Slippery Rock University

Naive Bayes Classifier Backend and Frontend Technical Manual

Trevor Hamilton (tjh1003) & Jonathan Stonebreaker(jds1018)

Data Mining - CPSC 405

Dr. Sam Thangiah

Table of Contents

**File Breakdown**

3............................................................................................Expected Input Data File Formating

5................................................................................................................Output Data File Format

**GUI**

8..................................................................................................................Input File Location Tab

9.................................................................................................................................Metadata Tab

9...............................................................................................................Output File Location Tab

**CLI**

11.....................................................................................Changing Input and Output File Locations

11................................................................................................................................Console Output

File Breakdown

Section 1: Expected Input Data File Formating

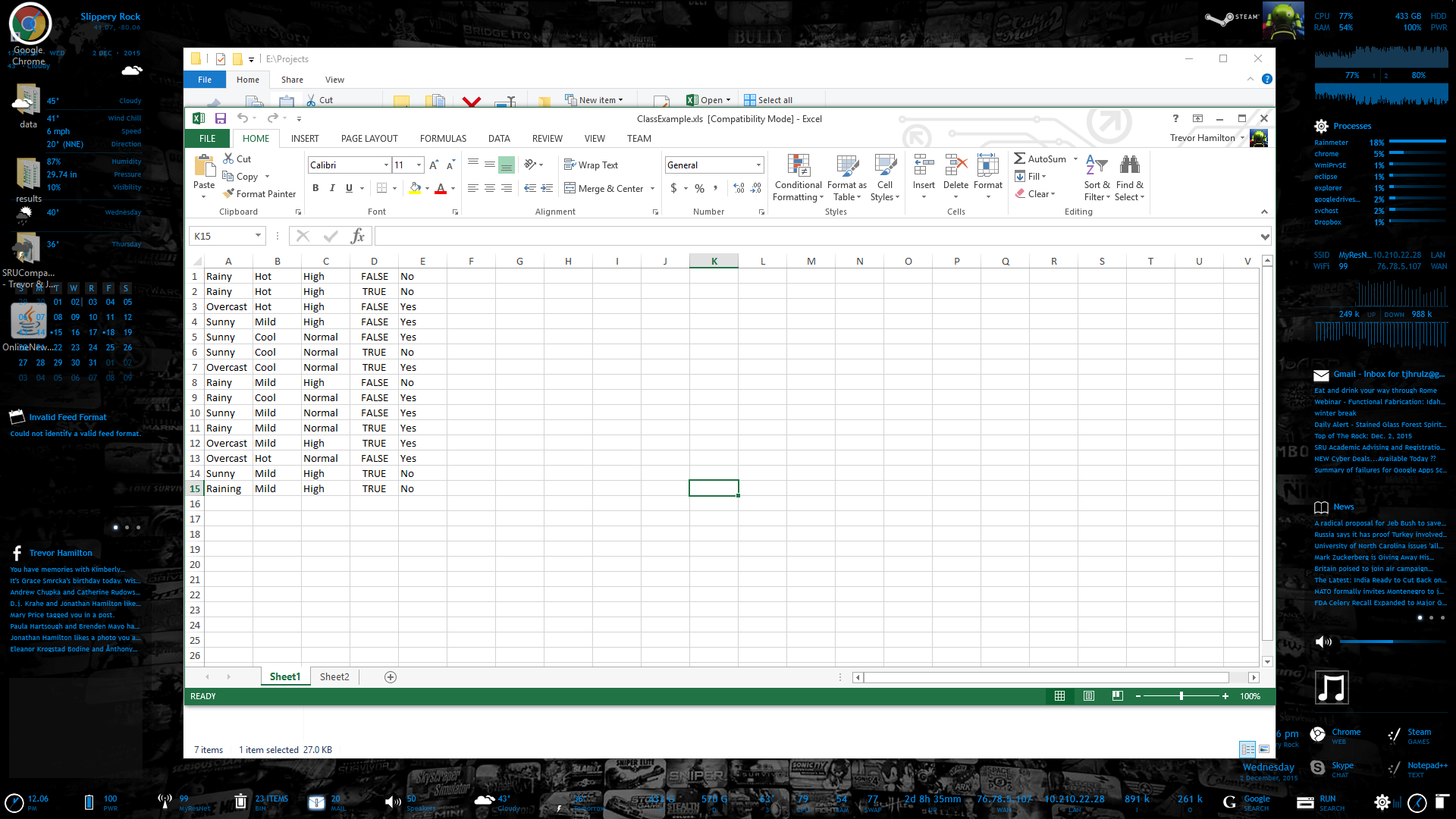
Sheet 1 Breakdown

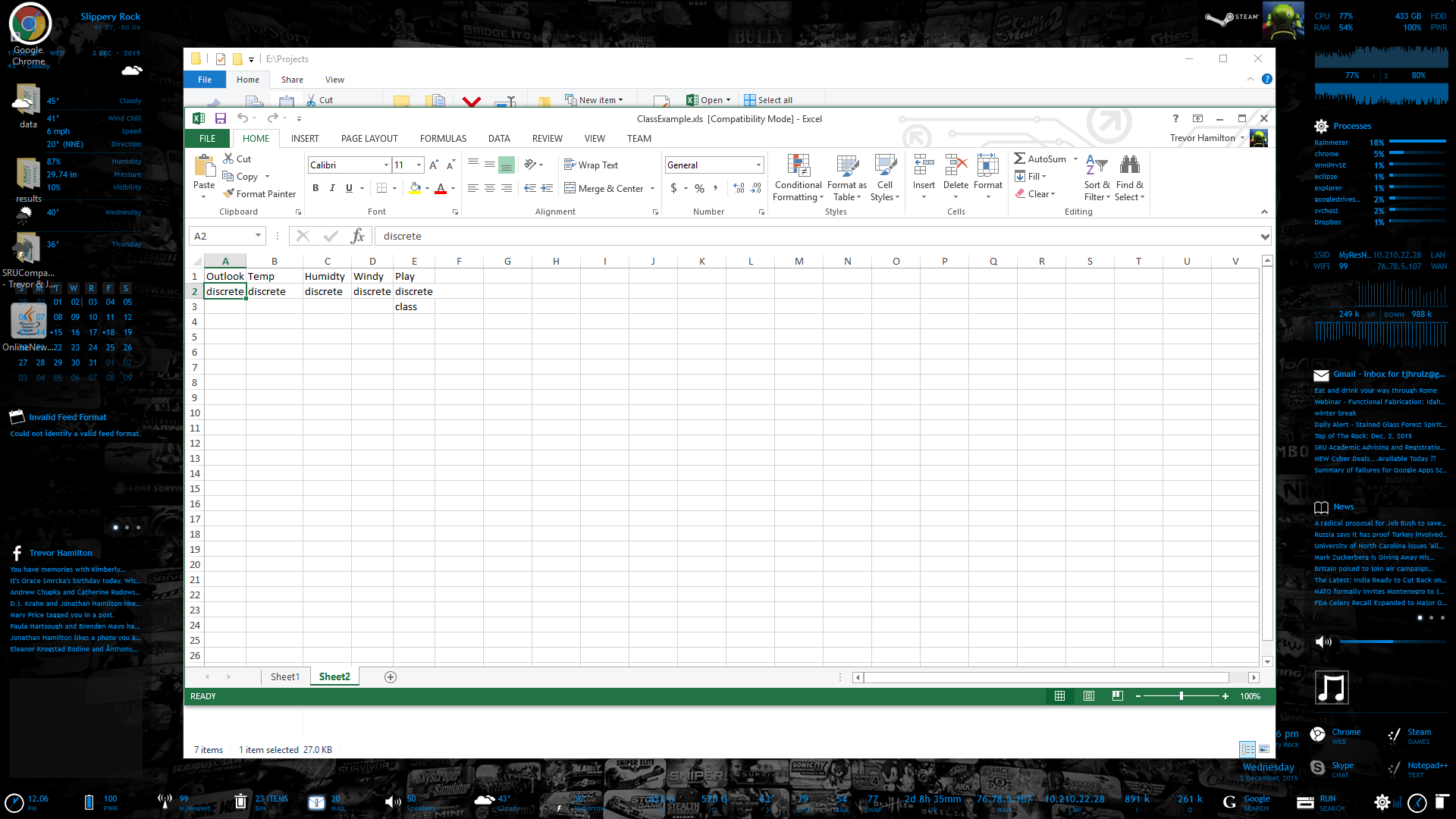
* Contains a list of all the data going to be used
  + Each column is an attribute
  + Each row is one data point

Sheet 2 Breakdown

* Must be listed in the same attribute order as the first sheet
  + Row 1
    - Contains a label for each attribute that you would like it the be referred to
  + Row 2
    - Contains if the data is discrete or continuous, unused but must contain a value
  + Row 3
    - Contains the word classify if this row is the classifier, can be any value but all other cells must be blank

Example expected Input file:





Section 2: Output Data File Format

Sheet 1 Breakdown

* Contains the testing data in the same format as input file
* Contains the label of each attribute along the top
* Contains the guessed classification on the right end
* Contains the percent right at the bottom right
  + Actual classification in yellow
  + Guessed classification in green if it matched the actual red if it didn't
  + Percent is green if above 90% else it is red

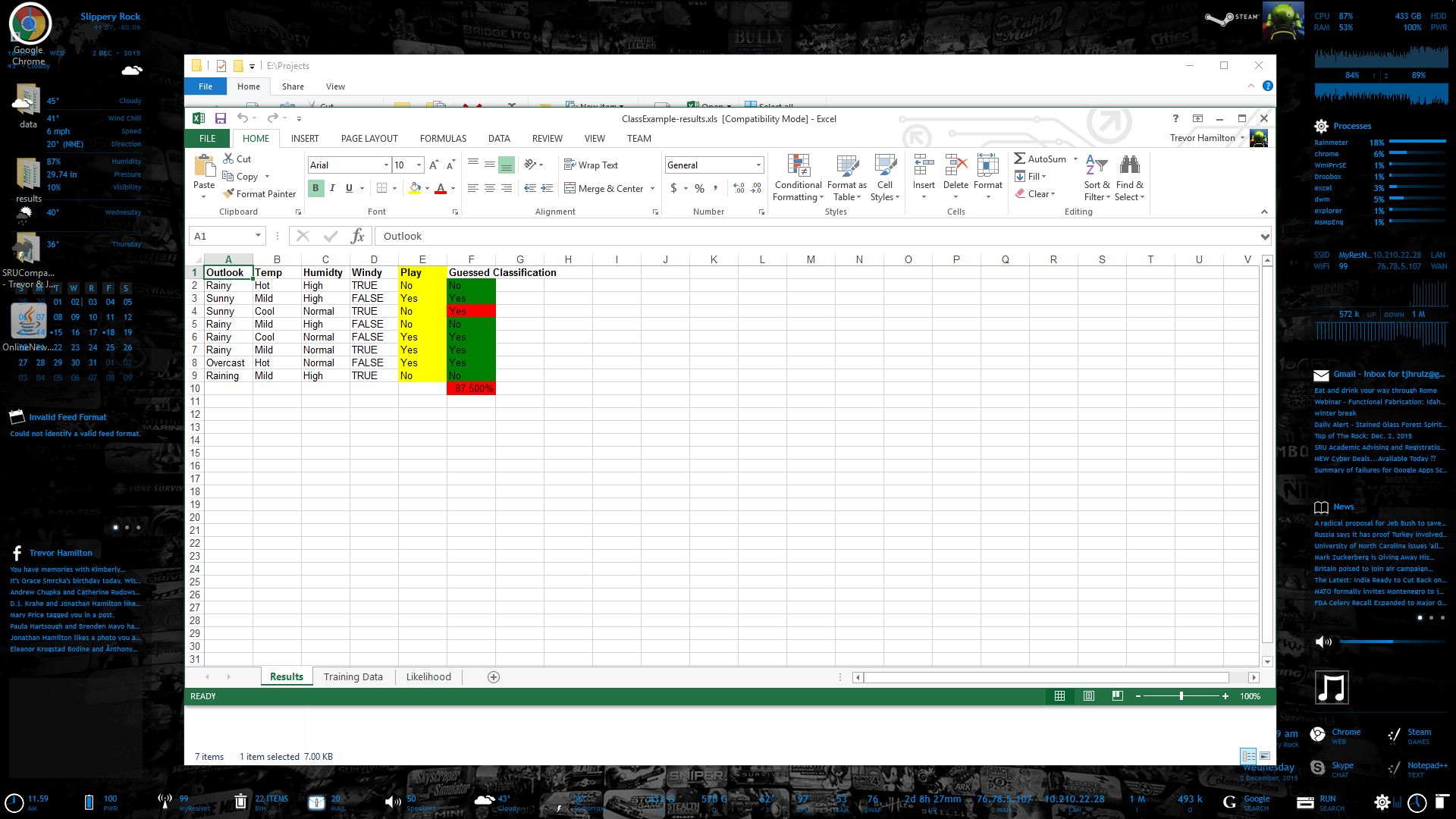
Sheet 2 Breakdown

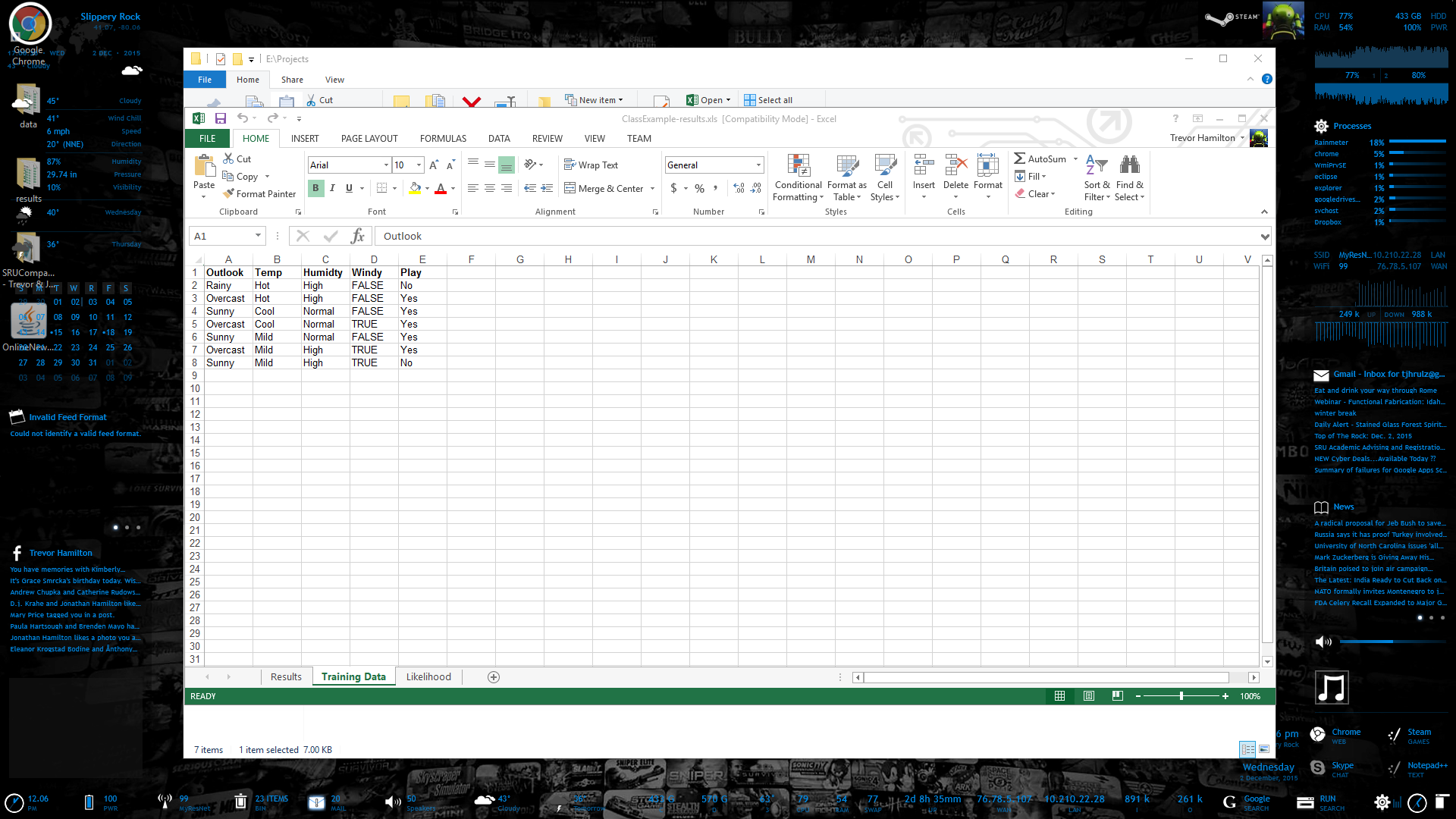
* Contains the training data in the same format as input file
* Contains the label of each attribute along the top

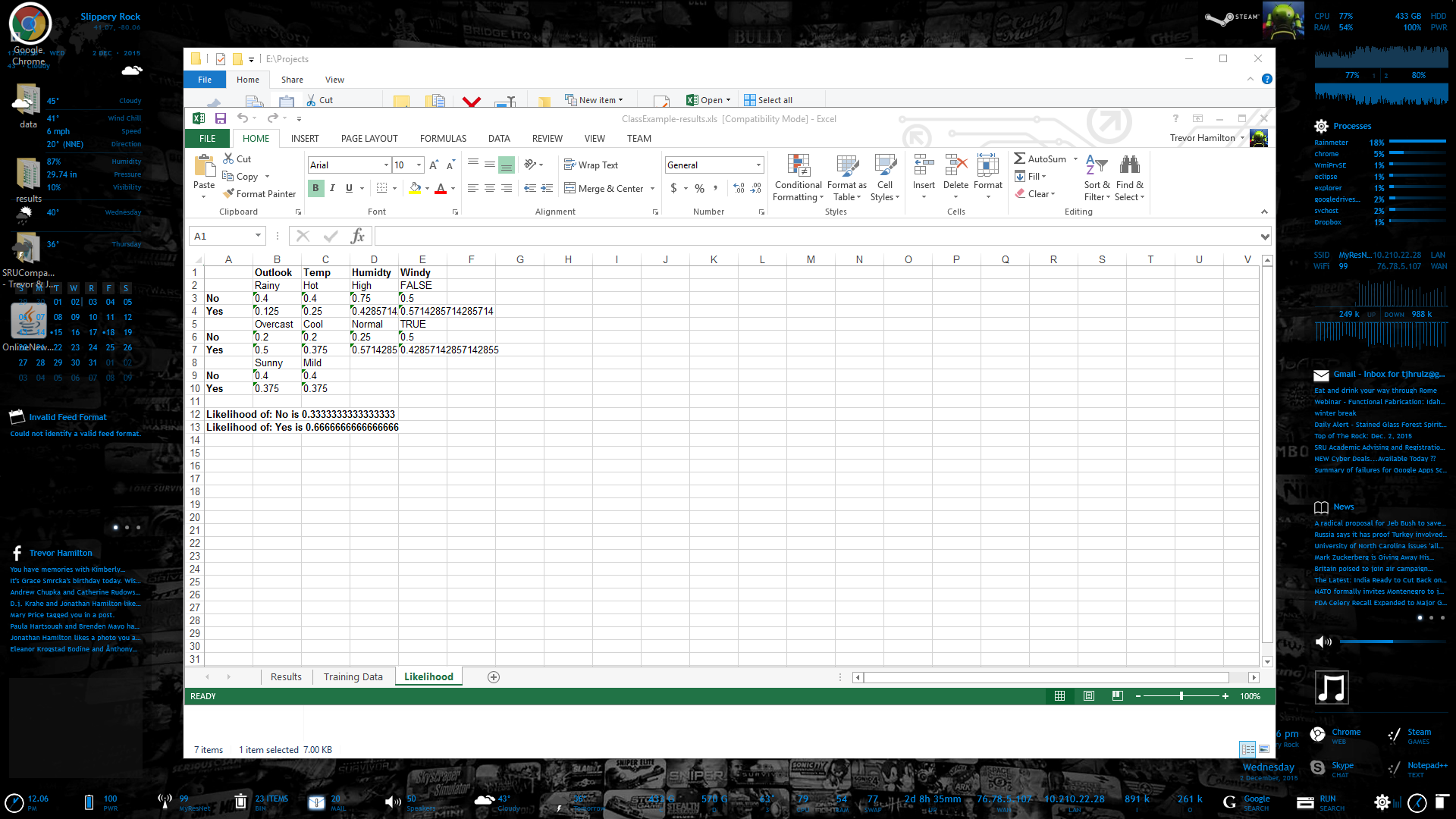
Sheet 3 Breakdown

* Contains the likelihood of each distinct occurrence in the training data for all known classifications
* Contains label of each attribute along the top
* Contains each distinct possible classification along the left
* Contains the likelihood of each possible classification on the bottom

Example Output file:



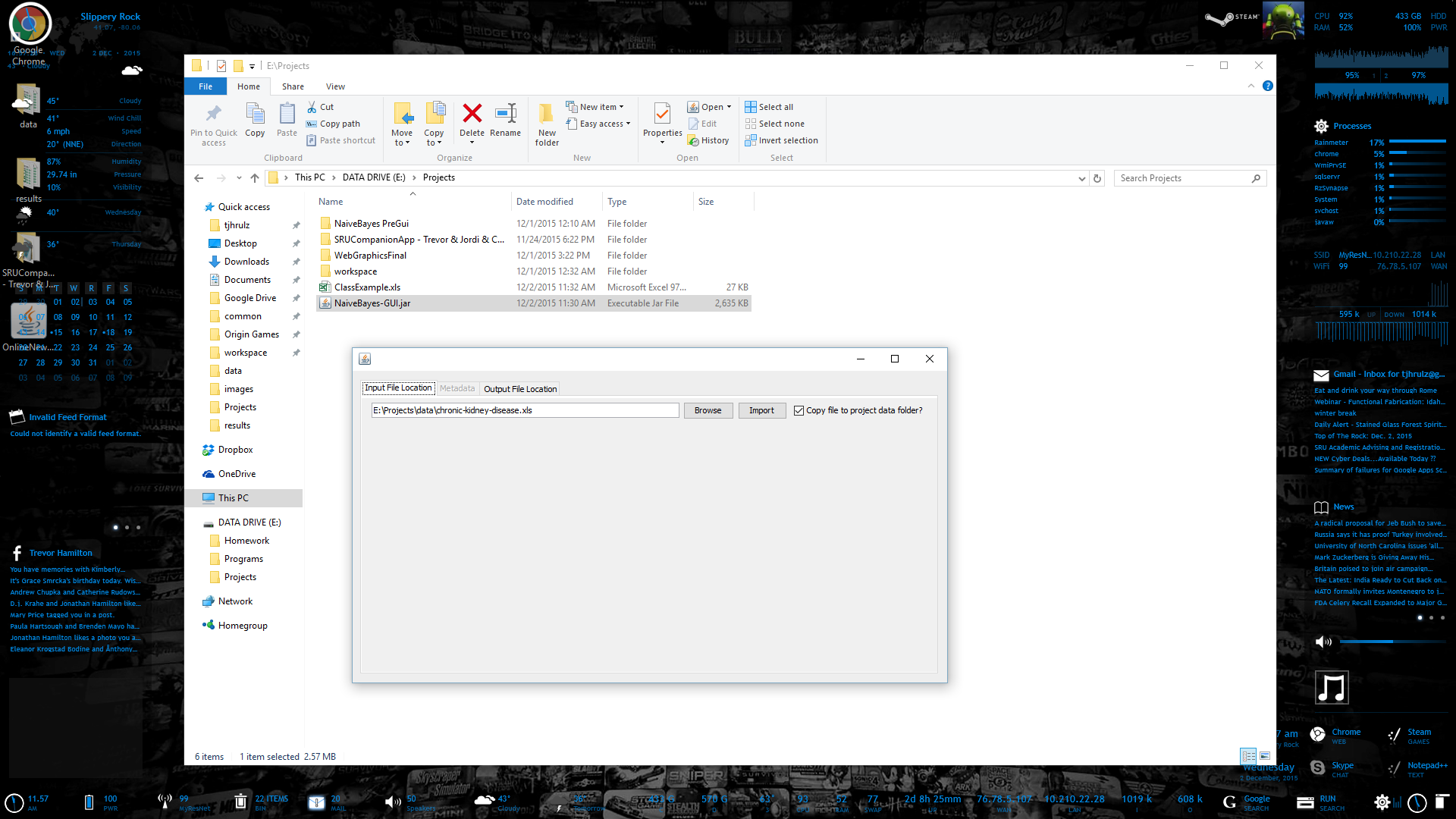




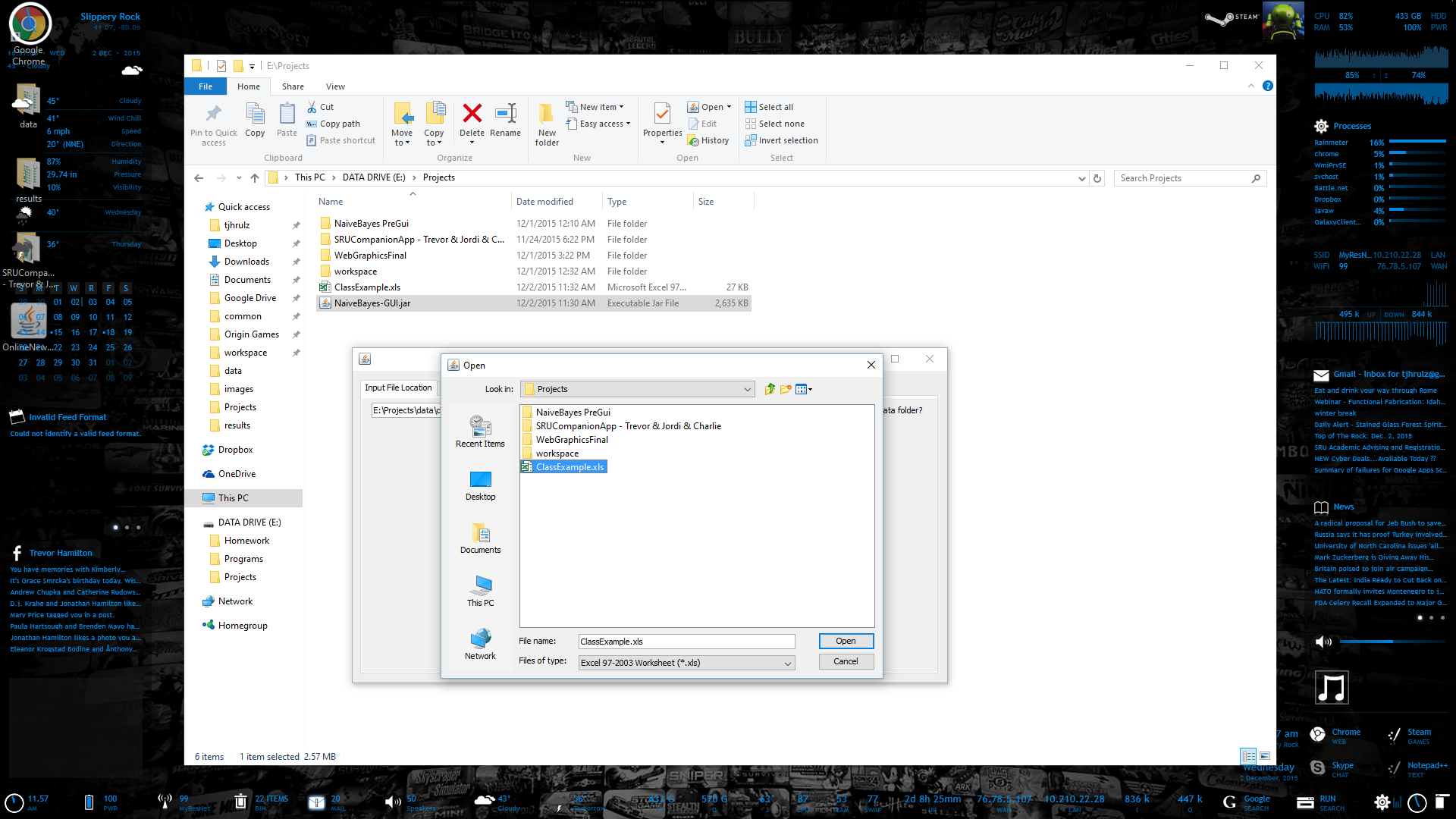
GUI

Section 3: Input File Location Tab

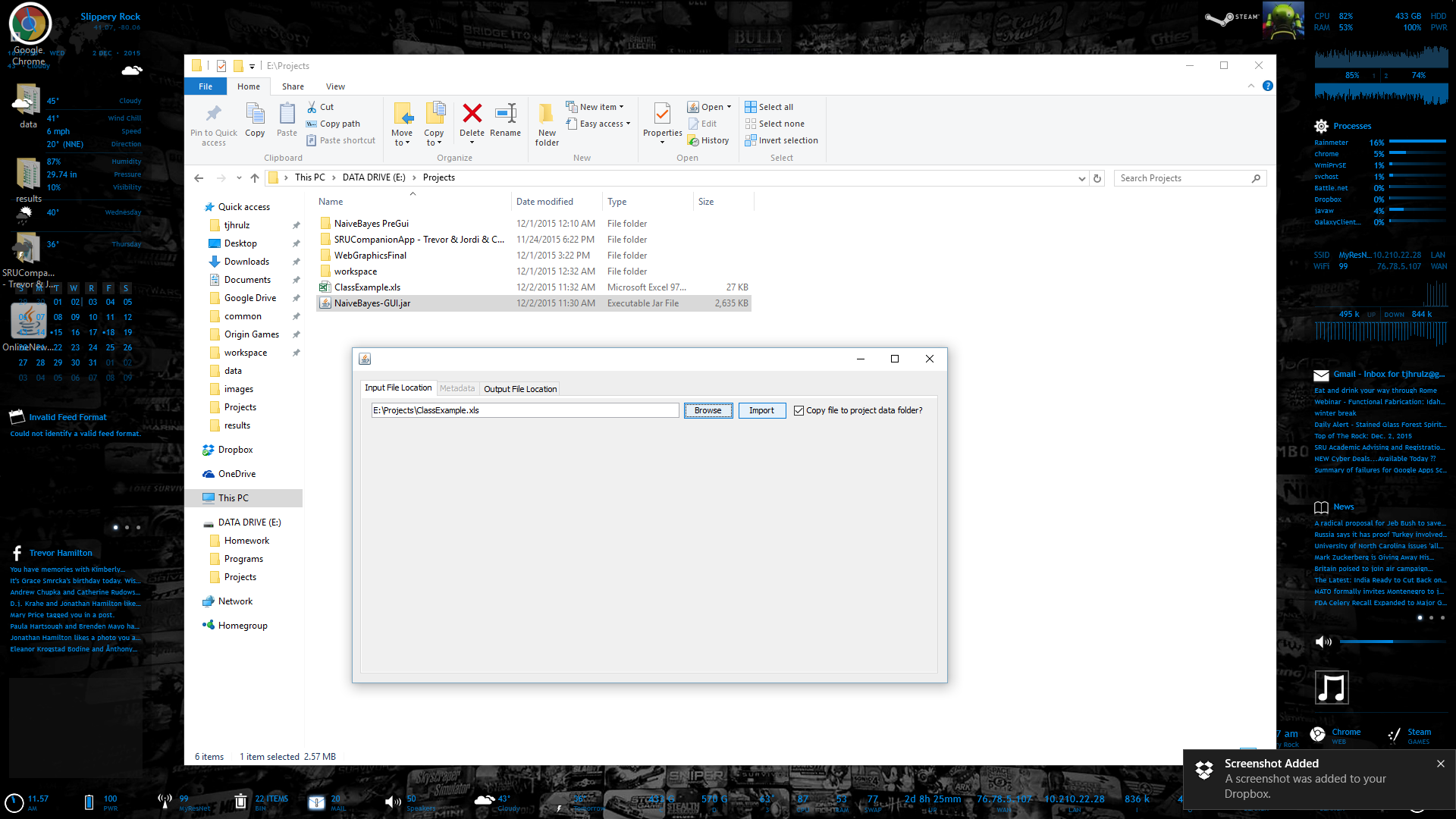
Breakdown



* Text box is the populated by default by the default data set and assumes it is in a data folder
  + If it does not exist hitting browse opens to the default directory on your computer
  + If that directory does exist it opens to that directory and if the file exists in there it opens to that file



* Hitting browse results in your native file browser
* When you select a file that exists and is a .xls it changes the textbox to that directory
  + If you were to hit browse again it would open to that selected files directory



* Copy file copies the selected file to the project's data directory if it exists when you go to import the file

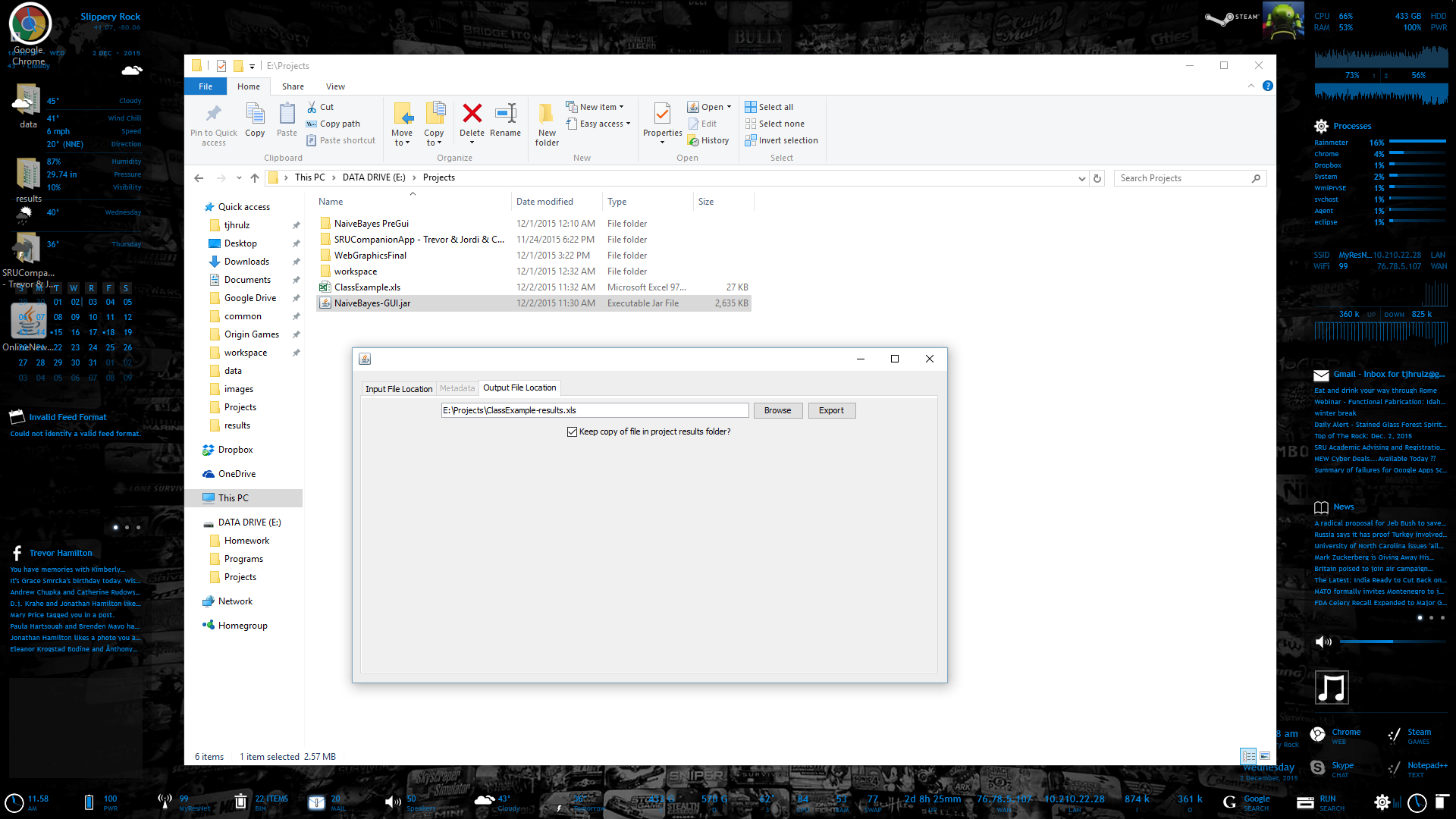
Section 4: Metadata Tab

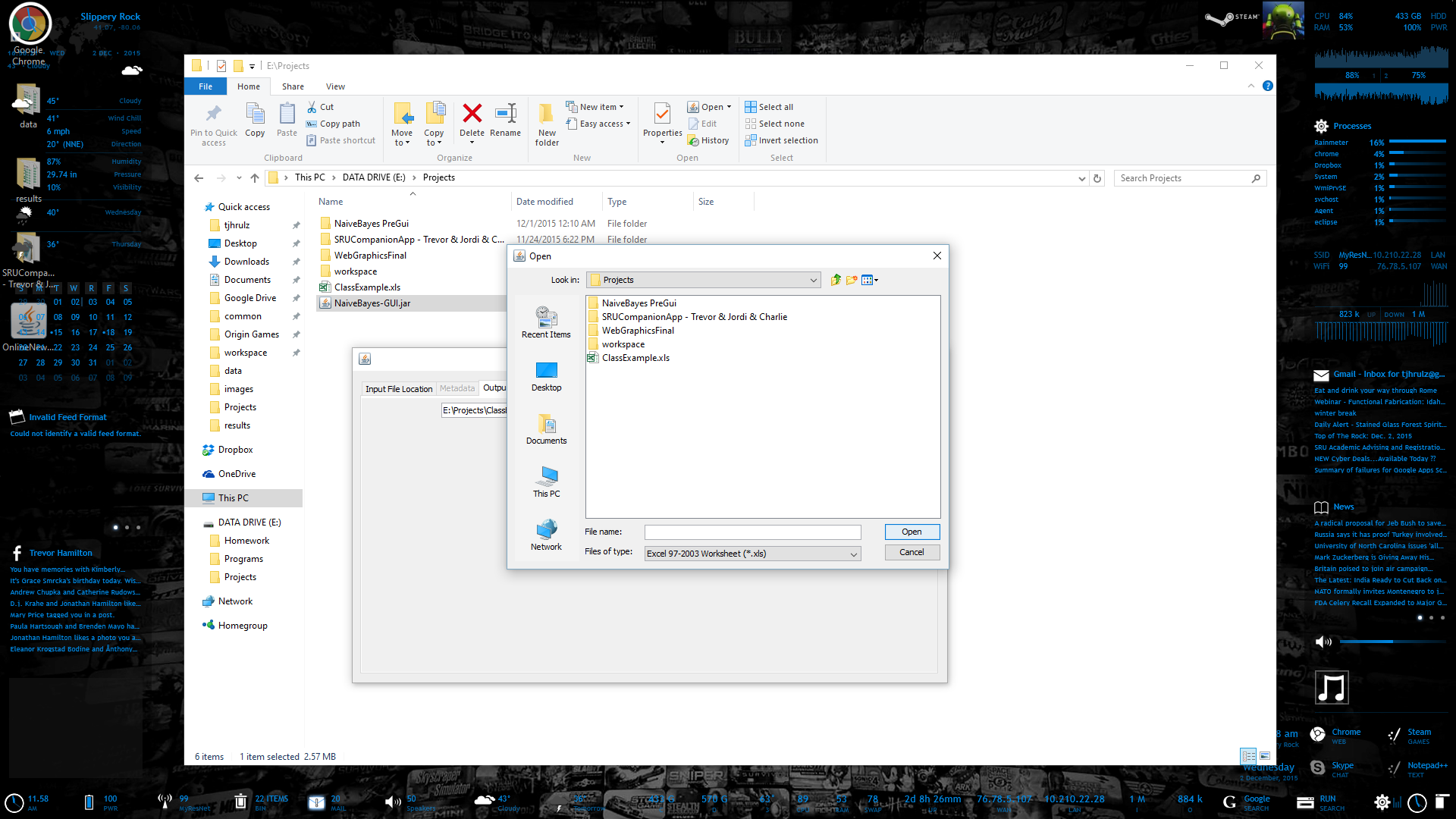
Breakdown

* Was disabled due to not being implemented was intended to duplicate sheet 2 of the input data and allow user to modify data before building classier and classifying data which it would have a button for. Button action moved to output button and the rest were unimplemented

Section 5: Output File Location Tab

Breakdown



* Text box takes your same file and location that you imported and adds -results to the end.
  + If you want to overwrite a different existing fill you can hit browse to choose a different file to overwrite (You can overwrite your original data but this is not recommended)
  + File browse should be the same as before

CLI

Section 6: Changing Input and Output File Locations

Breakdown

* Change intputFileName (Found at the first line in the main class which is the last defined function) to a full path to the data file or a relative path from the naiveBayes root directory. If you move your data into the data folder contained in the project just change everything after .data/ to your file.
  + Note: only accepts .xls (97-2005) excel files
* Change outputFileName (Found at the second line in the main class which is the last defined function) to a full path to where you want to file to be outputed and the name for the file. If you want to have it output to the contained results folder just change the name of the file after ./results.
  + Note do not output to input file as it will overwrite it also overwrites any file with the same name in that directory
  + Note if you do not put .xls on the end it does not add it for you but the file will still be a .xls file

Section 7: Console Output

Output Breakdown

* Breakdown of the output and function order of the main class with descriptions of what nonprinting function class do in the next section
  + Strings for the input and output excel files to be used
  + Do the import of the input excel file
  + Generate the training data
  + Notify use training data has been generated and print both the training and testing data
  + Generate the model (Likelihood of all the possible values of each attribute)
  + Inform use model has been generated and print the possible classifications and the model
  + Classify everything in the testing data
  + Inform user the guessed classifications have been generated and print of the guessed and acutal classifications
  + Export results to output excel file
* Output symbols mean the following
  + ::
    - Label was printed and this group of data listed belongs to it
  + :
    - Next data point for this grouping
  + ,
    - New group of data
  + [
    - Start of new dimension of data
  + ]
    - end of dimension of data