CREDIT CARD FRAUD DETECTION

PHASE 3:SUBMISSION

❖ Loading the data:

➤ The first step in any machine learning project is to load the data. This can be done using a variety of tools, such as Python libraries like pandas or NumPy. When loading the data, it is important to pay attention to the following:

✓ File format:

◆ What file format is the data in? The most common file formats for machine learning data are CSV, JSON, and XML.

✓ Data types:

◆ What are the data types of the different columns in the dataset? This information is important for ensuring that the data is loaded correctly.

✓ Missing values:

◆ Does the dataset contain any missing values? If so, how will you handle them?

❖ Preprocessing the data:

➤ Once the data has been loaded, it needs to be preprocessed before it can be used to train a machine learning model. This process involves cleaning and transforming the data to make it more suitable for machine learning. Some common preprocessing tasks include:

✓ Handling missing values:

◆ Missing values can be handled in a variety of ways, such as dropping the rows with missing values, imputing the missing values with a default value, or using a more sophisticated method like k-nearest neighbors (KNN) imputation.

Encoding categorical variables:

- Categorical variables, such as country or product category, need to be encoded before they can be used by machine learning algorithms. This can be done using a variety of methods, such as one-hot encoding or label encoding.

Scaling the data:

- Scaling the data ensures that all of the features are on the same scale, which can improve the performance of machine learning algorithms. Some common scaling methods include standard scaling and min-max scaling.

The following is an example of how to load and preprocess data for a credit card fraud detection project using Python:

```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.model_selection import train_test_split

import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import seaborn as sns
```