# \*\*\*Protein-Predictor\*\*\*

#Clear Environment

rm(list=ls())

#Import Library

library(devtools)

#Import Files

database=read.csv("C:/Users/Trinity/Desktop/Protein-Predictor-Final/Input\_Files/database.csv", header=TRUE)

sample=read.csv("C:/Users/Trinity/Desktop/Protein-Predictor-Final/Input\_Files/sampledata.csv", header=TRUE)

#Make Empty Data Frames for New Tables

compressed.database <- NULL

compressed.sample <- NULL

matched.data <- NULL

#Identify Relevant Parameters from Database

acession.number <- database[,1]

pI <- database[,2]

length <- database[,3]

weight <- database[,4]

sequence <- database[,10]

#Identify Relevant Parameters from Sample

acession\_number<- sample[,2]

#Compress Sample and Database to Relevant Info

compressed.database <- data.frame(acession.number, pI, length, weight, sequence)

compressed.sample <- data.frame(acession\_number)

#Find Matching Acession Numbers

sampleindatabase <- match(acession\_number, acession.number)

#Create for loop to Convert Row Numbers w/ Matches to the Values Inside the Column

for (i in sampleindatabase){

  acession.number.match <- compressed.database[sampleindatabase,1]

     pI.match <- compressed.database[sampleindatabase,2]

    length.match <- compressed.database[sampleindatabase,3]

  weight.match <- compressed.database[sampleindatabase,4]

sequence.match <- compressed.database[sampleindatabase,5]

}

#Put Matched Values Into New Data Frame

matched.data <- data.frame(acession\_number, pI.match, length.match, weight.match)

# Windows PowerShell

Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users\Trinity> ls

Directory: C:\Users\Trinity

Mode            LastWriteTime     Length Name

----            -------------     ------ ----

d-----    6/27/2019   9:02 AM       .anaconda

d-----     7/5/2019  11:33 AM           .atom

d-----    6/28/2019   3:59 PM       .conda

d-----    6/27/2019   1:47 PM       .idlerc

d-----    6/27/2019   9:13 AM       .ipynb\_checkpoints

d-----    6/27/2019   9:13 AM       .ipython

d-----    6/27/2019   4:05 PM       .jupyter

d-----     7/5/2019  11:53 AM           .matplotlib

d-----     7/5/2019  11:58 AM           .pylint.d

d-----    6/27/2019   1:49 PM       .vscode

d-r---     7/1/2019   8:57 AM         3D Objects

d-----    6/26/2019   4:28 PM       Anaconda3

d-----    6/26/2019   3:23 PM       Anaconda4

d-r---     7/1/2019   8:57 AM         Contacts

d-r---     7/3/2019   9:04 AM         Desktop

d-r---     7/2/2019  10:08 AM           Documents

d-r---     7/5/2019  10:57 AM           Downloads

d-r---     7/1/2019   8:57 AM         Favorites

d-r---     7/1/2019   8:57 AM         Links

d-----     7/3/2019   9:10 AM         lpthw

d-r---     7/1/2019   8:57 AM         Music

d-r---    6/27/2019   1:53 PM       OneDrive

d-r---     7/3/2019   3:08 PM         Pictures

d-----     7/3/2019   9:09 AM         same

d-r---     7/1/2019   8:57 AM         Saved Games

d-r---     7/1/2019   8:57 AM         Searches

d-----     7/3/2019   9:27 AM         temp

d-----     7/3/2019   9:09 AM         Test

d-r---     7/1/2019   8:57 AM         Videos

d-----     7/3/2019  10:45 AM           vim

-a----    6/28/2019  10:21 AM     193 .bash\_history

-a----    6/28/2019   3:59 PM     60 .condarc

-a----    6/28/2019   9:45 AM   188 .gitconfig

-a----    6/27/2019   9:05 AM 1100864 ENM-Protein-Predictor-master (1).zip

-a----    6/27/2019   4:04 PM   5889 estimator.py

-a----    6/27/2019   9:18 AM   160 prediction\_probability.csv

-a----    6/27/2019   9:17 AM 28940 Untitled.ipynb

PS C:\Users\Trinity> cd Desktop

PS C:\Users\Trinity\Desktop> ls

Directory: C:\Users\Trinity\Desktop

Mode            LastWriteTime     Length Name

----            -------------     ------ ----

d-----     7/2/2019  11:40 AM           ENM-Protein-Predictor-master

d-----    6/28/2019   1:04 PM       Ex\_Files\_Learning\_Python (1)

-a----     7/3/2019   9:04 AM   2253 Atom.lnk

-a----     7/2/2019  11:40 AM       985 ENM-Protein-Predictor-master - Shortcut.lnk

-a----    6/27/2019   3:41 PM   2427 GitHub Desktop.lnk

-a----    6/25/2019   5:48 PM   1446 Microsoft Edge.lnk

-a----    6/26/2019   4:17 PM   2144 Thonny.lnk

-a----    6/28/2019  12:55 PM   11536 validation\_utils.py

-a----    6/27/2019   1:49 PM   1408 Visual Studio Code.lnk

PS C:\Users\Trinity\Desktop> cd ENM-Protein-Predictor-master

PS C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master> ls

Directory: C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master

Mode            LastWriteTime     Length Name

----            -------------     ------ ----

d-----     7/5/2019  12:22 PM           ENM-Protein-Predictor-master

PS C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master> cd ENM-Protein-Predictor-master

PS C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master\ENM-Protein-Predictor-master> ls

Directory: C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master\ENM-Protein-Predictor-master

Mode            LastWriteTime     Length Name

----            -------------     ------ ----

d-----     7/2/2019  11:40 AM           Input\_Files

d-----     7/2/2019  11:40 AM           Output\_Files

d-----     7/5/2019  12:07 PM           \_\_pycache\_\_

-a----     7/2/2019  11:40 AM       34 .gitignore

-a----     7/5/2019  12:07 PM     23031 data\_utils.py

-a----     7/5/2019  12:22 PM     6258 estimator.py

-a----     7/5/2019  10:36 AM       22 helloworld.py

-a----     7/2/2019  11:40 AM     1072 LICENSE

-a----     7/5/2019  12:22 PM     44522 prediction\_probability.csv

-a----     7/5/2019  12:00 PM     3442 predictor\_utils.py

-a----     7/2/2019  11:40 AM     3572 README.md

-a----     7/2/2019  11:40 AM       68 requirements.txt

-a----     7/5/2019  11:58 AM     2665 statistic\_parser.py

-a----     7/5/2019  12:22 PM     1955 test.json

-a----     7/2/2019  11:40 AM     11536 validation\_utils.py

-a----     7/2/2019  11:40 AM     11259 visualization\_utils.py

PS C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master\ENM-Protein-Predictor-master> python estimator.py 2 test.json

python visualization\_utils.py

C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master\ENM-Protein-Predictor-master\data\_utils.py:432: FutureWarning: set\_value is deprecated and will be removed in a future release. Please use .at[] or .iat[] accessors instead

  categorical\_df.set\_value(key, v, 1)

Run Number: 0

Run Number: 1

#########Confusion Matrix over all 50 runs#########

unbound: 132.0, 110.0

bound: 51.0, 311.0

#########VALUES########

~Performance Metrics~

AUROC : 0.8094379252088946

Recall : 0.8591160220994476

Precision : 0.7388031062350444

F1 Score : 0.7943755709252251

Accuracy : 0.7334437086092715

~Feature Importances~

pI: 0.16431279536053978

% Positive: 0.12927072472405932

% Negative: 0.12102218995178268

Protein Weight: 0.1082367841421944

% Hydrophilic: 0.10335140691141526

% Aromatic: 0.0959682220109053

Protein Abundance: 0.08258167965544165

% Cysteine: 0.07623817629937636

Solvent NaCl Concentration\_0.0: 0.021281090238791052

Particle Size\_100: 0.014117796828675616

Particle Size\_10: 0.014048208982781531

Particle Charge\_0: 0.013317637619797883

Particle Charge\_1: 0.013192632208660252

Solvent Cysteine Concentration\_0.1: 0.011936310485503036

Solvent Cysteine Concentration\_0.0: 0.011884271146131338

Solvent NaCl Concentration\_0.8: 0.009724204061819758

Solvent NaCl Concentration\_3.0: 0.0095158693721248

########STD VALUES#######

~Performance Metrics~

AUROC: 0.004223551436007444

Recall: 0.008287292817679537

Precision: 0.005158246421960333

F1 Score: 0.000561137935534417

Accuracy: 0.0016556291390728561

~Feature Importances~

pI: 0.003863357919583893

Protein Weight: 0.00012999612352380258

% Aromatic: 0.00013472030631260512

% Negative: 0.0016884711952191364

% Positive: 0.0011322596427877224

% Hydrophilic: 0.0026615209311800433

% Cysteine: 0.0007291173428903008

Protein Abundance: 6.406534188375945e-05

Particle Size\_10: 6.295831913354089e-05

Particle Size\_100: 0.0003480878802699138

Particle Charge\_0: 0.00041806678028299873

Particle Charge\_1: 0.0004980584549342578

Solvent Cysteine Concentration\_0.0: 0.00019083834997165369

Solvent Cysteine Concentration\_0.1: 0.00029246050230751636

Solvent NaCl Concentration\_0.0: 0.00010537509160435561

Solvent NaCl Concentration\_0.8: 0.0003184196490902298

Solvent NaCl Concentration\_3.0: 0.0003908473490793129

PS C:\Users\Trinity\Desktop\ENM-Protein-Predictor-master\ENM-Protein-Predictor-master>

Thonny (script editor)

Go to line 126 in estimator.py: comment out

Go to line 92 in estimator.py and put a csv path there to predict data for samples