**Birla Institute of Technology & Science, Pilani**

**Work Integrated Learning Programmes Division**

**Introduction to Data Science (S2-21\_DSECSZG523)**

**Assignment – I**

**Instructions**

1. Download the open source statistical tool “JASP” from the URL - <https://jasp-stats.org/download/> and install locally in your system. You may choose to use JASP or Python to execute the statistical / machine learning portions of the assignment.
2. Download “Pima Indian Diabetes” dataset from the URL - <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

**Assignment Question**

1. Import the “Pima Indian Diabetes Dataset” in CSV format. The dataset contains features such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age which are Independent variables in the study and “Outcome” which is the Dependent variable. If the patient is diabetic, the ‘Outcome’ value will be 1, otherwise it will be 0. The dataset contains a total of 768 rows of labelled data. Perform the following operations on the dataset.

**I. Descriptive Analytics**

a. What are the mean, standard deviation and outlier values for Age and BMI variables?

**II. Diagnostic Analytics**

a. Explain the correlation between BMI and BloodPressure variables (positive, negative or neutral).

**III. Predictive Analytics**

Apply Decision Tree Classifier on the dataset. Consider the X variables as Age, BMI, Insulin, Blood Pressure, Pregnancies, SkinThickness, Glucose and DiabetesPedigreeFunction; and Y variable as Outcome. Keep the train-test split at 70-30 ratio.

a. What is the overall accuracy of the decision tree model?

b. Write the confusion matrix. Discuss the implications of Type-I error and Type-II error in this example.

c. What are the Precision, Recall, F1 score, and Error Rate values based on the confusion matrix?