**­­­­­PROJECT : WEATHER PREDICTION**

**Data Science** is nothing but getting meaningful insights from data**.**

Here my project deals with the Prediction of weather on the basis of certain parameters that can affect the weather conditions in real time. Making Prediction of weather on the basis of Machine Learning Algorithm.

**Machine Learning** is nothing but it is a subfield of Artificial Intelligence where software applications learn from Experience without being Explicitly Programmed**.**

**Input :** Precipitation, Min-max temperature, Wind speed.

**Algorithm used :** RandomForestClassifier

**Output :** Weather Prediction

**Application of model : Forecasting could be applied in air traffic, severe weather alerts, marine, agriculture, utility companies, private sector and military application.**

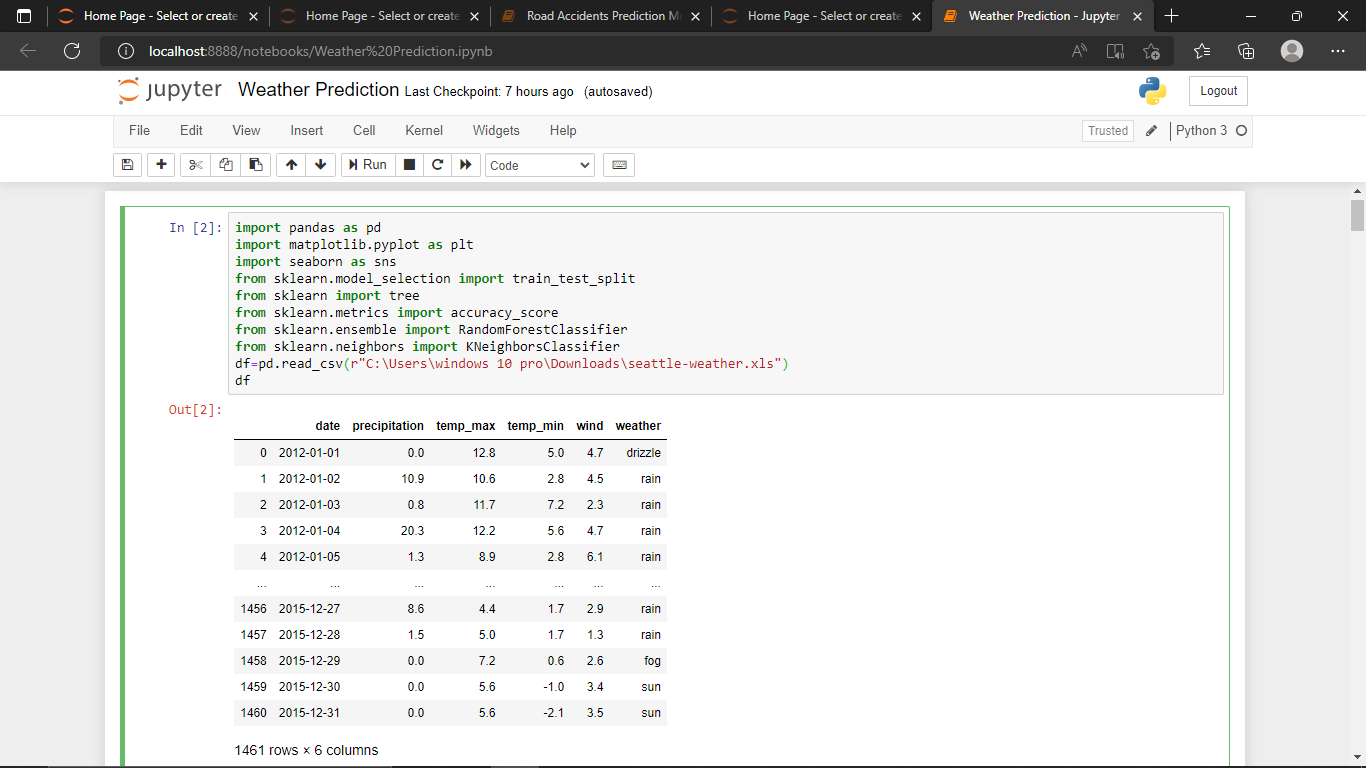
The goal of weather prediction is to provide information people and organizations can use to reduce weather-related losses and enhance societal benefits, including protection of life and property, public health and safety, and support of economic prosperity and quality of life.

**Step-1:**

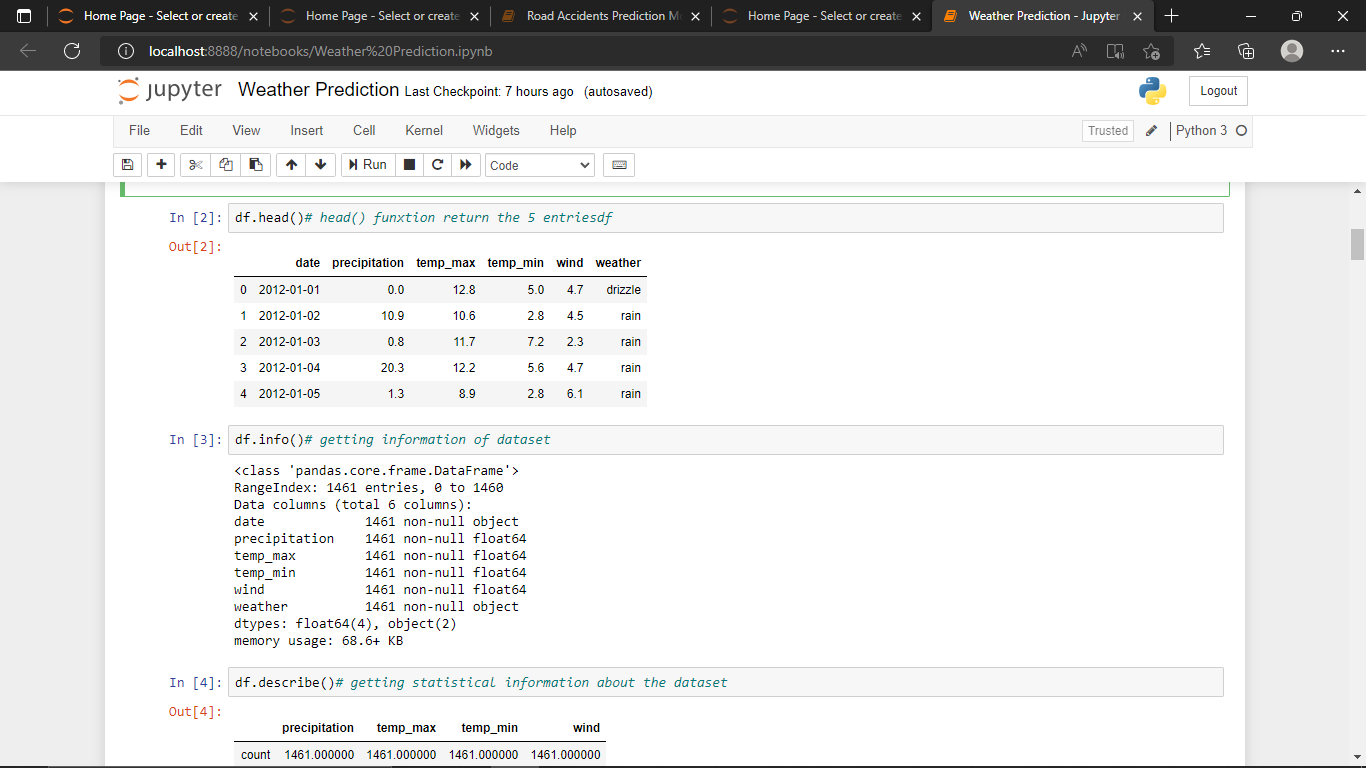
Importing the Dataset

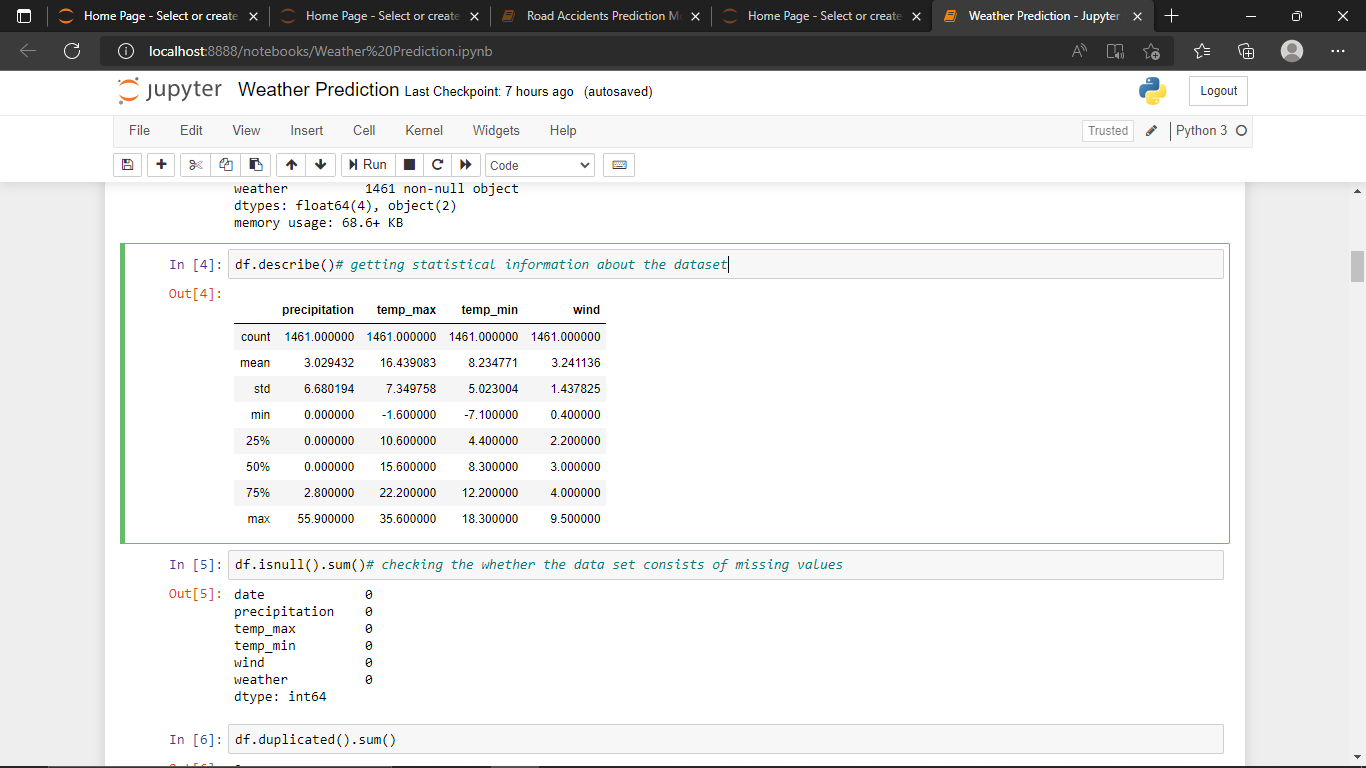
Importing a csv file named “seattle-weather.csv” that contains the data regarding weather factors with respect to the date ,temperature and other parameters.

Importing the package and then creating a Data Frame using the csv file we imported before.

·

* df.head() = #returns the first five rows of the DataFrame.
* df.info() = # getting information about the dataset.
* df.describe() = ## getting statistical information about the dataset.



****

**Step-2:**

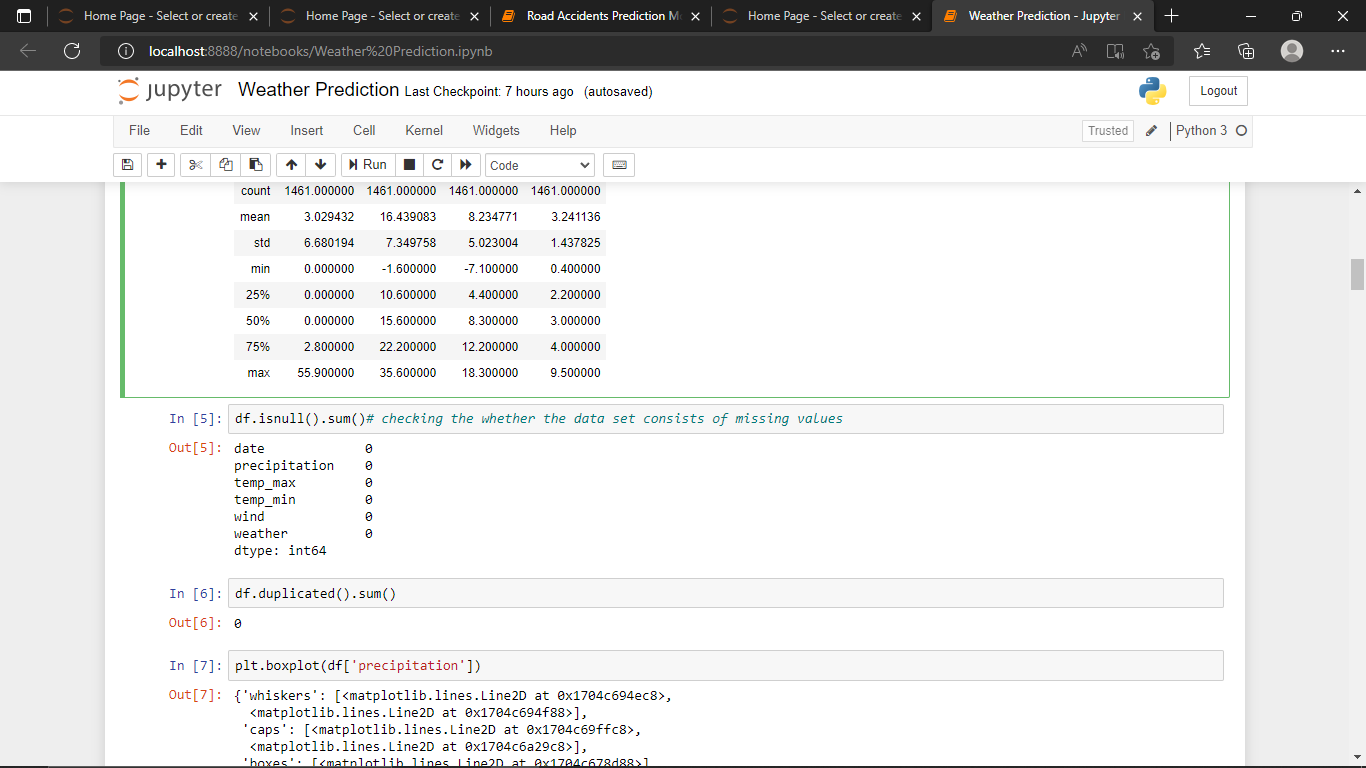
**Data Preprocessing:-**

It involves the manipulation of data that includes handling the missing values and removing the duplicate values and correcting the incorrect patterns etc. These all come to Data Cleaning.

1.Handling with missing values.

2.Removing duplicates.

* df.isnull().sum() = #gives the sum of null values presenting in each column.
* df.duplicated() = # Finding whether the dataset contains duplicated values or not.



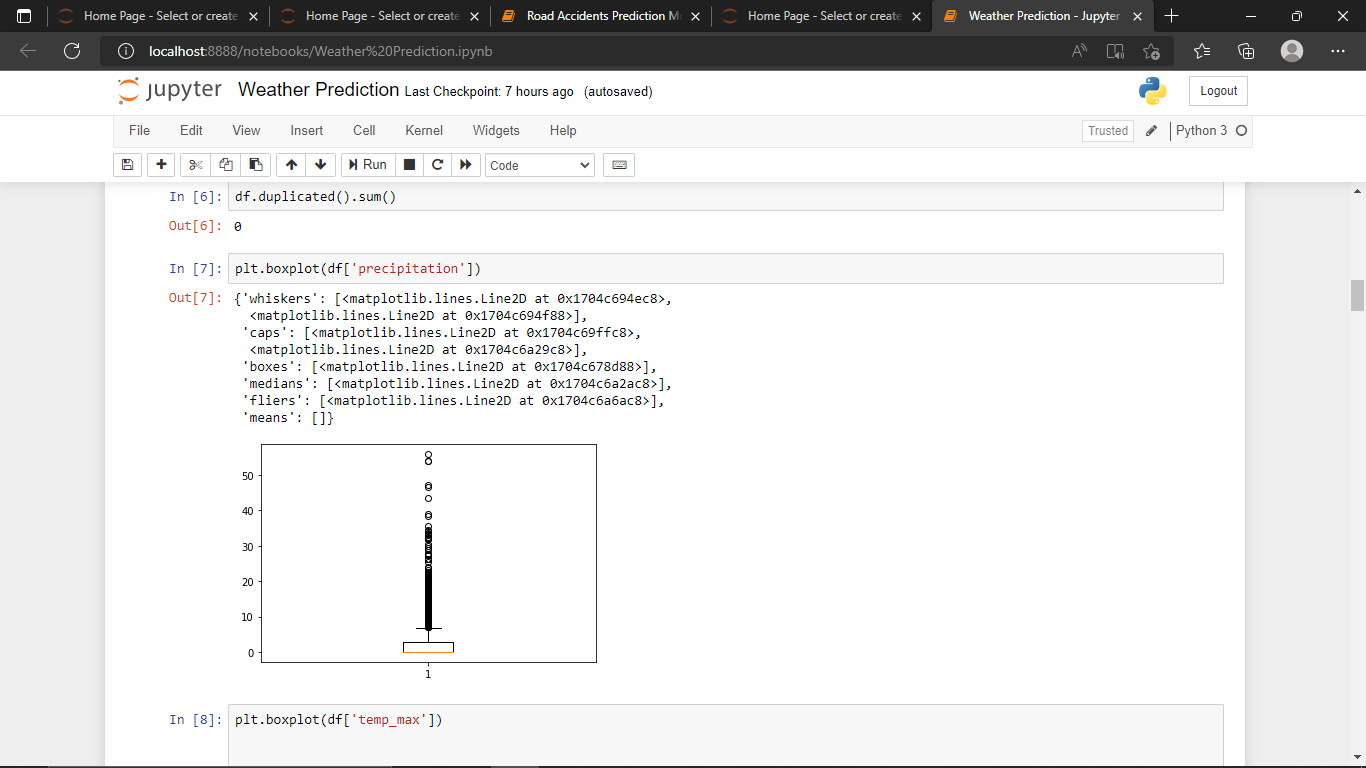
* Here this dataset doesn’t contain any of the duplicate values.
* Here this dataset doesn’t contain any of the null values.
* There is no incorrect data present in the data.
* Above data is cleaned.
* Data cleaning or Data cleansing is the process of detecting and correcting corrupt or inaccurate records from a record set, table or database.
* Hence our DataFrame looks perfect hence we have to visualize the data in different ways and find the relation between the variables present in the dataset and getting meaningful insights.

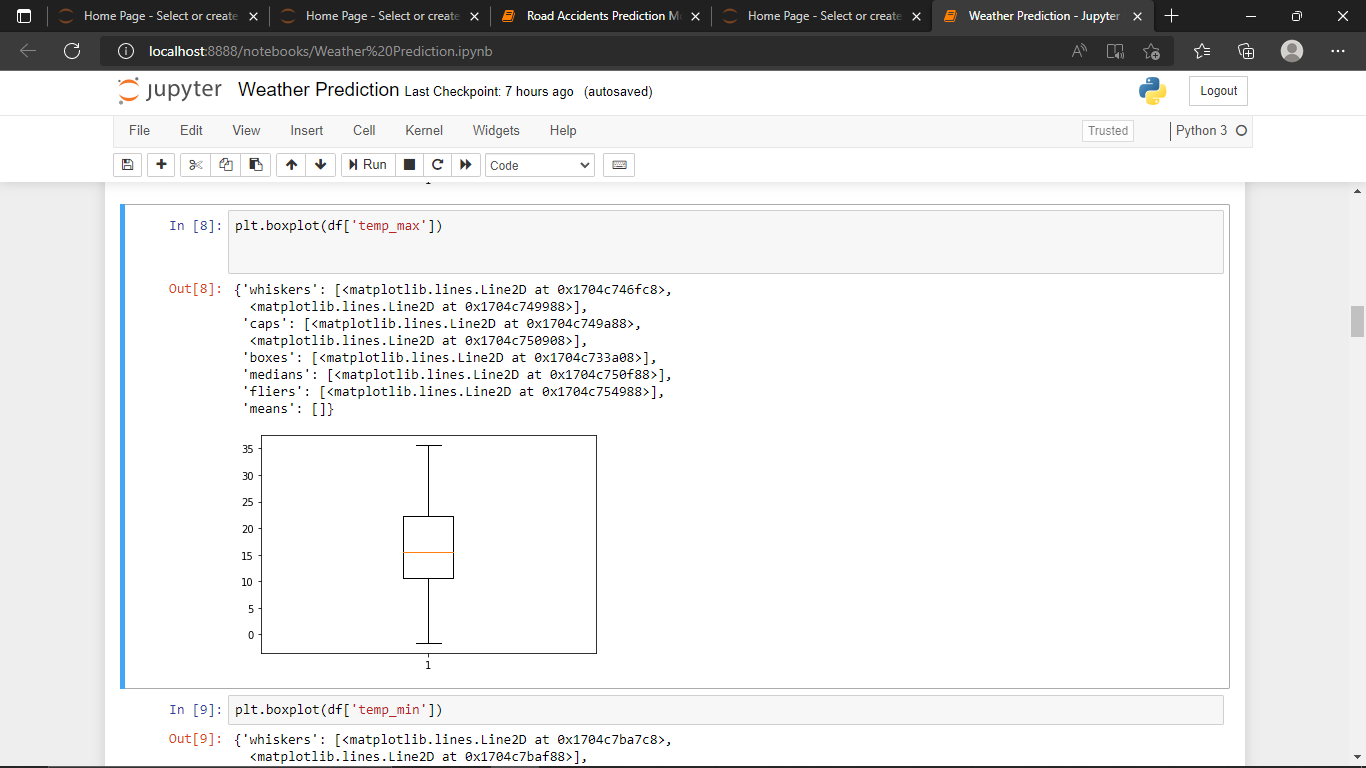
**Step-3:**

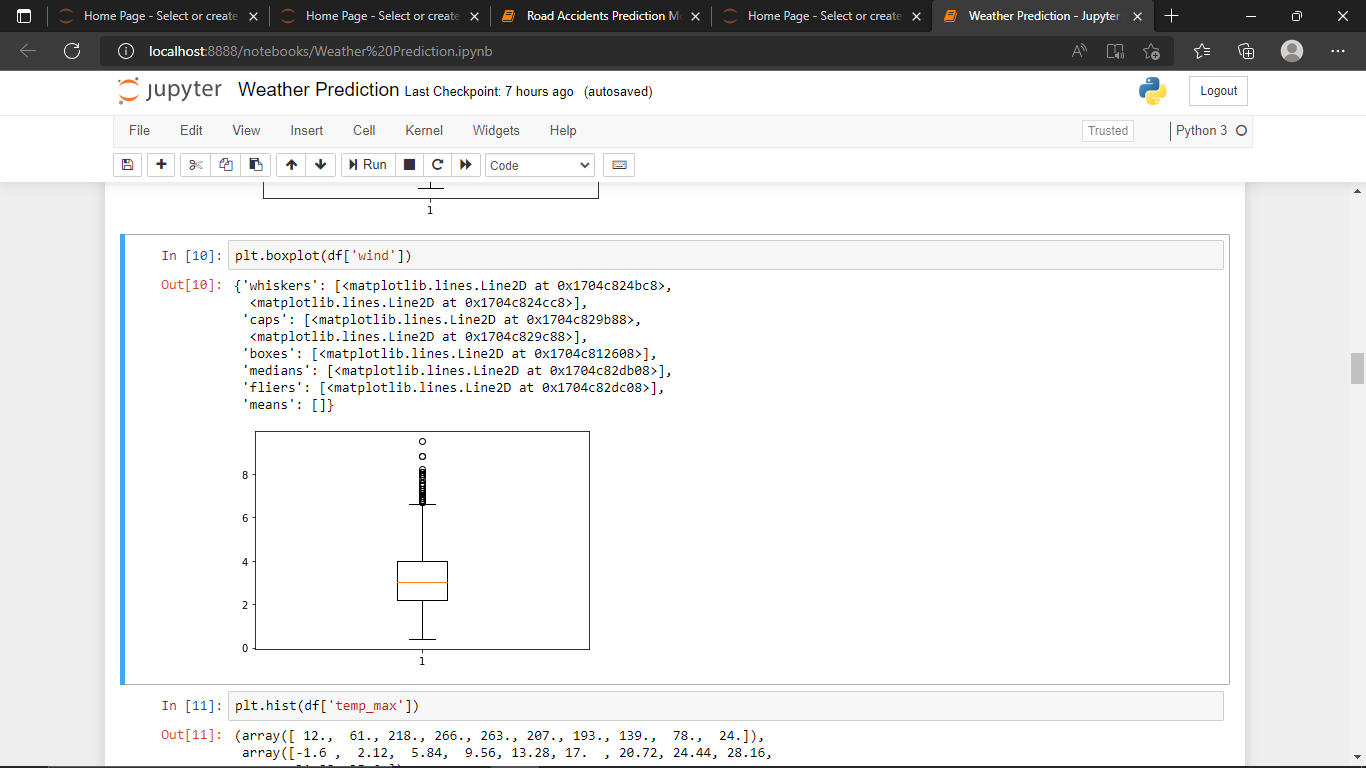
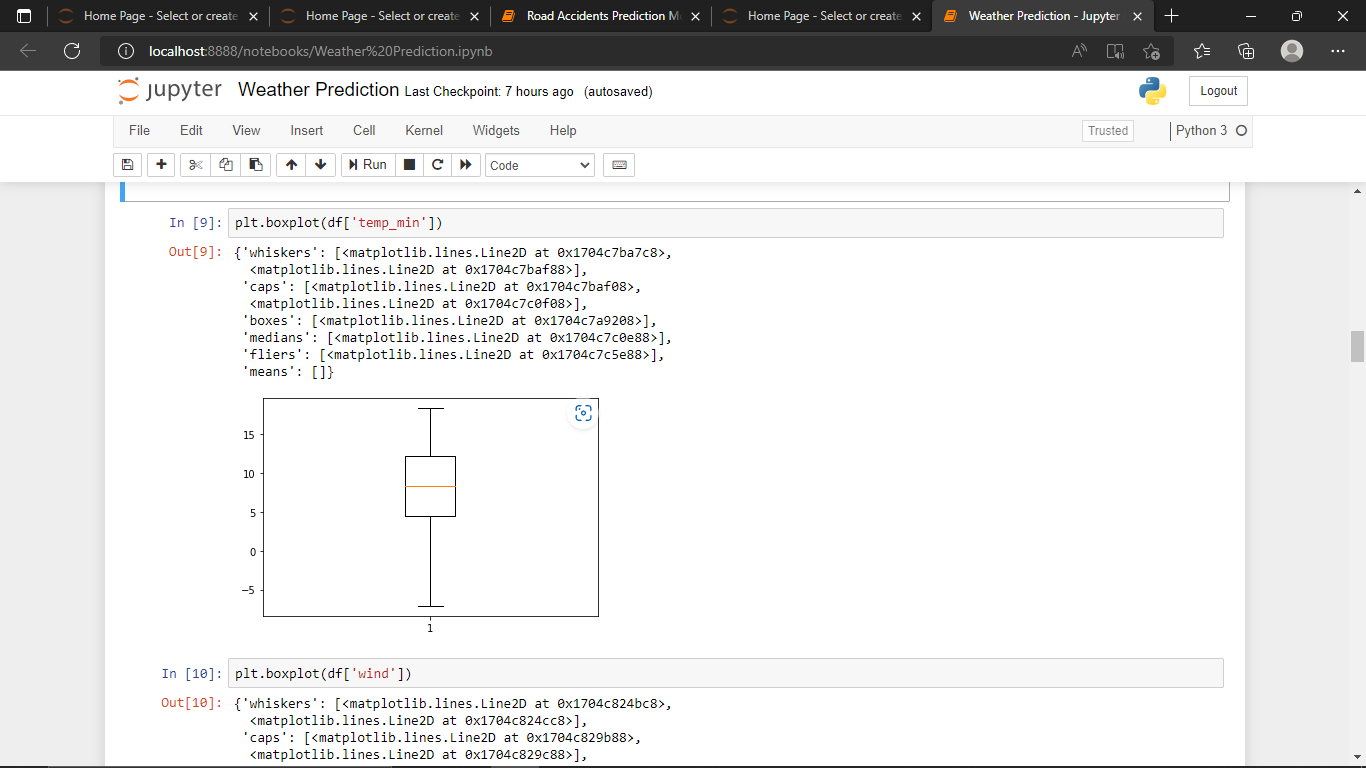
**Data Visualization:-**

* It is nothing but graphical representation of data and information is called Data Visualization.
* Importing the plotting Library called Matplotlib to visualize the data.

1. **How well the data has been distributed in the dataset using boxplot.**

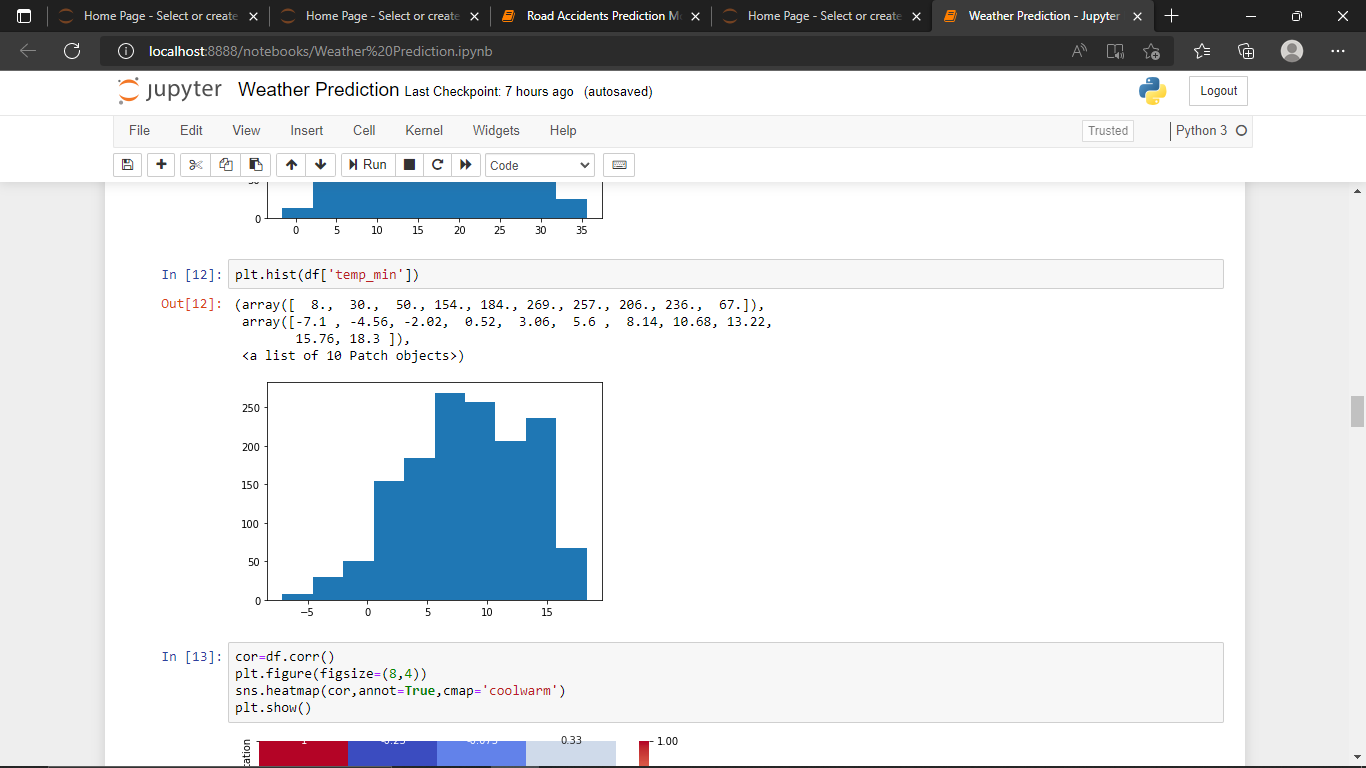
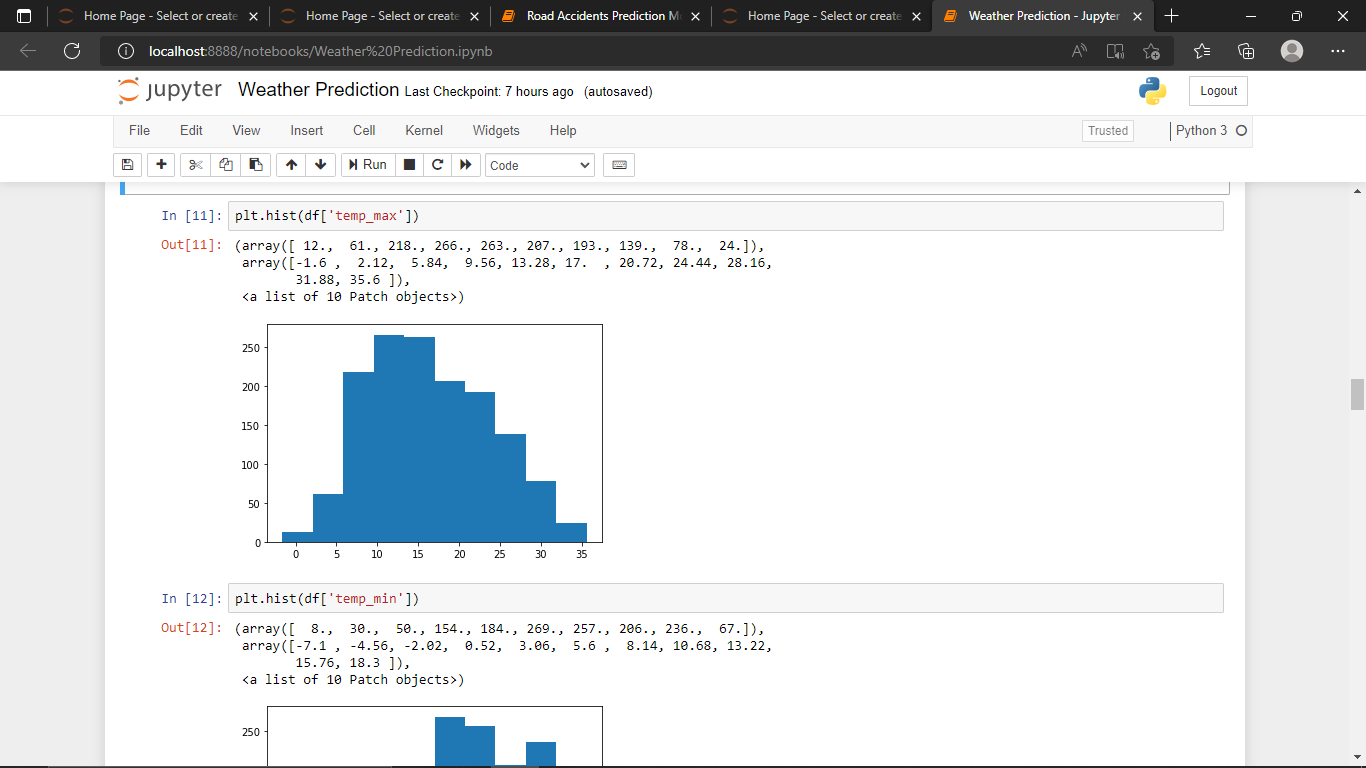
****

****

****

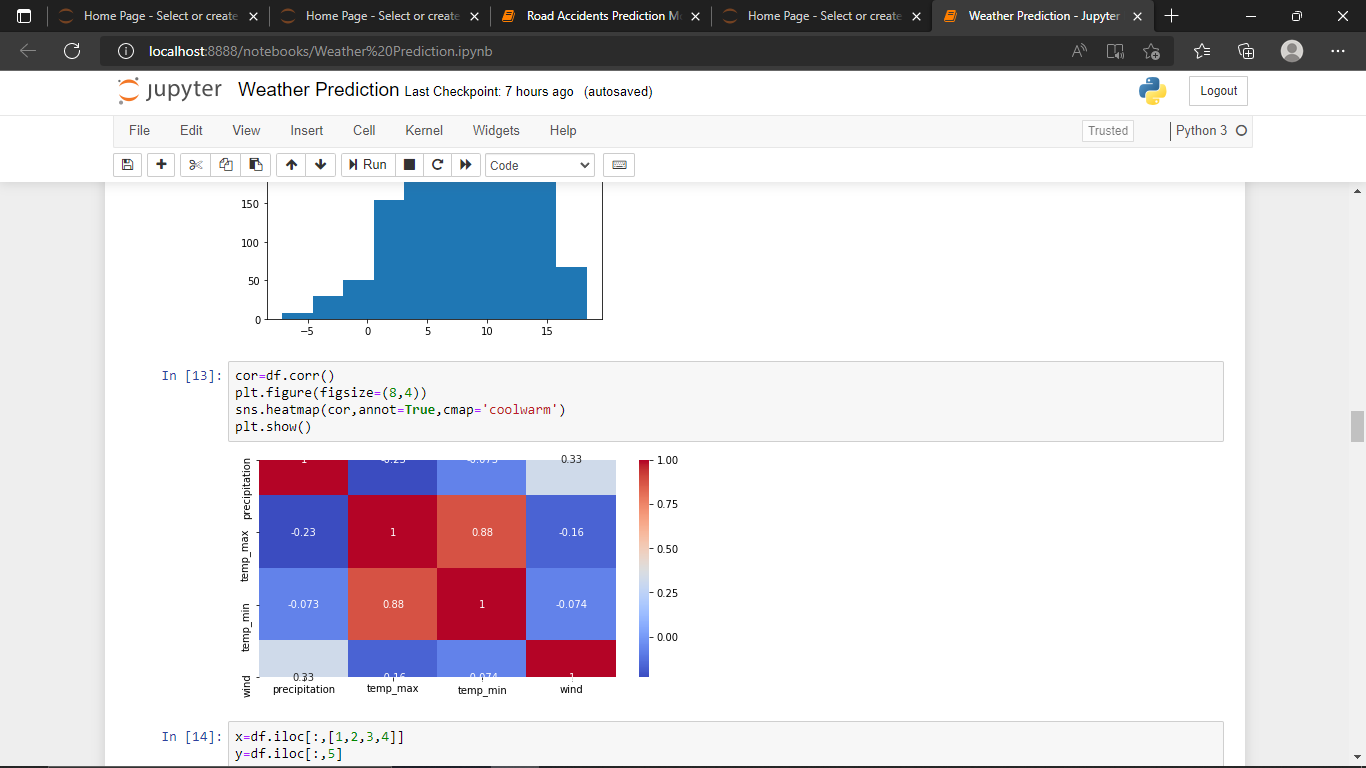
**2.Plotting the data using histogram.**

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions



**3. Correlation:A correlation matrix is a table showing correlation coefficients between variables. It ranges from -1 to 1.**

* It takes an input relation with numerical columns, and calculates Pearson Correlation Coefficient between each pair of its input columns.
* Positive correlation is measured on a 0.1 to 1.0 scale. Weak positive correlation would be in the range of 0.1 to 0.3, moderate positive correlation from 0.3 to 0.5, and **strong positive correlation from 0.5 to 1.0**.



* Here minimum and maximum temperature have a good correlation.
* Precipitation has a moderate correlation among the variables.

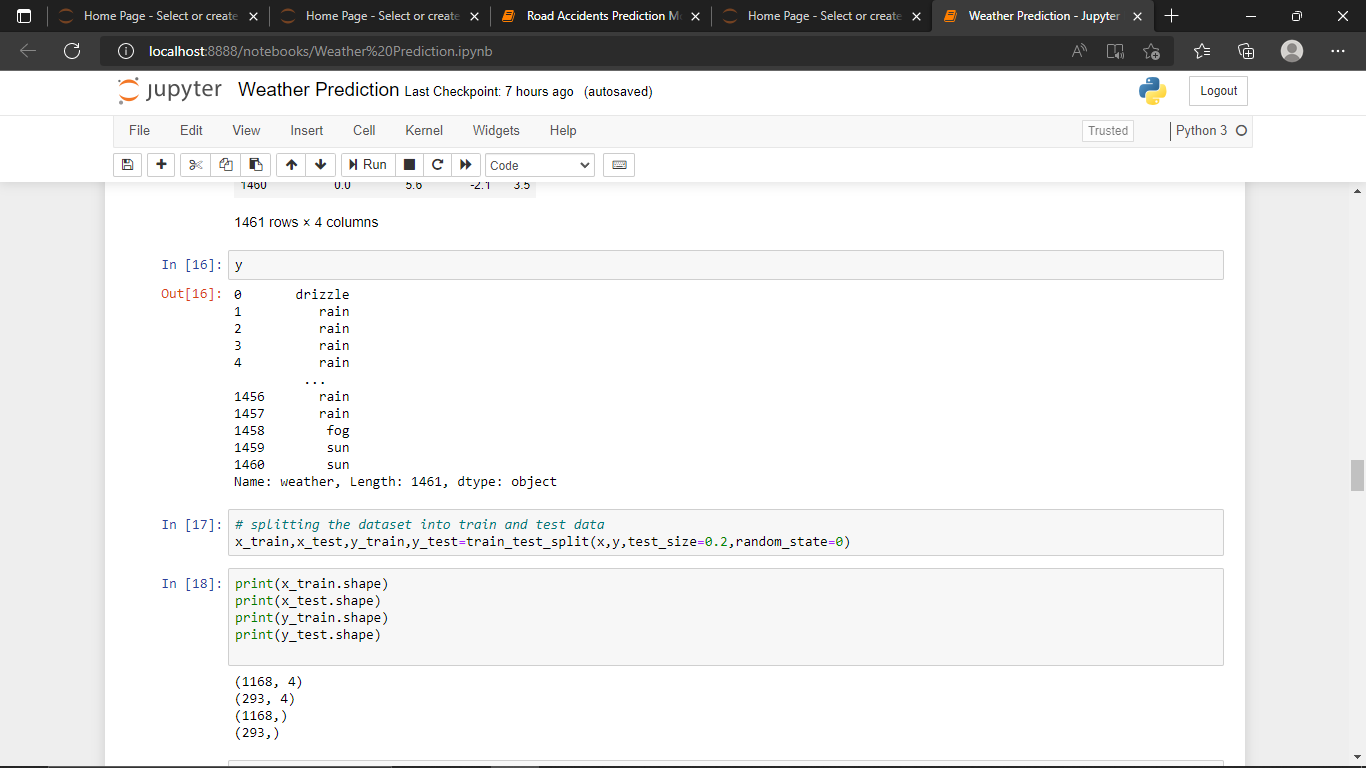
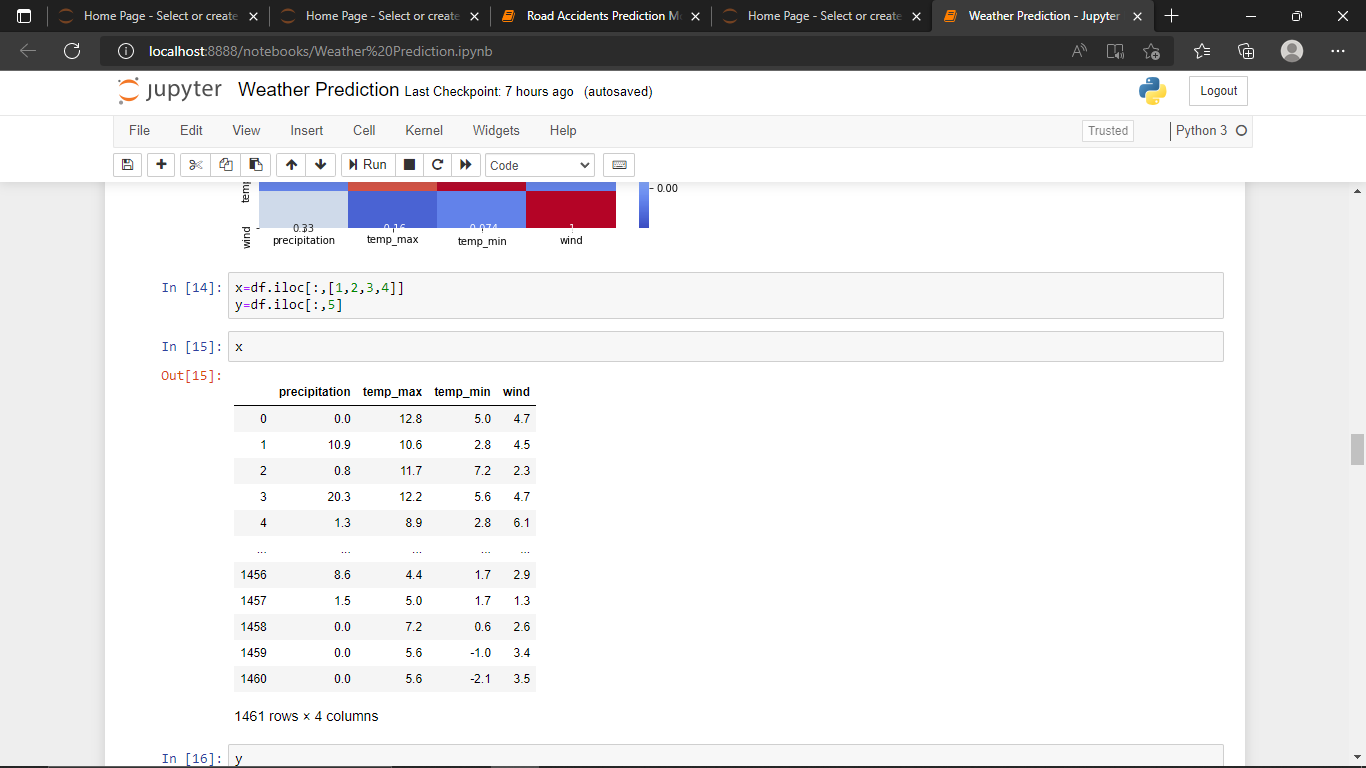
**Step-4:**

**MODEL MAKING:-**

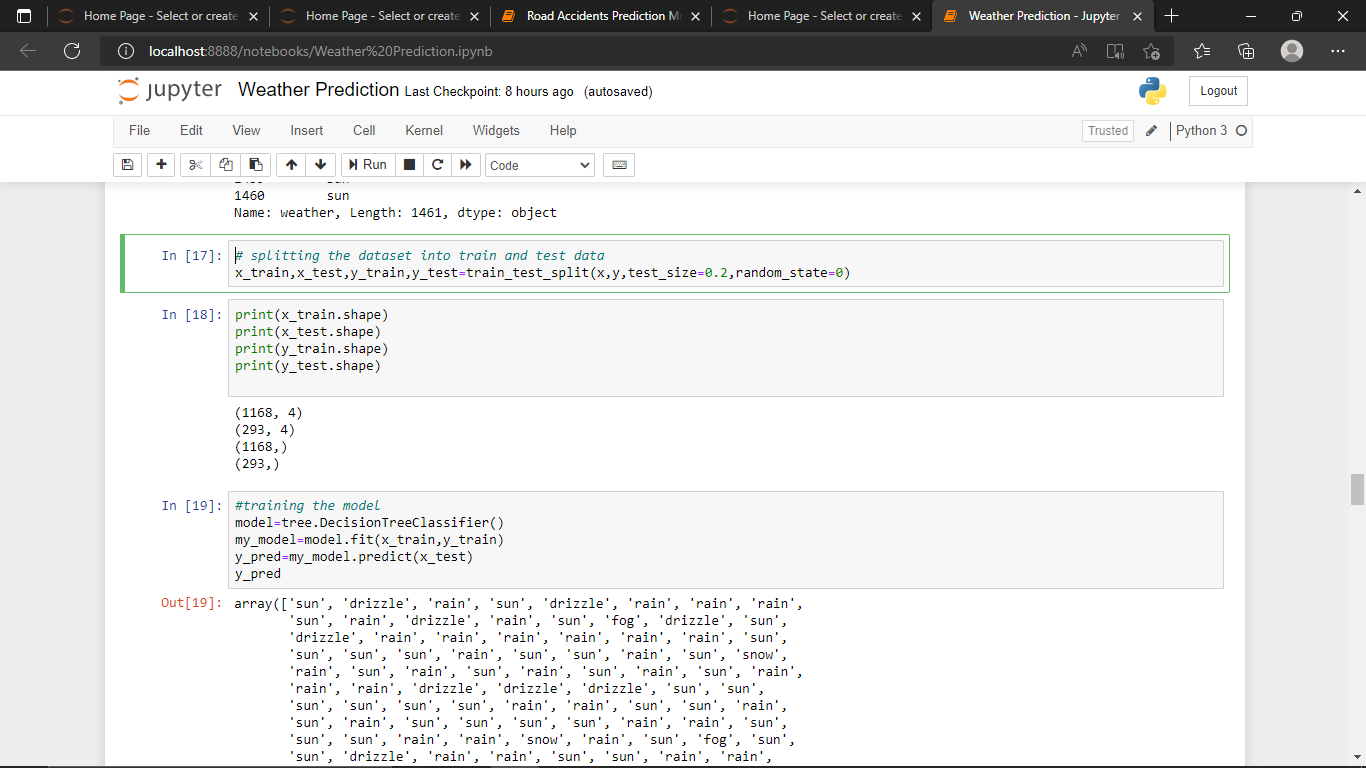
* Making Model using Machine Learning Algorithms

**1.Model making with Decision Tree.**

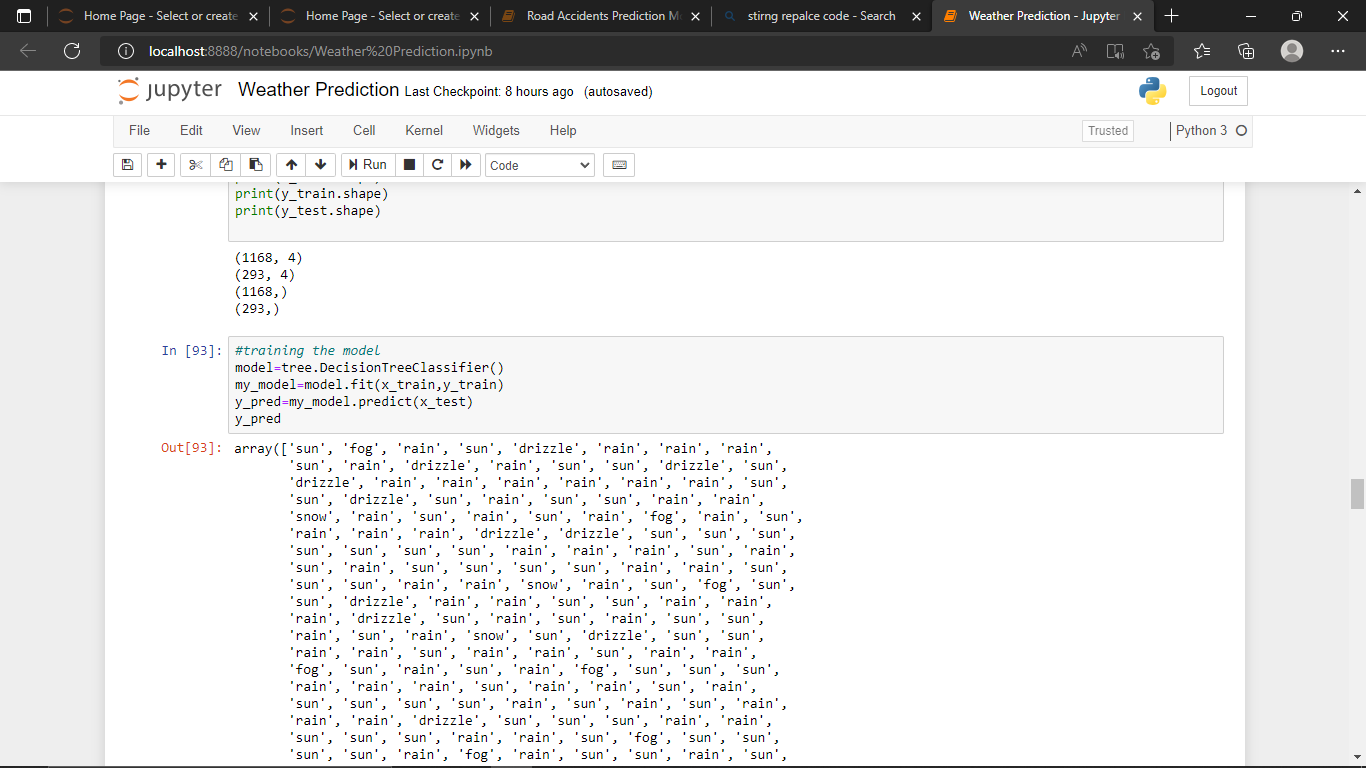
Assigning the values to x and y.



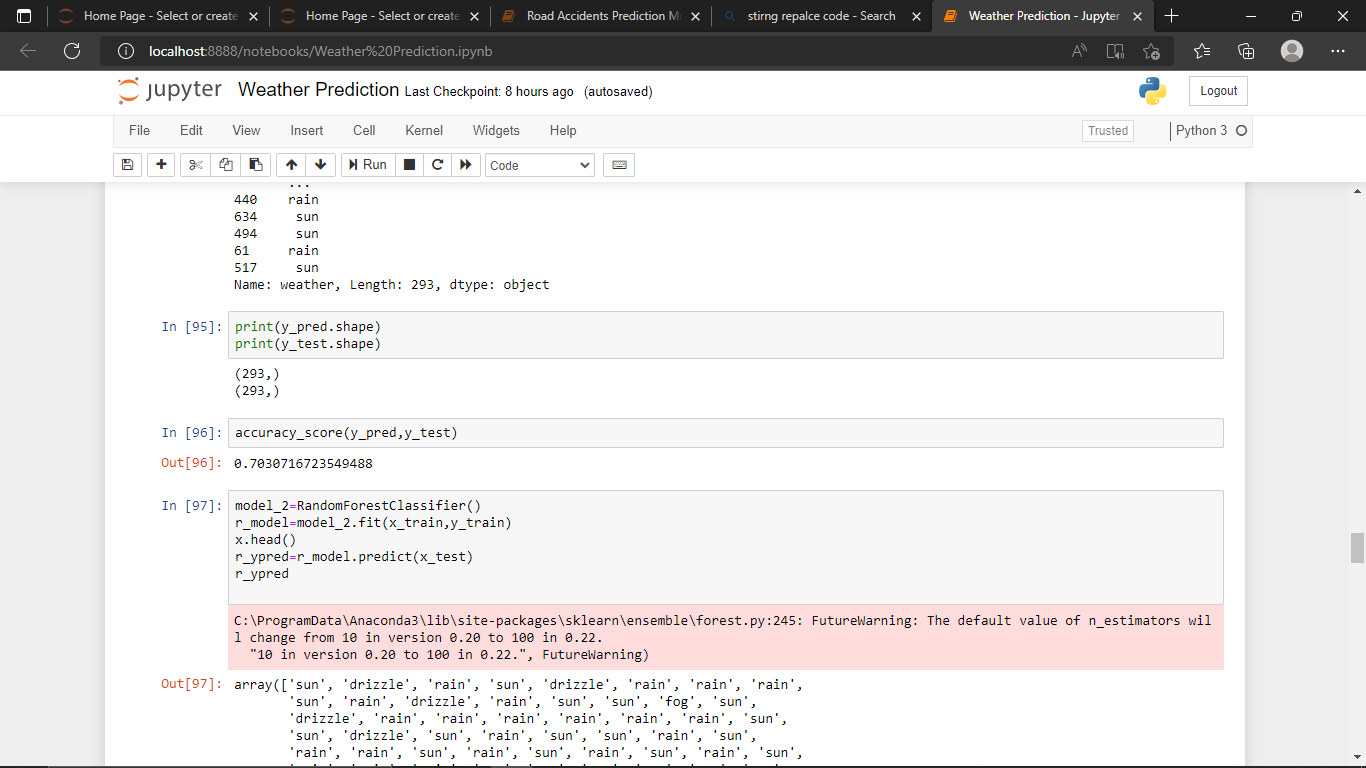
splitting the dataset into train and test data.



Training the model.

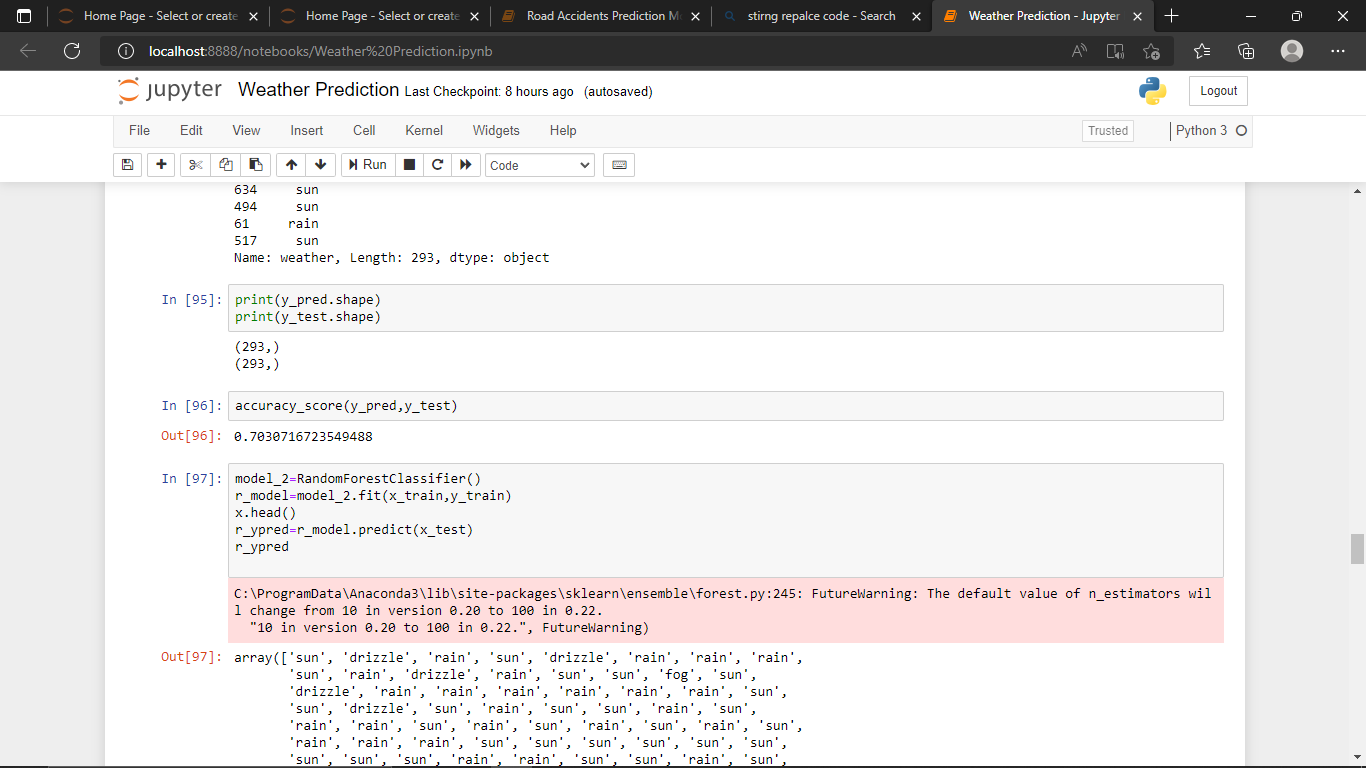


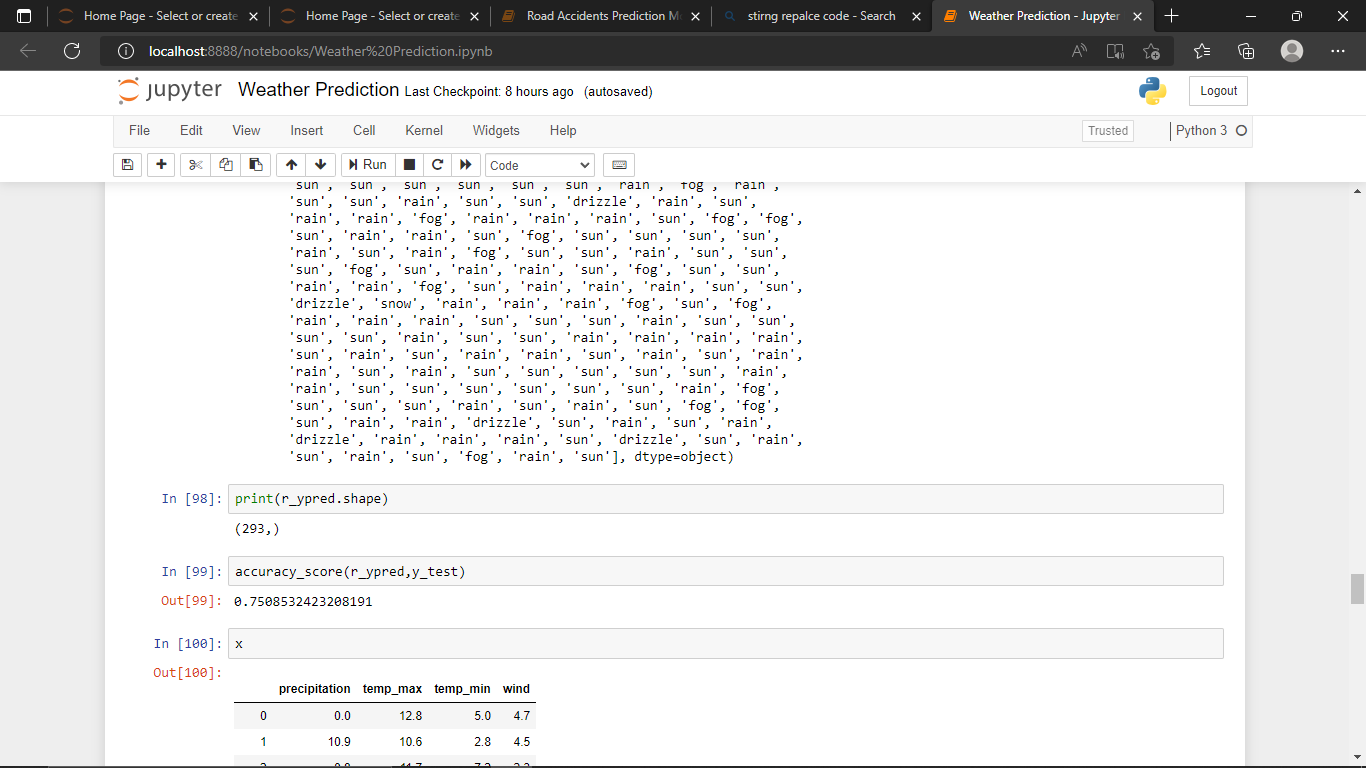
Checking the accuracy to the y\_predicted values to the y\_testing values.



**2.Model making with RandomForestClassifier.**

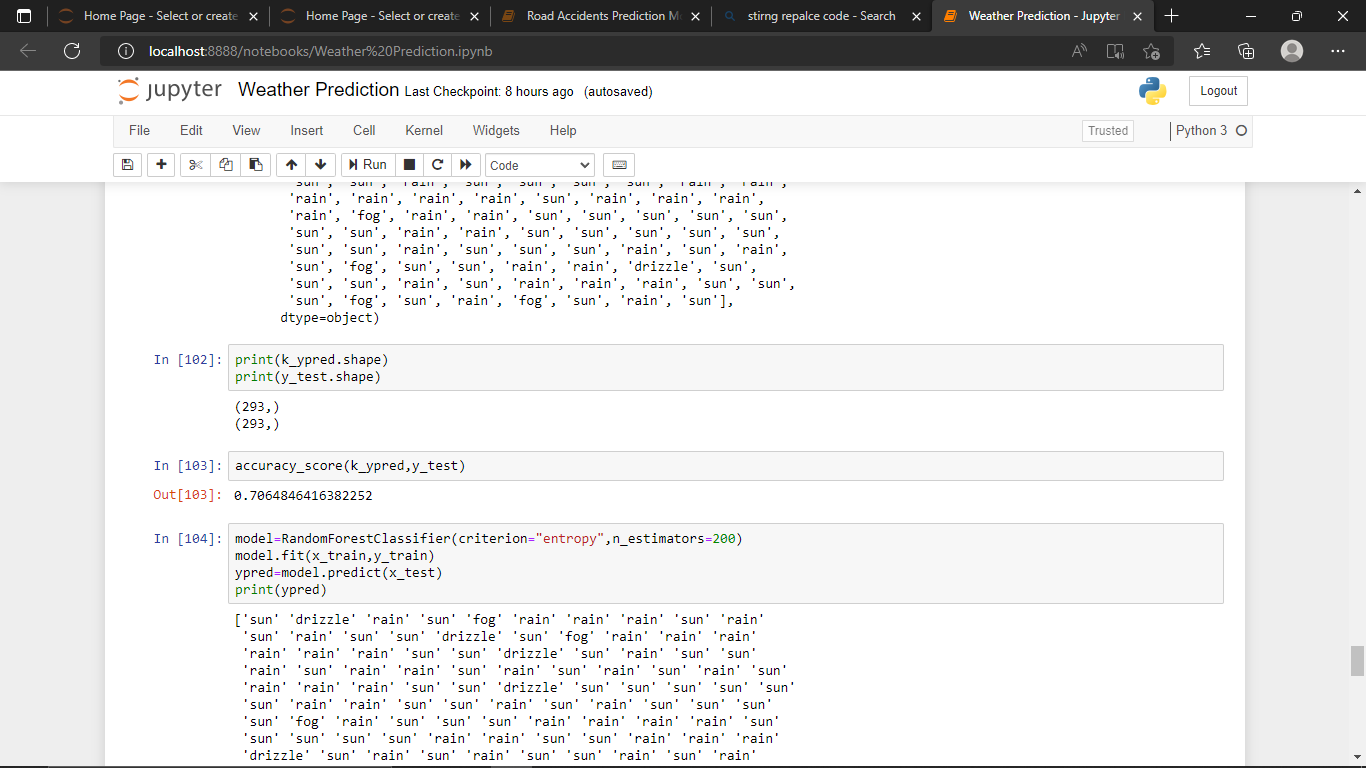
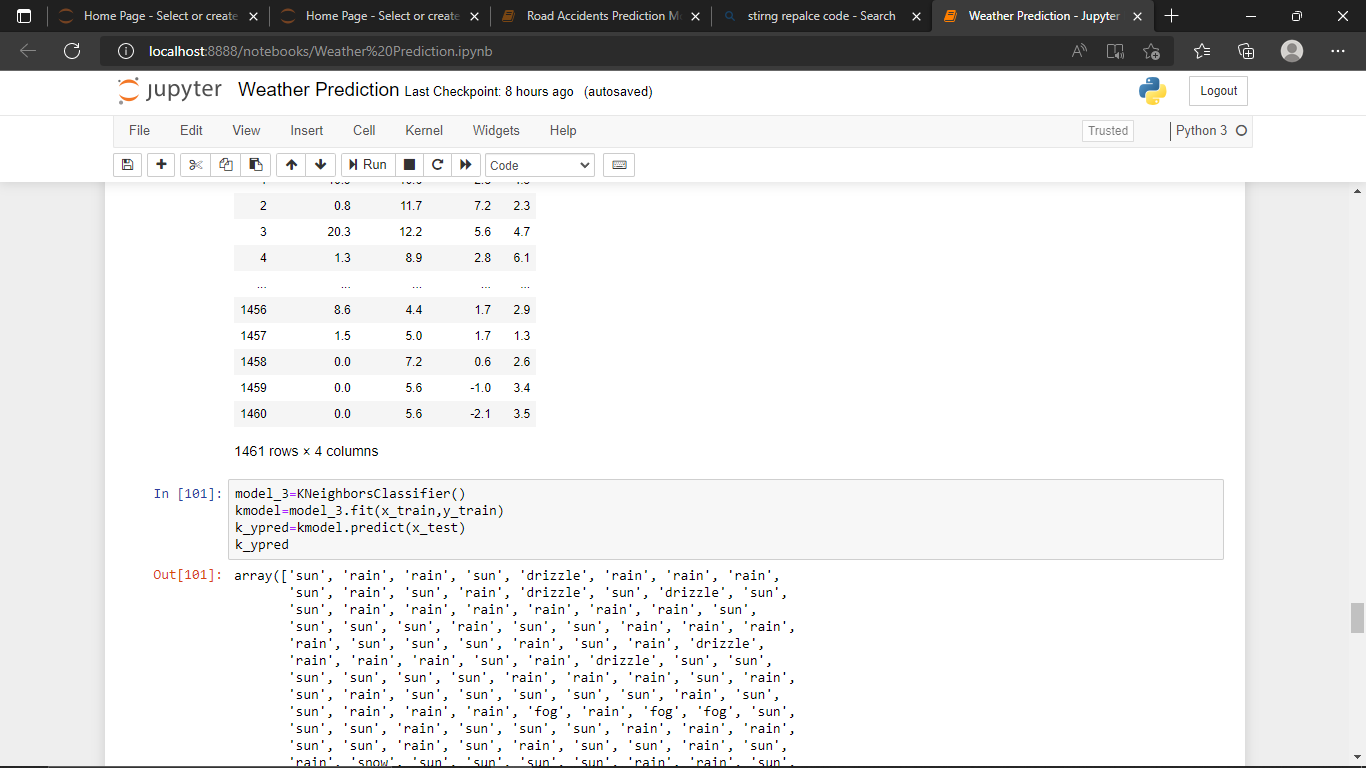
Training the model with RandomForest algorithm.





**3.Model making with KNeighborsClassifier.**

Training model with KNeighbors Algorithm.



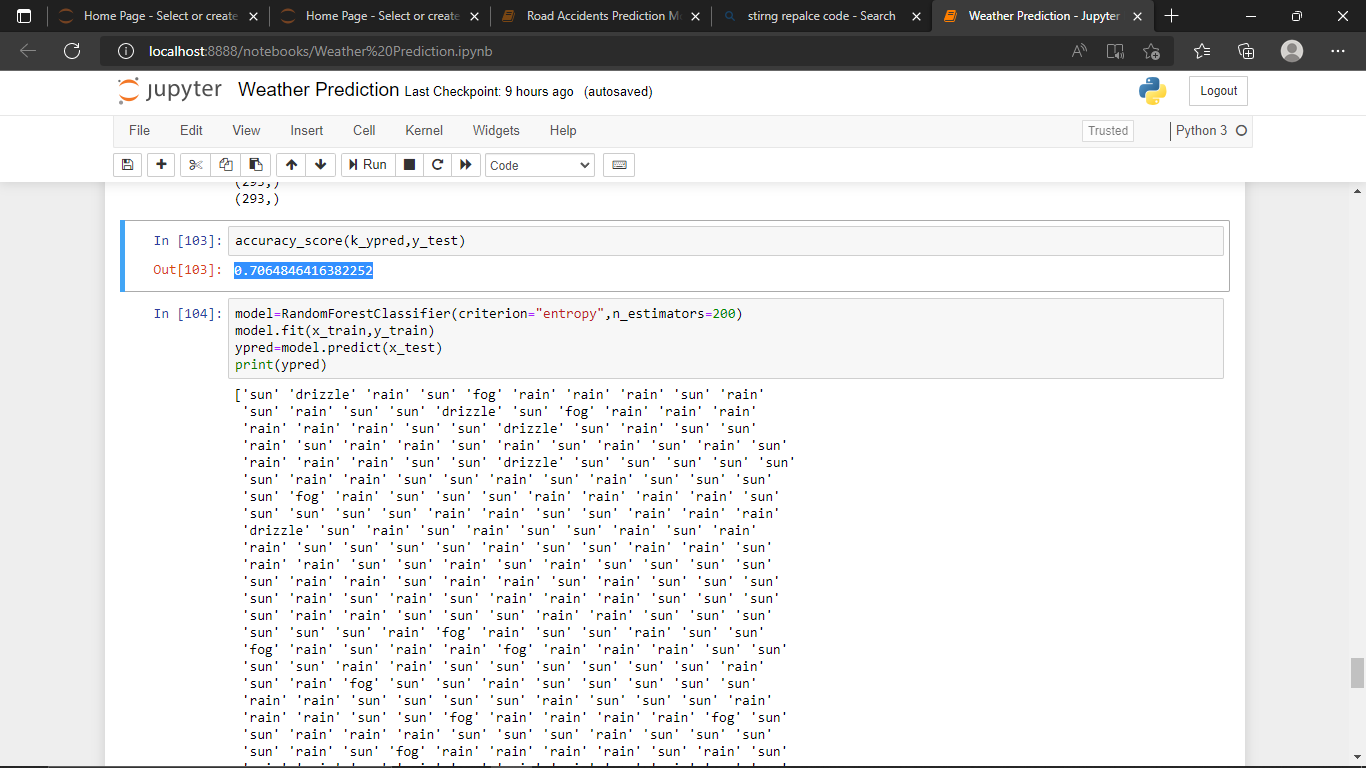
**4.Accuracy\_score of all the models:**

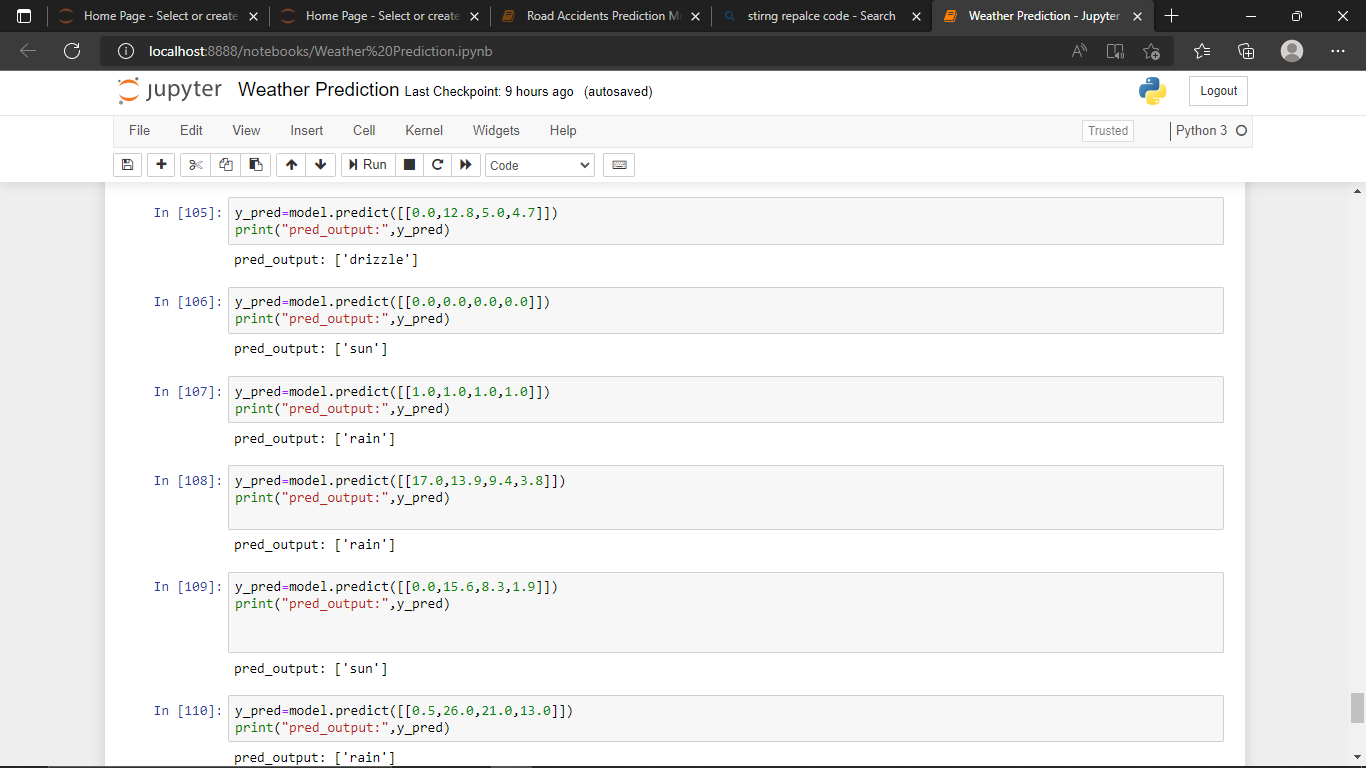
* accuracy\_score of(y\_pred,y\_test) in DecisionTreeClassifier : **0.7030716723549488**
* accuracy\_score of(y\_pred,y\_test) in RandomForestClassifier : **0.7508532423208191**
* accuracy\_score of(y\_pred,y\_test) in KNeighborsClassifier : **0.7064846416382252**

**Accuracy\_score in RandomForestClassifier is high, which means the model performance was great when compared to the other algorithms.**

**So I have decided to fit this model in RandomForestClassifier so that it can give accurate results.**

**4.Fitting and Testing model in RandomForestClassifier.**





**Conclusion:** These are the predicted output that I got when I have given the untrained data and trained data.

I have also taken the factors which are affecting weather like precipitation,min-max temperature,wind speed to calculate the weather conditions from past and future weather report.

**Presenting by:**

**Name : L.SUMA**

**Id : S180841**

**Class : CSE-2C**

**Sub : DSP**