School of Computer Science Engineering and Technology

Course- BTech
Course Code- 301
Year- 2022
Date- 26-01-2022

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

7 - Lab Assignment # No. (7.1)

Objective: To analyze the performance of different supervised machine learning in the classification and regression.

Regression Models:

Step 1: Go to UCI machine learning repository and download the Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set. The dataset was created with records of behavior of the urban traffic of the city of Sao Paulo in Brazil.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	135	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	18	**Date Donated**	2018-12-12
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	71028

One can download from the link:
https://archive.ics.uci.edu/ml/machine-learning-databases/00483/ (5)

Step 2: Apply the pre-processing steps on the interested datasets (10)

a) Remove unwanted features if any.
b) Handle Null value or enormous values if any.
c) Use one hot encoding (OHE) or multi hot encoding to labeling, if any.
d) Convert categorical value to numerical value if any.
e) Standardize the data into [0,1].

Step 3: Split the improved Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set after Step 2 into training and testing in the following ratio: (5)

a) 70:30
b) 30:70

Step 4: Train and Test the different regression model using sklearn library function (20)

a) Linear regression
b) Polynomial Regression
c) Lasso Regression
d) Ridge Regression

Step 5: Evaluate the performance of different regression models on various performance metrics: (10)

a) Confusion Metric
b) Mean Absolute Error (MAE)
c) Mean squared error (MSE)
d) Root mean squared error (RMSE)
e) R Square / Coefficient of Determination/ Goodness of fit

Step 6: Visualize the results of different regression models into different plots: (10)

a) regplot ().
b) FacetGrid using lmplot ().

Classification Models:

Step 1: Go to UCI machine learning repository and download the Mushroom Data Set. This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family.

Data Set Characteristics:	Multivariate	**Number of Instances:**	8124	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	22	**Date Donated**	1987-04-27
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	712556

One can download from the link:
https://archive.ics.uci.edu/ml/datasets/mushroom/ (5)

Step 2: Apply the pre-processing steps on the interested datasets (10)
a) Remove unwanted features if any.
b) Handle Null value or enormous values if any.
c) Use one hot encoding (OHE) or multi hot encoding to labeling, if any.
d) Convert categorical value to numerical value if any.
e) Standardize the data into [0,1].

Step 3: Split the improved Mushroom Data Set after Step 2 into training and testing in the following ratio: (5)
a) 70:30
b) 30:70

Step 4: Train and Test the different Classification model using sklearn library function (20)
a) Logistic Regression
b) Decision Tree
c) Random Forest
d) AdaBoost
e) Gradient Boost

Step 5: Evaluate the performance of different Classification models on various performance metrics: (10)
a) Confusion Metric
b) Accuracy
c) Precision
d) Recall
e) F Score

Step 6: Visualize the results of different Classification models into different plots: (10)
a) Heatmap () for Confusion matrix
b) Barchart for accuracy, precision, recall and f-score.

Output: Show the best classification model and regression model.