PHASE-3 PROJECT- PREPROCESSING THE DATA SET

PROJECT TITLE – CREDIT CARD FRAUD DETECTION.

PRE-PROCESSING:

INTRODUCTION:

Data preprocessing is the process of converting raw data into a format that can be analyzed by computers and machine learning. It's an important step in the data preparation stage.

The goal of data preprocessing is to make the data accurate, consistent, and suitable for analysis. It helps to improve the quality and efficiency of the data mining process.

Data preprocessing involves:

1. Detecting and correcting (or removing) corrupt or inaccurate records from a dataset
2. Identifying incorrect, incomplete, irrelevant parts of the data
3. Modifying, replacing, or deleting the dirty or coarse data
4. Converting the data into a suitable format for analysis

It is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

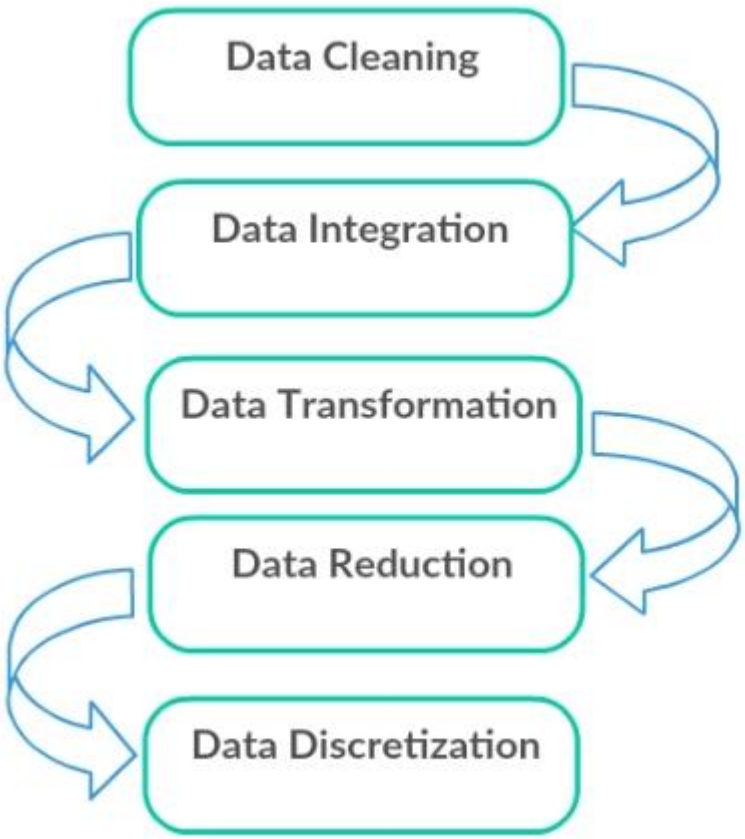
There are various approaches in Data Preprocessing:

- Aggregation.
- Sampling.
- Dimensionality Reduction.
- Feature Subset Selection.
- Feature Creation.
- Discretization and Binarization.
- Variable Transformation.

HOW TO PREPROCESS DATA :

- data in Pandas.
- Drop columns that aren’t useful.
- Drop rows with missing values.
- Create dummy variables.
- Take care of missing data.
- Convert the data frame to NumPy.
- Divide the data set into training data and test data

BLOCK DIAGRAM:



Overall, data preprocessing is essential to ensure the quality and reliability of data for further analysis and modeling. It can have a significant impact on the success of machine learning and data analysis projects.

CODING:

 # import the necessary packages

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from google.colab import files

uploaded=files.upload()
```

#Load the data set

```
dataset=pd.read_csv("creditcard.csv")
```

#printing dataset

```
dataset.info()
```

 # Print the shape of the dataset

```
print(dataset.shape)
```

```
print(dataset.describe())
# Determine number of fraud cases in dataset
fraud = dataset[dataset['Class'] == 1]
valid = dataset[dataset['Class'] == 0]
outlierFraction = len(fraud)/float(len(valid))
print(outlierFraction)
print("Fraud Cases: {}".format(len(dataset[dataset['Class'] == 1])))
print("Valid Transactions: {}".format(len(dataset[dataset['Class'] == 0])))
print(“Amount details of the fraudulent transaction”)
fraud.Amount.describe()
print(“details of valid transaction”)
valid.Amount.describe()
# Correlation matrix
corrmat = dataset.corr()
fig = plt.figure(figsize = (12, 9))
sns.heatmap(corrmat, vmax = .8, square = True)
plt.show()
#splitting data into training and testing sets
from sklearn.model_selection import train_test_split
# Split the data into training and testing sets
xTrain, xTest, yTrain, yTest = train_test_split( xDataset, yDataset, test_size = 0.2, random_state = 0)
```

ALGORITHM:

1) STEP 1 : Import the required python packages and libraries.
2) STEP 2 : Load the CSV data set using pandas.
3) STEP 3 : Print the dataset.
4) STEP 4 : Drop the unnecessary rows.
5) STEP 5 : Drop the unnecessary columns.
6) STEP 6 : Plot the correlation graph.
7) STEP 7 : Split the datasets into training and testing sets.
8) STEP 8 : View the output and correlation graph.

Conclusion:

Thus the given dataset for credit card fraud detection was analysed and pre-processed.