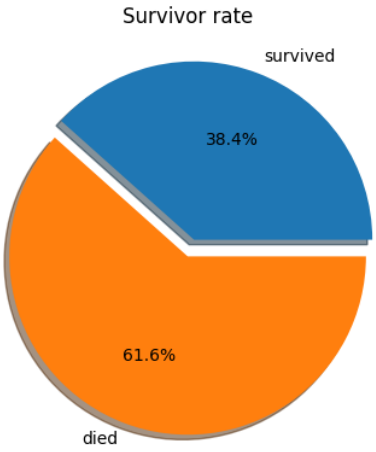
Titanic Survivors

Jens Kobler

February 3, 2025



Contents

1 Main Objective of Analysis

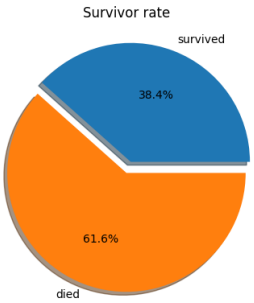
The main objective of this analysis is to find the best machine learning model for the competition "Titanic - Machine Learning from Disaster" for the stakeholders using only standard models (no hyperparameter optimization). It is assumed that the best model is the model with the highest F1-Score, since this metric balances the precision and recall.

2 Description of Dataset

- Name of dataset: Titanic - Machine Learning from Disaster
- Link to dataset: `https://www.kaggle.com/competitions/titanic/data`

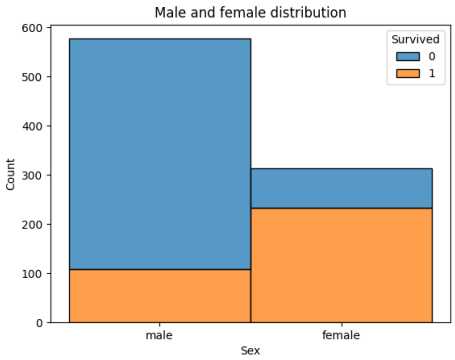
Target column

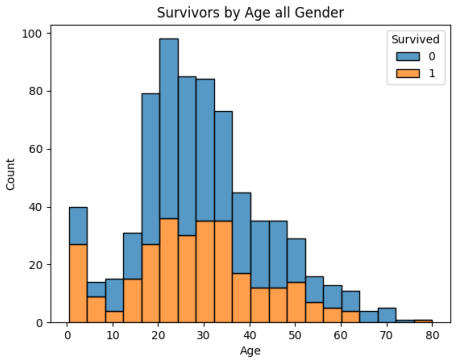
- Survived (1 = True, 0 = False)



Feature columns:

- *PassengerId*
- *Pclass*
- *Name*
- *Sex* (male and female)
- *Age*
- *SibSp*
- *ParCh*
- *Ticket*
- *Fare*
- *Cabin*
- *Embarked*





3 Data Cleaning and Feature Engineering

Data Cleaning

1. NaN age numbers to mean of age
2. delete *Cabin* column
3. delete *Name* column
4. delete *PassengerId* column
5. delete *Ticket* column

Feature Engineering

1. *Embarked* column to *C*, *Q*, *S* columns
2. *Sex* column to *female*, *male* columns
3. convert bool values (True, False) to int values (1,0)
4. re-scale features between 0 and 1 using the **Min-Max-Scaler** from *sklearn.preprocessing*, namely *MinMaxScalar*.
5. a stratified train-test split is done with test_size = 0.2

4 Variation of Classifier models

4.1 Logistic Regression

The standard logistic regression algorithm from the sklearn.linear_model library was used with solver *liblinear*, namely *LogisticRegression*.

4.2 K Nearest Neighbors

The standard k nearest neighbors algorithm from the sklearn.neighbors library was used with number of neighbors 10, namely *KNeighborsClassifier*.

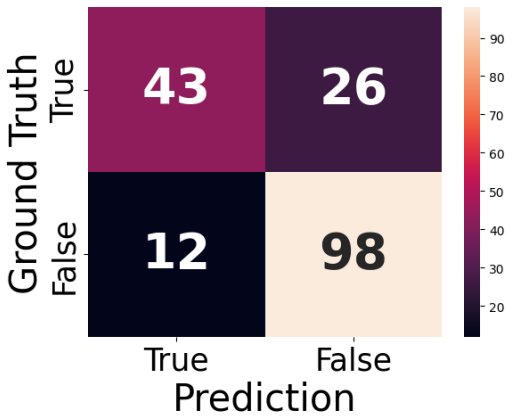


Figure 1: Result of classification for knn

4.3 Support Vector Machines

The standard support vector machines algorithm from the sklearn.svm library was used, namely *SVM*.

4.4 Decision Trees

The standard decision tree algorithm from the sklearn.tree library was used, namely *DecisionTreeClassifier* (DTC).

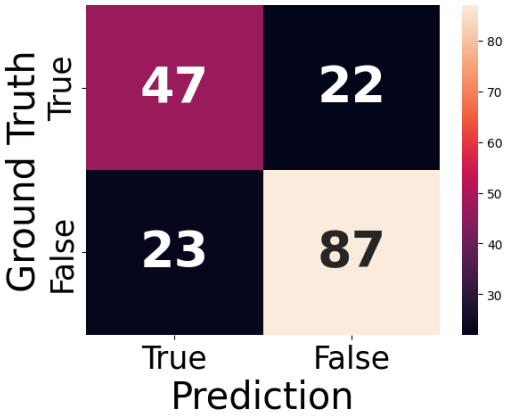


Figure 2: Result of classification for standard Decision Tree Classifier

Optimizing the DTC

The DTC is optimized via GridSearchCV.

4.5 Final Results

	Accuracy	Precision	Recall	Specificity	F1-Score
LR-standard	0.799	0.762	0.696	0.864	0.727
KNN-10-standard	0.788	0.782	0.623	0.891	0.694
SVM-standard	0.81	0.857	0.609	0.936	0.712
DTC-standard	0.749	0.671	0.681	0.791	0.676
DTC-optimized	0.805	0.804	0.652	0.9	0.72

The logistic regression model achieves the highest F1-Score.

5 Flaws in model and Revisiting Analysis

- You could include the following columns:
  - Name
  - Cabin
  - PassengerId
  - Ticket

- instead of using the mean of the age as replacement of the NaN values in the age column, you could use another approach
- improve models by hyperparameter optimization
  - SVM with regularization
  - for LR use other penalty term
- use Voting (Ensemble methods)
- you could start interpretation