CS 513 - B

KNOWLEDGE DISCOVERY AND DATA MINING

NEXT-DAY RAIN PREDICTION IN AUSTRALIA USING SEVERAL CLASSIFICATION MODELS

TEAM MEMBERS

ROUSHAN KUMAR

ID : 20009314

INTRODUCTION AND PROBLEM OVERVIEW

- Rain plays a vital role in our lives. Clouds are responsible for bringing rain to humans. In order to forecast when it will rain, the weather department tries to do some forecasting.
- Agriculture is a major industry in Australia, and rainfall plays a crucial role in determining crop yields. Accurate rainfall predictions can help farmers plan their planting and harvesting activities, and make decisions regarding irrigation and fertilization.
- Australia is the driest inhabited continent on earth; 70% of it is either arid or semi arid land, and water is a scarce resource. Predicting rainfall is essential for effective water management, including planning for water storage, allocation, and distribution.
- Heavy rainfall can cause flooding in many parts of Australia. Accurate predictions of rainfall can help emergency services and authorities prepare for and respond to potential floods.

OBJECTIVE AND GOAL

Build	Build a predictive model that predicts whether or not it will rain tomorrow in Australia based on their ten years of daily weather observations.
Predict	Predict next-day rain using different classification models/ machine learning algorithms /data mining techniques from 10 years of daily weather observations of many locations across Australia.
Classify and predict	Classify and predict possible causes of rain that falls the next day.
Compare	Compare different ML algorithms used for prediction.

ABOUT THE DATASET

- This dataset contains about 10 years of daily weather observations from many locations across Australia.

Dataset statistics

Number of variables	23
Number of observations	145460
Missing cells	343248
Missing cells (%)	10.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	25.5 MiB
Average record size in memory	184.0 B

Variable types

Categorical	5
Numeric	16
Boolean	2

Data Source References:
https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

DATA FIELDS

Column Name	Types	Description
Date	Categorical	The date of observation
Location	Categorical	The common name of the location of the weather station
MinTemp	Numeric	The minimum temperature in degrees celsius
MaxTemp	Numeric	The maximum temperature in degrees celsius
Rainfall	Numeric	The amount of rainfall recorded for the day in mm
Evaporation	Numeric	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	Numeric	The number of hours of bright sunshine in the day
WindGusDir	Categorical	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	Categorical	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Categorical	Direction of the wind at 9am
WindDir3pm	Categorical	Direction of the wind at 3pm
WindSpeed9am	Numeric	Wind speed (km/hr) averaged over 10 minutes prior to 9am

Column Name	Type	Description
WindSpeed3pm	Numeric	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Numeric	Humidity (percent) at 9am
Humidity3pm	Numeric	Humidity (percent) at 3pm
Pressure9am	Numeric	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Cloud9am	Numeric	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many
Cloud3pm	Numeric	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
Temp9am	Numeric	Temperature (degrees C) at 9am
Temp3pm	Numeric	Temperature (degrees C) at 3pm
RainToday	Boolean	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RainTomorrow	Boolean	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

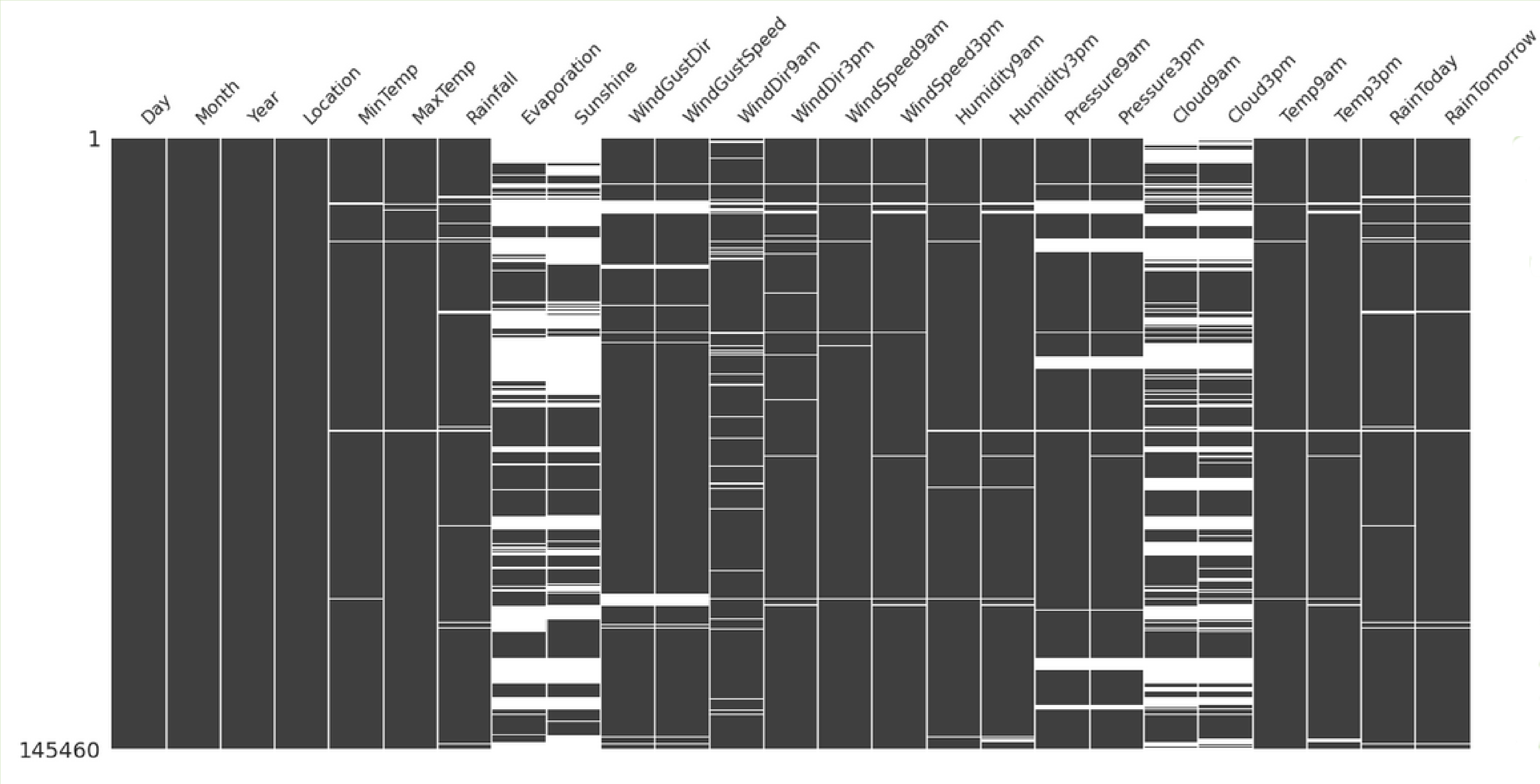
RainTomorrow: It is the target variable to predict next-day rain

ORIGINAL DATASET

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0
5	2008-12-06	Albury	14.6	29.7	0.2	NaN	NaN	WNW	56.0	W	W	19.0
6	2008-12-07	Albury	14.3	25.0	0.0	NaN	NaN	W	50.0	SW	W	20.0
7	2008-12-08	Albury	7.7	26.7	0.0	NaN	NaN	W	35.0	SSE	W	6.0
8	2008-12-09	Albury	9.7	31.9	0.0	NaN	NaN	NNW	80.0	SE	NW	7.0
9	2008-12-10	Albury	13.1	30.1	1.4	NaN	NaN	W	28.0	S	SSE	15.0

WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
24.0	71.0	22.0	1007.7	1007.1	8.0	NaN	16.9	21.8	No	No
22.0	44.0	25.0	1010.6	1007.8	NaN	NaN	17.2	24.3	No	No
26.0	38.0	30.0	1007.6	1008.7	NaN	2.0	21.0	23.2	No	No
9.0	45.0	16.0	1017.6	1012.8	NaN	NaN	18.1	26.5	No	No
20.0	82.0	33.0	1010.8	1006.0	7.0	8.0	17.8	29.7	No	No
24.0	55.0	23.0	1009.2	1005.4	NaN	NaN	20.6	28.9	No	No
24.0	49.0	19.0	1009.6	1008.2	1.0	NaN	18.1	24.6	No	No
17.0	48.0	19.0	1013.4	1010.1	NaN	NaN	16.3	25.5	No	No
28.0	42.0	9.0	1008.9	1003.6	NaN	NaN	18.3	30.2	No	Yes
11.0	58.0	27.0	1007.0	1005.7	NaN	NaN	20.1	28.2	Yes	No

MISSING VALUES

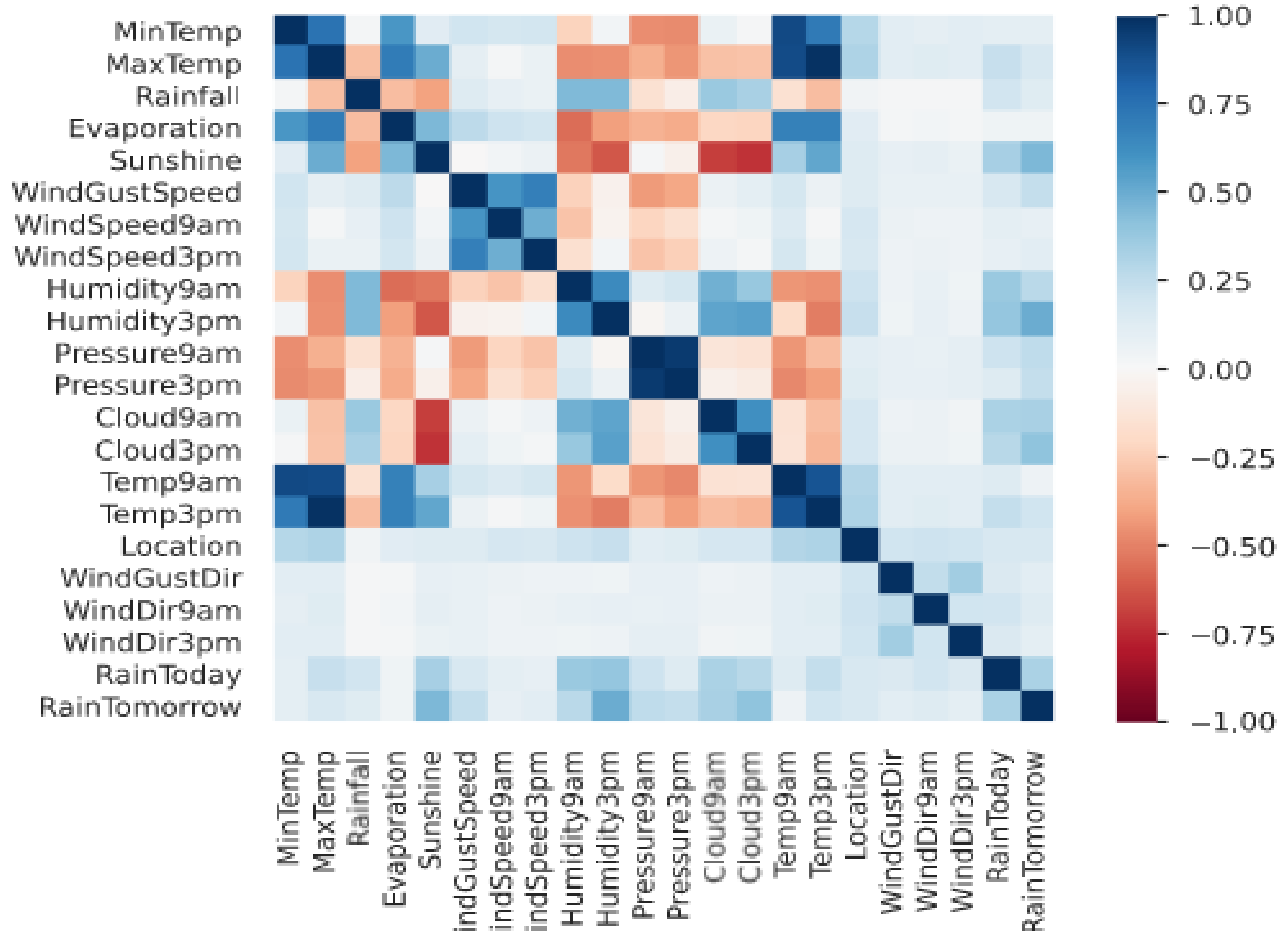


SKEWNESS

Day	0.009040
Month	0.030343
Year	-0.049357
MinTemp	0.022230
MaxTemp	0.224055
Rainfall	9.940909
Evaporation	5.177252
Sunshine	-1.070901
WindGustSpeed	0.923588
WindSpeed9am	0.786472
WindSpeed3pm	0.632461
Humidity9am	-0.491644
Humidity3pm	0.032054
Pressure9am	-0.098584
Pressure3pm	-0.045578
Cloud9am	-0.560375
Cloud3pm	-0.568990
Temp9am	0.090721
Temp3pm	0.247228

- Rainfall and Evaporation are highly positively skewed
- Sunshine is highly negatively skewed

CORRELATIONS HEATMAP



PREPROCESSING

- The feature **MaxTemp** has been removed from the training and testing datasets due to its high correlation with **Temp3pm**, **Temp9am**, and **MinTemp**.

- New features for **Year, Month,** and **Day** have been included by extracting their values from the original Date column. As a result, the Date feature has been removed from both the training and testing datasets.

- Reciprocal transformation was performed on **Rainfall** and square root transformation was performed on **Evaporation** because they were highly positively skewed. Additionally, the square of **Sunshine** was taken because it was highly negatively skewed.

- The values of No and Yes in the **RainToday** and **RainTomorrow** columns were substituted with 0 and 1 respectively.

PREPROCESSING

- The missing values in categorical features (**WindGustDir, WindDir9am, WindDir3pm , RainToday and RainTomorrow**) were filled by replacing them with the mode value of the corresponding feature.

- The missing values in numerical features were filled by replacing them with the median value of the corresponding feature

- Target encoding was utilized for the purpose of converting categorical variables of categorical features into numerical variables.

- **MinMaxScaler** was used to normalize features in the dataset.

- To address the bais in the training dataset, we utilized the **SMOTETomek** technique.

MODEL BUILDING

- The features **['Day', 'Month', 'Year', 'Location', 'MinTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', 'Temp3pm', 'RainToday']**
  were used for training the model.

- **['RainTomorrow']** is the feature that we aim to predict.

- The normalized and imputed data was divided into training and testing datasets in an 80:20 ratio.

- To validate the model against the validation dataset, we employed k-fold cross validation.

CLASSIFICATION ALGORITHMS/MODELS

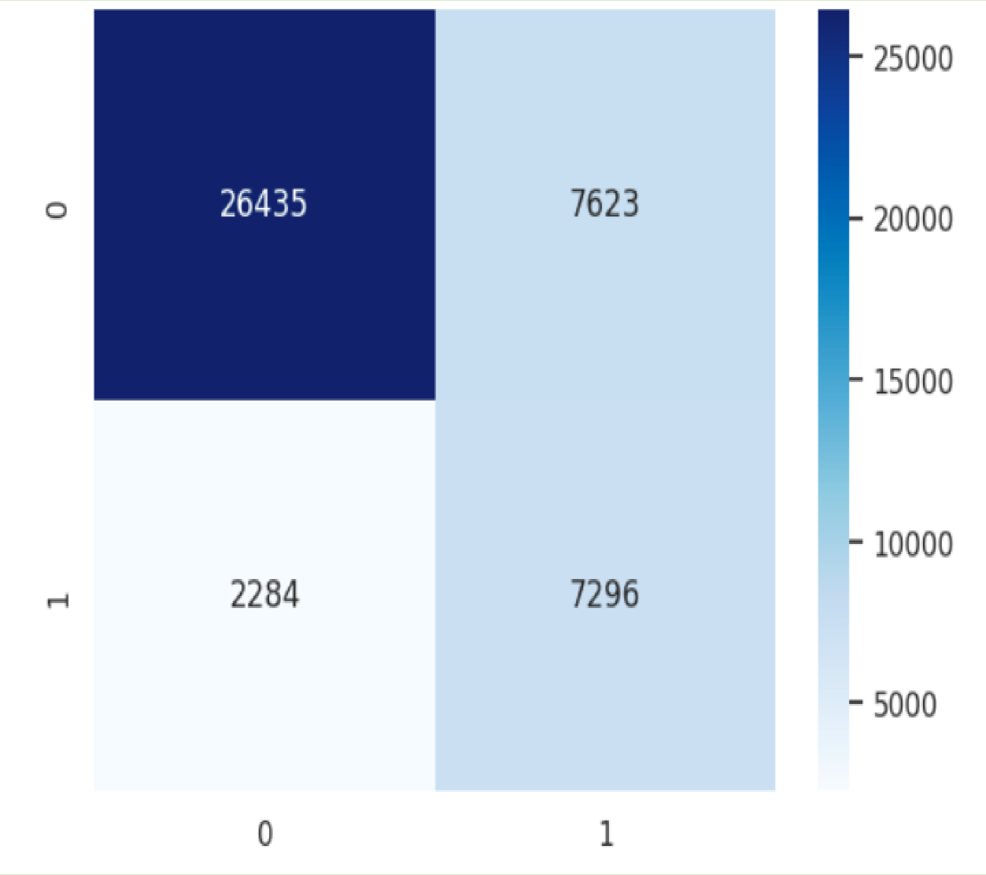
- LOGISTIC REGRESSION
- NAIVE BAYES
- DECISION TREE
- RANDOM FOREST
- KNEIGHBORS
- EXTREME GRADIENT BOOST
- SUPPORT VECTOR CLASSIFIER

LOGISTIC REGRESSION

```
Accuracy of Logistic Regression: 77.29730968422017

              precision    recall  f1-score   support

           0       0.92      0.78      0.84     34058
           1       0.49      0.76      0.60      9580

    accuracy                           0.77     43638
   macro avg       0.70      0.77      0.72     43638
weighted avg       0.83      0.77      0.79     43638
```

```
[ ] score1=cross_val_score(lr,X_valid, y_valid,cv=10)
    print(f"After k-fold cross validation score is {score1.mean()}")
```

```
After k-fold cross validation score is 0.8391998700429177
```



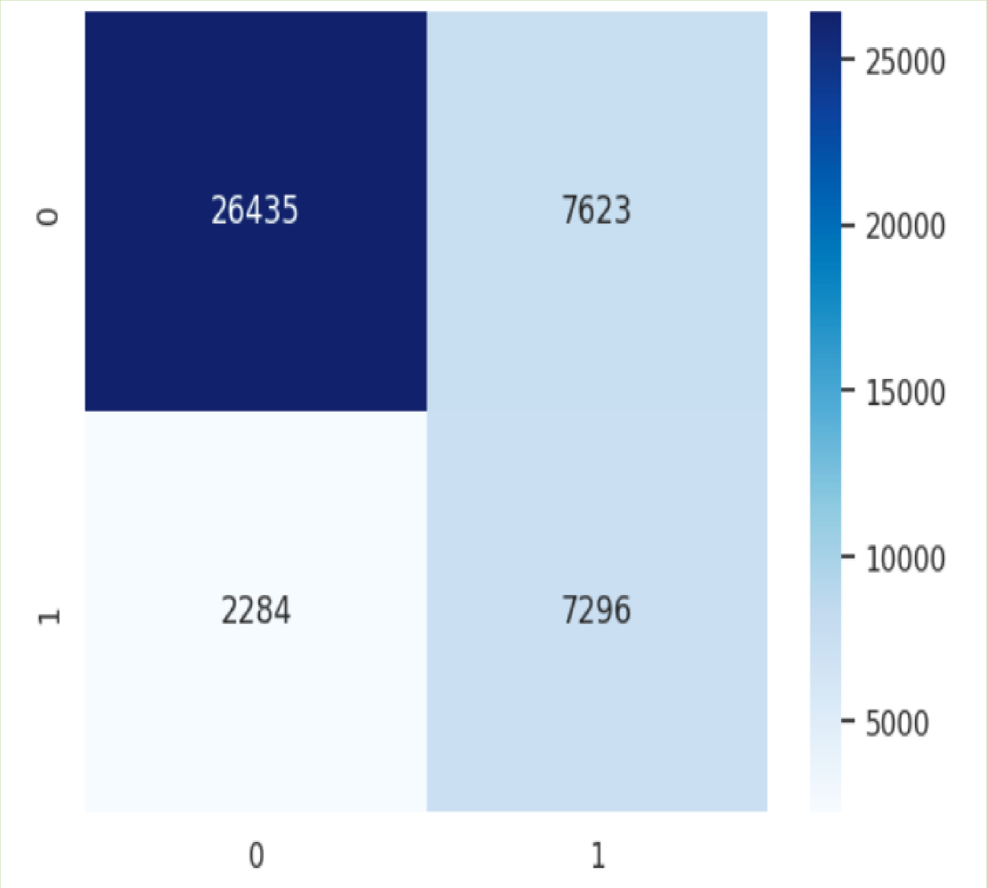
Confusion Matrix

NAIVE BAYES

```
Accuracy of Naive Bayes model: 73.87368807003071

              precision    recall  f1-score   support

           0       0.92      0.73      0.81     34058
           1       0.45      0.77      0.56      9580

    accuracy                           0.74     43638
   macro avg       0.68      0.75      0.69     43638
weighted avg       0.82      0.74      0.76     43638
```

```
[ ] score2=cross_val_score(nb,X_valid,y_valid,cv=10)
    print(f"After k-fold cross validation score is {score2.mean()}")
```

```
After k-fold cross validation score is 0.7823454165128687
```



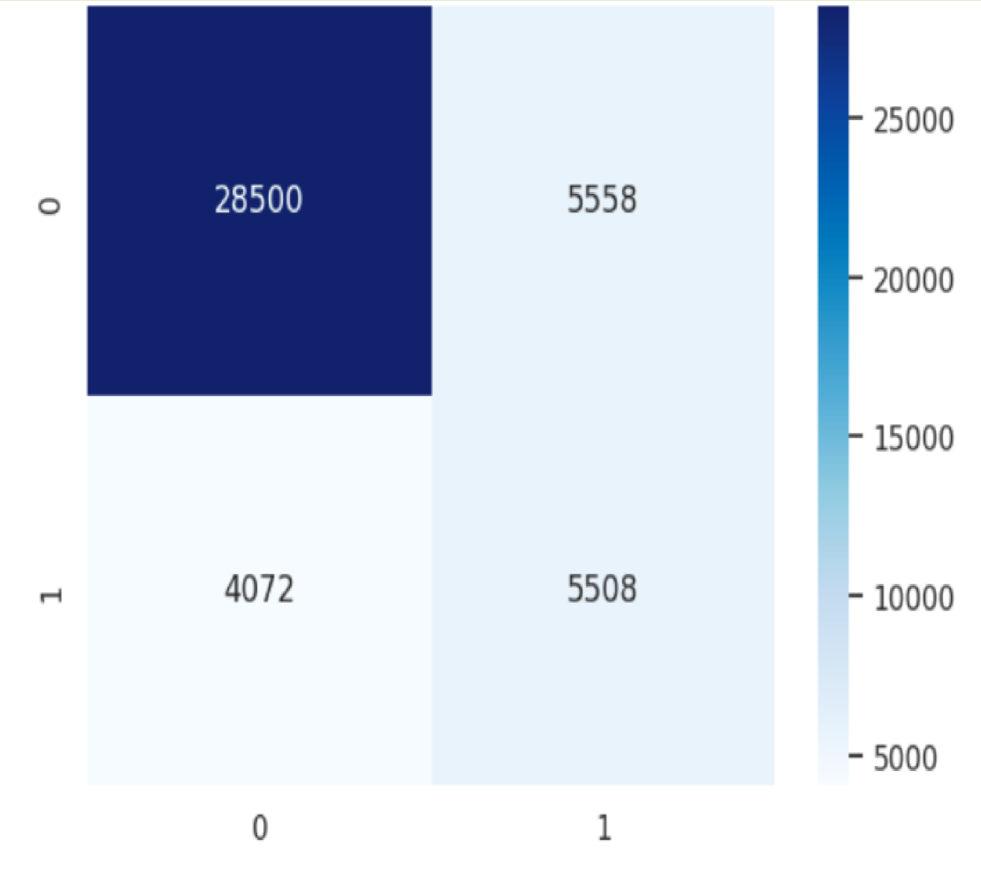
Confusion Matrix

DECISION TREE

```
Accuracy of DecisionTreeClassifier: 77.93207754709198

              precision    recall  f1-score   support

           0       0.87      0.84      0.86     34058
           1       0.50      0.57      0.53      9580

    accuracy                           0.78     43638
   macro avg       0.69      0.71      0.69     43638
weighted avg       0.79      0.78      0.78     43638
```

```
[ ] score6=cross_val_score(dt,X_valid,y_valid,cv=10)
    print(f"After k-fold cross validation score is {score6.mean()}")
```

```
After k-fold cross validation score is 0.7829871242489286
```



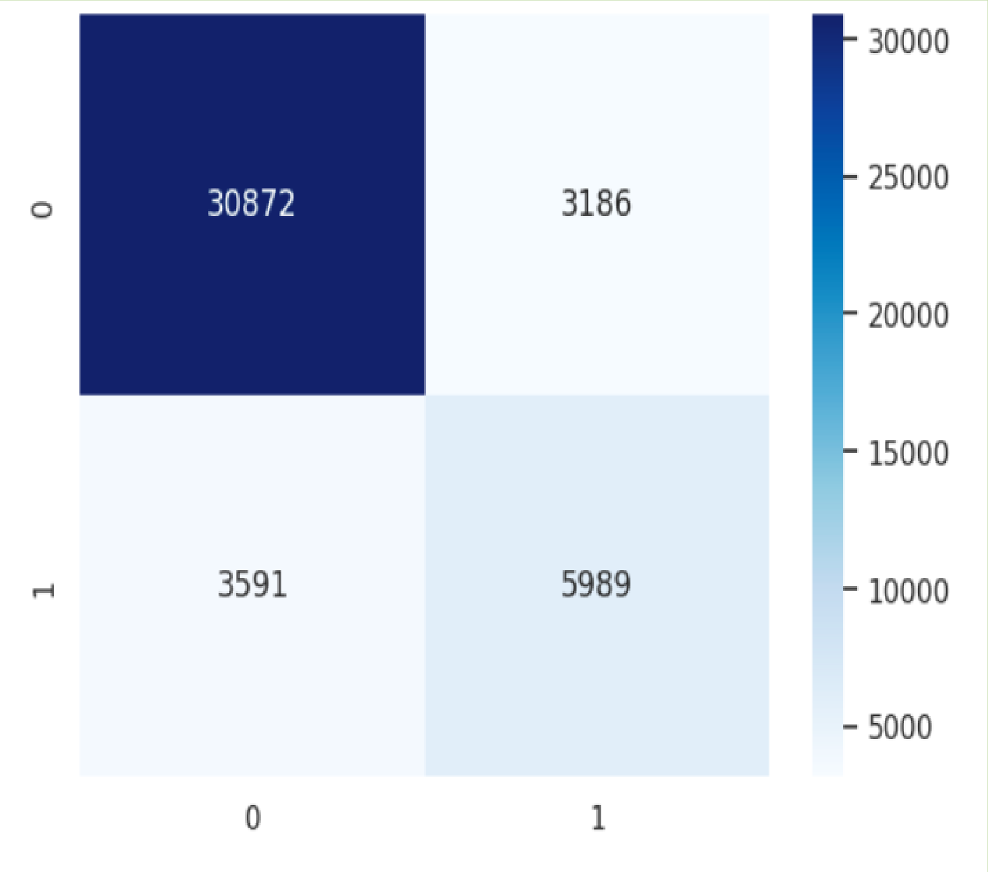
Confusion Matrix

RANDOM FOREST

```
Accuracy of Random Forest: 84.46995737659839

              precision    recall  f1-score   support

           0       0.90      0.91      0.90     34058
           1       0.65      0.63      0.64      9580

    accuracy                           0.84     43638
   macro avg       0.77      0.77      0.77     43638
weighted avg       0.84      0.84      0.84     43638
```

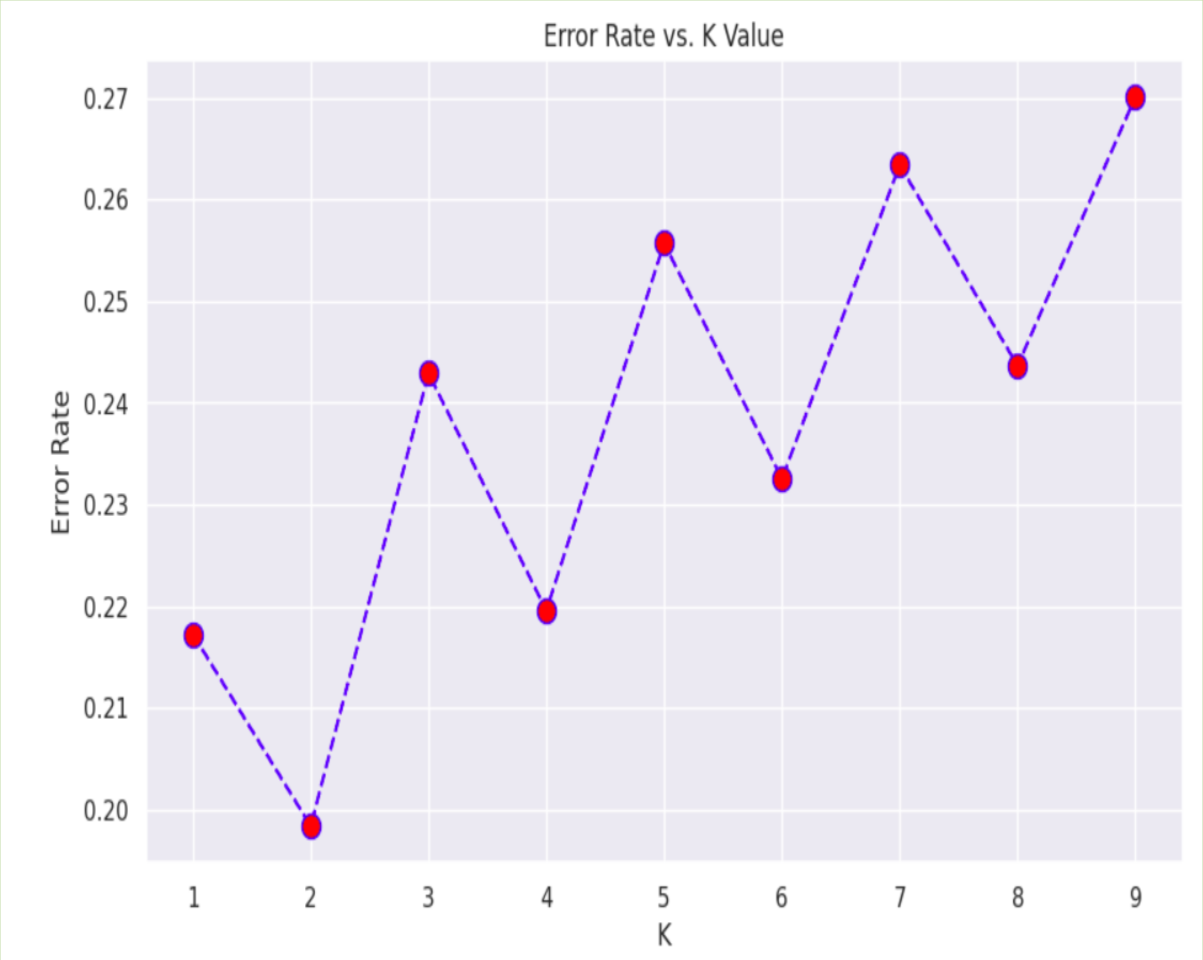
```
[ ] score3=cross_val_score(rf,X_valid,y_valid,cv=10)
    print(f"After k-fold cross validation score is {score3.mean()}")
```

```
After k-fold cross validation score is 0.8513450484481936
```

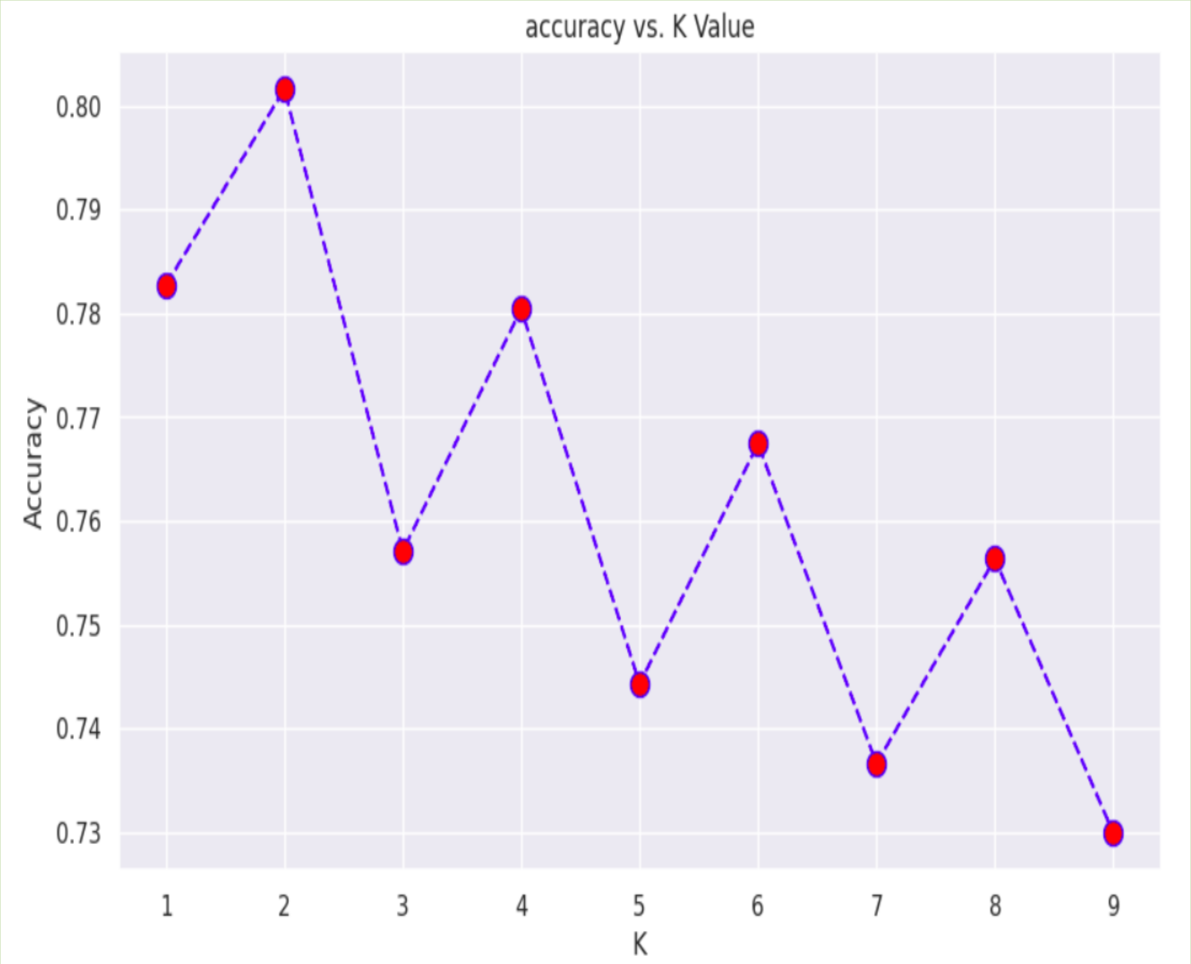


Confusion Matrix

KNEIGHBORS



Error Rate vs. K Value
Error Rate
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.20
1
2
3
4
5
6
7
8
9
K



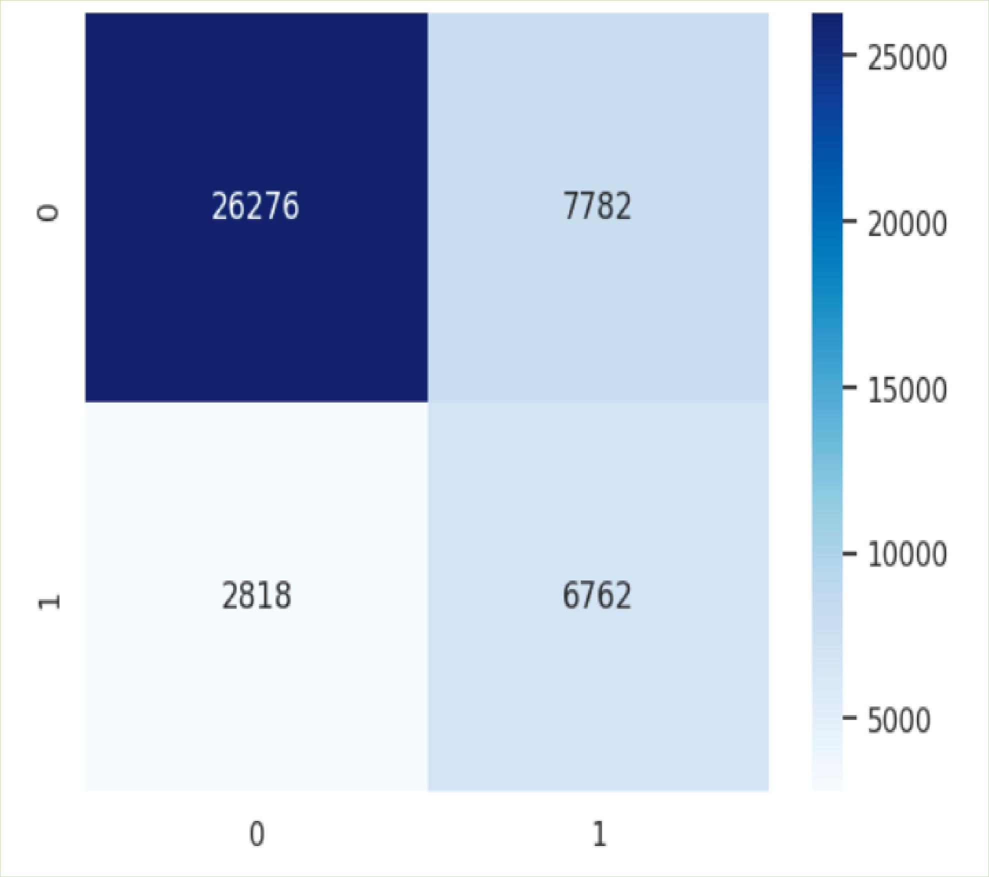
accuracy vs. K Value
Accuracy
0.80
0.79
0.78
0.77
0.76
0.75
0.74
0.73
1
2
3
4
5
6
7
8
9
K

KNEIGHBORS

```
Accuracy of K-NeighborsClassifier: 75.70924423667445

              precision    recall  f1-score   support

           0       0.90      0.77      0.83     34058
           1       0.46      0.71      0.56      9580

    accuracy                           0.76     43638
   macro avg       0.68      0.74      0.70     43638
weighted avg       0.81      0.76      0.77     43638
```

```
[ ] score5=cross_val_score(knn,X_valid,y_valid,cv=10)
    print(f"After k-fold cross validation score is {score5.mean()}")
```

```
After k-fold cross validation score is 0.796805284753278
```



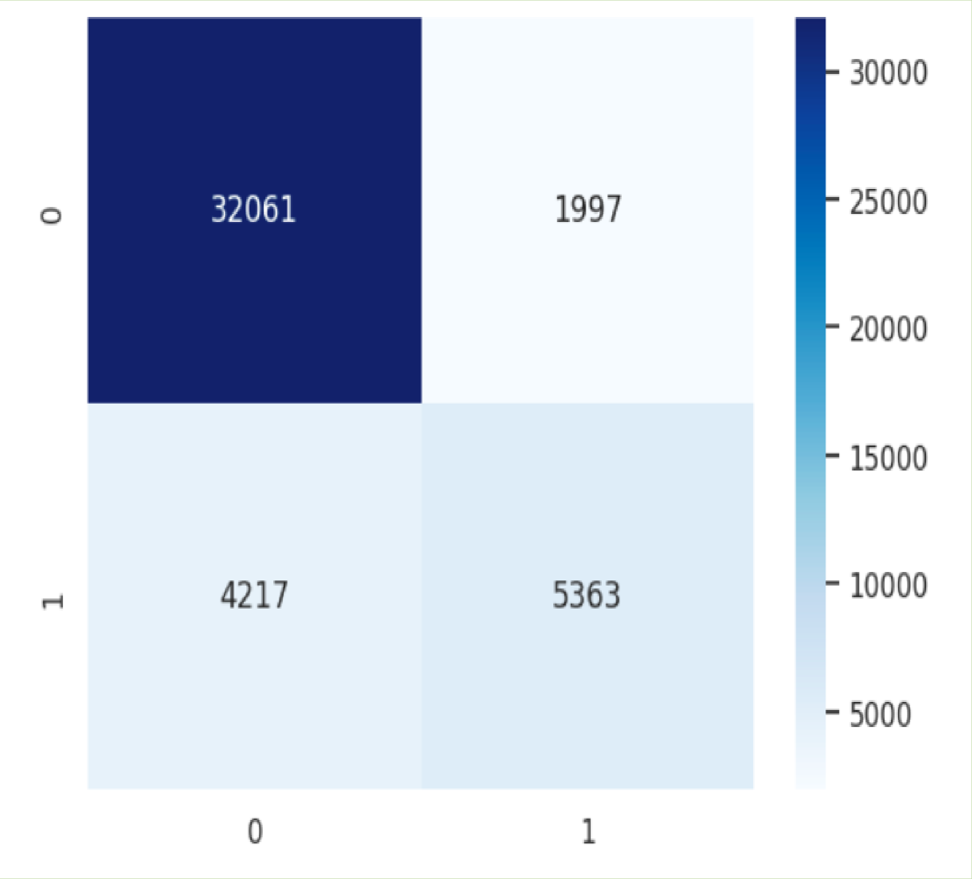
Confusion Matrix

EXTREME GRADIENT BOOST

```
Accuracy of Extreme Gradient Boost: 85.76011732893349

              precision    recall  f1-score   support

           0       0.88      0.94      0.91     34058
           1       0.73      0.56      0.63      9580

    accuracy                           0.86     43638
   macro avg       0.81      0.75      0.77     43638
weighted avg       0.85      0.86      0.85     43638
```

```
[ ] score4=cross_val_score(xgb,X_valid,y_valid,cv=10)
    print(f"After k-fold cross validation score is {score4.mean()}")
```

```
After k-fold cross validation score is 0.8526742356618116
```



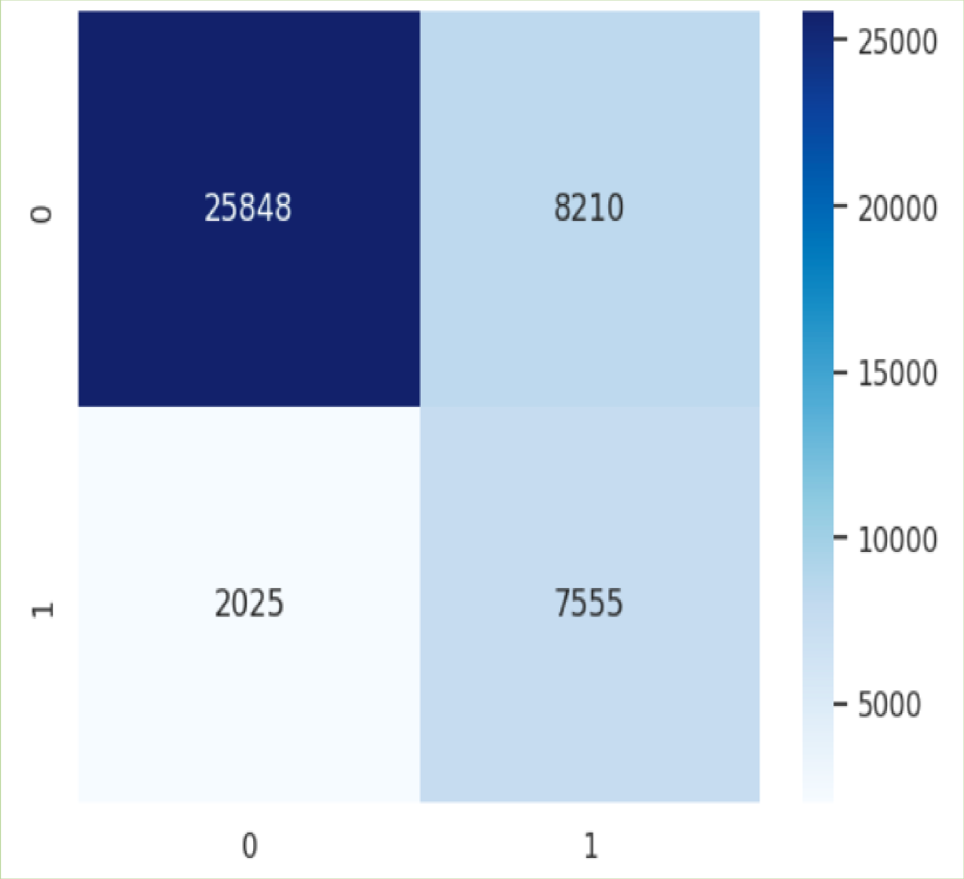
Confusion Matrix

SUPPORT VECTOR CLASSIFIER

```
Accuracy of Support Vector Classifier: 76.5456712039965

              precision    recall  f1-score   support

           0       0.93      0.76      0.83     34058
           1       0.48      0.79      0.60      9580

    accuracy                           0.77     43638
   macro avg       0.70      0.77      0.72     43638
weighted avg       0.83      0.77      0.78     43638
```

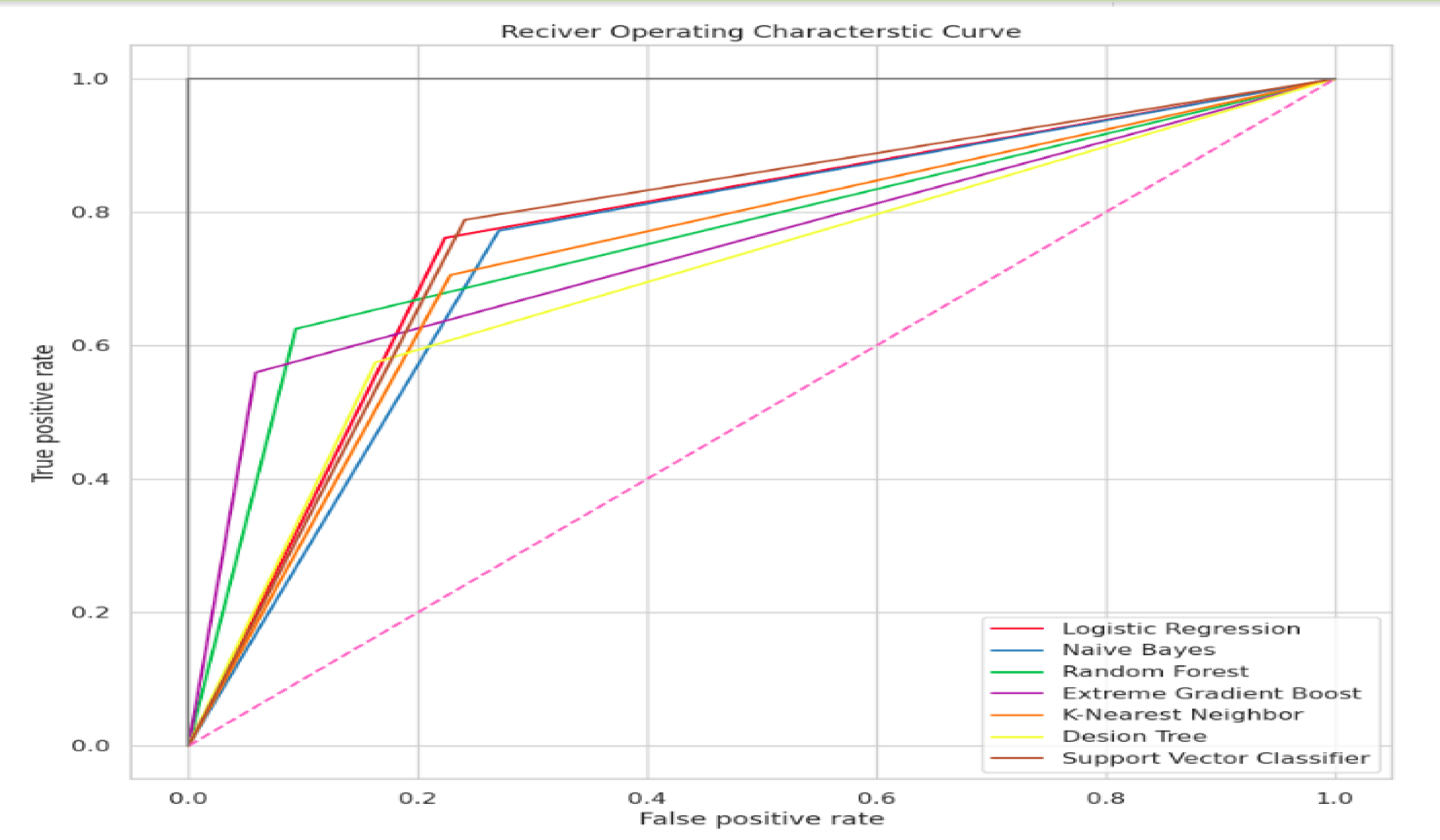
```python
score7=cross_val_score(svc,X_valid,y_valid,cv=10)
print(f"After k-fold cross validation score is {score7.mean()}")
```

```
After k-fold cross validation score is 0.8399560885397224
```



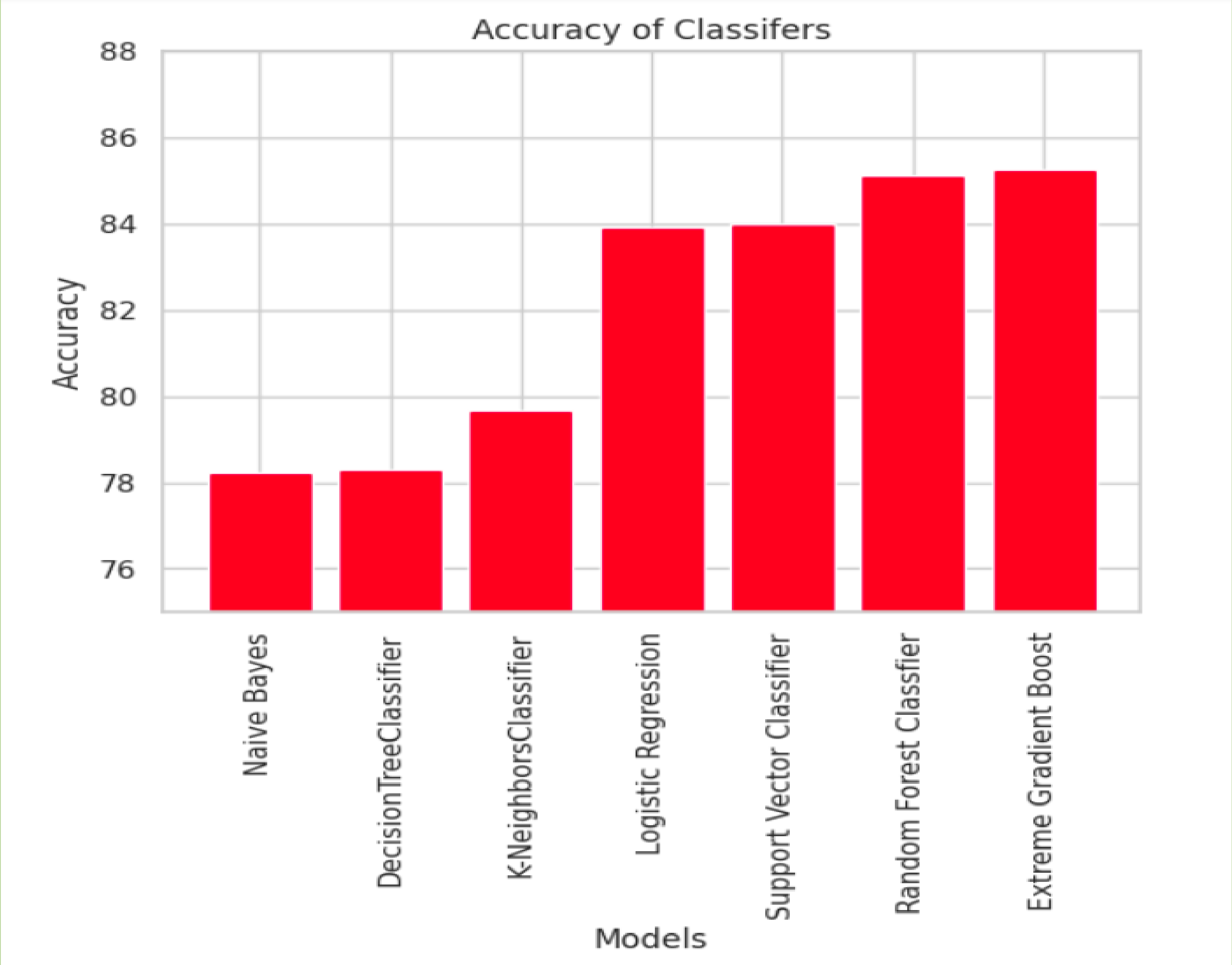
Confusion Matrix

ROC CURVE



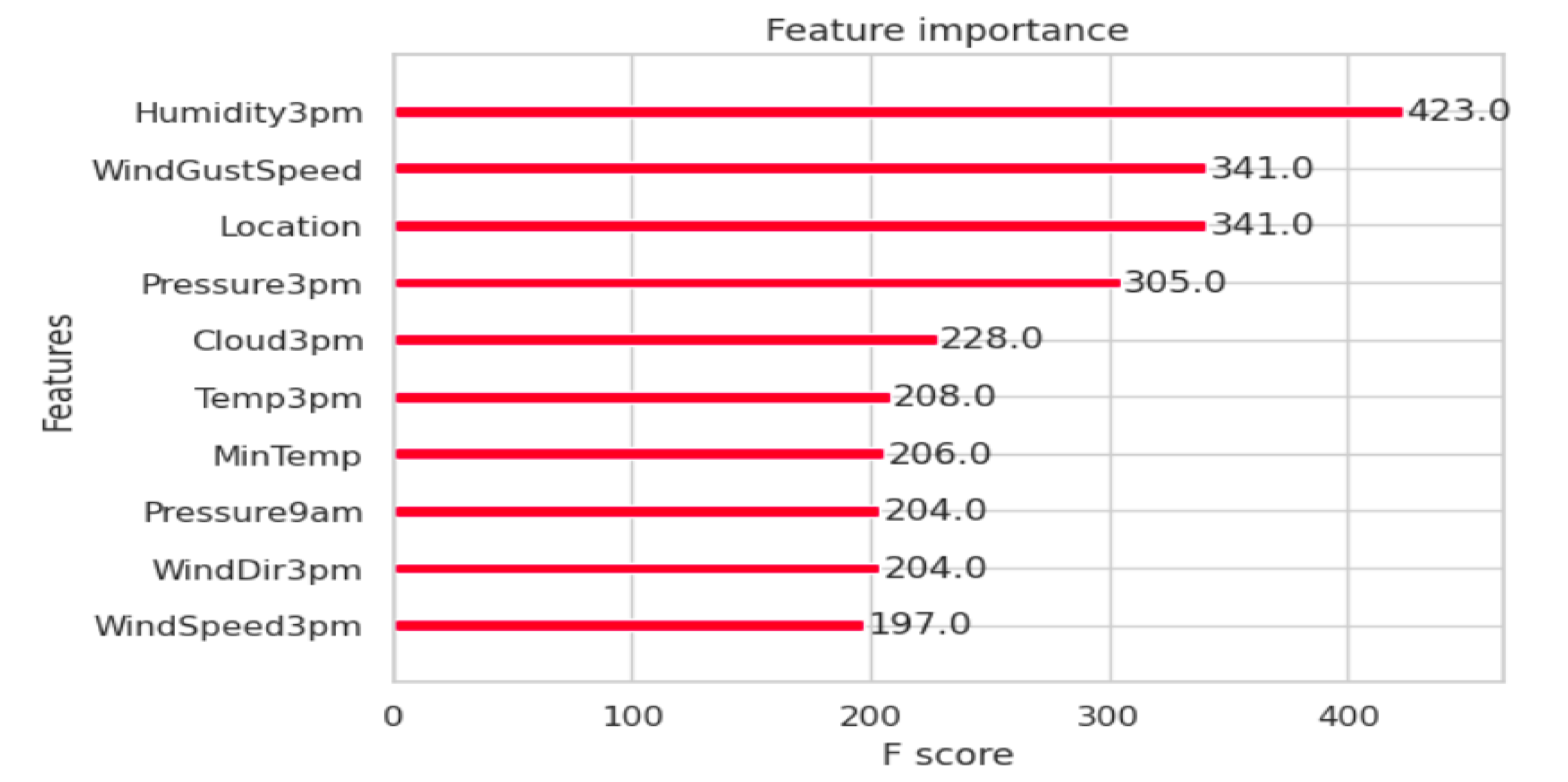
Reciver Operating Characterstic Curve
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Logistic Regression
Naive Bayes
Random Forest
Extreme Gradient Boost
K-Nearest Neighbor
Desion Tree
Support Vector Classifier

ACCURACY COMPARISON

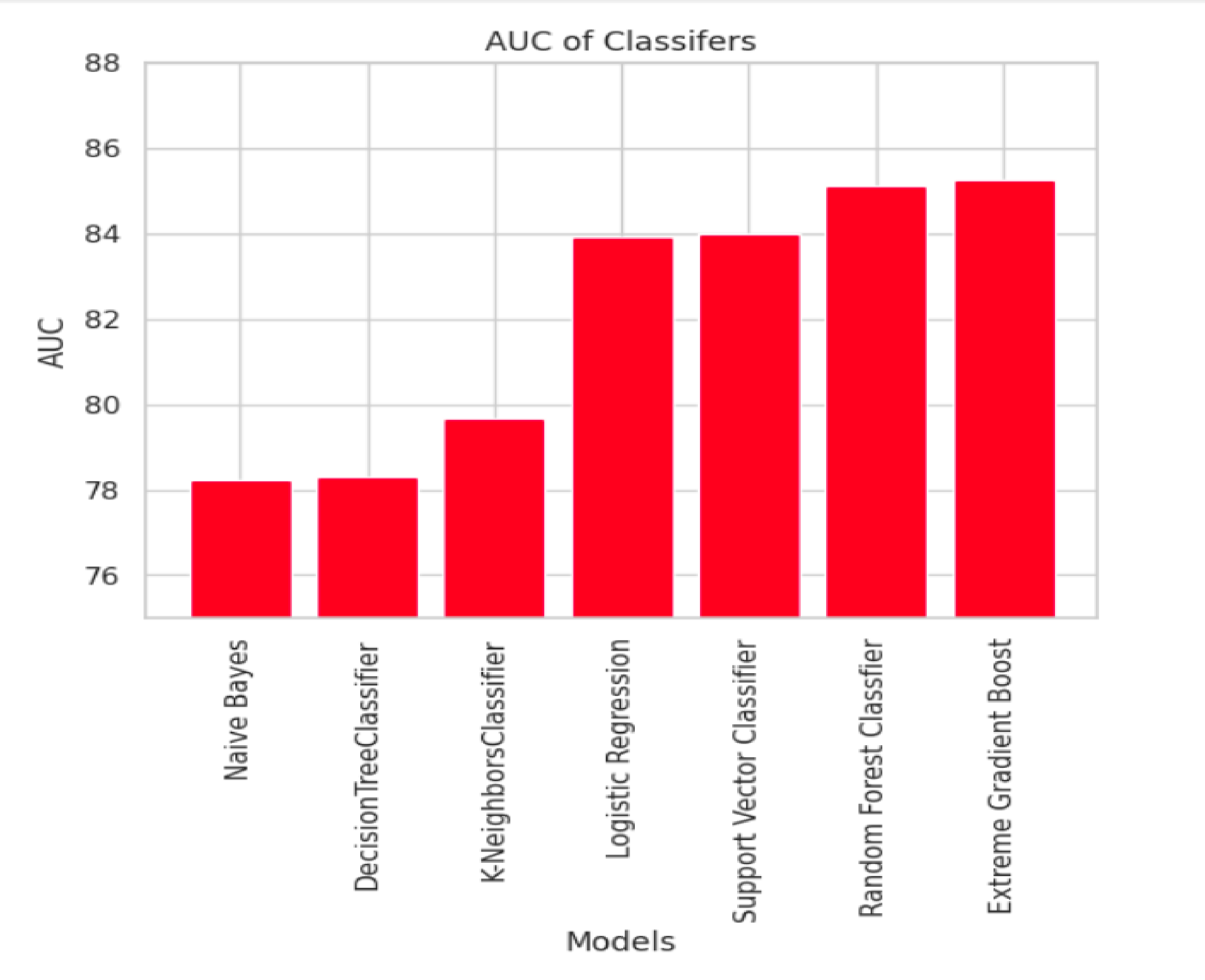


As we can see,
Extreme Gradient Boost
algorithm has highest
accuracy for this dataset.

IMPORTANT FEATURES OF EXTREME GRADIENT BOOST CLASSIFIER



AUC COMPARISON



As we can see, ***Extreme Gradient Boost*** algorithm has most AUC for this dataset.


Thank
You