The problem at hand is predicting the survival of passengers aboard the RMS Titanic. As public knowledge, the Titanic sank on the early morning of 15 April 1912 during her maiden voyage. She was carrying a total of 2,224 people aboard, with around 885 crew members and approximately 1317 passengers. The estimated number of survivors is 705, hence there was a survival rate of about 31.7%. My application of machine learning employs random forest trees to aid in the prediction of whether a passenger survives or not. The random forest is trained via several data parameters, which is initially derived from the original datasets. These datasets, one for training purposes and one for testing, describe a list of passengers with the following parameters (contained within a .csv file):

**PassengerID**: integer; easy reference for each unique passenger  
**Survival** (for training only): integer; 1 if survived, 0 if not  
**Pclass**: integer; proxy for passenger’s socio-economic status (SES), higher represents higher status  
**Name**: string; passenger’s title, first and last name  
**Sex**: string; whether passenger is male or female  
**Age**: float; passenger’s age (if < 1, represented by fraction)  
**SibSp**: integer; number of passenger’s direct (step)siblings or spouses aboard  
**Parch**: integer; number of passenger’s (step)parents or (step)children aboard (guardians do not count)  
**Ticket**: string; passenger’s ticket id  
**Fare**: float; amount paid by passenger for ticket  
**Cabin**: string; passenger’s cabin  
**Embarked**: char; passenger’s point of embarkation (between S, Q and C)

Above is a list of 12 parameters that I used to derive the following features in order to further train the random forest to perform better:

The “Sex” parameter is mapped to “Gender” to represent the genders as Booleans, where 0 = female and 1 = male. The “FamilySize” column represents the entire size of the passenger’s family, which is a combination of “SibSp”, “Parch” and the passenger himself/herself. The “Embarked” parameter is mapped to “BoardOrder” to represent the order of boarding as integers, where 1 = S (for Southampton), 2 = C (for Cherbourg) and 3 = Q (for Queenstown).

There are also certain weights on the parameters. We assume that families will be in their cabins during the time of impact (11:40 pm), hence it matters greatly which cabin the passenger is assigned. The Titanic is separated into 7 decks, A through G (“Deck” parameter), with each deck containing a variable number of berths (rooms). G is the lowest deck, which realistically would reach the waterline first when sinking, hence given the largest penalty. (Also Deck A is closest to the lifeboats.) Also, since we know that the bow of the ship sank first when separated, we penalise the rooms (“CabinNum” parameter) located near the front, which we assume are the smaller cabin numbers. The number of cabins that the passenger has booked does not affect his/her chances of survival (I only use the first cabin assigned to passenger). Next, we assume that Titanic enforced the “women and children first” policy for saving lives, hence males and adults are penalised further. This also lends itself to saving families with children first. Finally, there is also the notion of SES, which penalises those who are of lower status.

Note that there are values missing from the “Age” and “Cabin” columns, hence I estimated these values. For the passengers’ age, I deduced the age (mean and standard deviation) of the passengers relative to each socioeconomic class and gender. Then using a positively skewed normal distribution, I mapped the data with a computed skew value to predict the ages that were left blank. For the passengers’ cabin(s), we randomly assign the deck and cabin number based on their fare, socioeconomic class and point of boarding (cabins issued based on first come first serve).