School of Computer Science Engineering and Technology

Course-BTech	**Type - AI Core-1**
Course Code - CSET211	**Course Name - Statistical Machine Learning**
Year - Second	**Semester - ODD**
Date - 16/09/2024	**Batch - CSE 3rd Semester**

Lab Assignment - 6: Logistic Regression

CO- Mapping

Section	CO1	CO2	CO3
Section 1: Q1-Q6	✓		
Section 2: Q7-Q10	✓		✓

Dataset Download:

Pima Indians Diabetes Dataset is available at
Kaggle link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
Github link: https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset/blob/master/diabetes.csv

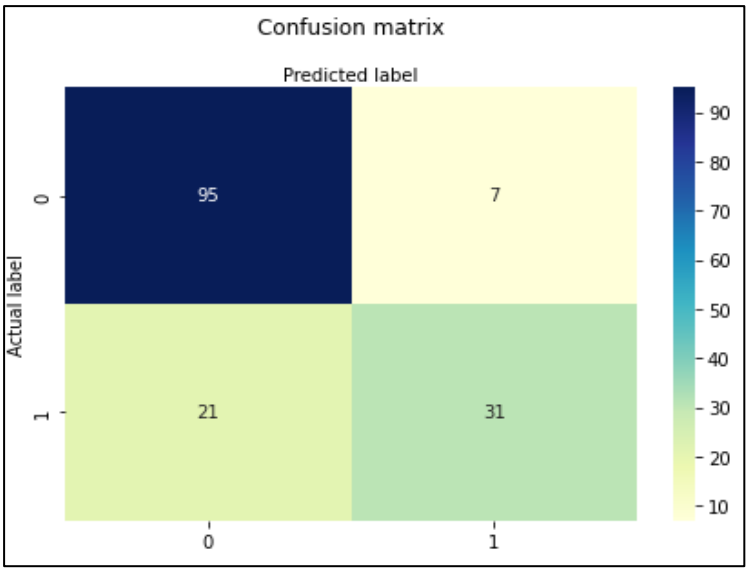
Section 1: Data Preprocessing on Dataset

1. Given a dataset ***diabetes.csv***, write a Python script to load and display the dataset.
2. Rename the columns accordingly: ***'Pregnancies'*** to ***'Pregnant'***, ***'BloodPressure'*** to ***'BP'***, ***'SkinThickness'*** to ***'Skin'*** and ***'DiabetesPedigreeFunction'*** to ***'Pedigree'*** and display again.
3. Use the describe() function to print the statistical summary of the data in the dataframe.
4. Consider the ***'Pregnant'***, ***'BP'***, ***'Insulin'***, ***'BMI'***, ***'Pedigree'*** and ***'Age'*** to be the feature columns and split the dataset into 80% train and 20% test data.
5. Create a scatterplot showing the relation between ***'BMI'*** and ***'Age'*** on the training data with ***'Outcome'*** as hue.
6. Perform Standardization using ***StandardScaler()*** function.

Section 2: Logistic Regression Model

7. Train a logistic regression model using an inbuilt ***LogisticRegression()*** function on train set.
8. Calculate the confusion matrix using inbuilt function ***confusion_matrix()*** and display it using heatmap.

The output will be something like this:



9. Calculate the accuracy and f1-score of the model using ***accuracy_score*** and ***f1_score*** inbuilt methods respectively.
10. Print the classification report using in ***classification_report()*** function with the target names ***'with diabetes'*** and ***'without diabetes'***.

The output will be like this:

	precision	recall	f1-score	support
without diabetes	0.82	0.93	0.87	102
with diabetes	0.82	0.60	0.69	52
accuracy			0.82	154
macro avg	0.82	0.76	0.78	154
weighted avg	0.82	0.82	0.81	154

Platform Required: Anaconda, Editor: Jupyter/Spyder/Pycharm/Google Colab

Submission Instructions:

- Submission required .ipynb file only
- Submission is through LMS only.