To get python installed:

Load python 3.7.0:

spack load -r python@3.7.0^gcc@6.3.0

Shows that python is loaded:

echo $LOADEDMODULES | sed "s/:/\n/g" | sort

Load pandas for use in python:

spack load -r py-pandas

Do not need to run but shows that pandas is downloaded now:

echo $LOADEDMODULES | sed "s/:/\n/g" | sort

To open python (like R very simple):

python

Now we can import pandas as usual:

import pandas as pd

IMPORTANT: BEFORE MANIPULATION OF THE DATA, WE FIRST HAVE TO OPEN THE .gz FILES USING LINUX COMMAND zcat FOR ALL FILES USED ALONG WITH USING THE > COMMAND TO MAKE THIS INTO A SUBSET. AN EXAMPLE CODE FOR ukb26867 IS SHOWN BELOW:

zcat /athena/elementolab/scratch/nib4003/ukbiobank/phenotypes/ukb26867.csv.gz > subset\_ukb

\* This does take a little bit of time since the files are so big, but then we can use them in python

The next step is to read in the columns we want for our dataframe to be able to diagnose prediabetes. We have to do this separately for the four HTML files.

first\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_26867', usecols = ['eid', '244

3-0.0', '2443-1.0', '2443-2.0', '2976-0.0', '2976-1.0', '2976-2.0', '4041-0.0', '4041-1.0', '4041-2.0'])

second\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_42385', usecols = ['eid', '30

740-0.0', '30740-1.0', '30741-0.0', '30741-1.0', '30750-0.0', '30750-1.0', '30751-0.0', '30751-1.0'])

Now we can merge these dataframes by eid (patient number).

merged\_prediabetes\_information = first\_step.merge(second\_step, on = 'eid')

Finally we can write this dataframe to a csv so that we can import it to our desktop using winscp.

merged\_prediabetes\_information.to\_csv(path\_or\_buf = '~nib4003/for\_winscp/merged\_prediabetes\_information')

1/18/2021:

The next step is to make a classifier that contains many columns that we want to cut down using feature selection. We first start by using Linux commands to create our 4 whole\_file\_42385 (one example of the 4) files that contain all the HTML features. We next need to combine all the features with the dataframe we created of all classified prediabetic patients so that the computation does not take a long time. First, we import our prediabetic patient dataframe.

all\_prediabetes = pd.read\_csv('for\_winscp/prediabetic\_all\_possible\_to\_classify\_final')

An additional column comes along with the above dataframe that we need to drop, shown below.

all\_prediabetes = all\_prediabetes.drop(columns = ['Unnamed: 0'])

Next we need to import the four large files with all the features. An example is shown below but we need to do this for all the files. This does take a long time for every file since they are so large.

first\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_26867')

Next we want to keep only the columns with a small percentage of NaN values in the columns. Therefore, we drop all columns with greater than 50% of values in the column being NaN.

less\_features\_first\_step = first\_step[first\_step.columns[first\_step.isnull().mean() < 0]]

We can look at the new number of features using the command below.

less\_features\_first\_step.shape

The complete list of commands for all four files is shown below (includes the previous 3):

first\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_26867')

less\_features\_first\_step = first\_step[first\_step.columns[first\_step.isnull().mean() < 0]]

less\_features\_first\_step.shape

second\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_33822')

less\_features\_second\_step = second\_step[second\_step.columns[second\_step.isnull().mean() < 0.5]]

less\_features\_second\_step.shape

third\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_41972')

less\_features\_third\_step = third\_step[third\_step.columns[third\_step.isnull().mean() < 0.5]]

less\_features\_third\_step.shape

fourth\_step = pd.read\_csv('/athena/elementolab/scratch/nib4003/documentation\_files/whole\_file\_42385')

less\_features\_fourth\_step = fourth\_step[fourth\_step.columns[fourth\_step.isnull().mean() < 0.5]]

less\_features\_fourth\_step.shape

We do this for all 4 large files and then merge the remaining 4 dataframes with many features to that of the prediabetic patient dataframe we defined on the desktop. The merge command is shown below.

all\_features\_prediabetics\_only = all\_prediabetes.merge(less\_features\_first\_step, on = 'eid').merge(less\_features\_second\_step, on = 'eid').merge(less\_features\_fourth\_step, on = 'eid')

Now we finally have the whole dataframe with all the features for the prediabetic patients. The next step is to save the dataframe as a csv, as we have done before, in order to send it back to the desktop using WinSCP.

all\_features\_prediabetics\_only.to\_csv(path\_or\_buf = '~nib4003/for\_winscp/all\_features\_prediabetics\_only')