The sinking of the RMS Titanic on April 15, 1912, is one of the most infamous maritime disasters in history, leading to the tragic loss of over 1,500 lives. This catastrophic event has since captivated public interest and inspired numerous studies to understand the factors that contributed to survival. The Titanic dataset, frequently used in predictive analytics, provides a rich collection of information about the passengers, including demographic details and socio-economic status, making it an ideal candidate for machine learning applications.

This report aims to leverage machine learning techniques to examine the relationship between passenger characteristics and their likelihood of survival. By utilizing various predictive models, we seek to uncover insightful patterns in the data that can illuminate the underlying factors influencing survival rates.

**Data Description**

The dataset used for this analysis is derived from the Titanic passenger manifest and comprises 891 rows and 12 columns. Each row represents an individual passenger, while the columns contain various attributes that provide insights into their demographics, travel details, and survival status. Below is a detailed description of the features in the dataset:

1. PassengerId: A unique identifier assigned to each passenger.
2. Survived: Indicates whether the passenger survived (1) or did not survive (0).
3. Pclass: The passenger class (1st, 2nd, or 3rd) representing the socio-economic status of the passenger.
4. Name: The full name of the passenger.
5. Sex: The gender of the passenger (male or female).
6. Age: The age of the passenger in years. This is a critical variable since age can influence survival chances, particularly among children and the elderly.
7. SibSp: The number of siblings or spouses the passenger had aboard the Titanic.
8. Parch: The number of parents or children the passenger had aboard the Titanic.
9. Ticket: The ticket number of the passenger.
10. Fare: The fare paid for the ticket, which can reflect economic status.
11. Cabin: The cabin number where the passenger stayed (if applicable).
12. Embarked: The port where the passenger boarded the ship (C = Cherbourg, Q = Queenstown, S = Southampton).

**Data Preprocessing**

In the Titanic dataset, missing values are a common challenge that can significantly impact the accuracy of any predictive model. Specifically, the **Age** and **Cabin** columns contain missing data, which can lead to biased results if not appropriately addressed.

After experimenting with different approaches, we found that the model achieved its best accuracy when we chose to drop rows with missing values, particularly focusing on the **Age** column. While some information was lost, the simplicity and effectiveness of this method in producing a more robust model highlighted the trade-offs between data completeness and predictive performance. Also the **Cabin** column has many missing values as well. Due to the large number of missing entries, we considered dropping this feature altogether

In our dataset, both the **Ticket** and **PassengerId** columns contained unique identifiers for each passenger. Therefore, we decided to drop both columns to simplify our dataset and focus on more impactful features.

However, we chose to retain the **Name** column because it contains valuable information that can be transformed into meaningful features. Specifically, we extracted a feature called **Title** from the names, representing the honorifics (e.g., Mr., Mrs., Miss) associated with each passenger.

Following this extraction, we created a new feature called **TitleCategory** by categorizing titles based on their correlation with survival rates. This categorization allowed us to group the titles into three distinct categories that provided further insight into the social status and potential survival likelihood of passengers and droped both the **Name** and **Title** columns from our dataset.

Additionally, we introduced a new feature called **AdultMale**, which is a binary indicator that assigns a value of **True** to male passengers over the age of 18. While the **Age** feature alone showed a relatively low correlation with survival, the **AdultMale** feature demonstrated a significantly higher correlation with the survival rate. This enhancement allowed us to capture important survival patterns related to adult males, which strengthened our model's predictive power.

In addition to this, we created several features such as **age group**, **family size**, and **family category**. However, we found that these features did not provide any meaningful contribution to the model and decided not to use them. We also manually removed some features with low correlation and explored feature engineering methods like PCA. Ultimately, due to a decrease in model accuracy, we opted not to include them in our final feature set.

For continuous features like **age** and **fare**, we employed various techniques to address outliers and skewness, utilizing methods such as **IQR** and **Z-score**. However, due to the distribution of these features resembling a normal distribution and the presence of numerous outliers, especially in the **fare** variable, these methods did not perform effectively.

Ultimately, to achieve a more normalized distribution, we applied the **quantile transform** method on both features. This approach effectively helped in transforming the distributions closer to normality and for the numerical features in the dataset, we used **StandardScaler** to standardize them.

**Model Implementation**

After preprocessing the data, we divided it into two sets: training and testing, using a ratio of 70% to 30%. This split allows us to train our models on a substantial portion of the data while reserving a portion for testing the model's performance on unseen data.

To build our predictive models, we experimented with several algorithms, including: **Random Forest**, **XGBoost**, **Decision Trees**, **Logistic Regression**

We also utilized the Stacking Classifier to improve the performance of our model. The stacking technique allows us to combine the predictions of several base models and leverage their individual strengths for better predictions.

After developing the initial models, we focused on improving their performance. We did this by performing hyperparameter tuning Using **Grid Search**, which involves adjusting the model parameters to find the best combination for enhancing accuracy and reducing overfitting.

**Result**

The highest accuracy among the individual models belongs to the XGBoost algorithm, which has an accuracy of **87.8%**.

And as expected, the best result obtained from all models is related to the stacking model, which predicts survival with an accuracy of **90%** on the test data.

Based on output classification report of the stacking classifier, we can see, the classification model shows strong and acceptable performance, confirmed by an overall accuracy of 0.90. This means that the model correctly classifies 90% of the cases in the data. The metrics of precision and recall provide more insight into the model's performance:

* Class 0 has a high precision of 0.89 and an excellent recall of 0.95, indicating that the model is successful in identifying true negatives and also correctly identifies the majority of positive cases.
* Class 1, although having a lower recall of 0.82, maintains a strong precision of 0.92, indicating that the model has high accuracy when predicting this class.

The F1 score is 0.92 for class 0 and 0.87 for class 1, showing a good balance between precision and recall. The results indicate a significant improvement in performance compared to any individual model, emphasizing the benefits of ensemble methods in tackling complex classification problems.

**Conclusion**

The analyses conducted on the Titanic data highlight factors that can significantly impact survival status. Utilizing various models and combining them can help improve predictions. Given the low correlation of features with the survival column, this problem has become one of the difficult ones to predict, and finding new features with relatively higher correlation seems to be a suitable approach for further improving prediction accuracy. Additionally, combining models or using more complex models like genetic algorithms and neural networks could be good options for more accurate predictions.