REPORT

This is a report of the titanic dataset in predicting the survival of passengers aboard.

By Andrew Marfo

DATA EXPLORATION

The train dataset consists of 712 individual row entries and 12 columns representing the features. The columns are:

1. PassengerId: It has 712 non-null values and is an integer data type.
2. Pclass: It has 712 non-null values and is an integer data type.
3. Name: It has 712 non-null values and is an object data type.
4. Sex: It has 712 non-null values and is an object data type.
5. Age: It 575 non-null values and is a float data type.
6. SibSp: It has 712 non-null values and is an integer data type.
7. Parch: It has 712 non-null values and is an integer.
8. Ticket: It has 712 non-null values and is an object data type.
9. Fare: It has 712 non-null values and is a float data type.
10. Cabin: It has 160 non-null values and is an object data type.
11. Embarked: It has 710 non-null values and is an object data type.
12. Survived: It has 712 non-null values and is an integer.

There were two (2) unique values for Sex which is a categorical data. The values were "male" and "female".

There were also (3) unique values for Embarked. The values were "S", "C" and "Q".

The statistics for the numerical data can be found in the image below

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Survived
count	712.000000	712.000000	575.000000	712.000000	712.000000	712.000000	712.000000
mean	444.405899	2.308989	29.807687	0.492978	0.390449	31.819826	0.383427
std	257.465527	0.833563	14.485211	1.060720	0.838134	48.059104	0.486563
min	1.000000	1.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	222.750000	2.000000	21.000000	0.000000	0.000000	7.895800	0.000000
50%	439.500000	3.000000	28.500000	0.000000	0.000000	14.454200	0.000000
75%	667.250000	3.000000	39.000000	1.000000	0.000000	31.000000	1.000000
max	891.000000	3.000000	80.000000	8.000000	6.000000	512.329200	1.000000

The count represents the total number of counts for the columns.

The mean represents the mean value.

The std represents the standard deviation.

The min represents the minimum value (i.e the smallest number).

The 25% represents the 25th percentile.

The 50% represents the median which is also known as the 50th percentile.

The 75% represents the 75th percentile.

The max represents the maximum value (i.e the largest number).

DATA CLEANING AND PREPROCESSING

Handling missing values:

There were 3 variables (Columns) which had missing values. The variables were Age, Cabin and Embarked.

The

The missing values in Age was replaced with the median because the distribution of this variable was skewed.

The missing values in Cabin was replaced with the modal value because it is a categorical variable, and the distribution also looked like a normal distribution.

The missing values in the Embarked variable was dropped. This was because the values represented a relatively small percentage of the dataset (0.28 %). The percentage is negligible so dropping it won't affect our model.

Encoding Categorical Variable:

Label Encoder was used to encode the categorical variables to numeric values

Normalization:

I used the standard scaler to normalize the features of the dataset. Standard scaler ensures the data have a mean value of zero and a unit variance.

Splitting the dataset into training validation:

The proportion for splitting the dataset into training and validation was 80:20. Having a reasonable large number for training improves the model.

FEATURE ENGINEERING

I generated 3 new features to help improve the model's accuracy when predicting the survival. These features are

1. Family Size: The number of people travelling together as a family. The family size is important because a passenger will always try to save a relative.

2. Is Child: A Boolean defining weather the passenger is a child or not. In situations like this, children are highly prioritized so the rate at which children survive is higher than that of adults.
3. Economic Status: How wealthy that passenger is. Being rich has certain advantages.
4. Title: The title of the passenger can also play role here. Title like Dr. for example will have a higher priority than Mr.

I dropped PassengerId and Name. This is because PassengerId and Name were just identifiers to uniquely identify passengers. It did not have any important influence on the target per my observations.

MODEL SELECTION AND TRAINING

Models trained

Five models were trained for the train.csv dataset. The models are:

1. Logistic Regression
2. Random Forest Classifier
3. K – Nearest Neighbors (KNN)
4. Support Vector Machines (SVC)
5. Decision Tree Classifier

Metrics Report Comparison

I used classification report to evaluate 4 metrics of all the models. Below is the table of the performance of the models on the validation set

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.77	0.76	0.75	0.76
Random Forest Classifier	0.85	0.85	0.84	0.84
K-Nearest Neighbors	0.81	0.81	0.78	0.79
Support Vector Classification	0.83	0.83	0.81	0.82
Decision Tree	0.82	0.81	0.80	0.80

From the table, Random Forest Classifier was the model with the highest Accuracy (0.85).

Model Optimization

I used the GridSearchCv to optimize all the models to improve their performance on the dataset.

Metrics Report Comparison on Optimized Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.77	0.76	0.75	0.76
Random Forest Classifier	0.87	0.86	0.86	0.86
K-Nearest Neighbors	0.83	0.83	0.81	0.81
Support Vector Classification	0.83	0.83	0.81	0.82
Decision Tree	0.83	0.83	0.80	0.81

The report shows an increase in performance of the Random Forest Classifier Model. The parameters of the best model were:

1. 'class_weight': 'balanced_subsample',
2. 'criterion': 'entropy',
3. 'max_depth': None,
4. 'max_features': 'sqrt',
5. 'min_samples_leaf': 2,
6. 'min_samples_split': 10,
7. 'n_estimators': 50

The accuracy was 0.87 with a precision of 0.86

This model was used to predict the survival of the test.csv dataset. The predicted result can be found in `Andrew_submission.csv` file which is the same directory as this.