**Homework Assignment #3**

**(Classification – Logistic Regression, K Nearest Neighbor, Support Vector Machine, K- SVM, Naïve Bayes, Decision Tree, Random Forest)**

**CS 156 Spring 2021**

**Problem Description**

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this assignment, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e., name, age, gender, socio-economic class, etc.).

**Overview**

The dataset called Data-Hw3.csv consists of 891 entries. This dataset needs to be split into two groups using 25% data for Test set

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class.

The test set should be used to see how well your model performs on unseen data. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

The dataset for this project contains information about the passengers in the Titanic and if they survived the historic accident. There are 8 column headers:

1. passenger ID An identifier for the passenger
2. name Name of the passenger
3. sex Male or Female
4. age Age in years
5. sibsp # of siblings / spouses aboard the Titanic
6. parch # of parents / children aboard the Titanic
7. pclass Ticket class. 1 = 1st, 2 = 2nd, 3 = 3rd
8. survived 0 = “no”, 1 = “yes”

**Variable Notes**

pclass: A proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower  
  
age: Age is fractional if less than 1. If the age is estimated, 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 travelled only with a nanny, therefore parch=0 for them.

**Part (A):** Data Import, Data Pre-processing

* 1. Read the file Data-Hw3.csv
  2. Replace Missing Data to make the data set complete
  3. Divide the data set into Training set and Test set

For each model in Parts (B) through (H), use the following data set to test your model:

**Data set to test your models:**

1. [sex = male, age = 4, sibsp = 0, parch = 0, pclass = 3]
2. [sex = male, age = 4, sibsp = 4, parch = 0, pclass = 3]
3. [sex = male, age = 4, sibsp = 0, parch = 5, pclass = 3]
4. [sex = male, age = 4, sibsp = 0, parch = 0, pclass = 1]
5. [sex = male, age = 40, sibsp = 0, parch = 0, pclass = 3]
6. [sex = male, age = 40, sibsp = 4, parch = 0, pclass = 3]
7. [sex = male, age = 40, sibsp = 0, parch = 5, pclass = 3]
8. [sex = male, age = 40, sibsp = 0, parch = 0, pclass = 1]
9. [sex = female, age = 4, sibsp = 0, parch = 0, pclass = 3]
10. [sex = female, age = 4, sibsp = 4, parch = 0, pclass = 3]
11. [sex = female, age = 4, sibsp = 0, parch = 5, pclass = 3]
12. [sex = female, age = 4, sibsp = 0, parch = 0, pclass = 1]
13. [sex = female, age = 40, sibsp = 0, parch = 0, pclass = 3]
14. [sex = female, age = 40, sibsp = 4, parch = 0, pclass = 3]
15. [sex = female, age = 40, sibsp = 0, parch = 5, pclass = 3]
16. [sex = female, age = 40, sibsp = 0, parch = 0, pclass = 1]

**Part (B):** Use **Logistic Regression** to predict if a passenger in the Test set will survive the accident

* 1. Print the prediction and the corresponding ground truth in the Test set
  2. Print the Confusion Matrix
  3. Compute Accuracy
  4. Print and Tabulate the result for the data set given above

**Part (C):** Use **K Nearest Neighbor Classification** with **7 neighbors** to predict if a passenger in the Test set will survive the accident

1. Print the prediction and the corresponding ground truth in the Test set
2. Print the Confusion Matrix
3. Compute Accuracy
4. Print and Tabulate the result for the data set given above

**Part (D):** Use **Support Vector Machine (SVM) Classification** to predict if a passenger in the Test set will survive the accident

1. Print the prediction and the corresponding ground truth in the Test set
2. Print the Confusion Matrix
3. Compute Accuracy
4. Print and Tabulate the result for the data set given above

**Part (E):** Use **Kernel Support Vector Machine (K-SVM) Classification** to predict if a passenger in the Test set will survive the accident

1. Print the prediction and the corresponding ground truth in the Test set
2. Print the Confusion Matrix
3. Compute Accuracy
4. Print and Tabulate the result for the data set given above

**Part (F):** Use **Naïve Bayes Classification** to predict if a passenger in the Test set will survive the accident

1. Print the prediction and the corresponding ground truth in the Test set
2. Print the Confusion Matrix
3. Compute Accuracy
4. Print and Tabulate the result for the data set given above

**Part (G):** Use **Decision Tree Classification** to predict if a passenger in the Test set will survive the accident

1. Print the prediction and the corresponding ground truth in the Test set
2. Print the Confusion Matrix
3. Compute Accuracy
4. Print and Tabulate the result for the data set given above

**Part (H):** Use **Random Forest Classification** with **10 Decision Trees** to predict if a passenger in the Test set will survive the accident

1. Print the prediction and the corresponding ground truth in the Test set
2. Print the Confusion Matrix
3. Compute Accuracy
4. Print and Tabulate the result for the data set given above

**Summarize your observations in terms of:**

1. Tabulate the result of prediction from each of the models for the 16 dataset points

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Logical Regression | K-Nearest Neighbors | Support Vector Machines | Kernel Support Vector Machine | Naïve Bayes | Decision Tree | Random Forest |
| [0] | [1] | [0] | [1] | [0] | [1] | [1] |
| [0] | [0] | [0] | [1] | [0] | [0] | [0] |
| [0] | [1] | [0] | [1] | [0] | [1] | [1] |
| [1] | [1] | [0] | [1] | [1] | [1] | [0] |
| [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| [1] | [1] | [1] | [1] | [1] | [1] | [1] |
| [1] | [0] | [1] | [1] | [0] | [1] | [0] |
| [1] | [1] | [1] | [1] | [1] | [1] | [1] |
| [1] | [1] | [1] | [1] | [1] | [1] | [1] |
| [1] | [0] | [1] | [0] | [1] | [0] | [0] |
| [0] | [0] | [1] | [0] | [0] | [0] | [0] |
| [1] | [0] | [1] | [0] | [1] | [0] | [0] |
| [1] | [1] | [1] | [0] | [1] | [1] | [1] |
| Accuracy Scores | | | | | | |
| 0.7847533632286996 | 0.7802690582959642 | 0.7802690582959642 | 0.6591928251121076 | 0.7757847533632287 | **0.9237668161434978** | **0.9282511210762332** |

1. Which predictive models performed the best – Top 3

In boldface above

1. What could possibly make the top 3 models outperform the rest?

Decision Trees and Random Forest are non-linear, whereas logistic regression works best with binary data.