**Fraud Detection**

**Steps:**

1. Import the dataset Fraud Detection.
2. Checking the null values.
3. Finding the correlation/dependency between the tables.
4. Preprocessing (LabelEncoding).
5. Splitting.
6. Standard Scaling.
7. Apply the regression algorithm on the dataset (Logistic, Decision Tree, Random Forest KNN).
8. Conclusion.

**Detailed Solution:**

1. First we input the Fraud.csv data in our dataframe.

After importing the dataset we got that dataset having 6.3M rows and 11 columns

1. After that we check the null values,and there is no null values in the whole dataset.
2. Then We plot the Heatmap for finding the correlation between the columns and we got the results:

a)OldbalanceOrg and NewbalanceOrg are highly correlated.

b) OldbalanceDest and NewbalanceDest are also highly correlated.

4)After that there are some columns (isFlaggedFraud,nameDest,nameOrig,step) which do not affect the response column(isFraud) while prediction,So we drop these columns.

In preprocessing, In the dataset we are having ‘type’ column which is categorical column so it is necessary to do the Label Encoding on this column for regression we need all the columns in continuous format.

5)For predicting the response we have to split the data in Training and Testing.

We split the data 80% for training and 20% for Testing.

6)After Preprocessing the the whole data is in the continuous format But it is seen that there is too much variation between the columns So for that purpose we done the standard scaling in the both training and testing data.

7)Then the data is fully prepared for applying the Regression Models

So we have the output in 0 and 1 format means we have to do the classification. So we applied Logistic, Decision Tree, KNN algorithms and calculated the accuracy score,confusion matrices and classification report for each of them and then compared.

**8)Conclusion:**

| **REGRESSION** | **ACCURACY\_SCORE (in %)** |
| --- | --- |
| Logistic | 99.89 |
| Decision Tree | 99.89 |
| KNN | 99.95 |
| Random Forest | 99.89 |

We have seen that accuracies of Logistic, Decision Tree,Random Forest are equal but KNN is having the highest Accuracy Score among all the regressions.

So we can go with anyone among all the algorithms because all of them have almost equal accuracy score.

But for more accurate prediction it is better to go with **KNN** Algo.

o ***ROC curve:***

o ROC Curve is plotted to visualize true-positive rate and true negative rate.

o AUC-ROC of KNN is 0.86 which is more than other models(Decision tree, logistic regression, Random Forest)

That's why for given problem statement KNN algorithm is best suitable

o ***Classification report:***

o It gives complete picture of precision-recall and f1-score

o KNN gives more recall rate and precision than other models.

o KNN is important in this problem as we want our model to spot as many real 1 as possible.

o Thus KNN is the best performing model.