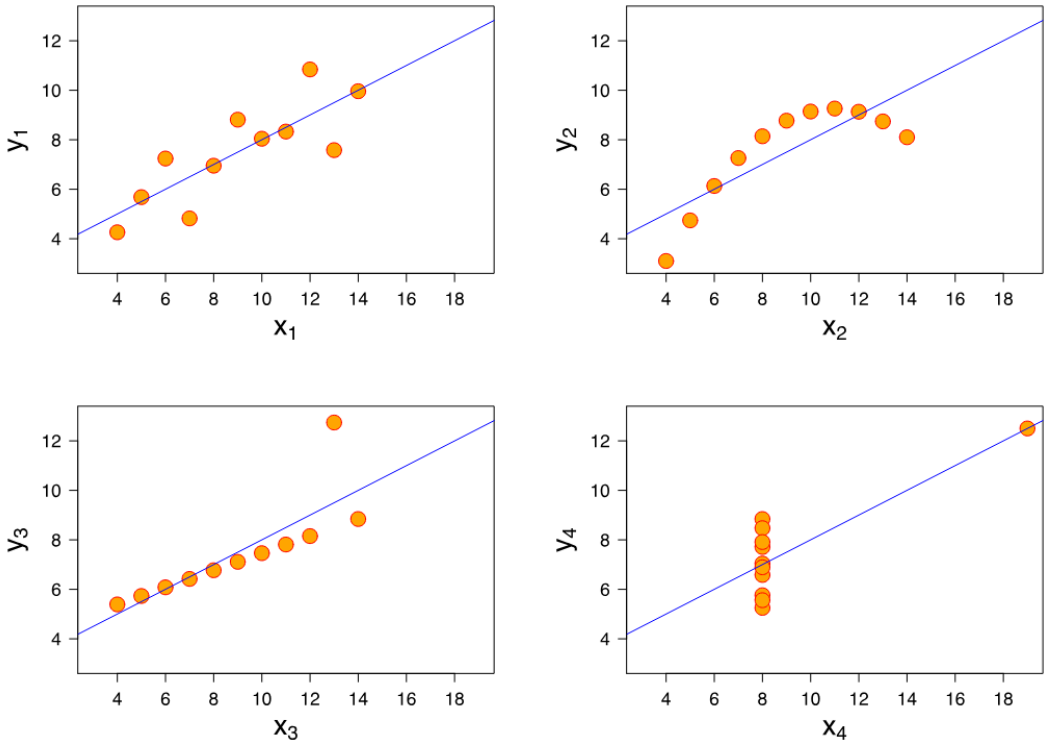
https://abcnews.go.com/Health/Fitness/story?id=5499878&page=1

California housing





Visualize your data



4 data sets

Nearly **identical** statistics

Very **different** plots

source:
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. **27** (1): 17–21.

California housing: **tasks**

- Import & visualize the data (datasets/housing.csv)
- Split the data set to a training set and a test set (stratified, 70:30)
- Compute/visualize correlations ("median_house_value", "median_income", "total_rooms", "housing_median_age")
- Prepare the training set for ML algorithms:
  - Add new features
  - Impute features with missing values
  - Scale the data
- Learning:
  - Choose appropriate algorithms
  - Use internal cross-validation to tune the parameters
  - Evaluate on training set
- Evaluate on test set