# Sinking of Titanic

Survival Analyzing through Machine Learning

# Introduction

## Background

- Titanic hit an iceberg on April 15th in 1992;
- 1500 people lost their lives;
- It becomes the deadliest commercial peacetime maritime disaster in modern history
- some groups with typical features were more likely to survive

## Purpose

- Identify the key factors which influence the survival rate mostly ;
- Using three Machine learning models (SVM, Decision Tree and Naïve Bayes) to find the best prediction model.

# Data Dictionary

| Variable | Definition | Key |
|---|---|---|
| Survival | Survival | 0=No; 1=Yes |
| Pclass | Ticket Class | 1=1st (Upper); 2=2nd (Middle); 3=3rd (Lower) |
| Sex | Sex | |
| Age | Age in years | |
| Sibsp | # of siblings / spouses aboard the Titanic | Sibling = brother, sister, stepbrother, stepsister<br>Spouse = husband, wife (mistresses and fiancés were ignored) |
| Parch | # of parents / children aboard the Titanic | Parent = mother, father<br>Child = daughter, son, stepdaughter, stepson<br>Some children travelled only with a nanny, therefore parch=0 for them. |
| Ticket | Ticket Number | |
| Fare | Passenger fare | |
| Cabin | Cabin Number | |
| Embarked | Port of Embarkation | C= Cherbourg, Q= Queenstown, S=Southampton |

# Some Findings

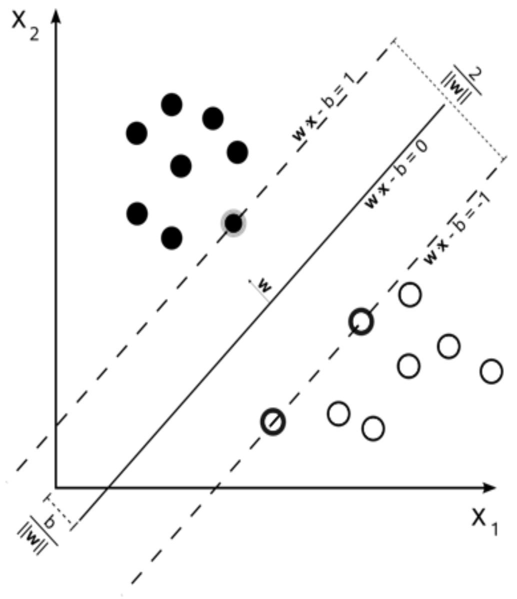
There are a total of 891 passengers in our training set.

The Age feature is missing approximately 19.8% of its values.

The Cabin feature is missing approximately 77.1% of its values.

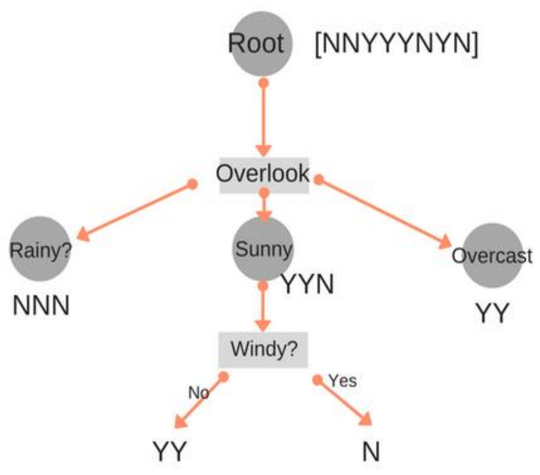
The Embarked feature is missing 0.22% of its values, which is harmless.

# Machine Learning Network



$x_2$
$x_1$
$\mathbf{w}\cdot\mathbf{x} - b = 1$
$\mathbf{w}\cdot\mathbf{x} - b = 0$
$\mathbf{w}\cdot\mathbf{x} - b = -1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

Support Vector Machine



Decision Tree

$$P(C(Class)|X(Features)) = \frac{P(X|C) \times P(C)}{P(X)}$$

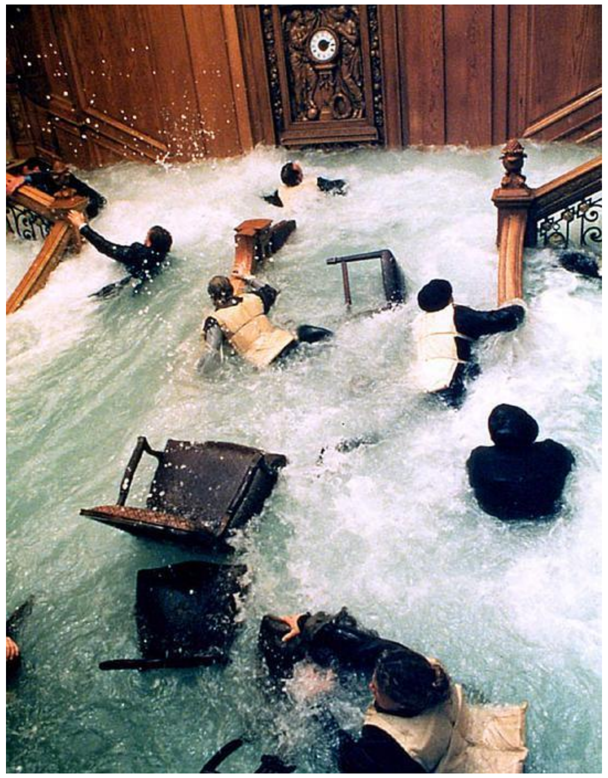
Naïve Bayes

# Data Exploring

Pclass ›

Sex ›

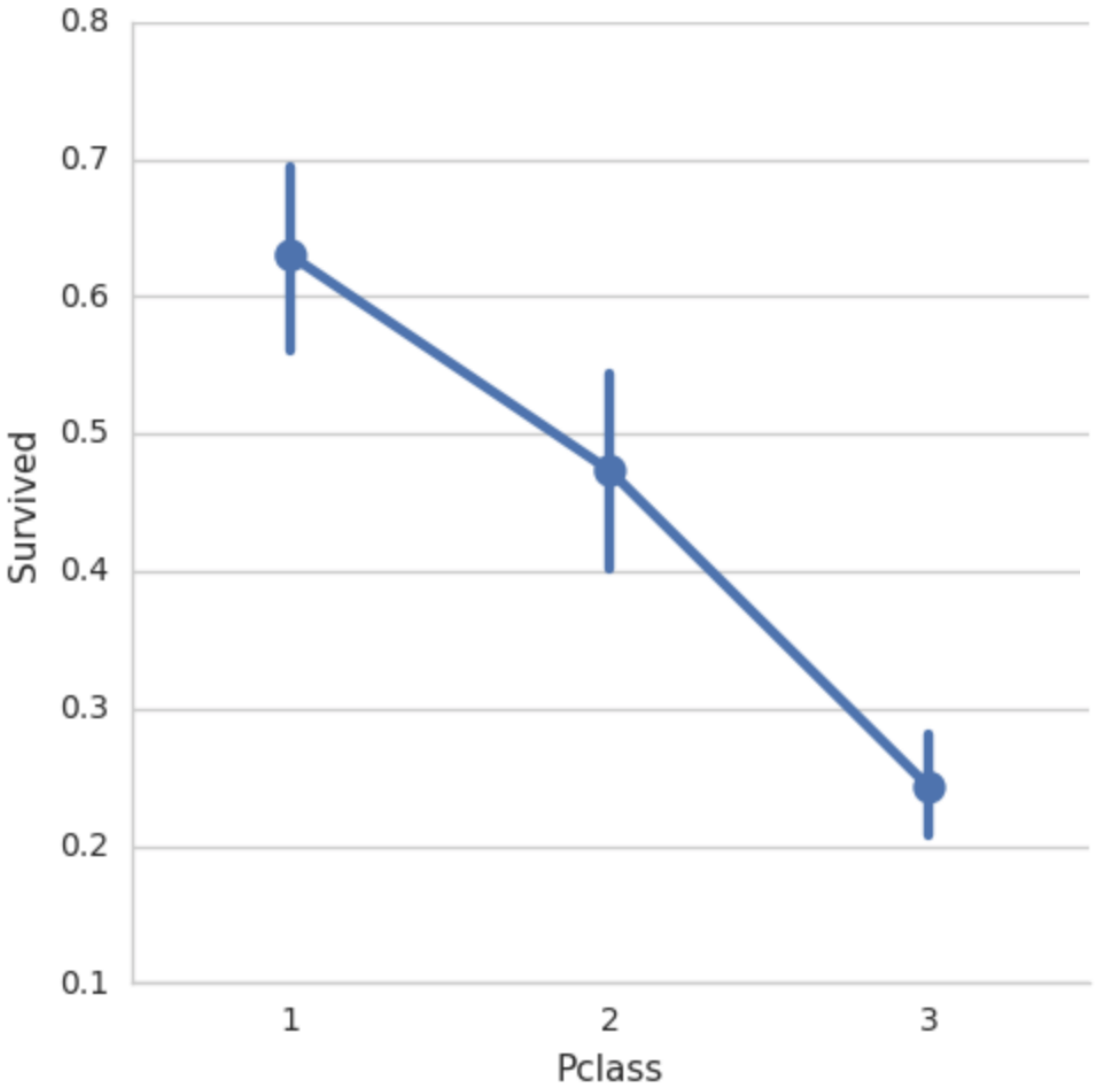
Sibsp & Parch ›



‹ Embarked

‹ Fare

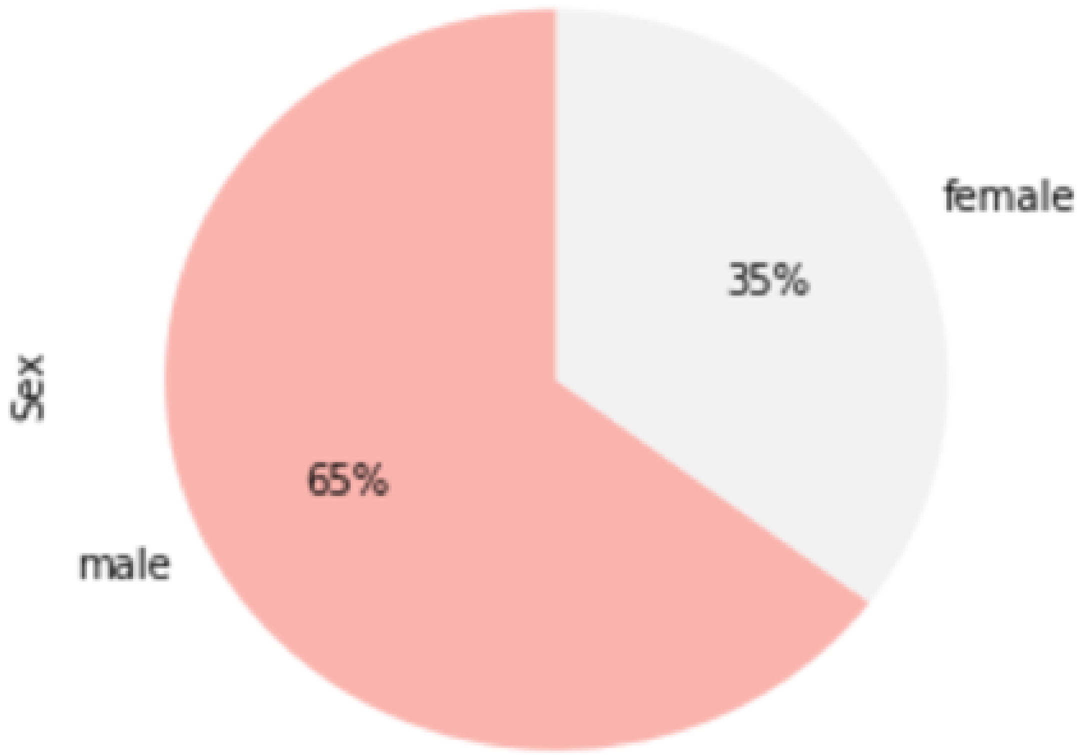
‹ Age



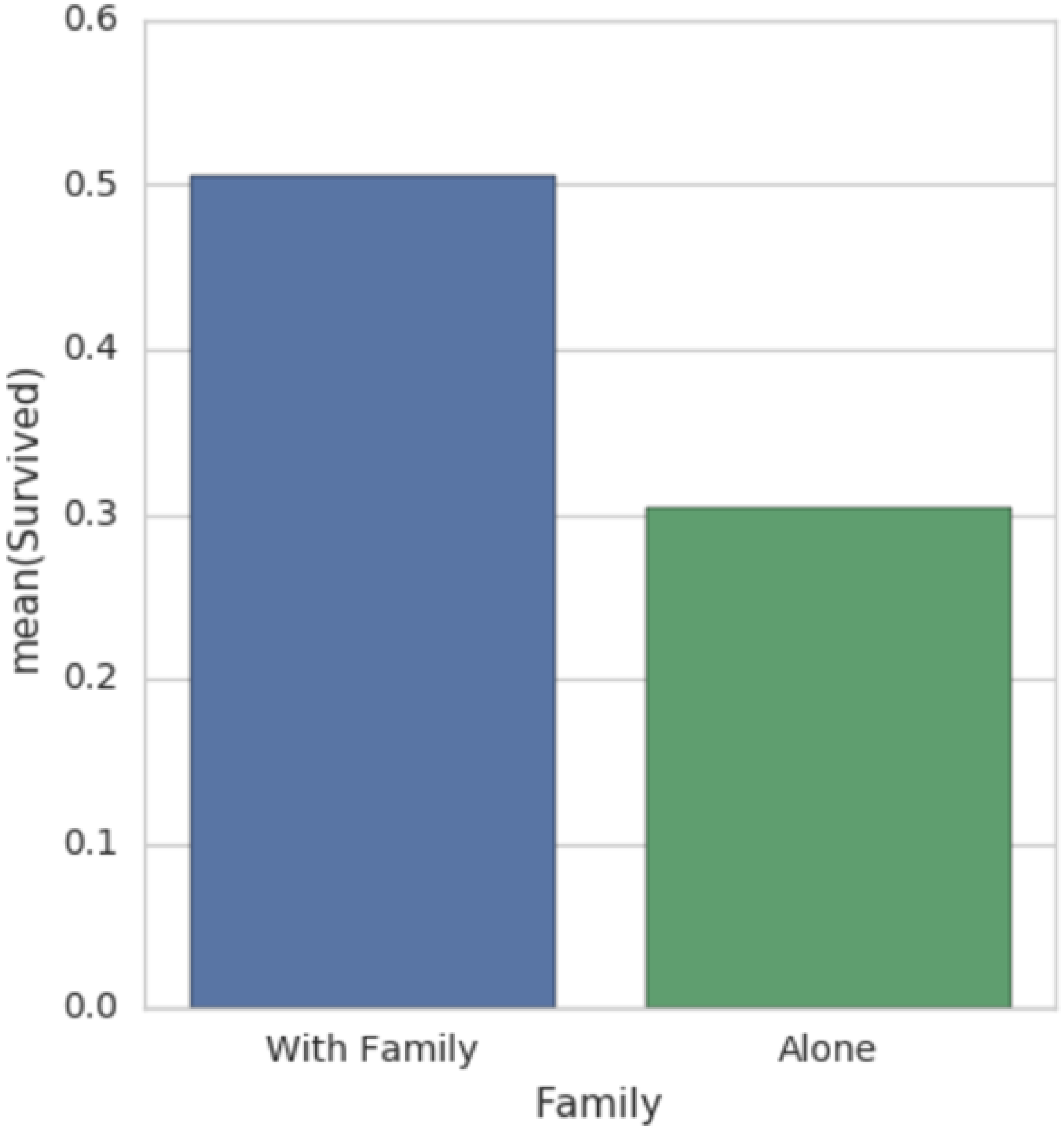
# Pclass

The upper level class customers win higher survival rates in the severe shipwreck.

# Sex

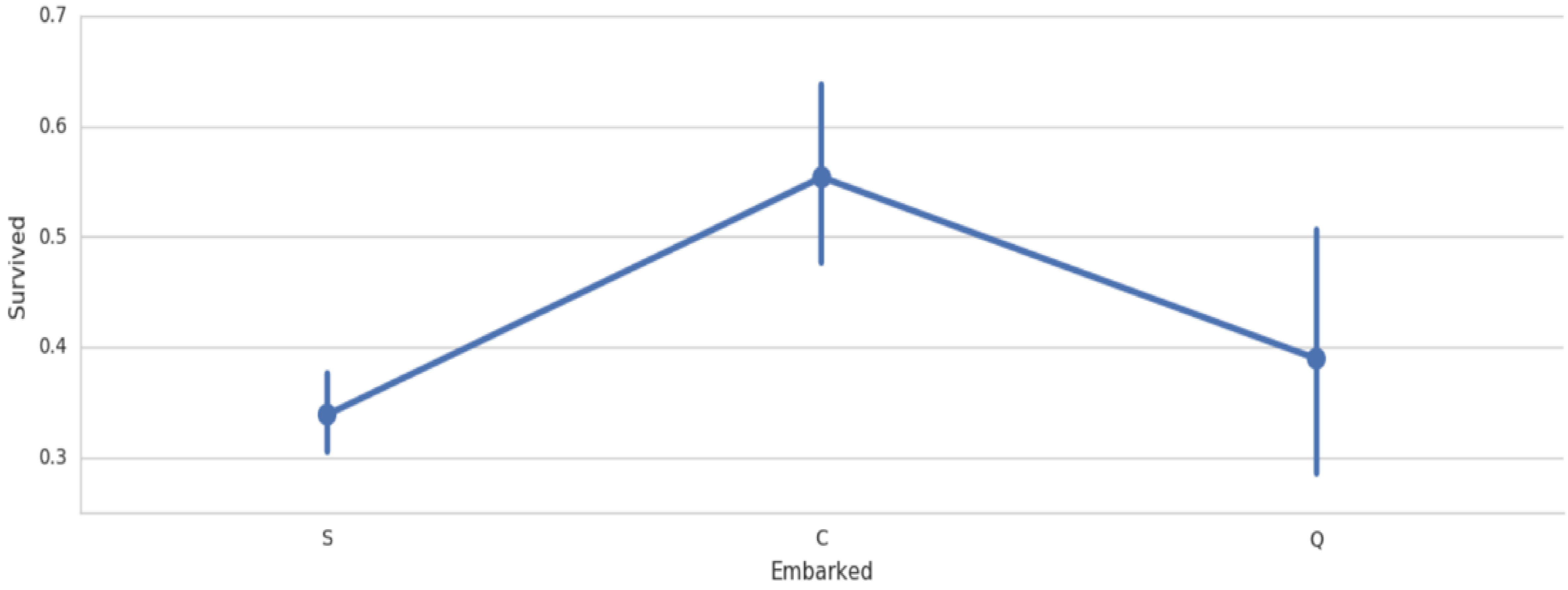


Female has a high probability to survive in the disaster. (The percent represents the death rate)



# Sibsp and Parch

Combing the variables of Sibsp and Parch Survival rate is higher if the passengers are on board with family.



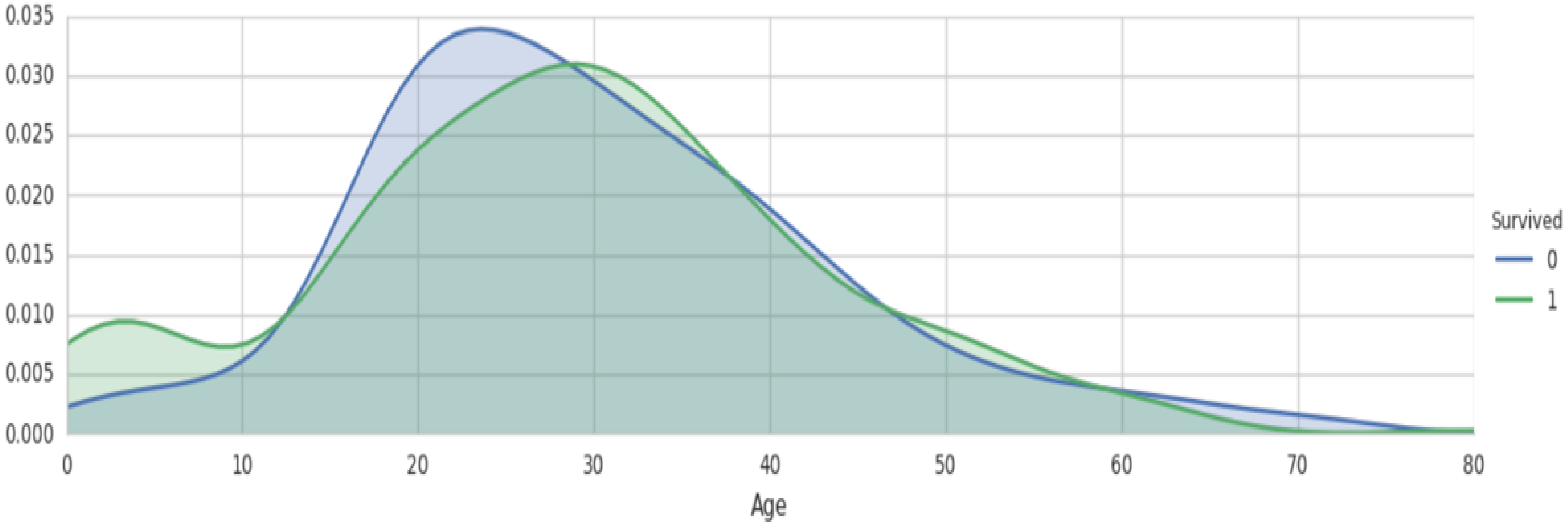
- Embarked

passengers embarked in Cherbourg are more likely to survive.

# Fare

| Fare | Survive |
|---|---|
| [0, 7.91] | 0.197309 |
| (7.91, 14.454] | 0.303571 |
| (14.454, 31] | 0.454955 |
| (31, 512.329] | 0.581081 |

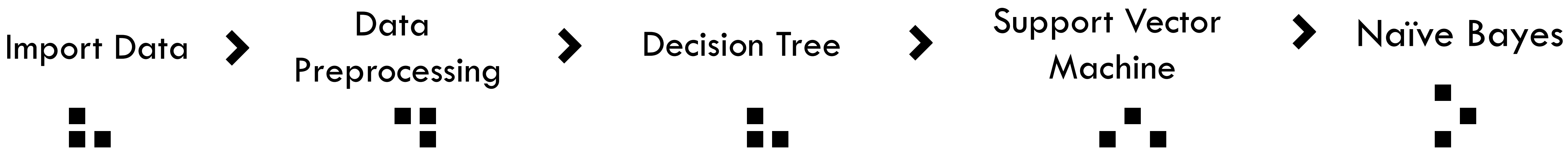
Divide the fare into 4 groups
Passengers who bought higher fare has a lightly higher possibility to survive.



- Age

Young people get lots of chance to survive comparing with the child and old people.

# Experiment setup



Import Data

Data Preprocessing

Decision Tree

Support Vector Machine

Naïve Bayes

# Results

# Decision Tree

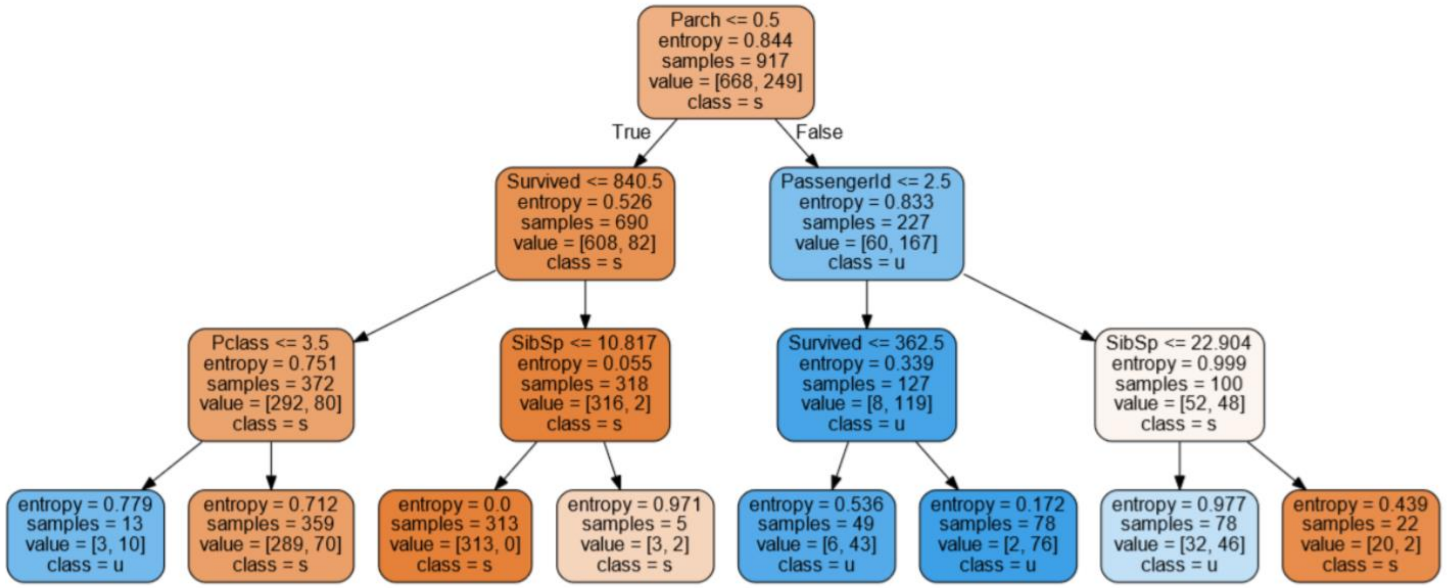


Figure 1 Decision Tree-Entropy

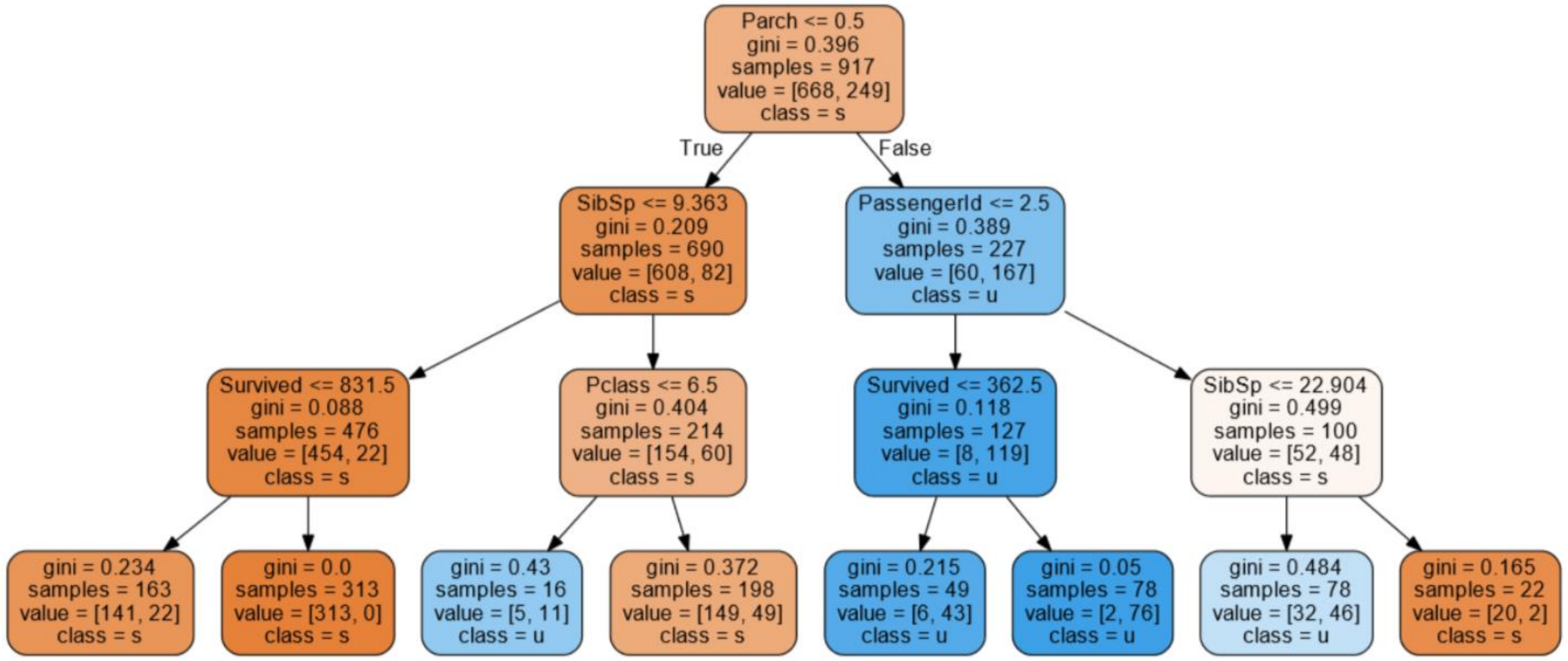


Figure 2 Decision Tree-GINI

Both of these two calculations provide quite similar information and would not have much differences in the results.

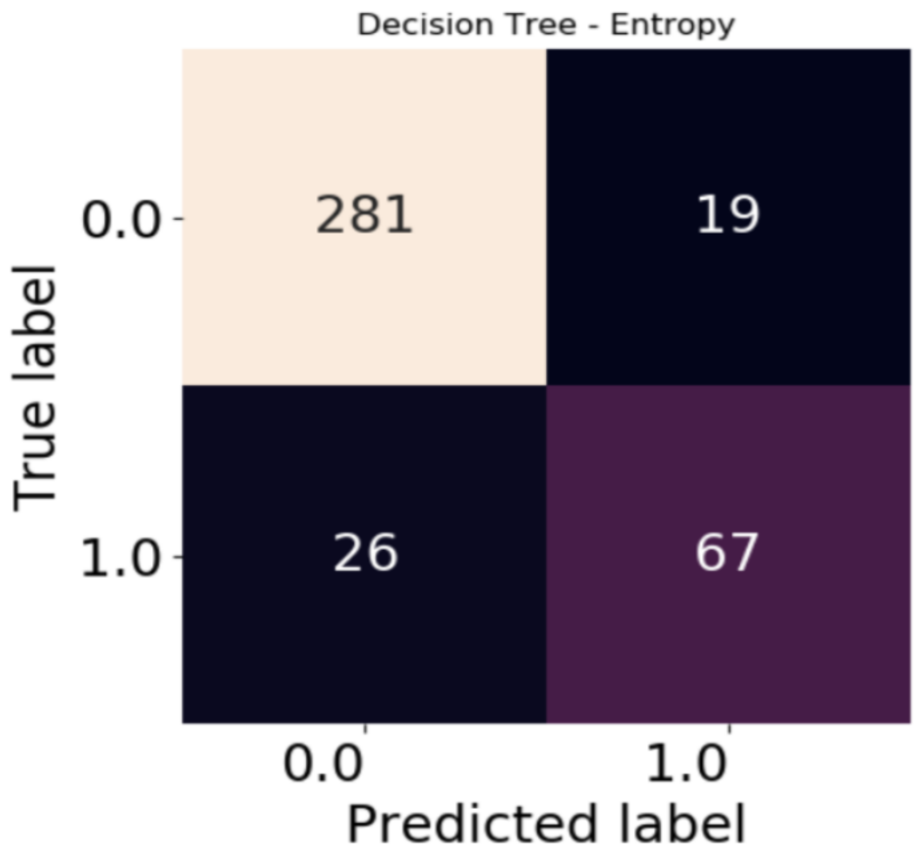


Figure 3 Entropy

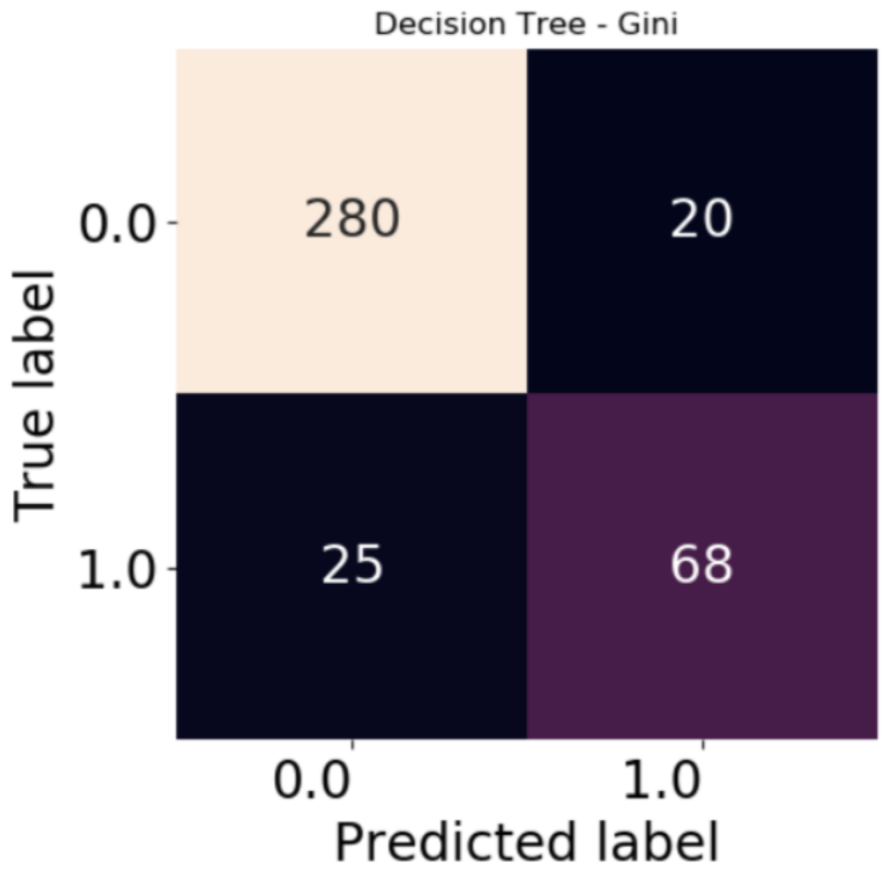


Figure 4 GINI

# Decision Tree

The performance of prediction with this two method, they were also the same (The prediction accuracy of both GINI and Entropy are 88.55% ).

## Difference:

1. Entropy mostly used on categorical data set (discrete data sets).
2. Gini mostly used on contagious data set.
3. Entropy is slower to compute, while Gini is faster.

# Support Vector Machine

AUC=0.9 which means SVM is a good classifier in this case.
(The AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example. )
ROC Curve were used in this part is because when the distribution of positive and negative samples in the test set changes, the ROC curves can remain unchanged.

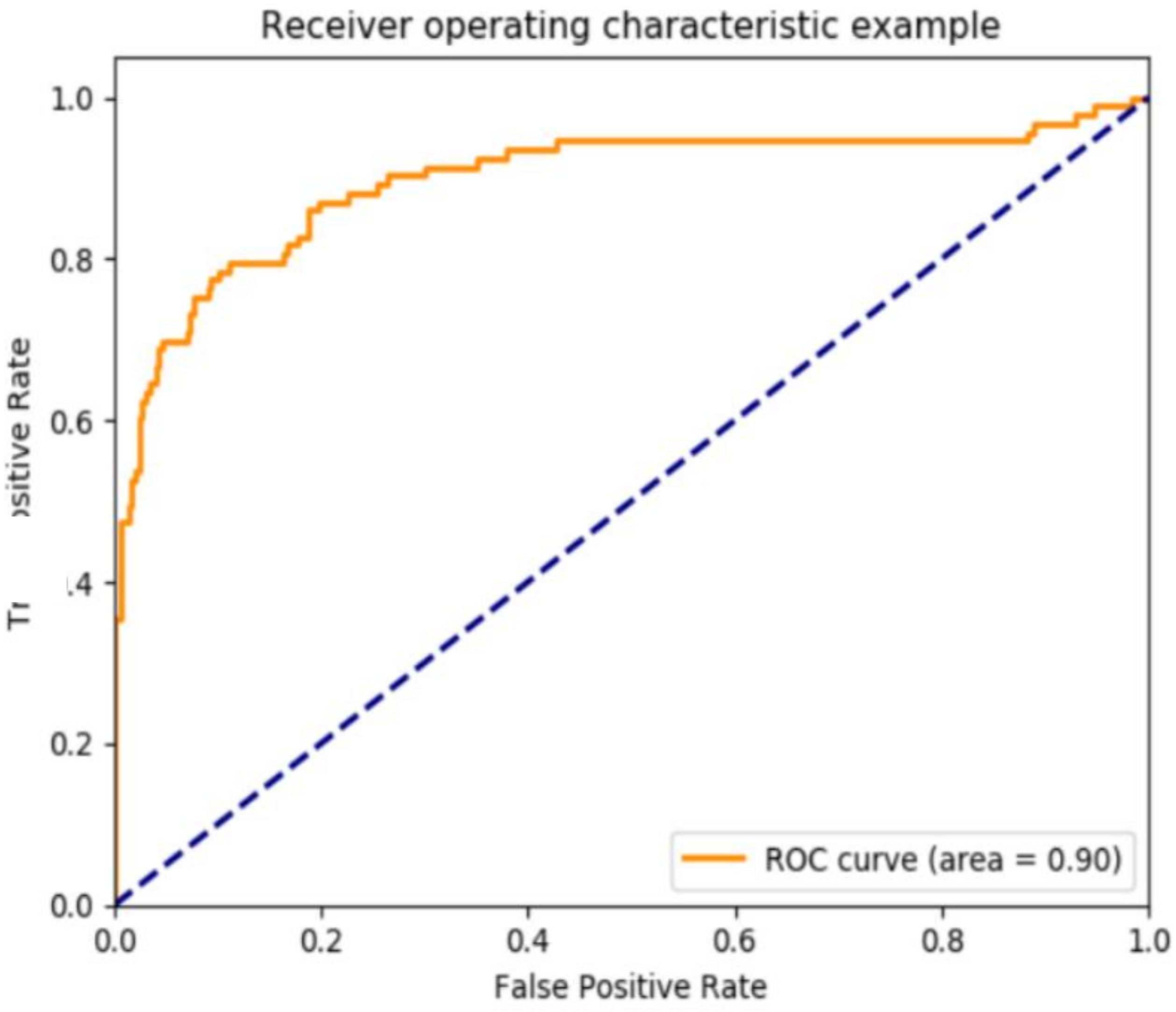


Figure 5 ROC Curve

# Support Vector Machine

The prediction accuracy of SVM is the same as Decision Tree (Entropy), which is about 87.78% Therefore, we may draw the conclusion that there is no significant difference between the performances of Decision Tree and SVM algorism in predicting the possibility of survival in shipwreck (Titanic case).

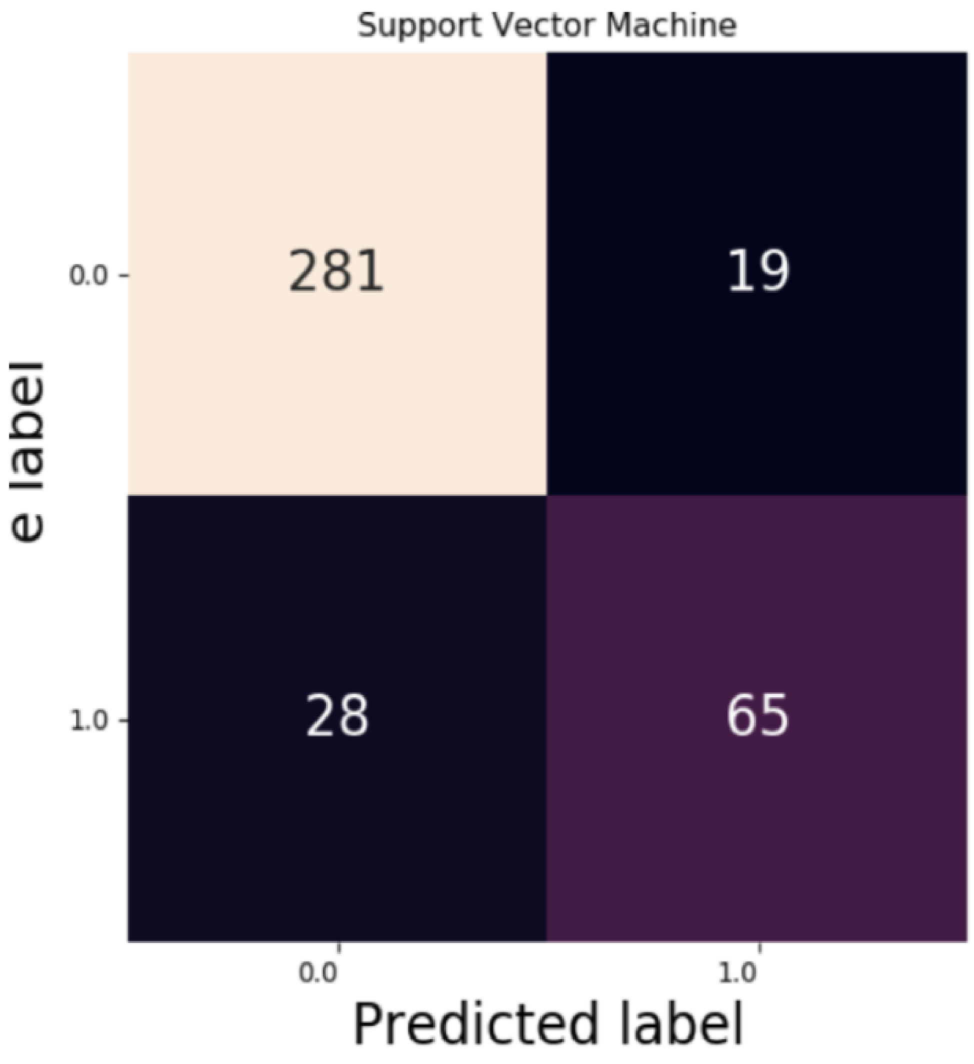


Figure 6 SVM

# Naïve Bayes

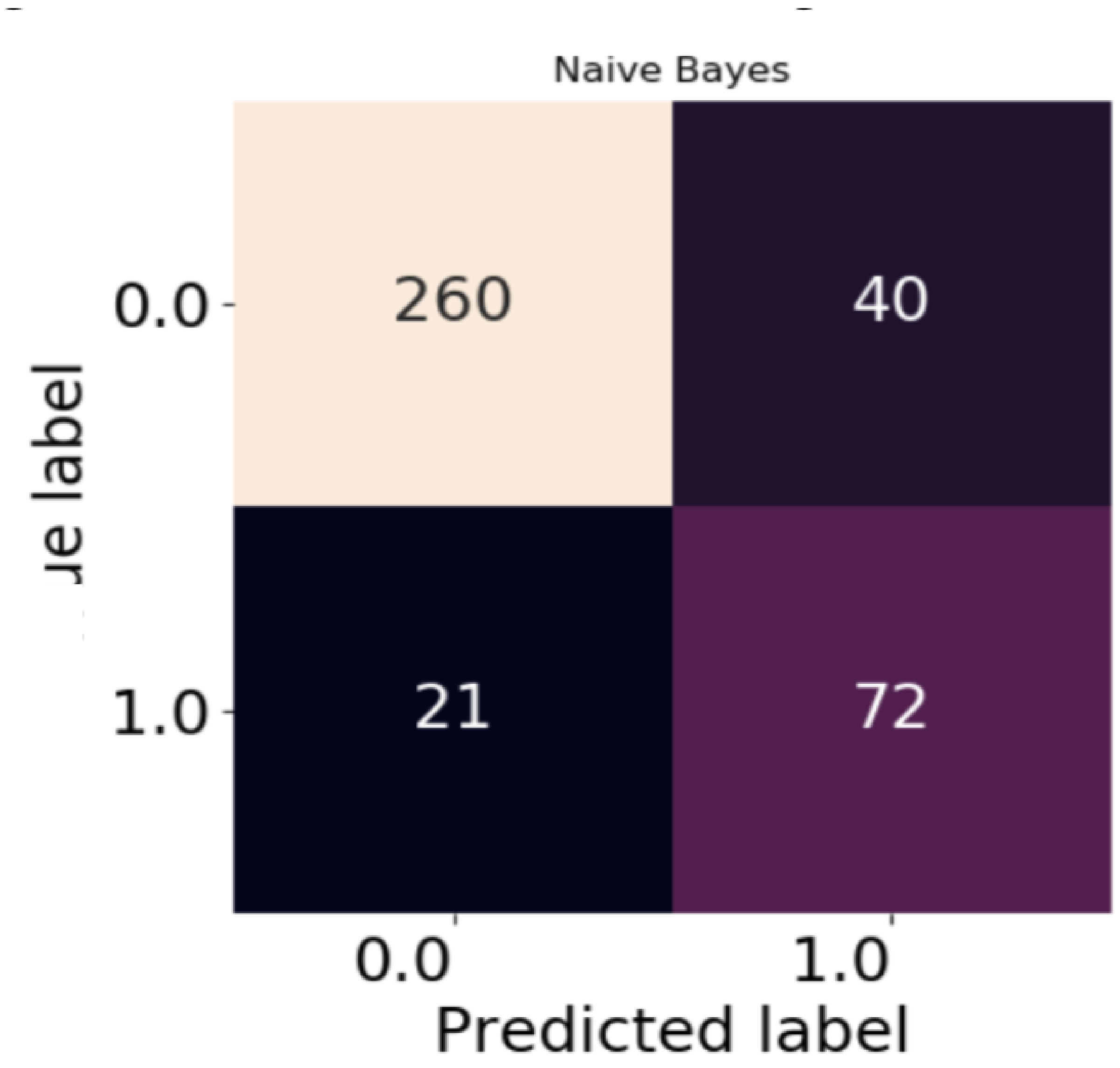


Figure 7 Naïve Bayes

The false positive rate is much higher than the other two algorisms (accuracy of prediction is only 84.47%), which means in real life of future, rescue workers would spend much time for recuing trapped person who may has died in the shipwreck, and it would absolutely waste much rescuing time and money if researchers doing such prediction by using Naive Bayes algorism.

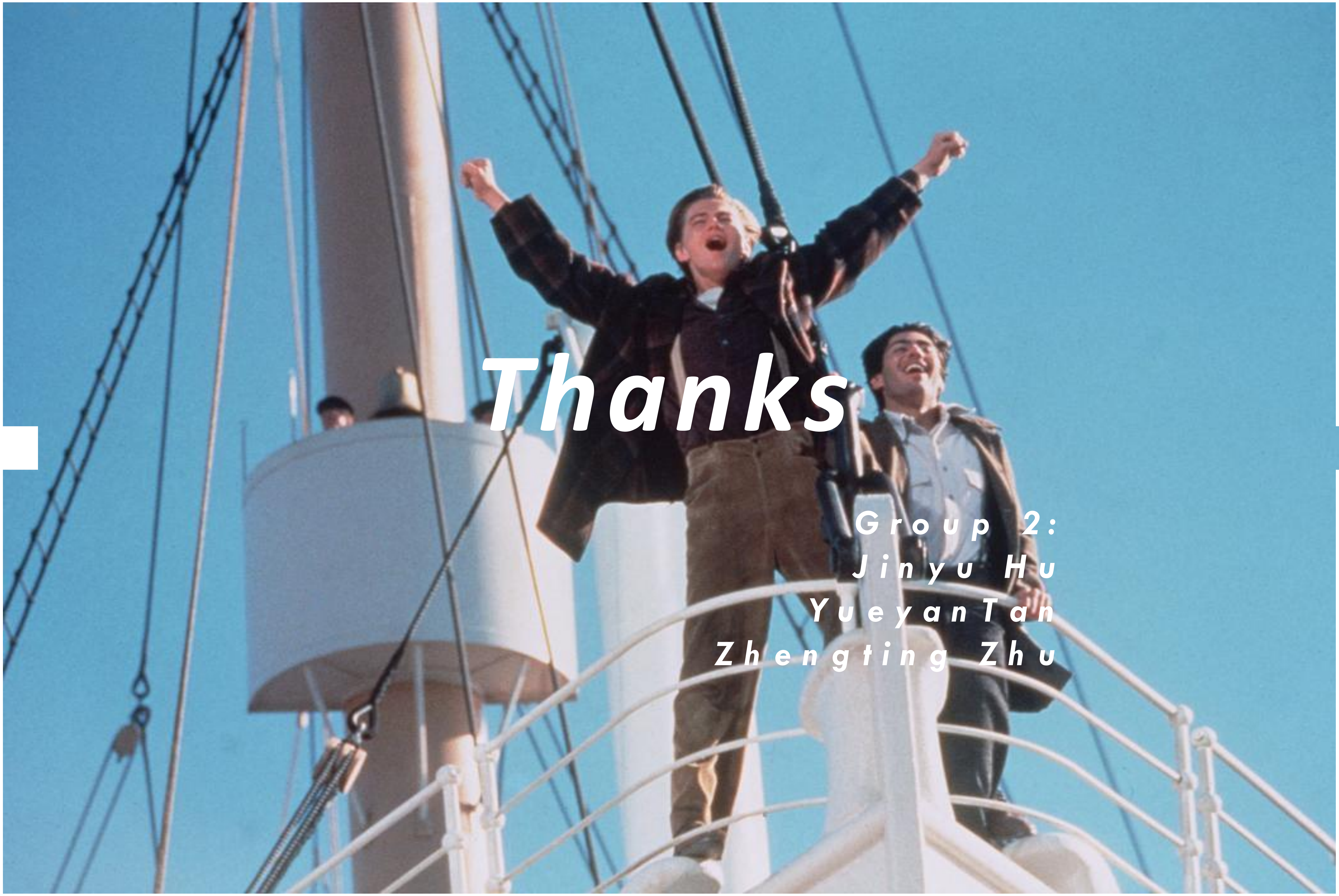
Taking all these above into consideration, we may draw the conclusion that although Naive Bayes has lots of advantages, it is not a appropriate algorism for prediction in this case.

# Conclusions

- Decision tree and support vector machine are more accurate than naive bayes based on metrics accuracy report.

- The most important features that impact passengers’ survival rates are parch(number of parents/children aboard the Titanic), sibsp(number of siblings/spouses aboard the Titanic) and pclass(ticket class).

- Young female passenger who in the upper level class and traveling with family is most likely to survive.



# Thanks

Group 2:

Jinyu Hu

YueyanTan

Zhengting Zhu