BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

II SEMESTER 2016-2017 CS C415/IS C415 – DATA MINING

- This assignment will be done in a group of exactly two students.
- The weightage of the assignment is 15%.
- You are required to submit a detailed written report on the due date 20/11/16. Write your report in your own words.
- The last date for taking approval for the dataset is October 18th. In case of the same dataset identified by two groups, the dataset will be assigned to a group by FCFS basis.
- For approval, submit details of the datasets at https://goo.gl/forms/PS2Lf427PynGCZv52
- Regarding any queries, you may contact TAs (WiSoc Lab 6012) or IC.
- You are required to do the following tasks:

Datasets

You will be using two datasets:

1. The **census- income dataset** from the US Census Bureau
   This dataset contains census information for 48,842 people. It has 14 attributes.
2. **Another dataset**, which you can choose depending on your own insterests. It should contain enough instances (at least 400 instances) and several attributes (at least 10). Ideally it should contain a good mix of numeric and nominal attributes. These datasets can be chosen from the following links
   UCI KDD ML Data Repository
   Datasets for Data Mining
   CMU's StatLib-Datasets Archive

You cannot choose any dataset included in the WEKA/IBM SPSS system.

Understand the datasets

Understand the datasets fully. Describe the dataset in terms of the attributes present in the data, the number of instances, missing values, and other **relevant** characteristics in your report and presentation.

Objectives of Data Mining Experiments

Figure out 3 to 5 specific, interesting questions **about the domain** that you want to answer with the data mining tasks and techniques. Try to choose questions that are about the domain not about the techniques or the experimental parameters. Use one or more data mining tasks (classification, association rules, clustering or anomaly detection) to answer a question. Try to apply different tasks to solve different questions to get the best possible solutions.

Performance Metrics

Explain what performance metric(s) will be used and why to evaluate the models that you have constructed in order to answer questions identified (e.g., accuracy, error rate, ...).

Preprocessing of the data

You should apply relevant preprocessing techniques to your dataset as needed before doing the mining. For instance, you may decide to remove apparently

irrelevant attributes, replace missing values if any, discretize attributes in a different way, transformation etc. Your report should contain a detailed description of the preprocessing of your dataset with **justifications** of all the steps. You can write filters/programs (you may incorporate them in the tool used) on your own if the tool does not provide the functionality that you need to include. You may use a separate tool as well.

Training and Testing Instances

Decide testing and training datasets for classification task. You can experiment with different approaches to decide testing and training datasets and explain the results affected by these experiments in your report.

Parameters and Settings

Describe what parameter values and other settings you used and why? Vary parameters and see the effect of these on the results/models. Also find the best settings of the parameters to answer the objectives.

Algorithms

Identify algorithms which you should apply to achieve the goals. You should **apply 2-3 algorithms** for each of the objectives. Compare algorithms and results for each objective and describe these in your report. Read the code that implements these techniques in your tool. In the report, describe the algorithms underlying the code in your own words. Explain these algorithms in terms of input they receive and output they produce and the main steps they follow to construct the model. Make sure to **describe the correspondence** between the step of the algorithm and the part of the code that implements it.

Resulting model

Describe the resulting model (e.g., size of the model, readability, accuracy…). Summarize the model in your own words focusing on the most interesting/relevant patterns. Elaborate on if and how the model answers the objectives identified.

Performance of the resulting model

- State the performance of the model. For example, elaborate on the confusion matrix and/or other relevant performance indicators.
- Compare the performance of the model with that of other models constructed.

Results

- What is/are the model(s) with the highest performance for each objective?
- Discuss how well a data mining technique worked on this dataset. What combination of parameters yielded particularly good results?
- Overall project conclusion: strength & weaknesses