**Assignment 8 - Web mining**

1. Implement the Crawler (DFS & BFS). The program should able to take seed input as any URLs/page and display all the links obtained by crawler.

2. Implement the PageRank algorithm to calculate the rank of each page in the file.The output should be the 10 pages with the highest rank, together with their rank values.

a. Tabulate the results containing adjacency matrix and rank of pages.

3. Implement the HITS algorithm to calculate the hub and the authority weight of each web page in the data set. The output should be the 10 most authoritative pages and 10 most hubby pages.

a. Tabulate the results containing adjacency matrix and rank of pages.

*Dataset -- Web Graph datasets from “Stanford Large Network Dataset Collection”*

[*http://snap.stanford.edu/data/#web*](http://snap.stanford.edu/data/#web)

**Assignment No. 7 - apriori**

1. Implement the Apriori algorithm for generating Association Rules

2. Experiment with different values of support, confidence, and maximum rule length.

3. Tabulate the results containing frequent item sets and, total number of rules generated for different support and confidence.

4. Find the interesting rules from above obtained rules using followingmetrics/measures

a. Lift

b. Chi-Square Test χ

c. All\_confidence measure

d. Max\_confidence measure

e. Kulczynski measure

f. Cosine measure

*Dataset --*

*https://archive.ics.uci.edu/ml/datasets/congressional+voting+records*

*b. Extended BAKERY dataset (1000, 5000, 20,000 and 75,000 transactions per set)*

[*https://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab01.html*](https://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab01.html)

**Assignment 6 – Clustering**

1. Design and implement the following clustering algorithm:

a) Hierarchical clustering - AGNES & DIANA. Plot Dendrogram.

b) k-Means

c) k-Medoids (PAM)

d) BIRCH

e) DBSCAN

2. Tabulate the results with cluster validation accuracy

*Dataset --*

*a) IRIS*

*b) Breast Cancer*

*c) For BIRCH / DBSCAN, use US Census Data (1990) Data Set*

[*https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)*](https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990))

**Assignment 5 – Bayes and knn classifier**

1. Design and implement the following classifiers:

a) Regression classifier.

b) Naïve Bayesian Classifier.

c) k-NN classifier (Take k = 1,3,5,7)

d) Three layer Artificial Neural Network (ANN) classifier (use back

propagation). Plot error graph (iteration vs error).

2. Tabulate the results in confusion matrix and evaluate the performance of above

classifier using following metrics :

a) Recognition rate

b) Misclassification rate

c) Sensitivity

d) Specificity

e) Precision & Recall

*Dataset – Iris and Breast Cancer*

**Assignment 4 – Rule Based Classifier**

Design the rule based classifier : Extract the rules from decision tree build in

assignment no. 3.

2. Tabulate the results and evaluate the performance of rules generated using

following metrics :

a. Coverage

b. Accuracy

c. Toughness (size)

*Dataset -- Balance Scale, Car evaluation and Breast-cancer dataset.*

**Assignment 3 – Decision Tree**

1. Implement the decision tree classifier using the following attribute selection

measures and graphically show/visualize the tree:

a. Information Gain

b. Gain Ratio

c. Gini Index

2. Tabulate the results in confusion matrix and evaluate the performance of above

classifier using following metrics :

a) Recognition rate

b) Misclassification rate

c) Sensitivity

d) Specificity

e) Precision & Recall

*Dataset -- Balance Scale, Car evaluation and Breast-cancer dataset.*

**Assignment No. 2 - Correlation, normalization**

1. Correlation analysis - Chi-Square Test

a. User should be able to choose any two attributes.

b. Display the contingency table

c. Show the chi-square value and conclusion whether the selected attributes are

correlated or not.

2. Correlation analysis – Correlation coefficient (Pearson coefficient) & Covariance

a) User should be able to choose any two attributes.

b) Show the calculated values and conclusion whether the selected attributes

are correlated or not.

3. Normalization using following techniques :

a. Min-max normalization

b. Z-Score normalization

c. Normalization by decimal scaling

i. User should be able to choose any attributes.

ii. Show the calculated values in tabulated form.

iii. Show the scatter plot for normalized attributes

**Assignment 1 – Five summary**

1. Data upload ( std. format like .csv, excel etc. ) and view

2. Calculate and show the measures of central tendency for uploaded data : mean ,

median , mode , midrange , variance and standard deviation

3. Calculate and show the dispersion of data : range , quartiles , interquartile range

, five-number summary

4. Graphical display of above calculated statistical description of data (provide the

facility - UI form to choose different attributes from uploaded data set) :

a. Quantile plot

b. Quantile-quantile (q-q) plot

c. Histogram

d. Scatter plot

e. Boxplot