Week 8 Titanic Report

Features created

* Age group

This feature was created to visualize the ages in the dataset better. Instead of using raw age values, grouping ages into categories (e.g., child, teenager, adult, senior) helps simplify the analysis and can capture non-linear relationships between age and the target variable.

* Family Size

Family Size can be derived from the number of siblings, spouses, parents, and children recorded for an individual. It provides a measure of how many people are traveling (or associated) together. This feature helps identify whether individuals in large or small family groups behave differently from those traveling alone.

* isAlone

This binary feature indicates whether the individual is traveling alone or not. Instead of relying on a range of family sizes, this binary feature captures the essence of being alone in a simpler form.

* Titles

This feature was modified to simplify the dataset and enhance the model's ability to generalize by grouping titles into broader, meaningful categories. Some titles (e.g., "Mlle," "Jonkheer") may appear very infrequently in the dataset, leading to sparse data. Sparse categories can introduce noise and prevent the model from effectively learning patterns. As such, combining less common titles into broader categories (e.g., "Mlle" → "Miss," "Jonkheer" → "Mr") increases the frequency of each category, improving the reliability of the data.

Model Selection and Training

Performance of the model before optimization

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score | ROC-AUC |
| Randon Forest | 0.7456 | 0.6458 | 0.7209 | 0.6813 | 0.8475 |
| Ada Boost | 0.7982 | 0.7174 | 0.7674 | 0.7416 | 0.8669 |
| Bagging Classifier | 0.7632 | 0.6739 | 0.7209 | 0.6966 | 0.8169 |
| Logistic Regression | 0.7544 | 0.6471 | 0.7674 | 0.7021 | 0.8464 |
| KNN | 0.7719 | 0.6977 | 0.6977 | 0.6977 | 0.8285 |
| Decision Tree | 0.7105 | 0.6087 | 0.6512 | 0.6292 | 0.6988 |
| SVM | 0.8070 | 0.7333 | 0.7674 | 0.7500 | 0.8795 |
| Naive Bayes | 0.8158 | 0.7115 | 0.8605 | 0.7789 | 0.8903 |

Model Optimization

Performance of the model after optimization

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score | ROC-AUC |
| Randon Forest | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| Ada Boost | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| Bagging Classifier | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| Logistic Regression | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| KNN | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| Decision Tree | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| SVM | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |
| Naive Bayes | 0.815789 | 0.711538 | 0.860465 | 0.778947 | 0.824599 |