Kernel SVM on Titanic Dataset

Remya Kannan

Introduction:

The “unsinkable” RMS Titanic struck tragedy on April 14, 1912, as it crashed into an ice burg and sunk into the Atlantic Ocean leaving many of its passengers to a solemn fate. There were 20 lifeboats on the ship, however, they were not large enough to save all the passengers. 705 passengers escaped safely, however, 1500 were aboard the sinking vessel.

While this tragedy struck over a 100 years ago, researchers are still considering the various causes and survival rates of the passengers on board. Various hypotheses have been stated and tested to understand the parameters that affected this tragedy, in hopes of gaining closure on whether its human error or nature’s force. Interesting questions arise from various studies to determine where the class of the passengers affected the survival, or the gender of the passenger influenced the chances of survival, was an adult or a child more likely to survive, etc. This study aims to focus and work on similar questions with the help of data mining technique such as Kernel SVM learning method. A comparative study on linear vs radial kernel SVM methods is done on the data. Furthermore, analysis is done to determine the characteristics of passengers that affect the chance of survival.

Previous literature uses various tools and data mining techniques such as k-nn clustering, random forest, neural networks, etc. to determine the survival rate of the passengers and the underlying features to predict the classes. A kernel is a similarity function that is provided to the data mining algorithm. Many data mining algorithms are written using only dot products and then replaced by kernels. This helps the learner to not use the feature set, thus reducing computational complexity and the cost. This results in an effective and efficient method that is relatively low dimensional and low-performance and thus, can improve the results of the model. This added to the existing SVM classifier algorithm enhances the results, thus improving the accuracy of the model.

This study uses the above-mentioned data mining technique to develop a model that can predict the survival chances of a passenger with relatively good accuracy given a set of underlying conditions. Firstly, the dataset is explored, cleaned, and balanced for use in the data mining process. Then some analysis is done on the dataset to learn the attributes that contribute to the survival chances of passengers. Kernel SVM (linear vs radial kernel) is then applied and compared for evaluation. Finally, a detailed evaluation is done on the effects of Radial Kernel SVM on the model.

This report is divided as follows: section 1 briefly describes the data highlighting the relevant features of the dataset used in this study, section 2 explains how the data was munged and cleaned in preparation for model building, section 4 highlights the experimental results that discusses the performance of the models, section 5 interprets and analyses the results followed by analysis conclusion in section 6.

1. Data Description:

The Titanic dataset used in this study is obtained from Kaggle repository and describes the survival status of individual passengers on the Titanic. The dataset consists of 1309 passengers and has 12 variables. The target class attribute features the survival: 1 or non-survival: 0. The class label determines the final model prediction. The dataset is divided into training and testing dataset with the training set having 891 instances and the testing set having 418 instances. The feature set consists of 12 attributes including the class variable and are described as follows: Passenger (passenger id), Pclass (denotes the passenger class (1=1st, 2=2nd, 3=3rd)), Survived (Survival of the passenger (1=Yes, 0=No)), Name, Sex, Age, Sibsp (Number of siblings/Spouses aboard), parch (Number of parents/children aboard), Ticket (Ticket Number), Fare (Passenger fare in British pound), Cabin and Embarked (port of embarkation: C,Q,S).

2. Data preprocessing:

The data was acquired from Kaggle repository and consisted of attributes that would determine the survival of the passengers aboard the Titanic. Data partition aims to split the data into subsets of training and testing data. It is a process by which mutually exclusive data is partitioned using 10-fold cross validation. To avoid bias, the records are randomly selected for partitioning. It helps reduce the computation time during model implementation. The training set consists of 891 instances and the testing set consists of 418 instances.

In this study, balancing the dataset is not a concern as SVM performs well due to the distribution free nature of the margins generated by SVM. When the focus is on the rare case class, it is essential that balancing not be done by under-sampling or over-sampling as it might affect the outcome of the result. This is handled and taken care of when the dataset is divided into training and testing sets.

The most important process of the data preprocessing stage, is data cleaning. Data cleaning takes raw data and cleans them off the missing values and transforms it to be used by the model. It takes place over individual attributes before they are fed to the classifiers. Missing values in this dataset could indicate missing on the information of the impact of certain features on the survival rate of the passenger and this could impact the analysis. There is thus, a need for missing value imputation, that are either replaced by the mean, or 0 or the median. In this dataset, there were several missing values for attributes such as Cabin, Fare, and Age. The missing values were either replaced by 0 or the mean and median value of the respective attributes. This helps us determine the factors to a certain accuracy during model evaluation. The data was not as noisy and there were very few outliers that affected the model.

Feature selection is yet another important aspect of data preprocessing and involves reduction of the number of attributes that don't contribute to the model. For this dataset, however, the 11 features (not including the class variable), contribute directly and as a combination of features towards the model building process.

The next section of the report, considers how the data analysis and feature selection techniques were applied to this dataset.

3. **Data Analysis:**

Following the data cleaning process, the next stage of the study moves on to the exploratory analysis of the data. There is a need to find what features contribute towards the model building process. In this section, each variable is examined in turn. Thus, exploratory analysis is divided into univariate analysis and bivariate analysis.

3.1. Univariate analysis:

3.1.1. Pclass:

From historical evidence, it is said that the first-class passengers had higher survivability than third class passengers. Fig. 1 confirms the theory and hence, we may conclude that Pclass could be a useful feature for the model.

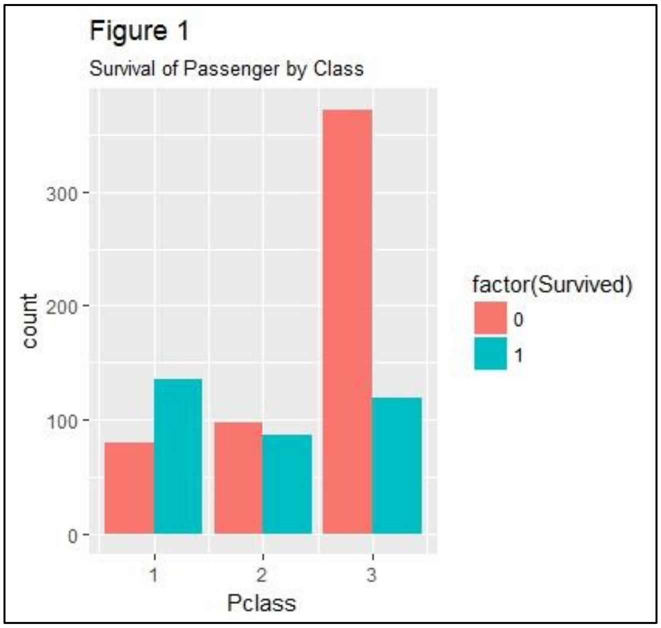


Fig.1. Survival of Passengers by class

3.1.2. Name, Sex, Age:

On taking a closer look at the data, we observe that the passenger name includes the title of the passenger, such as Mr., Mrs., Miss, and Master. This could aid in imputing the missing values of the attributes Sex and Age, thus keeping the data as precise as possible. Thus, the title is extracted from the name of the passenger and a new feature added to the dataset.

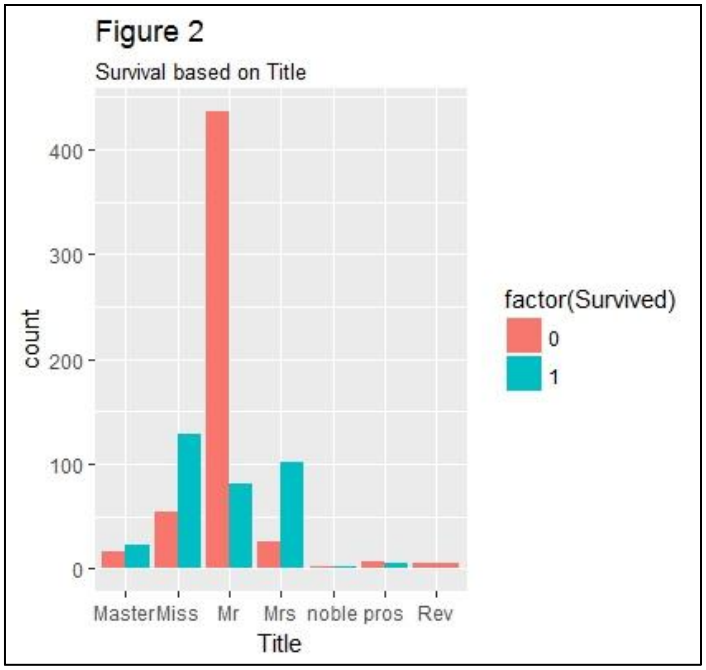


Fig.2. Survival of passengers based on their Title

The Sex attribute is factorized and the Age variable is imputed off the NAs using the median age by Title.

3.1.3. Ticket:

Looking at the dataset, we observe that the ticket variable is rather messy. Thus, this variable requires cleaning before analysis. Ticket is thus split into the ticket string and ticket number.

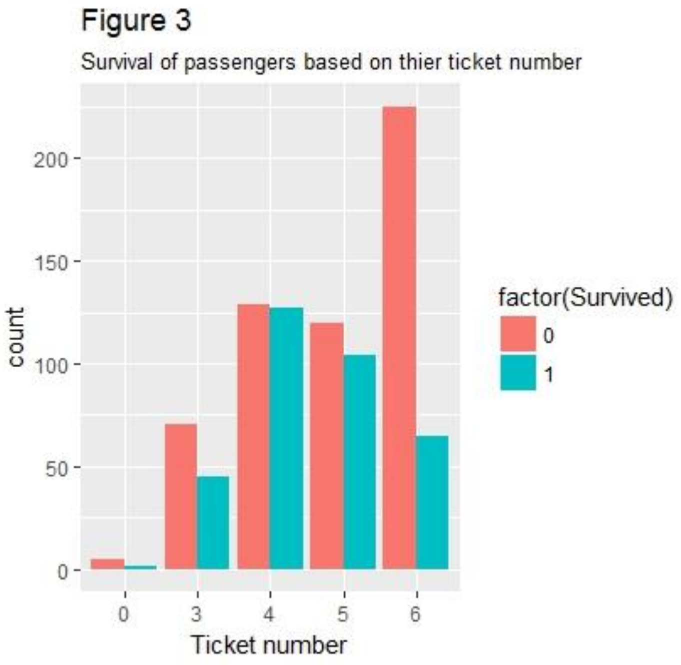


Fig.3. Survival by ticket number

There is somewhat of a pattern that can be observed between the plot for survival by ticket number and survival by Pclass.

3.1.4. *Fare:*

The NA values for fare is imputed by replacing it with the median fare value.

3.1.5. *Cabin:*

The cabin variable has many missing values and hence is not useful for analysis.

3.1.6. *Embarked:*

The missing values in embarked is automatically assumed to embark from Southampton.

3.2. Bivariate analysis:

3.2.1. Do families sink or swim together?

Th attributes SibSp and Parch are the key information to learn how many people travelled together with the passenger. This can be combined to form a family variable for analysis.

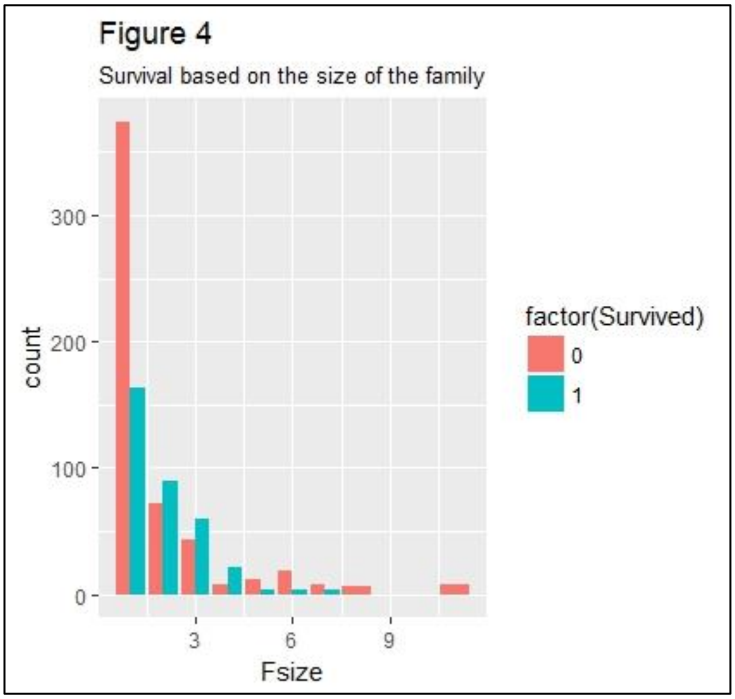


Fig.4. Survival count of the passengers based on the size of the family

Based on the output in Fig.4, we observe that passengers with large families or passengers who travelled alone didn’t have much chance of survival compared to passengers with small families and this is further confirmed by the plot in Fig.5.

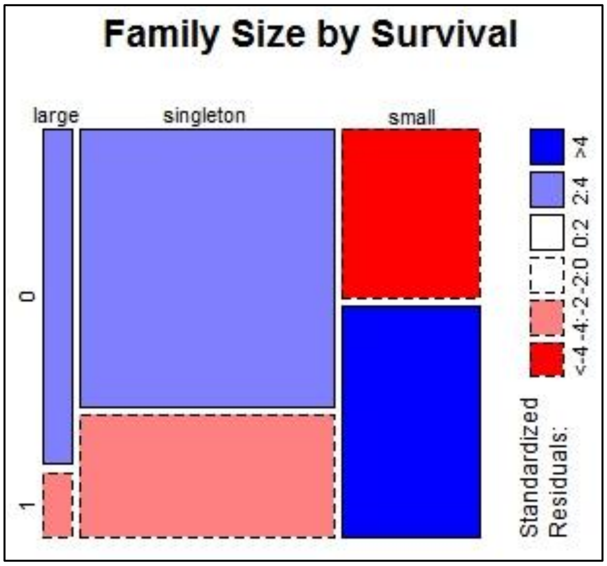


Fig.5. Mosaic plot showing survival of family by the size

3.2.2. Age vs survival:

Now that the age variable has been imputed and explored, we create age-dependent variables: child and mother. We consider a passenger to be a child if their age is less than 18, and a mother as a female passenger over 18 and with more than 0 children and does not have the title 'Miss'. First, we look at how the survival count stands per age for both the sexes.

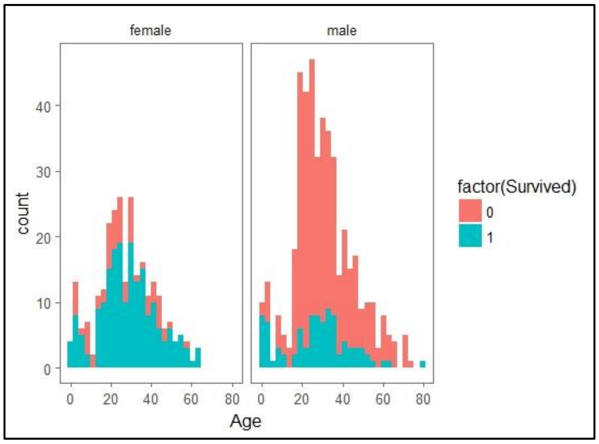
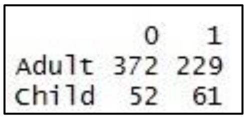


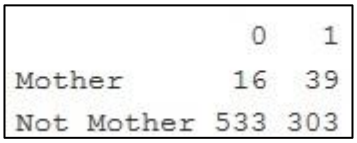
Fig.6. Age vs Sex vs Survival

We observe on analysis, that being a child betters your chance of survival, but doesn’t necessarily leave you safe as seen from the output below.



```
        0   1
Adult 372 229
Child  52  61
```

It is interesting to learn the survival rate of mothers aboard the Titanic. The output is as given below:



```
              0   1
Mother       16  39
Not Mother  533 303
```

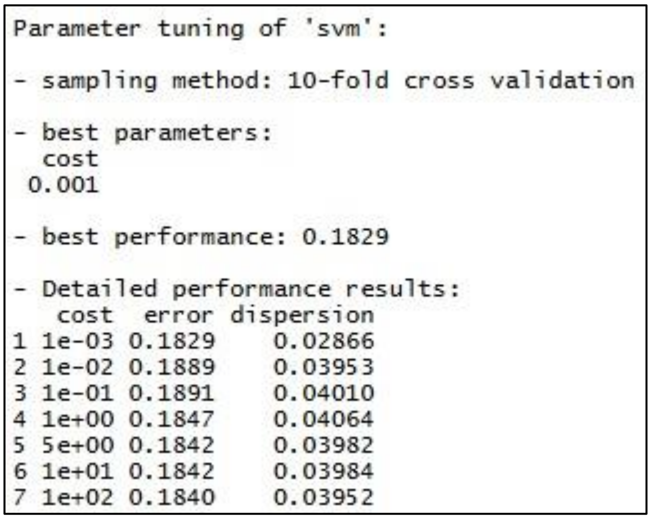
4. Experimental results:

This section highlights the results obtained from the Kernel SVM learning technique used in the study. The focus lies on the two techniques of kernel SVM learning, linear kernel SVM and radial kernel SVM to compare the features of both methods and highlight the results of the best model.

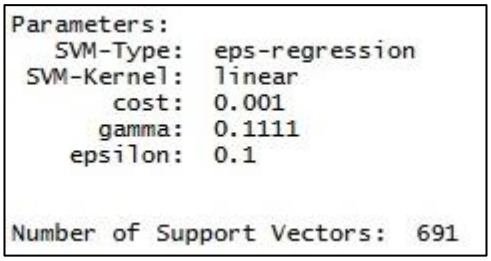
4.1. Linear Kernel SVM:

We know that the SVM method works by grouping feature points per the classifiers. In linear kernel, the SVM algorithm finds the largest linear margin that separates the regions of the feature points.

The results on implementing the linear kernel model is as follows:



```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
  cost
 0.001

- best performance: 0.1829

- Detailed performance results:
   cost  error dispersion
1 1e-03 0.1829    0.02866
2 1e-02 0.1889    0.03953
3 1e-01 0.1891    0.04010
4 1e+00 0.1847    0.04064
5 5e+00 0.1842    0.03982
6 1e+01 0.1842    0.03984
7 1e+02 0.1840    0.03952
```



```
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  0.001
      gamma:  0.1111
    epsilon:  0.1


Number of Support Vectors:  691
```

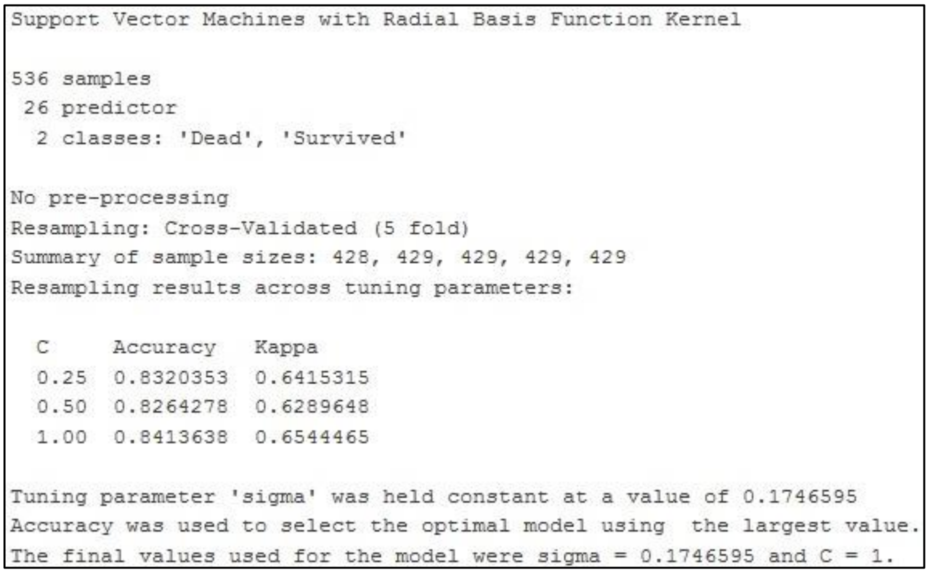
The accuracy of the linear kernel model is 76.555%.

4.2. Radial Basis Function Kernel SVM:

Another popular kernel method is the radial basis function kernel SVM. The value of the RBF kernel decreases with the distance and is in the range of 0 and 1 and can be used as the similarity measure resulting in a feature space with infinite number of dimensions.

4.2.1. First iteration:

For the first iteration of the radial kernel, the default tuning parameters are used. The output is as follows:



```
Support Vector Machines with Radial Basis Function Kernel

536 samples
 26 predictor
  2 classes: 'Dead', 'Survived'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 428, 429, 429, 429, 429
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.8320353  0.6415315
  0.50  0.8264278  0.6289648
  1.00  0.8413638  0.6544465

Tuning parameter 'sigma' was held constant at a value of 0.1746595
Accuracy was used to select the optimal model using  the largest value.
The final values used for the model were sigma = 0.1746595 and C = 1.
```

The highest accuracy of this model is noted to be 84.1% for sigma value of 0.1746595.

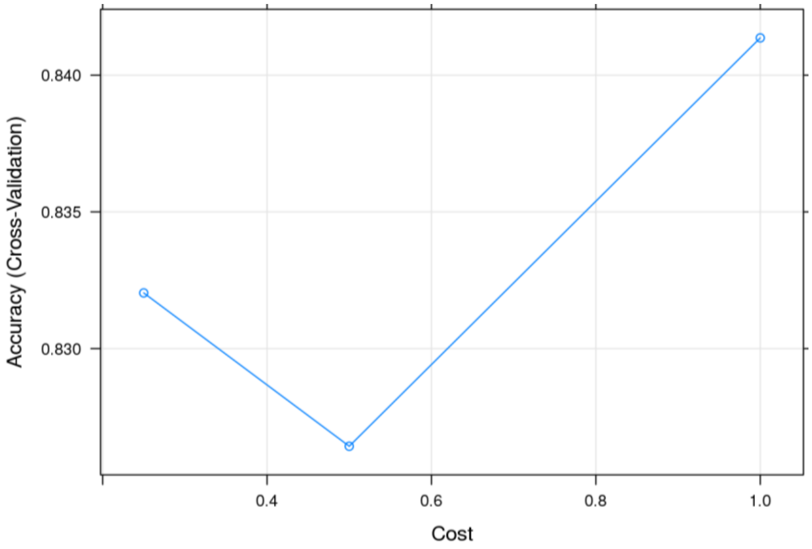
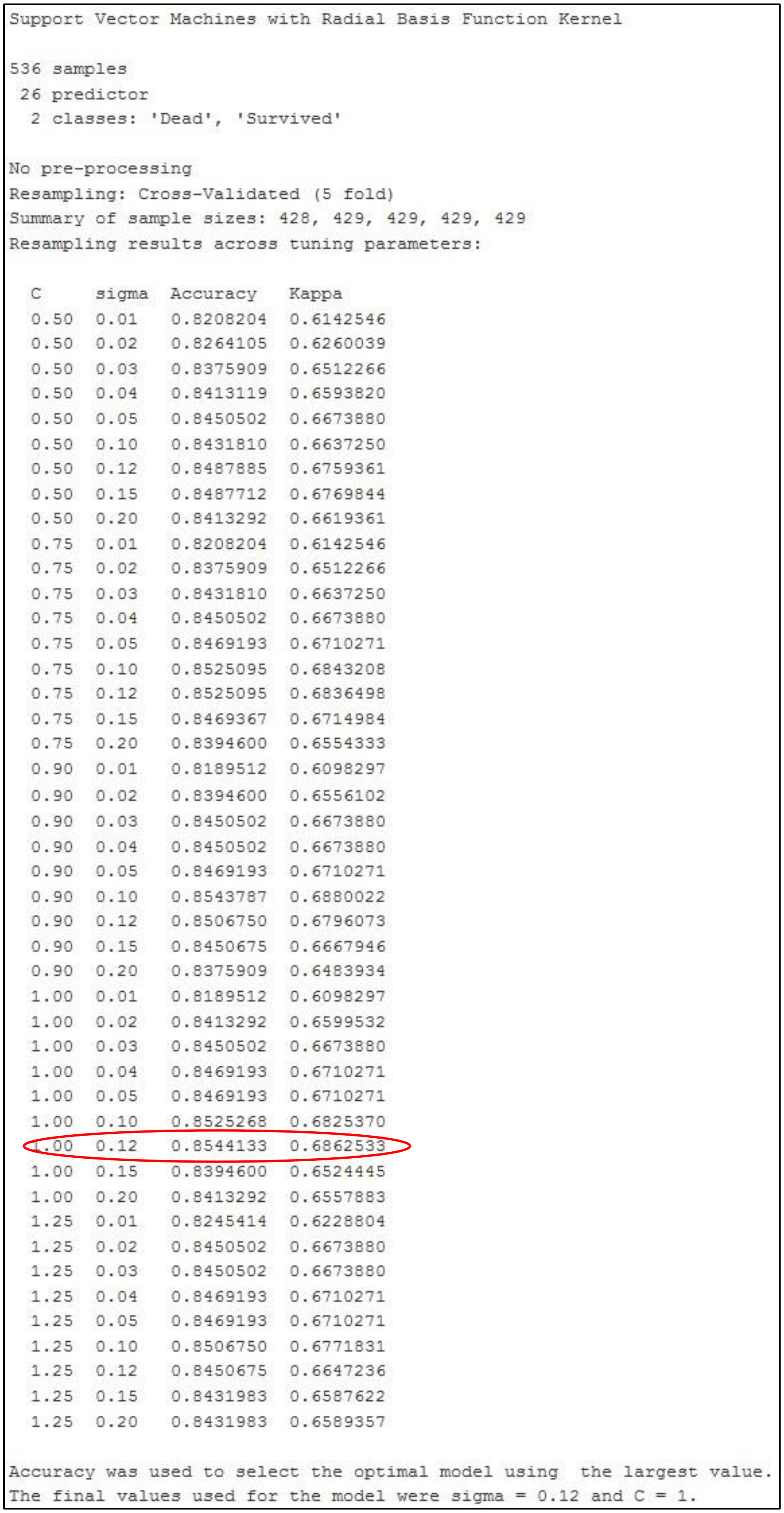


Fig.7. Cost vs accuracy for first iteration

4.2.2. Second iteration:

For the second iteration, the values are manually tuned to compare the performance. The output is as follows:



```
Support Vector Machines with Radial Basis Function Kernel

536 samples
 26 predictor
  2 classes: 'Dead', 'Survived'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 428, 429, 429, 429, 429
Resampling results across tuning parameters:

  C     sigma  Accuracy   Kappa
  0.50  0.01   0.8208204  0.6142546
  0.50  0.02   0.8264105  0.6260039
  0.50  0.03   0.8375909  0.6512266
  0.50  0.04   0.8413119  0.6593820
  0.50  0.05   0.8450502  0.6673880
  0.50  0.10   0.8431810  0.6637250
  0.50  0.12   0.8487885  0.6759361
  0.50  0.15   0.8487712  0.6769844
  0.50  0.20   0.8413292  0.6619361
  0.75  0.01   0.8208204  0.6142546
  0.75  0.02   0.8375909  0.6512266
  0.75  0.03   0.8431810  0.6637250
  0.75  0.04   0.8450502  0.6673880
  0.75  0.05   0.8469193  0.6710271
  0.75  0.10   0.8525095  0.6843208
  0.75  0.12   0.8525095  0.6836498
  0.75  0.15   0.8469367  0.6714984
  0.75  0.20   0.8394600  0.6554333
  0.90  0.01   0.8189512  0.6098297
  0.90  0.02   0.8394600  0.6556102
  0.90  0.03   0.8450502  0.6673880
  0.90  0.04   0.8450502  0.6673880
  0.90  0.05   0.8469193  0.6710271
  0.90  0.10   0.8543787  0.6880022
  0.90  0.12   0.8506750  0.6796073
  0.90  0.15   0.8450675  0.6667946
  0.90  0.20   0.8375909  0.6483934
  1.00  0.01   0.8189512  0.6098297
  1.00  0.02   0.8413292  0.6599532
  1.00  0.03   0.8450502  0.6673880
  1.00  0.04   0.8469193  0.6710271
  1.00  0.05   0.8469193  0.6710271
  1.00  0.10   0.8525268  0.6825370
  1.00  0.12   0.8544133  0.6862533
  1.00  0.15   0.8394600  0.6524445
  1.00  0.20   0.8413292  0.6557883
  1.25  0.01   0.8245414  0.6228804
  1.25  0.02   0.8450502  0.6673880
  1.25  0.03   0.8450502  0.6673880
  1.25  0.04   0.8469193  0.6710271
  1.25  0.05   0.8469193  0.6710271
  1.25  0.10   0.8506750  0.6771831
  1.25  0.12   0.8450675  0.6647236
  1.25  0.15   0.8431983  0.6587622
  1.25  0.20   0.8431983  0.6589357

Accuracy was used to select the optimal model using  the largest value.
The final values used for the model were sigma = 0.12 and C = 1.
```

As seen from the output above, the best model has an accuracy of 85.44% with a sigma value of 0.12.

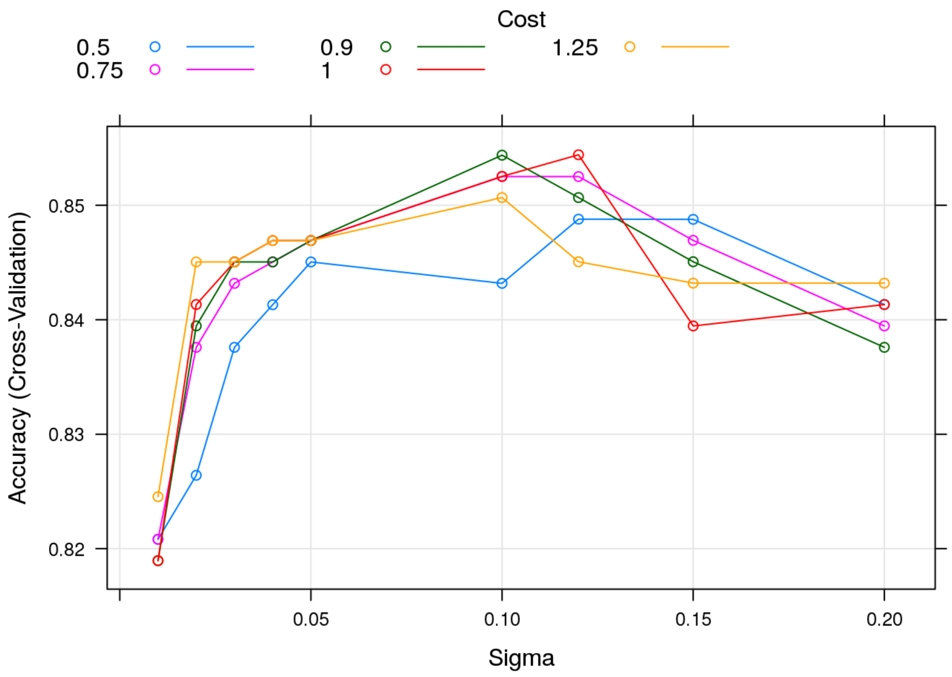
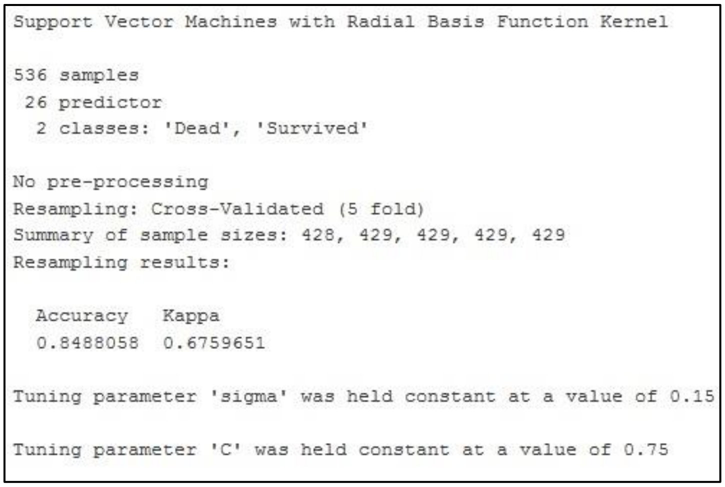


Fig.8. Sigma values for accuracy

From the above plot, we observe that to obtain a robust model, the sigma value is between 0.10 to 0.15.

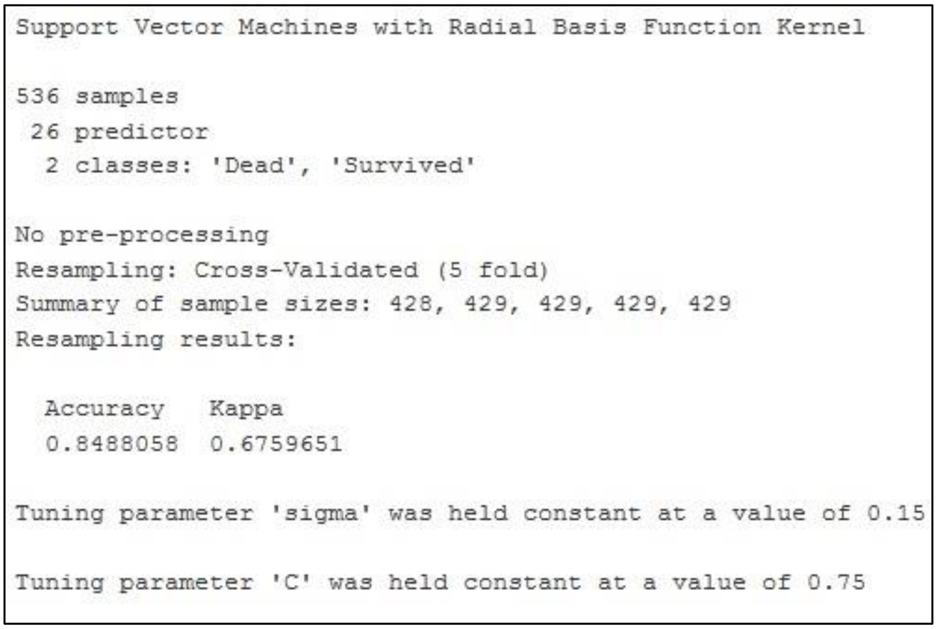
4.2.3. Third iteration:

In the third iteration, the values are chosen to build a robust model for prediction. The output is as follows:

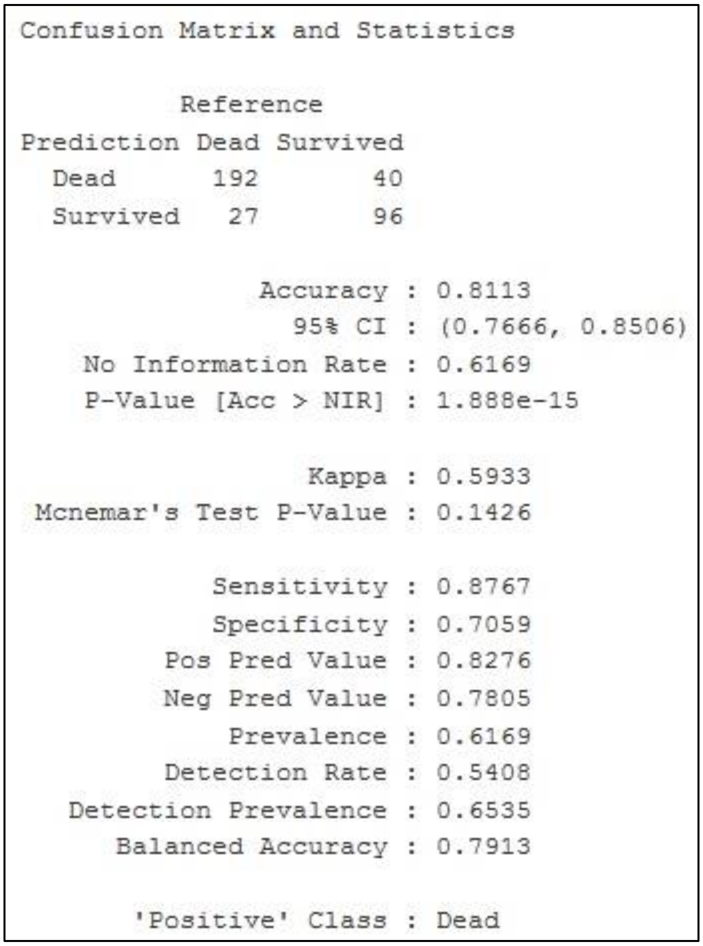


```
Support Vector Machines with Radial Basis Function Kernel

536 samples
 26 predictor
  2 classes: 'Dead', 'Survived'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 428, 429, 429, 429, 429
Resampling results:

  Accuracy   Kappa
  0.8488058  0.6759651

Tuning parameter 'sigma' was held constant at a value of 0.15

Tuning parameter 'C' was held constant at a value of 0.75
```

From the R output above, we observe that the accuracy of the model is improved to 84.88% for the sigma value of 0.15. The next step is to predict the values and test the performance. The output is as shown below:



```
Support Vector Machines with Radial Basis Function Kernel

536 samples
 26 predictor
  2 classes: 'Dead', 'Survived'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 428, 429, 429, 429, 429
Resampling results:

  Accuracy   Kappa
  0.8488058  0.6759651

Tuning parameter 'sigma' was held constant at a value of 0.15

Tuning parameter 'C' was held constant at a value of 0.75
```



```
Confusion Matrix and Statistics

          Reference
Prediction Dead Survived
  Dead      192       40
  Survived   27       96

               Accuracy : 0.8113
                 95% CI : (0.7666, 0.8506)
    No Information Rate : 0.6169
    P-Value [Acc > NIR] : 1.888e-15

                  Kappa : 0.5933
 Mcnemar's Test P-Value : 0.1426

            Sensitivity : 0.8767
            Specificity : 0.7059
         Pos Pred Value : 0.8276
         Neg Pred Value : 0.7805
             Prevalence : 0.6169
         Detection Rate : 0.5408
   Detection Prevalence : 0.6535
      Balanced Accuracy : 0.7913

       'Positive' Class : Dead
```

4.2.4. Model evaluation:

The ROC curve in Fig.9., shows the robustness of the model. The AUC of the model was calculated to be 0.91 and the accuracy of the model is 84.8%.

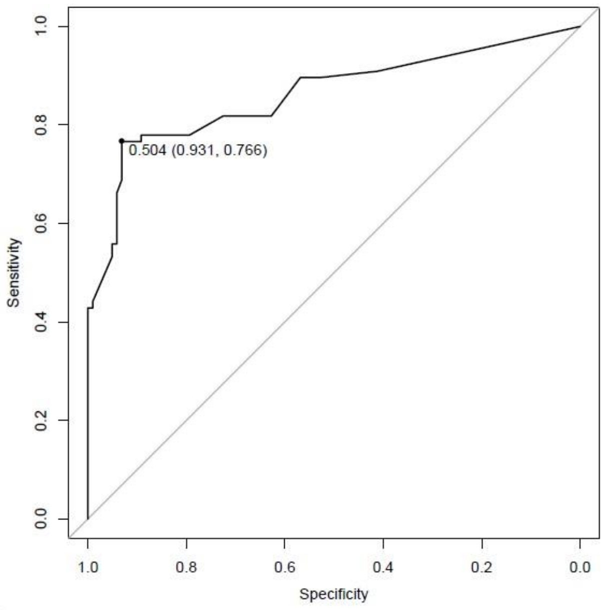
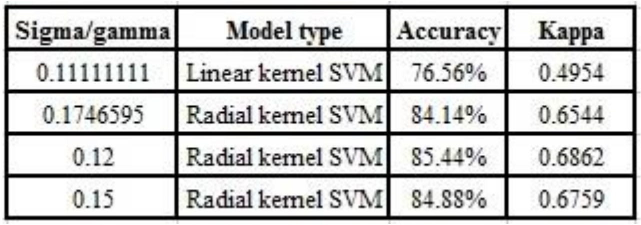


Fig. 9. ROC curve of the model

4.3. Model comparison and summary:



Sigma/gamma	Model type	Accuracy	Kappa
0.11111111	Linear kernel SVM	76.56%	0.4954
0.1746595	Radial kernel SVM	84.14%	0.6544
0.12	Radial kernel SVM	85.44%	0.6862
0.15	Radial kernel SVM	84.88%	0.6759

5. Experimental Analysis:

The objective of this study is to find a model using kernel SVM techniques that can predict the survivability of the passenger aboard Titanic, based on the features in the data. This could help possibly avoid future disasters and solve the mystery behind the ‘unsinkable’.

This study started with the exploration of the dataset. After initial data analysis of missing values imputation, feature selection, univariate and bivariate analyses the data is deemed good for model building.

In this study, various methods of the kernel SVM techniques are incorporated into the model. The model is used to predict the survivability of passengers on the Titanic. The first model built is based on linear kernel SVM. With the cost values set to 0.001, 0.01, 0.1, 1, 5, 10 and 100, the SVM model is tuned to be built using linear kernel and 10-fold cross validation is applied to produce results with an accuracy of 76.555%. Previous literature in the study have reported better results and that inspired to work with radial based function kernel SVM, to build a model with better accuracy.

The next model built is based on radial kernel SVM. In the first iteration, default tuning parameters were used to understand the functioning of the model with the data. The best model in the first iteration, has the sigma value of 0.1746595 at the cost of 1 and an accuracy of 84.136%. A second iteration of the model is done, where the tuning parameters are entered manually and the performance is compared. The sigma values on which the model is built are: 0.01 through 0.05, 0.1, 0.12 and 0.15 with the varying costs of: 0.5, 0.75, 0.9, 1 and 1.25. The bets model in this iteration has a sigma value of 0.12 with cost being 1 and the accuracy of the model is 85.441%. On plotting the model, we observe that the model looks robust between 0.1 and 0.15 and the cost is 0.75 as is seen in Fig.

8. So, for the third iteration, we consider the cost to be 0.75 and a sigma value of 0.15 and the model has an accuracy of 84.88% with a Kappa value of 0.6759.

The model is then run on the testing set for prediction and the performance of the model is calculated. The model returned an accuracy of 83.5% with the Kappa value of 0.6435. We now calculate the ROC and AUC of the model as the metric for the model's performance. Looking at the ROC curve in Fig. 9, we observe that the plot is towards the true positive side, which means that very few/none of the classes are misclassified. As the ROC gets closer to the optimal point of perfection, AUC gets closer to 1. The AUC value for the model is 0.91 which further indicates that the model is adequate for prediction. **Thus, we may conclude that we are 95% confident that the model can predict the survivability of the passenger aboard the Titanic with an accuracy of 83.5% as the p-value <<0.05.**

6. Conclusion:

The focus of this study is to find an efficient and effective model using the kernel SVM techniques to predict the survivability of passengers aboard the Titanic based on the features in the dataset with relatively good accuracy. The Titanic dataset is acquired from the Kaggle repository and is used in this analysis. The dataset has 1309 instances and 12 attributes which was split into training and testing sets for model building.

After data preprocessing techniques of missing value imputations and standardization were performed on the data, kernel learning methods such as Linear kernel SVM and Radial based function kernel SVM are implemented. The model using linear SVM could predict the survivability of the passengers with 76.56% accuracy. Radial based function kernel learning using SVM is then applied through three iterations to find the best model. In the first iteration, the best model achieved an accuracy of 84.4% using default tuning parameters. In the second iteration, manual tuning parameters was used to build the model and resulted in 85.44%. However, on plotting the model, it was found that the optimum robustness of the model is enhanced when the cost is set to 0.75 and sigma has a value of 0.15. Thus, in the third iteration, the model parameters were tuned to the optimum setting and it could predict the outcome with 84.88% accuracy and when tested has an AUC value of 0.91. Hence, we may conclude that an effective and efficient model is built using the Radial based function kernel SVM set to sigma value of 0.15 and the cost function value of 0.75 to achieve an accuracy of 84.88% with 95% confidence interval to predict the survivability of passengers aboard the Titanic.

This study has only inspired further questions to be asked in terms of other factors that could affect the survivability using better feature selection mechanisms. Various options in kernel learning techniques can be implemented on the model to compare the results. This study can also extend over datasets of varying fields to find interesting analysis. There is a never-ending search for a model that can provide results with high accuracy with minimum computational complexity that can be used in different fields.