William Seery and Nahom Ogbazghi

5/3/17

Final Project Project Report

https://www.drivendata.org/competitions/2/warm-up-predict-blood-donations/ = challenge page

https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center = dataset description

The data for this competition was gathered from random participants in a blood donation site in Taiwan. There are various attributes related to gathering blood, as well as a binary variable for classification for whether or not the test subject donated blood in March 2007. The goal is to use the training dataset to try and classify the subjects in the test dataset to see if they will donate blood in the future. The challenge asks for the probability that they will donate, but we will be just giving the most likely outcome, donate or not donate. Our goal is to help predict the amount of donors they will get. This allows them to more efficiently allocate their resources for having the right amount of capacity. It also allows them to test the successfulness of promotional campaigns more accurately by monitoring changes in the amount of donors compared to what they previously predicted. These are very important for the overall quality of operations and operational efficiency.

This is a classification problem, so we have been exposed to various classification methods in this class. Specifically, we have used the naïve Bayesian and c4.5 classification methods in our classification homework. We also considered using k nearest neighbors, but decided that a decision tree was the best way to go in the end.

For this project, we modified our code for our classification homework. My homework had decent accuracy, but the way I implemented it was specific to the dataset we were given, so we modified it to be modified to work with this dataset. We tried to change the string used to store the values into an array and change the classification variable values. The main issue we had was making the data set usable with our code and giving a relevant result. Since the data was very fragmented, the values for the attributes rarely matched up exactly, since they were numeric values, we decided to separate the data values into 10 ranges. We found the range of values for each attribute and then split them into 10 values. Then we changed the value of each data entry to the closest range value above it. Since we were not given any test data, we used random to select a random value for each attribute from the 10 range values we generated for each attribute. We made 200 random data entries for our test data.

There is only a training dataset provided for the challenge for blood donations. We wanted to use the classification accuracy to evaluate our results, but we weren't given a test set. Instead, we made our own test set and classified it with our code. We tried to use Weka to compare to our results, but we could not get weka to accept the test data we made. Also, when we print out the results, it does not print all the classifications of our results.

The work was be completed together. So there will be pair programming, and where one had to work alone, the other compensated by progressing on the report. Some of the things we learned were how to deal with data that is fragmented. Initially, we thought we might have to use k nearest neighbors because none of the data values matched up. For C4.5, you need the data values to match up to some extent to be able to have a relevant result from the tree. If there is

only one instance of each data value, that does not give a definitive result. If we had more time, we could implement more of the classification algorithms.