**INTEL INTERNSHIP PROJECT**

**REPORT ON**

**Predictive Analysis of Survivors on Titanic Disaster**

**Submitted By**

**Akshit Negi Abhishek Lingwal Abhishek Raj**

**500045597 500045517 500045328**

**R110215016 R110215006 R110215007**

***Under the guidance of***

**Mr. Himanshu Sahu**

Assistant Professor (SS)

C:\Users\mrabha\Downloads\upesLogo.png

**Department of Virtualization,**

**School of Computer Science and Engineering,**

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**Dehradun-248007**

**June 2018- July 2018**

**INDEX**

**S.No. Contents Page No**

1. **Abstract 2**
2. **Introduction 2**
3. **Problem Statement 4**
4. **Literature Review 4**
5. **Objectives 5**
6. **Methodology 6**
7. **System Requirement 9**
8. **Pert Chart 10**
9. **References**   **10**

**ABSTRACT**

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. The RMS Titanic was the largest ship afloat at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast.

Hence, it can be easily summarized that there exists a co relation between the various aspects of a passenger and their chances of survival. It is therefore necessary to define a certain mathematical model which can convert such parameters to find a numerical co relation factor which defines near approximate correct probability as to whether a person survived. This kind of comparison is necessary for defining parameters in order to generalize the factors that affect the casualties in an accident. This would help in defining relations between disasters and various factors which affect the probability of survivors. The problem becomes further more interesting when we study the various factors which play a major role in such disasters. If we can establish certain co relation between the nature of passengers and their survival rates, it would help in defining better standards for the chances of survival.

**INTRODUCTION**

The major reasons for this high number of deaths were the lack of enough lifeboats for the people onboard and poor evacuation management. Most of the people who jumped or fell into the water either drowned or died of cold shock. The disaster shocked the world and caused widespread outrage over the lack of lifeboats and other safety measures. The unequal treatment of the three passenger classes during the evacuation also came under heavy scrutiny. Hence, we aim to study and predict, by applying machine learning, what sort of people were likely to survive the disaster.

The project aims to predict about who were the people who were likely to survive during the titanic disaster based upon machine learning algorithm. Machine learning is used because upon basic observation it is quite evident that the chances of survivor was a combination of various factors.

The available data consists of various attributes pertaining to a particular passenger. These attributes are taken into account to draw a co-relation between the passenger’s chances of survival and their attribute values. Binary classification based on Support Vector Machine algorithm is used to predict whether a person survived or not.

**Titanic Facts: Statistics Involving the Survivors and Victims**

**Number of survivors**: 705 (broken down in 492 passengers and 213 crew).

**Number of victims**: 1523

**Percentage of passengers that survived**: Approximately 31.6 percent of the passenger on the Titanic survived the disaster.

Survivors broken down into numbers:

**First Class**:  61 percent survived.

**Second Class**: 42 percent survived.

**Third Class**: 24 percent survived.

**Female passengers**: 75 percent of the survivors were female.

**Male passengers**: 20 percent of the survivors were male.

**Titanic Facts: Statistics About the Crew**

**Total number of crew who survived**: 214

**Total percentage of crew who survived**:  22 percent of the crew survived the disaster.

**Total percentage of female crew members who survived**: 77 percent

**Total percentage of male crew members who survived**: 22 percent

**Total percentage of the Navigation officers who survived**: 50 percent, (4 officers out of 8)

**Total percentage of the Engineering officers:** 0 percent (every single Engineering officers was below decks trying to keep the power on so distress signals could be sent)

The RMS Titanic sank in the early morning hours of April 15th, 1912. Though it has been almost 102 years, interest in the Titanic is just as strong now as it was back then. Many people search tirelessly for information about the disaster, and hopefully these Titanic facts will aid them in their quest to find out exactly what happened that terrible, freezing cold, April night.

Number of **entries in the test set**: **891**

Number of **entries in the test set**: **417**

The list of features upon which prediction is to be performed are:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| **Survival** | Whether the passenger survived or not | **0** = No (**Dead**), 1 = Yes (**Survived**) |
| **pclass** | A proxy for socio-economic status (SES) | 1st = Upper, 2nd = Middle,  3rd = Lower |
| **sex** | Gender of the passenger |  |
| **Age** | Age in years |  |
| **sibsp** | Number of siblings / spouses aboard the Titanic |  |
| **parch** | Number of parents / children aboard the Titanic |  |
| **ticket** | Ticket number |  |
| **fare** | Passenger fare |  |
| **cabin** | Cabin number |  |
| **embarked** | Port of Embarkation | **C** = Cherbourg, **Q** = Queenstown,  **S** = Southampton |

**PROBLEM STATEMENT**

Efficient prediction and analysis of passenger details in order to find the degree of co-relation and predict as to whether a person would survive or not in the infamous titanic disaster.

**LITERATURE REVIEW**

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems and as such, this is what we will focus on in this post. SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the image below. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set. **[1]** As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data. Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it. Hence, when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it. SVM is used for text classification tasks such as category assignment, detecting spam and sentiment analysis. It is also commonly used for image recognition challenges, performing particularly well in aspect-based recognition and colour-based classification. SVM also plays a vital role in many areas of handwritten digit recognition, such as postal automation services. **[6]**

Seaborn is a library for making statistical graphics in Python. It is built on [matplotlib](https://matplotlib.org/) and closely integrated with [pandas](https://pandas.pydata.org/) data structures. It offers:

* A dataset-oriented API for examining [relationships](https://seaborn.pydata.org/examples/scatter_bubbles.html#scatter-bubbles) between [multiple variables](https://seaborn.pydata.org/examples/faceted_lineplot.html#faceted-lineplot)
* Specialized support for using categorical variables to show [observations](https://seaborn.pydata.org/examples/jitter_stripplot.html#jitter-stripplot) or [aggregate statistics](https://seaborn.pydata.org/examples/pointplot_anova.html#pointplot-anova)
* Options for visualizing [univariate](https://seaborn.pydata.org/examples/distplot_options.html#distplot-options) or [bivariate](https://seaborn.pydata.org/examples/joint_kde.html#joint-kde) distributions and for [comparing](https://seaborn.pydata.org/examples/horizontal_boxplot.html#horizontal-boxplot) them between subsets of data
* Automatic estimation and plotting of [linear regression](https://seaborn.pydata.org/examples/anscombes_quartet.html#anscombes-quartet) models for different kinds [dependent](https://seaborn.pydata.org/examples/logistic_regression.html#logistic-regression) variables
* Convenient views onto the overall [structure](https://seaborn.pydata.org/examples/scatterplot_matrix.html#scatterplot-matrix) of complex datasets
* High-level abstractions for structuring [multi-plot grids](https://seaborn.pydata.org/examples/faceted_histogram.html#faceted-histogram) that let you easily build [complex](https://seaborn.pydata.org/examples/pair_grid_with_kde.html#pair-grid-with-kde) visualizations
* Concise control over matplotlib figure styling with several [built-in themes](https://seaborn.pydata.org/tutorial/aesthetics.html#aesthetics-tutorial)
* Tools for choosing [color palettes](https://seaborn.pydata.org/tutorial/color_palettes.html#palette-tutorial) that faithfully reveal patterns in your data

Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. **[2]**

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. **Numeric**, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. **[3]**

Pandas enables us to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R. Combined with the excellent [IPython](https://ipython.org/) toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate. Pandas does not implement significant modelling functionality outside of linear and panel regression. **[5]**

**OBJECTIVES**

To predict the probability of survival of passengers on board the titanic and to draw similar statistical co-relations between the probability of survival of a passenger and various attributes.

**METHODOLOGY**

The available data is divided into two data sets, namely:

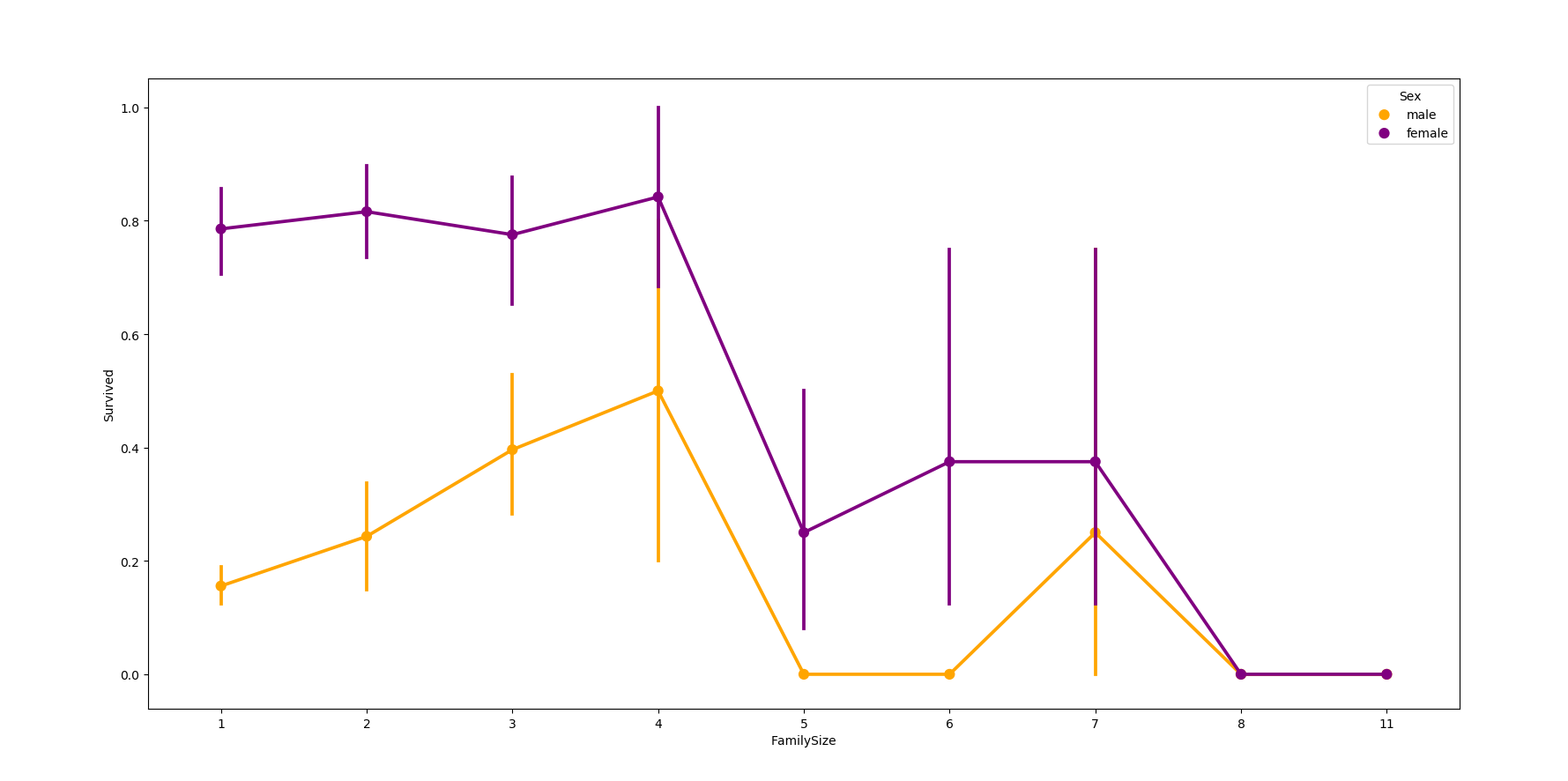
1. **Training Data Set:** The training data set is stored in the **train data frame** using the **read\_csv()** function. It is used as the data set to train the machine learning model.
2. **Test Data Set:** The test data set is stored in the **test data frame** using the **read\_csv()** function. It is used as the data set to test the machine learning model to predict if a passenger would survive or not.
3. **Feature Engineering and Data Cleaning:** A co-relation is found between the various features and the survival feature using the **groupby()** function:
   * 1. to combine feature **sibsp** (i.e. the number of siblings or spouses onboard the titanic) and **parch** (i.e. parents and number of children onboard) into a **new feature called family size** (whose minimum value is 1) which is then grouped with survival feature.
     2. To combine sex of the passenger and whether the passenger survived or not.
     3. To combine passenger class feature and survival feature.
     4. A new feature named **isAlone** is created to draw co relation between it and survival.

1. **Filling Empty Data:**

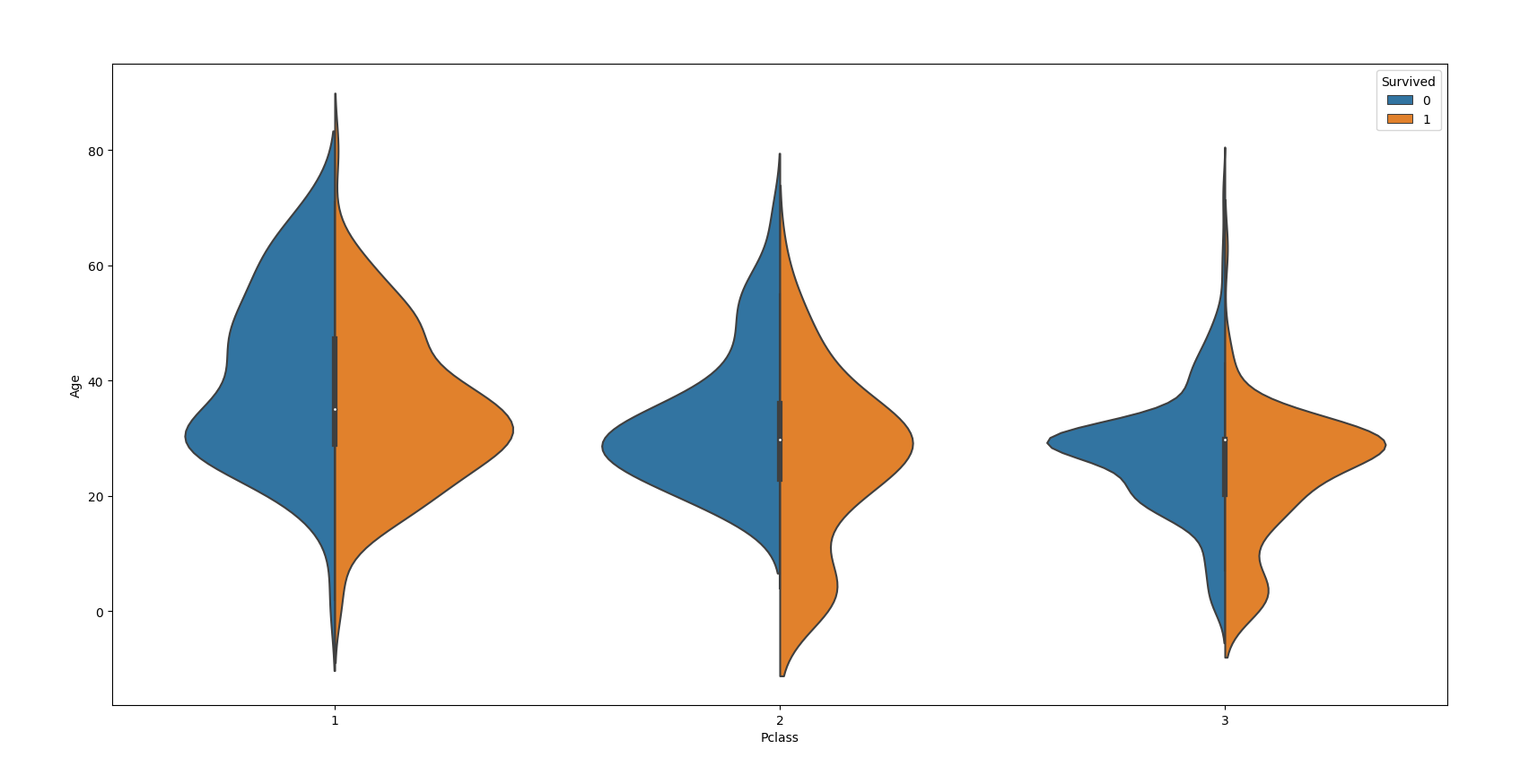
**fillna()** function is used to fill values where no value or 0 as a data entry results in ambiguity. Thus, the value of such features for that passenger is set to **NaN (Not a Number).**

1. **Data cleaning** is done to clean the title feature where **.replace()** function is used to replace redundant titles with specified categories. Then insignificant features, i.e. the features which were of no relevance and which have been combined via feature engineering to form a new feature are dropped using the **.drop()** function.
2. **Data Visualization and EDA**: The **seaborn data visualization library** is used to graphically represent the co-relation between the various features and their effect on the survival rate of the passenger. This is done via various graphical representations using **pointplot, barplot, heatmap, grid** representations.
3. **Modelling of Data:** Before modelling of data is done, the **categorical** data is **encoded** because the machine learning algorithm to be used takes numerical value as input, hence it is required to convert categorical data into numerical values. This is achieved using the **label encoder and one hot encoder**. The **label encoder .fit\_transform()** function is used to assign numerical value from **0 to number of categories-1**. Then data is divided into X and y using iloc for accessing rows and columns. Then one hot encoder is then used for binary classification and conversion of values.
4. **Feature Scaling** is then performed on the training and test data set using the **standard Scaler.**
5. Then **fitting in the SVM** model is performed.
6. **Grid Search** is applied to find the best model and the best parameters using **GridSearchCV().**
7. Finally, prediction is performed on the test set and the predicted output is converted to dataset and saved in a csv file.

**OUTPUT:**

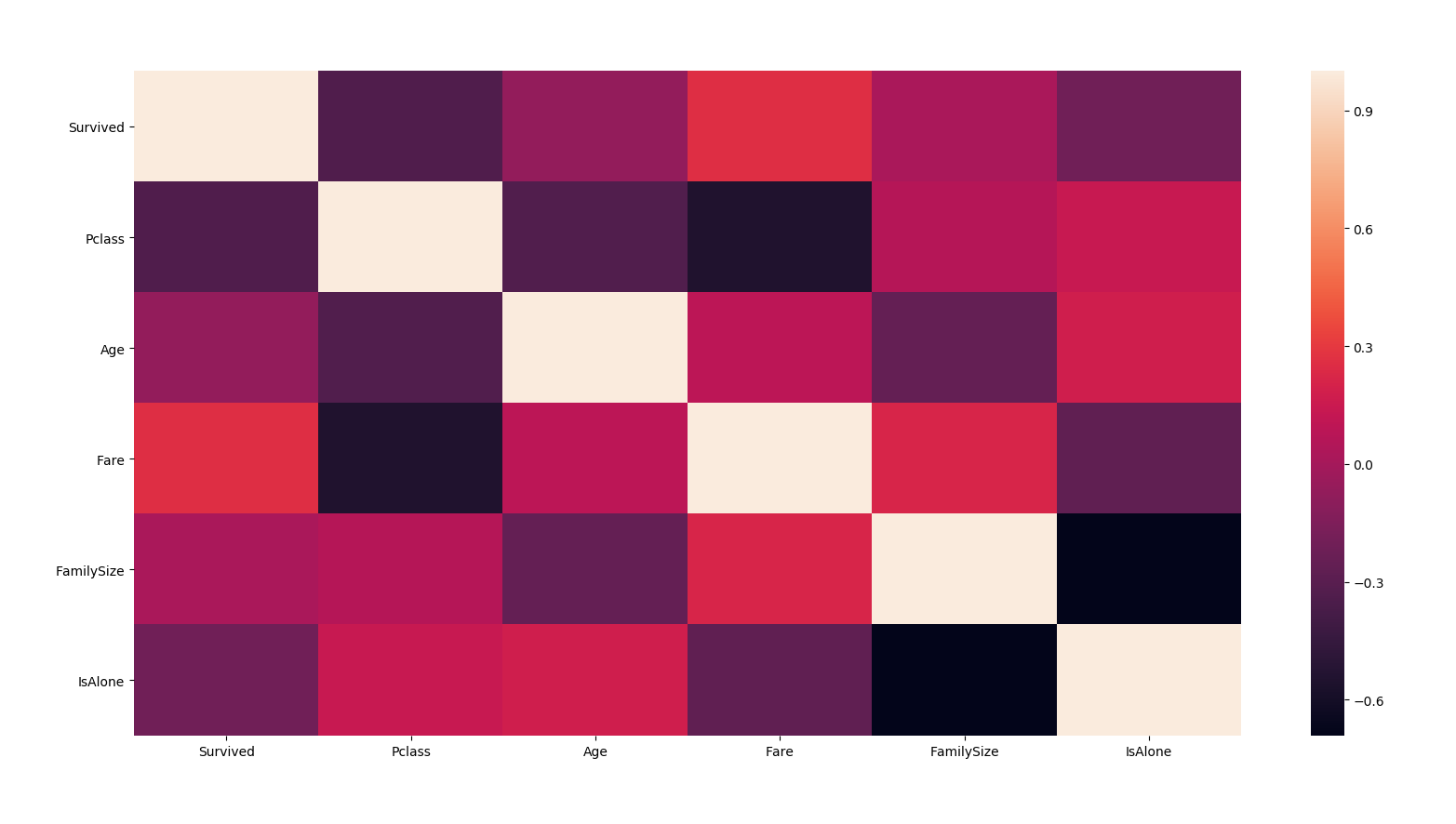
**1.** 

**Dependency of family size on survival rate**

**2. **

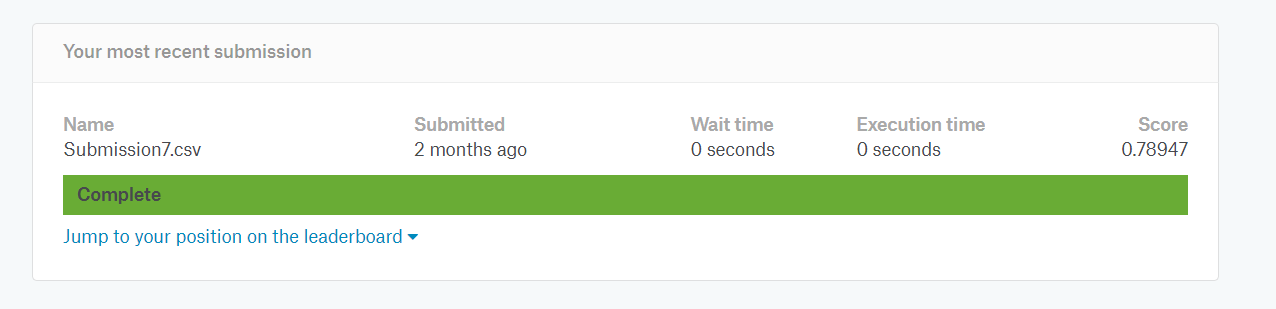
**Chances of survival as per age and pclass (0 represents death, 1 represents survival)**

**3.**

****

**Heat Map showing probability of survival based on various parameters**

**Execution Result:**

****

The data set for the passengers on board the Titanic, is used to draw relations between the various parameters pertaining to passengers and their chances of survival are predicted based on the degree of co-relation which exists between these parameters or a group of parameters and survival rate.

**SYSTEM REQUIREMENT**

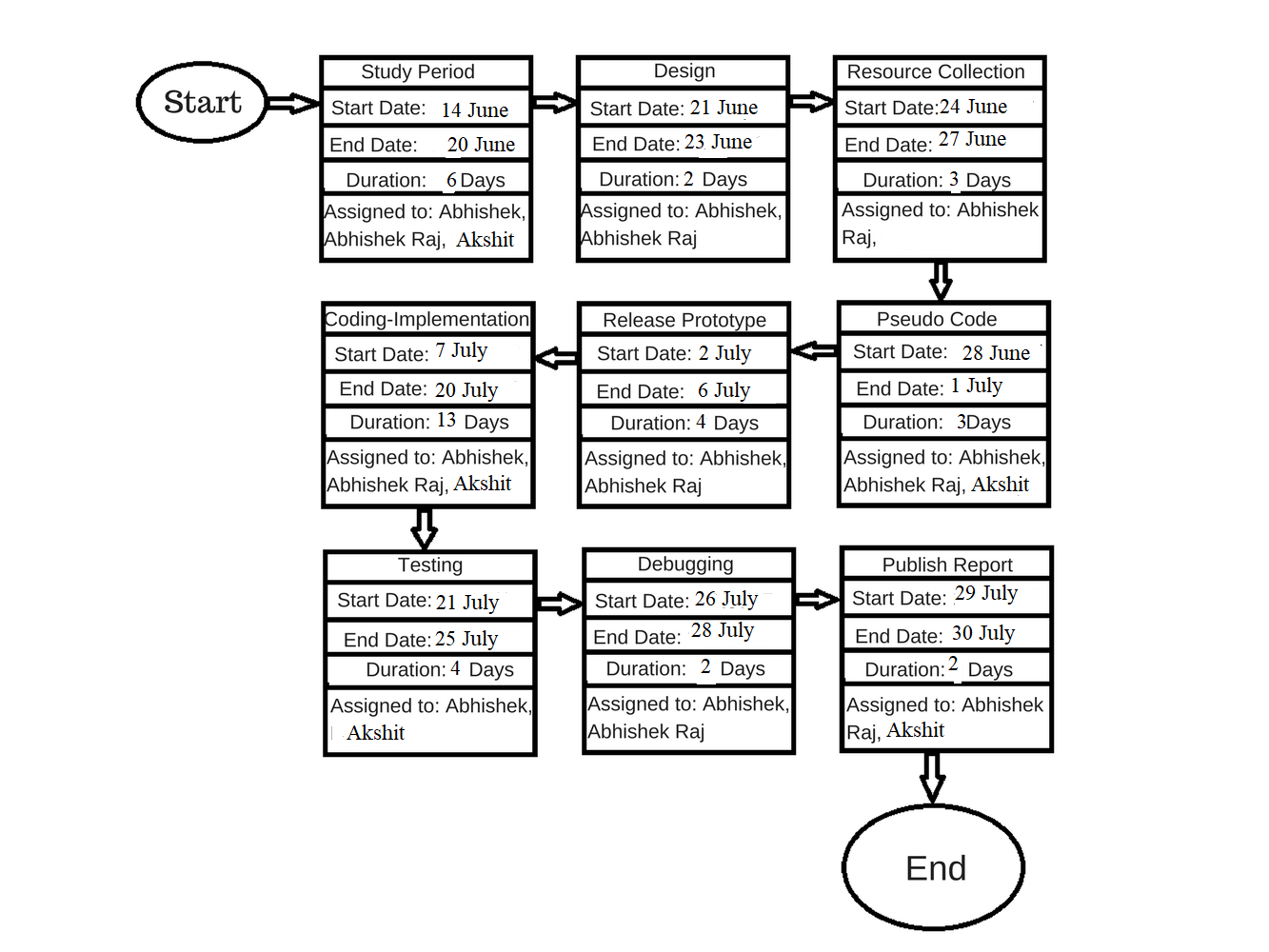
**Hardware Requirements:**

* 2 GHz x86 processor
* 8 GB of system memory (RAM)
* 1 GB of hard-drive space
* Monitor to display output
* Keyboard/Mouse for data input

**Software Requirements:**

* Python Environment (Jupyter, Pycharm IDEs)
* Numpy v1.12.0
* Pandas v0.19.2
* Scikit-learn v0.18.1
* Scipy v0.18.1
* Sklearn v0.0
* Dataset
* MS Word(Documentation)

**PERT CHART**

****

**REFERENCES**

1. [**https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python**](https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python)
2. **https://www.kaggle.com/c/titanic**
3. [**https://seaborn.pydata.org**](https://seaborn.pydata.org)
4. [**https://docs.scipy.org/doc/numpy-1.15.1**](https://docs.scipy.org/doc/numpy-1.15.1)
5. [**https://pandas.pydata.org/**](https://pandas.pydata.org/)
6. **“DATA CLASSIFICATION USING SUPPORT VECTOR** **MACHINE**”, **1.** DURGESH K. SRIVASTAVA,

**2.** LEKHA BHAMBHU[2014]

1. **http://web.cs.iastate.edu/~honavar/hearst-svm.pdf**