CS 5402 – Intro to Data Mining
Fall 2019
HW #8

- This assignment is **due in class on Monday, Dec. 2, 2019**.
- This assignment is worth **200 points**.
- You are **REQUIRED to work as part of a team of 2-3 people**, each of whom must be a person enrolled in this course. That includes distance ed students! If you want to **make a change to your HW #7 team**, contact Dr. Leopold (leopoldj@mst.edu) as soon as possible; otherwise, we will assume that you will maintain the same team for this assignment.

Project Description

For this assignment you are to use techniques discussed throughout this semester to **analyze** the dataset that is posted on Canvas along with this assignment; we would prefer that you **ONLY** use techniques/methods that were discussed in this class. You will **NOT** be given **specific instructions** about what procedures to perform on the dataset; your grade on this assignment will be based on your determination of **what data mining methods should be done and what information can be deduced from the results of those methods.**

You can use any software applications and/or programming languages to do the analysis. For every method that you do, you must **provide brief technical documentation** explaining **how** you did it. For example, if you used a particular method in Weka, include a screenshot of the Explorer/KnowledgeFlow program that you set up to perform that method (including specification of parameter settings). If you wrote a program to perform some method, include the source code for that program. Do **NOT** submit lengthy/verbose explanations of what you did – we won't read (or grade) more than the first 3 sentences you write! If you do not provide **concise, precise documentation for each method** you used, you will not receive full credit!

The first **1-5 pages** of your homework submission should simply be a **table** containing about 3 columns: (1) a column (briefly) saying **what** you did, (2) a column saying **how** you did it, and (3) a column (briefly) saying what were the **main results** from performing that task. You also could include a fourth column that gives a reference to another page in your report where the source code, Explorer/KnowledgeFlow screenshot, etc. can be found that provides additional (more verbose) documentation for the method used.

Your homework submission must **also include a summary** of (what you consider to be) the most important/interesting results from your analysis. You should take into consideration the accuracy of the methods you tested as well as the novelty (balanced with the practicality) of the results your data mining methods produced. In addressing the latter, consider what you know about this dataset (i.e., it's a secure water treatment plant that is being researched for cyber security breaches). **This summary can be no more than 2 pages long**; after that, we will stop reading – it's not that we're not interested; we've just got a lot of reports to grade!

What To Submit for Grading

You should submit a **paper** copy of your report in class the day it is due. Do **NOT** turn in a printed copy of the data file! **The grader reserves the right to contact you and ask to see anything you did for this project.** It is your responsibility to have your **source code and data files** available to show him upon demand (i.e., if he contacts you, you can't say "the system ate my files and I don't have them anymore"); if you don't have the files to show him when he asks for them, you will get a zero on this assignment! You should have every member of your team make backups of everything!

You are **also required to submit ONLINE (via Canvas) a survey/evaluation of your team members**. It will ask what tasks you and each member of your team did on this assignment, and what percentage of credit (e.g., full vs. partial) you think each team member deserves. This will be taken into consideration when determining your grade on this assignment. If the tasks/credit that you claim for yourself vastly differ from what your team members state for you, then a meeting will be held with your instructor (Dr. Leopold) and possibly the Computer Science Department Chair (Dr. McMillin) to determine if academic dishonesty has taken place; so be honest! **If you do not complete the online survey/evaluation, you will receive zero on this assignment, even if you worked on the project.** The survey will be posted on Canvas as an online "quiz" called **HW #8 Evaluation**.