Research Question

I have chosen to work with the Titanic dataset.

The sinking of the RMS titanic is one of the most infamous shipwrecks in history. On April 15, 1912 during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the entire world and the international community.

A policy of "women and child first" on the Titanic was clearly demonstrated but appears that having a large family was also not good for chances of survival. My research question would be ***"Who are the people most likely to survive in such an accident, where the survival was just a matter of luck and what sorts of people were likely to survive?"***

Data set Description (fields):

Dataset I am using has the complete passenger information with 892 rows. The following are the parameters

- Passenger_id: Unique identification number given to the passengers.
- Survived: Boolean value if the passenger is survived (0=NO 1=YES)
- Pclass: What class the passenger is (1^{st} Class, 2^{nd} Class, 3^{rd} Class)
- Name: Passenger's name
- Sex: Passenger's sex

- Age: Passenger's age
- sibsb: Number of siblings /spouses of the passenger aboard
- parch: Number of parents/children of the passenger aboard
- Ticket: Ticket number
- Fare: Passenger fare
- Cabin: Cabin number given to the passenger
- Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S =Southampton)

Project Description and Machine Learning Algorithm:

The machine learning algorithm I will follow is Supervised Learning. A supervised learning algorithm analyzes the training data (In my case the titanic data is labeled data) and produces an inferred function, which can used for mapping. The titanic disaster was famous for saving "women and children first", so firstly I will try to figure out and take a look at the Sex and Age variables to see if any patterns are evident. I am still in the process of EDA (exploratory data analysis) which will help me explore the formal modeling and hypothesis testing task. I may also try using feature Engineering (Optional). This process attempts to create additional relevant features from the existing raw features in the data, and to increase the predictive power of the learning algorithm.

There are 891 observations (rows) in the dataset with 12 variables each. To build the model I will make Feature Engineering that could feed predictions or create additional variables. I have

also thought of using Decision trees where that data inherently supports to the model. But going forward I will also use Random Forest model which will give more accuracy when compared to decision tree on the dataset and finally predict which sort of people were more likely to survive.