Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603 110
(An Autonomous Institution, Affiliated to Anna University, Chennai)

UCS2612 Machine Learning Laboratory

Academic Year: 2023-2024 Even **Batch: 2021-2025**
Faculty In-charges: Y.V. Lokeswari & Nilu R Salim **VI Semester A & B**

A. No. : 7 . **Predicting Diabetes using decision tree**

Download the diabetics dataset from the link given below:

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

Develop a python program to predict diabetics using Decision Tree Model. Visualize the features from the dataset and interpret the results obtained by the model using Matplotlib library. **[CO1, K3]**

Use the following steps to do implementation:

1. Loading the dataset.
2. Pre-Processing the data (Handling missing values, Encoding, Normalization, Standardization).
3. Exploratory Data Analysis.
4. Feature Engineering techniques.
5. Split the data into training, testing and validation sets.
6. Train the model.
7. Test the model.
8. Measure the performance of the trained model.
9. Represent the results using graphs.

………………………………………………………………………………………………

Upload the code in GitHub and include the GitHub main branch link in the assignment PDF.

Hints to do the assignment:

Do the following:

1. Load the dataset.
2. Pre-Processing the data (Handling missing values, Encoding, Normalization, and Standardization).
3. Exploratory Data Analysis
4. Feature Engineering techniques.
   Refer to
   https://machinelearningmastery.com/feature-selection-machine-learning-python/
   https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

https://www.datacamp.com/tutorial/feature-selection-python

5. Apply Decision Tree algorithm on the input dataset and perform classification.
   Use Entropy and Gini-index as impurity measure. Construct Decision Tree model and compare both results.
   https://scikit-learn.org/stable/modules/tree.html
   https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
   https://www.datacamp.com/tutorial/decision-tree-classification-python
   https://statisticallyrelevant.com/decision-trees-in-python-predicting-diabetes/
   https://github.com/bushra-ansari/Diabetes-Prediction-by-Decision-Tree-Algorithm

6. Upload python project in GitHub and explore all git commands. Git Commands Tutorial : https://git-scm.com/docs/gittutorial

   Upload IPython to GitHub
   https://reproducible-science-curriculum.github.io/sharing-RR-Jupyter/01-sharing-github/

   Additional Reference:
   https://www.youtube.com/watch?v=LlrKTV4-ftI

Upload the code in GitHub and include the GitHub main branch link in the assignment PDF.

Upload python project in GitHub and explore all git commands.

Git Commands Tutorial : https://git-scm.com/docs/gittutorial