**Predicting Rain with Machine Learning**

Will it rain tomorrow? Let’s build an AI model to answer that.

## *Introduction Natural processes on Earth can be classified into several categories, including hydrological processes like storm waves and groundwater; biological processes like forest growth; atmospheric processes like thunderstorms and rainfall; human processes like urban development; and geological processes like earthquakes. The field of physical geography seeks to investigate the distribution of the different features/parameters that describe the landscape and functioning of the Earth by analyzing the processes that shape it. These features/parameters have been referred to as geophysical parameters in the literature (Karimi, 2014). Rainfall is a key geophysical parameter that is essential for many applications in water resource management, especially in the agriculture sector. Predicting rainfall can help managers in various sectors to make decisions regarding a range of important activities such as crop planting, traffic control, the operation of sewer systems, and managing disasters like droughts and floods (Htike and Khalifa, 2010). A number of countries such as Malaysia and India depend on the agriculture sector as a major contributor to the economy (Htike and Khalifa, 2010; Parmar et al., 2017) and as a source of food security. Hence, an accurate prediction of rainfall is needed*

## 

## Methodology

## *. The review covers several aspects which relate to the input into, output from, and methods used in the various systems devised in the literature for this purpose. The review specifically focuses on studies that use supervised learning for both regression and classification problems. 4 Rainfall Prediction Using Machine Learning Models: Literature Survey 77 Fig. 4.1 Pie chart showing proportions by publication year for papers in this review Google scholar was used to collect papers from 2016 to 2020, with the following key words: (“machine learning” OR “deep learning”) AND (“precipitation prediction” OR “rainfall prediction” OR “precipitation nowcasting”). Almost 1240 results were obtained, and of these only supervised rainfall prediction papers that used meteorological data from, e.g., radar, satellites, and stations were selected, while papers that used data from normal cameras, e.g., photographs were excluded. Even though this review focuses on the prediction of rainfall, the methods used to achieve this can be extended and applied to other geophysical parameters like temperature and wind. Hence, the conclusions and discussions of this chapter can be adapted to other parameters. The total number of reviewed papers are 66, which are a combination of conferences and journal papers published from 2016–2020, except for one paper (Shi et al., 2015) which was published in 2015 and is a seminal work in this field. Figure 4.1 shows the reviewed studies per year. Tables which summaries the reviewed paper can be found in Appendices 1 and 2. Figure 4.2 shows the generic structure of supervised ML models. This structure was used as a guideline to construct a set of questions used to systematically categorize and analyze the 66 papers*

## Input Data Pre-processing

## *Before ML tools are applied to make predictions on the available data, the input data is usually pre-processed to reformat the data into a form that will make training of, and prediction by, the ML tool(s) easier and faster. The pre-processing techniques usually applied in geophysical parameter forecasting can be broken down into three broad categories, namely data imputation; feature selection/reduction; and data preparation for classification*

## Data Imputation

## *Data sets are regularly found to have missing data entries, which is caused by a range of factors such as data corruption, data sensor malfunction, etc. This is a serious issue faced by researchers in data mining or analysis and needs to be addressed as part of pre-processing before feature selection/preparation and training. The techniques used to infer and substitute missing data are collectively referred to as data imputation techniques. Data imputation is challenging and is an on-going research area*

## Feature Selection/Reduction

## *Feature selection/reduction aims to determine and use salient features in the data, and disregard irrelevant features in the data. This helps to reduce training time, decrease the model complexity and increase its performance. In the papers in this review, it is observed that feature selection is carried out either automatically or manually*

## Data Preparation for Classification

## *When attempting to carry out classification into discrete classes, it is either necessary to use a data set in which the desired output variable is discrete or to convert a desired continuous-valued output variable into discrete classes. This involves setting the desired number of classes, which is usually done manually and arbitrarily, followed by determining the range of values represented by each class, i.e., determining the thresholds that divide the continuous scale into the desired classes. Finally, where the number of instances across classes is imbalanced, it is necessary to balance them.*

## Problem Statement

*The weather has a significant impact on the agricultural industry and because of that, being able to predict it helps farmers in their day-to-day decisions such as how to plan efficiently, minimize costs and maximize yields.*

*A major agricultural company needs you to help them maximize growth efficiency, save resources and optimize their production.*

*To achieve these things, the company needs to have an accurate weather prediction algorithm that will improve their decision-making on typical farming activities such as planting and irrigating.*

*Using historical weather information from their region,****can you predict what the weather will be like in the next few days****?*

**The goal 🥅**

Predict the next day’s weather based on three labels.

* N— No rain
* L— Light rain
* H — Heavy rain

Let’s now look at the data

**The data**[**💾**](https://emojipedia.org/floppy-disk/)

📂 **train**  
 ├── region\_A\_train.csv  
 ├── region\_B\_train.csv  
 ├── region\_C\_train.csv  
 ├── region\_D\_train.csv  
 ├── region\_E\_train.csv  
 ├── solution\_format.csv  
 └── solution\_train.csv📂 **test**  
 ├── region\_A\_test.csv  
 ├── region\_B\_test.csv  
 ├── region\_C\_test.csv  
 ├── region\_D\_test.csv  
 └── region\_E\_test.csv

The data has been conveniently split into train and test datasets.

In each train and test, you’re given weather data which consists of anonymized locations named region A through region E, which are all neighboring regions.

Here’s a look at the first five rows of region\_A\_train.csv

The first thing you should notice is that the date column isn’t a date but was anonymized to be some random value.

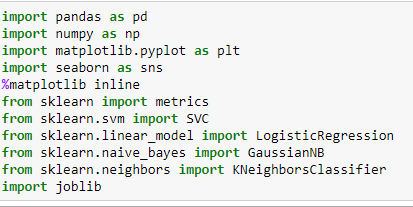
There is a total of 10 features, which are composed of temperature, precipitation, wind speed, wind speed direction, and atmospheric pressure

Then, looking at solution\_format.csv

We can utilize the date column to join it with the training data and build a model.

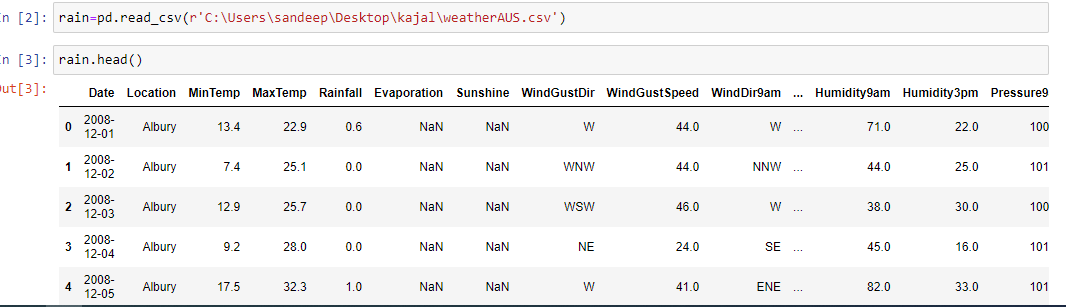
## Load Libraries

Next, we load up some essential libraries for visualizations and machine learning.



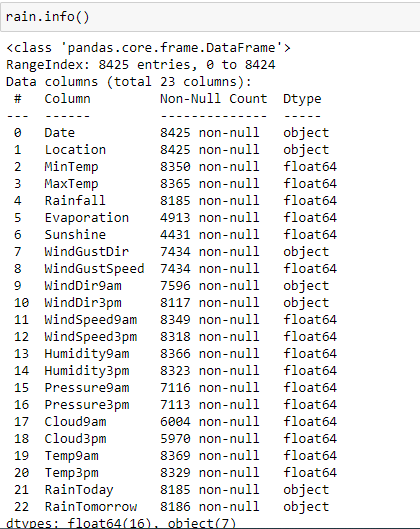
## Load the data

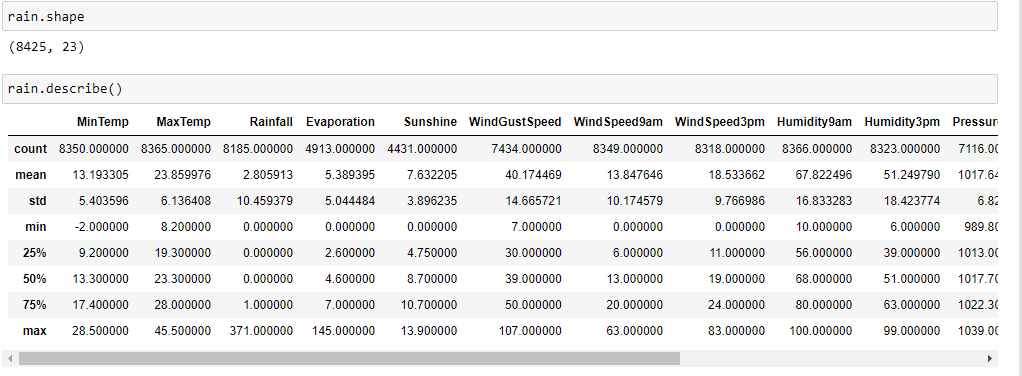
Let’s read in all the data we have.

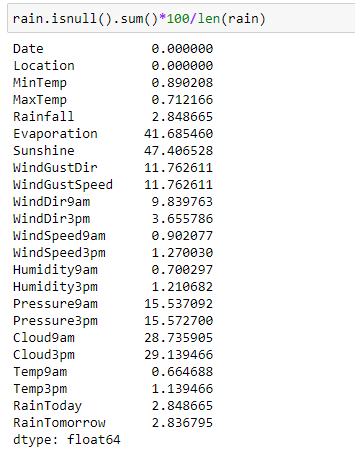


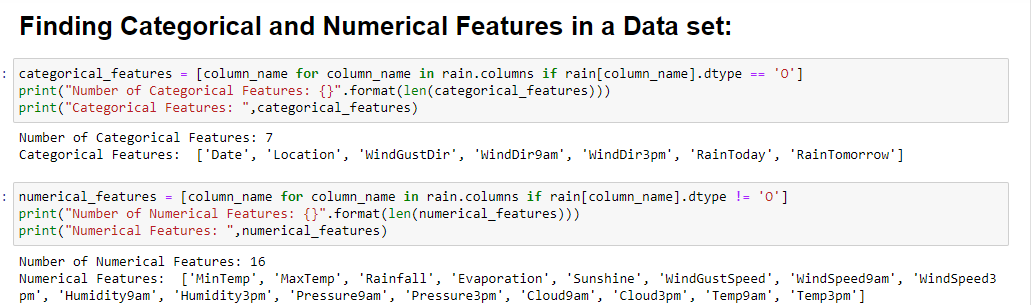
# Exploratory Data Analysis

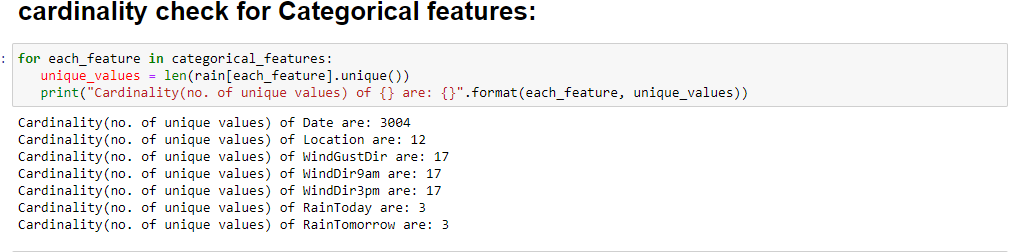
It’s time for the fun part, visualizing the data.

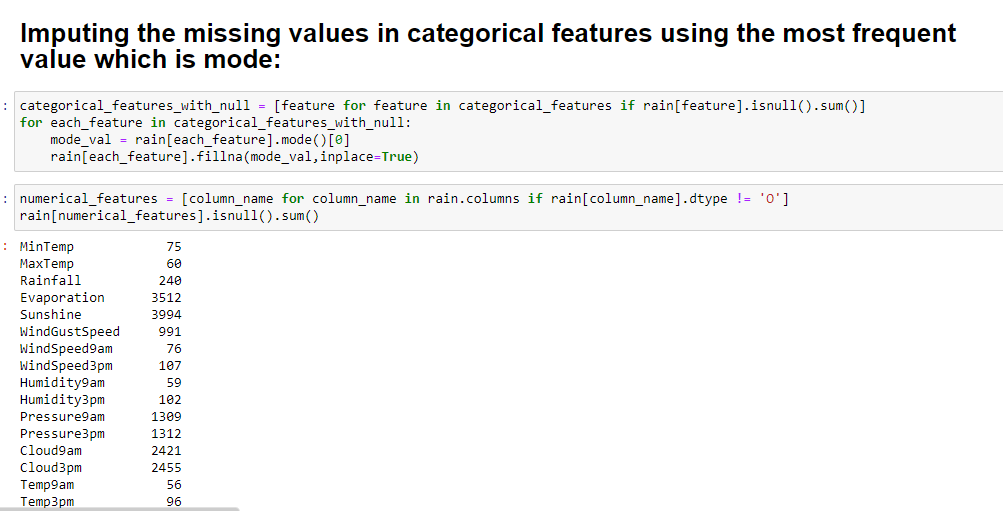


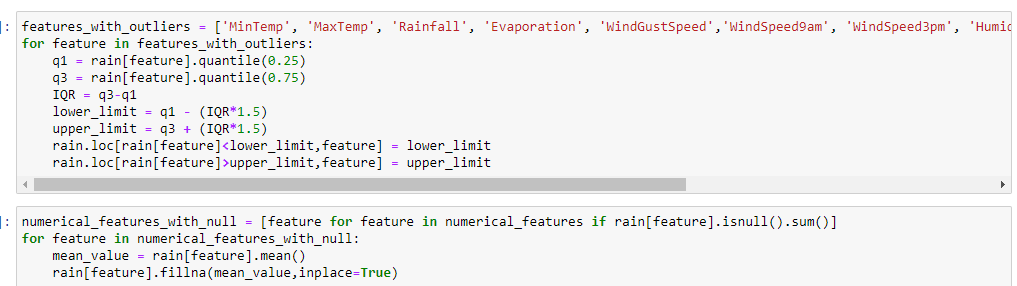


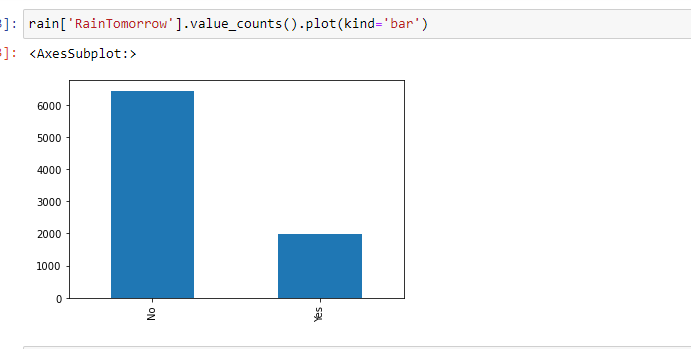




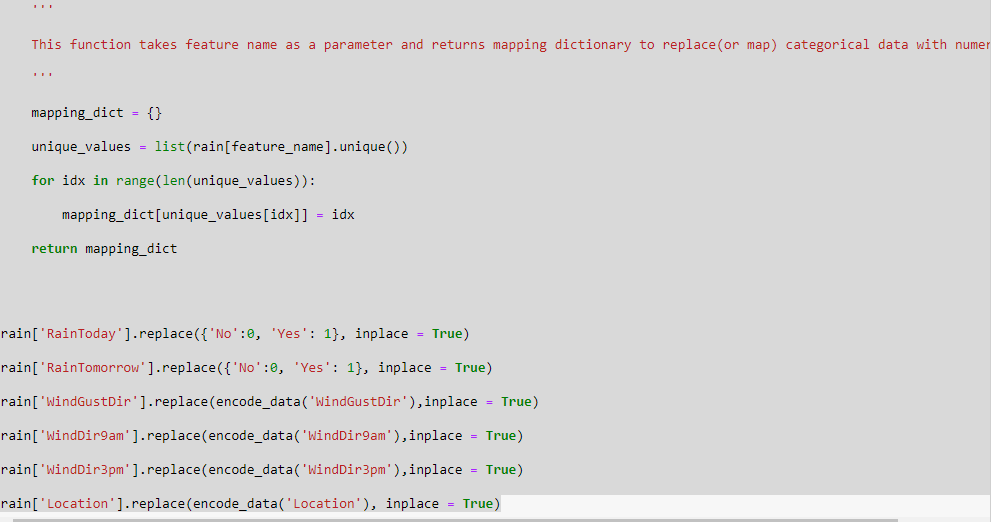




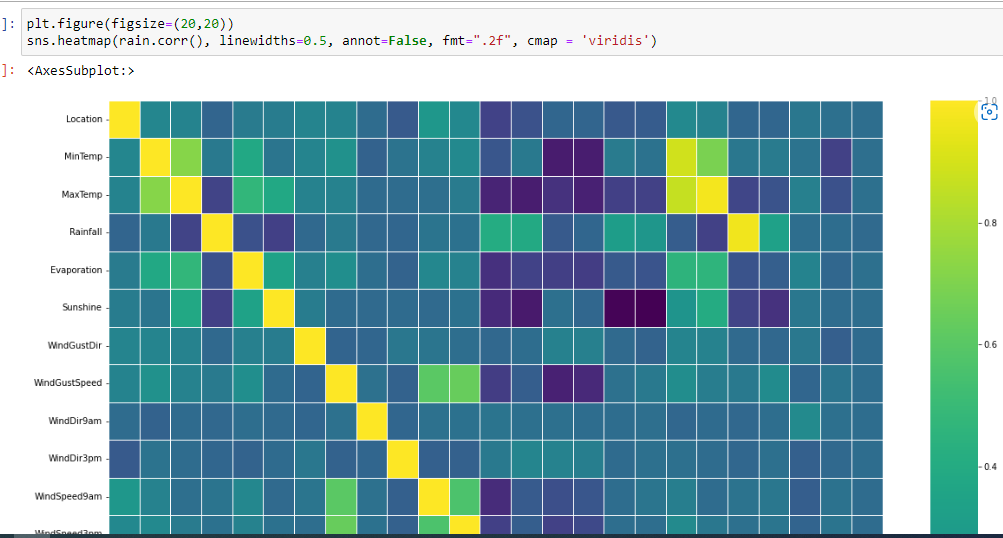




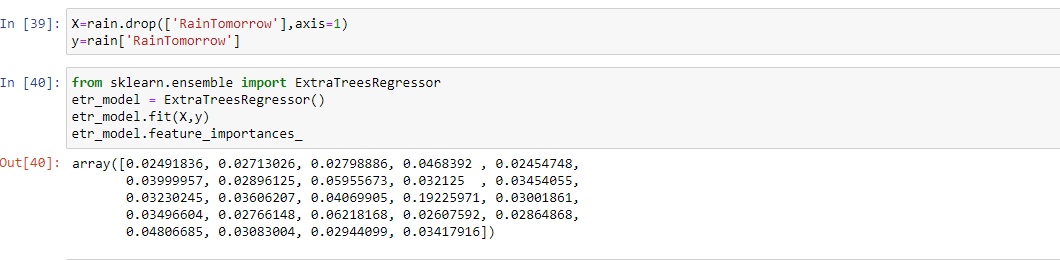
# Encoding of Categorical Features:[¶](http://localhost:8888/notebooks/rainfall%20prediction.ipynb#Encoding-of-Categorical-Features:)



**CORRELATION**

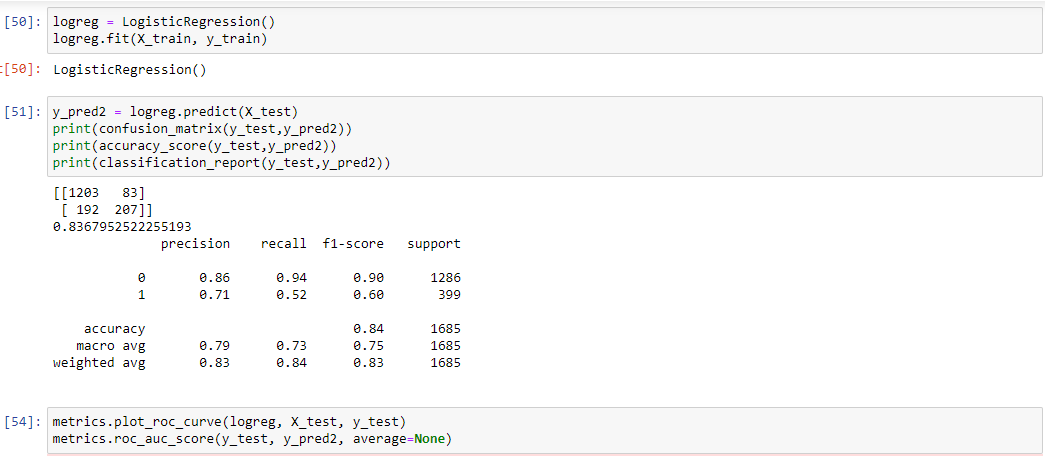
****

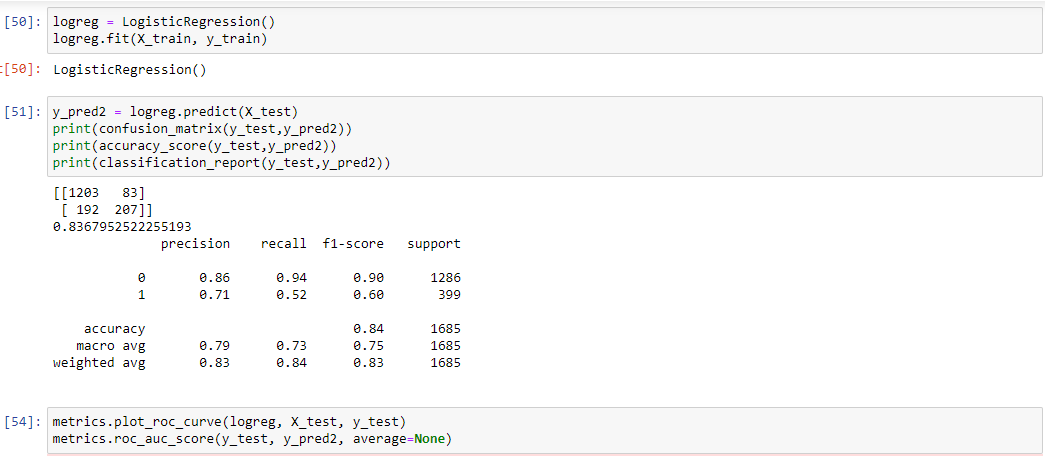
**SPLIT THE DATA**

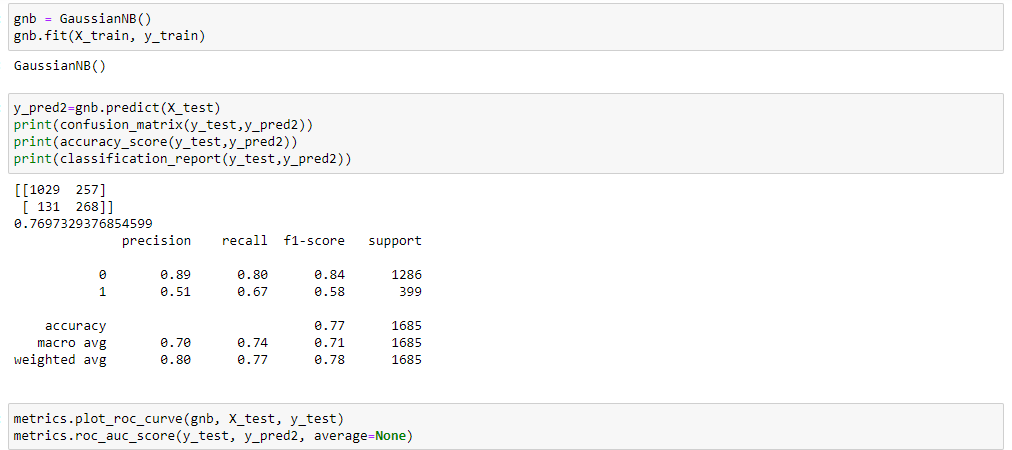
****

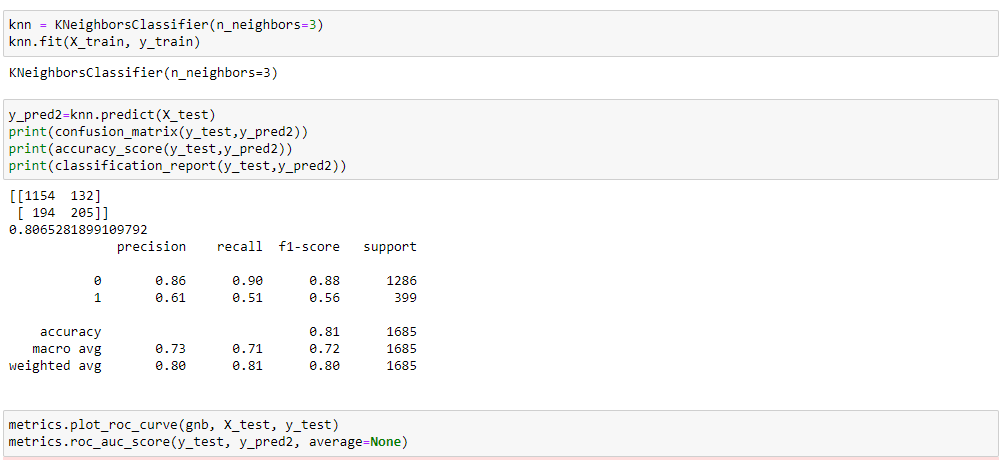
****

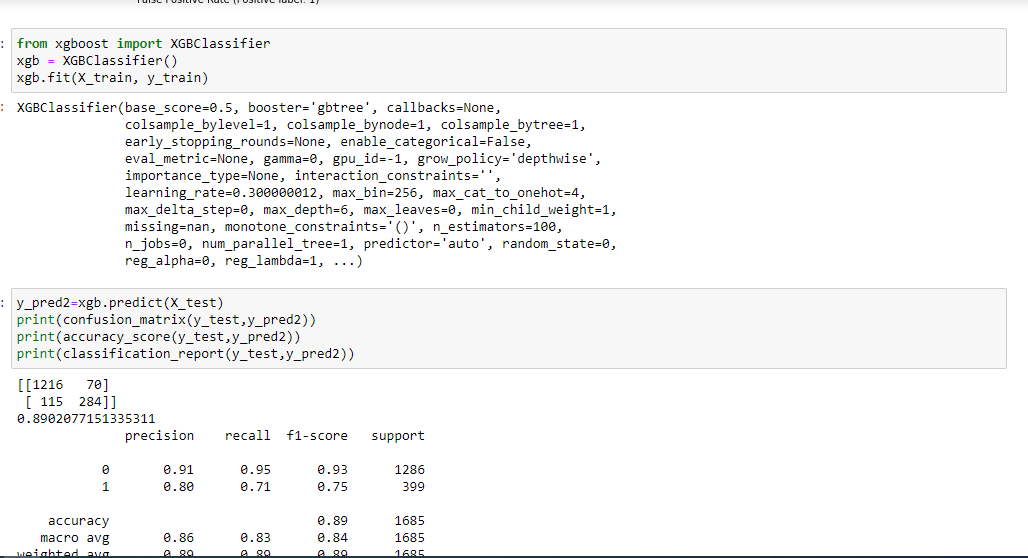
**MODELS USED FOR PREDICTION**

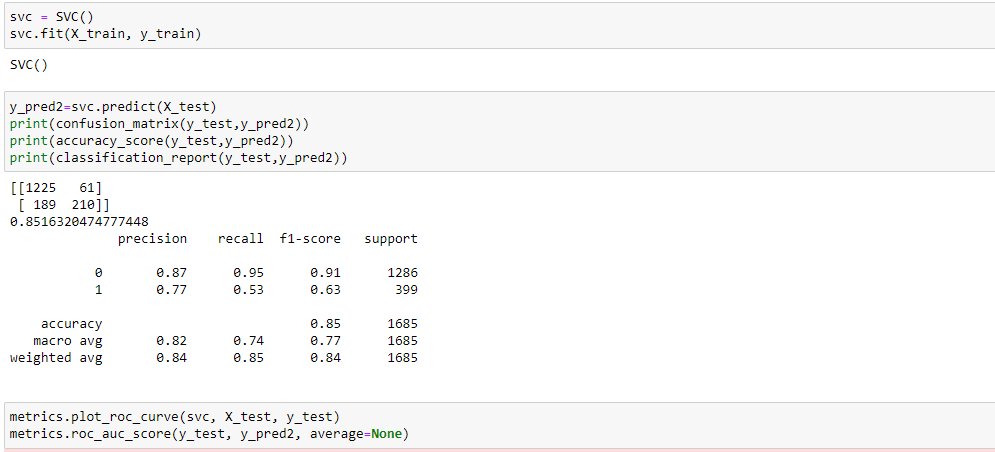












**SAVE THE MODEL**

