CSE 3024: Web Mining **Slot: L39 + L40**

Online Submission Deadline: 11th May 2021

Web Content and Usage Mining

[3 + 3 + 3 + 6]

- **This assignment can be carried out individually or in a group of 2.**
- **In case of group assignment, please mention the name and registration number of both the members in the title page and both the members must upload the same copy.**
- **Upload your code and result as a single PDF file in VTOP [Mandatory] and MS Team Assignment [optional] on or before the deadline.**
- **No other form of submission will be acceptable.**
- **If you failed to upload in VTOP on or before the deadline, but successfully uploaded in MS Team Assignment, then 2 marks of penalty will be imposed on the secured marks.**
- **If you fail to upload your assignment in both VTOP and MS Team Assignment, then your assignment will not be evaluated and ZERO (0) mark will be awarded.**
- **File should contain**
  - **Question**
  - **Code**
  - **Result / Output screen**

---

1. Write a python program to show the implementation of Decision Tree and Naïve-Bayes techniques using the below mentioned dataset.
   - Handle missing values, If any
   - Use 5-fold cross validation technique
   - Prepare the confusion matrix, find out the precision, recall value, F-measure and prediction accuracy.
   - Prepare ROC and AUC curve based on the result obtained.
   - Compare the results obtained using these two techniques in order to assess their performance for the considered dataset.

   The detailed description of the dataset is given in the below link:
   https://archive.ics.uci.edu/ml/datasets/Facebook+Large+Page-Page+Network

   [*Note:*- Consider only the musae_facebook_target.csv from the downloaded zip file for classification.]

2. Write a program to show the implementation of agglomerative hierarchical clustering (single, complete and average linkage) using the below mentioned dataset. Show the resultant clusters using graph and dendrogram.
   - Consider Euclidean distance as measure
   - Handle missing values, if any

https://drive.google.com/open?id=1FGHlK1Ffn6RvxfMFMhBIYr6o7c-JfXCt

The detailed description of the dataset is given in the below link:
https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

3. Write a program to show the implementation of apriori algorithm using web log usage data for web usage mining purpose. (Consider any publicly available web log data to show the implementation.)

4. Consider the COVID-19 dataset for India given in the following link.

   https://www.kaggle.com/sudalairajkumar/covid19-in-india

   Analyze the dataset by extracting the past 15 days records for each states to cluster them with respect to *number of active cases, death rate and recovery rate ratio*. [Use any **TWO** clustering algorithms of your choice and provide the performance analysis of the techniques used w.r.t. results obtained.]

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)