Titanic Survival Prediction Using Machine Learning

The RMS Titanic was known as the unsinkable ship and was the largest, most luxurious passenger ship of its time. Sadly, the British ocean liner sank on April 15, 1912, killing over 1500 people while just 705 survived.

In this article, we will analyze the Titanic data set and make two predictions. One prediction to see which passengers on board the ship would survive and then another prediction to see if we would’ve survived.

# Data Set Column Descriptions

* **pclass:** Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
* **survived:** Survival (0 = No; 1 = Yes)
* **name:** Name
* **sex:** Sex
* **age:** Age
* **sibsp:** Number of siblings/spouses aboard
* **parch:** Number of parents/children aboard
* **fare:** Passenger fare (British pound)
* **embarked:** Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

## ****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:

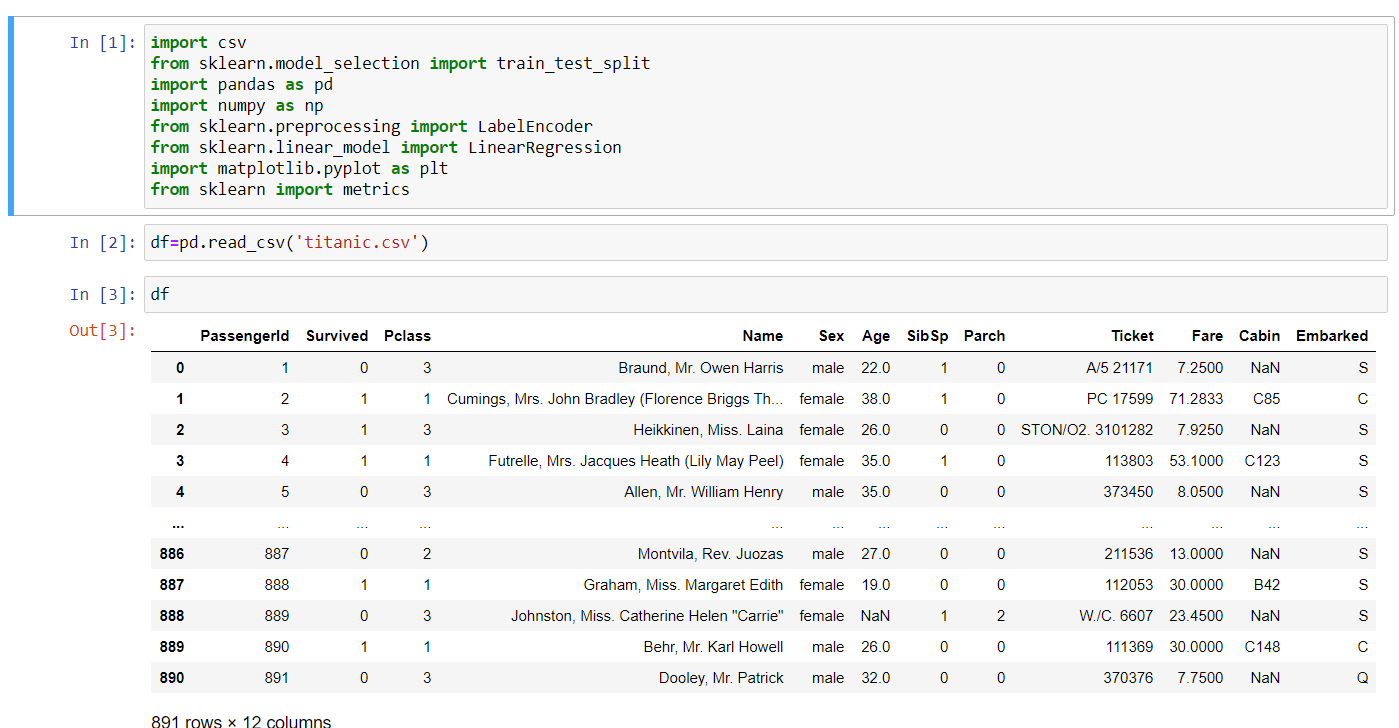
* Parent= mother, father
* Child= daughter, son, stepdaughter, stepson  
  Some children traveled only with a nanny, therefore parch=0 for them.

Start Programming

The first thing that I like to do before writing a single line of code is to put in a description in the comments of what the code does. This way, I can look back on my code and know exactly what it does.

# Description: This program predicts if a passenger will survive on the titanic

Now import the packages /libraries to make it easier to write the program. And load the data from where you have saved the csv file and print a few rows.



Now, I will analyze the data by getting counts of data, survival rates, and creating charts to visualize the data.

Getacount of the number of rows and columns in the data set. Note that each row is a passenger onboard the ship and the columns are data points for each passenger.

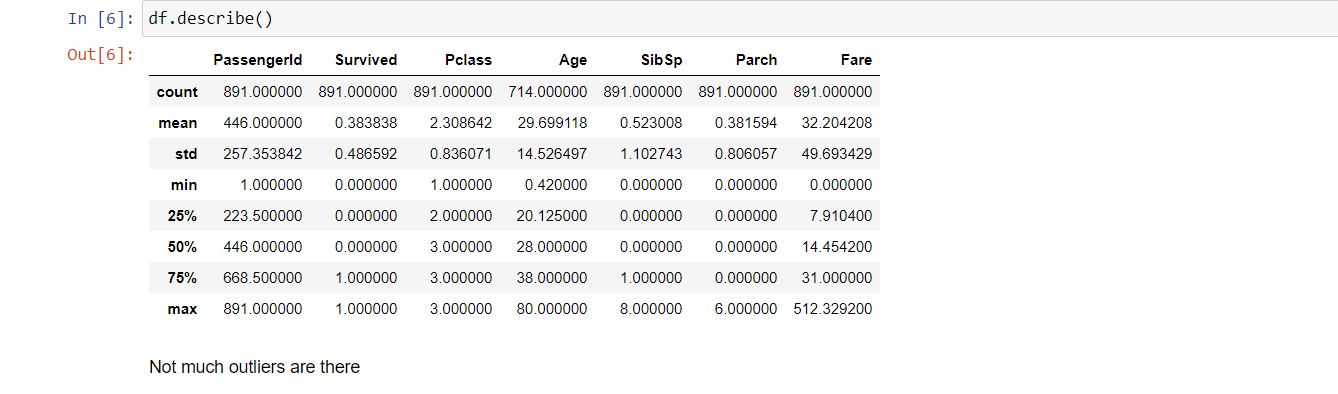
There are 891 rows/passengers and 15 columns/data points in the data set.

Now Get some statistics on the data set, such as the count, mean, standard deviation, etc.

After getting these statistics, I see the max price/fare a passenger paid for a ticket in this data set was 512.3292 British pounds, and the minimum price/fare was 0 British pounds.

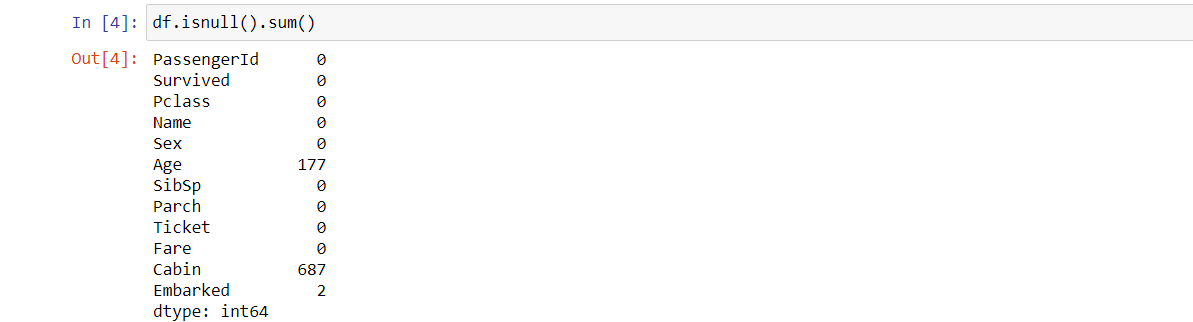
We can also see that there is some missing data for the age column as it’s less than 891 (the number of passengers in this data set).

The mean age is 29.699 and the oldest passenger in this data set was 80 years old, while the youngest was only .42 years old (about 5 months).

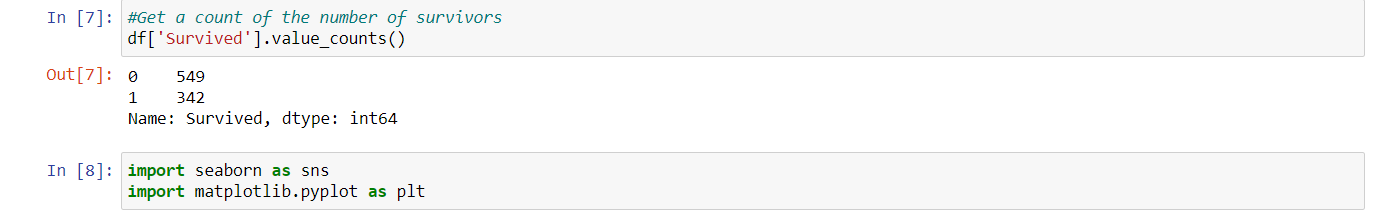


Check which columns contain empty values (NaN, NAN, na). Looks like columns age, embarked and cabin are missing some values.

All the other columns are not missing any values.

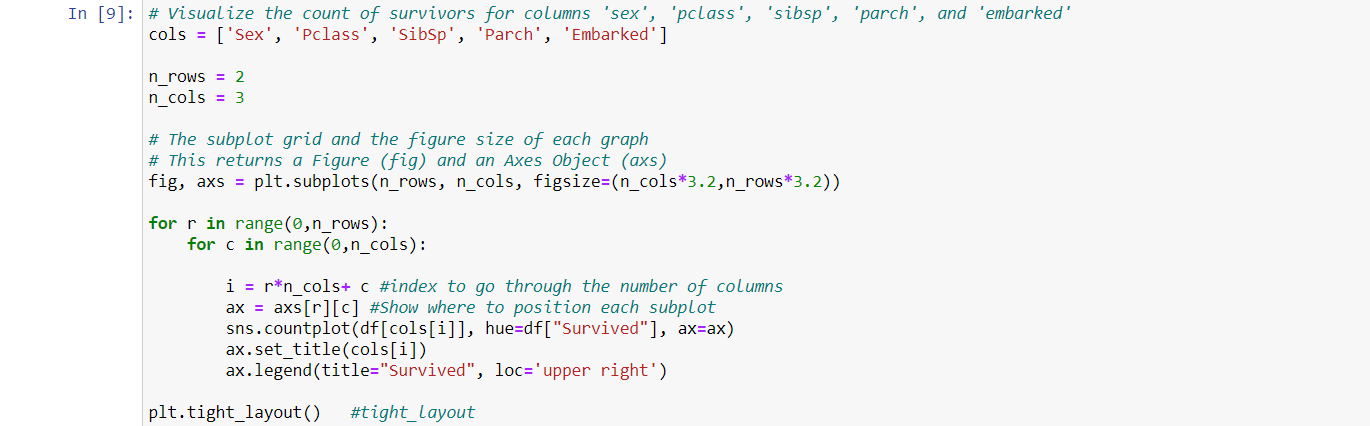


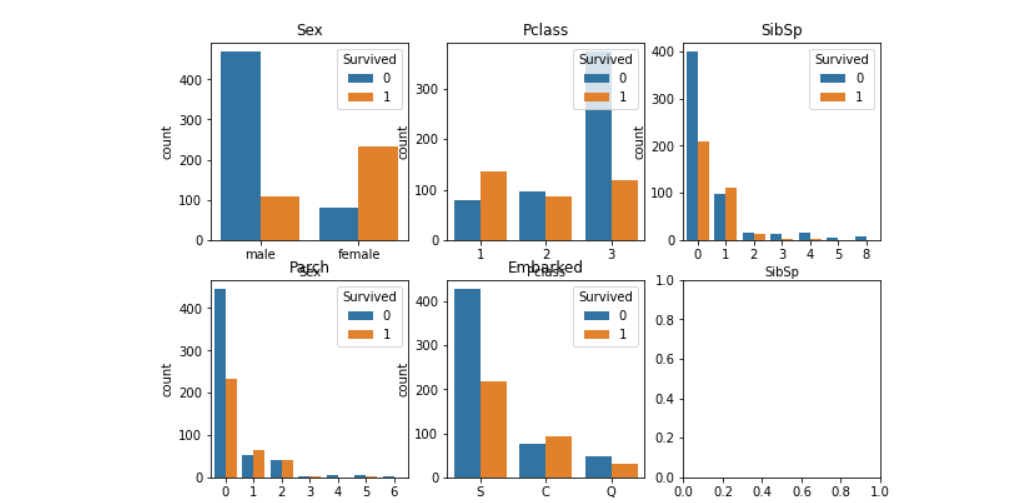
Get acount of the number of survivors on board the Titanic in this data set. Notice that, in this data set, there were more passengers that didn’t survive (549) than did (343).



Visualize the count of survivors for the columns who*,*sex*,*pclass*,*sibsp*,*parch*,*andembarked*.*

* From the charts below, we can see that
* Females are most likely to survive from the chart sex.
* Third class is most likely to not survive by chart pclass.
* If you have 0 siblings or spouses on board, you are not likely to survive according to chart sibsp.
* If you have 0 parents or children on board, you are not likely to survive according to the parch chart.
* If you embarked from Southampton (S), you are not likely to survive according to the embarked chart.





Next, I want to take a look at the survival rate by sex.

From the table below, we can see that about 74.2% of females survived and about 18.89% of males survived. And I want to Look at thesurvival rate by sex and class.

From the pivot table below, we see that females in first class had a survival rate of about 96.8%, meaning the majority of them survived.

Males in third class had the lowest survival rate at about 13.54%, meaning the majority of them did not survive.



Plot the prices paid for each class.



Next, I will drop the redundant columns that are non-numerical and remove rows with missing values.

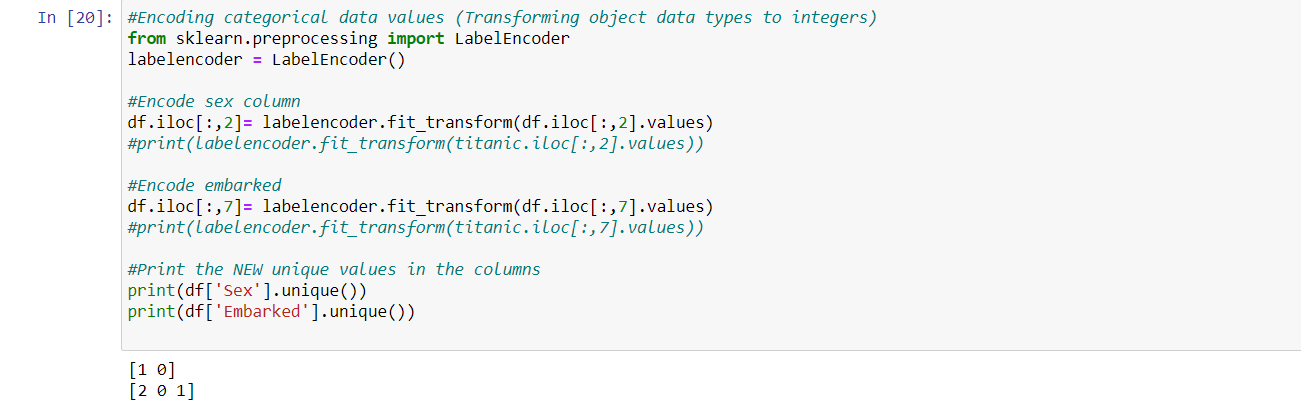
I also decided to drop the column called cabin because it’s missing 688 rows of data which means 688/891 = 77.22% of the data is missing for this column.

Then, let’s see the new number of rows and columns in the Titanic data set.

Then Look at the data types to see which columns need to be transformed/encoded to a number. It looks like column sex and embarked are the only two columns that need to be transformed.



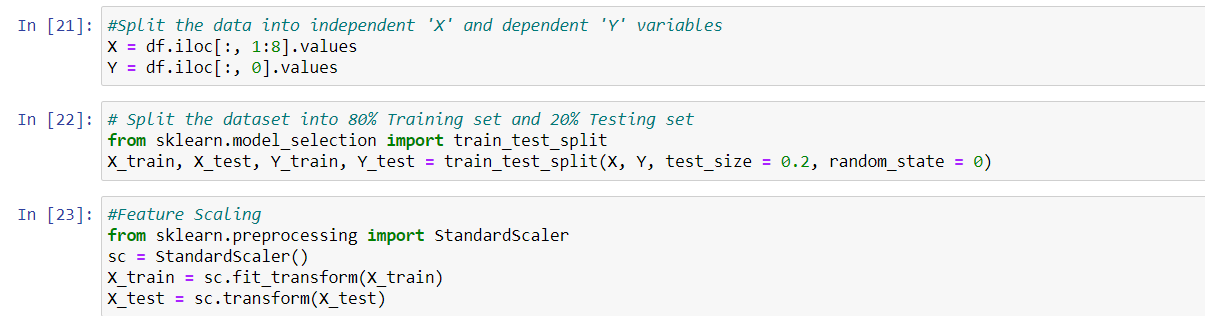
Change the non-numeric data to numeric data through label encoder , and print the new values



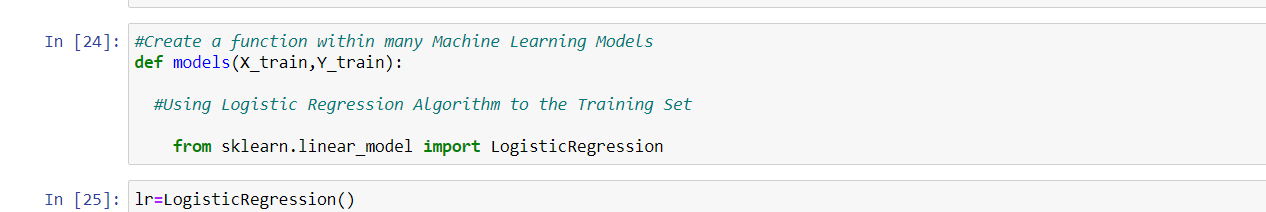
Split the data into independent ‘X’ and dependent ‘Y’ data sets.

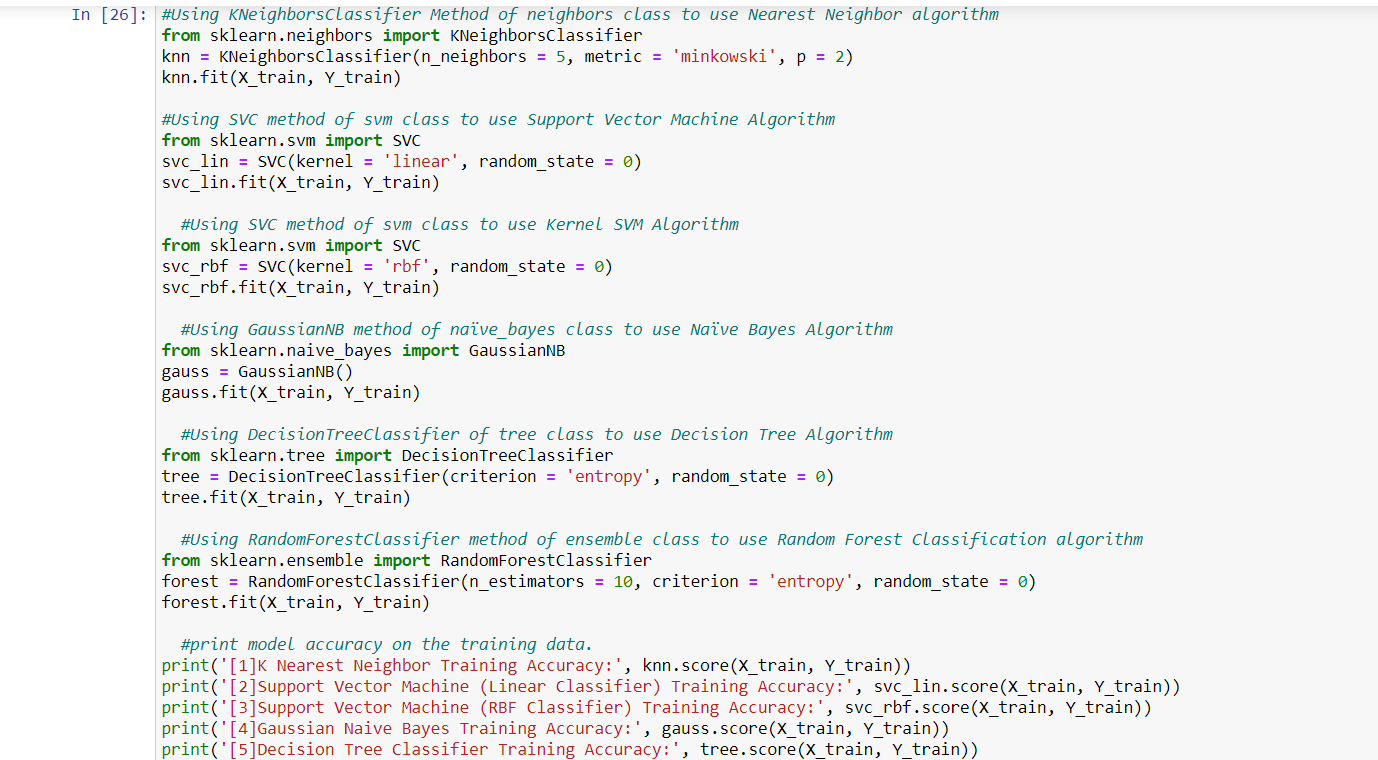
Split the data again, this time into 80% training (X\_train and Y\_train) and 20% testing (X\_test and Y\_test) data sets.

Then Optionally, we can scale the data,meaning the data will be within a specific range, for example 0–100 or 0–1.



Create a function that has within it many different machine learning models that we can use to make our predictions.





The scores are as follows:

[1]K Nearest Neighbor Training Accuracy: 0.8664323374340949

[2]Support Vector Machine (Linear Classifier) Training Accuracy: 0.7768014059753954

[3]Support Vector Machine (RBF Classifier) Training Accuracy: 0.8506151142355008

[4]Gaussian Naive Bayes Training Accuracy: 0.8031634446397188

[5]Decision Tree Classifier Training Accuracy: 0.9929701230228472

[6]Random Forest Classifier Training Accuracy: 0.9753954305799648

Get and train all the models and store them in a variable called model.

The model that was most accurate on the training data was the Decision Tree Classifier with an accuracy of 99.29%, according to fig

Now just save the model by importing pickle and then check and test the decision tree model. The codes are as follows:



# Conclusion and Resources

That is it, you are done creating your program to predict if a passenger would survive the Titanic or not!