IBM Developer

April 29, 2021

Predicting Titanic Shipwreck Survivors

Today's Agenda

1 Introduction to the Problem

2 Data Utilized

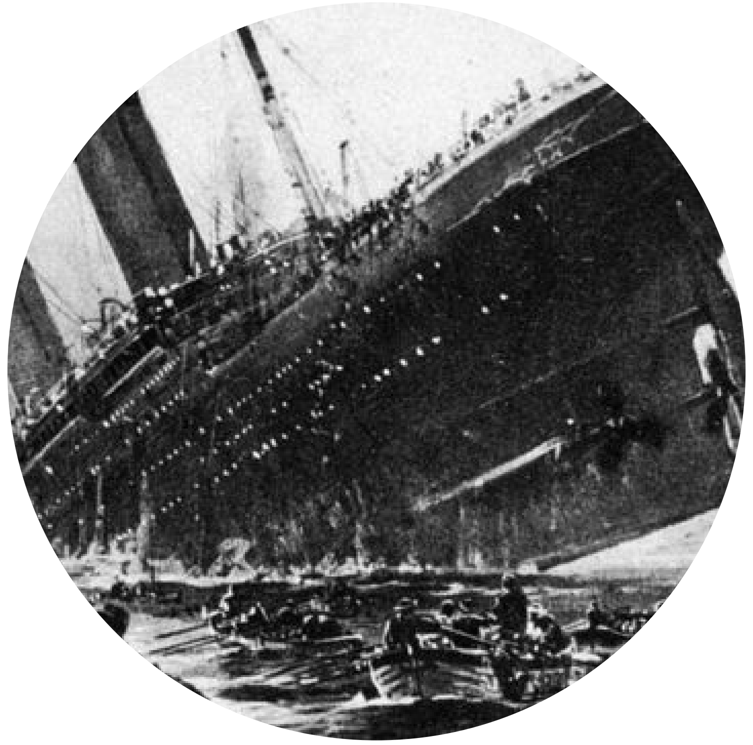
3 Analysis

4 Conclusions

Introduction

The Titanic Shipwreck is the most infamous one in history

The goals are to utilize data science and machine-learning to successfully predict the probability of surviving the titanic shipwreck.
This project was promoted by Kaggle for their annual data-science competition.



Are you ready?

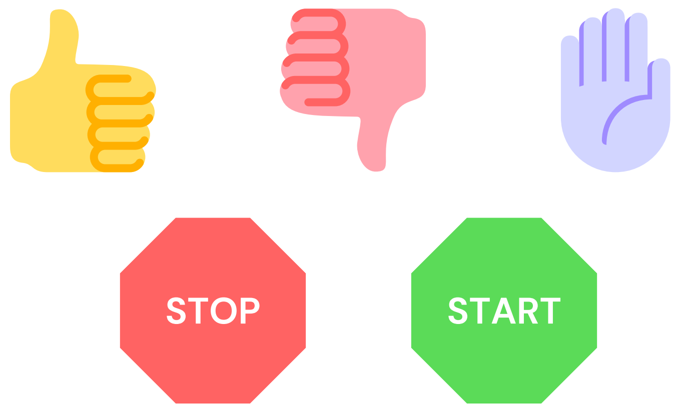
Let's Begin!

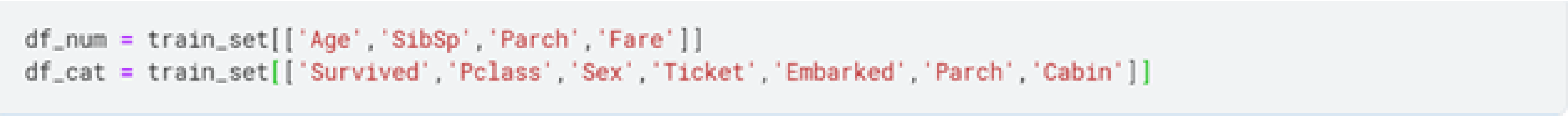
Data Analyzed

The data is provided from Kaggle, separating in three CSVs. A train set, test set and gender_submission, to analyze and correctly predict the target mentioned above. The data is classified in the following columns:

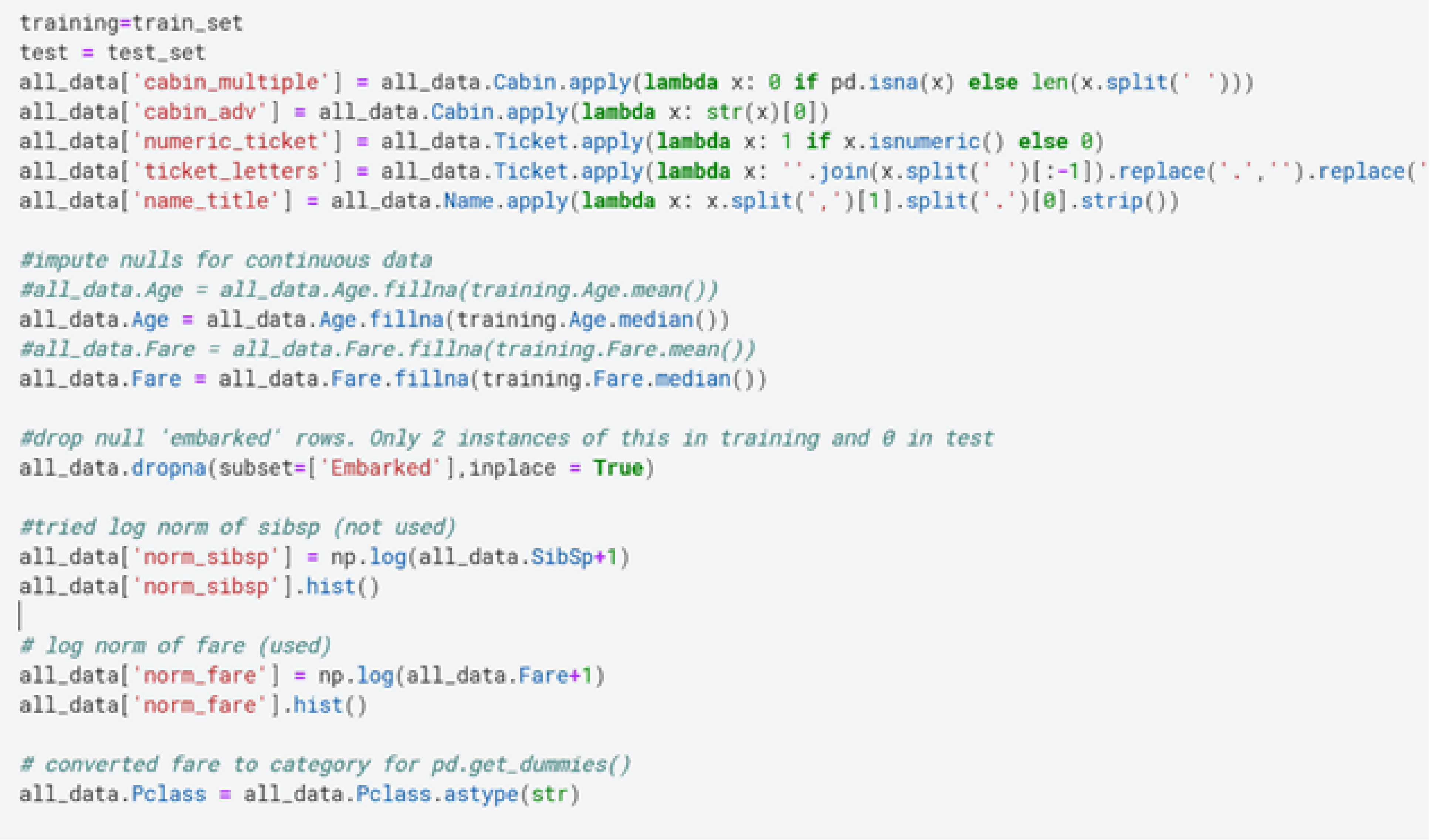
1. Separate into numerical and categorical variables
2. Data Cleansing

3. IMplementing analysis and machine learning algorithms





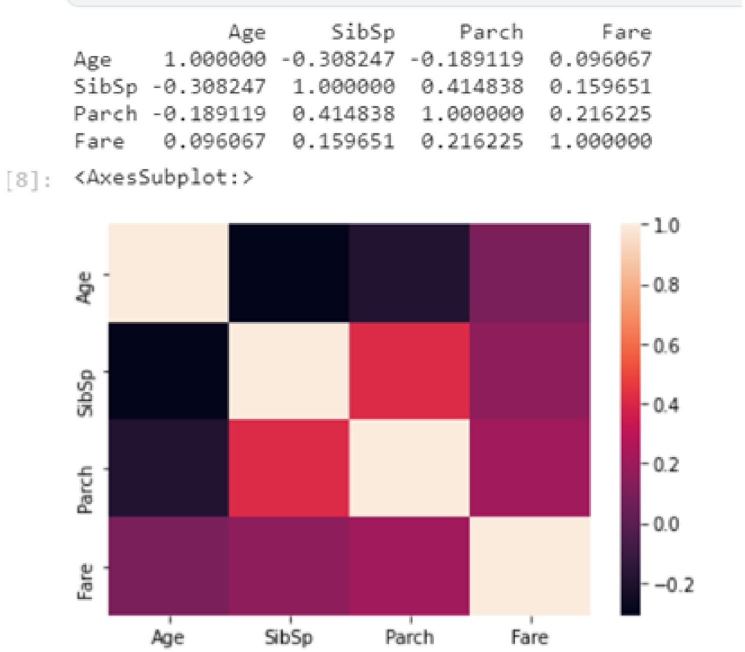
```
df_num = train_set[['Age','SibSp','Parch','Fare']]
df_cat = train_set[['Survived','Pclass','Sex','Ticket','Embarked','Parch','Cabin']]
```



```
training=train_set
test = test_set
all_data['cabin_multiple'] = all_data.Cabin.apply(lambda x: 0 if pd.isna(x) else len(x.split(' ')))
all_data['cabin_adv'] = all_data.Cabin.apply(lambda x: str(x)[0])
all_data['numeric_ticket'] = all_data.Ticket.apply(lambda x: 1 if x.isnumeric() else 0)
all_data['ticket_letters'] = all_data.Ticket.apply(lambda x: ''.join(x.split(' ')[:-1]).replace('.','').replace('
all_data['name_title'] = all_data.Name.apply(lambda x: x.split(',')[1].split('.')[0].strip())

#impute nulls for continuous data
#all_data.Age = all_data.Age.fillna(training.Age.mean())
all_data.Age = all_data.Age.fillna(training.Age.median())
#all_data.Fare = all_data.Fare.fillna(training.Fare.mean())
all_data.Fare = all_data.Fare.fillna(training.Fare.median())

#drop null 'embarked' rows. Only 2 instances of this in training and 0 in test
all_data.dropna(subset=['Embarked'],inplace = True)

#tried log norm of sibsp (not used)
all_data['norm_sibsp'] = np.log(all_data.SibSp+1)
all_data['norm_sibsp'].hist()

# log norm of fare (used)
all_data['norm_fare'] = np.log(all_data.Fare+1)
all_data['norm_fare'].hist()

# converted fare to category for pd.get_dummies()
all_data.Pclass = all_data.Pclass.astype(str)
```

Analysis

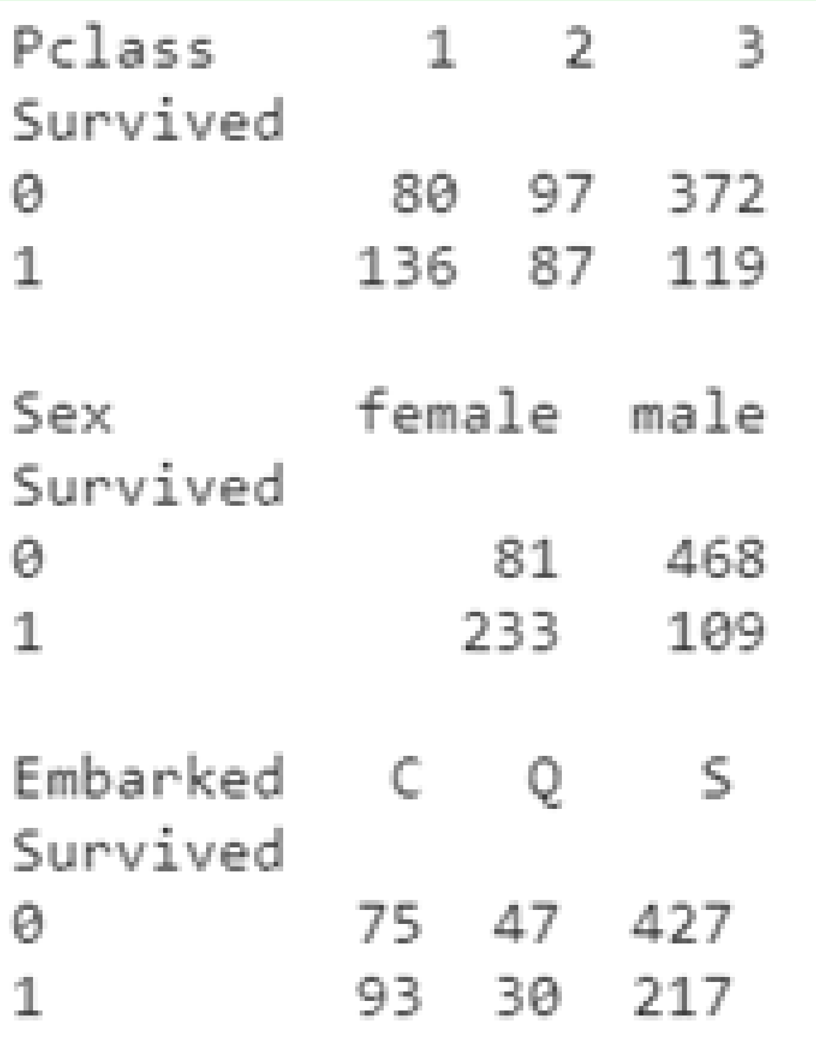
We utilized several variables against the target variable to see if there was any evident relationship between them.



```
          Age     SibSp     Parch      Fare
Age    1.000000 -0.308247 -0.189119  0.096067
SibSp -0.308247  1.000000  0.414838  0.159651
Parch -0.189119  0.414838  1.000000  0.216225
Fare   0.096067  0.159651  0.216225  1.000000
```

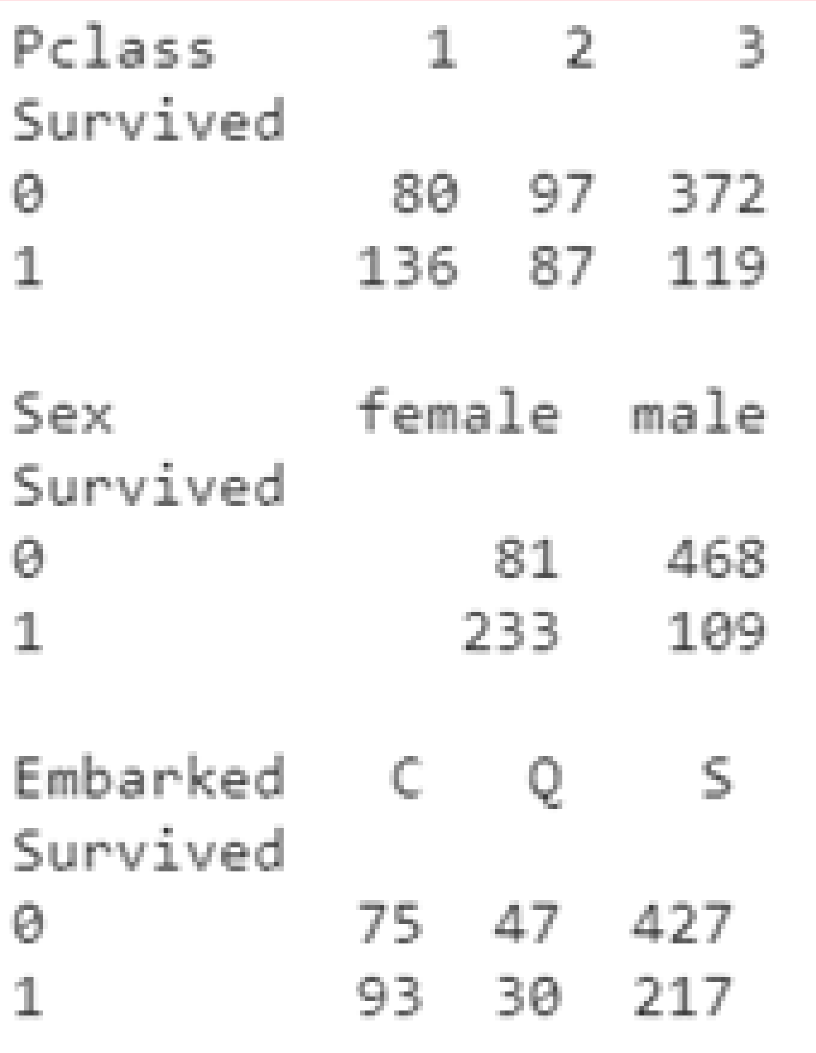
[8]: <AxesSubplot:>

AGE



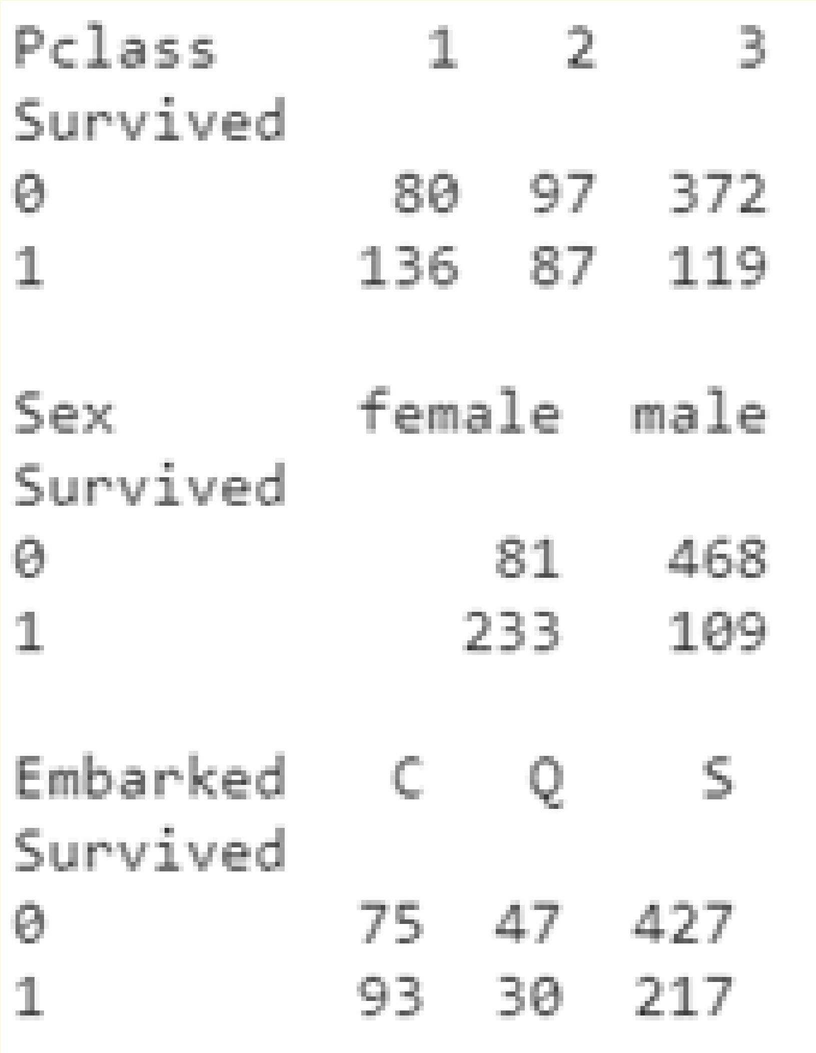
```
Pclass      1    2    3
Survived
0          80   97  372
1         136   87  119

Sex       female  male
Survived
0             81   468
1            233   109

Embarked   C   Q    S
Survived
0         75  47  427
1         93  30  217
```

GENDER



```
Pclass      1    2    3
Survived
0          80   97  372
1         136   87  119

Sex       female  male
Survived
0             81   468
1            233   109

Embarked   C   Q    S
Survived
0         75  47  427
1         93  30  217
```

TICKET CLASS



```
Pclass      1    2    3
Survived
0          80   97  372
1         136   87  119

Sex       female  male
Survived
0             81   468
1            233   109

Embarked   C   Q    S
Survived
0         75  47  427
1         93  30  217
```

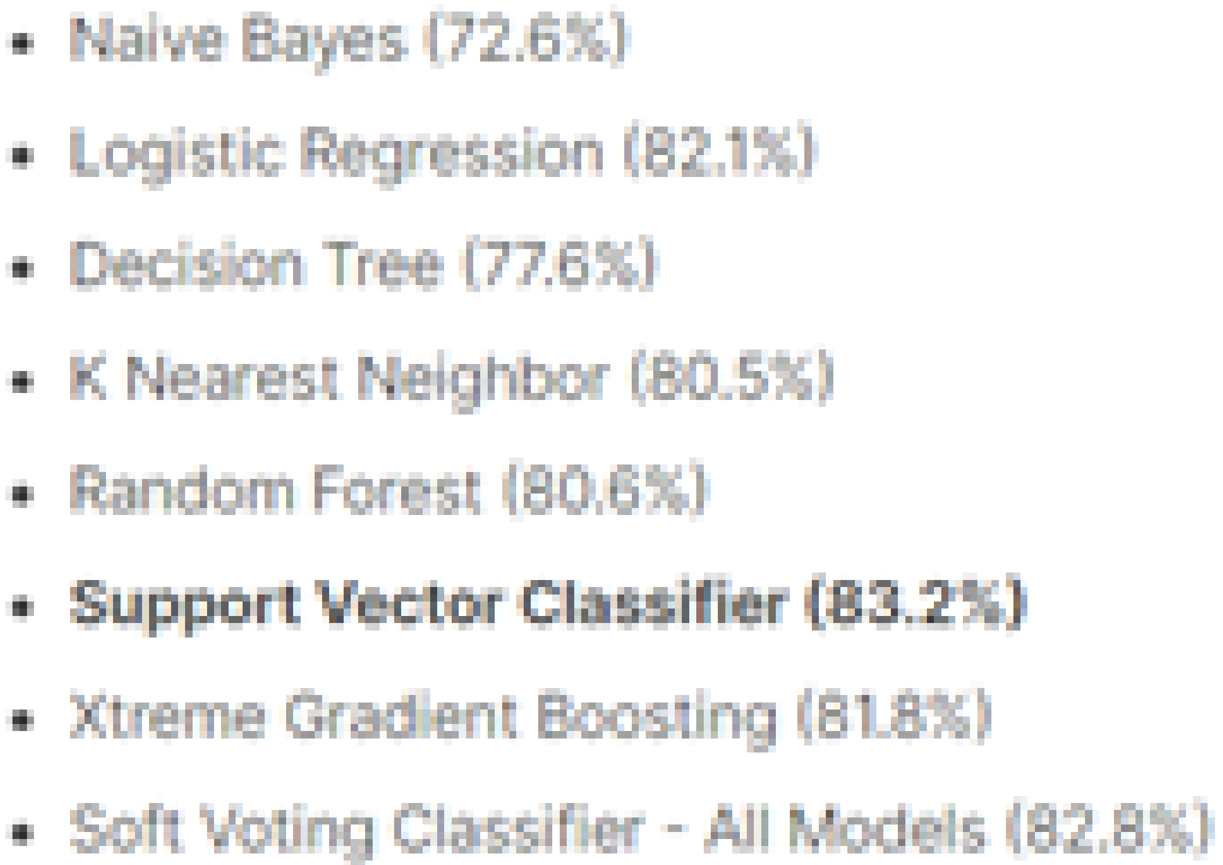
Machine-Learning

We utilized distinct models to see which one fitted the best in the data.

1 Choose a Machine-Learning Approach

2 Analyze and utilize the one that is the most effective

Results



- Naive Bayes (72.6%)
- Logistic Regression (82.1%)
- Decision Tree (77.6%)
- K Nearest Neighbor (80.5%)
- Random Forest (80.6%)
- **Support Vector Classifier (83.2%)**
- Xtreme Gradient Boosting (81.8%)
- Soft Voting Classifier - All Models (82.8%)

Conclusions

In this study, I was successfully able to predict and learn the different factors that could affect the survival probability of a shipwreck. At first glance, factors like sex or age can be thought of as very intuitive, but their degree of impact was proved to be extremely high. Apart from that, other factors that were discussed in the study also influenced the survival probability, proving that data can seriously help us improving current technology and their uses.

Thank you

Have a great day ahead.