# PROJECT NAME : TITANIC

--------------------------------------------------------------------

**1. Problem Statement :**

The Titanic Problem is based on the sinking of the ‘Unsinkable’ ship Titanic in early 1912. It gives you information about multiple people like their ages, sexes, sibling counts, embarkment points, and whether or not they survived the disaster. Based on these features, you have to predict if an arbitrary passenger on Titanic would survive the sinking or not.

**2. Data Analysis :**

Data analysis involves manipulating, transforming, and visualizing data in order to infer meaningful insights from the results. Individuals, businesses,and even governments often take direction based on these insights.

Data analysts might predict customer behavior, stock prices, or insurance claims by using basic linear regression. They might create homogeneous clusters using classification and regression trees (CART), or they might gain some impact insight by using graphs to visualize a financial technology company’s portfolio.

**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)

Adult\_male: A male 18 or older (0 = No, 1=Yes)

Deck: Deck of the ship

Who: man (18+), woman (18+), child (<18)

Alive: Yes, no

Embarked\_town: Port of embarkation ( Cherbourg, Queenstown, Southampton)

Class: Passenger class (1st; 2nd; 3rd)

Alone: 1= alone, 0= not alone ( you have at least 1 sibling, spouse, parent or child on board)

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.

**3. EDA Concluding Remark.**

Roughly 20 percent of the Age data is missing. The proportion of Age missing is likely small enough for reasonable replacement with some form of imputation. Looking at the Cabin column, it looks like we are just missing too much of that data to do something useful with at a basic level. We’ll probably drop this later, or change it to another feature like “Cabin Known: 1 or 0”

**4. Pre-Processing Pipeline.**

In order to prepare our data for training in our Naïve ﻿ Bayes classifier, we remove or replace blank values and ﻿ determine bin sizes for each feature. For instance, fare can ﻿ range a large number of values, so we group fares ﻿ together. The same is done for cabin data since the data is ﻿ divided into cabin sections (A,B,C,D). A similar grouping ﻿ is done to the ticket data. ﻿ For our SVM, we do not need to bin values together. ﻿ Instead we simply turn all the values into numerical ﻿ values. In order to do this, we’ll interpret the bit ﻿ representation of strings and characters as float ﻿ represented numbers.

We see the breakdown of the data to get a better sense of what features might be good indicators of our classification problem. First, we notice that out of all the passengers in the test data, 36.38% survived. If we breakdown the group into sex, we see that a significant difference in survival between Females (74.20%) and males (18.89%). This is a strong Indicator that sex would probably be a good feature to use. Continuing our analysis, we see that of the females in first Class and second class (first class can be thought of as Upper class, second class as middle class, and third class as Lower class), more than 90% survived. Third class fared much worse with a 50% survival rate. Of the males, first Class had a much higher survival rate (36.89%) than Second (15.74%) or third (13.54%). Interestingly, there was not significant variation in survival given a person’s Age in any subgroup except for youths in second class.

Approach/Method basic Naïve Bayes classification in [1] is used as a baseline to see what is achievable. More sophisticated techniques like SVM in [2] and decision tree analysis is to see if improvements can be made in the Classification test. We experimented with using different Feature sets of each method and found the optimal feature combination on the test group.

For the Naïve Bayes model, e considered the following Features: 1) sex, 2) passenger class, 3) age, and 4) fare. We chose to use the multinomial event model and Laplace smoothing. First we had to find P(died) and P(survived) by tallying up the number of passengers that survived and

Dividing by the total number of samples. Both gender and

Passenger class took on discrete values. Gender has two values, male and female, and passenger class has three

Values, 1st, 2nd, and 3rd class. Next, we had to estimate the conditional probabilities of these features given whether a passenger survived. The same can be done for other features and values.

Before computing the parameter estimates for the features age and fare, they first need to be discretized. For age, we grouped the ages into buckets of size 5. We used a default value of -1 for samples for which age information was not provided. Similarly, for fare, we grouped the fare prices into buckets of size 20 and used -1 for samples with no fare information. After discretizing age and fare, we found the estimate for conditional probabilities of those features given whether a passenger died or survived the same way as before. This gives us a conditional probability distribution based on survival.

Using this conditional probability distribution, we can compute the probability that a test point survives given the feature set .

We multiply the probability of each feature given a negative outcome and compare that with the probability of each feature given a positive outcome. We make a prediction based on which probability is greater.

Basic Naïve Bayes classification in [1] is used as a Baseline to see what is achievable. More sophisticated Techniques like SVM in and decision tree analysis is to see if improvements can be made in the classification test. We experimented with using different feature sets of each method and found the optimal feature combination on the test group.

Unlike Naïve Bayes, no extra data cleaning was needed. Iterating through all possible feature combinations, we were able to achieve an accuracy rate of 77.99% on the test data set using only three features. The three features that achieve this rate were class, sex and place of embarkation. Using age, fare, and place of embarkation resulted in the worst accuracy of 58.13%. It is interesting to note that this accuracy would be less if we had just guessed that all test points died (accuracy of 63.23%).

This suggests that perhaps class and sex are strong indicators of survival whereas age and fare are weaker indicators of survival. In figure 2, we see the SVM learning curve using the features class, sex and place of embarkation. At around 400 samples, the training curve has reached its asymptotic value of 77.99% and any additional sample does not improve the accuracy.

We built our decision using the following features –

gender, passenger class, age, and fare. We first split the data into males and females because it was most correlated with the chance of survival. From just using a single feature, we achieved an accuracy of 76.79%, which is the same percentage as Naïve Bayes with just the gender feature. This is expected since with just the gender feature, both classifiers are labeling test samples the same way (which is marking all females as survived and all males as died). Then we split both males and females into passenger classes. Even after splitting the data into passenger class, males in each class are more likely to die, and passengers in each class, other than class 3, are more likely to survive.

If we choose the hard decision that female passengers in class 3 all survive, it will still produce an accuracy of 76.79% because the classifier hasn’t changed from the earlier process of labeling all males as died and females as survived. However, if we choose the hard decision that all females in class 3 will die, our accuracy improves to 77.27% on the test data.

**5. Building Machine Learning Models.**

Next, we look at the feature age. Since the domain of age is continuous, we have to find a good decision boundary to split our data. After plotting the age and survival of passengers in each gender and passenger class, we decided to use a binary decision because in most cases, older passengers were more likely to die than younger ones. Instead of using the same age boundary for each gender and passenger class, we considered each gender and passenger class, case by case and found different boundary thresholds for each. To find our boundary threshold, we tried to minimize the classification error on our training set. This means that we chose the age boundary for each gender and passenger class such that if we classify all samples below the age boundary as survived and all above as died, we minimize the classification error on the training set. After including age in the decision tree, we achieve a classification error of 78.94%. Of the three methods, Naïve Bayes performed the worst and decision tree performed the best. However, the best and worst performance only differs by 2.64% so all the methods have roughly the same performance on our data set. This is probably because there was one feature that was strongly correlated with whether a passenger survives. The Naïve Bayes model assumes that all features are independent but the decision tree does not make this Assumption. Even though the decision tree considers correlation between features, it only performs marginally

Better than Naïve Bayes. So this shows that assuming that features are independent is not necessarily a bad assumption for our problem. It offers a summary of the achievable accuracy using Naïve Bayes, SVM, and Decision tree analysis.

Even though we were given many features of passengers in our data, we found that most of the features were not useful in classification. For example, the number o sibling/spouses and the number of parents/children did not help with classification in any of the three models.

Knowing the number of relatives aboard did not help with cassification, but perhaps, if we were given the links between passengers then we’d be able to infer more about the survival rate. Since family units tend to all die or all survive, knowing the family links would have been useful.

**6. Concluding Remarks.**

There were not significant differences in accuracy between the three methods we experimented with. Even using every combination of features, we were still not able to produce an accuracy rate that was much different than simple Naïve Bayes classifier using only sex as a feature.

It appears that the other features were only weakly indicative of survival, as sex seemed to dominate the others in terms of being able to accurately predict survival. Even with more sophisticated algorithms, we were not able to achieve much improvement. This shows the

Importance of choosing important features and obtaining good data. It would be interesting to continue this analysis with other possible features or with other machine learning algorithms like random forests or other variants of Naïve Bayes.

We have also used some other models and we can see their accuracy and efficiency for the same including random forest, and others as well. After that ROC curve was used to see where the data and prediction goes, in simple words, to see if these match or not or we can say to check the accuracy.