

SPM4500: FUNDAMENTALS OF DATA ANALYTICS -
FINAL ASSIGNMENT REPORT

**Performance of various Data Analytics
Techniques on Kaggle's Problem Set 'Titanic:
Machine Learning from Disaster'**

Authors:

Nidhi SINGH
4242246
n.singh-2@student.tudelft.nl
MSc. Computer Science

K. CHAITANYA AKUNDI
4242246
k.c.akundi@student.tudelft.nl
MSc. Computer Science

List of Figures

Titanic Data Set

Problem Description

For our final assignment, we have taken up a challenge from Kaggle ‘Predict survival on the Titanic’. The dataset includes details of people who travelled on RMS Titanic which sank in 1912 killing 1502 out of 2224 passengers. The aim of the Kaggle challenge is to complete the analysis of what sorts of people were likely to survive. In order to do so, we will apply different predictive models to the dataset and will finally evaluate their performance against each other. Kaggle also supports Leaderboards which evaluate the submitted results, but since this evaluation is based on only 50% of the test data, it makes sense to do performance evaluation of all the models.

Since we are given both training and test data set, this problem’s predictive models will fall under the umbrella of Supervised Learning Algorithms. Also we have to decide whether a passenger survived or not, this makes it a classic Classification problem.

Data Exploration

Before diving deep into prediction making on test data, we will explore the dataset. We are given two sets of data, training (data containing attributes and known outcomes [survived or perished] for a subset of the passengers) and test (data containing attributes without outcomes for a subset of passengers). The given training data set has 891 observations of following 12 variables:

- PassengerId - Unique generated Id for each passenger
- Survived - Survival(0 = No; 1 = Yes)
- Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- Name - Name of the person
- sex - Sex
- Age - Age
- Sibsp - Number of Siblings/Spouses Aboard
- Parch - Number of Parents/Children Aboard
- Ticket - Ticket Number

- Fare - Passenger Fare
- Cabin - Cabin in the ship
- Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Let us start by looking at the type of these variables

```
> str(train_csv)
```

```
'data.frame':      891 obs. of  12 variables:
 $ PassengerId: int   1  2  3  4  5  6  7  8  9 10 ...
 $ Survived   : int   0  1  1  1  0  0  0  0  1  1 ...
 $ Pclass     : int   3  1  3  1  3  3  1  3  3  2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",...: 109 191 358 277 16 559 520 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num   22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int   1  1  0  1  0  0  0  3  0  1 ...
 $ Parch      : int   0  0  0  0  0  0  0  1  2  0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",...: 525 596 662 50 473 276 86 396 3 ...
 $ Fare       : num   7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "", "A10", "A14",...: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "", "C", "Q", "S": 4 2 4 4 4 3 4 4 4 2 ...
```

Here Factor refers to categorical data, since all the names are unique, we have 891 levels equal to number of observations.

```
> prop.table(table(train_csv$Survived))
```

```
      0      1
0.6161616 0.3838384
```

This shows that 61.6% of the passengers perished and only 38.3% survived. Running the same code for Sex, we find 35.2% females and 64.7% in the training data set.

```
> summary(train_csv$Age)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.
0.42  20.12   28.00   29.70   38.00
Max.    NA's
80.00    177
```

Summary results on Age shows that this variable is missing for 177 passengers and the minimum age is 0.42 or 5 months and maximum is 80, while 90% of the passengers were below 50.

Feature Engineering

Chapter 1

Prediction Models

Since the ‘Predict survival on the Titanic’ challenge is a classification problem, we will start with ANOVA and Linear Classifiers and then go further with methods like Decision trees, Random Forests and Ensembles of classifiers. In the following sections we will explore each model in detail and will also report its evaluation on Kaggle.

1.1 ANOVA