MACHINE LEARNING FOR
SOCIAL SCIENCES

FRAUD DETECTION

Group 14

Ahmad Nawaz and Muhammad Jawad Raza

ML helps us understand and recognize patterns in fraudlent data, detecting such transactions in real life

WHY ML IS NEEDED

PATTERN RECOGNITION

ADAPTABILITY

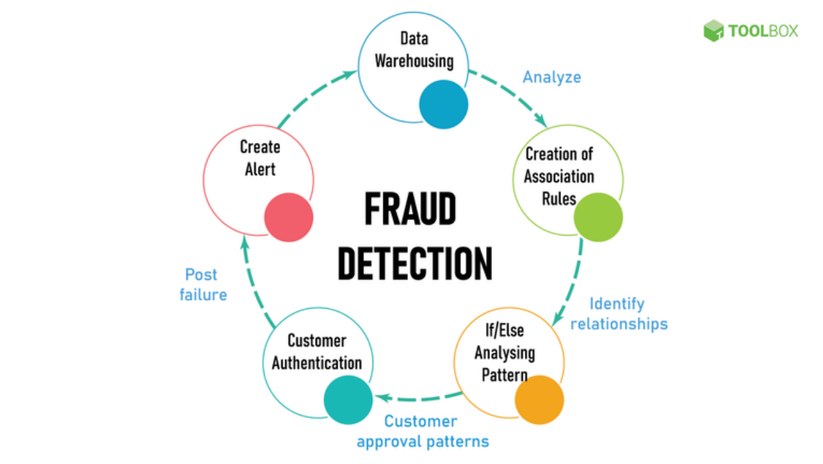
SCALABILITY

UNSUPERVISED LEARNING

REDUCED FALSE POSITIVES

REAL TIME DETECTION

PROBLEM STATEMENT



Our basic goal is to reduce fraud with the help of machine learning, to help firms and people catch such frauds easily

Manual detection of fraud is impractical due to the volume and complexity of transactions. Machine learning can automate the detection process, learning from historical transaction data to identify patterns and anomalies that indicate fraudulent activities.

FRAUD DETECTION

Fraud detection in financial transactions is crucial for maintaining the integrity of banking systems and protecting customers' assets. Rapid detection and prevention of fraudulent activities can save substantial financial losses.

DATASET

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.00	0.00	0.0	0.0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.00	0.00	0.0	0.0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.00	0.00	1.0	0.0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.00	0.00	1.0	0.0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.00	0.00	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
989559	45	CASH_OUT	166662.12	C1059301307	24944.0	0.00	C1429108522	50324.42	216986.54	0.0	0.0
989560	45	PAYMENT	1698.76	C858991144	12230.0	10531.24	M1179748519	0.00	0.00	0.0	0.0
989561	45	PAYMENT	17198.40	C863675783	272.0	0.00	M298238508	0.00	0.00	0.0	0.0
989562	45	CASH_IN	26243.43	C990828630	30955.0	57198.43	C197399765	3503704.84	3477461.41	0.0	0.0
989563	45	CASH_OUT	151185.85	C1909488416	590127.0	43.00	NaN	NaN	NaN	NaN	NaN

989564 rows × 13 columns

FEATURES

step: Simulated time in hours since the start, where each step represents an hour of time.

type: Type of transaction (e.g., PAYMENT, TRANSFER).

amount: Amount of the transaction.

nameOrig: Customer who started the transaction.

oldbalanceOrg: Initial balance before the transaction.

newbalanceOrig: New balance after the transaction.

nameDest: Recipient of the transaction.

oldbalanceDest: Initial recipient balance before the transaction.

newbalanceDest: New recipient balance after the transaction.

isFraud: This is the target variable where '1' means the transaction is fraudulent.

isFlaggedFraud: Flags illegal attempts to transfer more than 200,000 in a single transaction.

DATASET EXPLANATION

Target Variable

ISFRAUD

Independent Variable

step
type
amount
nameOrig
oldbalanceOrg
newbalanceOrig
nameDest
oldbalanceDest
newbalanceDest
isFlaggedFraud

PREPROCESSING

RENAMING VARIABLES

we renamed all variables in a lowercase standard way

MISSING VALUES

3 of the variables had missing values 1 each, these missing values were removed and hence the data set now has complete values

EDA

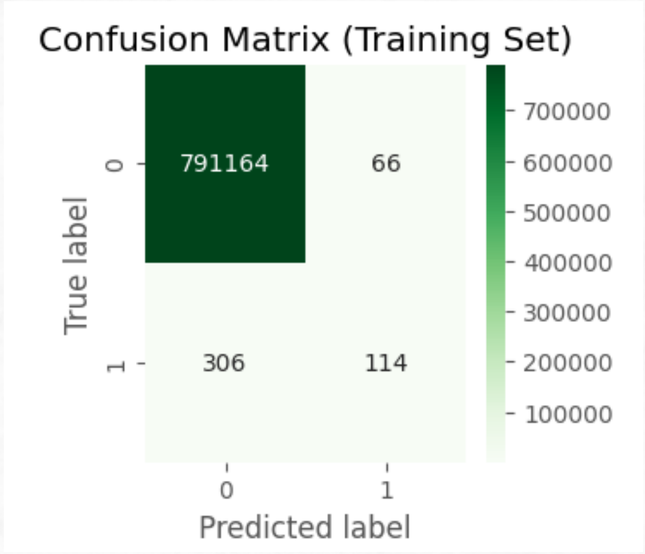
Checked for missing values and imputed/dropped where necessary.
Removed irrelevant features like nameOrig, nameDest and isFlaggedfraud as they are specific identifiers not useful for pattern recognition.

ENCODING

Encoded the type variable

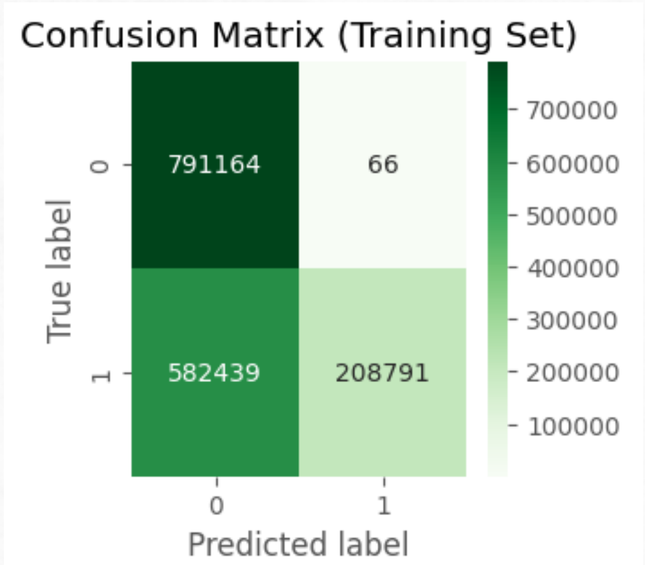
LOGISTIC REGRESSION

Training Set



Before Smote

```
Classification Report (Training Set):
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00    791230
         1.0       0.63      0.27      0.38       420

    accuracy                           1.00    791650
   macro avg       0.82      0.64      0.69    791650
weighted avg       1.00      1.00      1.00    791650

ROC AUC Score (Training Set): 0.8979249336325661
```

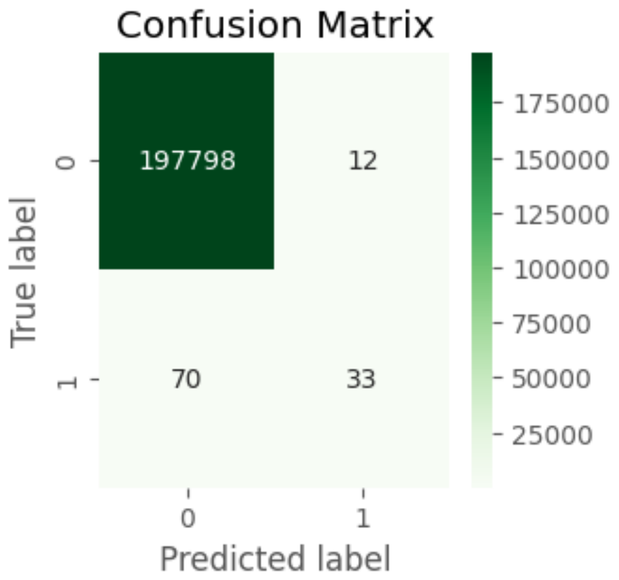


After Smote

```
Classification Report (Training Set):
              precision    recall  f1-score   support

         0.0       0.58      1.00      0.73    791230
         1.0       1.00      0.26      0.42    791230

    accuracy                           0.63   1582460
   macro avg       0.79      0.63      0.57   1582460
weighted avg       0.79      0.63      0.57   1582460

ROC AUC Score (Training Set): 0.9055592154033778
```

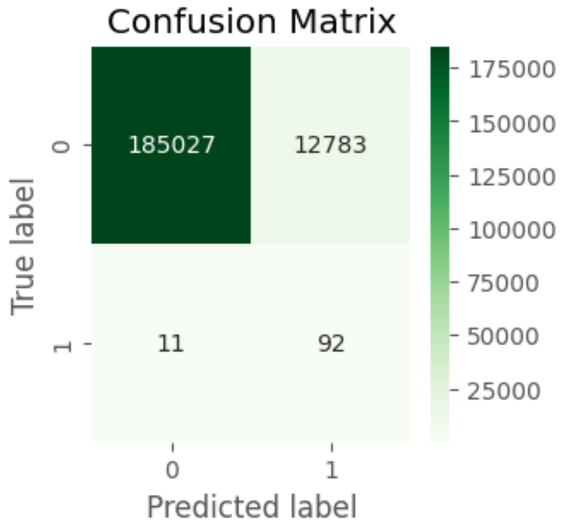
LOGISTIC REGRESSION

Test Set



Before Smote

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

         0.0       1.00      1.00      1.00    197810
         1.0       0.73      0.32      0.45       103

    accuracy                           1.00    197913
   macro avg       0.87      0.66      0.72    197913
weighted avg       1.00      1.00      1.00    197913

ROC AUC Score: 0.885986454590386
```

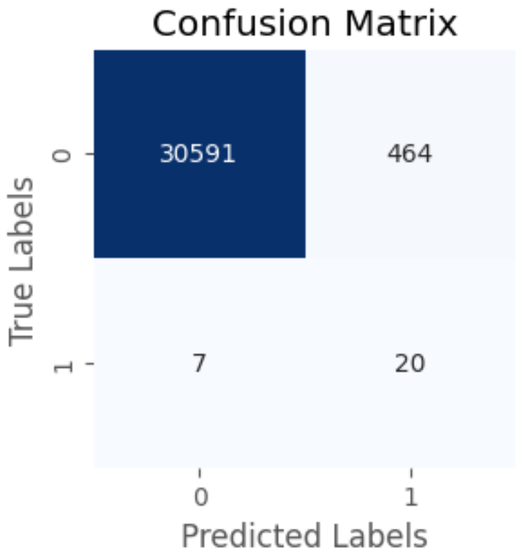


After Smote

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

         0.0       1.00      0.94      0.97    197810
         1.0       0.01      0.89      0.01       103

    accuracy                           0.94    197913
   macro avg       0.50      0.91      0.49    197913
weighted avg       1.00      0.94      0.97    197913

ROC AUC Score: 0.9333302575826662
```

KNN

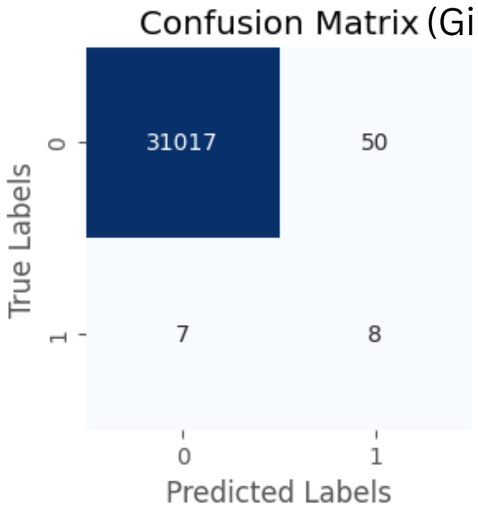
Test set



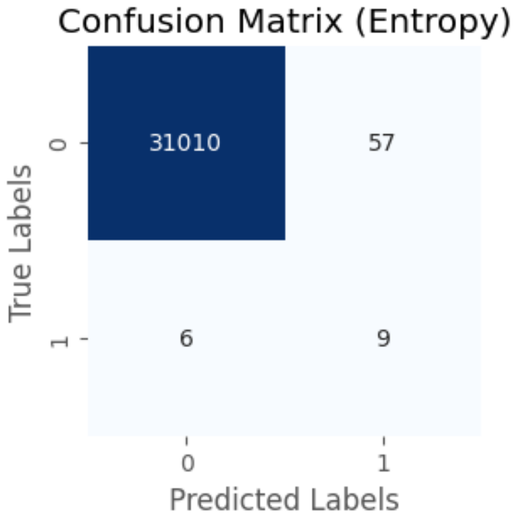
After Smote

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

         0.0       1.00      0.99      0.99     31055
         1.0       0.04      0.74      0.08        27

    accuracy                           0.98     31082
   macro avg       0.52      0.86      0.54     31082
weighted avg       1.00      0.98      0.99     31082

ROC AUC Score: 0.8641734795494254
```

DECISION TREE



After Smote



Test set

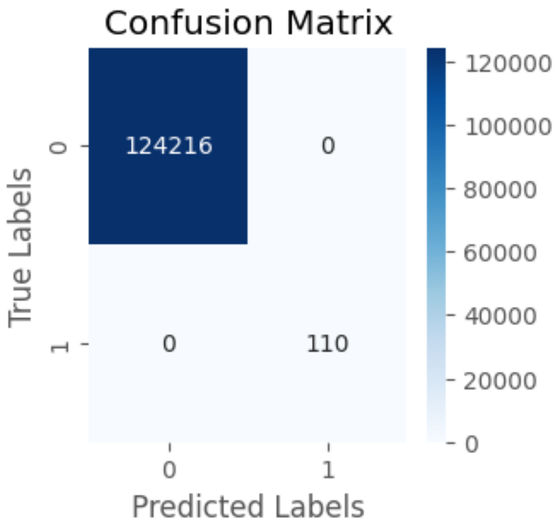
Classification Report: (Gini)

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	31067
1.0	0.14	0.53	0.22	15
accuracy			1.00	31082
macro avg	0.57	0.77	0.61	31082
weighted avg	1.00	1.00	1.00	31082

Classification Report (Entropy):

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	31067
1.0	0.14	0.60	0.22	15
accuracy			1.00	31082
macro avg	0.57	0.80	0.61	31082
weighted avg	1.00	1.00	1.00	31082

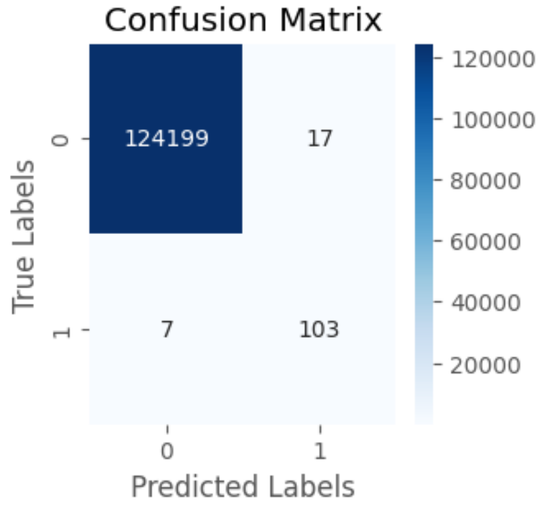
RANDOM FOREST



Before Smote

Training Set

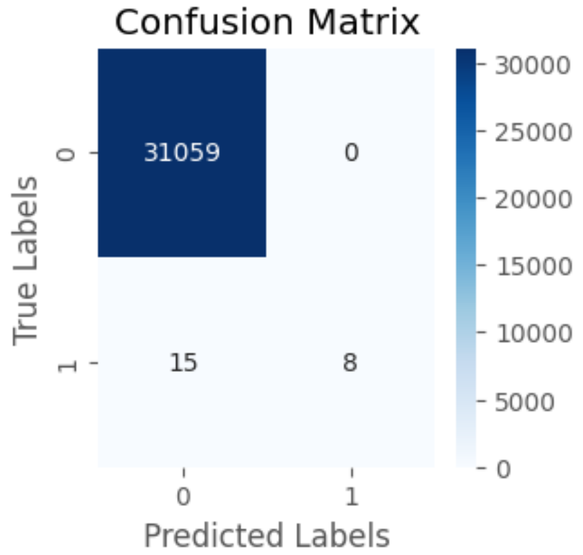
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	126932
1.0	1.00	1.00	1.00	112
accuracy			1.00	127044
macro avg	1.00	1.00	1.00	127044
weighted avg	1.00	1.00	1.00	127044



After Smote

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	124216
1.0	0.86	0.94	0.90	110
accuracy			1.00	124326
macro avg	0.93	0.97	0.95	124326
weighted avg	1.00	1.00	1.00	124326

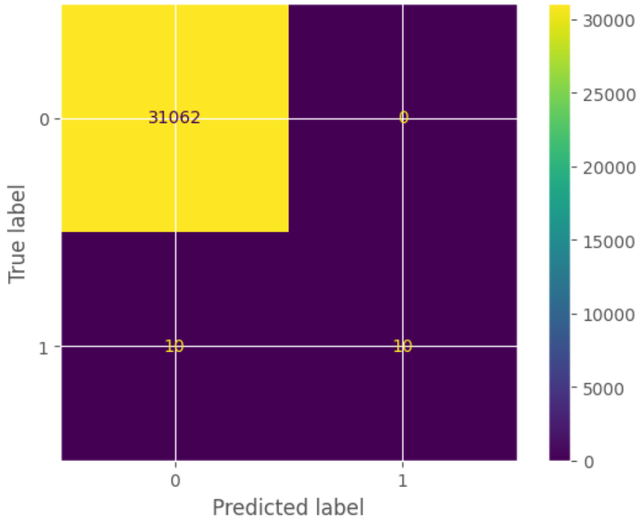
RANDOM FOREST



Before Smote

Test Set

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	31059
1.0	1.00	0.35	0.52	23
accuracy			1.00	31082
macro avg	1.00	0.67	0.76	31082
weighted avg	1.00	1.00	1.00	31082

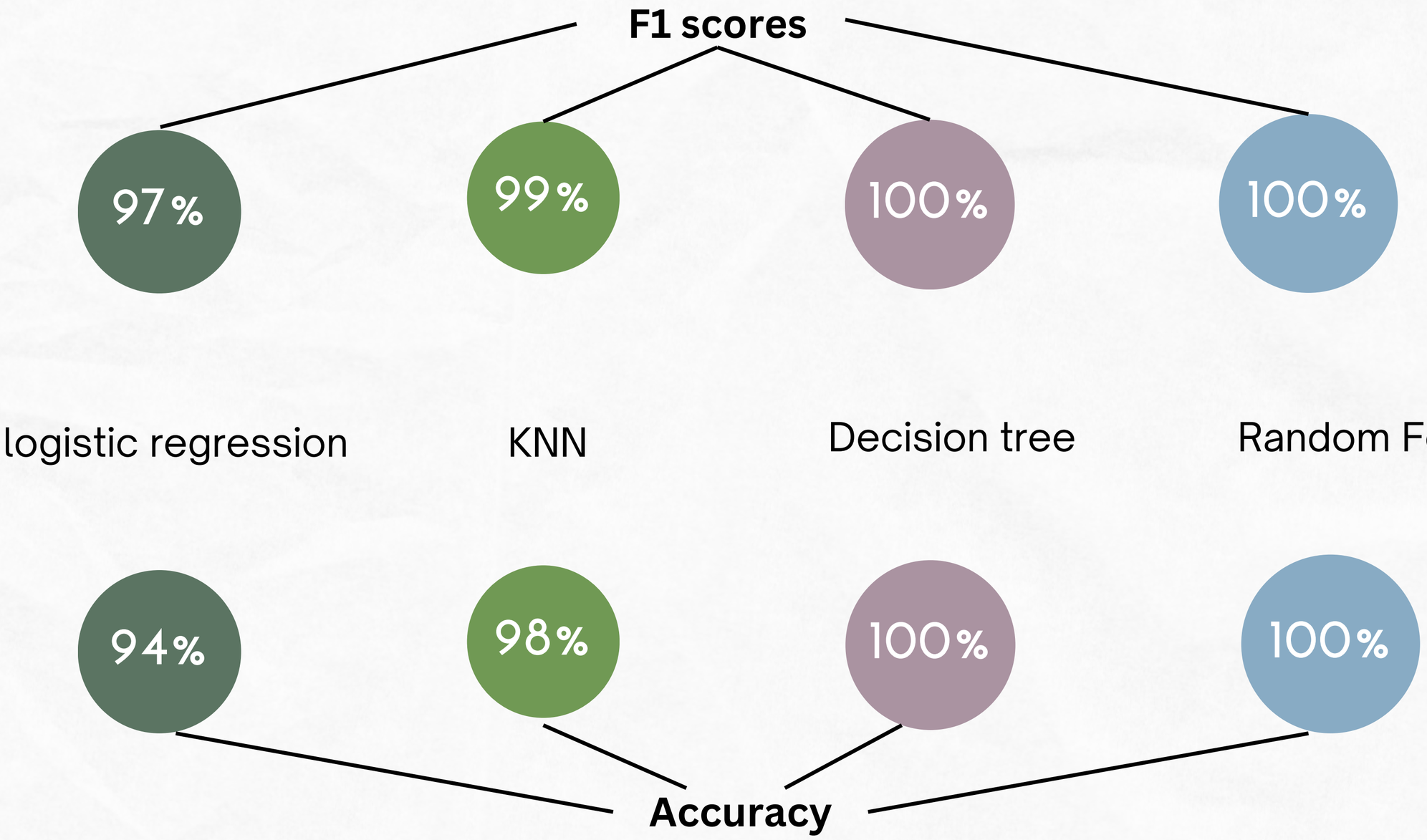


After Smote

Classification Report:

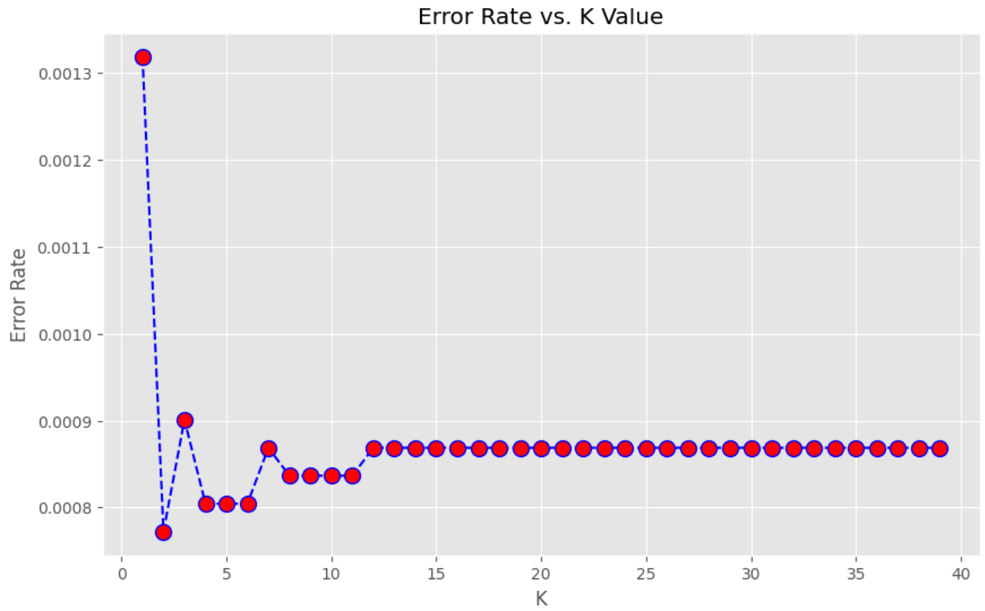
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	31062
1.0	1.00	0.50	0.67	20
accuracy			1.00	31082
macro avg	1.00	0.75	0.83	31082
weighted avg	1.00	1.00	1.00	31082

MACHINE LEARNING MODELS



Hyper parameter tuning

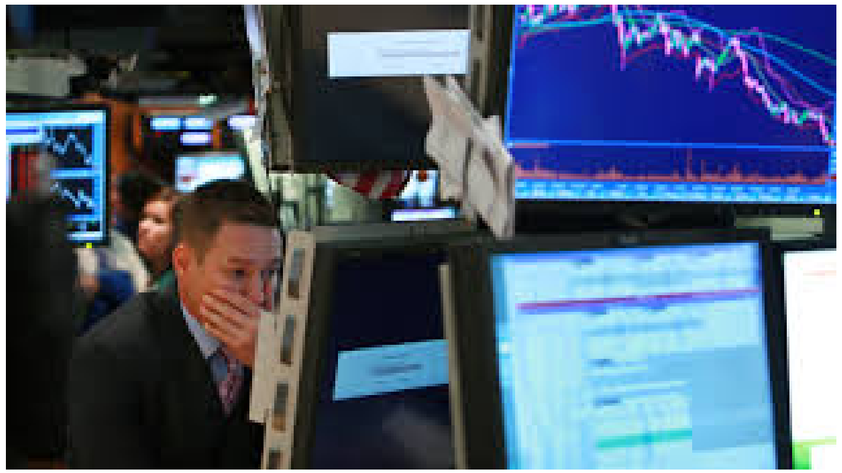
BEST HYPERPARAMETERS: {'MAX_DEPTH': 12, 'N_ESTIMATORS': 120}



Final results

Looking at the F1 scores and accuracy it proves that the model is successful and effective in identifying fraudulent transactions so hence this model can be used to avoid financial fraud.

2008 FINANCIAL CRISIS



from 2003 fraudulent transactions paved the path to 2008's stock market crash, slowly these transactions in various forms added to the ticking time bomb until it exploded and everything came crashing down

ROLE OF FRAUD IN CRASH

Fraudulent activity leading up to the market crash was widespread: mortgage originators commonly deceived borrowers about loan terms and eligibility requirements, in some cases concealing information about the loan like add-ons or balloon payments

Why ML is important

The federal reserve bank was unable to anticipate this massive crash but machine learning could’ve somewhat understood this pattern and give a better insight into this, it may not potentially indicate towards a crisis but would’ve been beneficial in understanding the situation at that time

FUTURE IMPLICATIONS

fraud reduction machine learning would be revolutionary in predicting and catching frauds such as credit card, insurance, healthcare and cyber security threats. Moreover financial frauds, ecommerce and supply chain. Organizations and governments can use these tools to prevent such frauds to minimize financial loss and protect customers.

THANK YOU