PROJECT TITLE : Air Quality Analysis and prediction in Tamil Nadu

Algorithm:

#import libraries

import pandas

#Dataset reading and activities

pandas.set\_option("display.max.rows",None)

pandas.set\_option("display.max.columns",None)

file\_data = pandas.read\_csv(r"Air quality analysis dataset.csv")

print(file\_data)

print(file\_data.head(200))

print(file\_data.tail(100))

print(file\_data.describe())

print(file\_data.info())

#Dataset cleaning

from sklearn.preprocessing import LabelEncoder

my\_le=LabelEncoder()

#Train and Testing

from sklearn.datasets import make\_classification

value1, y = make\_classification(

n\_features=6,

n\_classes=2,

n\_samples=800,

n\_informative=2,

random\_state=66,

n\_clusters\_per\_class=1,

)

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

value1, y, test\_size=0.33, random\_state=125

)

from sklearn.naive\_bayes import GaussianNB

model = GaussianNB()

model.fit(X\_train, y\_train)

predicted = model.predict([X\_test[6]])

print("Actual Value:", y\_test[6])

print("Predicted Value:", predicted[0])

from sklearn.metrics import (

accuracy\_score,

confusion\_matrix,

ConfusionMatrixDisplay

)

#Accuracy prediction

y\_pred = model.predict(X\_test)

accuray = accuracy\_score(y\_pred, y\_test)

print("Accuracy:", accuray)

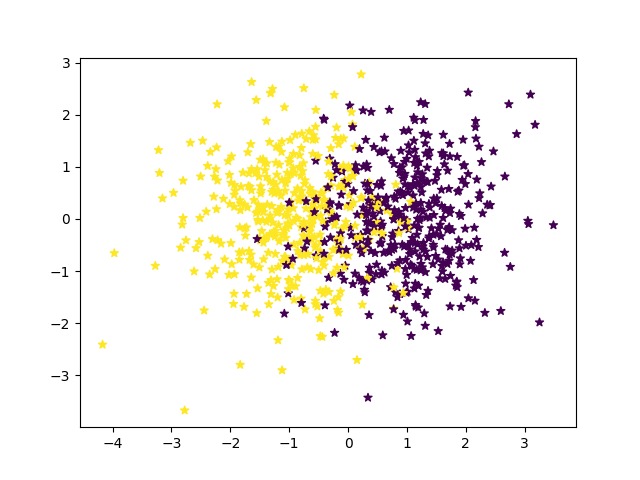
import matplotlib.pyplot as plt

#visuailzation

plt.scatter(value1[:, 0], value1[:,1], c=y, marker="\*")

plt.show()

Output:



Stn Code SO2 NO2 RSPM/PM10 PM 2.5

count 2879.000000 2868.000000 2866.000000 2875.000000 0.0

mean 475.750261 11.503138 22.136776 62.494261 NaN

std 277.675577 5.051702 7.128694 31.368745 NaN

min 38.000000 2.000000 5.000000 12.000000 NaN

25% 238.000000 8.000000 17.000000 41.000000 NaN

50% 366.000000 12.000000 22.000000 55.000000 NaN

75% 764.000000 15.000000 25.000000 78.000000 NaN

max 773.000000 49.000000 71.000000 269.000000 NaN

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2879 entries, 0 to 2878

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Stn Code 2879 non-null int64

1 Sampling Date 2879 non-null object

2 State 2879 non-null object

3 City/Town/Village/Area 2879 non-null object

4 Location of Monitoring Station 2879 non-null object

5 Agency 2879 non-null object

6 Type of Location 2879 non-null object

7 SO2 2868 non-null float64

8 NO2 2866 non-null float64

9 RSPM/PM10 2875 non-null float64

10 PM 2.5 0 non-null float64

dtypes: float64(4), int64(1), object(6)

memory usage: 247.5+ KB

None

Actual Value: 1

Predicted Value: 1

Accuracy: 0.8977272727272727