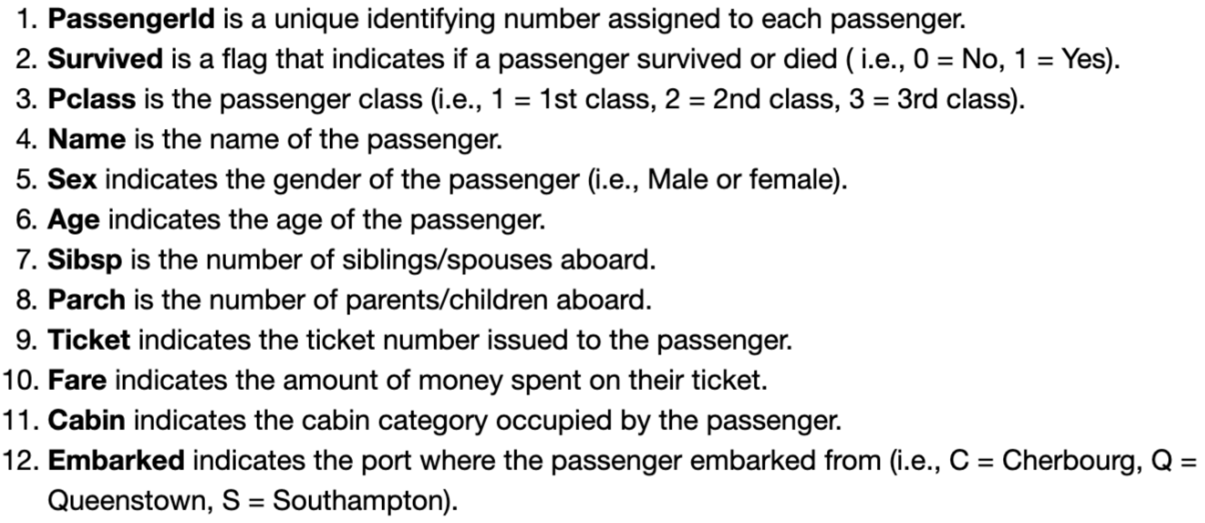
Titanic: survival prediction

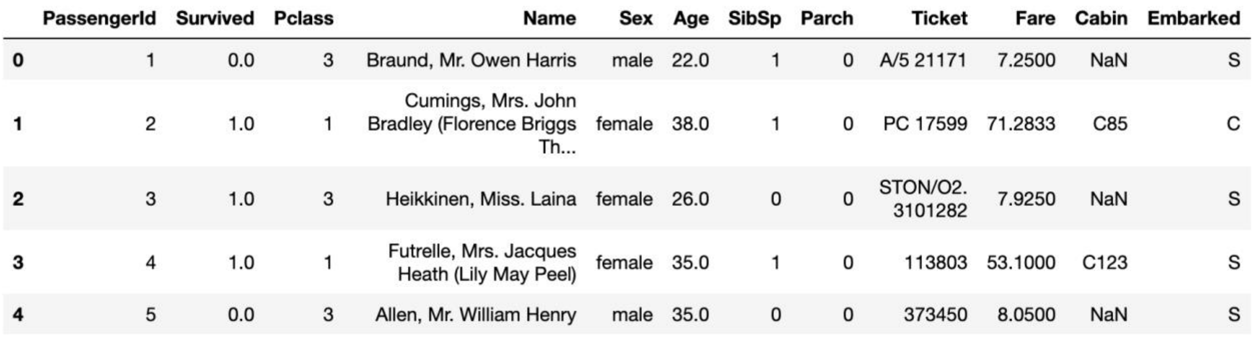
Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. The analysis of what sorts of people were likely to survive is developed. In particular, the tools of machine learning are applied to predict which passengers survived the tragedy.

Data



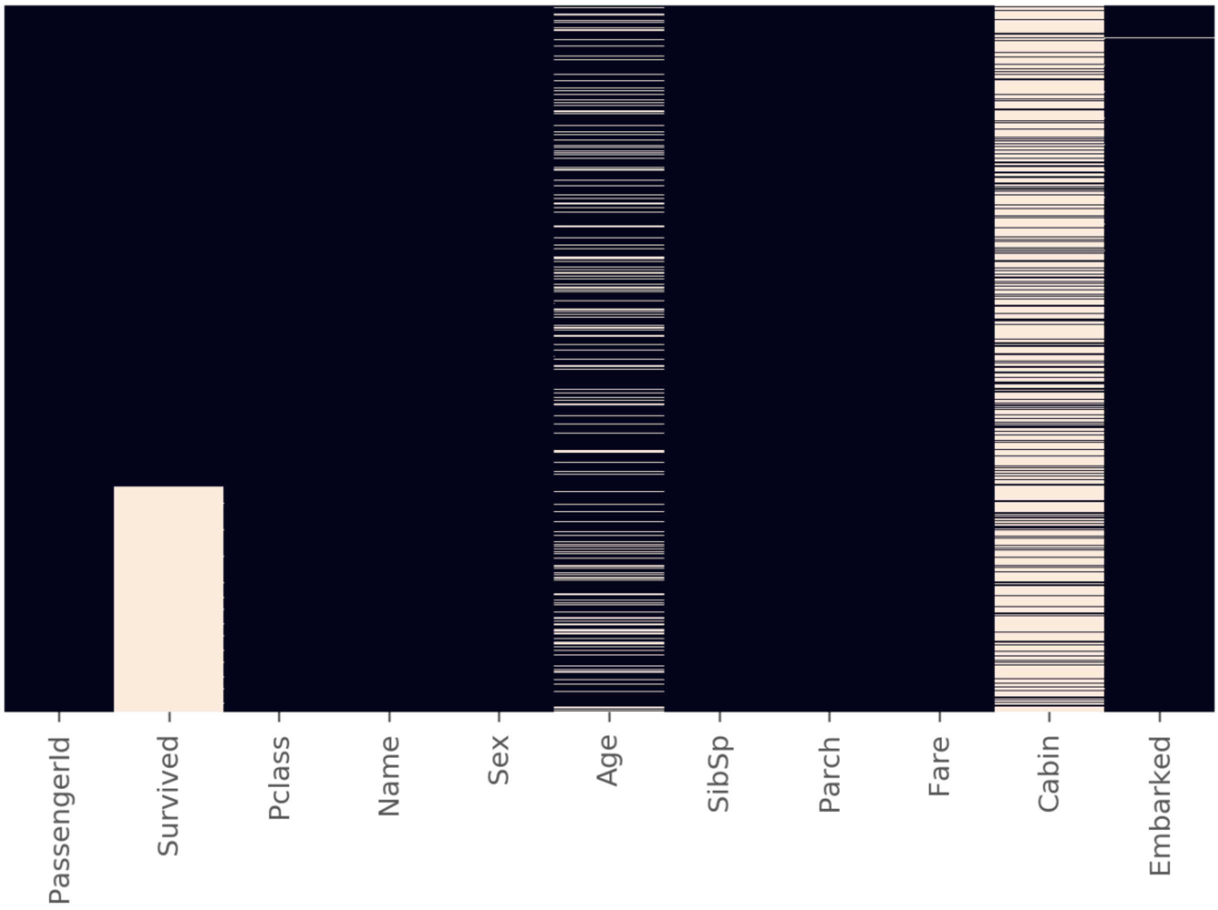
1. **PassengerId** is a unique identifying number assigned to each passenger.
2. **Survived** is a flag that indicates if a passenger survived or died (i.e., 0 = No, 1 = Yes).
3. **Pclass** is the passenger class (i.e., 1 = 1st class, 2 = 2nd class, 3 = 3rd class).
4. **Name** is the name of the passenger.
5. **Sex** indicates the gender of the passenger (i.e., Male or female).
6. **Age** indicates the age of the passenger.
7. **Sibsp** is the number of siblings/spouses aboard.
8. **Parch** is the number of parents/children aboard.
9. **Ticket** indicates the ticket number issued to the passenger.
10. **Fare** indicates the amount of money spent on their ticket.
11. **Cabin** indicates the cabin category occupied by the passenger.
12. **Embarked** indicates the port where the passenger embarked from (i.e., C = Cherbourg, Q = Queenstown, S = Southampton).



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Methodology

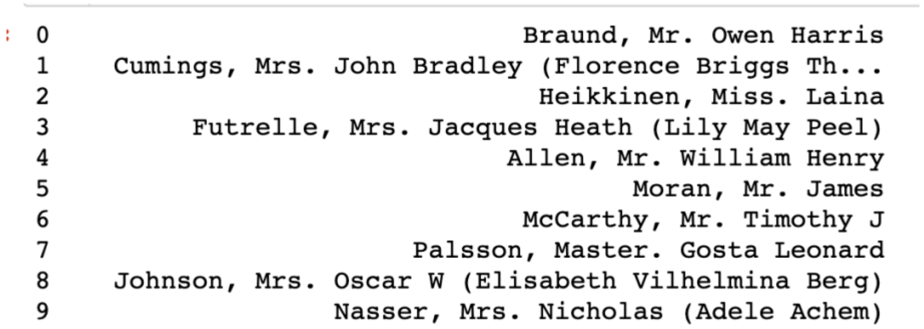
The first thing to do in handling the missing variable. By plotting an heatmap we get:



Due to the many Cabin missing that feature is directly dropped.

The Parch and SibSp are used to create a new feature called ‘family_size’.

The name has been processed to get information about the age, in particular:



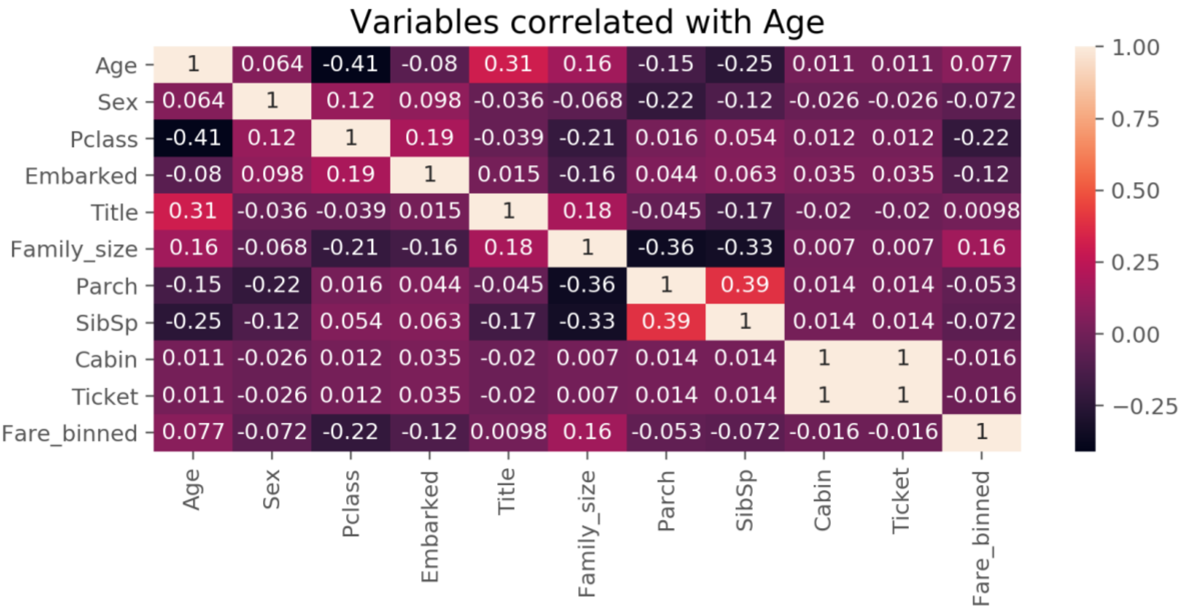
```
0                                  Braund, Mr. Owen Harris
1    Cumings, Mrs. John Bradley (Florence Briggs Th...
2                                   Heikkinen, Miss. Laina
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                                 Allen, Mr. William Henry
5                                         Moran, Mr. James
6                                  McCarthy, Mr. Timothy J
7                           Palsson, Master. Gosta Leonard
8    Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
9                     Nasser, Mrs. Nicholas (Adele Achem)
```

- This column contains string titles such as Mr, Mrs, Master etc.

- These titles give us useful information about sex and age for example Mr=Male, Mrs=Female and married, miss= Female and young.
- Now we want to extract these titles from Name to check if there is any association between these titles and Survived.
- We can see there are several titles with the very least frequency. So, it makes sense to put them in fewer buckets.
- Professionals like Dr, Rev, Col, Major, Capt will be put into ‘Graduated’ bucket.
- Titles such as Dona, Jonkheer, Countess, Sir, Lady, Don were usually entitled to the aristocrats.
- We would also replace Mlle and Ms with Miss and Mme by Mrs as these are French titles.

All the variables are made categorical to increase the performance of the decision tree algorithm.

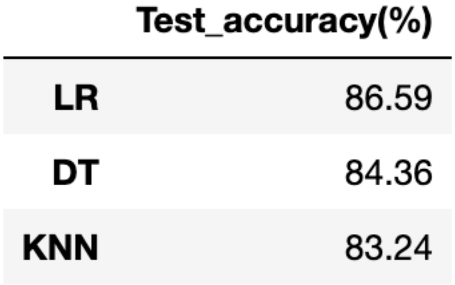
The age has been filled with the median of the Title and Pclass feature according to their correlation:



The algorithms used are: Decision Tree, KNN and Logistic Regression.
This choice have been made because it is a classification problem.

Results:

The algorithm that performs better is the Logistic Regression (LR).
The train dataset has been split in 80-20 because the test set is not labelled.



	Test_accuracy(%)
LR	86.59
DT	84.36
KNN	83.24

Conclusion and next step:

The Logistic Regression has an accuracy of 86.59% on the test set that is quite good.
To improve the model reliability and its accuracy a k-fold cross-validation can be performed and also an hyper-parameter scan.