Titanic - Machine Learning from Disaster

The task: The sinking of the Titanic is one of the most infamous shipwrecks in history. Unfortunately, in that incident there weren't enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew. Our aim in Titanic Dataset is to predict the survival status of passengers.

What type of problem: Since we have to find the given passenger survived or not, thus this is a categorical problem which goes to classification.

Dataset: The dataset consists of in total 12 features**.** Namely: ['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'].

#	Column	Non-Null Count	Dtype
0	PassengerId	891 non-null	int64
1	Survived	891 non-null	int64
2	Pclass	891 non-null	int64
3	Name	891 non-null	object
4	Sex	891 non-null	object
5	Age	714 non-null	float64
6	SibSp	891 non-null	int64
7	Parch	891 non-null	int64
8	Ticket	891 non-null	object
9	Fare	891 non-null	float64
10	Cabin	204 non-null	object
11	Embarked	889 non-null	object

Data Analysis using Visualization: We first have to find out what features are affecting the survival columns. So for this, we find the correlation between each feature and the target column. Thus by checking we found that the Pclass, Sex, SibSp, Parch and embarkment were correlated to the survival column. Few observations from data analysis are listed below:-

- The majority of passengers were in Pclass=3, however, the majority did not survive. Confirms our classifying assumption. The majority of infant passengers in Pclass=2 and Pclass=3 survived. This qualifies our classification assumption even further. The majority of passengers in Pclass=1 made it out alive.
- Female passengers scored significantly better than male passengers in terms of survivability. Confirms the classification. Males exhibited a greater survival rate than females in Embarked=C. This might indicate a link between Pclass and Embarked, and then Pclass and Survived, rather than a straight link between Embarked and Survived.

Pclass=3 and male passengers had different survival rates depending on the port of departure.

- Unwanted features were dropped off.

The algorithm used: Decision tree classifier, Decision tree works by breaking down the dataset into small subsets. This breaking down process is done by asking questions about the features of the datasets. The idea is to unmix the labels by asking fewer questions necessary. As we ask questions, we are breaking down the dataset into more subsets. Once we have a subgroup with only the unique type of labels, we end the tree in that node. For Eg let's say we are asking to find out whether a certain person in the test dataset survived or not. We may ask a question like, is the person "male" or "female." Let's say the answer is "female." Then the algorithm might ask about the person's P-class. And likewise, it will compute further.

Reason for using this algorithm: It's easy to comprehend and interpret. It is possible to visualize trees. Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
The cost of using the tree that is predicting data is logarithmic in the number of data points used to train the tree. Able to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable.

Parameters: Used GridSearch, the hyperparamerter chosen are: criterion='entropy', max_depth=8, max_features='auto'

Accuracy: 89% on validation set.(Score on Kaggle- 0.78)