MULTIMEDIA UNIVERSITY
TM

TDS 3301 Group Assignment

---

1. The assignment is to be completed by a group of maximum **2** students.
2. The dataset: *KDD Cup 1999*. Detail description of each dataset can be found together with the dataset from MMLS.
3. This project is to be conducted fully within the WEKA environment. Java coding or external interfaces are optional.
4. In your report, you must report in details on the following items:

    a. **Individual Contribution.**
        1. Indicate clearly each group member's contributions to the project.
    b. **Introduction**.
        1. Write a clear introduction to your work.
        2. Search the related work that has used the similar or related dataset as what you are choosing. Example:

            XYZ (2006) used association rule mining for dataset x. The results indicated that association rule mining is suitable for dataset x. ABC (2007) implemented k-nearest neighbour clustering technique and the results ….
    c. **Formulating the *learning* problem**.
        1. Identify the *type* of problem described by the dataset. That is, the problem might be well described as a *classification problem*, a *clustering problem*, *frequent pattern mining* problem, etc.
    d. **Understand your data**
        1. Describe your dataset.
        2. Analyze your dataset using statistics, plots, histogram, etc.
    e. **Preprocessing**
        1. Maintain the attack classes as it is or convert them into 5 classes or convert them into only 2 classes: attack and non-attack.
        2. The process such as discretization, normalization, and the like may be included.
        3. Perform appropriate feature selection algorithm/method to identify the most appropriate set of features, etc.

f. **Algorithm choice based on the chosen learning problem**.
    1. Identify the appropriate algorithms for the dataset. You must report on at least THREE (3) data mining techniques.
    2. Ex: if you choose classification problem, you select let's say Decision Tree, Multilayer Perceptron and Bayesian Network. Ex: If you choose association rule, you select let's say Apriori, FP-tree, and Class-based association rule.

g. **Performance criterion**.

For example, if you choose classification problem, then answer:

1. How should the performance be measured? Error rate? Expected misclassification cost? Cross-validation Likelihood?
2. Overall performance vs performance for each class.
3. Which data mining technique produces highest accuracy?
4. Do the preprocessing / features selected influence the model's accuracy?
5. , etc.

If you choose association rule mining, then answer:

1. How many rule do you generate?
2. How do you evaluate the rules? Ex: clusters to classes evaluation.
3. Which algorithm give you better rules in terms of interpretation, usefulness?
4. Do the preprocessing / features selected influence the rule generation?
5. , etc.

If you choose clustering, then answer:

1. How do you determine the optimal number of clusters?
2. How should the performance be measured?
3. Which clustering algorithm gives you the best result?
4. Do the preprocessing / features selected influence the rule generation?
5. , etc.

5. **Grading:** The grading will be based on the report you provide and presentation (conducted during tutorial session).
6. **The report** should be submitted to your **tutor** in **hard copy** before the **due date (Friday, Week 12)**. You must use the cover template given. Sample format of report is given as well.
7. **Attach in your report a CD** that contains i) softcopy of your report ii) datasets at different stages of the data mining task.
8. Points will be assigned as follows:

Report (group-based)	**%**
Introduction	1
Formulating the problem	1
Understand your data	2
Preprocessing	2
Algorithm choice & Performance criterion	3
Conclusion	1
Presentation (individual-based)	**%**
Clarity in presentation	2
Q & A	3

Group Assignment

TDM 3301

(Data Mining)

Put Your Title here

Prepared by

Name, ID, Tutorial Section, email, contact no.
Name, ID, Tutorial Section, email, contact no.

TABLE OF CONTENTS

1.0 INTRODUCTION (FONT SIZE = 14)

Type the introduction of your manuscript in 12-point Times New Roman, single spaced. Do not bold the text and be sure that your text is fully justified. All paragraphs should be indented ½ inch. Please do not place any additional blank lines between paragraphs.

Introduction should describe the dataset that you use, preliminary examination on the data (how many attributes, if its qualitative or quantitative type, no missing values or outliers that can be identified etc.). It must also contain the objective that you want to achieve at the end of your mining process.

Tables and figures must be numbered separately. Table and figure captions should be 12-point boldface Times New Roman, aligned center. Initially capitalize only the first word of each figure caption and table title. Bold the header of each column. For table contents, use left alignment for texts and centre alignment for numbers. Set the vertical alignment of each cell to center.

TABLE 1. Title of the first table in boldface type

Column 1	**Column 2**	**Column 3**	
		Sub-Column	**Sub-Column**
Text	Text	Number	Number
Text	Text	Number	Number
	Text	Number	Number

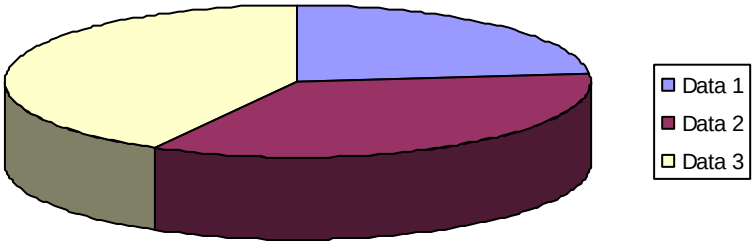

FIGURE 1. Title of the first figure in boldface type

…

8.0 REFERENCES

This section is compulsory as well. Don't forget to cite your dataset source (you can get it from you dataset desc file, and any paper or book that you use as a reference.
End references should be listed in alphabetical order. Use 12-point Times New Roman, fully-justified. Indent the subsequent line(s) ½ inch from the left.

Alahakoo, D., Halgamuge, S. K. and Srinivasan, B. 2002. Dynamic self-organizing maps with controlled growth for knowledge discovery. *Journal of IEEE Transactions on Neural Networks* 11(3): 601-604.

Mallach, E. G. 2002. *Decision support and data warehouse system*. Boston: McGrawhill.