The original training set values, training set labels and testing set values are all .csv files downloaded from <https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/data/>

Note: all of the input datasets were written as file paths referenced to group members’ own computer. If you want to run the code, please change the paths.

As we used multiple datasets and multiple classification methods, we will detail how everything can be done.

The subfolder included contains only the files needed to run a decision tree on a dataset cleaned by our first method.

(NOTE: While, we have provided both the cleaning methods and the outputted data, a bug stemming from a package mismatch makes the cleaning code annoying to run (The output, when it works, is valid. But it can take 10+ tries for the code to run.)

1.data cleaning and preprocessing:

The data was cleaned by separate three methods, also outlined in the report. The first method replaces missing values with the attribute mean/mode. The second replaces missing values with an ‘unknown’. The third drops rows with missing values. All three methods drop 5 attributes, and convert all categorical values to numerical values.

Files:

Cleaning.py

Cleaning2.py

Cleaning3.py

To run Cleaning.py or Cleaning2.py, code is in python and requires the numpy and scikit packages (I used an anaconda environment).

The code reads in two files, training data and test data, from its current directory with the names TrainingSetValues.csv and TestSetValues.csv (these contain the data as downloadable from the competition site.)

run code with command:

python Cleaning.py

The output will be written to cleaned\_train\_values.csv and cleaned\_test\_values.csv or cleaned\_train\_values\_2.csv and cleaned\_test\_values\_2.csv

Cleaning3.py is different in that it needs the training data to have the labels attached (train\_data\_with\_labels.csv). The output’s last column will be the labels, and should be removed to its own file.

An extra corrective step in our process, removing attribute names and altering a date attribute. This is run on any given pair of outputs from one of the cleaning methods

Files: timeconvert.py

python timeconvert.py

outputs a column of numbers that give the count of days from the 2010/12/31

2. cleaned data files (different versions)

cleaned\_train\_values.csvcleaned\_train\_values2.csv,cleaned\_train\_values3.csv,

(Test files are named similarly)

(for labels: label\_classify.csv is used for all classifiers unless dataset3 is being used, in which case labels3.csv is used)

3. Classification

Naive Bayes: naive\_baye.py

Single Decision Tree: decisionTree.py

Adaboost: boost\_final.py

1 random forest tree: 1randomForest.py

10 random forest tree: 10ranodmForest.py

100 random forest tree including voting scheme: 100\_rf.py

All of the above classification methods can be run by command

python xxx.py

We made use of the sklearn package in these methods.