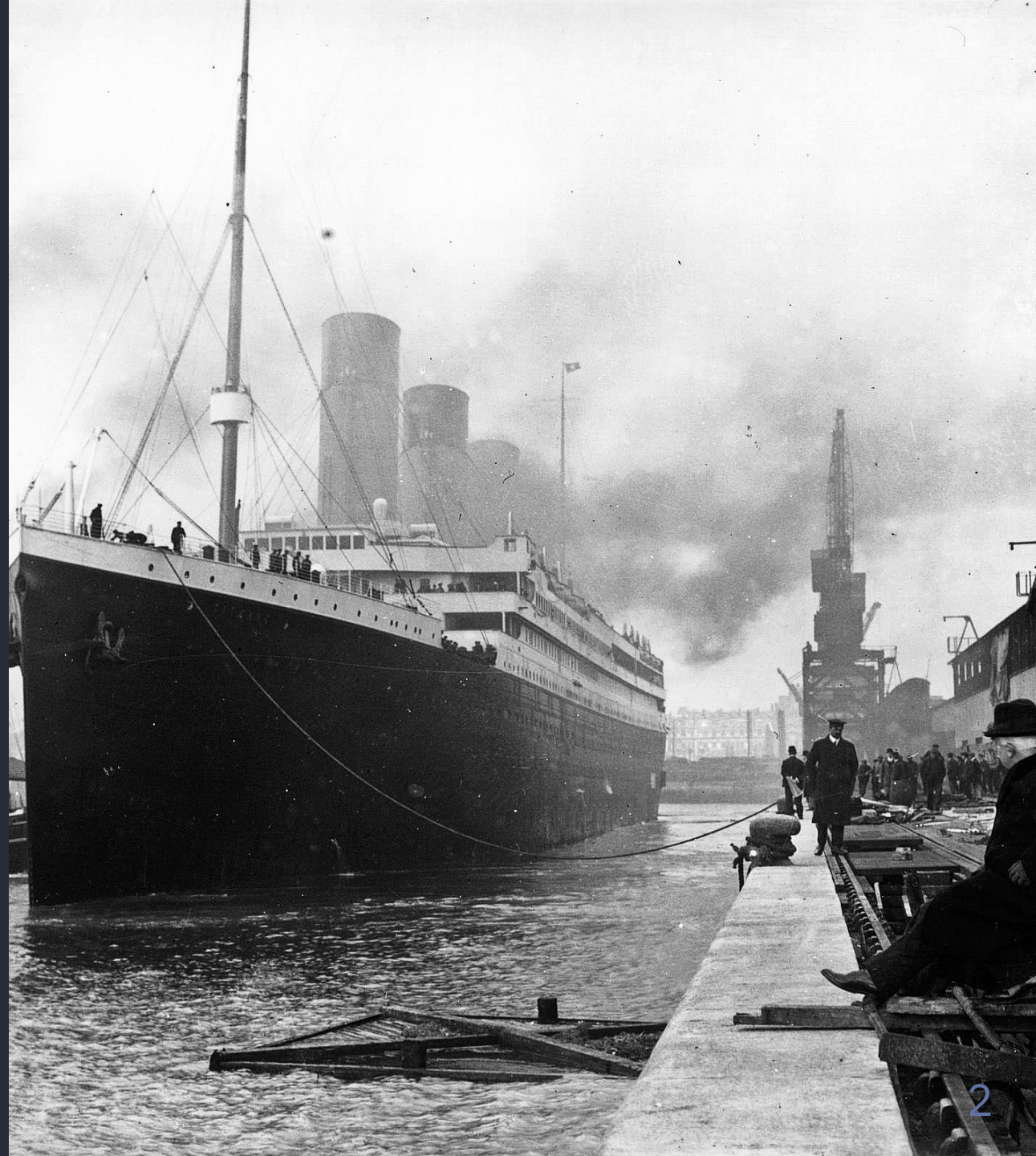
Titanic Challenge *

Qzer - 2022

* Amine, Carlo, Davide, Enrico, Federico, Giulio

Il Titanic

Il **Titanic** è stato un transatlantico britannico naufragato nelle prime ore del 15 aprile 1912, durante il suo viaggio inaugurale, a causa della collisione con un iceberg.



Qualche numero

I passeggeri del Titanic erano teoricamente **2224**, di cui:

- 324 in prima classe;
- 284 in seconda;
- 709 in terza;
- 906 membri dell'equipaggio

Morti

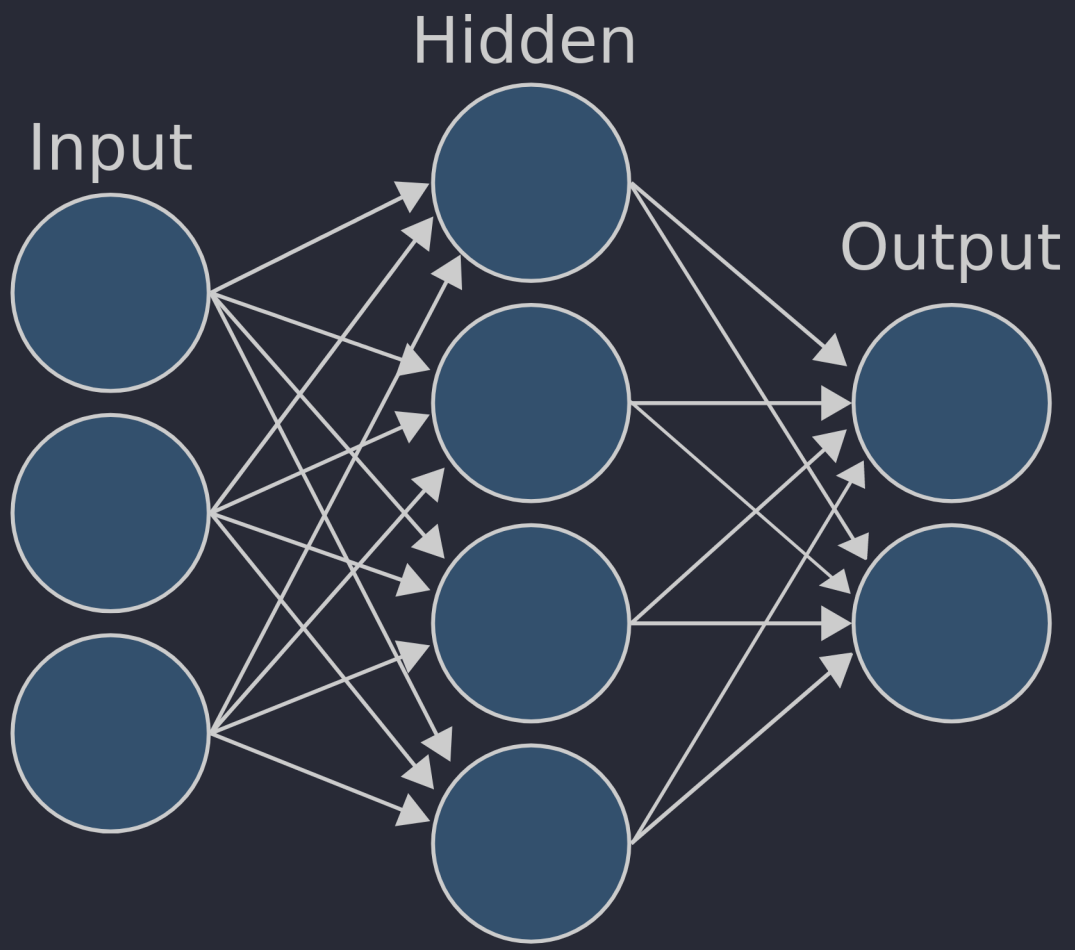
Stando ai numeri ufficiali, nel disastro persero la vita ben 1502, **67,54%**

Challenge

Costruire un modello predittivo che risponda alla domanda: **“che tipo di persone avevano maggiori probabilità di sopravvivere?"**

La **Titanic Challenge** è un esempio classico di problema di classificazione che viene spesso utilizzato per mostrare come funzionano le reti neurali

Rete neurale



Dati

In questa competizione avremo accesso a due set di dati simili che includono informazioni sui passeggeri come:

- nome;
- età;
- genere;
- classe socio-economica;
- ecc.

Train.csv

`Train.csv` conterrà i dettagli di un sottoinsieme dei passeggeri a bordo (891 per l'esattezza) e, soprattutto, conterrà l'informazione relativa al loro destino

```
training_set = pd.read_csv('/kaggle/input/titanic/train.csv')
training_set.info()

RangeIndex: 891 entries, 0 to 890
 #   Column       Non-Null Count  Dtype
 #   ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
```

Test.csv

Il file `test.csv` contiene informazioni su altri 418 passeggeri, non rivelando la loro sorte.

Il compito della challenge è, utilizzando i dati del file `train.csv`, prevedere se questi 418 passeggeri a bordo sopravviveranno

```
testing_set = pd.read_csv('/kaggle/input/titanic/test.csv')
testing_set.info()

RangeIndex: 418 entries, 0 to 417
 #   Column       Non-Null Count  Dtype  
 #   ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Pclass       418 non-null    int64  
 2   Name         418 non-null    object 
 3   Sex          418 non-null    object 
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64  
 6   Parch        418 non-null    int64  
 7   Ticket       418 non-null    object 
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object 
 10  Embarked     418 non-null    object
```

Data selection

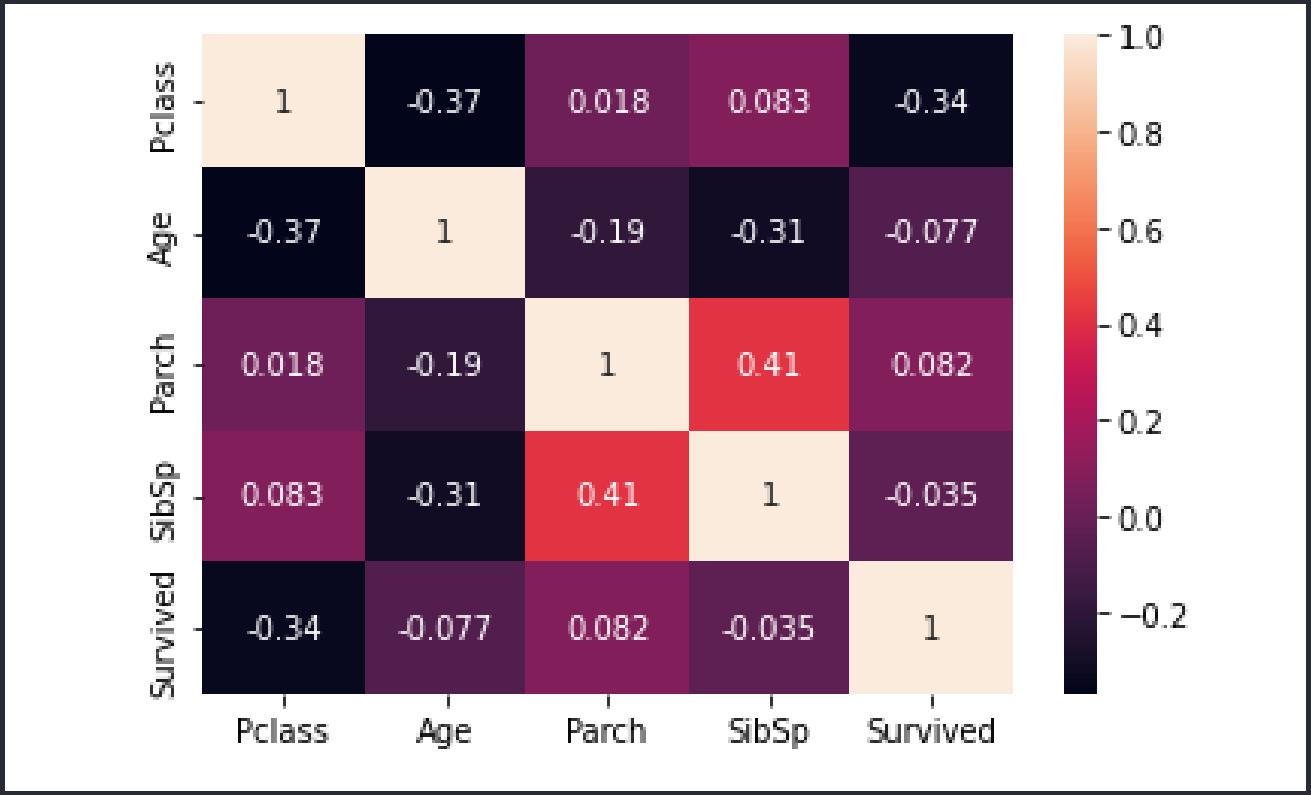
Alcune colonne non sono utili per raggiungere il nostro obiettivo. Il primo passo consiste quindi nel **selezionare le colonne** coi dati che, ipoteticamente, possono avere un'influenza sulla sopravvivenza di un passeggero

```
clean_training_set = training_set[["Pclass", "Sex", "Age", "Parch", "SibSp", "Embarked", "Survived"]]
```

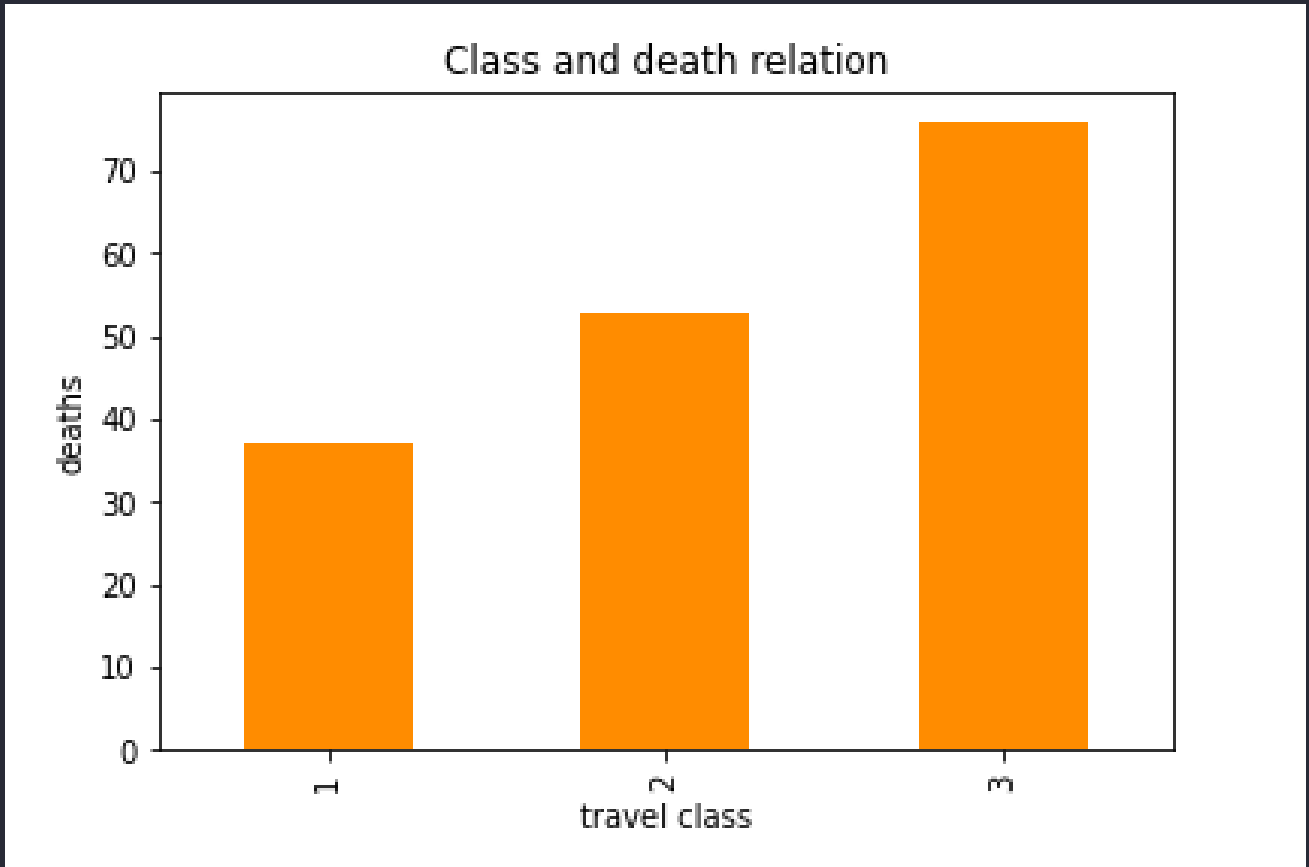
Queste le colonne trattenute dal file `train.csv`

```
RangeIndex: 891 entries, 0 to 890
 #   Column    Non-Null Count  Dtype
 #   ------    --------------  -----
 0   Pclass    891 non-null    int64
 1   Sex       891 non-null    object
 2   Age       714 non-null    float64
 3   Parch     891 non-null    int64
 4   SibSp     891 non-null    int64
 5   Embarked  889 non-null    object
 6   Survived  891 non-null    int64
```

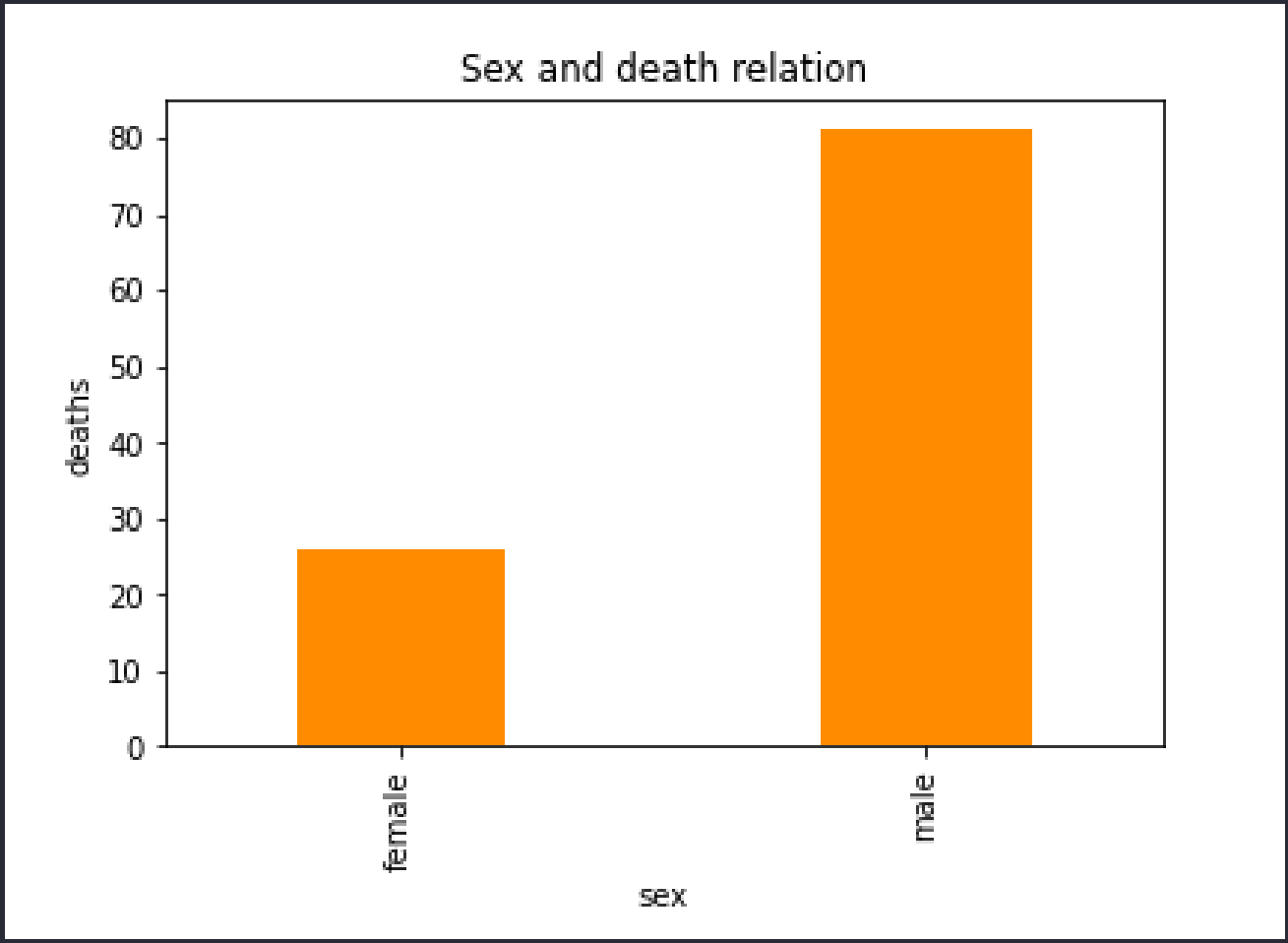
Correlazioni



Classe di viaggio



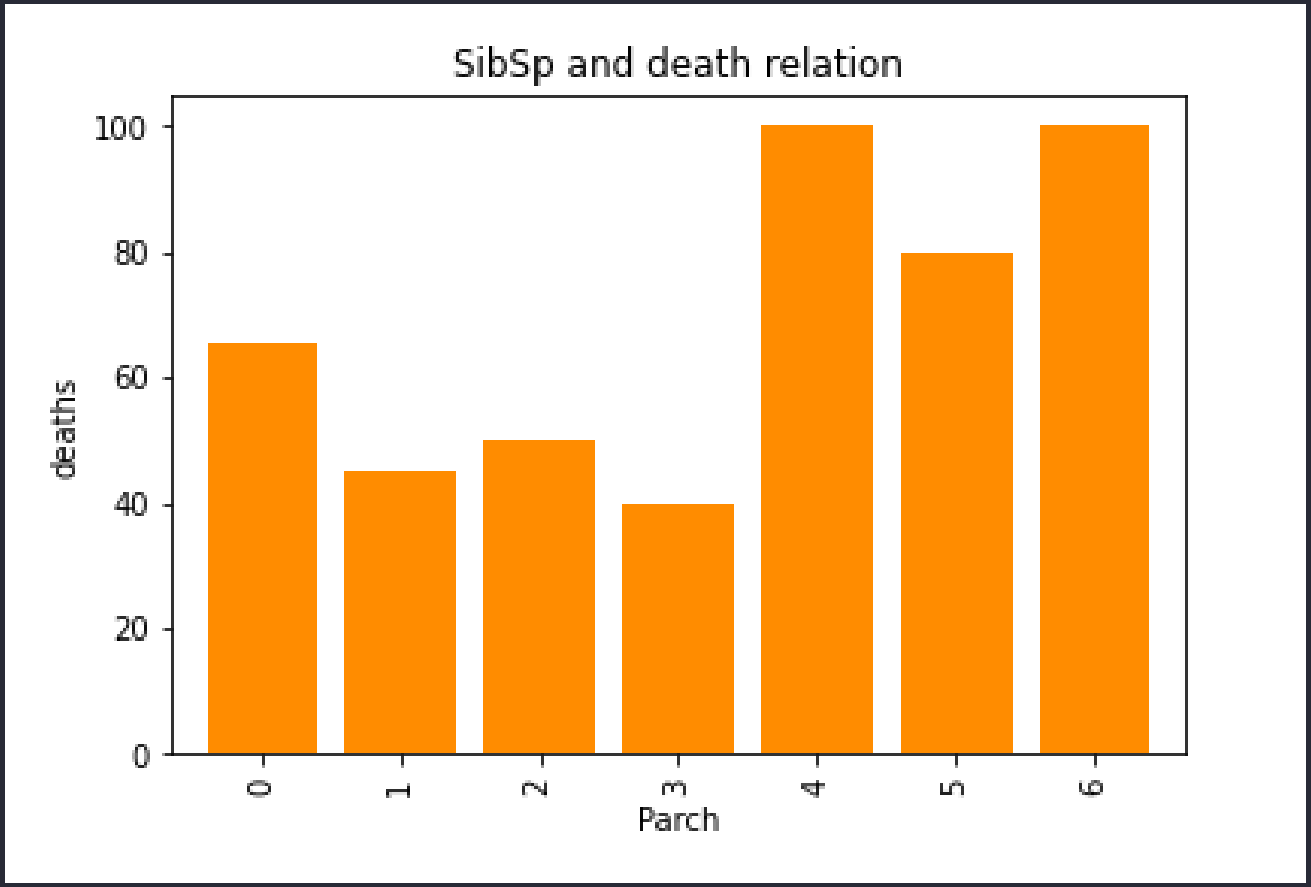
Sesso


Sex and death relation
deaths
0
10
20
30
40
50
60
70
80
female
male
sex

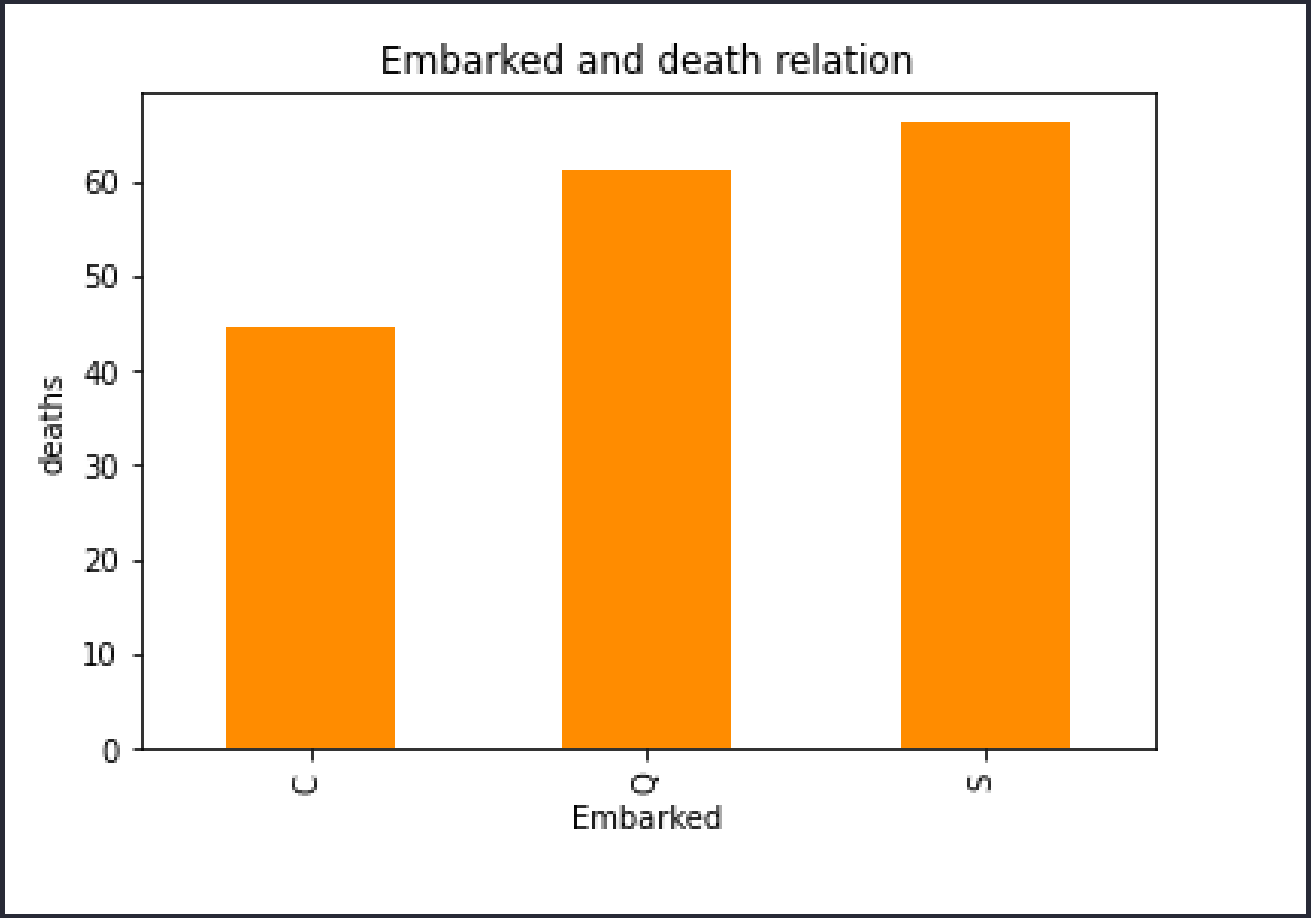
Età


Age and death relation
survived
age
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80

Matrimonio


SibSp and death relation
deaths
100
80
60
40
20
0
0
1
2
3
4
5
6
Parch

Luogo d'imbarco



Normalizzazione

	PassengerId	Pclass	Sex	Age	Parch	SibSp	Embarked
0	892	3	0	34.5	0	0	Q
1	893	3	1	47.0	0	1	S
2	894	2	0	62.0	0	0	Q
3	895	3	0	27.0	0	0	S
4	896	3	1	22.0	1	1	S
...	...	...	...	...	...	...	...
413	1305	3	0	NaN	0	0	S
414	1306	1	1	39.0	0	0	C
415	1307	3	0	38.5	0	0	S
416	1308	3	0	NaN	0	0	S
417	1309	3	0	NaN	1	1	C

Normalizzare il DataSet

Convertire i **maschi** nel valore **0** e le **femmine** nel valore **1**

Sostituire i valori **NaN** della colonna Age con **l'età media**, ossia 30 anni

Nella colonna **Embarked** sostituire coi numeri **0**, **1** e **2** le lettere **C**(herbourg), **Q**(ueenstown) e **S**(outhampton):

```
# Transforms from numbers to strings
df["Embarked"]=df.Embarked.map({"C":0,"Q":1, "S":2})
```

Normalizzare il DataSet

Dividere la colonna Age in **Child**, **Adult** e **Elderly**

```
child_list = df['Age'].apply(lambda x: 1 if x < 18 else 0)
df.insert(4, "Child", child_list, True)

adult_list = df['Age'].apply(lambda x: 1 if x >= 18 and x < 50 else 0)
df.insert(5, "Adult", adult_list, True)

elderly_list = df['Age'].apply(lambda x: 1 if x > 50 else 0)
df.insert(6, "Elderly", elderly_list, True)
```

DataSet Normalizzato

	PassengerId	Pclass	Sex	Child	Adult	Elderly	Parch	SibSp	Embarked
0	892	3	0	0	1	0	0	0	1
1	893	3	1	0	1	0	0	1	2
2	894	2	0	0	0	1	0	0	1
3	895	3	0	0	1	0	0	0	2
4	896	3	1	0	1	0	1	1	2
...	...	...	...	...	...	...	...	...	...
413	1305	3	0	0	1	0	0	0	2
414	1306	1	1	0	1	0	0	0	0
415	1307	3	0	0	1	0	0	0	2
416	1308	3	0	0	1	0	0	0	2
417	1309	3	0	0	1	0	1	1	0

Algoritmo

Multi-layer Perceptron classifier (MLP)

L'algoritmo **MLP** è un metodo per addestrare le reti neurali multistrato. Consiste nel modificare i pesi delle connessioni tra i neuroni della rete neurale in modo da ridurre l'errore tra l'output della rete neurale e l'output desiderato. L'algoritmo viene ripetuto finché l'errore non raggiunge un livello accettabile.

L'accuratezza dell'algoritmo è del **77.04**

Github

Per maggiori informazioni

TitanicKaggle

Tecnologie utilizzate

- VSCode
- Git
- Markdown, Marp e CSS3
- Python
- Pandas
- Jupyter Notebook

Bibliografia

Titanic - Wikipedia
Passeggeri del Titanic - Wikipedia
Titanic Challenge - Kaggle

Grazie