Machine Learning

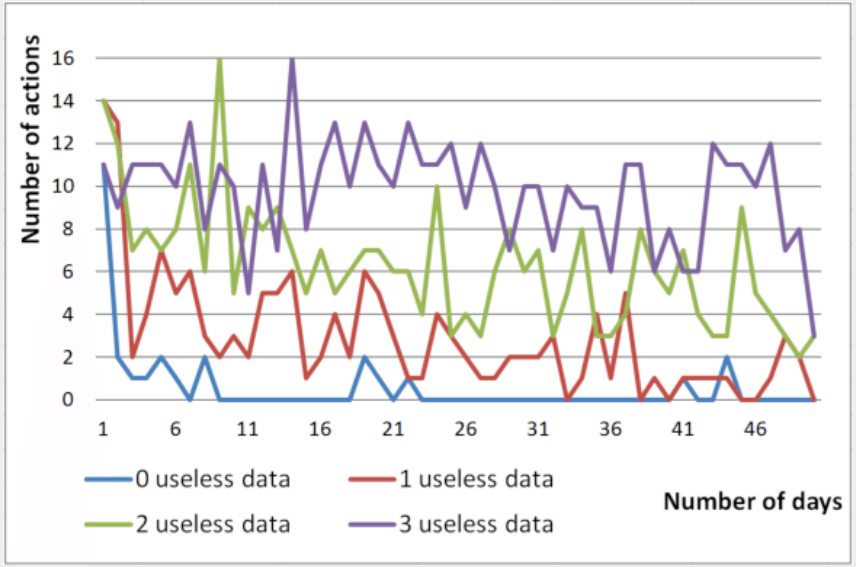
Data science pentru machine learning

Performanța depinde de date

Dacă datele sunt gunoi, atunci și rezultatele vor fi gunoi. De acesta în machine learning datele au așa de multă importanță.

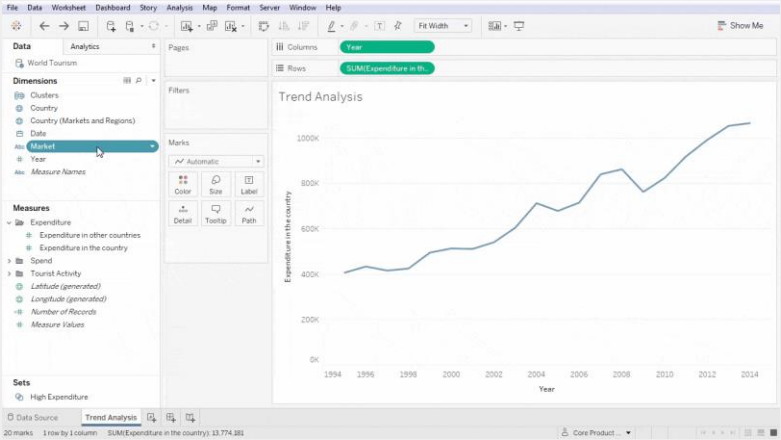
Nu toate datele sunt utile

Datele de intrare trebuie să aibă o relație cu datele de ieșire.



Metode de vizualizare a datelor

Sunt o mulțime de programe și unelte pentru aceasta
Ex: Matlab, Matplotlib, Seaborn, și alte unelte online și offline.

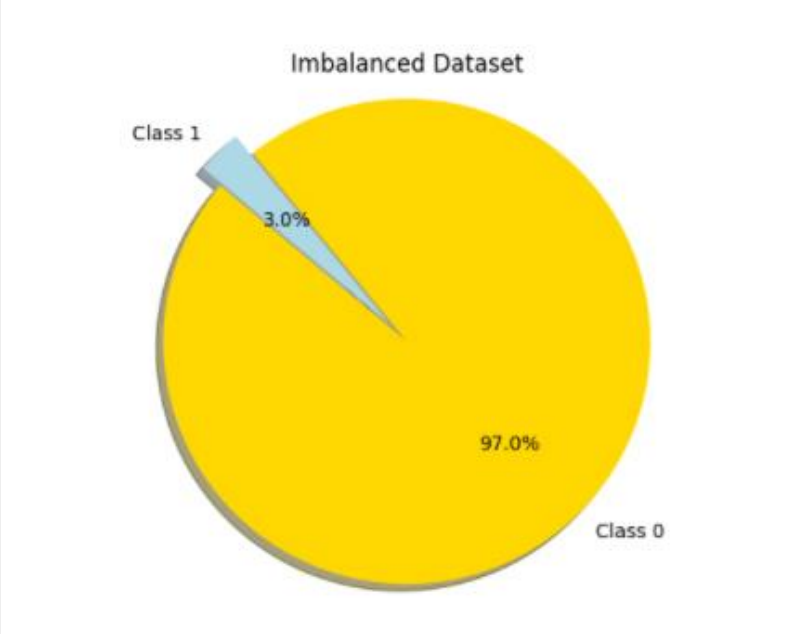


Identificarea datelor bune

1) Identificați scopurile pe care le-ar putea îndeplini datele.
2) Identificați dacă datele pot avea o corelație cu rezultatul așteptat.
3) Identificați utilitatea datelor.
4) Identificați calitatea datelor.

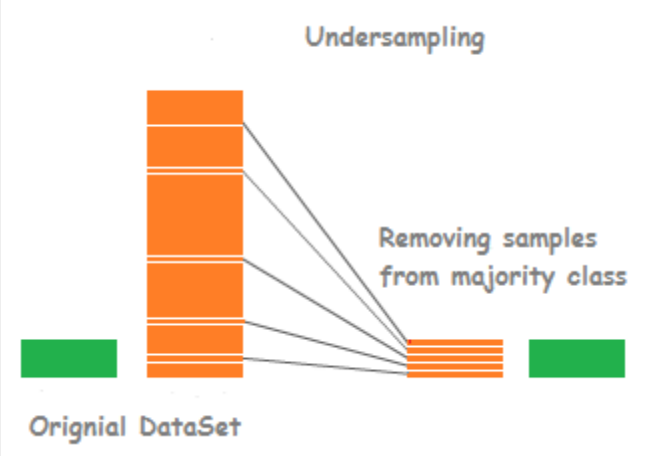
Balansarea datelor

Învățarea trebuie să fie făcută pe date balansate, implicând că numărul de rezultate negative și rezultate pozitive trebuie să fie comparabil.



Metode de balansare a datelor

1. Undersampling. Utilizarea doar a subsetului balansat de date



Metode de balansare a datelor

2. Oversampling la datele ce sunt in minoritate. Prin oversampling, se înțelege mărirea artificială a unui set de date.

Codificarea datelor

Modelele de machine learning sunt niște funcții matematice, cum să transformăm datele in numere.

Imagine → [0.1,1,-0.3,...]
Text → [1,0,0,0,...]
🤔 → [-0.1,0.8,0.3,...]

Codificarea datelor de tip text

1. Codificare One Hot
   Maparea unui vector cu indicele obiectului unui dicționar.

Human-Readable

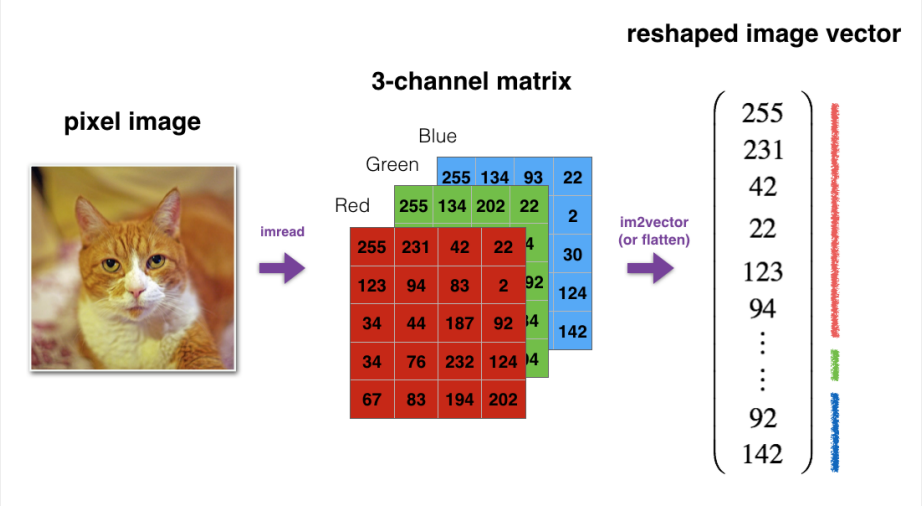
Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

Codificarea datelor de tip imagine

Imaginile deja sunt numere, necesita doar schimbarea dimensiunii.



Referinte

https://www.kaggle.com/learn/data-cleaning

https://www.datarobot.com/wiki/data-preparation/

https://www.tableau.com/learn/articles/data-visualization

http://tylervigen.com/spurious-correlations

https://towardsdatascience.com/stop-one-hot-encoding-your-categorical-variables-bbb0fba89809