**Statistical Learning Lab**

**Assignment - 2**

**Logistic Regression Assignment**

**Show the code snippets and the corresponding output for the following:**

1. Load the dataset “diabetes.csv”. Display first few rows of the dataset.
2. Perform preliminary analysis to show how the variables are related to each other. Use scatter plot, box plot etc. to visualize how different variables impact the “Outcome” variable.
3. Randomly sample 80% of the data as training data and rest as test data. Fit a Logistic Regression model with all the predictors on training data. From the summary which factors seem to be significant? Explain how the predictors impact the log-odds ratio of diagnosed with diabetes (Outcome)
4. From the model fitted in problem 3, derive confusion matrix, accuracy, and F1-score on test data.
5. Let’s call the model fitted in problem 3 M1. Now choose predictors “Pregnancies”, “Glucose” and “BMI” and fit a model (M2). Compare the deviances among these two models and perform hypothesis test to show whether M1 is significantly more informative than M2.

Data can be downloaded from: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

Description of the study:

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (p. 261). American Medical Informatics Association.