DMAT – Assignment 2

Name 1: Student Number: Name 2: Student Number:

Course MSCBD-DMAT **Stage / Year** 1 **Module** Data Mining Algorithms & Techniques **Semester** 2 **Assignment** Assignment 2 **Date of Title Issue** 16th April **Assignment Deadline** 25 April at 23:55 **Assignment Submission** Upload to Moodle **Assignment Weighting** 25% of module

Objective

1. To successfully apply a set of data mining skills imparted in this module to a previously unseen datasets to achieve knowledge discovery.
2. Evaluate a well-regarded peer reviewed paper or journal article which concerns the application of one of the techniques covered in this module and comment on its relevance to your dataset.

Deliverables

A single zip called firstName1_lastName1_studentNumber1_firstName2_lastName2_studentNumbe r2_assignment2.zip to be uploaded to Moodle containing the following files:

- This file edited to contain the results of your investigation. Each of the **NUMBERED HEADINGS IN RED** should be expanded to satisfy the requirements of the section.
- A set of supporting files including but not limited to the following, which should be clearly referenced from your documentation. You only need to submit the files relevant the techniques you have explored.
    - The original dataset file
    - dataset.arff
    - trainigSet.arff
    - testingSet.arff
    - j48tree.arff

- associationrules.arff
- kmeans.arff
- dbscan.arff
- mlp.arff
- The research paper.

Choosing Your Dataset

1. Your dataset should concern a real-world problem that lends itself to easy understanding by your classmates. 2. It should not be identical to the dataset you used in assignment1. 3. It should have >1000 tuples/rows/instances. 4. It should have >=10 attributes 5. It should have attributes which can serve as labels so that the accuracy of your data analysis can be determined. 6. If you cannot find one dataset which is suitable for use with all techniques then you may choose 2. Please clearly indicate which dataset was used in which case.

The list below should help you on your search, student please share additional sources on Moodle discussion form.

- **UCI Machine Learning Repository** - A repository of more than 200 data sets for machine learning and data mining
- **Movie Ratings Data** - Real movie ratings data from **www.movielens.org** Web site. Contains ratings on 1600+ movies by 1000 users
- **Kaggle.com Competition Data Sets** - Data sets from a variety of competitions.
- **Stanford Large Network Dataset Collection** - A variety of network data sets, including data from social networks, product reviews, online communities, etc.
- **Yelp Data Set Challenge** - Reviews and check-in data on thousands of businesses.
- **Million Song Dataset** - Freely-available collection of audio features and metadata for a million contemporary popular music tracks.
- **Public Data sets on Amazon Web Services** - Large public data sets (including data sets for US Census, Wikipedia, Freebase, human genome project), ready for big data analytics on the cloud.
- **Data.gov** - Publically available data sets from Federal, State, and local government, including economic, geological, demographic and many other types of data sources. This site also includes a list of other **Open Data Sites** with similar publicly available data sources from various cities, states, and countries.

• **KDnugget's list of data sets for data mining**

• **Infochimps Data Market** - Thousands of data sets, including data from various social networks and collaborative tagging sites such as Twitter, Delicious, Last.fm, MusicBrainz, as well as data sets from many other domains.

Initial Tasks

1. Description of your dataset(s) and findings – 20%

• **Title**: Brief title to capture the data and objective of your assignment

• **Data description:** A description of the data in detail under the following subheadings:

- The problem domain
- The source of the data
- The agencies working with the data
- The intended use of the data
- The attribute types of the data

Please include screen shots (with one or two sentences of summary) of the dataset and also of the data summaries that are available though Weka.

• **Objective**: Your objective. You can update this as you progress through your assignment revising it and making it more specific.

• **Summary of Findings**: This should be written following the application of your data mining techniques.

2. Preprocessing – 10%

In this section you should

1. Identify the set of preprocessing techniques that can be applied to your data and clearly indicate which techniques are appropriate and which ones are not. 2. Provide evidence through screenshot of the effects of preprocessing the data along with a short explanation. 3. Generate a file called dataset.arff which is the outcome of the preprocessing.

3. Divide your dataset into training and test set

Divide the test into a training and testing set in the ration of (9:1). The files generated as part of this process should be saved and submitted as the

following

- trainingSet.arff and - testingSet.arff Screen shots of these files should be included.

Data Mining Techniques

Classification / Association

For each of the following classification techniques 1. Train your model using trainingSet.arff 2. Test your model using testingSet.arff 3. Write a few paragraphs analyzing the results. Be sure to vary parameters at least 3 times in each case. Support this analysis with screenshots of the following

a. The model or a visualization of the model b. The results of the model c. Any additional output of the model including but not limited to

i. Rules ii. Confidence Values iii. Confusion Matrixes iv. etc d. Simple references to the notes or URL links to online resources

complete with a sentence or two of explanation.

4. Classification/ Association: J48 Tree or Association Rules – 10%

5. Classification: MLP or a similar advanced technique from Weka – 15% If you are using a similarly advanced technique please clearly identify the technique and the steps you are taking. You may reference online tutorials and videos.

Clustering

For one of the following 2 clustering techniques

1. Use dataset.aff as input. If adaptions are necessary clearly indicate them.
2. Write one or two paragraph analyzing the results of the clustering. Be sure to vary parameters at least 3 times in each case. Support this analysis with screenshots of the following

a. The clusters and/or a visualization of the clusters b. The results of the clusters c. Any additional output of the clustering process d. Simple references to the notes or URL links to online resources

complete with a sentence or two of explanation. e. Evaluate the clusters using the “classes to clusters evaluation”. A

worked example may be found here
http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-ex3.html

6. Clustering: K-Means or DBSCAN – 10%

Time Series Forecasting

For the following task

1. Use dataset.aff as input. If adaptions are necessary clearly indicate them.
2. Write one or two paragraph analyzing the results of the forecasting.

   Support this analysis with screenshots of (whenever possible)

   a. The regression equation b.
   Diagram of the historical values c.
   Diagram of the predictions

7. Time Series Forecasting – 15%

Research publication

8. Research Publication Summary and relevance / potential relevance to your work – 20% (2-4 pages)

Please discuss under the following headings

a. Publication and researchers
b. Dataset
c. Technique (mention any adaptions)
d. Major Findings
e. Relevance / potential relevance to your work

NB: softweare requirement for this assignment is weka do not use any other softweare apart from weka fail to do so it will result as zero mark!.