**Module 20 Report:**

The purpose of the model is to generate a model that will be able to correctly identify healthy loans and high-risk loans. The application is based on loan size, interest rate, borrower income, debt to income, number of accounts, derogatory marks and total debt.

**Process:**

**Split the Data into Training and Testing Sets**

Step 1: I have read the CSV file into a Pandas Data Frame.

Step 2: Then separated labels and features as y variables and x variables.

Step 3: I have used train test split technique to divide datasets into two subsets.

**Create a Logistic Regression Model with the Original Data**

Step 1: In the first step I have used Logistic Regression algorithm.

Step 2: In the second step I have used KNeighbour algorithm.

Step 3: In this step I have generated confusion matrix algorithm to visualize and summarize.

Result:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.99 | 1.00 | 1.00 | 18681 |
| 1 | 0.93 | 0.87 | 0.90 | 703 |
|  |  |  |  |  |
| accuracy |  |  | 0.99 | 19384 |
| macro avg | 0.96 | 0.93 | 0.95 | 19384 |
| Weighted avg | 0.99 | 0.99 | 0.99 | 19384 |

Summary:

 The Logistic regression model does very well at predicting healthy loans with accuracy, Precision is close to 99% and recall is close to 100%, it does well. But it does not do well with identifying high-risk loans where Precision is close to 91% and recall is close to 84%. this suggest that this model is less classify high-risk loans than healthy loans.