**DATATRAINED**

**BLOG**

**TITANIC PROJECT**

**INTRODUCTION:**

The ships were constructed by the Belfast shipbuilders [Harland and Wolff](https://en.wikipedia.org/wiki/Harland_and_Wolff),. **Titanic** was 882 feet 9 inches (269.06 m) long with a maximum breadth of 92 feet 6 inches (28.19 m). Her total height, measured from the base of the keel to the top of the bridge, was 104 feet (32 m). She measured 46,328 [gross register tons](https://en.wikipedia.org/wiki/Gross_register_tonnage) and with a draught of 34 feet 7 inches (10.54 m), she displaced 52,310 tons. It was one of the largest and most opulent ships in the world. It was also considered unsinkable, due to a series of compartment doors that could be closed if the bow was breached. However, four days into its maiden voyage in 1912, the Titanic struck an iceberg, and less than three hours later it sank.

**PROBLEM STATEMENT:**

 The Titanic Problem is based on the sinking of the ‘Unsinkable’ ship Titanic in early 1912. It gives you information about multiple people like their ages, sexes, sibling counts, embarkment points, and whether or not they survived the disaster. Based on these features, you have to predict if an arbitrary passenger on Titanic would survive the sinking or not.

PROCEDURE DONE

1.Exploratory Data analysis (EDA).

2.Preprocessing of the data.

3.Training process of the dataset to make best model (include various steps).

4.Hyper parameter tuning

5.ROC AUC curve 6. Saving the model

**DATA ANALYSIS and EDA CONCLUDING REMARKS**:

* This is a classification problem as our target variable (Survived) is categorical.
* Data set has **891 rows × 12 columns**.
* **0** duplicate values.
* Few null values which were treated using **imputation** **techniques**. While treating null values column cabin was dropped as it was of no use and most of the values were null.
* **Input variables:** ('PassengerId', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',

'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked')

* **Target variable:** Survived

* Visual analysis of the data using various types of plots (count plot, violine plot, dis plot, scatter plot, bar plot, and hist plot) imported from seaborn and matplot.lib libraries for better understanding of the data
* **Sex** – Visualization of this column tells that sex of most of the people on the ship was male than of the female. Most of the survived passengers were female.
* **Pclass**(Ticket Class) – Visualization of this column tells that Pclass (Ticket Class) of most of the people on the ship was **3** than **1** than at last **2**. But most of the people survived were from class 1 than 2 and at last 3.
* **Age** – Visualization of this column tells that age of most of the people on the ship was between the range of **20** to **40**.
* **Embarked** – This column tells the port of embarkation. Visualization of this column tells that most of the people on the ship embarked form **Southampton** then from **Cherbourg** than at last from **Queenstown** but most of the people that survived were from **Cherbourg** than **Queenstown** than **Southampton.**
* **Fare**- It is in the range of **0** to **100**.
* **Parch**- It tells the children that boarded the ship or the parents that bought their children. Most of the people that boarded the ship didn’t bring their children.
* **Survived**- This column has values as 0 and 1 where 0 means no and 1 means yes. Most of the people died.

**PRE-PROCESSING**

* Correlation of the columns with the target variable
* sex is highly correlated with survived.
* Pclass is the second thing most correlated with the class but not as much as sex.
* Embarked is fine in correlation with survived.
* Ticket is also fine.
* Age, Name, SibSp, PassengerId and parch these columns are less correlated than others.
* The worst correlated column is fare.
* Encoding the data using Label Encoder so that data is readable by the machine. Label encoder converts label into numeric form so as to convert them into machine readable form.
* Outliers were present in the dataset which were removed with the data loss of **7.182940516273851** percentage.
* Skewness was also present in data. Various methods such as SQUARE ROOT METHOD and CUBE ROOT METHOD was used but they were not successful so skewness was removed using YEO JOHNSON method at last.
* Balancing the data using **SMOTE**.

**BUILDING MACHINE LEARNING MODELS**

This is a classification problem as the target variable is categorical so after the EDA and Pre-processing of the data the next step is to build and train the model for which we’ll call important libraries. Steps performed in building and training the model were: -

1. Calling important libraries.
2. Finding best random state where maximum accuracy is received.
3. Then Creating train and test split.
4. Using various algorithms with that created split to see which works the best. Algorithms used were: -

* Decision tree classifier
* Random forest classifier
* K neighbors classifier

5.Doing cross validation. Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation **to detect overfitting, ie, failing to generalize a pattern**.

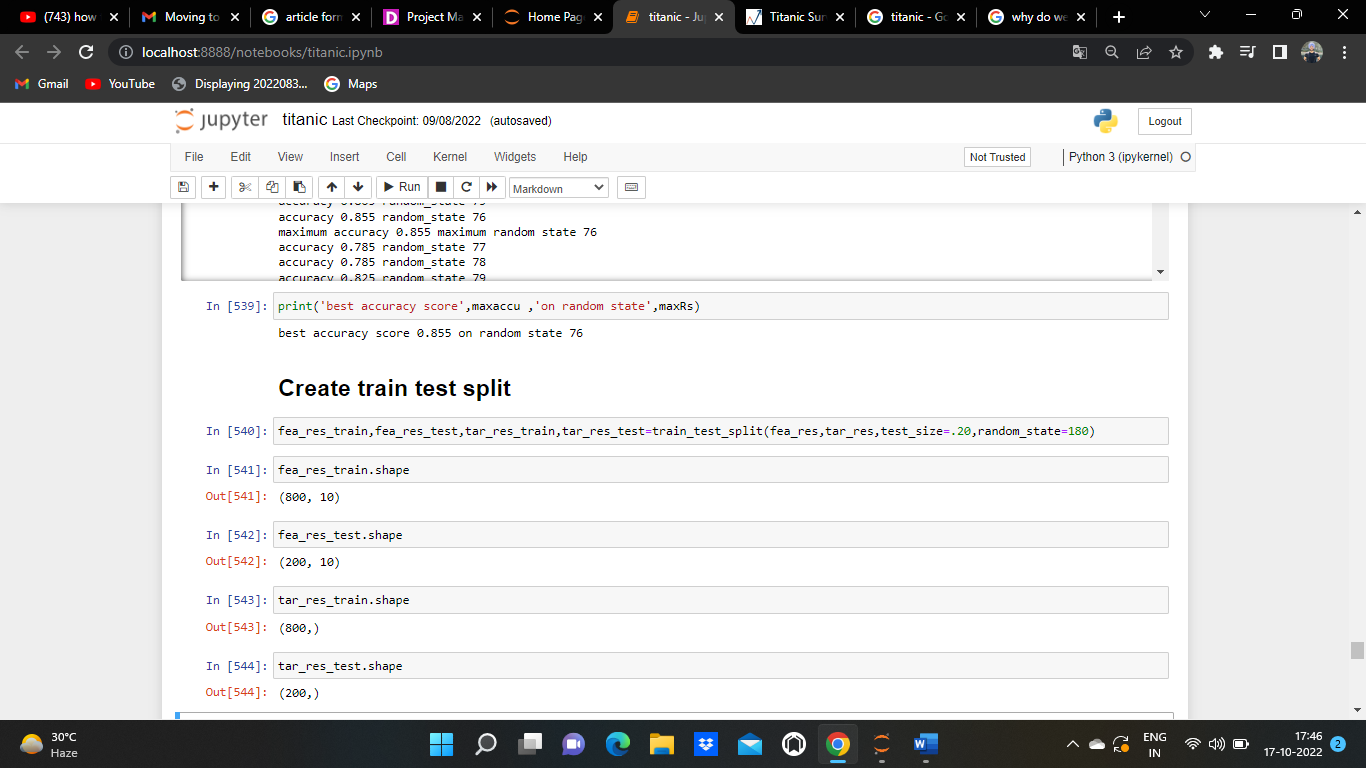
6. After cross validation of all the models the best one is the **kneighbors classifier** with which the further processing is continued.

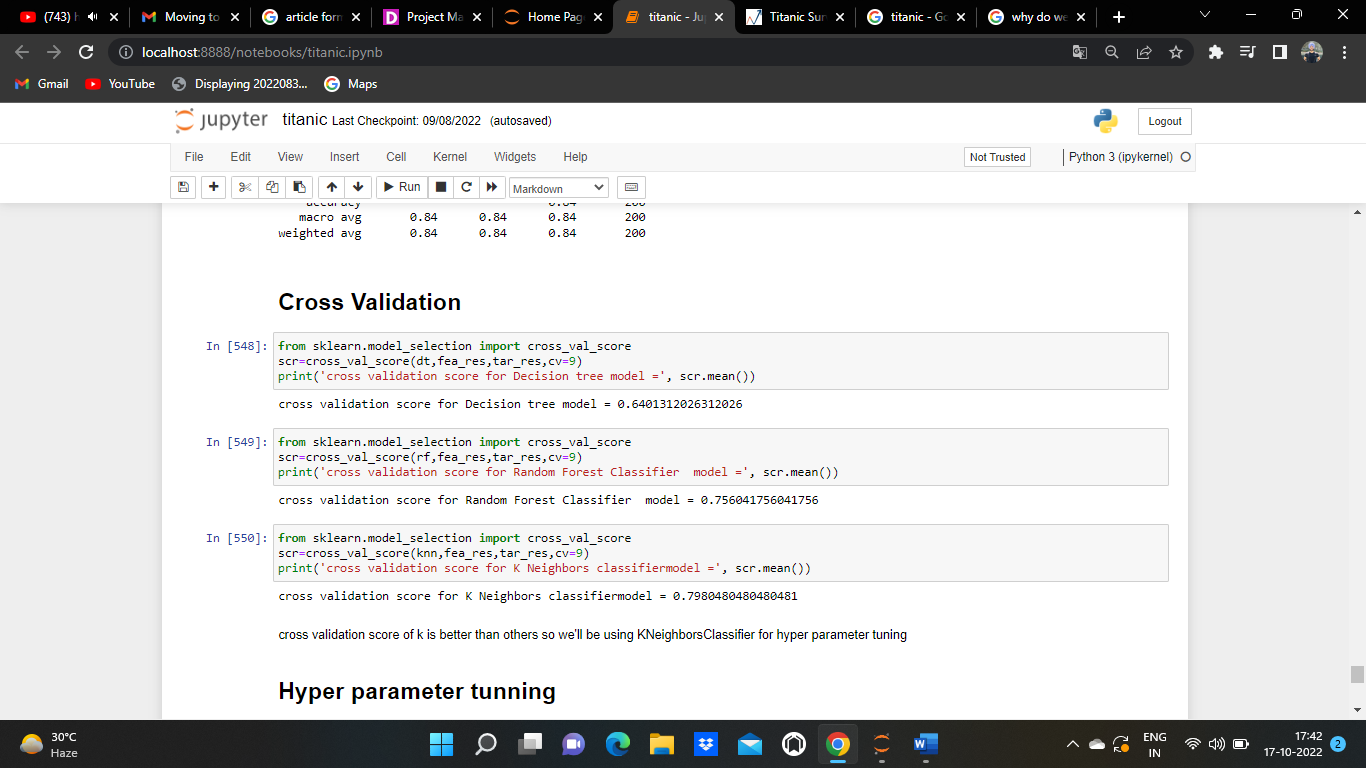
7.Hyper parameter tuning using grid search CV and finding best parameters and best estimators and then predicting the accuracy score.

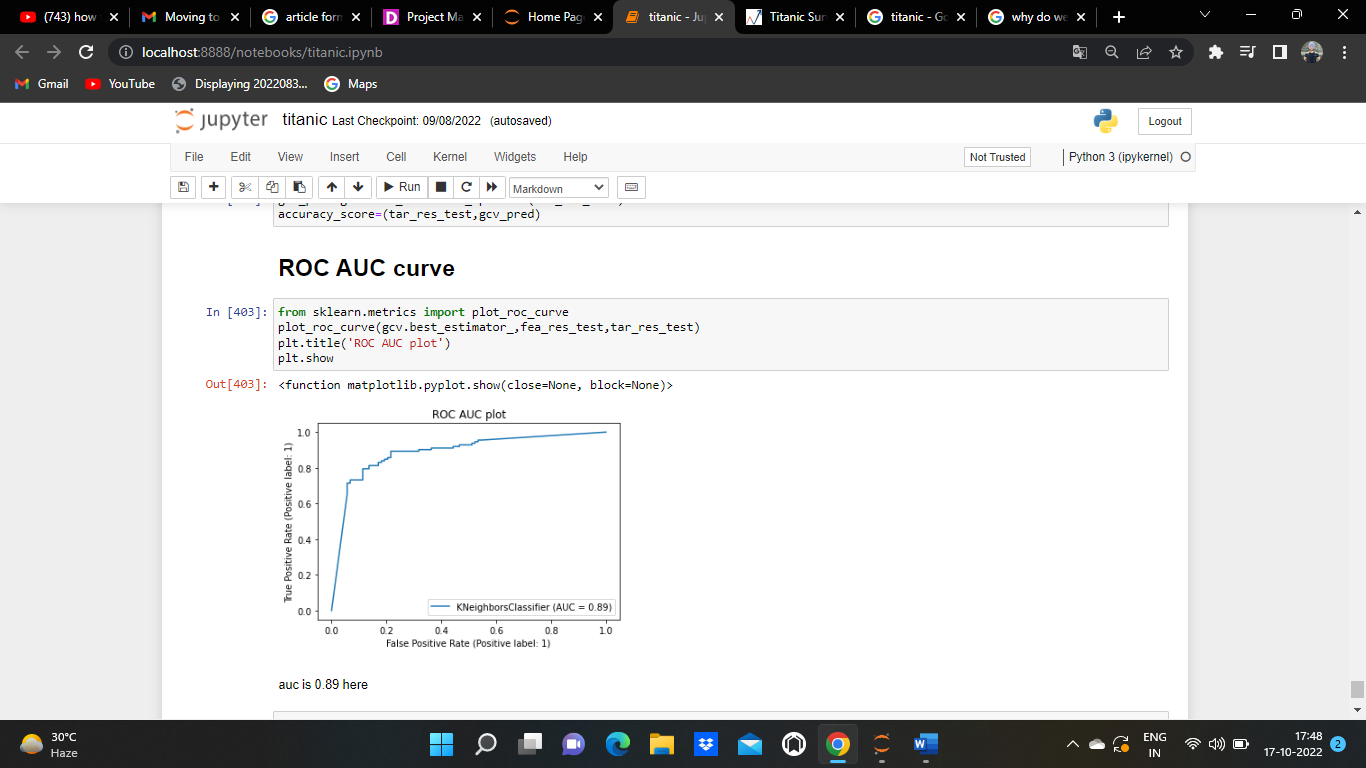
8.Ploting ROC AUC plot where AUC is 0.89.

**SNAPSHOTS**

Sharing snapshots of cross validation , train and test split , and ROC AUC curve.







**SAVING THE MODEL**

Saving the model using joblib.

**CONCLUDING REMARKS**

**Titanic** project was a classification problem as it’s target variable or output variable (Survived) was a categorical one. In this the prediction was done and models were made using various type of classification algorithms and the best among them was selected. After studying this data set we get to know that males were more in number as passengers but very few passengers survived among which most of them were women and were of 1st class then followed by other ticket classes this was it because passengers of 1st class had easy and more access to life boats and were evacuated easily as compared to other 2 classes . Most of the people that survived embarked from Cherbourg. Among various classification models the best working one was the **knn classifier** with

* **accuracy score** of 84.0 and.
* **cross-validation score** of 0.7980480480480481 and.
* while in ROC AUC CURVE the **AUC** comes to be 0.89.