Predicting Titanic Passenger Survival Using Support Vector Machines (SVM)

GITHUB Repository Link: https://github.com/rk24aao/Predicting-Titanic-Passenger-Survival

Table of Contents

Introduction

Overview of the Tutorial

One of the supervised ML algorithms where Support Vector Machines (SVM) is used mostly for regression and classification problems. “SVM is the best hyperplane to classify data into two categories in the best way such that the classification will be possible in the maximum margin. Also, SVM has one clear advantage of being able to work fairly well in high-dimensional spaces and being quite resistant to overfitting particularly for small datasets (Tampinongkol *et al.*, 2024). The given fields are text classification, image recognition and others. The versatility of SVM is due to its reusability of the kernel that can be used to make linear or nonlinear data separation.

Objective of the Tutorial

This report tries to have a grasp over the whole story of how Support Vector Machines (SVM) value assigned to result is calculated and its application in classifying a real-life classification problem. To do this Titanic dataset is used, which contains information about the passengers like age, sex and class, and it can be used to show how to apply SVM to predict if the passengers survived the Titanic disaster. The tutorial kicks off with an explanation of the high-level mechanics of SVM and goes on to provide step-by-step instructions for extracting data, transforming it, training models with hyperparameters and measuring performance.

Literature Review

Introduction to SVM

Support Vector Machines (SVM) is a supervised machine learning algorithm that is mainly used for classification and regression”. SVM was created by Vladimir Vapnik and other authors in the 1990s and is founded on statistical learning theory (Veisi, H., 2023.). SVM has been utilized to a very large extent to solve linear and nonlinear classification issues by projecting data onto high-dimensional spaces. Unlike the universal classifiers like logistic regression, SVM seeks to discover the optimal decision boundary with the maximum margin between classes, thus a great tool for pattern recognition. Unlike other classifiers like logistic regression, SVM seeks to discover the optimal hyperplane with the maximum margin between classes to achieve generalization. It is very efficient in pattern recognition, image classification, and outliers detection. SVM can also employ a variety of kernel functions like linear, polynomial, radial basis function (RBF), and sigmoid, thus very adaptable in dealing with complex data. Although it is very robust in performance, SVM is computationally costly with very large datasets and is parameter-sensitive and needs careful tuning to yield the best outcome.

Theory of SVM

The fundamental idea of “SVM is to discover the optimal hyperplane to distinguish data points of disparate classes. For linearly separable data, SVM discovers the hyperplane that maximizes the margin of the closest data points, or the support vectors. Margin is the distance from the hyperplane to the closest data points of both classes. This optimization problem is addressed by the Lagrange multiplier method and quadratic programming. For non-linearly separable data, SVM employs the kernel trick, which transforms input features to a higher-dimensional space where a separating hyperplane can be discovered (Liu, Z.L., 2025). Popular kernels are linear, polynomial, radial basis function (RBF), and sigmoid. An effective kernel function has a significant impact on SVM performance.

Key Concepts:

- ***Support Vectors*** – Data points closest to the decision boundary that influence its position.
- ***Hyperplane*** – The optimal decision boundary separating different classes.
- ***Margin*** – The distance between the hyperplane and the closest support vectors; a larger margin results in better generalization.
- ***Kernel Trick*** – A mathematical function that transforms non-linearly separable data into a higher-dimensional space.

Applications of SVM:

SVM finds widespread application in various applications because it excels at solving classification issues:

- ***Text Classification*** – Spam filter, sentiment classification, and document categorization.

- ***Image Recognition*** – Digit recognition, handwritten, and facial recognition.
- ***Bioinformatics*** – Detection of cancer and gene classification.

Advantages and Limitations of SVM:

Advantages:

- Performs well in high-dimensional space.
- Effective for situations with a small data set but lots of features.
- Resistant to overfitting, particularly with appropriate regularization.

Disadvantages:

- Computationally costly with large data sets.
- Needs accurate tuning of hyperparameters and choosing the kernel.
- Sensitive to noisy data and overlapping classes.

Methodology

- **Dataset Description**

The Titanic data set has 891 rows and 12 columns with information on passengers of the RMS Titanic. The main columns are PassengerId, Survived (binary: 0 for dead, 1 for alive), Pclass (class of ticket: 1st, 2nd, 3rd), Name, Sex, Age, SibSp (number of siblings/spouses on board), Parch (number of parents/children on board), Ticket (ticket number), Fare, Cabin, and Embarked. However, Age contains missing values (714 rows out of 891), while Cabin contains considerable missing data (just 204 rows). It is commonly used for machine learning classification problems, especially survival prediction based on passenger information and ticket type. It is best suited to model exploration such as Support Vector Machines (SVM), capable of classifying survival outcomes on the basis of non-linear interactions between features. Considering the significance of variables such as “Pclass, Sex, and Fare, feature selection and preprocessing” (encoding categorical variables and missing data handling) are very important prior to training an SVM model. The dataset supports trying out various machine-learning strategies for binary classification issues.

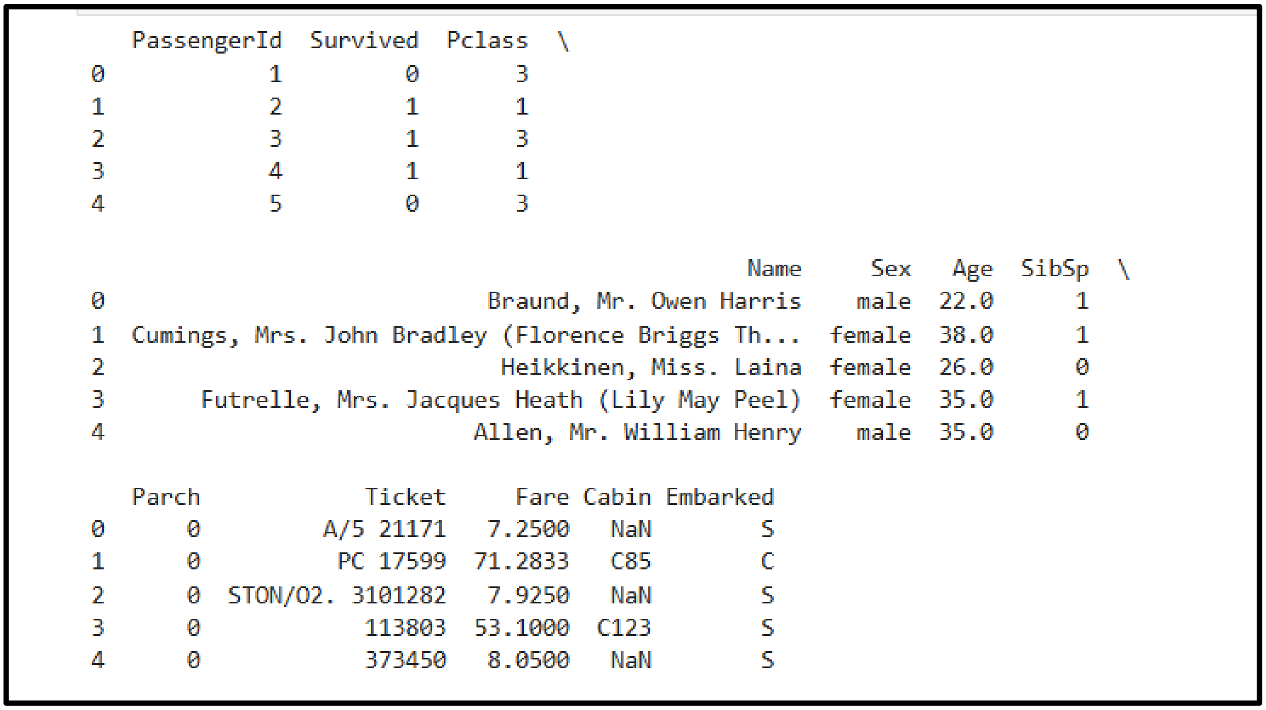
- **Data Preprocessing**

Missing values for Age are replaced with the median, and Cabin column is excluded because there are too many missing values. One-hot encoding is applied to categorical features such as Sex and Embarked. New features are not generated. Feature scaling is needed for SVM as it uses distance-based operations; StandardScaler is applied for numerical feature normalization. The data is divided into 80% training and 20% testing to maintain a balanced test. This preprocessing enhances model performance by addressing inconsistencies, normalizing data, and maximizing classification accuracy in predicting Titanic passenger survival with SVM.

- **SVM Implementation**

The SVM model is applied by preprocessing the data, dealing with missing values, converting categorical variables, and scaling features. The data is divided into training (80%) and testing (20%) sets. The RBF kernel is first used due to its capability of dealing with non-linear relationships, but the linear kernel is also compared. GridSearchCV is applied to optimize hyperparameters such as C and gamma. The model is trained on the training set and tested using accuracy, precision, and recall on the test set. Accurate kernel selection allows the model to successfully distinguish survivors from non-survivors according to passenger features.

Results and Evaluation



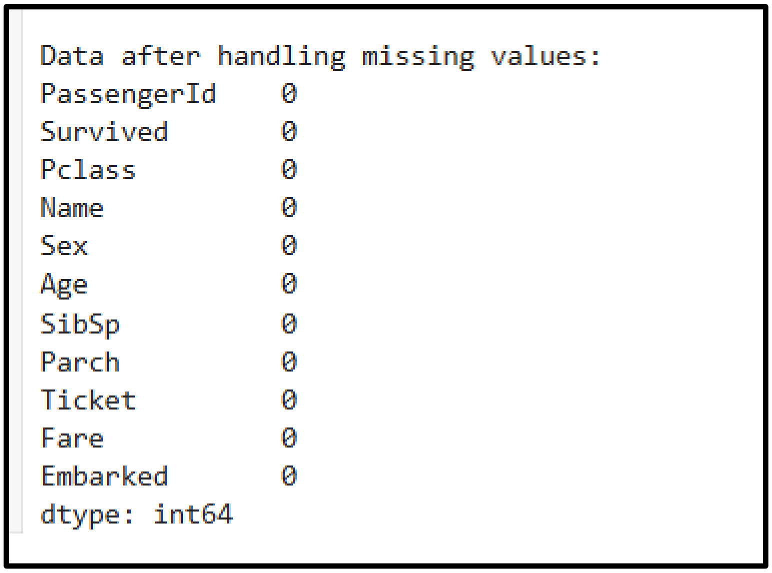
```
   PassengerId  Survived  Pclass  \
0            1         0       3   
1            2         1       1   
2            3         1       3   
3            4         1       1   
4            5         0       3   

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1   
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1   
2                             Heikkinen, Miss. Laina  female  26.0      0   
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1   
4                           Allen, Mr. William Henry    male  35.0      0   

   Parch            Ticket     Fare Cabin Embarked  
0      0         A/5 21171   7.2500   NaN        S  
1      0          PC 17599  71.2833   C85        C  
2      0  STON/O2. 3101282   7.9250   NaN        S  
3      0            113803  53.1000  C123        S  
4      0            373450   8.0500   NaN        S  
```

Figure 1: Data Loading

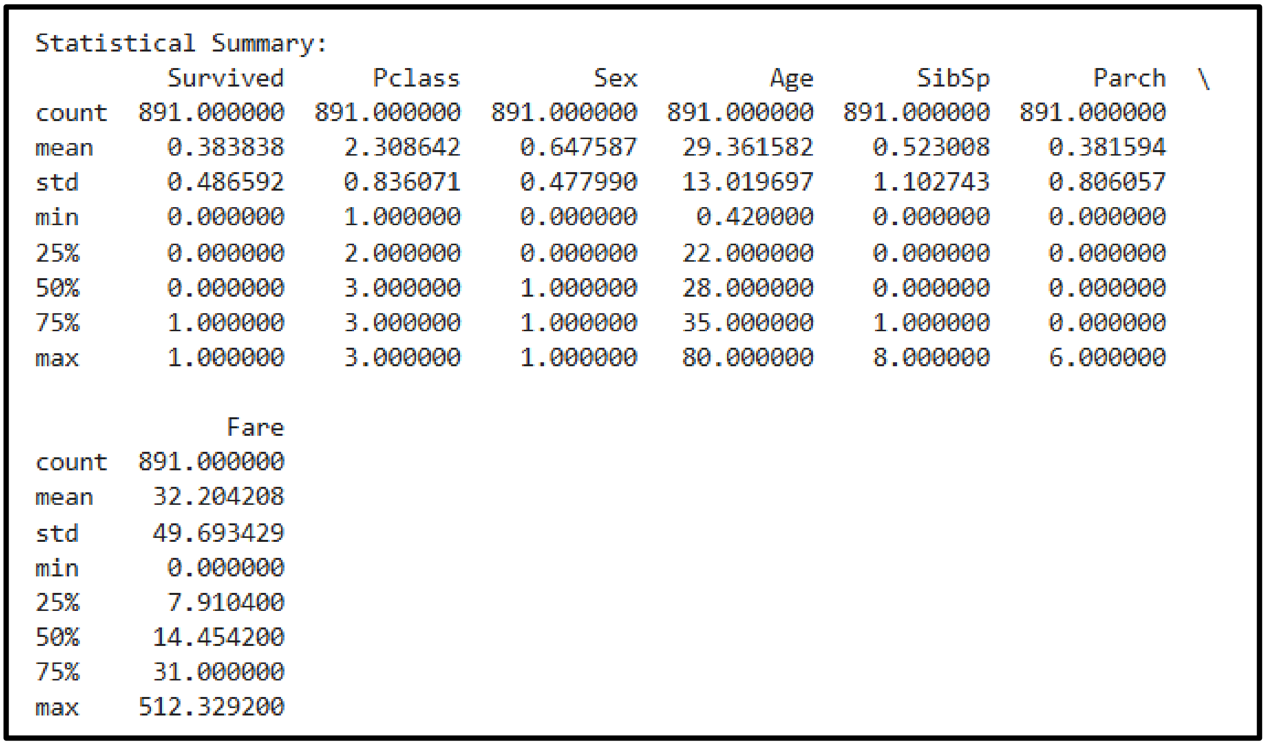
Here is one of the snapshots from the Titanic dataset with relevant information as first 5 rows: PassengerId, Survived (target variable), Pclass (passenger class), Name, Sex, Age, SibSp (siblings/spouses), Parch (parents/children).



```
Data after handling missing values:
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

Figure 2: Handling missing values

The figure shown here is the dataset after dealing with missing values. For Age, Embarked, and so on, This imputed details to the missing values or dropped them depending on the requirement. The dataset was then updated with zero missing values on all columns, signifying that the dataset is now clean and ready for training.



```
Statistical Summary:
         Survived      Pclass         Sex         Age       SibSp       Parch  \
count  891.000000  891.000000  891.000000  891.000000  891.000000  891.000000
mean     0.383838    2.308642    0.647587   29.361582    0.523008    0.381594
std      0.486592    0.836071    0.477990   13.019697    1.102743    0.806057
min      0.000000    1.000000    0.000000    0.420000    0.000000    0.000000
25%      0.000000    2.000000    0.000000   22.000000    0.000000    0.000000
50%      0.000000    3.000000    1.000000   28.000000    0.000000    0.000000
75%      1.000000    3.000000    1.000000   35.000000    1.000000    0.000000
max      1.000000    3.000000    1.000000   80.000000    8.000000    6.000000

             Fare
count  891.000000
mean    32.204208
std     49.693429
min      0.000000
25%      7.910400
50%     14.454200
75%     31.000000
max    512.329200
```

Figure 3: Summary Statistics

This shows the summary statistics of the Titanic dataset and then gives the insights about the central tendency, spread and the shape for the distribution of the numerical features such as Age, Fare and SibSp. Important measures of data distribution such as mean, median, standard deviation, and percentiles are used for understanding the data, and identify outliers or if the data is skewed.

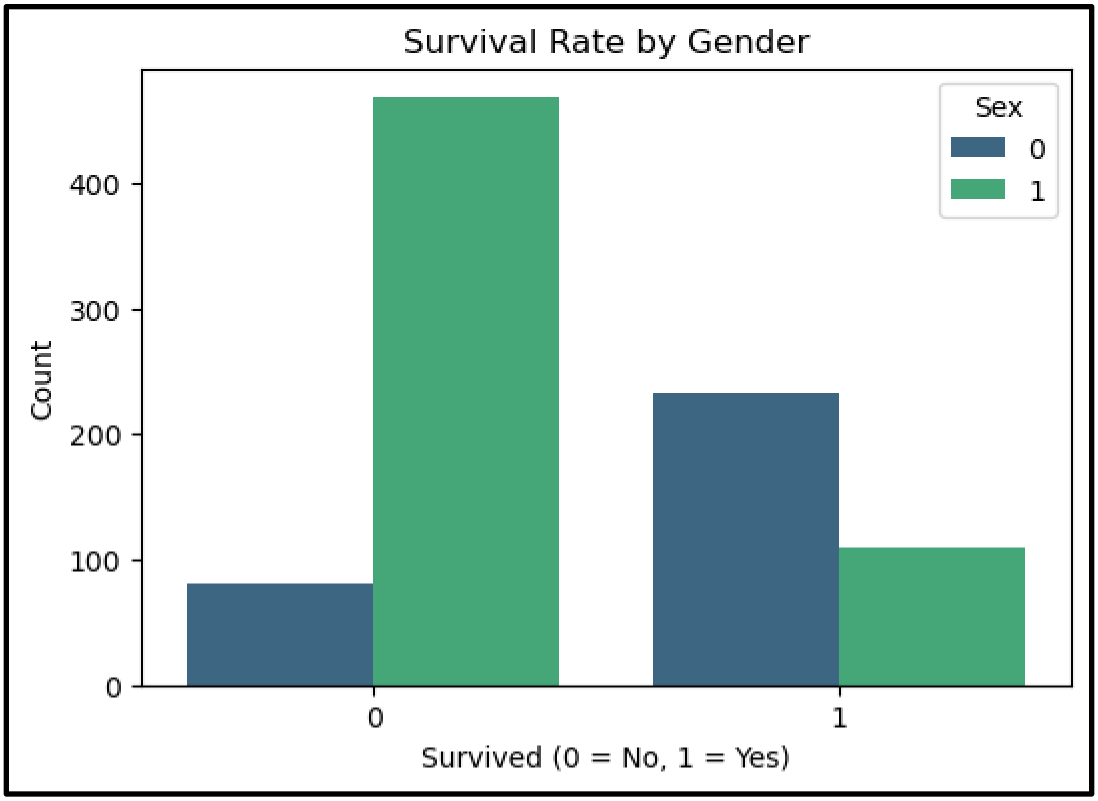


Figure 4: Survival Rate by Gender

This graph depicts the survival rate and market by gender. This plot is a counterplot of male passengers (1) with a lower survival count than that of female passengers (0) with a higher survival rate. The significance of this visualization lies in how it can assist in visualizing gender-based disparities in survival in the Titanic disaster and in a common trend that is found in many historical datasets.

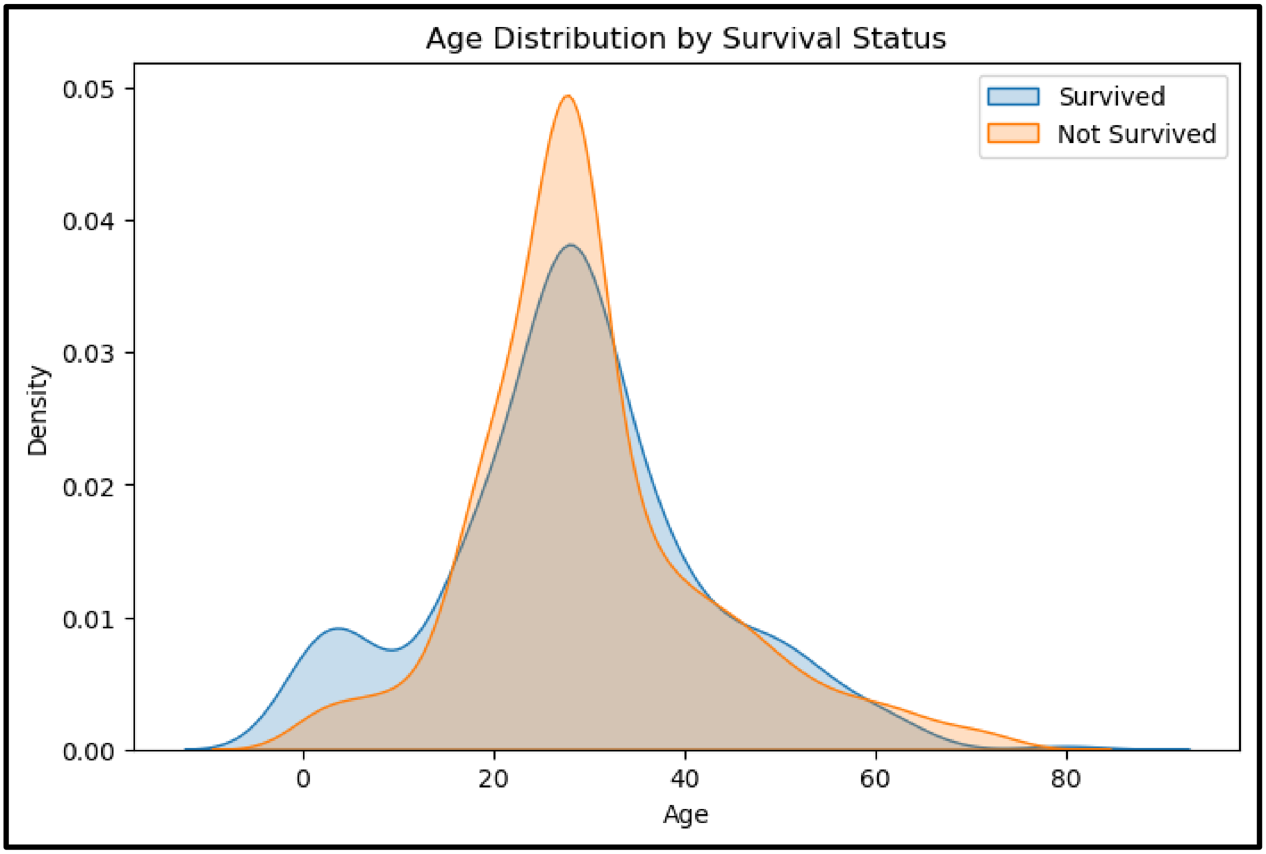


Figure 5: Age Distribution by Survival Status

The above figure is a density plot of the age distribution of passengers that survived versus passengers who didn’t. The concentration of survivors in younger age groups is less than that of non-survivors, indicated by the blue curve representing the survivors, while the orange curve indicates the broad distribution of the survivors in all ages.

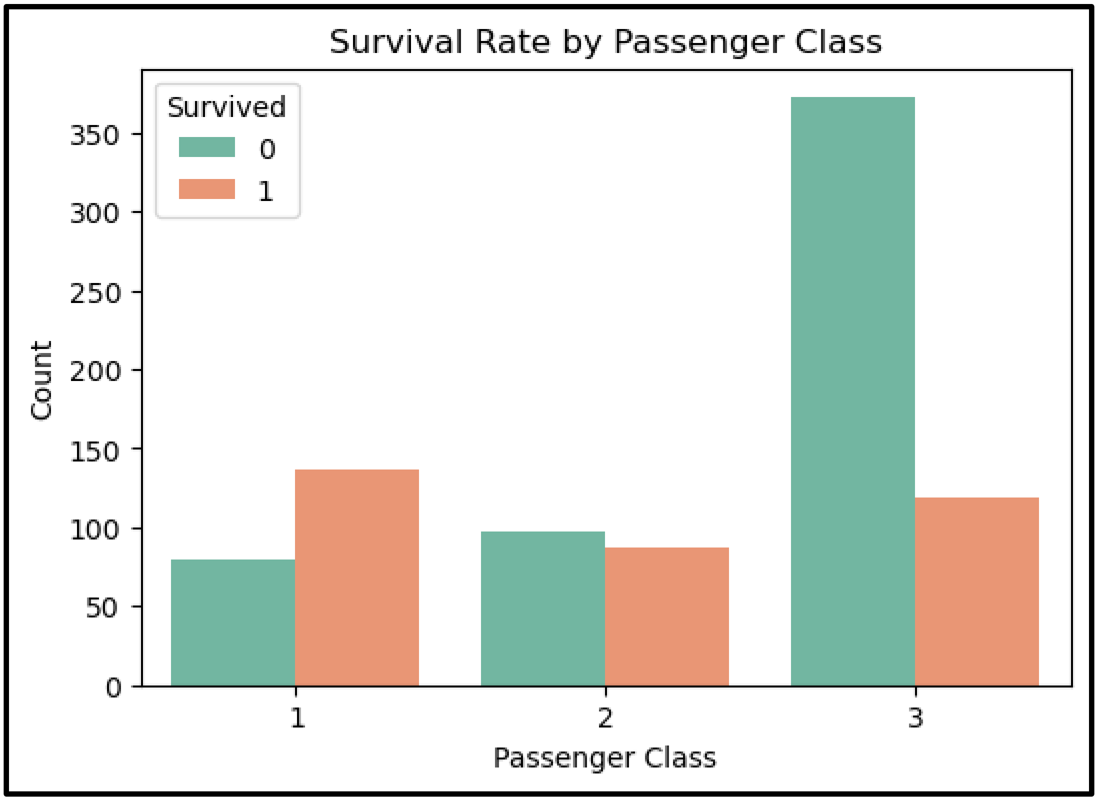
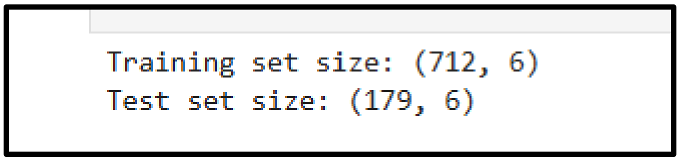


Figure 6: Survival Rate by Passenger Class

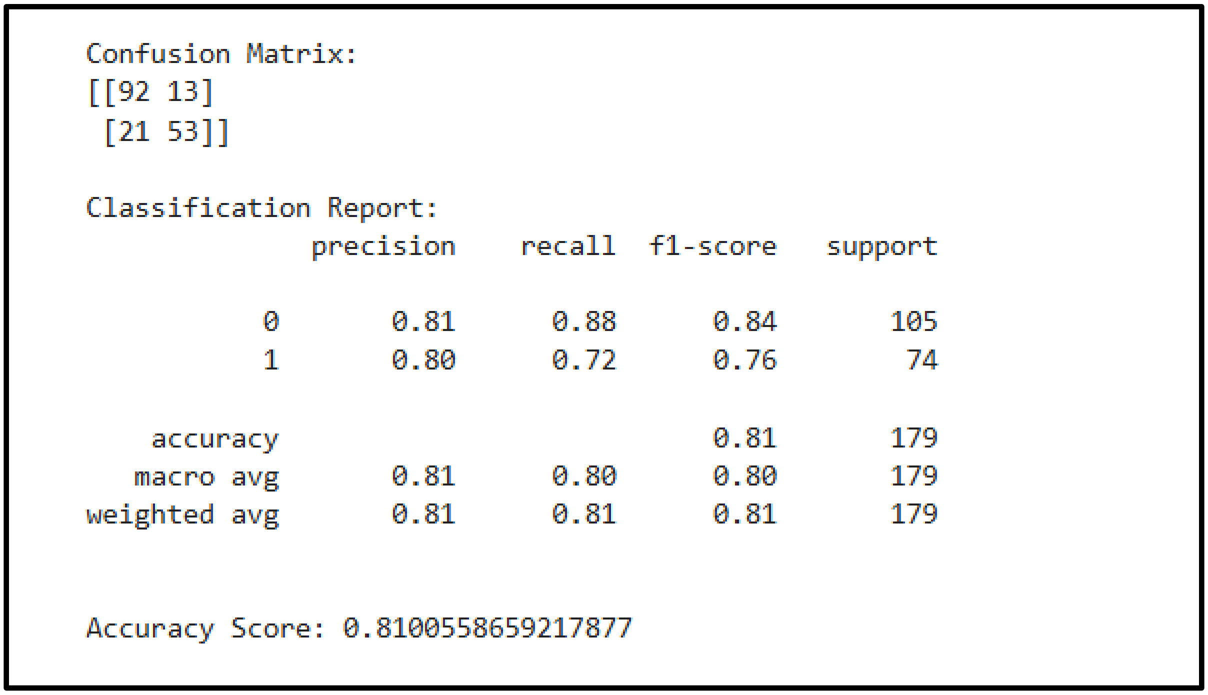
This shows the survival rate by the passenger class in a counterplot. Without a doubt, the chart shows that the highest survivors were of First Class and then Second Class and the poorest survivors were in the Third Class.



```
Training set size: (712, 6)
Test set size: (179, 6)
```

Figure 7: Data Splitting

After splitting Titanic dataset, the training and test set size are shown in Figure 7. The training set contains a sample of size of 712 and the test set has a size of 179. This split also ensures that the model is trained on most of the data and tested on another unseen set to check how the model does.



```
Confusion Matrix:
[[92 13]
 [21 53]]

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.88      0.84       105
           1       0.80      0.72      0.76        74

    accuracy                           0.81       179
   macro avg       0.81      0.80      0.80       179
weighted avg       0.81      0.81      0.81       179


Accuracy Score: 0.8100558659217877
```

Figure 8: Model Performance Evaluation

The confusion matrix and classification report of the evaluation for the SVM model" are shown in this figure. It gives a view on the model's performance and the precision, recall and the f1 score of both classes. About 81% accuracy score was the accuracy of the model to predict the survival.

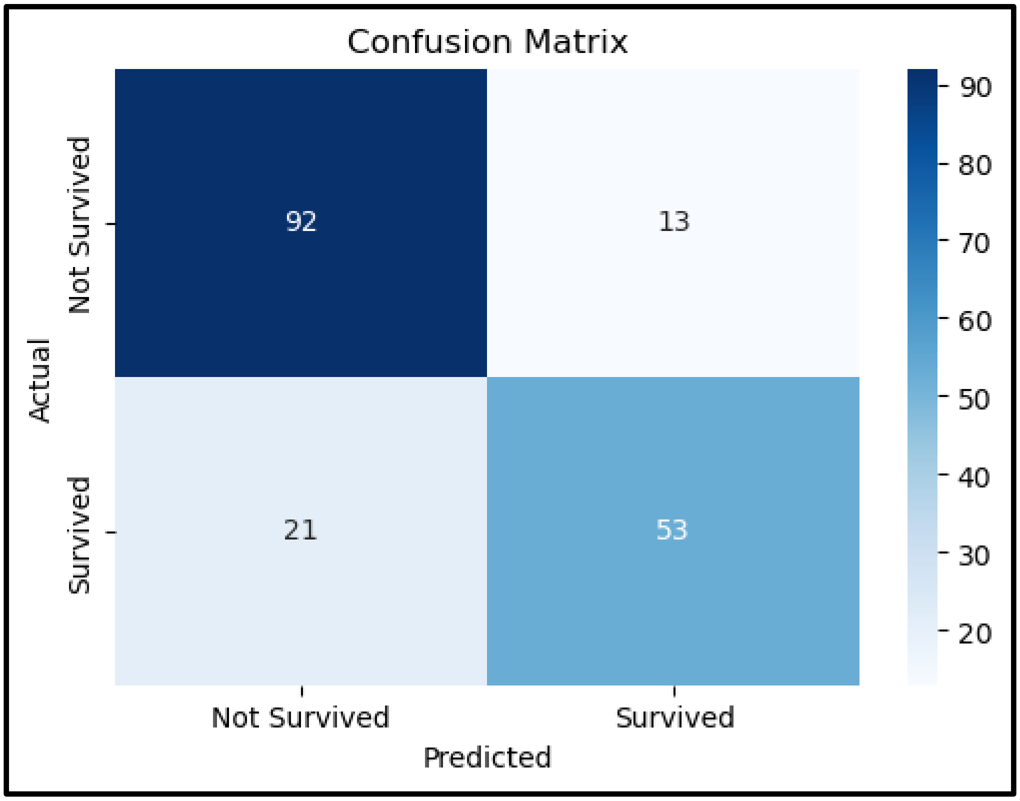
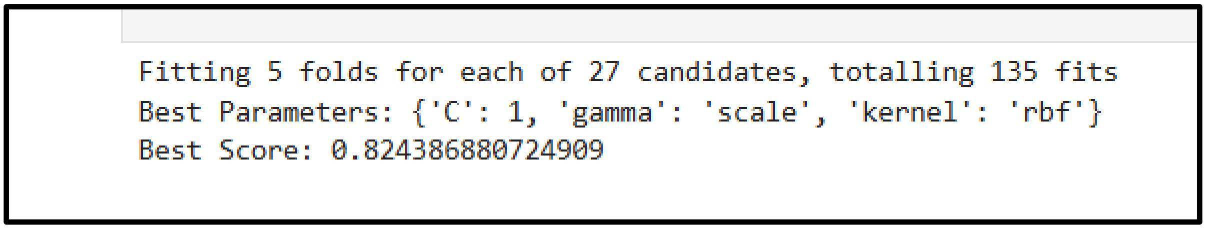


Figure 9: Confusion Matrix

This figure shows the confusion matrix of the performance of the SVM model. It displays the number of true positives, false positives, true negatives and false negatives. Since there are 13 false positives and 21 false negatives, the model does not predict 21 passengers as Not Survived and 13 passengers as Survived.



```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
Best Parameters: {'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}
Best Score: 0.824386880724909
```

Figure 10: Hyperparameter Tuning

GridSearchCV results are displayed in Figure 10. From the above, the best parameters for the SVM model are the C =1; gamma in scale; rbf kernel. The tuning process achieves the best score of 0.824, which signifies better model performance after the tuning process.

Methodological Challenges and Limitations

Challenges

Some challenges were encountered in implementing support vector machines (SVM). On the one hand, kernel selection was one of the primary difficulties. For the dataset, SVM needs to choose an appropriate kernel (linear, polynomial, or radial basis function). Linear kernels are simple and efficient but may not meet the requirements when there are complex patterns in the non-linear data (Reza *et al.*, 2023). Performance improved when an RBF kernel was chosen and the complexity of tuning the hyperparameters, in particular that of C (regularization) and gamma, was significantly increased by the choice of an RBF kernel.

Another challenge was data imbalance. The Titanic dataset had a lot of class imbalance as the number of survivors (1) was smaller compared to non-survivors (0). Distortion of this imbalance might create a bias when the model misclassifies the minority class (survived) which can affect the model's ability to predict correctly. Resampling or class weight adjustment in SVM might help tackle this problem but had to be carefully addressed.

Limitations of SVM

However, SVM has some limitations. Computational complexity is one such limitation and is especially limited while dealing with large datasets. As the size of the dataset increases, the complexity of SVMs also increases, along with the requirement of large memory and computational power. Additionally, SVM struggles with noisy data and outliers. As SVM tries to maximize the margin space between classes, outliers can have a very negative effect on the model's performance, making it fit the data or not classify the one correctly (Mohammadi *et al.*, 2021).

SVMs also have another limitation that it is difficult to separate the non-linear shape. However, the kernel trick enables mapping non-linearly separable data to a high dimensional space but the process becomes computationally expensive as well as can still be suboptimal in terms of performance on high dimensional or non linearly complex datasets.

Conclusion

The SVM model scored 81% in survival prediction for Titanic passengers, with performance enhanced after hyperparameter optimization (C=1, gamma=scale, kernel=RBF, accuracy=82.4%). Survival was significantly impacted by gender and class of passengers. Class imbalance and kernel selection posed challenges, and computational complexity and outlier sensitivity were SVM's limitations. Aside from Titanic data, SVM is generally applied in image classification, medical diagnosis, and detecting fraud. Future enhancements might include experimenting with various kernels, employing ensemble strategies, and considering deep learning approaches for improved predictive accuracy in sophisticated, real-world data sets.

References

Liu, Z.L., 2025. Support vector machines. In *Artificial Intelligence for Engineers: Basics and Implementations* (pp. 129-140). Cham: Springer Nature Switzerland. Available at: https://link.springer.com/chapter/10.1007/978-3-031-75953-6_5

Mohammadi, M., Sarmad, M. and Arghami, N.R., 2021. An Extension of the Outlier Map for Visualizing the Classification Results of the Multi-Class Support Vector Machine. *Malaysian Journal of Computer Science*, *34*(3), pp.308-323. https://jupidi.um.edu.my/index.php/MJCS/article/view/20105

Reza, M.S., Hafsha, U., Amin, R., Yasmin, R. and Ruhi, S., 2023. Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. *Computer Methods and Programs in Biomedicine Update*, *4*, p.100118. https://www.sciencedirect.com/science/article/pii/S2666990023000265

Tampinongkol, F.F., Kamila, A.R., Wardhana, A.C., Kusuma, A.W.C. and Revaldo, D., 2024. Implementation of Random Forest Classification and Support Vector Machine Algorithms for Phishing Link Detection. *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, *7*(1), pp.127-137. https://journal.ittelkom-pwt.ac.id/index.php/insta/article/view/1588

Veisi, H., 2023. Introduction to SVM. In Learning with Fractional Orthogonal Kernel Classifiers in Support Vector Machines: Theory, Algorithms and Applications (pp. 3-18). Singapore: Springer Nature Singapore. Available at: https://link.springer.com/chapter/10.1007/978-981-19-6553-1_1