**DEPARTMENT OF INFORMATION TECHNOLOGY**

**Course: Data Mining & Business Intelligence Lab (ITL601)**

**B.Tech. (Information Technology) – Semester VI**

**Academic Year: 2023-24 (Even Semester)**

**PRACTICAL 5**

**Aim:** Implement and evaluate any classification algorithm in Python.

**Lab Objective:** To learn how to gather and analyse large sets of data to gain useful business understanding.

**Theory:**

Classification is a fundamental task in supervised machine learning, where the objective is to categorize input data points into predefined classes or categories based on their features. This process involves training a model on a labeled dataset, where each input is associated with a known output class. During the training phase, the algorithm learns the relationship between the features of the input data and their corresponding classes.

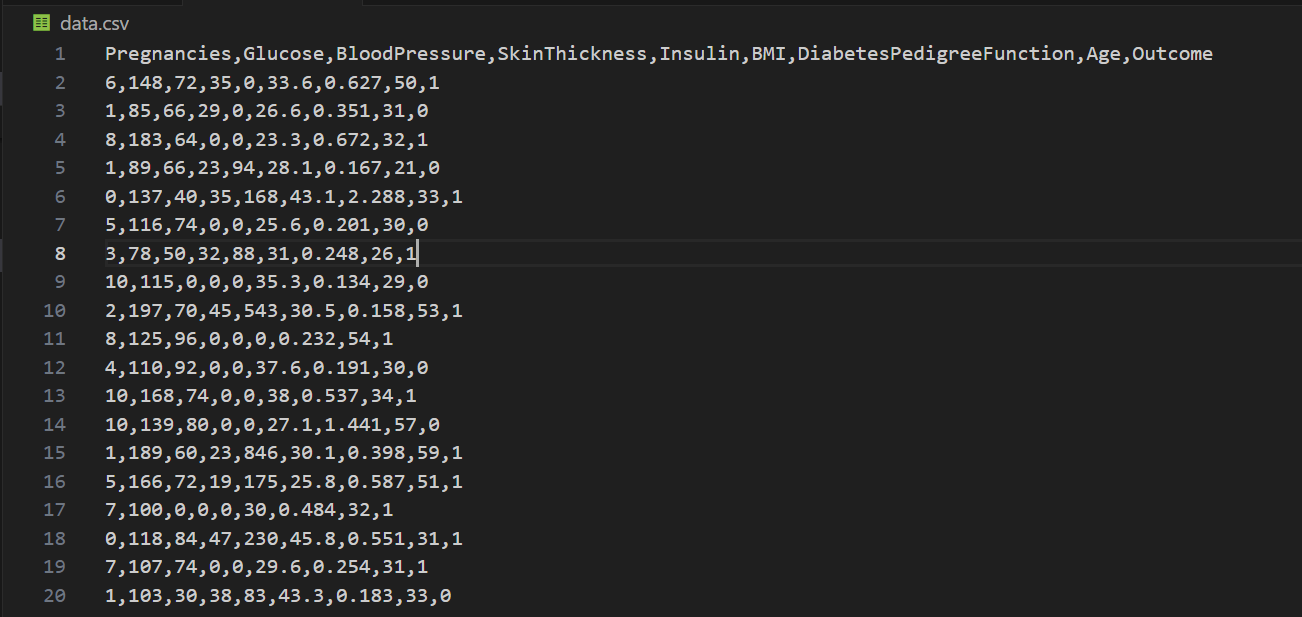
**In the training phase**, the dataset used typically comprises labeled data, where each input sample is annotated with its respective class label. Features, which are the characteristics or attributes describing each input, are used by the model to discern patterns and make predictions. Through model training, the algorithm derives insights from the labeled data to accurately classify new, unseen data in subsequent stages.

**During the testing phase**, the trained model is evaluated on unlabeled data, where the true class labels are unknown. The model then predicts the class labels for the test data based on the patterns it learned during training. Performance metrics such as accuracy, precision, recall, and F1 score are employed to assess how effectively the model generalizes to new, unseen data.

Classification tasks encompass various types, including binary classification, which involves categorizing instances into two classes, and multiclass classification, which extends the classification to multiple classes. Additionally, multi-label classification assigns multiple labels to each instance, accommodating scenarios where an instance may belong to multiple categories simultaneously.

A range of classification algorithms exists to address diverse data and problem domains. These include logistic regression, decision trees, random forests, support vector machines (SVM), k-nearest neighbors (KNN), naive Bayes, and neural networks. Each algorithm offers distinct advantages and is suited to different types of data and problem complexities.

**Dataset :**

****

**Code :**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score

# Read the data

data = pd.read\_csv('data.csv')

# Split data into features and target

X = data.drop('Outcome', axis=1)  # Features

y = data['Outcome']  # Target

# Split data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize decision tree classifier with hyperparameters

clf = DecisionTreeClassifier(max\_depth=5, min\_samples\_split=2, min\_samples\_leaf=1)

# Fit the classifier on the training data

clf.fit(X\_train, y\_train)

# Predictions for training set

y\_train\_pred = clf.predict(X\_train)

# Predictions for testing set

y\_test\_pred = clf.predict(X\_test)

# Calculate accuracy on training set

accuracy\_train = accuracy\_score(y\_train, y\_train\_pred)

print("Accuracy on Training Set:", accuracy\_train)

# Calculate accuracy on testing set

accuracy\_test = accuracy\_score(y\_test, y\_test\_pred)

print("Accuracy on Testing Set:", accuracy\_test)

# Test with a new patient

# Example: [Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age]

new\_patient = [[6, 148, 72, 35, 0, 33.6, 0.627, 50]]

predicted\_outcome = clf.predict(new\_patient)

print("Predicted Outcome for New Patient:", predicted\_outcome[0])

**Output:**

****

**Prediction**



**Conclusion:**In this Python implementation, a classification algorithm was applied to a dataset, and its performance was evaluated using accuracy as the metric. Classification involves categorizing data into predefined classes based on their features. The chosen algorithm demonstrated its ability to learn patterns from the training data and make accurate predictions on unseen data. Further optimization and fine-tuning may enhance the model's performance and generalization capabilities for real-world applications.

**Lab Outcome:** Learned to gather and analyse large sets of data to gain useful business understanding.

**Submitted Details -**

**Name of Student:** Parth Malviya

**Roll No.: 22**

**Date of Performance:** 06/02/2024

**Date of Submission:** 28/02/2024