Titanic: Machine Learning from Disaster

Data Overview

The data consists of several key attributes per passenger:

- PassengerId
- Survived
- Pclass (Ticket class)
- Name
- Sex
- Age
- SibSp (Siblings/Spouses aboard)
- Parch (Parents/Children aboard)
- Ticket (Ticket number)
- Fare, Cabin (Cabin number)
- Embarked (Port of Embarkation). C = Cherbourg, Q = Queenstown, S = Southampton

Data Preprocessing

1. Missing values are filled using mean or median for continuous variables.

2. Categorical data such as Sex and Embarked are encoded into numerical formats.

3. Relevant features selected for model training include Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked.

Machine Learning Algorithms

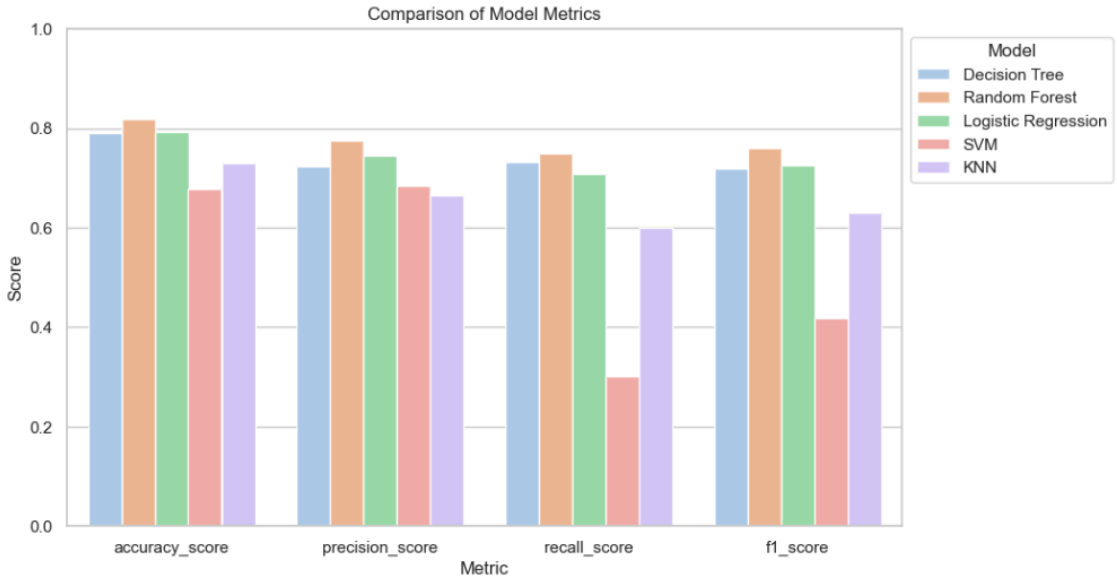
Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, and K-nearest neighbors are the models used.

Model Evaluation

Models are assessed based on accuracy, precision, recall, and F1 score.

Project Results & Evaluation

- **Random Forest** stands out as the most effective model across all metrics for this dataset, suggesting robustness in handling varied data characteristics.
- **Decision Tree** and **Logistic Regression** provide competitive alternatives with decent accuracies and balanced precision-recall scores.
- **SVM** and **KNN** might require parameter tuning or may not be as suitable for the specific characteristics of the Titanic dataset, as evidenced by their lower performance metrics.


Comparison of Model Metrics
Score
1.0
0.8
0.6
0.4
0.2
0.0
accuracy_score
precision_score
recall_score
f1_score
Metric
Model
Decision Tree
Random Forest
Logistic Regression
SVM
KNN