DMAT – Project 1

Name 1: Student Number:

Name 2: Student Number:

|  |  |
| --- | --- |
| **Course** | MSCBD-DMAT |
| **Stage / Year** | 1 |
| **Module** | Data Mining Algorithms & Techniques |
| **Semester** | 2 |
| **Assignment** | Project |
| **Date of Title Issue** | 13th April |
| **Assignment Deadline** | 6th May |
| **Assignment Submission** | Upload to Moodle |
| **Assignment Weighting** | 30% of module |

# Group Project

You will be working in groups of two to complete this project. Email [alex.cronin@griffith.ie](mailto:alex.cronin@griffith.ie) with the names of the people in your group and cc the other group member. I suggest that you work through Google Drive with this

Group members:

1) Pronoy D’sa -2952742

2) Gaurav Singh - 2957104

document being stored as a Google Doc so that you can both work together.

# 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.
3. Concisely communicate your findings of the above activities in a 10 min presentation, which will be followed by questions.

# Deliverables

A single zip called firstName1\_lastName1\_studentNumber1\_firstName2\_lastName2\_studentNumber2\_project.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**](http://archive.ics.uci.edu/ml/)- A repository of more than 200 data sets for machine learning and data mining
* [**Movie Ratings Data**](http://facweb.cs.depaul.edu/mobasher/classes/ect584/data/movielens.zip) - Real movie ratings data from [**www.movielens.org**](http://www.movielens.org/) Web site. Contains ratings on 1600+ movies by 1000 users
* [**Kaggle.com Competition Data Sets**](http://www.kaggle.com/competitions) - Data sets from a variety of competitions. Also a good source for class project ideas.
* [**Stanford Large Network Dataset Collection**](http://snap.stanford.edu/data/)- A variety of network data sets, including data from social networks, product reviews, online communities, etc.
* [**Yelp Data Set Challenge**](http://www.yelp.com/dataset_challenge/) - Reviews and check-in data on thousands of businesses.
* [**Million Song Dataset**](http://labrosa.ee.columbia.edu/millionsong/) - Freely-available collection of audio features and metadata for a million contemporary popular music tracks.
* [**Public Data sets on Amazon Web Services**](http://aws.amazon.com/publicdatasets/) - 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**](http://catalog.data.gov/dataset)- 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**](http://www.data.gov/opendatasites) with similar publicly available data sources from various cities, states, and countries.
* [**KDnugget's list of data sets for data mining**](http://www.kdnuggets.com/datasets/)
* [**Infochimps Data Market**](http://www.infochimps.com/datasets) - 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 and findings – 20%

* **Title**: Health Insurance and Hours Worked By Wives
* **Data description:** A cross-section from 1993, number of observations : 22272, observation : individuals, country : United States, description of the data in detail under the following subheadings:
  + The problem domain: A comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wives
  + The source of the data:

Olson, Craig A. (1998) “A comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wiwes”, Journal of Applied Econometrics, 13(5), september–october, 543–565.

* + The agencies working with the data

The JAE Data Archive, which is hosted by a server belonging to the Economics Department of Queen's University, by Olson, Craig A. (1998).

* + 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 project 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 – 0%

Follow the instructions presented in the link below divide the test into a training and testing set in the ration of (9:1).

<https://weka.wikispaces.com/How+do+I+divide+a+dataset+into+training+and+test+set%3F> The files generated as part of this process should be saved and submitted as the following

* trainingSet.arff and
* testingSet.arff

Screen shorts of these files should be included.

#### Data Mining Techniques

# Classification

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
   1. The model or a visualization of the model
   2. The results of the model
   3. Any additional output of the model including but not limited to
      1. Rules
      2. Confidence Values
      3. Confusion Matrixes
      4. etc
   4. Simple references to the notes or URL links to online resources complete with a sentence or two of explanation.

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

## 5. Classification: MLP or a similar advanced technique from Weka – 10%

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
   1. The clusters and/or a visualization of the clusters
   2. The results of the clusters
   3. Any additional output of the clustering process
   4. Simple references to the notes or URL links to online resources complete with a sentence or two of explanation.
   5. 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
   1. The regression equation
   2. Diagram of the historical values
   3. Diagram of the predictions

## 7. Time Series Forecasting – 10%

#### Research publication

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

Please discuss under the following headings

1. Publication and researchers
2. Dataset
3. Technique (mention any adaptions)
4. Major Findings
5. Relevance / potential relevance to your work

#### Presentation

## 9. Presentation – 10 %

If slide are used they may be submitted but all marks are going for the presentation quality. A clear division of the marks will be given

#### Division of Labor

## 10. Division of Labor

Please complete the sections below with regard to the estimate of the division of work between the two partners

### Summary of division of work

If the work was split in the range of 45% to 55% per partner then that is fine and simply say “Work was evenly divided”. If this was not the case then state with a summary sentence. This is the important statement of this file.

Division of work : *work was evenly divided*

### Percentage of work completed by each partner on each class / task

Some area requires more work than others so this is only for reference. An average of these values will not be calculated.

Please ensure that both students contribute to the research publication & presentation sections. The remaining topics can be divided among the students, but each student must apply at least 1 techniques.

|  |  |  |
| --- | --- | --- |
| **Filename / Task** | **Student Name 1** | **Student Name 2** |
| Selection of dataset |  |  |
| Cleaning of dataset |  |  |
| Task 1 etc….. |  |  |
| Paper selection |  |  |
| Paper Review | 40% | 60% |
| Presentation Preparation | 60% | 40% |
| Presentation Delivery | Etc. |  |