School of Computer Science Engineering and Technology

Course- BTech
Course Code- 301
Year- 2022
Date- 20-01-2022

Type- Core
Course Name-AIML
Semester- Even
Batch- 4th Sem (SPL)

3 - Lab Assignment No. 3.2

Objective: To implement Logistic Regression (using Scikit-learn) to predict Diabetes (as Positive or Negative) and handling of numerical and categorical features.

1. **Download** the dataset from http://archive.ics.uci.edu/ml/machine-learning-databases/00529/diabetes_data_upload.csv . The first 16 attributes are the symptoms all of which are categorical values except the 'Age' which contains numerical values between [16, 90]. The last attribute contains the diabetes class (positive or negative). (10)

2. **Load** the data and print first 10 and last 10 rows using a suitable function. (5)

3. **Transform** all the categorical features into numerical features using label encoding. (15)

4. Since the **'Age'** feature is in a larger range, it can be pre-processed into a smaller scale by using different methods such as Standardization (using StandardScaler()), Scaling (using MinMaxScaler()) or Normalization (using normalize()). (10)

5. Define **X** matrix (independent features) and **y** vector (target feature). (5)

6. **Split** the dataset into **80% for training** and rest **20% for testing** (sklearn.model_selection.train_test_split function) (5)

7. **Train** Logistic Regression Model using built-in function on the training set (sklearn.linear_model.LogisticRegression class). (10)

8. Use the trained model to **predict** on the **test set** and then (15)
   a. Print 'Accuracy' obtained on the testing dataset i.e. (sklearn.metrics.accuracy_score function)
   b. Confusion matrix (sklearn.metrics.confusion_matrix),
   c. Precision, Recall and F1 scores (sklearn.metrics.precision_recall_fscore_support)

9. Compare and analyse the **test accuracy** for different train-test splits of data such as 60-40, 70-30, 80-20 and 90-10 with the help of **suitable graphs**. (15)

Additional fun (will not be evaluated)

10. Implement Logistic Regression from scratch by modifying the Linear Regression implementation. **You just need to modify the hypothesis function.**