PITCH DECK-
IBMZ DATATHON

DATAVENGERS
Team 107

Contents

- Problem Statement
- Dataset
- Data Analytics
- Model building
- Use of IBMz tool
- Target Market
- Competitive Advantages

Problem Statement - Diabetes class detector

The ability to predict a disease in its early stages is a field of active research. Diabetes, specifically is a difficult disease to predict with a decent level of accuracy. **Early stage detection and prediction** of the disease might be an effective way to understand the biochemical pathways linked to the disease.

Moreover, compared to analyzing the pathway using a single specific omics dataset, integrating different omics datasets (**multi-omics approach**) can help us in understanding the disease comprehensively.

The ability to make a computer system understand this very comprehensive learning is what we expect to achieve by building a **machine learning model** for the multi-omics data analytics approach. We expect to achieve this concept using multi-omics datasets obtained from **diabetic patients**. In this way, it can **classify a patient in one of the 4 classes: Diabetic, Prediabetic, Control and Crossover.**

DATASETS

Source:

The multi-omic datasets are taken from Stanford Medicine

Constituents:

It comprises three omics datasets namely: Proteomic, Lipidomic and Metabolomic

Metabolomics:

- Obtained from blood plasma
- LC-MS intensity measures (log-transformed)
- The data had 729attributes and 176 sample records

Proteomics:

- Obtained from blood plasma
- LC-MS intensity measures (log-transformed)
- The data had 261 attributes and 176 sample records.

Lipidomics:

- Tissue samples
- MS intensity normalized
- The data had 711 attributes and 174sample records

DATA ANALYTICS

Single Omics - Preprocessing

The individual omics datasets were preprocessed by missing value imputation, feature selection using corrplots.

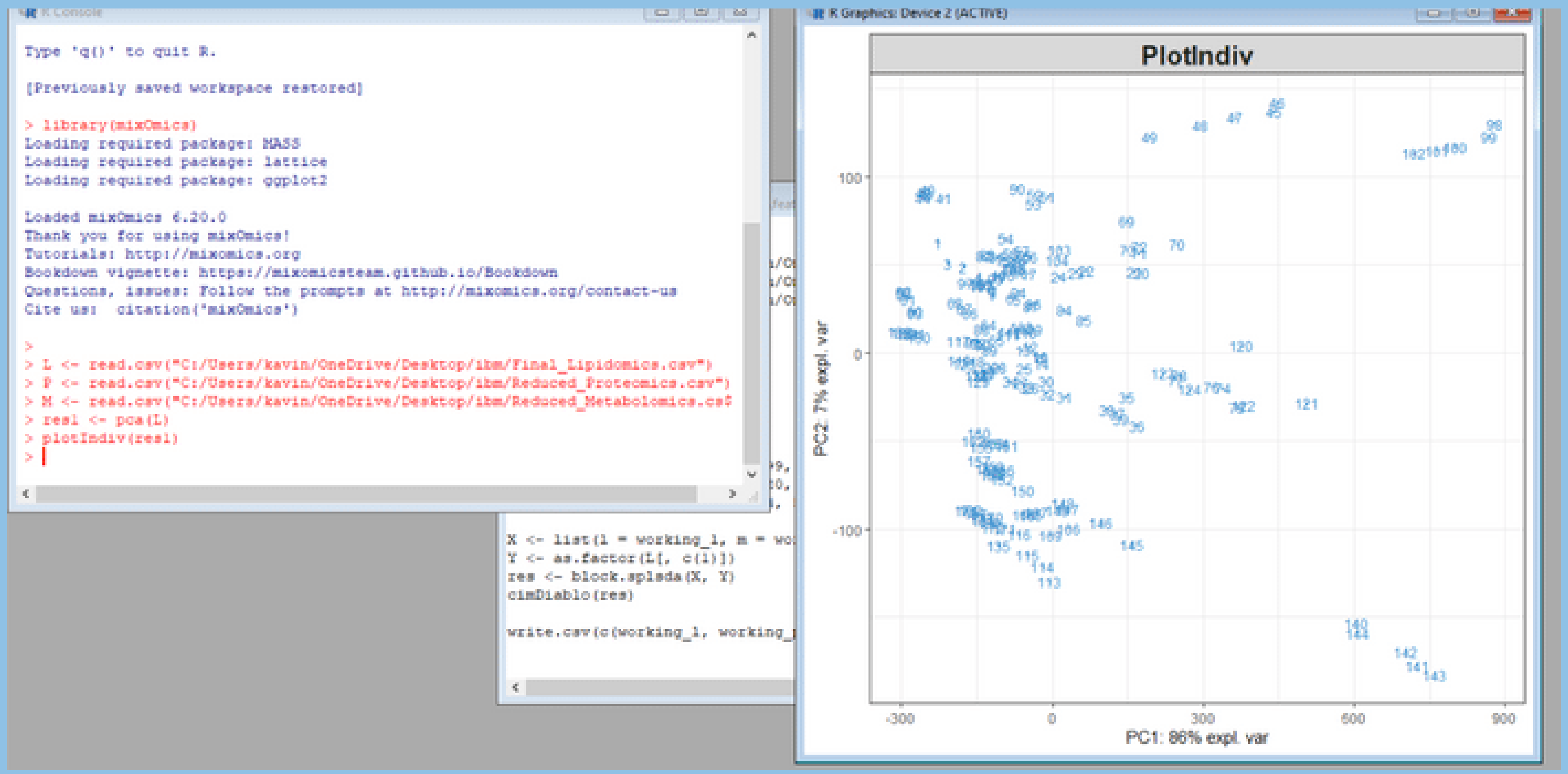
Single Omics - PCA

The preprocessed datasets were then individually plotted using PCA and the top differentially expressed features were collected.

Multi omics - CIM

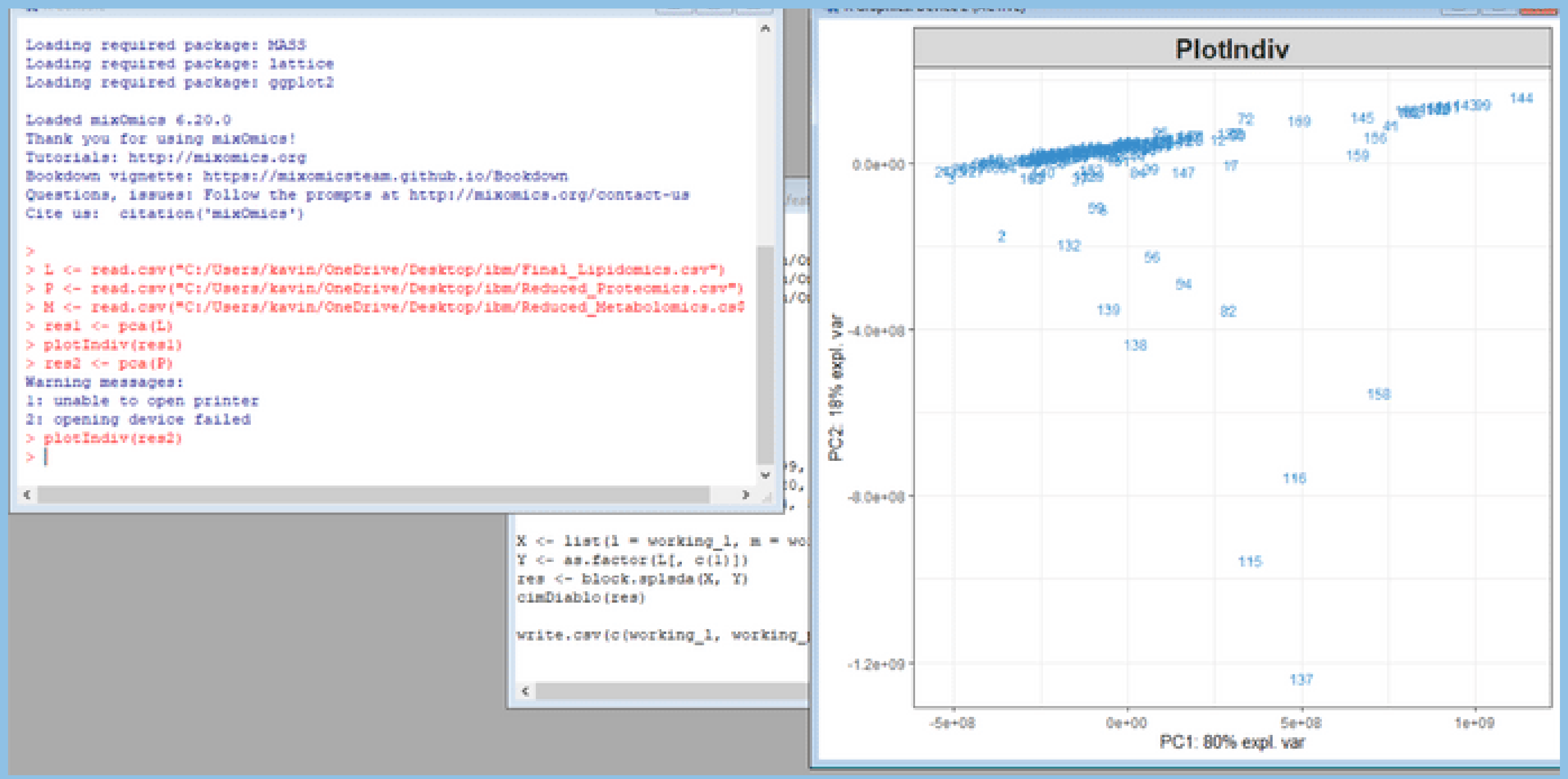
The features selected from each dataset were integrated using the concept of multi-omics data integration approaches and it was found that the features from each omics dataset were independent of the other datasets.

PCA Plots



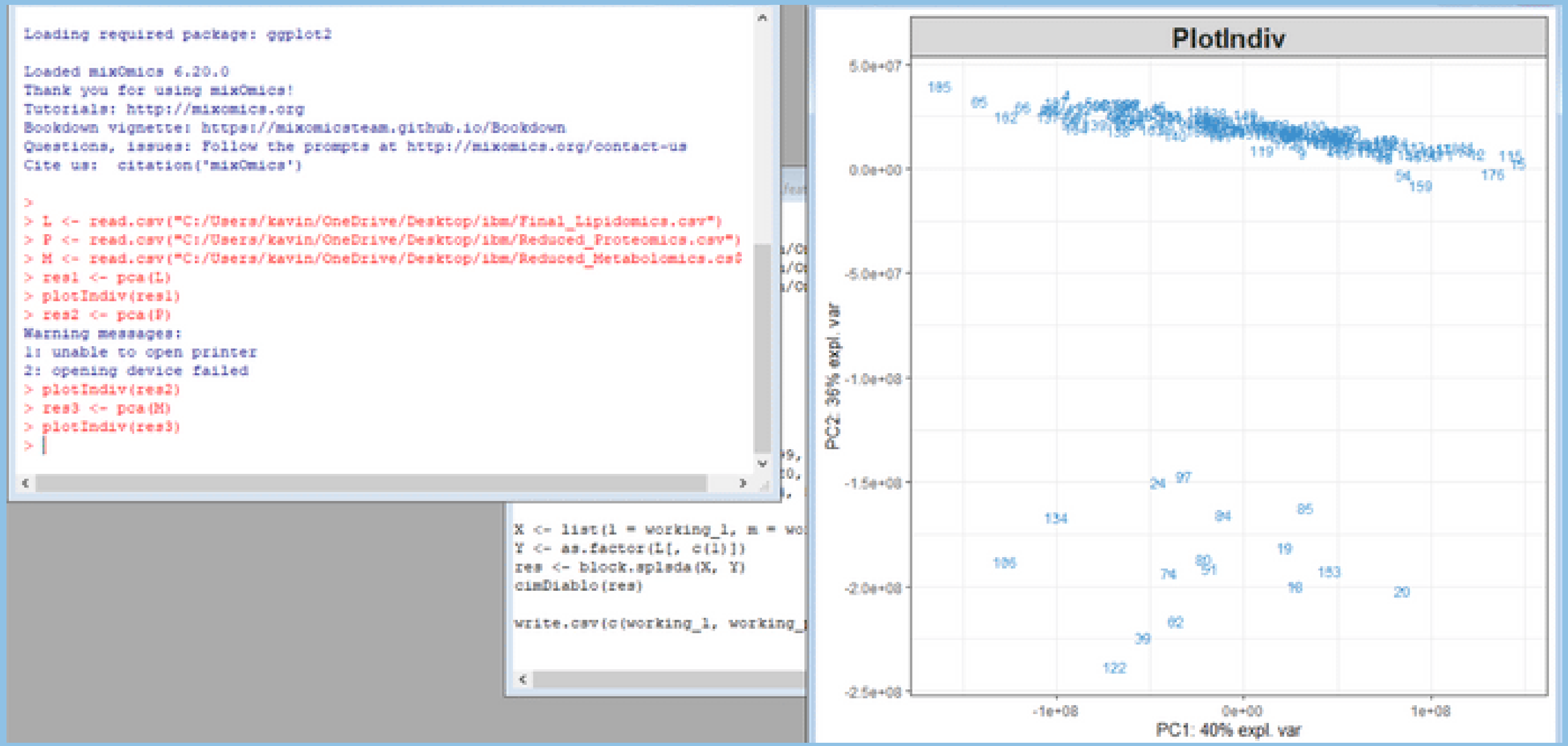
```
Type 'q()' to quit R.

[Previously saved workspace restored]

> library(mixOmics)
Loading required package: MASS
Loading required package: lattice
Loading required package: ggplot2

Loaded mixOmics 6.20.0
Thank you for using mixOmics!
Tutorials: http://mixomics.org
Bookdown vignette: https://mixomicsteam.github.io/Bookdown
Questions, issues: Follow the prompts at http://mixomics.org/contact-us
Cite us:  citation('mixOmics')

>
> L <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Final_Lipidomics.csv")
> P <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Reduced_Proteomics.csv")
> M <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Reduced_Metabolomics.cs$
> res1 <- pca(L)
> plotIndiv(res1)
>
```

```
X <- list(l = working_l, m = wo
Y <- as.factor(L[, c(1)])
res <- block.splsda(X, Y)
cimDiablo(res)

write.csv(c(working_l, working_
```



```
Loading required package: MASS
Loading required package: lattice
Loading required package: ggplot2

Loaded mixOmics 6.20.0
Thank you for using mixOmics!
Tutorials: http://mixomics.org
Bookdown vignette: https://mixomicsteam.github.io/Bookdown
Questions, issues: Follow the prompts at http://mixomics.org/contact-us
Cite us:  citation('mixOmics')

>
> L <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Final_Lipidomics.csv")
> P <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Reduced_Proteomics.csv")
> M <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Reduced_Metabolomics.cs$
> res1 <- pca(L)
> plotIndiv(res1)
> res2 <- pca(P)
Warning messages:
1: unable to open printer
2: opening device failed
> plotIndiv(res2)
>
```

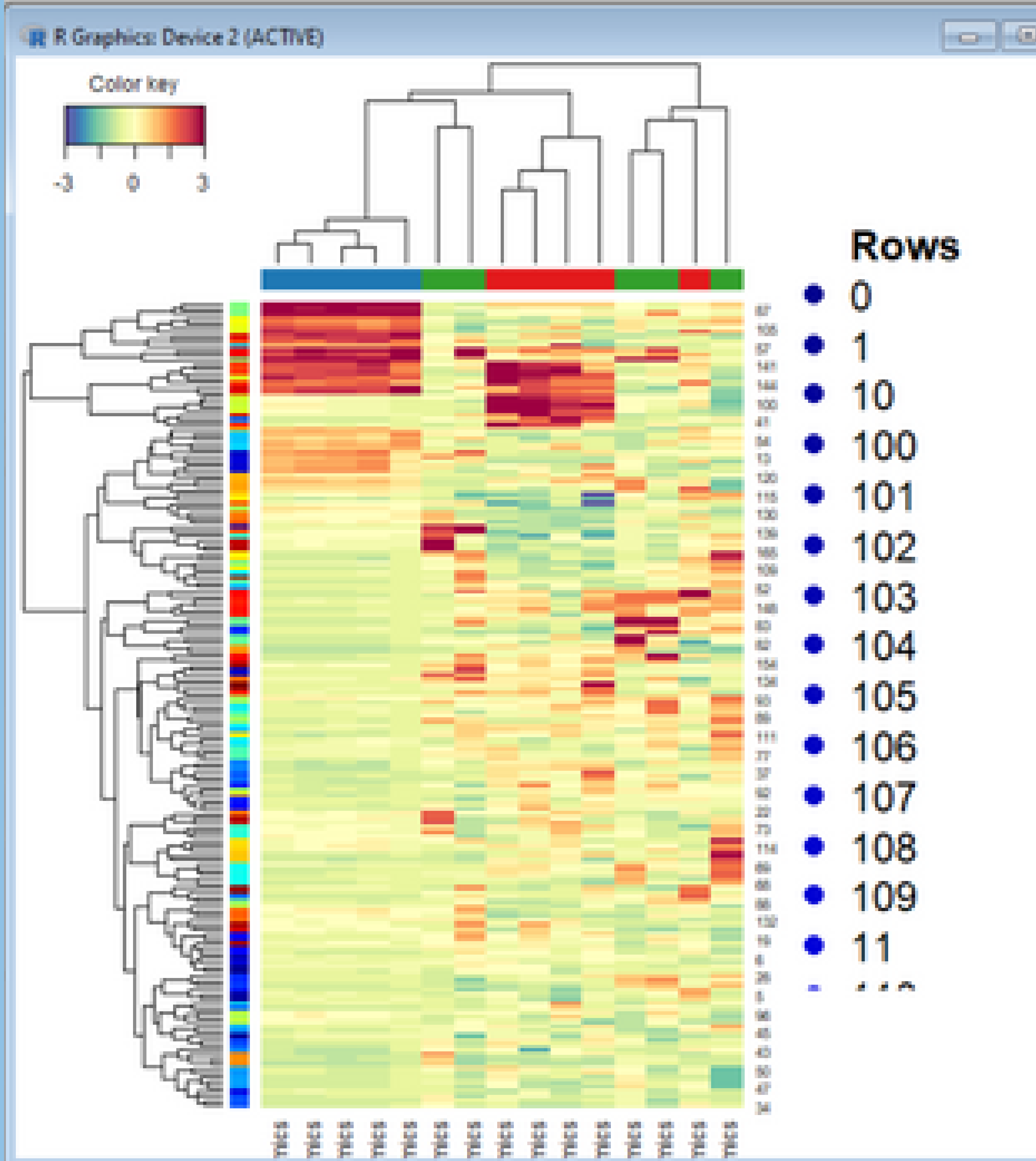
```
X <- list(l = working_l, m = wo
Y <- as.factor(L[, c(1)])
res <- block.splsda(X, Y)
cimDiablo(res)

write.csv(c(working_l, working_
```



```
Loading required package: ggplot2

Loaded mixOmics 6.20.0
Thank you for using mixOmics!
Tutorials: http://mixomics.org
Bookdown vignette: https://mixomicsteam.github.io/Bookdown
Questions, issues: Follow the prompts at http://mixomics.org/contact-us
Cite us:  citation('mixOmics')

>
> L <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Final_Lipidomics.csv")
> P <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Reduced_Proteomics.csv")
> M <- read.csv("C:/Users/kavin/OneDrive/Desktop/ibm/Reduced_Metabolomics.cs$
> res1 <- pca(L)
> plotIndiv(res1)
> res2 <- pca(P)
Warning messages:
1: unable to open printer
2: opening device failed
> plotIndiv(res2)
> res3 <- pca(M)
> plotIndiv(res3)
>
```

```
X <- list(l = working_l, m = wo
Y <- as.factor(L[, c(1)])
res <- block.splsda(X, Y)
cimDiablo(res)

write.csv(c(working_l, working_
```

Clustered Heatmap



Model Building

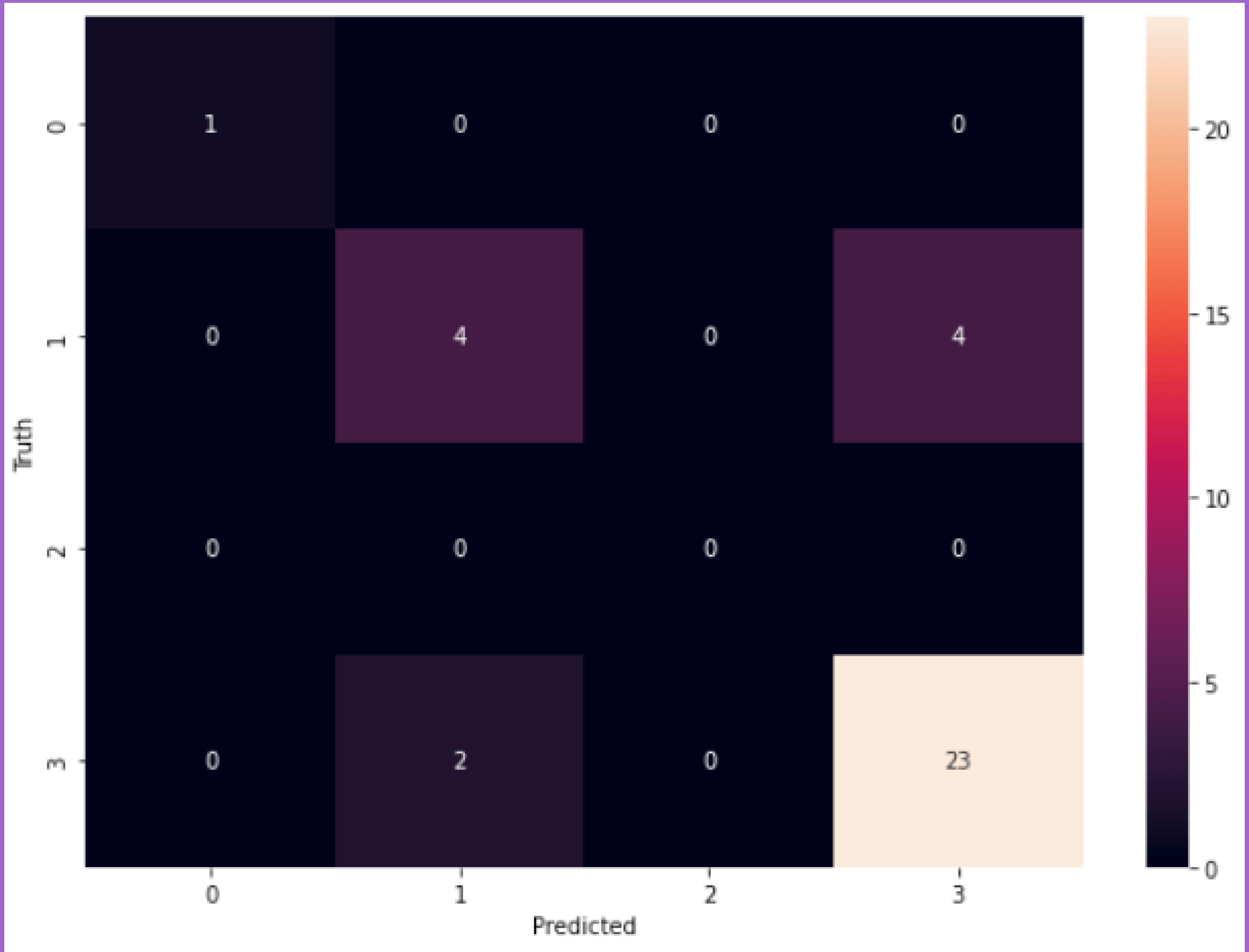
Creating ANN to predict diabetes using our pre-processed dataset

- A simple ANN model is built which classifies the patient record into 4 classes - diabetic, pre-diabetic, control and crossover

- It has a 4 layer architecture. Each of the hidden layer consist of 100 units with 'ReLU ' activation function. The final output layer consists of a 'softmax' activation which gives us the probablity of the sample belonging to one of the classes.

- Post training with 100 epochs, the model resulted in 82% accuracy and a loss of 0.42

Confusion matrix



	0	1	2	3
0	1	0	0	0
1	0	4	0	4
2	0	0	0	0
3	0	2	0	23

Usage of IBMz tool

Collaborative Coding

Highly efficient in terms of updating the status of the files that are being worked on by multiple users

Security

Access to the notebook instance is only allowed through SSH key pairs and thus prevents unknown accesses to the server. This guarantees absolute safety of our files and data.

Speed

Execution of code happens at a very fast pace and hence helped us in training our model in a very short amount of time .

Management Of Assets

Assets are easier to upload and manage on the L1CC platform and do no get destroyed even when the runtime is disconnected. Thus this feature didn't require us to upload files multiple times on to the Notebook instance.

Usage of IBMz tool

Preinstalled Libraries

The IBMz Platform consists an abundance amount of pre installed libraries and eases load off the user to keep track of dependencies and this definitely helped us big time.

Docker & Container User Format

IBMz L1CC platform makes of Docker and deploys its Jupyter notebook instance on a container.This gives way to just pulling an FS image and deploying it on a Cloud Server.

Flexibility

On the IBMz L1CC Platform,users can scale services to fit their needs, customize applications and access cloud services from anywhere with an internet connection. This fundamentally helped us to access the services remotely and work collaboratively.

Tech stack used:

- Tensorflow
- Python
- R Programming

TARGET MARKET

Who are the customers we want to cater to?

Hospital Management

It efficiently uses the data from the hospital database to predict diabetes in patients.

General Public

Acts as a tool for adults who want to take susceptibility tests for Diabetes.

Competitive Advantages

Research Potential

Existing diagnosis for diabetes involves mere clinical tests. However, combining two powerful concepts like ANN and multi-omics integration provides great research potential.

Comprehensive omics study

The model makes use of not one but 3 different datasets corresponding to different omics study of the human body. This enhances the quality of tracing the pathway and hence classification of prediabetic and diabetic patients.

Predictive analysis

Our model works with an accuracy of 88% on the prediabetic class, proving to be an efficient forecasting strategy for diabetes susceptible patients.

Different perspective to diabetes prediction

As opposed to the usual way of measuring sugar levels in the blood to detect diabetes, we try to draw lines between diabetes and lipids(fats), metabolites and protein content in the human body obtained via mass spectroscopy and liquid crystallization of the tissue and blood samples.

Github link

https://github.com/jsv1604/IBMz_Datathon

Meet our Team

Shrinithi Natarajan

Rhea Sudheer

Jeffrey S Varghese

Shreyas Battula

THANK YOU!