A blog post about Global Power, by Sharad Yadav

Photo credit: - nsenergybusiness.com

Problem Statement: The Global Power Plant Database is a comprehensive, open source database of power plants around the world. The database covers approximately 35,000 power plants from 167 countries and includes thermal plants (e.g. coal, gas, oil, nuclear, biomass, and waste, geothermal) and renewables (e.g. hydro, wind, solar). Each power plant is geolocated and entries contain information on plant capacity, generation, ownership, and fuel type. It will be continuously updated as data becomes available. We will do analysis and prediction of Primary fuel type from 'Fuel Type' attribute.

Current scenario of India in power generation: - Across the world India is the third largest power producer and third largest consumer of electricity. The national electric grid in India has an installed capacity of

383.37 GW as of 31 May 2021. Out of which 95.7 GW of renewable energy capacity, and represents ~ 25% of the overall installed power

India is targeting about 450 Gigawatt (GW) of installed renewable energy capacity by 2030 – about 280 GW (over 60%) is expected from solar. 500 billion USD investment required to meet 450 Gigawatt(GW)

2. Data Analysis: - In this study I am going to explore data of India's power generation using primary fuel generation type's insight with the help of sklearn's library.

```
#import imp library
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split ,GridSearchCV
from sklearn.metrics import r2_score,mean_absolute_error,confusion_matrix,classification_report,accuracy_score,precision_
import statsmodels as sm
import matplotlib.pyplot as plt
import scikitplot as skplt
from matplotlib import pyplot
from collections import Counter
import warnings
import math
warnings.filterwarnings('ignore')
```

```
1.This Question contains 2 target variable s per problem statement first i will solve classification for fuel type.
2.second is regression problem to predict capacity_mw(megawat)  so here our target variable is capacity_mw
```

```
#read csv file and see top data overview to attemp classification problem first then will go for regression
power=pd.read_csv(r"C:\Users\INPshy\Desktop\DATA Science\database_IND.csv")
power.head()
```

Out[2]:

	country	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	...	geolocation_source	wepp_i
0	IND	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	NaN	...	National Renewable Energy Laboratory	Nal
1	IND	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	...	WRI	Nal
2	IND	India	AES Saurashtra	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	...	WRI	Nal

Image: - screenshot of sklearn library

You can find the code on git hub, click on below link

https://github.com/sharadyadav1988/Evaluation-/blob/main/Global%20Power%20plant.ipynb

The original dataset contains 25 features columns and 908 rows after cleaning data 20 features columns and 908 rows left for analysis including target variable.

Moving forward analysis of power generation with primary fuel types as Coal and hydro power major contribution in power generation followed by Solar and wind.

Keeping in mind to reduce carbon footprint from the world, moving towards clean and green sustainable energy that is solar and wind

energy. Biomass is also an option to generate power through waste management.

India's target of about 450 Gigawatt (GW) of installed renewable energy capacity by 2030 – about 280 GW (over 60%) is expected from solar.

As per pie chart analysis of 2017 data solar contribution is 14% only

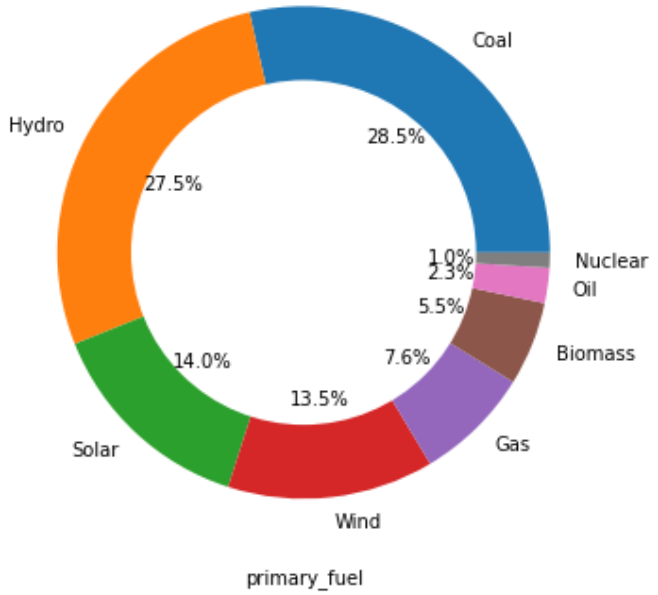


Image:- Piechart of primary Fuel types

Which will incrase upto 60% of total power generation by 2030, Currently Adani Green ,Tata Power ,JSW energy few others company are inline with government polices to achive this target.New investment in clean energy in the country reached 11.1 billion US dollar in 2018.

EDA Concluding Remarks: - analysing data types found that some columns are object type some are numerical, object data treated with encoding techniques, further checking null values of each columns

Plotting distribution plot shows cpacity_megawat column data right skewed ,latitude seems normal distribution , longitude slightly right skewed,commissioning_year left skewed,owner right skewed,source

right skewed,url right skewed,geological source only 2 category,generation_gwh,2013,2014,2015,2016,2017 data right skewed

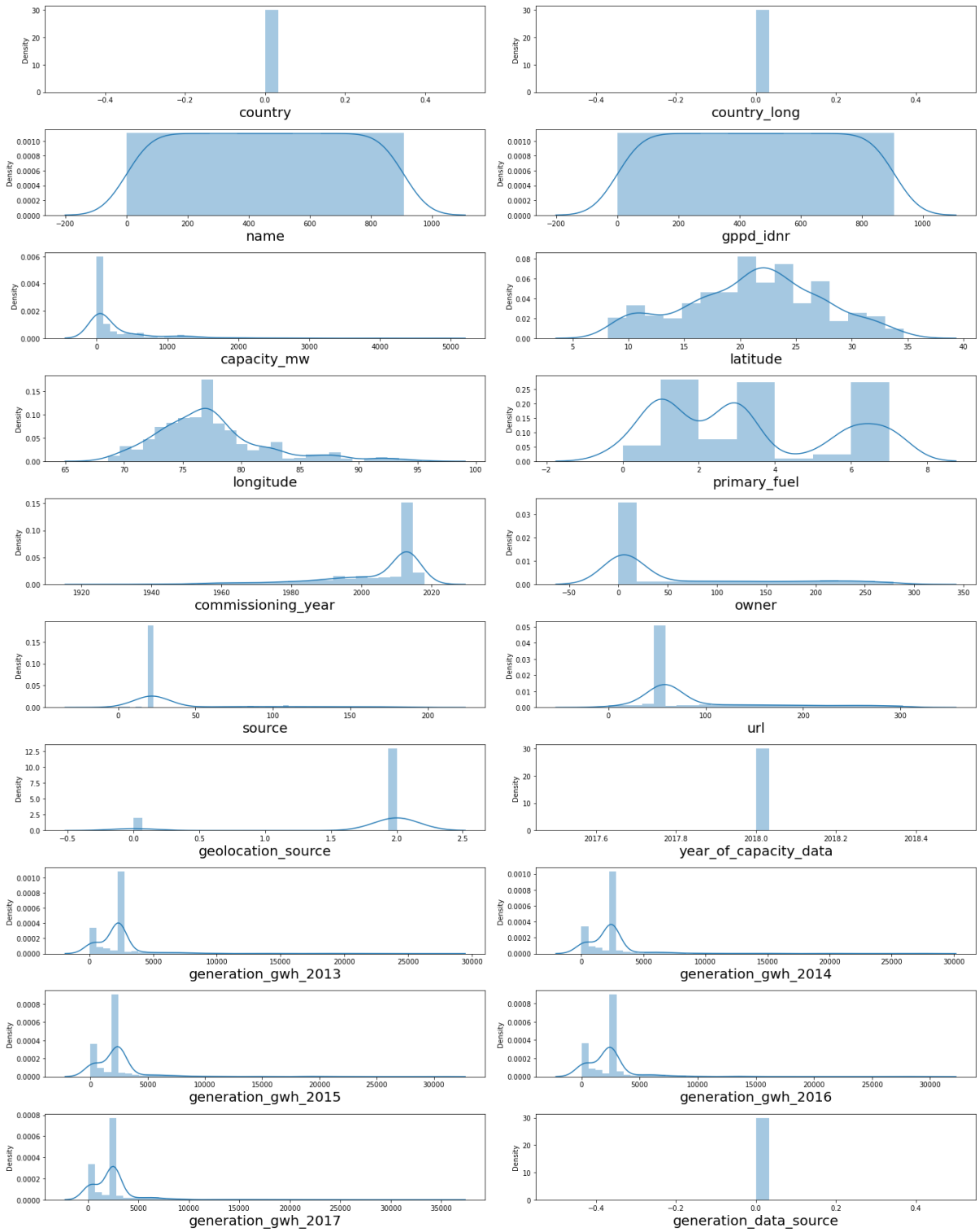


Image: - Distribution plot of dataset

After visualization of distribution plot I have got idea about skewness and outliers present in continuous dataset. Then I have reconfirms with boxplot to check exact which columns have how much outliers in dataset.

Moving forward to remove outliers from dataset with Zscore method. After that heatmap plot to check multicollinearity between features columns

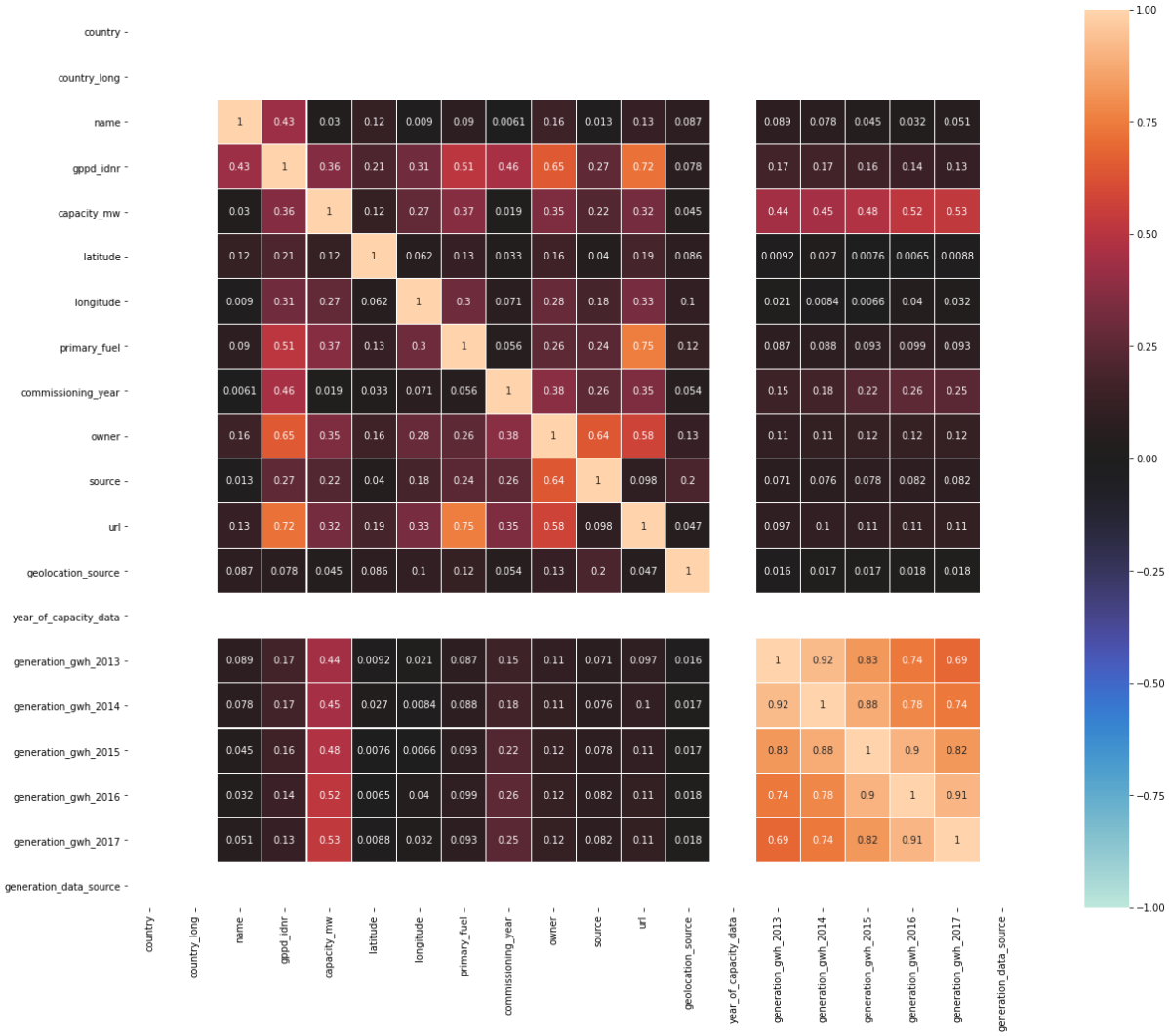


Image:- Heatmap to check features correlation

Further visualization with help of scatter plot between ‘generation_gwh_2013 , 2014, 2015, 2016, 2017 refer below scatter plot visualization ,find out that not so strong positive linear relationship between features ,multicollinearity nullified.

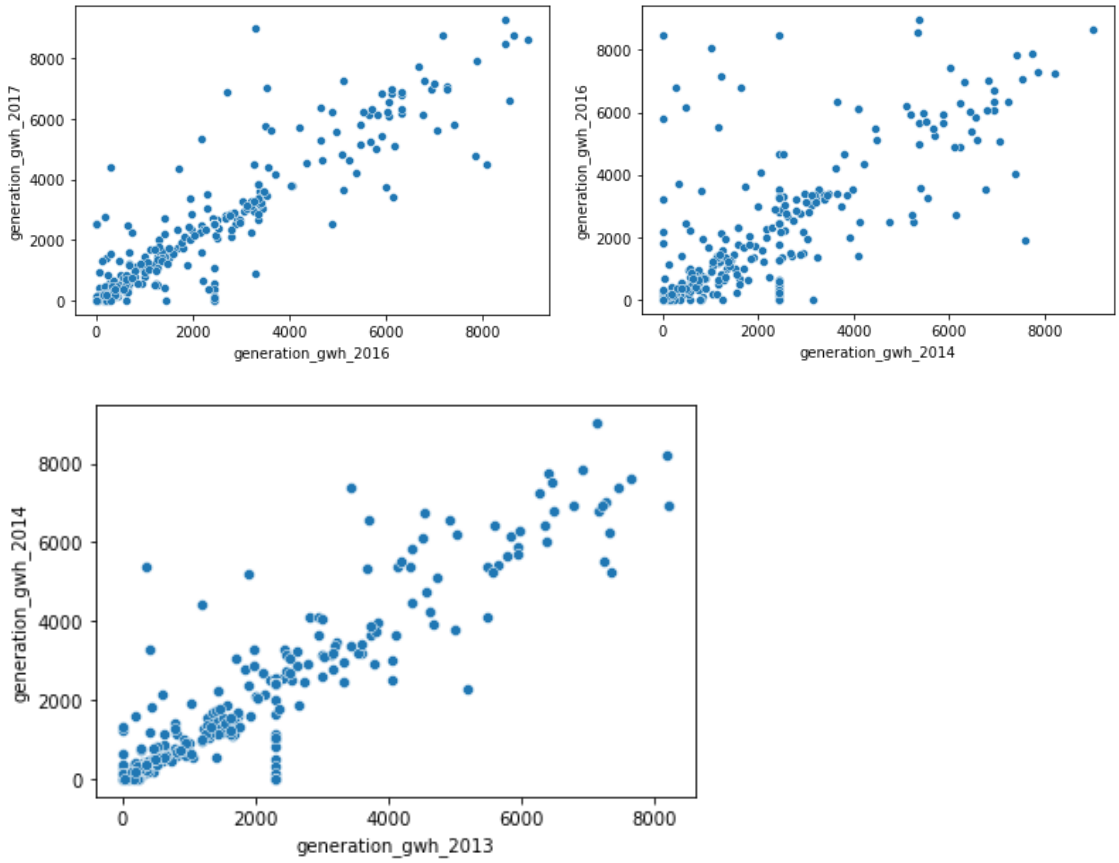


Image: - Scatter plot to identify features mutual relationship

Further I investigated class i.e. 'primary fuel' with help of countplot so total 8 types fuel present in dataset (Coal,Hydro,Solar,Wind,Gas,Biomass,oil,Nuclear) class is imbalanced Coal, Hydor ,Solar ,Wind are in majority which is balanced by SMOTE

method which imported from imblearn.over_sampling librar

```
6]:
#resample of imbalanced dataset
oversample = SMOTE()
X, Y = oversample.fit_resample(X, Y)
# summarize distribution
counter = Counter(Y)
for k,v in counter.items():
    per = v / len(Y) * 100
    print('Class=%d, n=%d (%.3f%%)' % (k, v, per))
# plot the distribution
pyplot.bar(counter.keys(), counter.values())
pyplot.show()
```

```
Class=6, n=217 (12.500%)
Class=7, n=217 (12.500%)
Class=1, n=217 (12.500%)
Class=3, n=217 (12.500%)
Class=2, n=217 (12.500%)
Class=0, n=217 (12.500%)
Class=5, n=217 (12.500%)
Class=4, n=217 (12.500%)
```

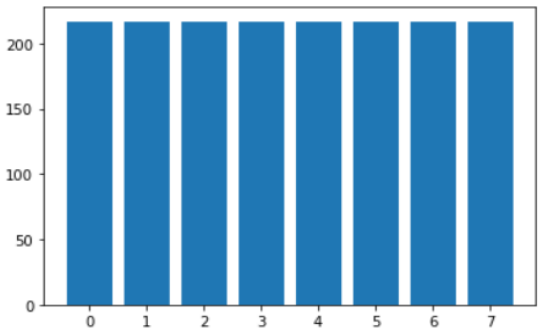


Image: - balanced class of primary fuel types

Features Selection: - As per problem statement I have to predict primary fuel types(classification) present in dataset Class is label with Y and independent features label with X, then features data scale and train test split with test size of 25 % data, and train data 75% taken from original dataset.

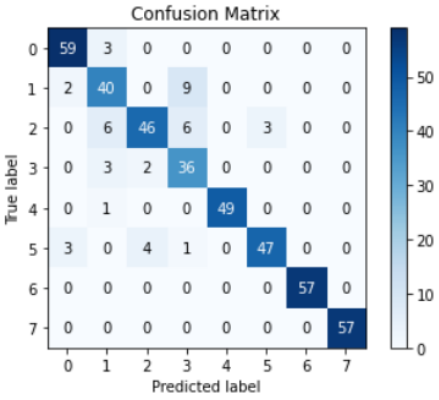
```
#scale feature data
scaler=StandardScaler()
X_scaler=scaler.fit_transform(X)
```

```
#train test split
x_train,x_test,y_train,y_test=train_test_split(X,Y, test_size=.25,random_state=10)
```

Image: - screenshot of features scaling and traintestsplit

Model Building: - From sklearn library import DecesionTreeClassifier building model, DecesionTree model often issue of overfitting DecesionTree score is 1, but after hyper parameter tuning DT model score is 80% and cross validation score is 78%.

Confusion matrix and classification report of DecesionTreeClassifer both are giving quite good result




```
print(classification_report(y_pred,y_test))
```

	precision	recall	f1-score	support
0	0.92	0.95	0.94	62
1	0.75	0.78	0.77	51
2	0.88	0.75	0.81	61
3	0.69	0.88	0.77	41
4	1.00	0.98	0.99	50
5	0.94	0.85	0.90	55
6	1.00	1.00	1.00	57
7	1.00	1.00	1.00	57
accuracy			0.90	434
macro avg	0.90	0.90	0.90	434
weighted avg	0.91	0.90	0.90	434

Image: - DecesionTreeClassifier Confusion matrix &classification report

Note: - This analysis is multiclass classification where Target variable (Class) is more than two class.

Similarly I have used sklearn library to build Logistic regression and GradientBoostingClassifier to achieve more accuracy, high model score finally I have found GradientBoostingClassifier perform better among all model

Below image of Confusion matrix reflecting that model is quite accurate and classification report also give precision recall f1-score good score for each class my model is not biased for any class, means its good trained for each class.

```
print('**** confusion matrix post tuning****')
print(confusion_matrix(y_test,y_pred))
```

```
**** confusion matrix post tuning****
[[64  0  0  0  0  0  0  0]
 [ 3 41  6  1  1  1  0  0]
 [ 0  1 48  1  0  2  0  0]
 [ 0  8  4 40  0  0  0  0]
 [ 0  0  0  0 49  0  0  0]
 [ 1  1  0  0  0 48  0  0]
 [ 0  0  0  0  0  0 57  0]
 [ 0  0  0  0  0  0  0 57]]
```

```
print('********************Classification Report************
print(classification_report(y_pred,y_test))
```

```
********************Classification Report********************
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        68
           1       0.77      0.80      0.79        51
           2       0.92      0.83      0.87        58
           3       0.77      0.95      0.85        42
           4       1.00      0.98      0.99        50
           5       0.96      0.94      0.95        51
           6       1.00      1.00      1.00        57
           7       1.00      1.00      1.00        57

    accuracy                           0.93       434
   macro avg       0.93      0.93      0.93       434
weighted avg       0.94      0.93      0.93       434
```

Image: -confusion matrix & classification report of GradientBoostingClassifier Model

After analysis of DecesionTree ,Logistic Regression,and GradientBoostingClassifier plotting ROC for each model and cross validation score decides that GradientBoostingClassifier is best model among all ,save GradientBoostingClassifer to use in future for analysis

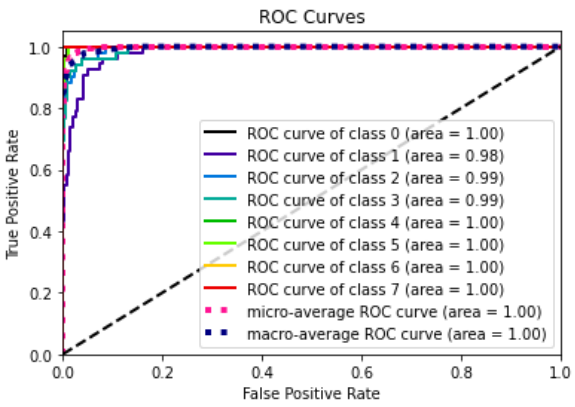


Image: - ROC plot of GradientBoostingClassifier

I hope that I have done my bit to analysis of power generation Primary fuel type's classification, you can do further more investigation and apply different techniques approach to improvise and contribute towards future clean and green energy to make the world a better place.

Conclusion: - Power is essential component of infrastructure, crucial for the economic social growth. The world has focused on ESG(Environmental, Social, Governance) parameters now. Electricity demand in the country has increased rapidly and is expected to rise further in the years to come. In order to meet the increasing demand for electricity in the country, massive addition to the installed generating capacity is required.

The government focused on clean energy to promote sustainable industrial growth, for that by 2030 share of renewable energy generation would increase from 18% to 44%while thermal is expected to reduce from 78% to 52%.

AI has the potential to cut energy waste, lower energy costs, and facilitate and accelerate the use of clean renewable energy sources in power grids worldwide. AI can also improve smart energy meter to control of power system.

Thank you for reading my blog!