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. : 8. **Applications of Random Forest and AdaBoost Ensemble Techniques**

Download the Wisconsin Breast Cancer Diagnostic dataset from the link given below:

https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at http://www.cs.wisc.edu/~street/images/

There are 569 instances and 30 features. Target is Diagnosis.
Develop a python program to diagnose breast cancer using Ensemble Models. 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. Compare the results of each ensemble model using graphs.
10. Represent the ROC of training and test results in the 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 Ensemble techniques such as Bagging, Random Forest and AdaBoost on the input dataset and perform classification.

   Refer to the following sources.

https://scikit-learn.org/stable/modules/ensemble.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html

   Construct Ensemble models and compare both results.

https://www.kaggle.com/code/faressayah/ensemble-ml-algorithms-bagging-boosting-voting

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