

**B.Sc. II Year IV Semester (CBCS): Data Science Syllabus
(With Mathematics Combination)
(Examination at the end of Semester - IV)
Practical - 4: Machine Learning (Lab)**

Objective:

The main objective of this laboratory is to put into practice the various machine learning algorithms for data analysis using Python and Weka.

ML Toolkits

Students are expected to learn

1. Scikit-learn(<https://scikit-learn.org/>) an open source machine learning Python library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.
2. Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) is another widely used ML toolkit.

Datasets

1. The sklearn.datasets package embeds small toy datasets. It includes utilities to load these datasets. It also includes methods to load and fetch popular reference datasets and features some artificial data generators. Students are expected to study and make use of these datasets.
2. Weka also has provides various data sets.

References:

1. scikit-learn user guide.https://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf
2. Ian Witten, Eibe Frank, and Mark Hall, Chris Pal. DATA MINING: Practical Machine Learning Tools and Techniques, 4th Edition. Morgan Kaufmann.

Exercises

1. Write a Python program using Scikit-learn to split the iris dataset into 70% train data and 30% test data. Out of total 150 records, the training set will contain 120 records and the test set contains 30 of those records. Print both datasets
2. Write Python program to use sklearn's DecisionTreeClassifier to build a decision tree for the sklearn's datasets. Implement functions to find the importance of a split (entropy,

information gain, gini measure)

3. Write a Python program to implement your own version of the K-means algorithm. Then

apply it to different datasets and evaluate the performance.

4. Design a perceptron classifier to classify handwritten numerical digits (0-9).

Implement

using scikit or Weka.

5. Write a Python program to classify text as spam or not spam using the Naïve Bayes Classifier

6. Use WEKA and experiment with the following classifiers: Association Rule Mining (Apriori), Agglomerative and Divisive Clustering.