School of Computer Science Engineering and Technology

Course- BTech
Course Code- CSET301
Year- 2022
Date- 26-09-2022

Type- Core
Course Name-AIML
Semester- Odd
Batch- V Sem

Lab Assignment No. 6.1.1

Exp. No.	Name	CO-1	CO-2	CO-3
6.1.1	Decision Tree Classifier	✓	✓	--

Objective: To implement Decision Tree Classifier (DT) (using Scikit-learn) and perform binary classification after suitable pre-processing steps.

Download the dataset from:
https://archive.ics.uci.edu/ml/datasets/Raisin+Dataset (10)

About Dataset:

Data Set Characteristics:	Multivariate	Number of Instances:	900	Area:	Life
Attribute Characteristics:	Integer, Real	Number of Attributes:	8	Date Donated	2021-04-01
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	1532071

Images of Kecimen and Besni raisin varieties grown in Turkey were obtained with CVS. A total of 900 raisin grains were used, including 450 pieces from both varieties. These images were subjected to various stages of pre-processing and 7 morphological features were extracted. These features have been classified using three different artificial intelligence techniques.

Attribute Information:

1. Area: Gives the number of pixels within the boundaries of the raisin.
2. Perimeter: It measures the environment by calculating the distance between the boundaries of the raisin and the pixels around it.
3. MajorAxisLength: Gives the length of the main axis, which is the longest line that can be drawn on the raisin.
4. MinorAxisLength: Gives the length of the small axis, which is the shortest line that can be drawn on the raisin.
5. Eccentricity: It gives a measure of the eccentricity of the ellipse, which has the same moments as raisins.
6. ConvexArea: Gives the number of pixels of the smallest convex shell of the region formed by the raisin.
7. Extent: Gives the ratio of the region formed by the raisin to the total pixels in the bounding box.
8. Class: Kecimen and Besni raisin.

1. **Data Pre-processing step: (40)**
    a) Read Raisin_Dataset using Pandas and display First 5 rows.
    b) Check the presence of Null Values/Missing Values. If present handle them with suitable approach.
    c) Covert the Class value into discrete: Kecimen as '0' and Besni raisin as '1' class.
    d) Check Feature importance using Chi-Square (Hint: sklearn.feature_selection.chi2)
    e) Discard the least important features using chi-square value.
2. Split the dataset into 80% for training and rest 20% for testing (sklearn.model_selection.train_test_split function) (5)
3. Train DT classifier **using** built-in function on the training set with default parameters (sklearn.tree.DecisionTreeClassifier)(10)
4. Evaluate the train model using testset with the help of confusion matrix, Accuracy, Precision and Recall.
5. Set the criteria as entropy and log_loss and train the model and evaluate it on testset.
6. Parameter Tuning:
    a. Try with max_depth as [10, 100]
    b. Min_samples_split as [4, 6,8]
    c. max_features {"auto", "sqrt", "log2"}
7. Compare the results and find the best suitable model

Suggested Platform: Python: Jupyter Notebook/Azure Notebook/Google Colab.