Machine Learning

Pas cu pas

Pas 1: Problema

Ce problema dorești să rezolvi?

Problema este că nu știi machine learning

Soluția este să înveți machine learning.

Regula #1 a Machine Learning de la Google

Nu utiliza machine learning dacă nu trebuie.

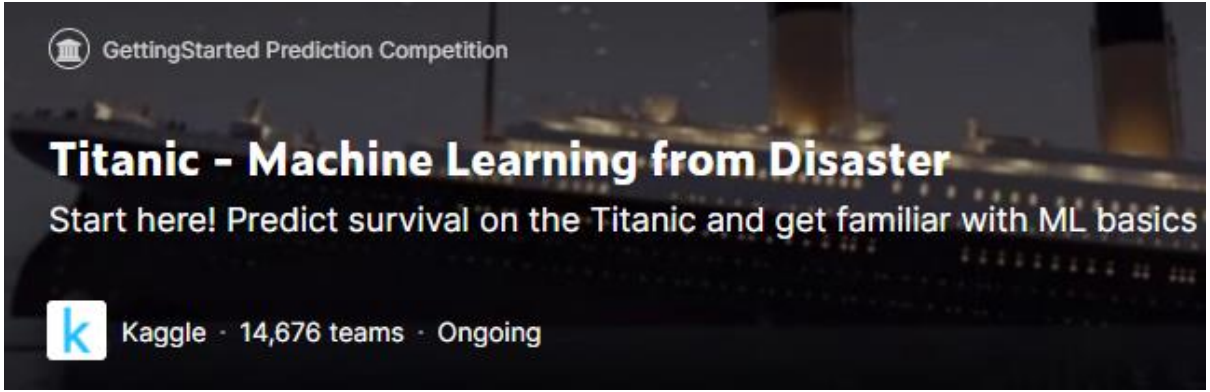
Pas 2: Caută date necesare

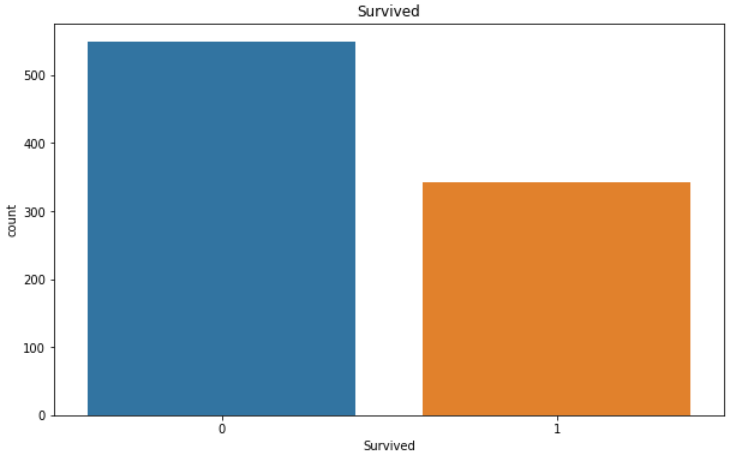
Pentru problema noastră orice tip de date se potrivește.

Setul de date Titanic

Cum am spus în prezentarea trecută, data setul titanic este un bun set de date pentru incepatori, și acesta este ca un ”hello world” pentru data science.

https://www.kaggle.com/c/titanic





Pas 3: Înțelege datele

Verifică ce conținut au și ce oferă datele

Exemplu de date

Acestea sunt primele 5 intrari din datele care ne le ofera data setul titanic.

Observăm că sunt niște date tabulare, formate din 12 coloane.

Unele coloane sunt înțelese:
Ex: Name, Survived, Age

Altele nu prea:
Ex: Pclass, SibSp, Embarked

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Există explicație

Pe kaggle, majoritatea data seturilor vin cu explicație a datelor.
Originea datelor, Contextul datelor, Variabilele datelor și alte lucruri.

În aplicarea reală a machine learning, nimeni nu garantează explicație datelor care le găsești.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Note aditionale

Pclass reprezinta un proxy pentru statusul socio-economic

Age poate să fie fractional

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

Pas 4: Curățarea datelor

Determinarea datelor utile, eliminarea datelor inutile.

Din datele curente nu toate coloanele sunt de folos, probabilitatea ca numele unei persoane să determine dacă acesta a supraviețuit sau nu este foarte mică.

Din motive similare vom ignora coloanele PassengerId, Name, Ticket și Cabin

Cabin va fi ignorat din cauză că datele nu sunt calitative.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Eliminarea valorilor lipsă

În imaginea din dreapta este numărul de date lipsa pentru fiecare coloană.

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

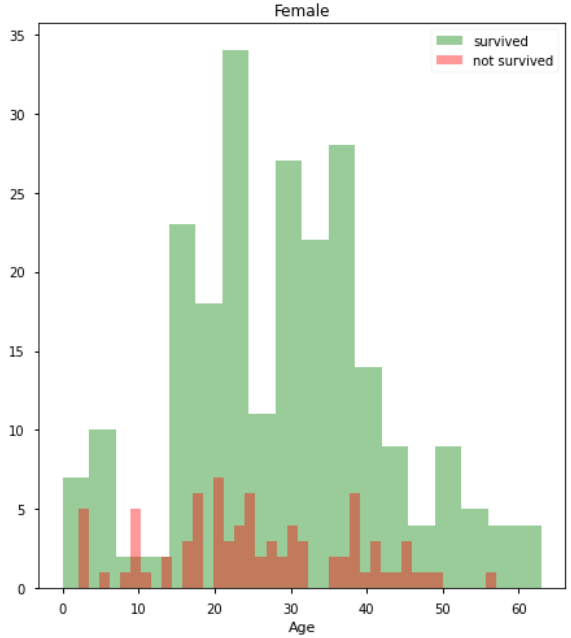
Eliminarea valorilor lipsa din coloana Age

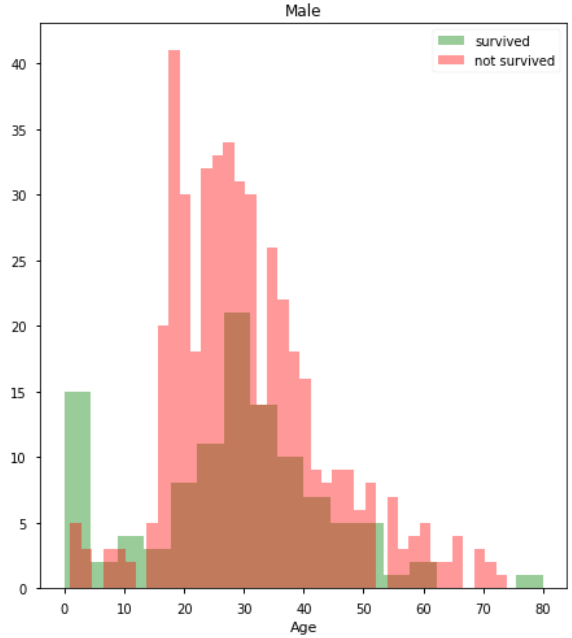
Se poate să facem valorile null să fie zero, dar aceasta va influența negativ acuratețea prezicerilor, deoarece sunt date reale în data set care au valoarea funcțională, și este aproape de 0.

Se poate să oferim valorii Age o valoare numerică imposibilă, ex: -1, Și putem în pasul următor să rezolvăm așa valori.

Putem să stergem rândurile care nu au valoare pentru Age.

Putem să extrapolăm valoarea age după valorile care există.





Pentru antrenare eliminăm NULL

Putem introduce valori artificiale, dar acestea pot împiedica antrenării.

Am decis să eliminăm randurile care nu au valoarea Age.

Pentru coloana Embarked, am utilizat backwards fill, care copie valoarea văzută precedent pentru un rând cu valoare NULL. Sunt prea puțini oameni care nu au valoarea Embarked, așa că nu contează ce punem aici.

```
import pandas as pd

raw_data = pd.read_csv('train.csv')
data = raw_data.dropna(subset=['Age'])
data['Embarked'].fillna(method='bfill', inplace=True)
raw_data.info()
data.info()
```

Așa arată setul de date

După ce am făcut modificările

```
Int64Index: 714 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  714 non-null    int64
 1   Survived     714 non-null    int64
 2   Pclass       714 non-null    int64
 3   Name         714 non-null    object
 4   Sex          714 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        714 non-null    int64
 7   Parch        714 non-null    int64
 8   Ticket       714 non-null    object
 9   Fare         714 non-null    float64
 10  Cabin        185 non-null    object
 11  Embarked     714 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 72.5+ KB
```

Referinte

https://developers.google.com/machine-learning/guides/rules-of-ml

https://www.kaggle.com/mgamal91/titanic-model-first-kaggle-challenge

https://www.kaggle.com/c/titanic