Machine Learning Project – In this article, I will take you through a very interesting case study for machine learning practitioners which is to predict titanic survival with Machin Learning. I will first introduce you to this case study and then I will show you how we can build a predictive model to predict survival with Machine Learning.

**Machine Learning Case Study: Titanic Survival Analysis**

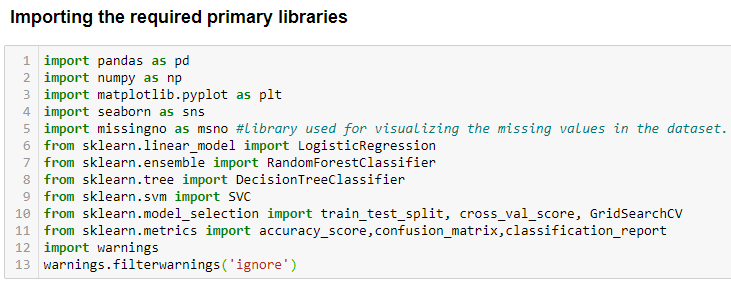
The sinking of the Titanic is one of the most infamous wrecks in history. On Aprli5, 1912, during her maiden voyage, the RMS Titanic, widely considered ‘unsinkable’, sank after hitting the iceberg.

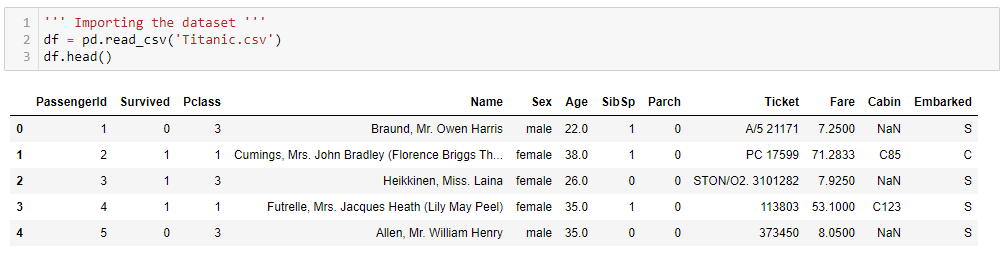
Unfortunately, there were not enough lifeboats for everyone on board, resulting in the deaths of 1,502 out of 2,224 passengers and crew. While there was an element of luck in survival, it appears that certain groups of people were more likely to survive than others.

Here, our challenge is to build a predictive model that can give a solution to the question, “What types of people were more likely to survive?” using passenger data (i.e. name, age, sex, social-economic class, etc.).

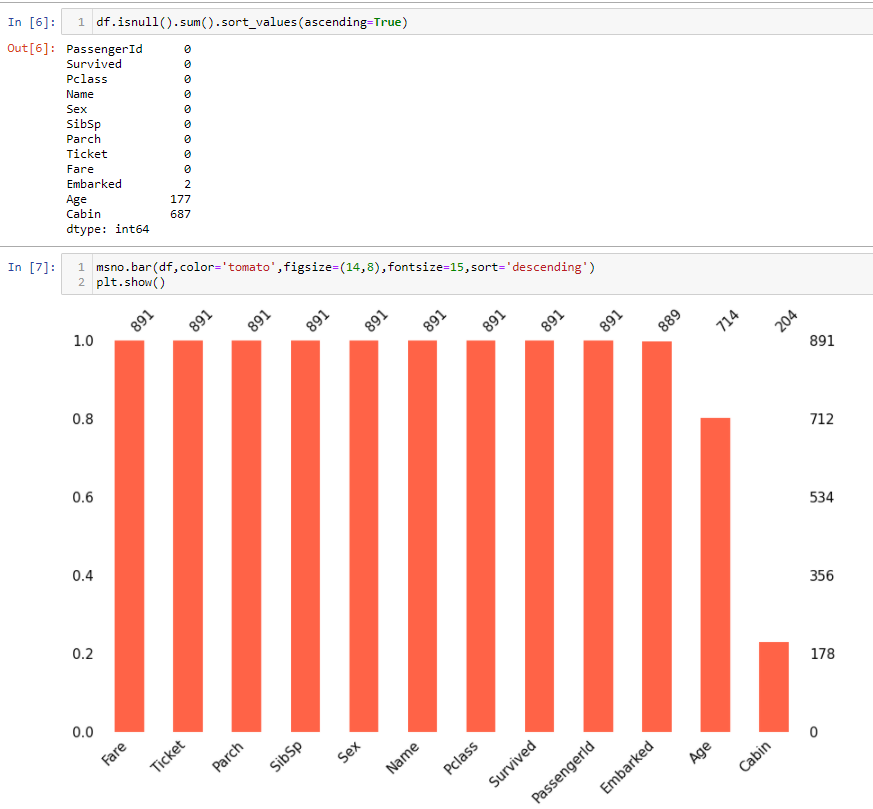
**Predict Titanic Survival with Machin Learning**

Now, as a solution to the above case study for predicting titanic survival with machine learning, I’m using a Titanic Dataset, which relates to passenger survival rates on the Titanic, which sank in 1912. I’ll start this task by loading the test and training dataset using pandas:



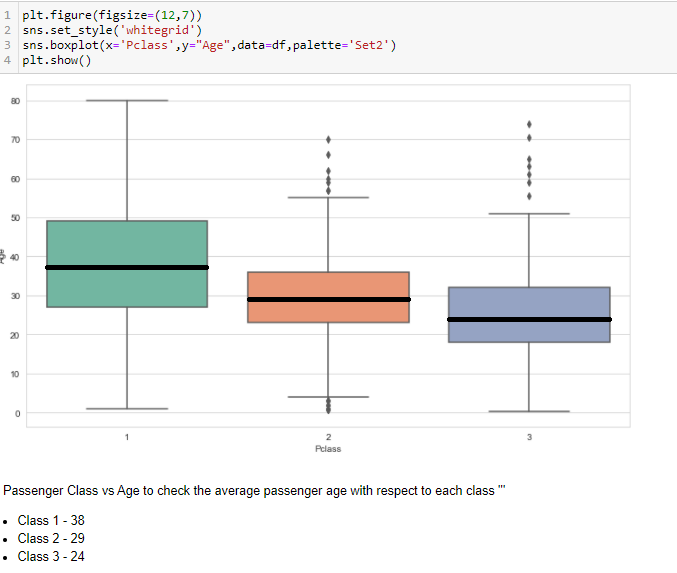


Scikit-learn’s algorithms generally cannot be powered by missing data, so I’ll be looking at the columns to see if there are any that contain missing data:

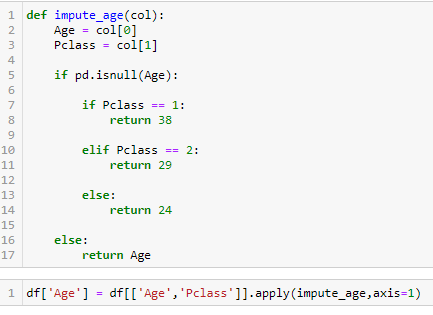


**Handling the Missing Values (Age)**

“We need to fill all the missing values in Age as we notice some sort of relationship between Age and other features. One way to do this is by filling in the mean age of passenger”

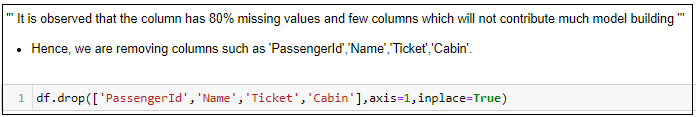


Here, we found mean value for all the P-classes which we are going to replace with the missing values.



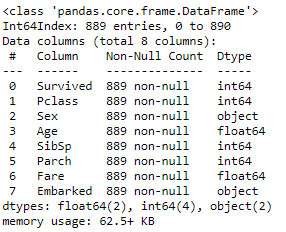
**Feature Engineering**

Now, we handled missing values in Age column, it’s time to drop couple of attributes which are not necessary to build the model. The columns that we are going to drop are PassengerID, Name, Ticket, Cabin (cabin has 80% missing value so it’s better to drop it rather filling missing values).



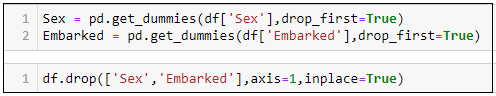
**Handling categorical Features**

We have 2 object datatypes Sex and Embarked which needs to be converted as numerical values before we build model for our datasets.

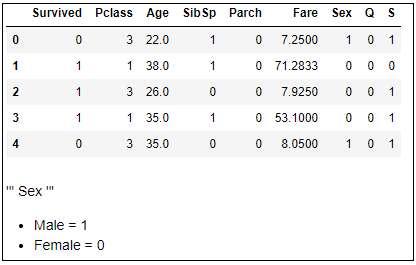
****

Let’s create the dummies (It adds a column by splitting the unique values in the referred column) for Sex and Embarked variables.

Post creating the dummies we need to drop the parent columns from the dataset.



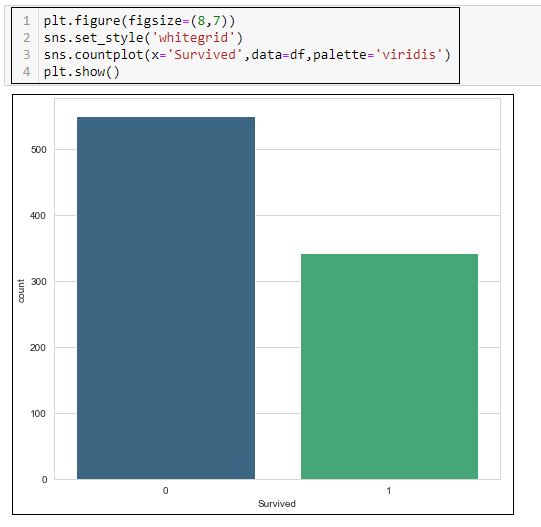
Now after all the EDA this is how the dataset looks like



**Visualization**

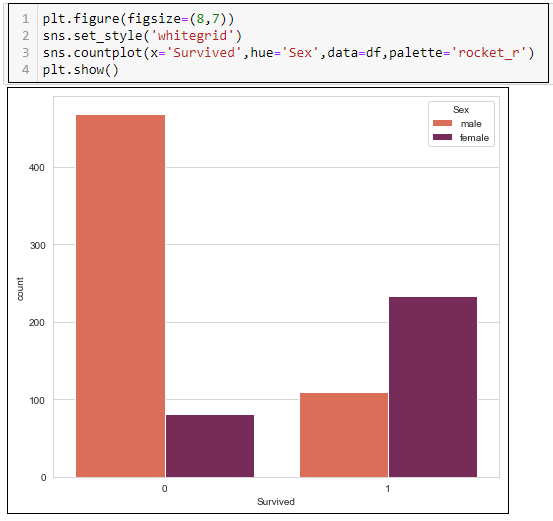
Let us check for the data visualization to under the dataset structure.

Survivors

****

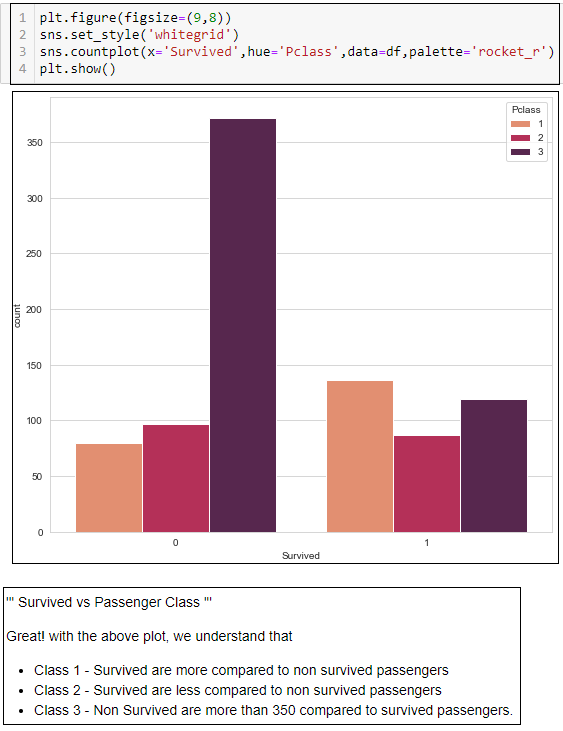
Here, the dataset i.e. target variable is quite balanced which is a good sign for us in building a model.

Survived vs Sex

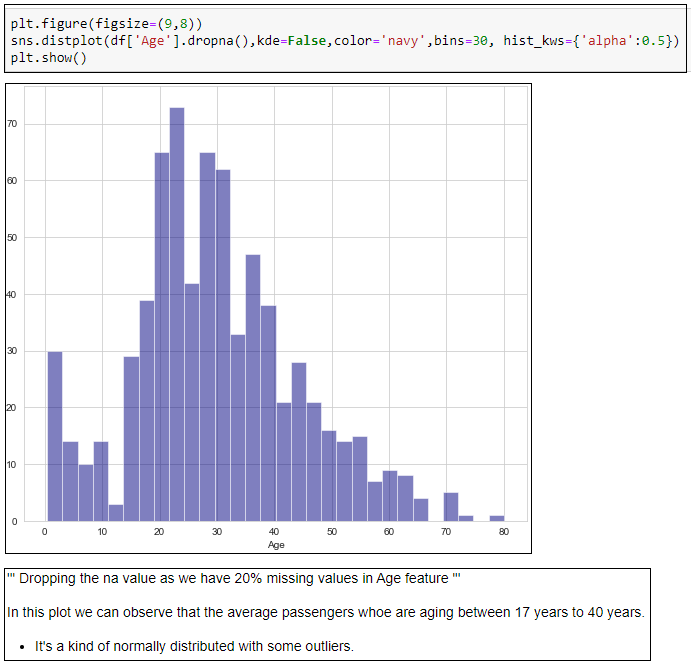


With the above plot we can understand that the female survivors are more compared to male survivors.

Survived vs Pclass

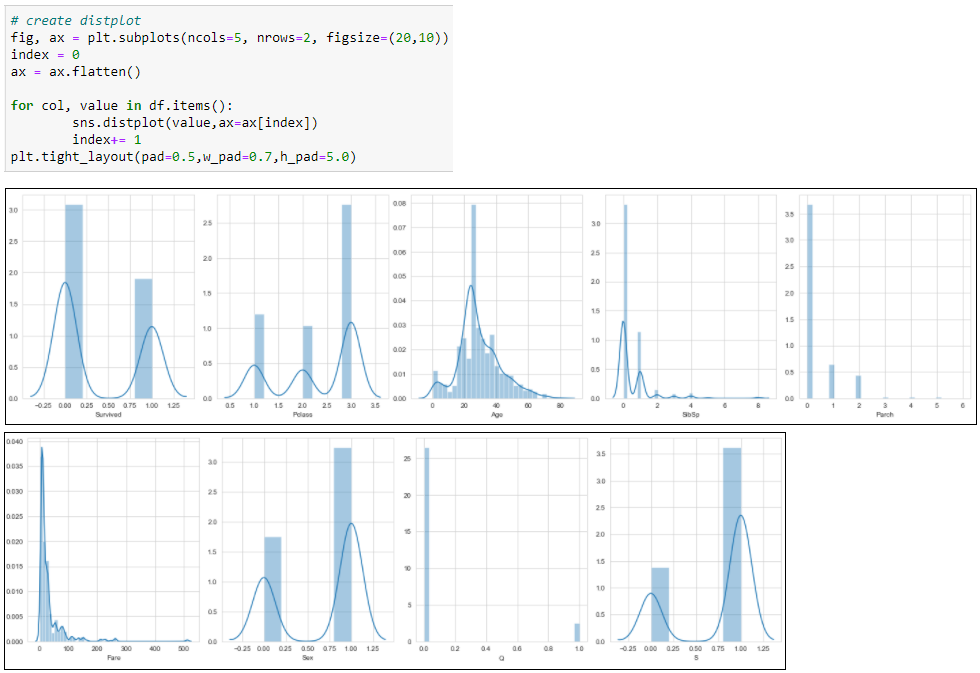


Age

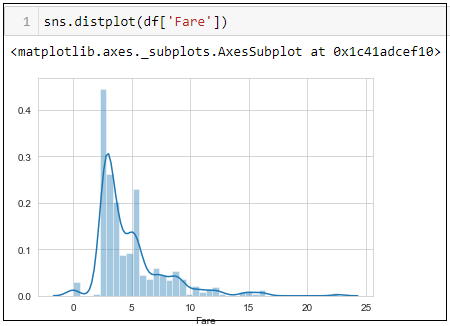


The above plot helps us to understand that age category among all the passengers in Titanic Ship.

Distribution Plot

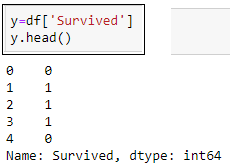
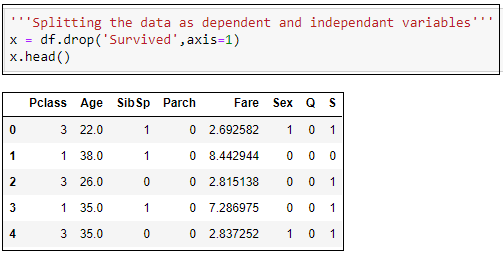


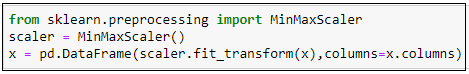
Here, we can observe that the data is quite normally distributed except feature Fare, Hence, we need to treat the column Fare.



Now the data is ready for the Model Building

* We also need to process scaling the dataset for better prediction.

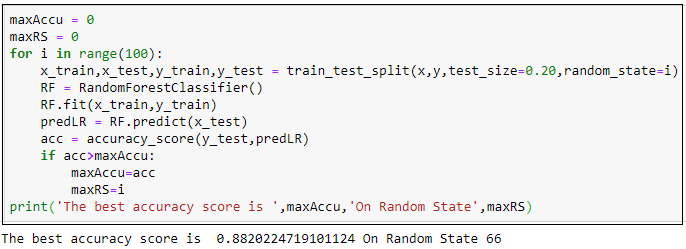




Finally, we have completed scaling the dataset X, which doesn’t include Target variable.

**Machine Learning Model to Predict Titanic Survival**

**Finding the best Random State**

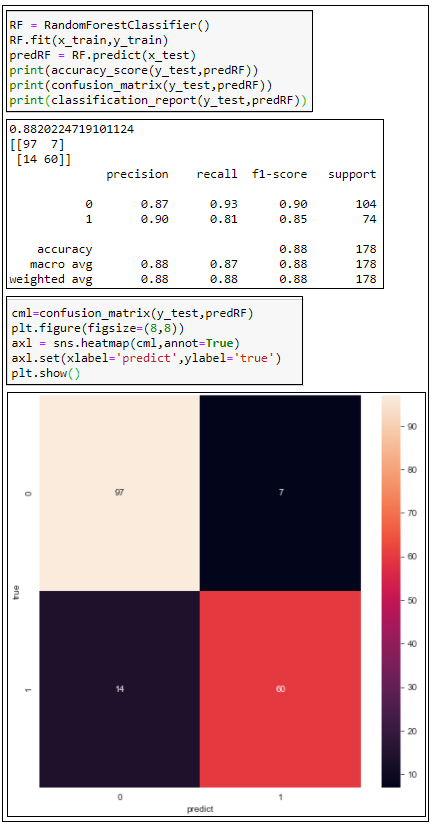
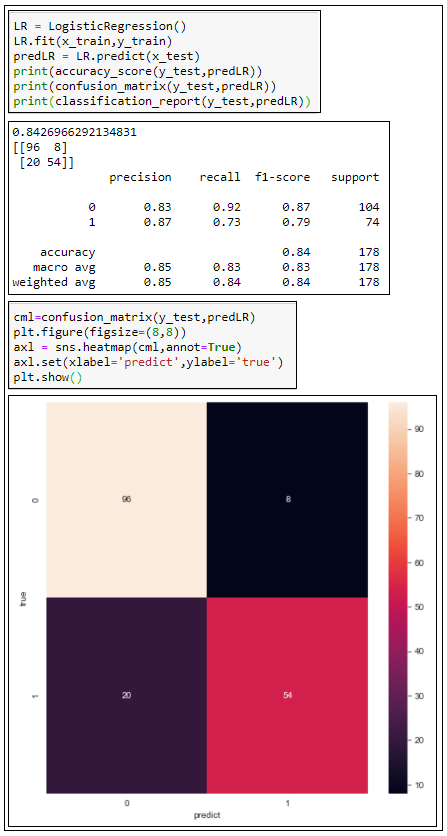
****

The best random state for this dataset is 66

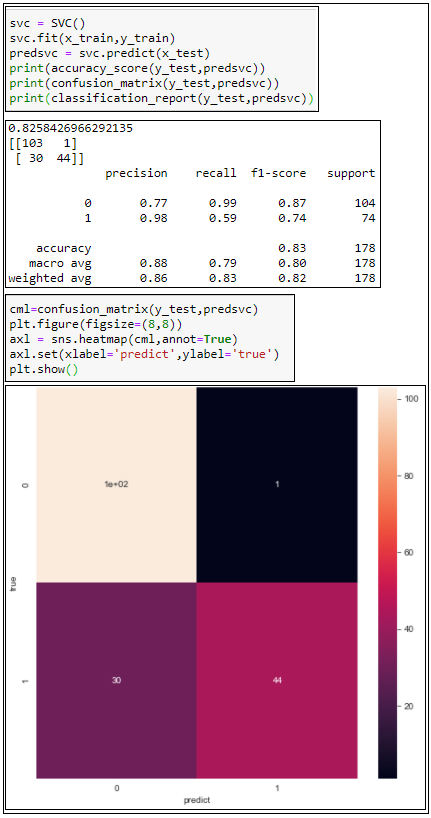
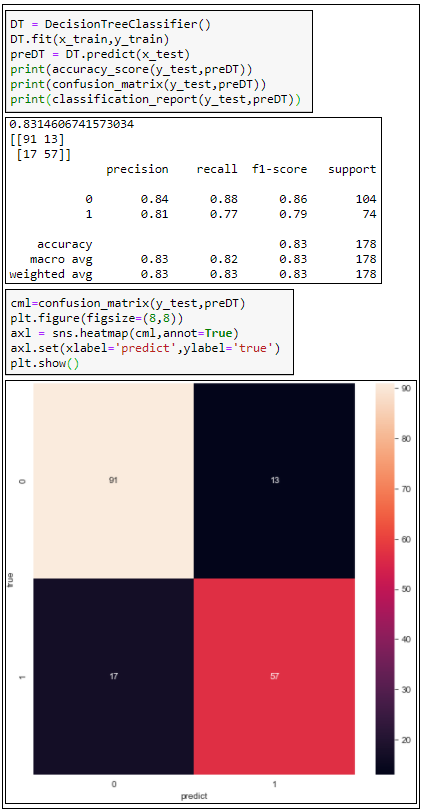
Now, I’m going to use the datasets to build models with the below modelling techniques

* LogisticRegression
* RandomForestClassifier
* DecisionTreeClassifier
* SupportVectorMachine

Logistic Regression & RandomForestClassifier



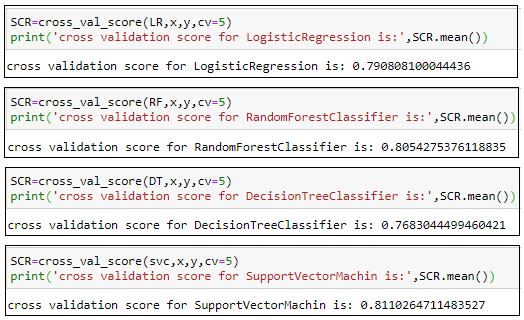
DecisionTreeRegression & SupportVectorMachine



With the above model building, we get to know that the Support vector machine is the best model with the accuracy score of 83%.

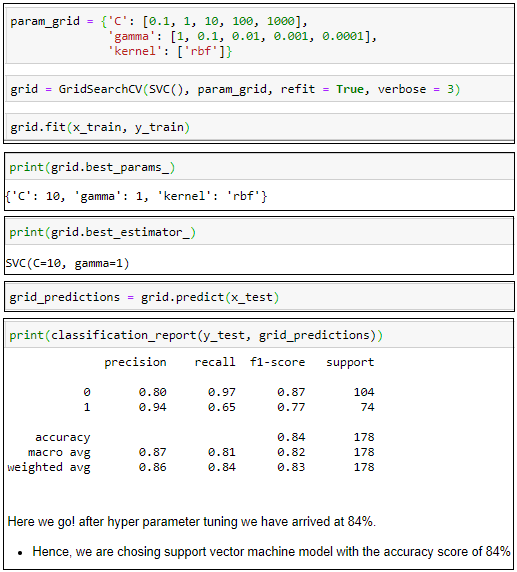
**Cross-Validation Score**

Now we need to check for cross validation score, in order to understand that none of the models are over fitted or under fitted.

****

**HyperParameter Tuning:**

Hyper parameter is used to enhance the accuracy score for every model. Hence, I am using the parameter tuning for support vector machine model to improve the accuracy score.



We could see that through Hyper parameter tuning we have increased the score 84% which is not a bad score to start into Data science world.

Conclusion

We made the entire journey in a small data science project. We explored the data, cleaned up the data, then we modified features and created new ones and in the last step we made a prediction with a support vector machine.

But there is still a lot to do, next you can test the following things?

* Try other Algorithms which can perform better?
* Can you chose the bins for Age and Fare Better?
* Can the ticket variable be used?
* Is it possible to further adjust the survival rate?
* Do we really need to keep all the features or is it okay we excluded in this dataset.