CS 5402 – Intro to Data Mining
Fall 2020
HW #7

- This assignment is **due by 11:59 p.m. on Friday, Dec. 11, 2020**.
- This assignment is worth **145 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.

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. The decision attribute in the dataset is named *class*.

First, you must **clean/preprocess** the dataset. This would include dealing with: missing values, inconsistent data, noise, redundant attributes, non-independent attributes, binning, grouping values, normalization, aggregation, etc.

Next, you must **analyze** the data. This includes using both supervised and unsupervised methods to uncover meaningful information. Listed below are the methods you are **required** to use:

- **Decision tree** (ID3/J48 or CART) **with conditions dropped** using the predicted error rate method
- **Association rules with conditions dropped** using the predicted error rate method
- **Clustering** (k-Means, DBSCAN, or EM, etc.)
- **Support Vector Machine**
- **Bayesian Network (Simple)**
- **Gradient Boosting Method** (with some reasonable attempt made to find optimal parameters)
- **Stacking** (you must use at least 3 different methods, you can't make it primarily different kinds of decision trees, and you can't use decision stump at all)
- **One other supervised method**, chosen from: **Linear Regression**, **Prism, kd-Tree (KNN), Artificial Neural Network, (Non-Simple) Bayesian Network**, or something we didn't cover in lecture

You are expected to use **k-fold cross-validation** when testing the methods.

Finally, you must **write a short summary** about your findings. This should be **no more than 1 page** (including figures). It should explain which method produced the **best model for predicting the class** for the dataset, and what was used as the basis for your decision (e.g., accuracy, ROC, correlation coefficient, TP/FP/TN/FN rates, examination of the confusion matrix to see where errors are being made, etc.). It should also state the most interesting findings yielded by the unsupervised methods like clustering and association rule mining (i.e., **interesting relationships** between attributes, not necessarily the decision attribute). **Note: If you do not include a meaningful summary, you will not get ANY credit for this assignment!** There's no point in running the methods if you do not try to derive any meaning from the results.

Special Note

You are expected to use what you have learned in this course to make decisions about what to do on this assignment. There may not be a single correct way to do something; they may be several viable alternatives. Do **NOT** ask your instructor how to do things; for example, do **NOT** ask questions like the following:

- What should I do about the missing values for a particular attribute?
- Which attributes should I compare when clustering or doing association rule mining?
- Should I bin/normalize/aggregate/group/… the values for a particular attribute?
- Do I need to keep this attribute?
- What does this attribute represent?
- Have I done enough for …?

These are things that YOU must resolve within your group!

What To Submit for Grading

You should submit a single **pdf** file containing your summary on the **first** page, followed by documentation of **everything** you did to clean/preprocess and analyze the dataset. The latter could be Weka screenshots, Python source code, written description of things you did in Excel, etc. The key is **documentation**! If you do not provide clear, sufficient documentation for a task, you will **NOT** get credit for having done it!

We reserve 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 file** available to show us upon demand (i.e., if we contact 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 us when we ask 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. **Note: No one in the team can only do the cleaning/preprocessing; everyone must be involved in some of the mining!** 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 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 #7 Evaluation**.