Methodology

Before training the decision tree, there were many steps taken to improve performance and reduce overfitting, including preprocessing and feature selection. Some predictors in the dataset were not realistically related to a passenger's chance of survival and were therefore removed. These included *PassengerId*, *Name*, *Ticket*, *service_id*, *booking_reference*, *name_length*, and *name_word_count*. While these variables may be useful in other contexts, they do not relate to where a passenger was located on the ship or whether they were prioritized for lifeboats. Additionally, the majority of these variables are unique to each passenger, so they are not useful for a classification problem.

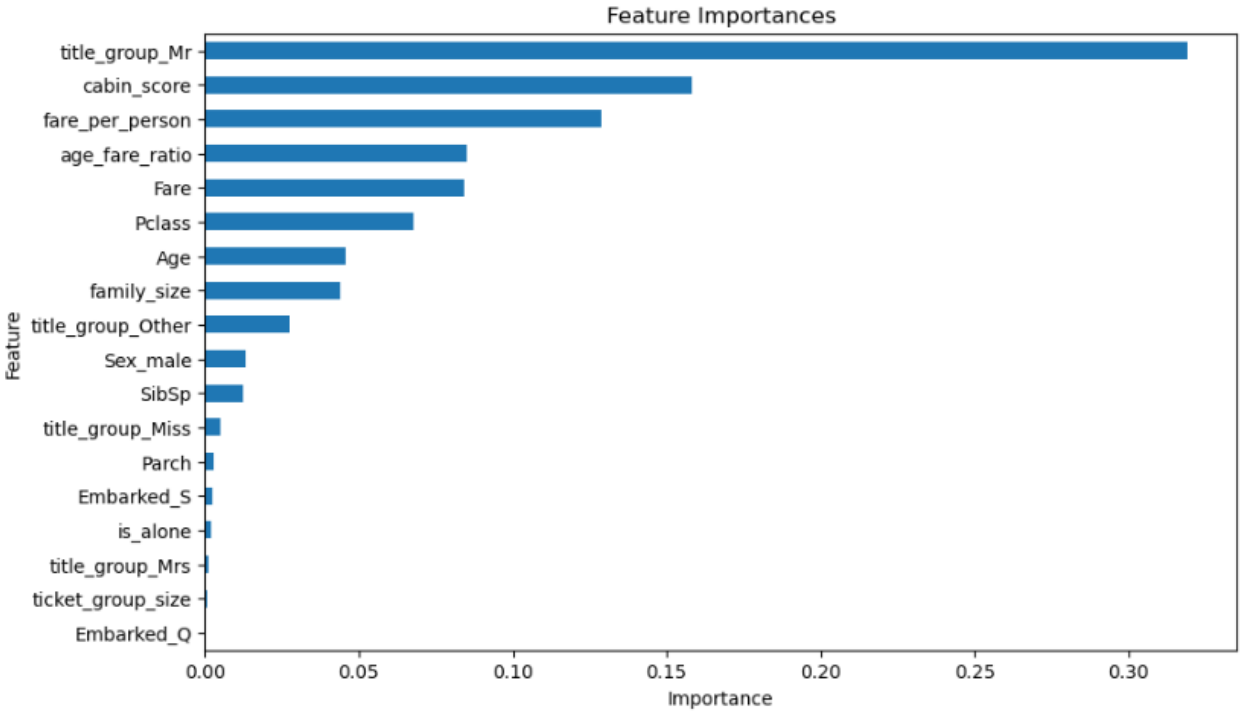
Additionally, some predictors were removed because they contained a large number of missing values that cannot be reasonably estimated. These included *Cabin*, *cabin_room_number*, and *cabin_deck*. Theoretically, these missing values could be filled with the median or mode, but there is no way to verify the accuracy of the estimated value so this would introduce a large amount of inaccurate information into the dataset. The variable *title* was also removed, as it provides the same information as *title_group*.

For predictors that were relevant and contained relatively few missing values, the missing values were filled with either the median or mode. Missing values for *Embarked* were filled using the mode. Missing age values were filled using the median age within each title group (Mr, Mrs, Miss, etc.), which provides a more realistic estimate than using a single overall median. After these steps, no missing values remained.

To preserve class balance and ensure reproducibility, the data was split into 75% training and 25% testing and a random seed of 42 was set. An initial decision tree model was trained using all remaining predictors and the accuracy score was calculated. Feature importance scores were then calculated to help with dropping unimportant features through a trial-and-error process. To choose the appropriate model complexity, the testing different values for maximum tree depth and minimum samples per leaf was conducted. To ensure generalizability and avoid overfitting, cross-validation was conducted on the three trees that shared the same accuracy score. The final tree was chosen due to its generalizability and predictive performance.

Results

After preprocessing and the initial feature reduction, the first decision tree model produced a test accuracy score of approximately 74.44%. Though this indicates a good model, additional steps were taken to improve performance.

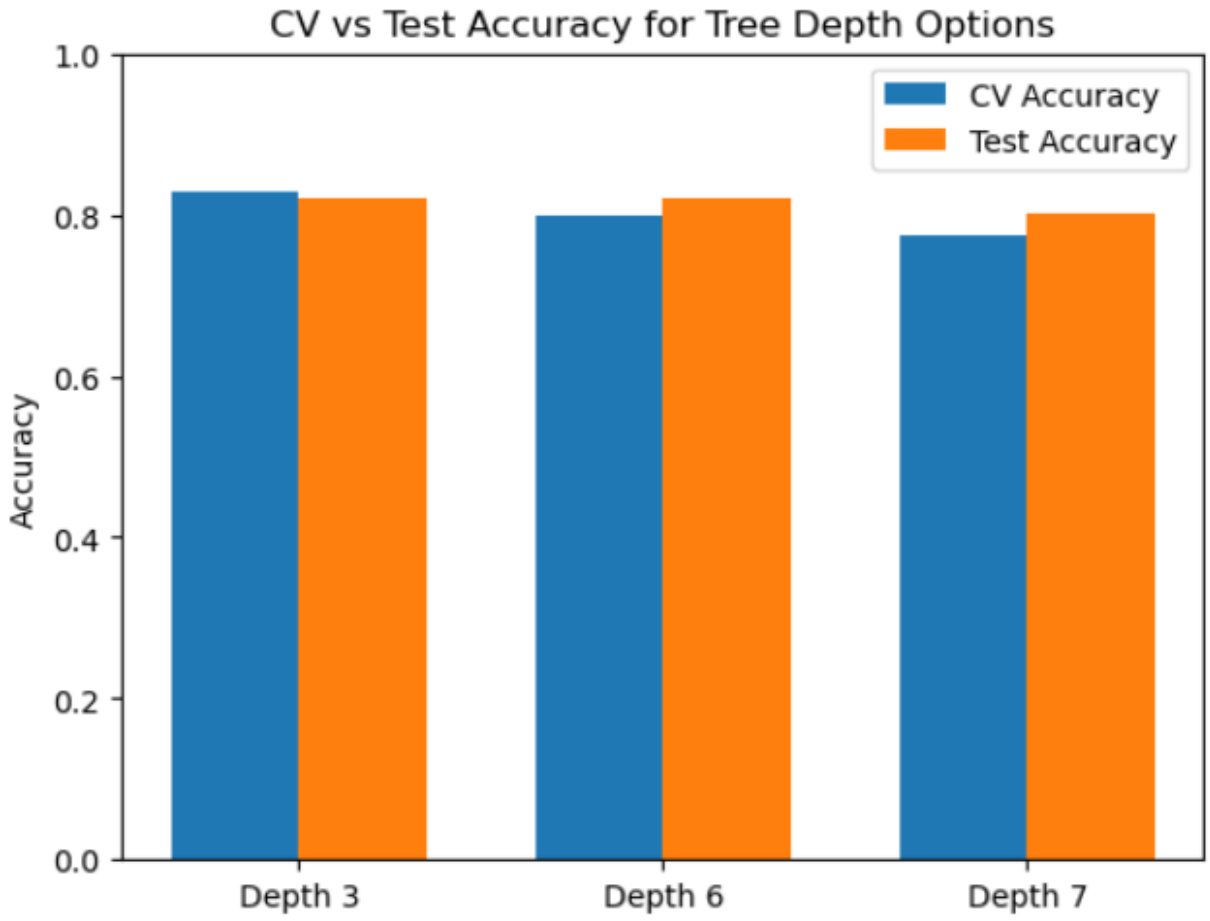


Feature importance scores were examined to determine which predictors contributed the least to the model. As shown in Figure 1, *Embarked_Q*, *ticket_group_size*, and *_title_group_Mrs* are of the lowest importance. Trial-and-error was used to temporarily remove the predictors with low importance while the test accuracy was recalculated until the highest score, 75.34%, was achieved with only these three predictors removed. Removing additional predictors did not improve accuracy and in some cases reduced performance, so no other features were removed.

Next, hyperparameter tuning was performed by training decision trees with maximum depth values ranging from 3 to 8 and minimum samples per leaf ranging from 1 to 10. Test accuracy was calculated for each combination. Trees with the highest test accuracy, 82.06%, were discovered to be at *max_depth* = 6 and *min_samples_leaf* = 1, *max_depth* = 7

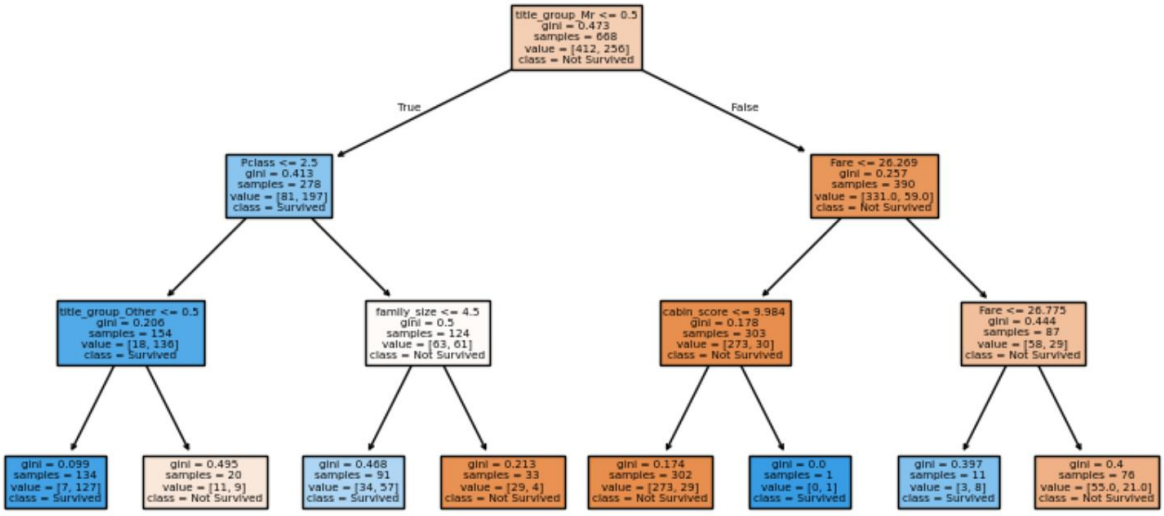
and *min_samples_leaf* = 3, and *max_depth* = 3 and any *min_samples_leaf*. The relationship between maximum depth and test accuracy is shown in Figure 2.

To assess which of these trees performs the best on unseen data, cross-validation was performed on these three models. The cross-validated accuracies for models with *max_depth* 3, 6, and 7 are 82.78%, 79.94%, and 77.39% respectively. Given the mutual test accuracy of 82.06%, this indicates that the model with the smallest difference between test accuracy and cross-validation accuracy is the tree with *max_depth* = 3.



Based on these results, the final model selected was a decision tree with a maximum depth of 3. This model provides a strong balance between accuracy and interpretability.

The structure of the final tree is shown in Figure 4, and its decision rules are summarized in Table 1. From these rules, it is clear that a passenger's chance of survival was mainly influenced by their title group, class, family size, fare price, and cabin score. These patterns align with true historical information, which increases the validity and reliability of this model.



Title Mr	Passenger Class	Title Other	Family Size	Fare	Cabin Score	Survived?
No	≤ 2.50	No	-	-	-	Yes
No	≤ 2.50	Yes	-	-	-	No
No	> 2.50	-	≤ 4.50	-	-	Yes
No	> 2.50	-	> 4.50	-	-	No
Yes	-	-	-	≤ 26.27	≤ 9.98	No
Yes	-	-	-	≤ 26.27	> 9.98	Yes
Yes	-	-	-	26.27-26.77	-	Yes
Yes	-	-	-	> 26.77	-	No