**CS612 - Lab 4**

**Question 1: Do the following questions**

1. Give at least two advantages to working with data stored in text files instead of in a binary format.

A)Text can easily be understood after opening. but we cannot understand the content of binary format

B) Since content of text file is understandable for us, so we can manipulate the content of file which is not easy for binary file

c) Text files are easy and convenient to read and open , we can open text file with anything but binary files should be opened by the application that created(coded) them

1. Can you think of a situation in which identification numbers would be useful for prediction? One example is that Student IDs are a good predictor of graduation date.

Yes, another example is credit card that is a good predictor of bank that we have that credit card.

I was heard that social security number is also a predictor of State of US that we are from.

1. An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each.
   1. What type of attributes would you have

the type of attributes would be Binary. Because Association rules are "if-then" statements that is either True or False (1 or 0)

* 1. How many of them are there?

4\*100=100

**Question 2: Do the following**

* Watch the video to find out how you an extract data from the Binding Database
* Go to the virtual machine in the cougarapps (cougarapps.csusm.edu)
* Go to the binding database <https://www.bindingdb.org/bind/index.jsp>
* Search for “gamma”
* Select “gamma-secretase”. This represents data for Alzheimer disease
* Add all the pages and make a .sdf and .tsv data file as shown in the video
* Go to .tsv data file and delete all columns except the IC50 column
* Go back to where you picked the .tsv and this time take sdf content
* Go to the apps and search for Dragon-7 and use the sdf file to calculate the descriptors. Save them into a file called Xes
* Create an excel file, drag the Xes into that excel file, delete the first column (it is index) and replace the first column with the IC50 value from the Y columns
* Manual cleaning:
  + Remove the rows that have no IC50 value . To do this you can sort the data based on IC50 and the ones with no value appear on the top
  + Remove the rows with no name. This time sort based on the name and the ones with no value appear on the top and you can remove them
* Email the data file to yourself after doing manual cleaning
* Write an ***object-oriented*** program in python that does the following:
  + get rid of the columns that have more than 200 junks. Otherwise their junks should be changed to zero
  + get ride of the columns that have all zero
* Save this file as reference file. Make a copy. Get rid of the name column in the copy file and the first row that is the names of the properties
* Do rescaling of the copy file

Your program should print the following:

* Total number of descriptors (columns) that have been deleted
* Total number of descriptors (columns) that sum to zero and have been deleted
* A file that is called “RescaledFile.xlxs” that includes all the data that are re rescaled.

**What to submit:**

* **A file that includes the result of question1. Call it Lab4-Q1**
* **The myGamma.xlxs file**
* **The “RescaledFile.xlxs” file**
* **Your code. Call it Lab04-Code.py**
* **The snapshot of your output of your program. Call it Lab04-Output.txt**
* **Place all the files in a folder and call the folder based on your name and your partner’s name (ex:Jack-and-Nancy-Lab4)**
* **Naming Convention: Zip the folder and only one person in the team is required to submit the work.**