**Prepared for: ICT Academy at IITK**

**Prepared** **by: Paras Laul, Kkishita Jain, Aditya Narayan Singh**

**Project** **Name**: Machine Learning from Disaster

Predicting the Titanic Survival Rate

**20 June 2020**

# **Introduction**

The sinking of the Titanic is one of the most historic shipwrecks of all time. The tragedy killed thousands, 1502 out of 2224 passengers, and led many wondering what could have been done better. One of the most important reason is that there was not enough lifeboats, and although there was probably quite amount of luck involved, there were some groups of people that were more likely to survive than others. In this paper a data analytical study will be conducted with the passenger’s data from the Titanic dataset to find out about this survival likelihood. For the data analytical approach, we apply the theory of machine learning and predicting the survival of passengers by applying Machine Learning models.

## Requirements

1. Jupyter Notebook
2. The following libraries are required to run the code in Jupyter Notebook

* Pandas
* Numpy
* Seaborn
* Matplotlib
* Sklearn

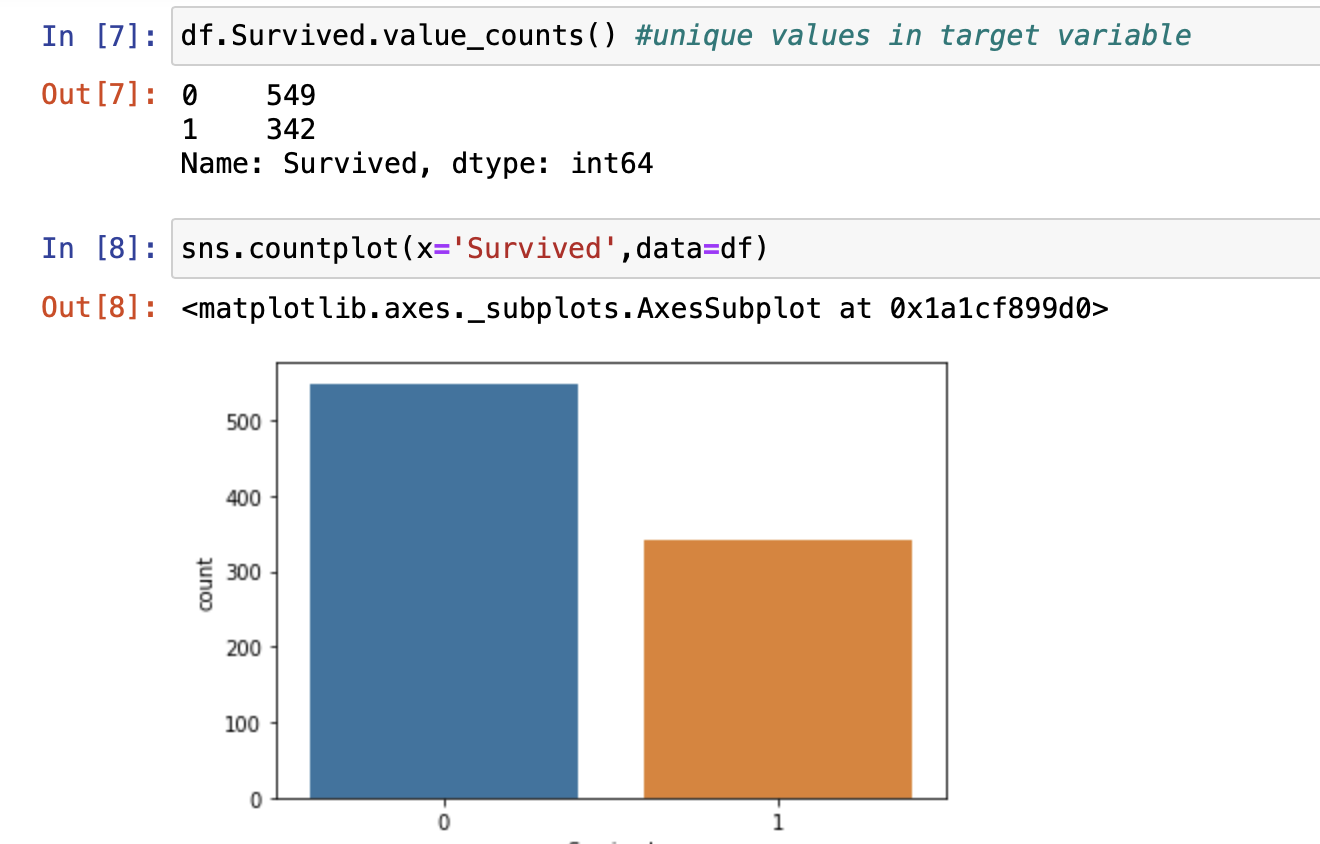
**Procedure**

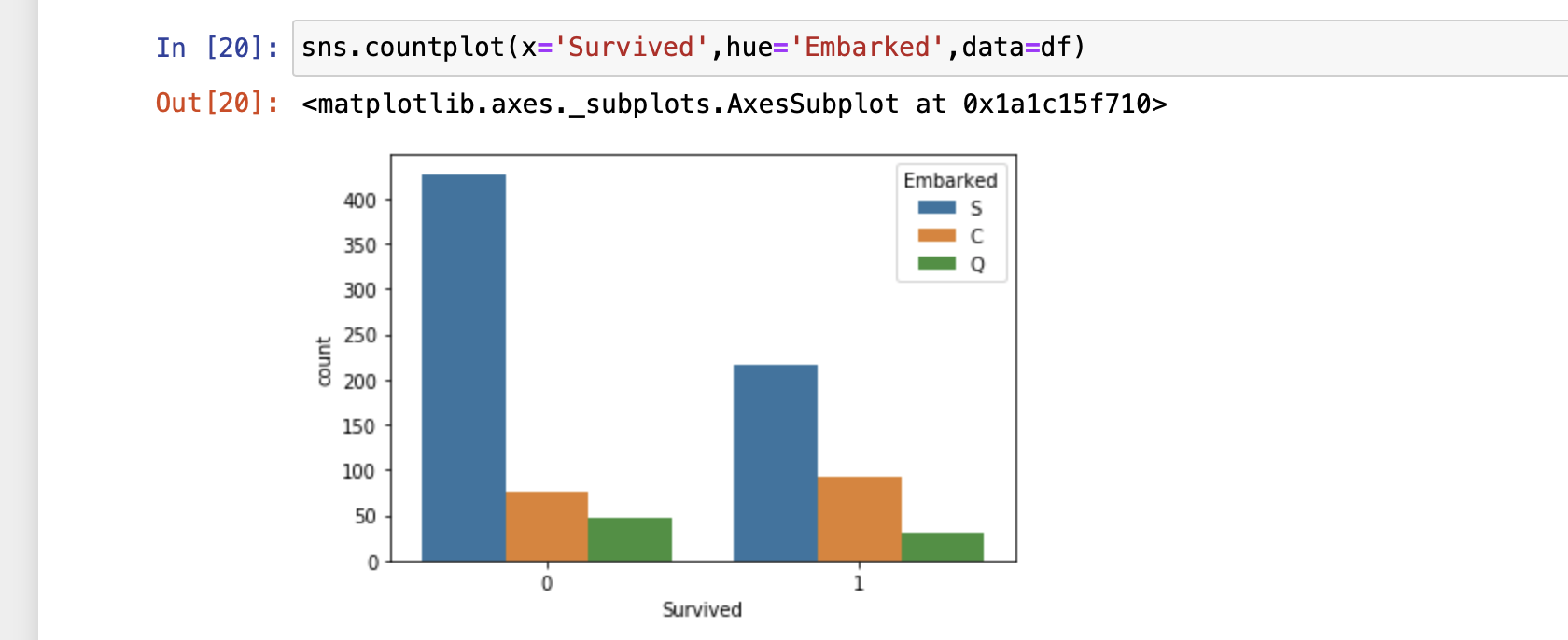
All the Lifecycle In A Data Science Projects¶¶

* Data Analysis
* Feature Engineering
* Feature Selection
* Model Building
* Model Deployment

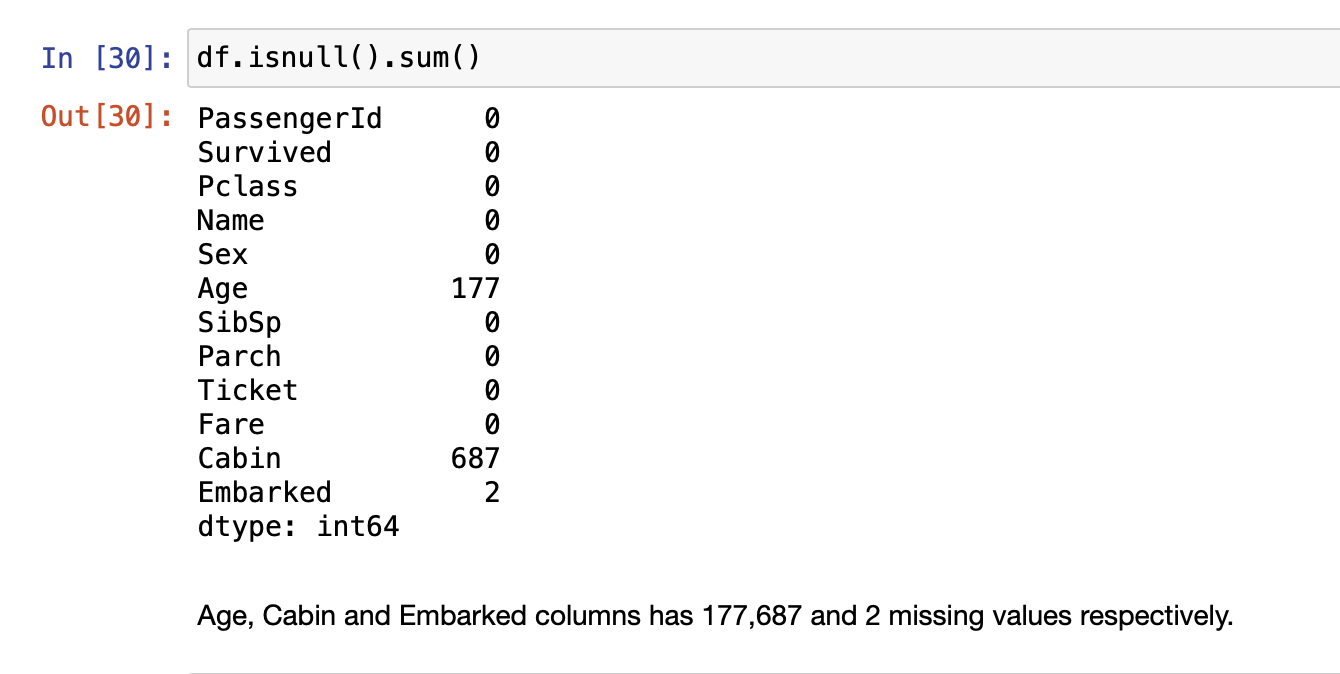
1. **Data Analysis**

In Data Analysis we will find the relationship of different features of titanic dataset with survived attribute and visualise these relationships using seaborn library.

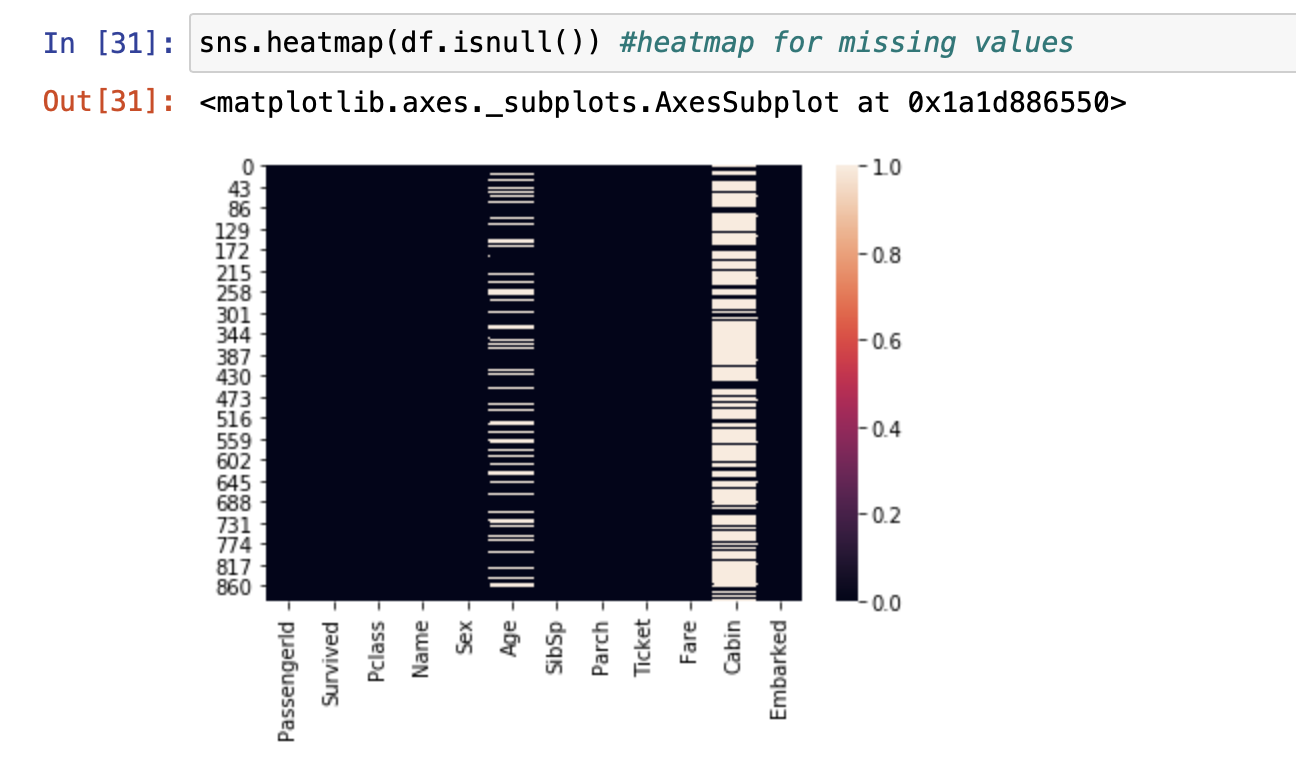
* Find the count for passenger who survived. Here 0 stands for Survived and 1 stands for not Survived.
* Survived and Sex, Survived and PClass, Survived and Parch, Sex and Embarked



2. **Data wrangling**

Processing the data and removing missing values in dataset by either substituting them with mean, mode, median or most frequent value.

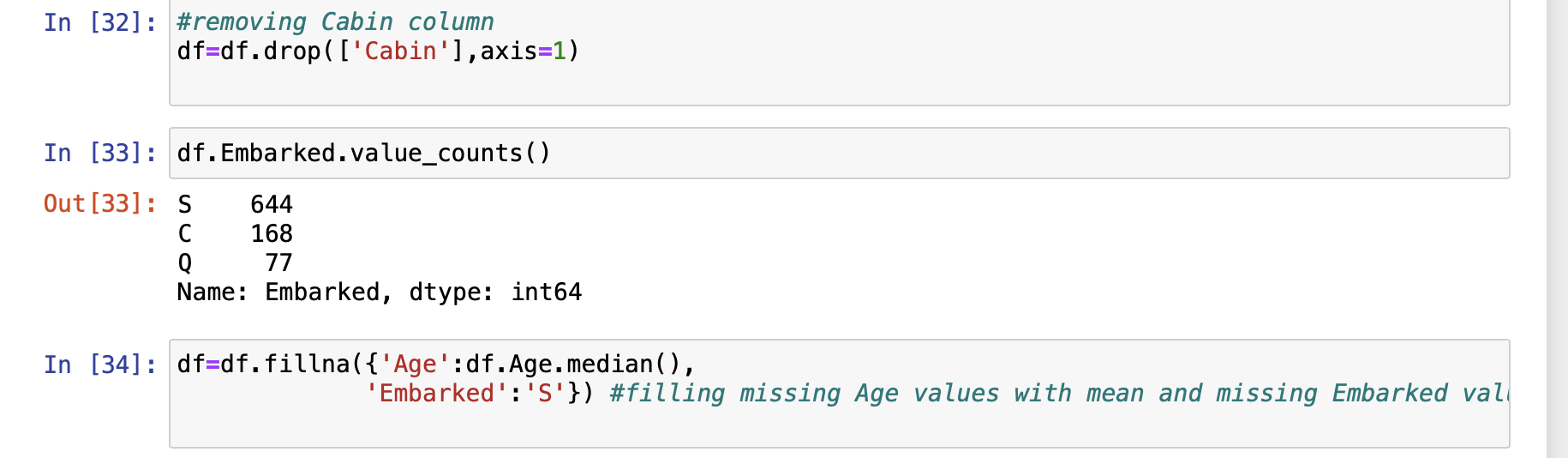
* Age has 177 missing values whereas Cabin and Embarked has 687 and 2 missing values respectively

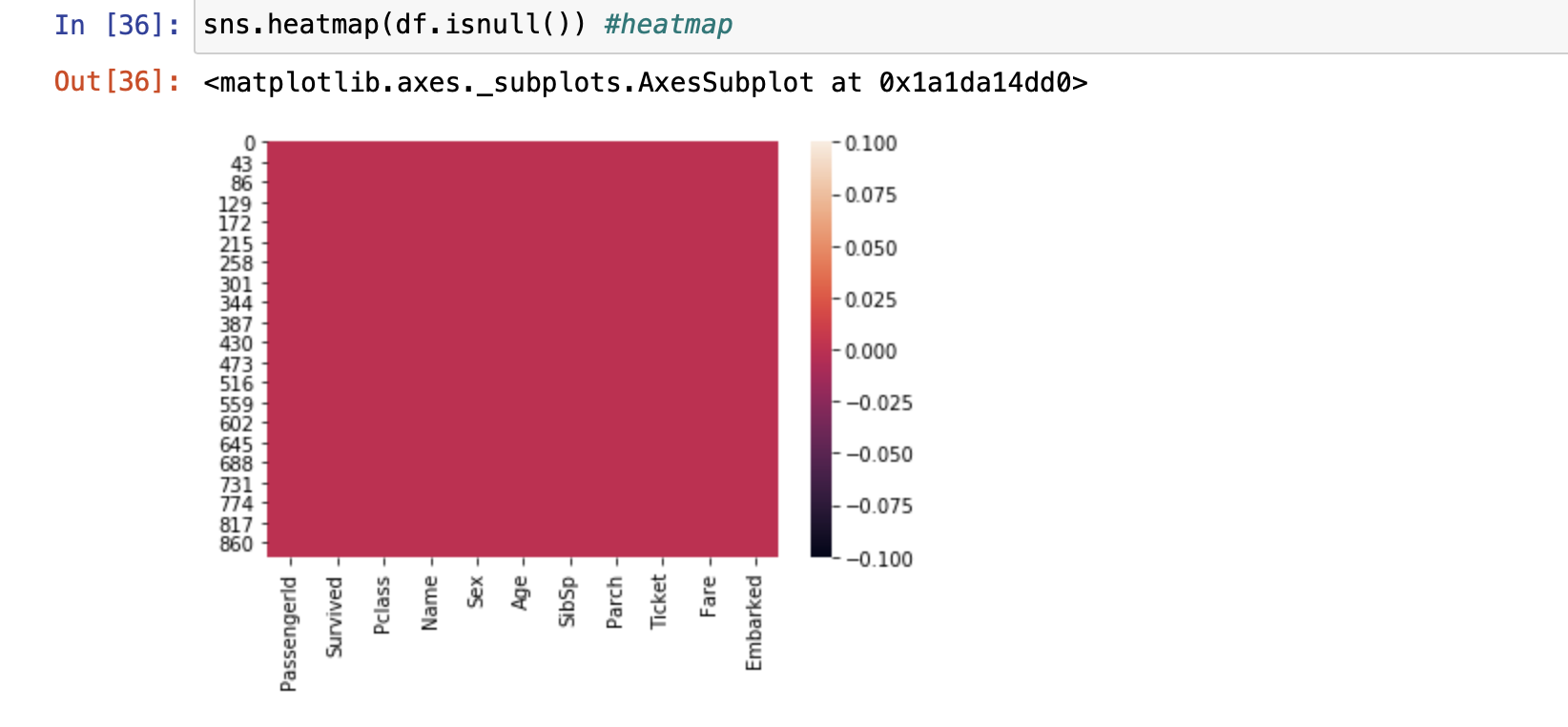


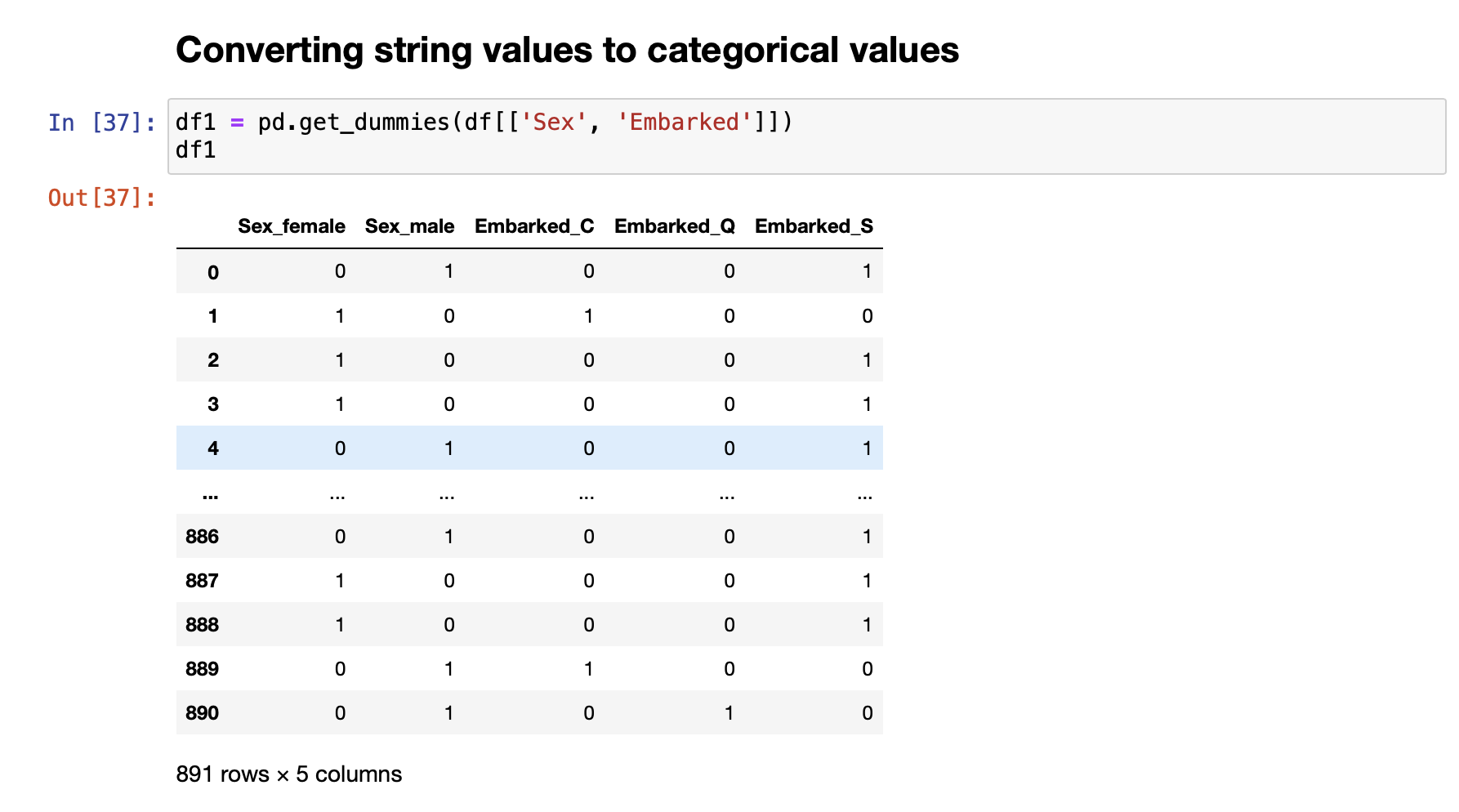
* Heat Map for missing values. Heat map shows that Cabin has maximum no. of missing values.

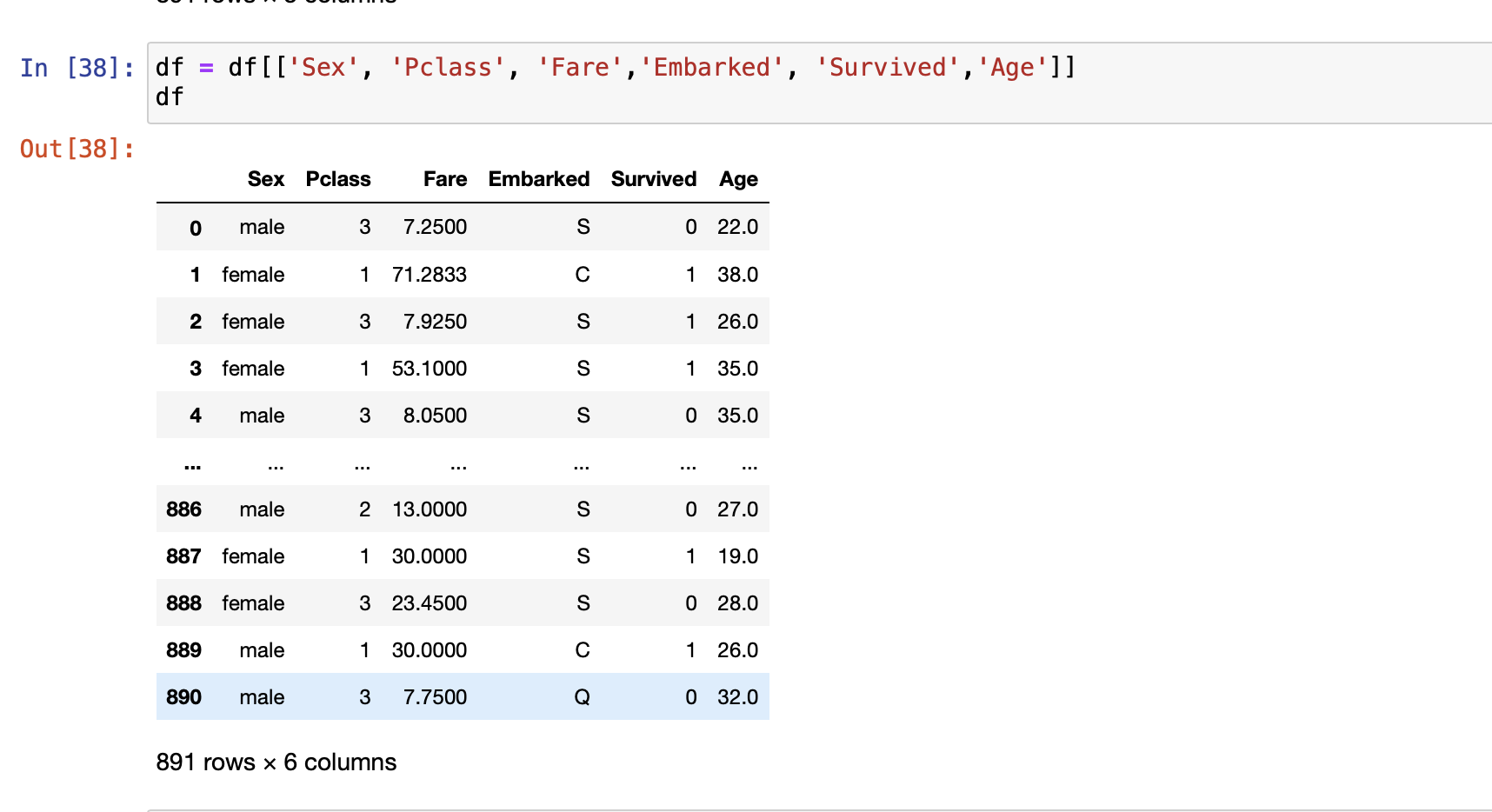
3. **Feature Engineering(Feature Scaling)**

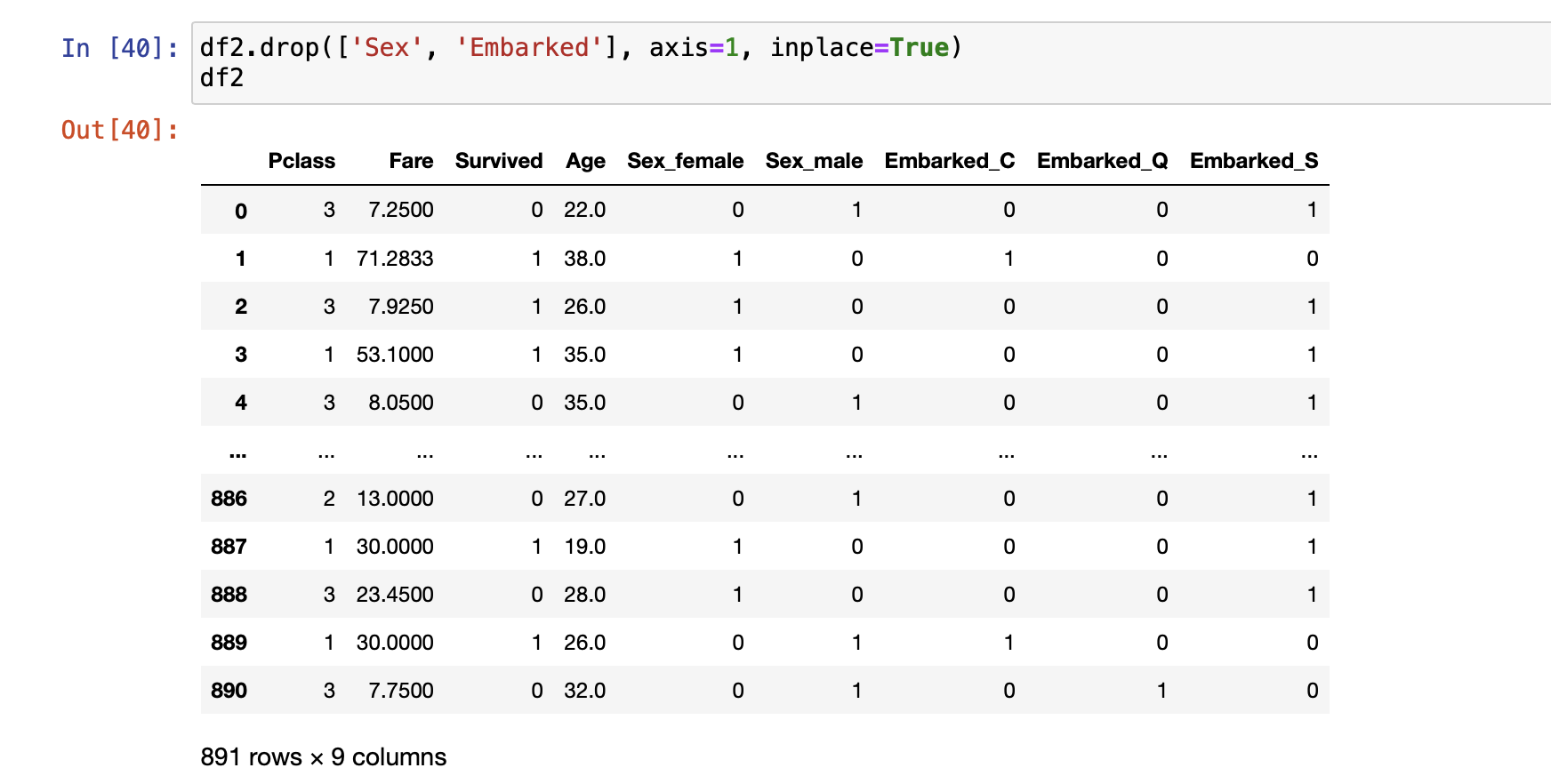
We will be performing all the below steps in Feature Engineering

* Missing values
* Temporal variables
* Categorical variables: remove rare labels
* Standardise the values of the variables to the same range.
* Dropping the Cabin column with maximum no. of missing values.
* Replacing the missing values in Age column with median value of Age column.
* Replacing the missing values in Embarked column with most frequent values in Embarked column i.e ’S’.



* Heat map after removing missing values.
* **Converting string values in categorical values** 
  + Converting string values of Sex and Embarked column to Categorical values and then concatenating old and new data to get new Dataframe.



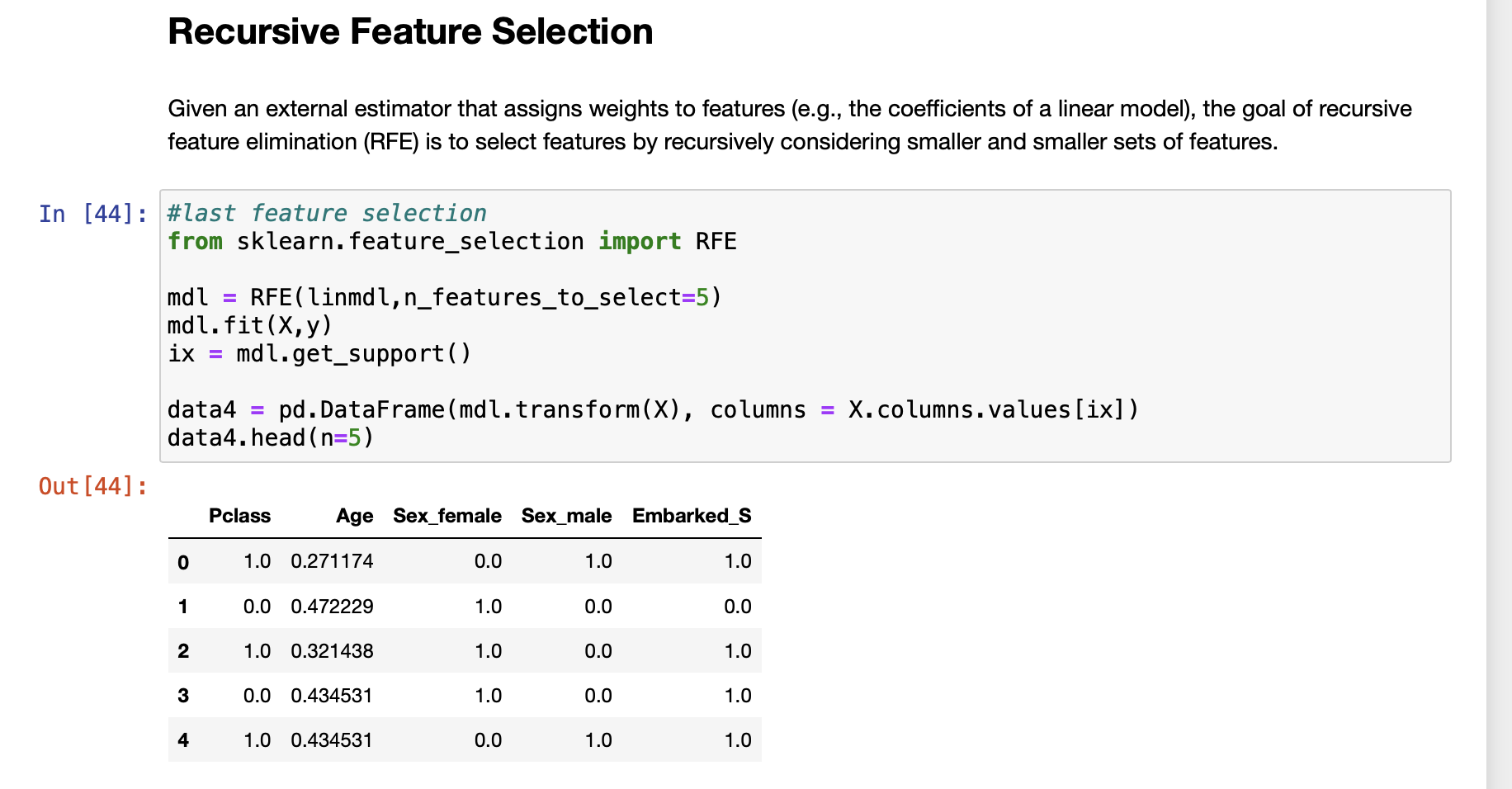
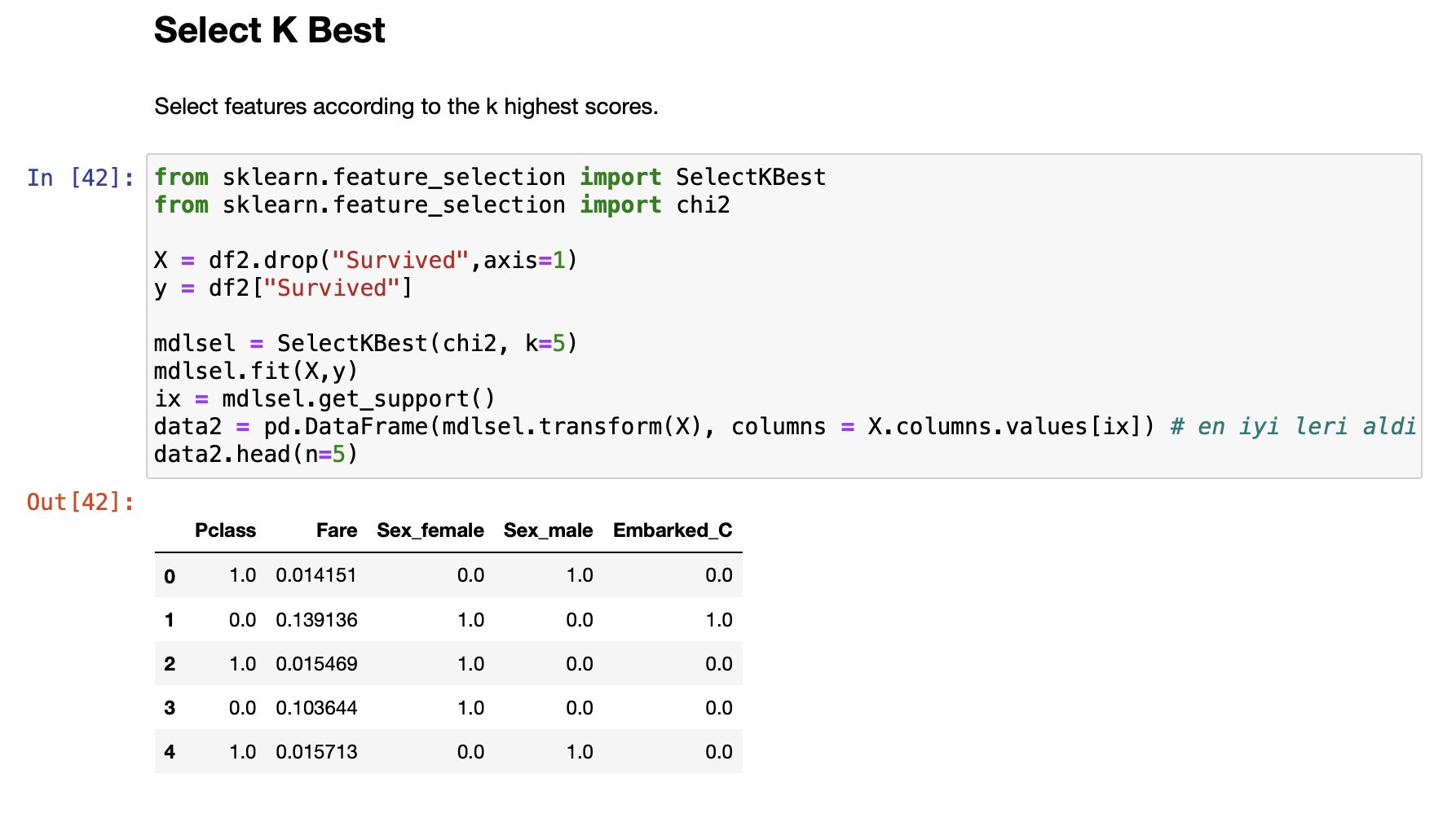
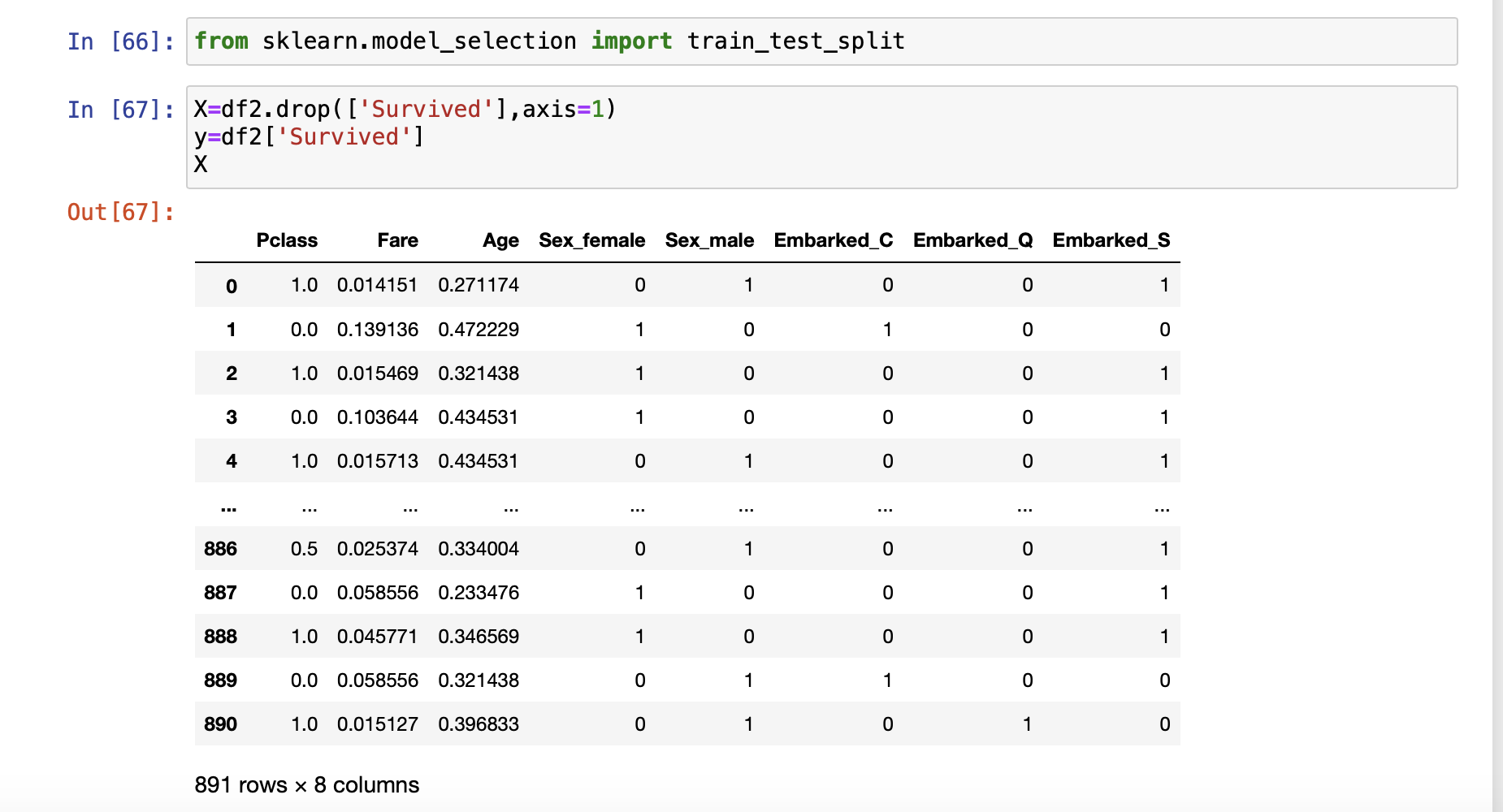


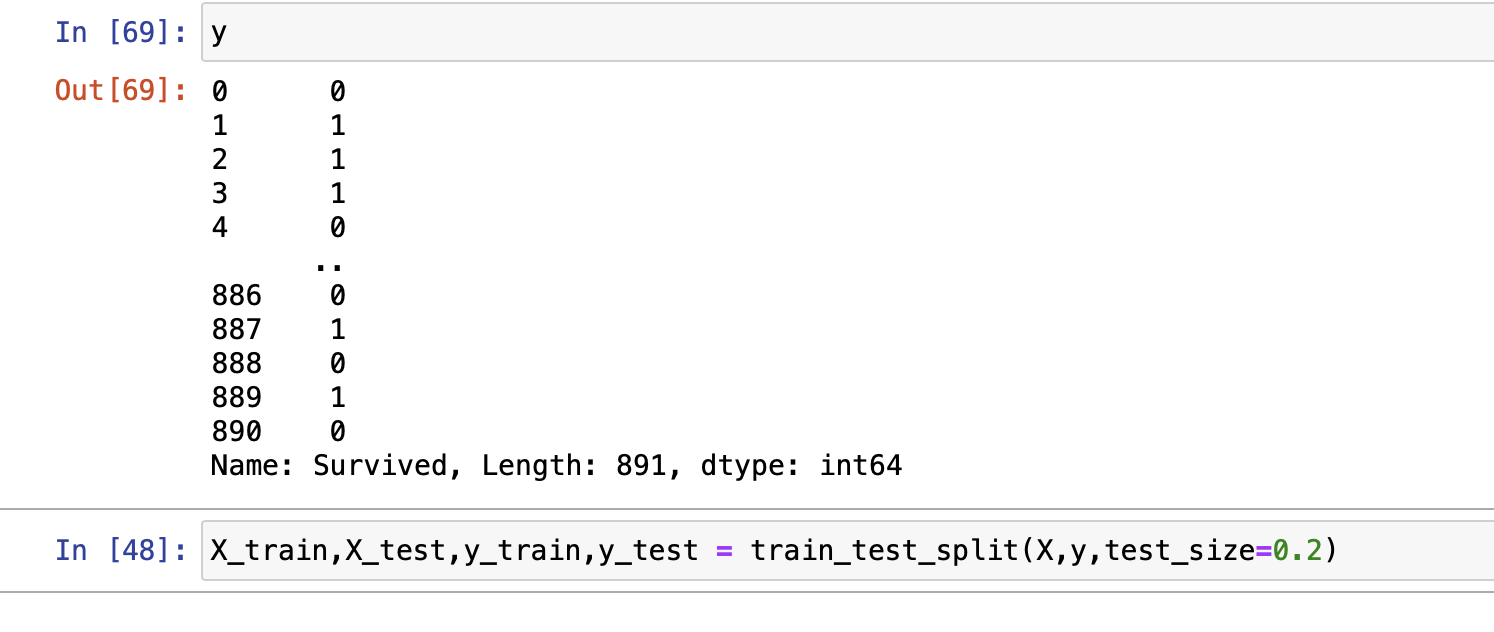
* Scaling the values in Age , PClass and Fare columns between 0 and 1 using Min max scaler.

4. **Feature Selection**

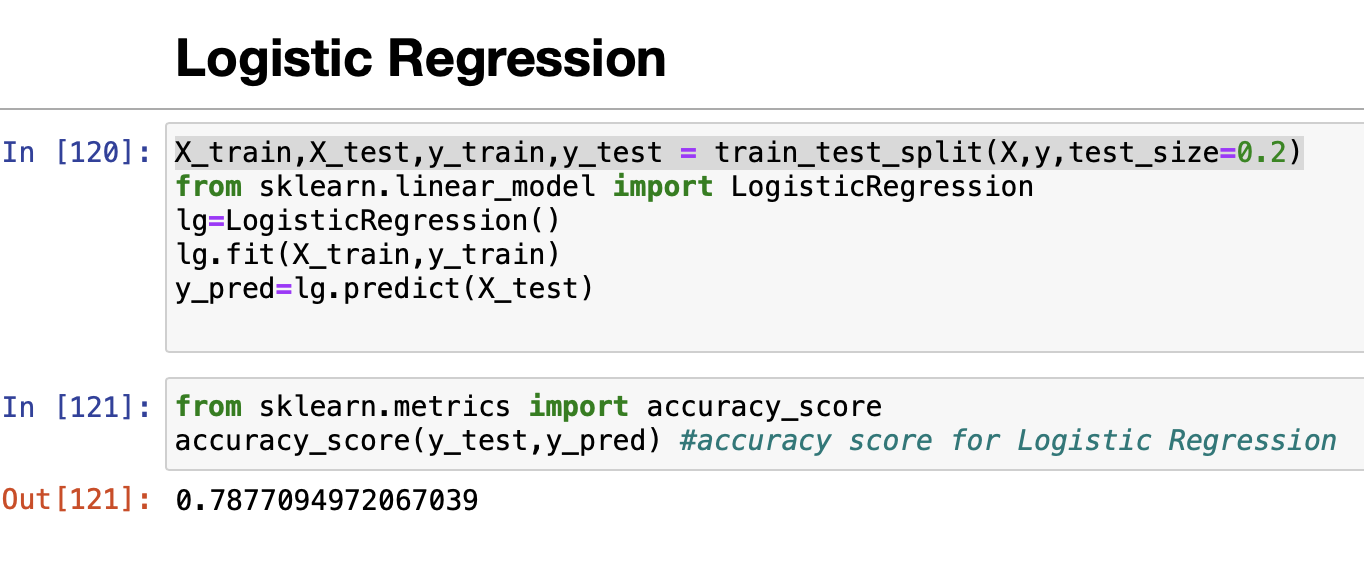
Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

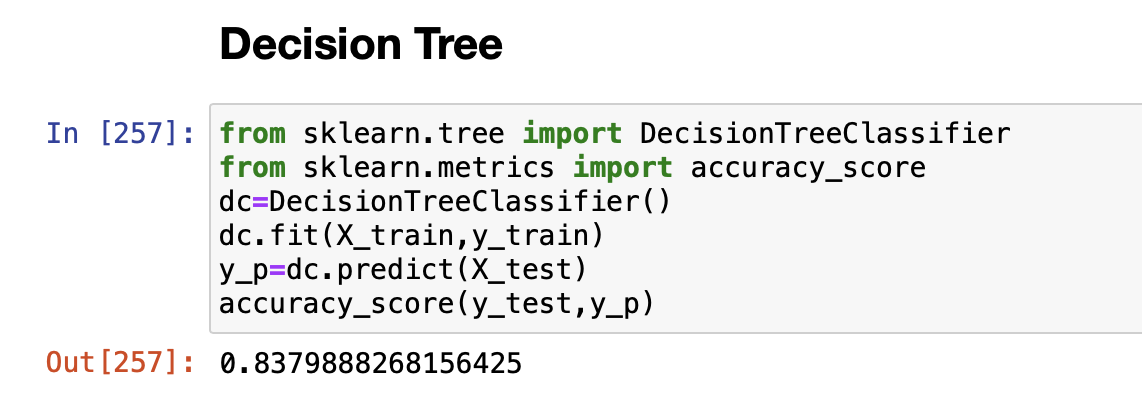
Applying some feature selection methods to extract important features.

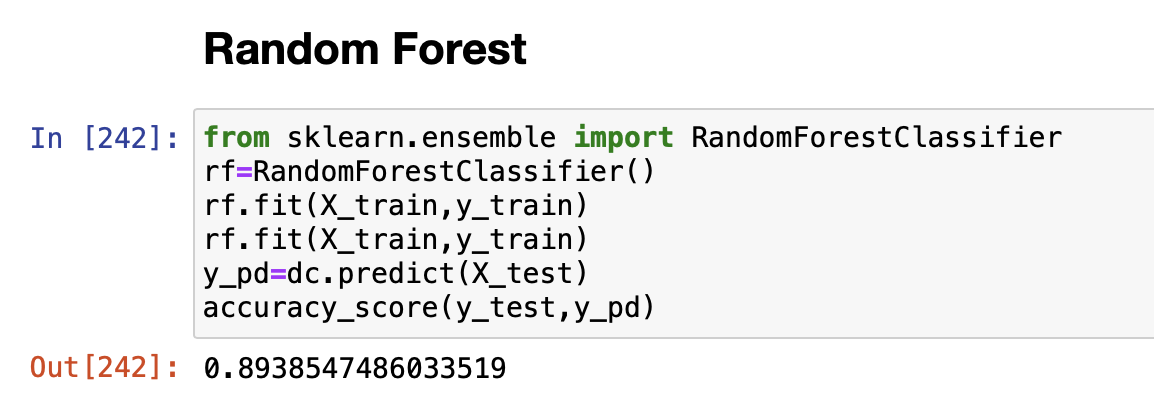
* K Best
* Model for Logistic Regression
* Recursive Feature Selection
* Splitting the data into training and test.
* Declaring X as input variable and y as target variable.

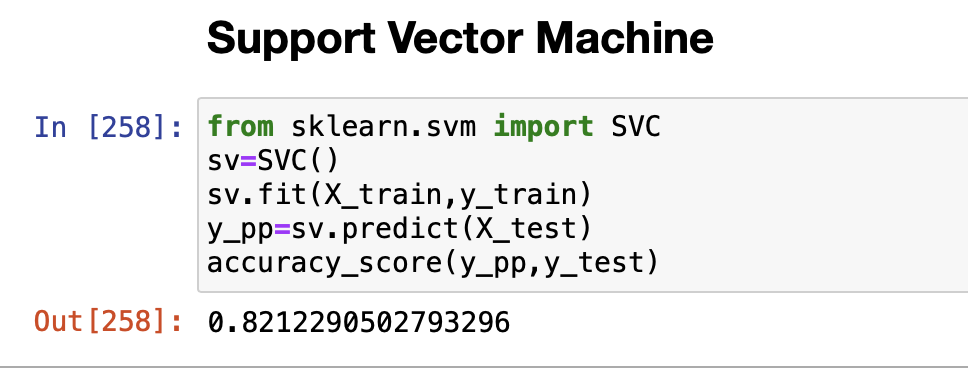


5. **Model Building**

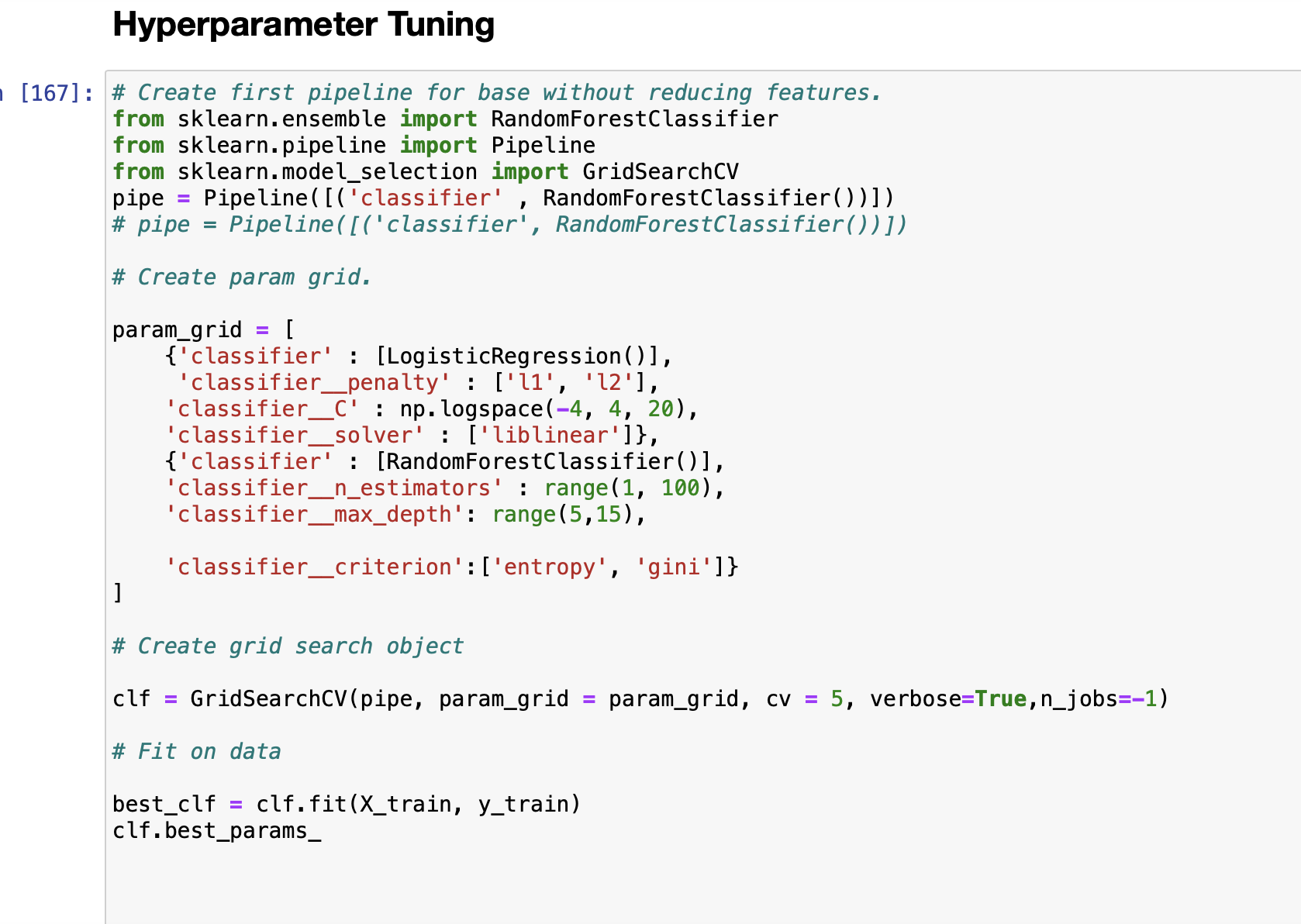
In model building we apply different machine learning algorithms and select models according to accuracy score.

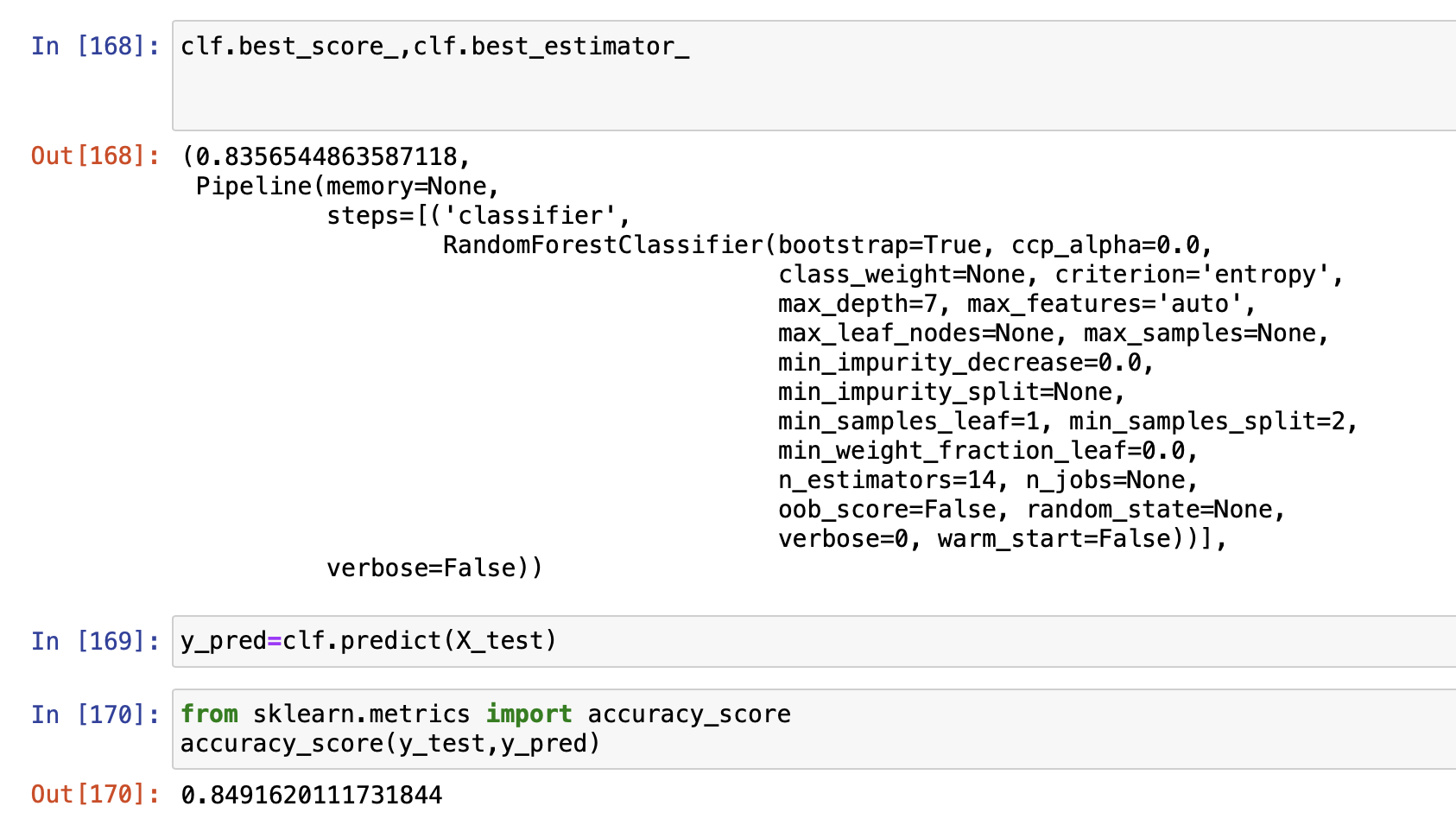
****

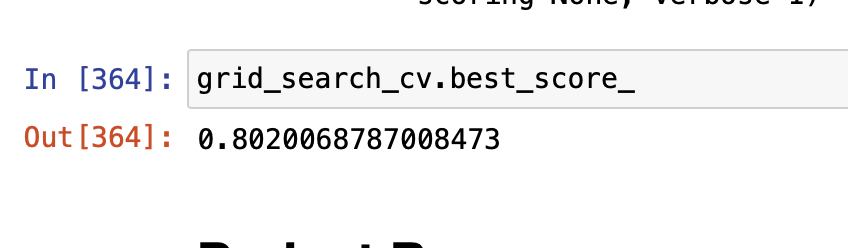
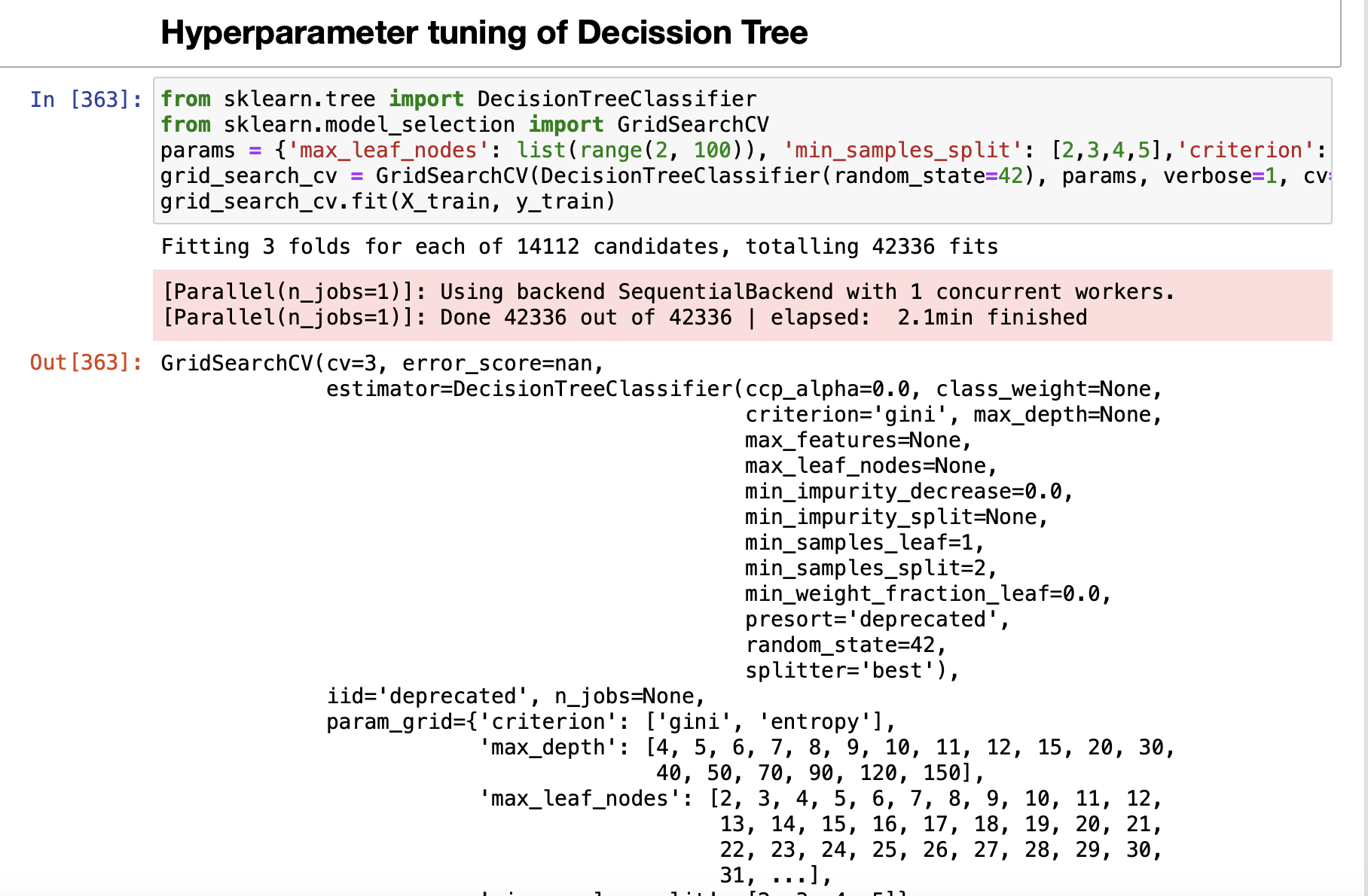
****

****

6. **Hyperparameter Tuning**

Tuning of Hyperparameters are important in order to achieve high accuracy of model. Using GridsearchCV from sklearn best hyperparameters are achieved.

* Hyperparameter tuning for RandomForest.



**RESULTS**

Accuracy scores of different Machine Learning algorithms on Titanic dataset.

**RESULTS after HyperParameter Tuning**