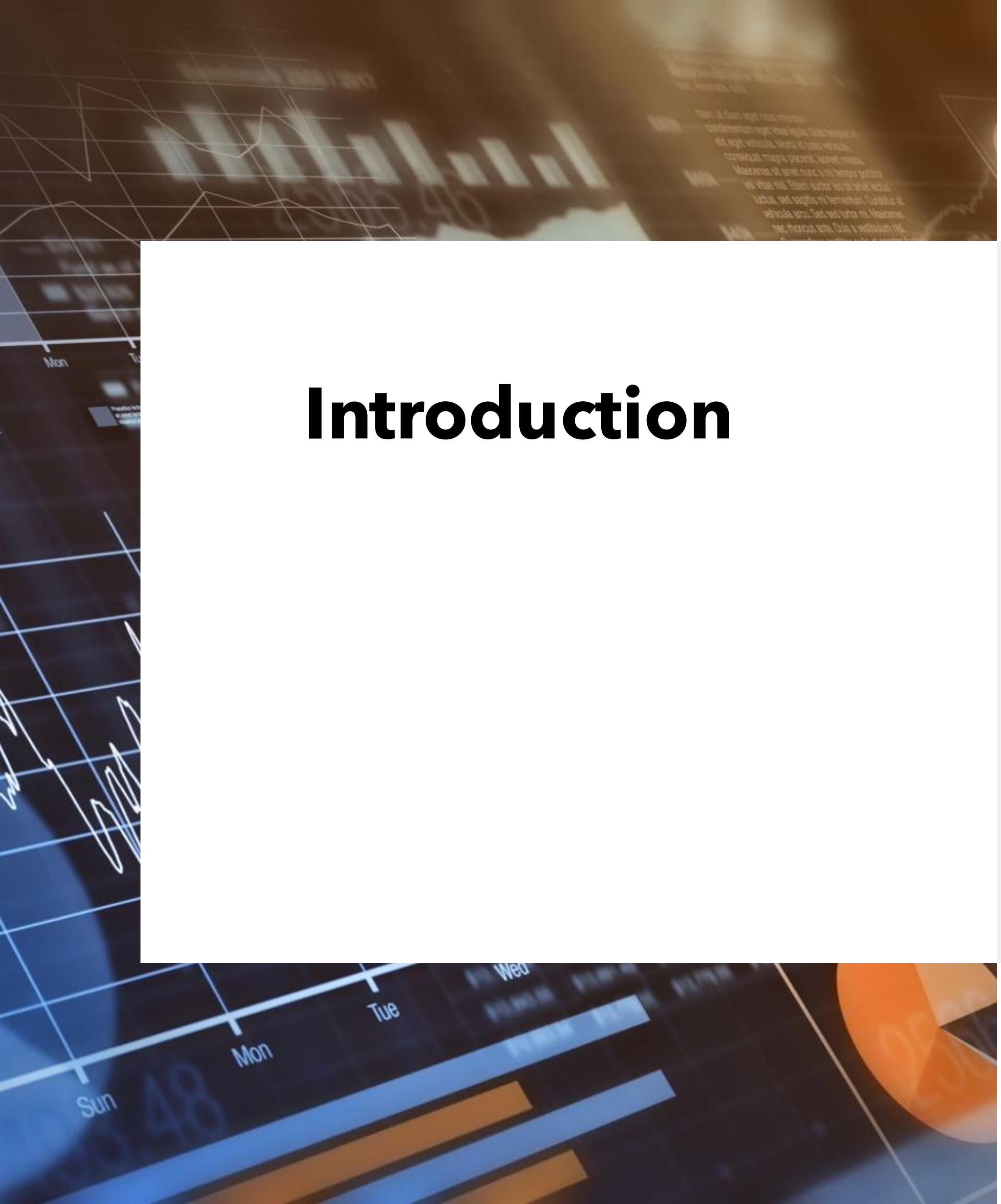
Predicting Health Outcomes Using Nutritional Intake, Socioeconomic Factors, and Sleep Data

HUY DUC PHAM

HANUSH KUMAR

NEHA ANIL CHEDE



Introduction

This study aims to develop a model for disease prediction using NHANES data.

It applies PCA analysis and association rule mining to identify patterns and classifies individuals into risk groups.

Why?

Enhanced Health Predictions: Improves the ability to predict health outcomes and identify at-risk individuals.

Personalized Healthcare: Grouping individuals by risk levels allows for optimized treatment.

Data-Driven Insights: Uncovers hidden relationships between health factors.

Public Health Impact: Identifying health trends and risk factors helps shape public health initiatives and policies.

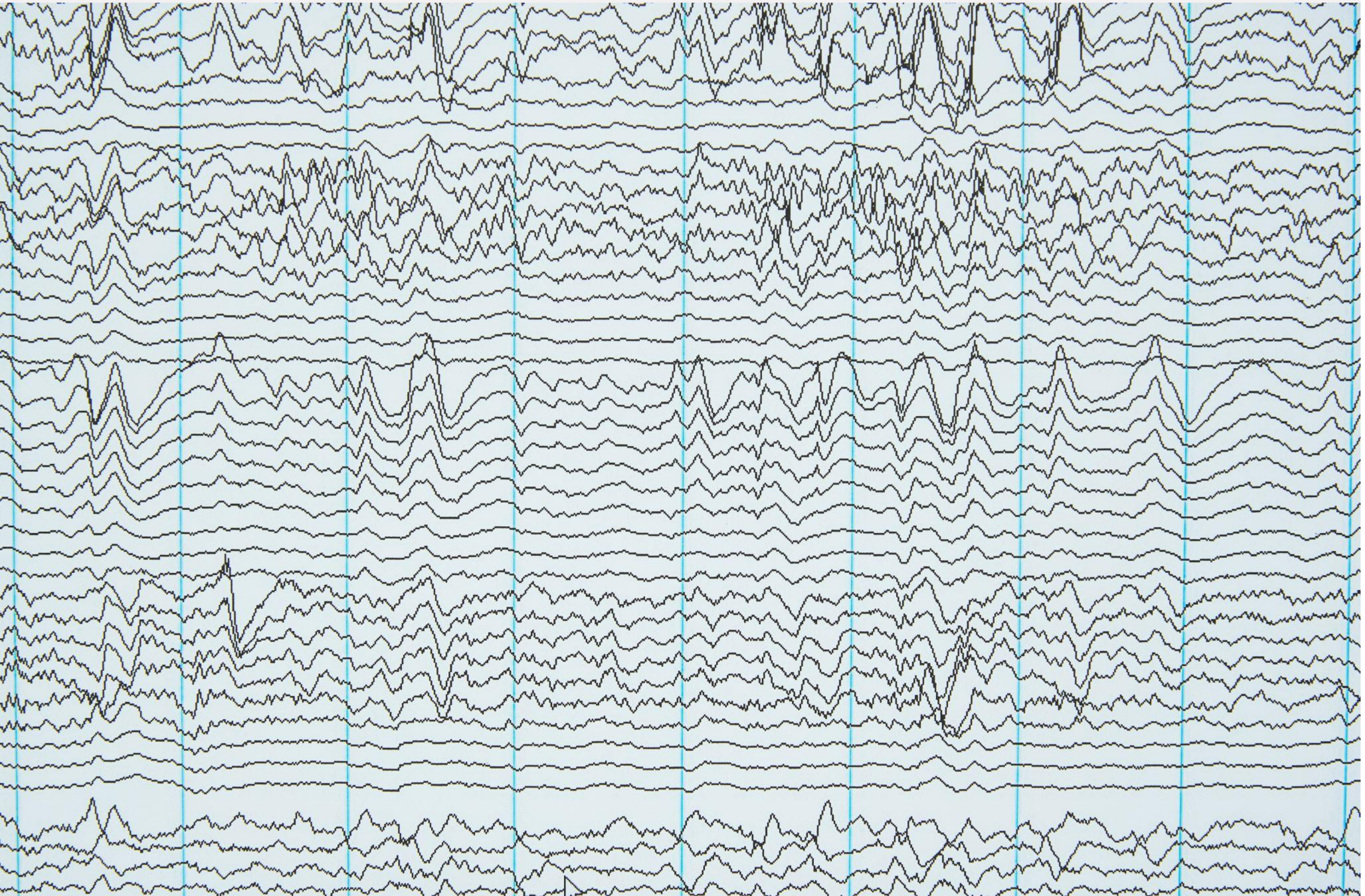
The Dataset: NHANES 2013-2014 (National Health and Nutrition Examination Survey)

NHANES 2013-2014 is a nationally representative survey conducted by the CDC (Centers for Disease Control and Prevention) to assess the health and nutritional status of the U.S. population.

It combines interviews, physical examinations, and laboratory tests to collect data on a wide range of topics, including chronic conditions, dietary habits, physical activity, and environmental exposures.

The datasets are broken into 6 different categories.

Data Categories and Key Variables (Part 1)



Demographics:

Age, gender, race, education, and household structure.

Data Types: Categorical, Continuous

Dietary:

Nutrient intake (e.g., sodium, B12), meal patterns, and food frequency.

Data Types: Continuous, Categorical

Examination:

Blood pressure, weight, height, BMI, and body dimensions (e.g., waist, leg, arm measurements).

Data Types: Continuous, Categorical

Data Categories and Key Variables (Part 2)

Laboratory:

Blood and urine tests, including cholesterol, glucose, and iron levels.

Data Types: Continuous

Medications:

Prescription drug use and primary/secondary diagnoses.

Data Types: Categorical

Questionnaire:

Alcohol and Smoking Habits: Frequency and quantity of use.

Health Conditions: Diagnoses of chronic illnesses (e.g., asthma, hypertension).

Physical Activity and Sedentary Time: Daily minutes reported.

Socioeconomic Factors: Monthly family income.

Data Types: Continuous, Categorical

Analysis Performed

ASSOCIATION ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

CLASSIFICATION

Association Analysis (Ex. Using the Medication and Examination Datasets)

SEQN	BPXSY1	BPXDI1	BPXSY2	BPXDI2	BPXSY3	BPXDI3	BMXWT	BMXHT	BMXBMI
73557	122	72	114	76	102	74	78.3	171.3	26.7
73558	156	62	160	80	156	42	89.5	176.8	28.6
73559	140	90	140	76	146	80	88.9	175.3	28.9
73560	108	38	102	34	104	38	32.2	137.3	17.1
73561	136	86	134	88	142	86	52	162.4	19.7
73562	160	84	158	82	154	80	105	158.7	41.7

Association Analysis Process

Data Loading and Preprocessing:

Categorize continuous variables (e.g., blood pressure, BMI) into discrete bins (e.g., Low, Medium, High, Very High).

Data Transformation:

Convert continuous and categorical variables into **binary formats.**

Merge medication and examination data on a common identifier (SEQN).

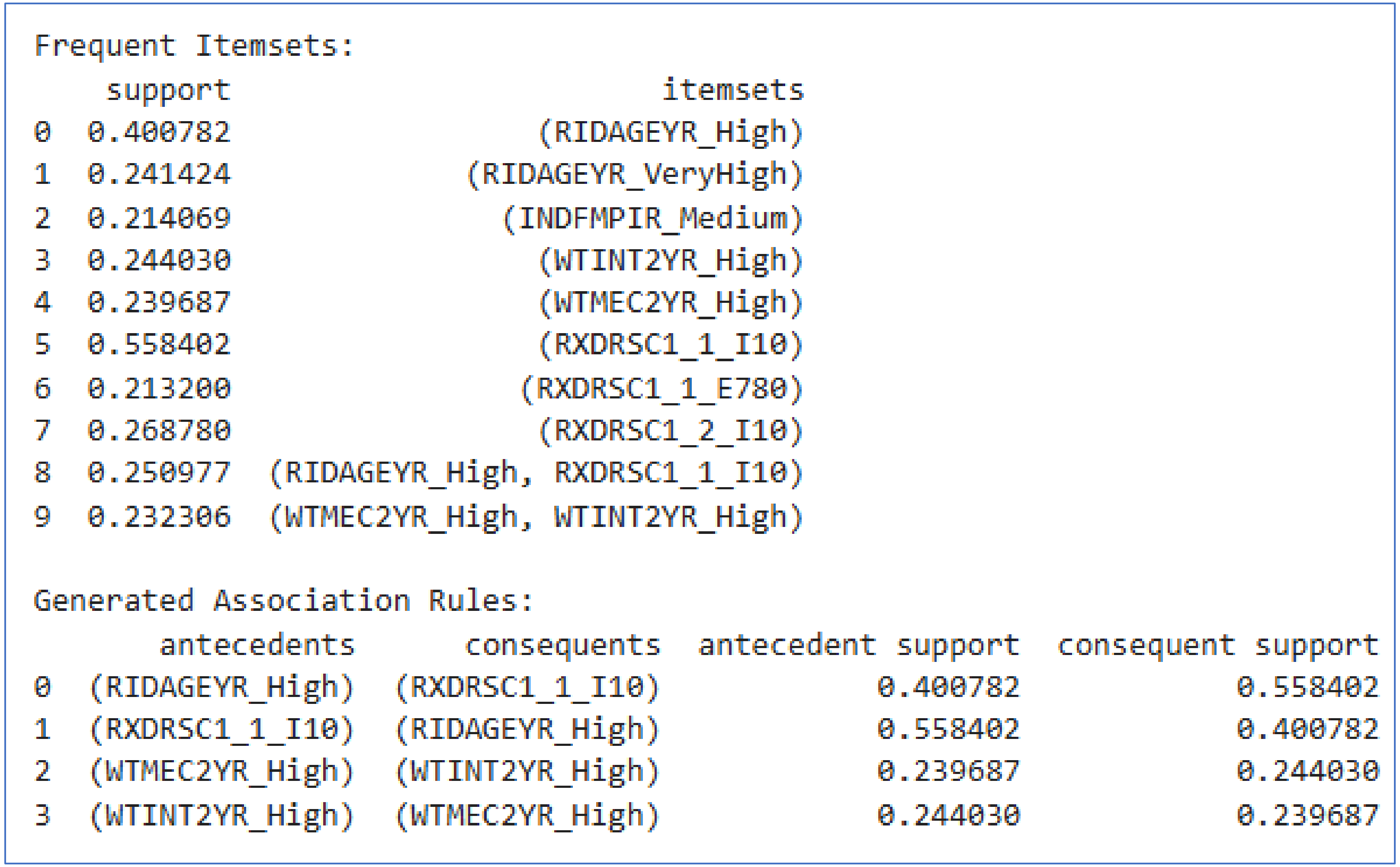
Frequent Itemset Mining (Apriori):

Identify frequent itemsets with a minimum support level.

Association Rule Generation:

Generate association rules based on the frequent itemsets with minimum lift of 1.0

Association Analysis Results (Truncated Example)

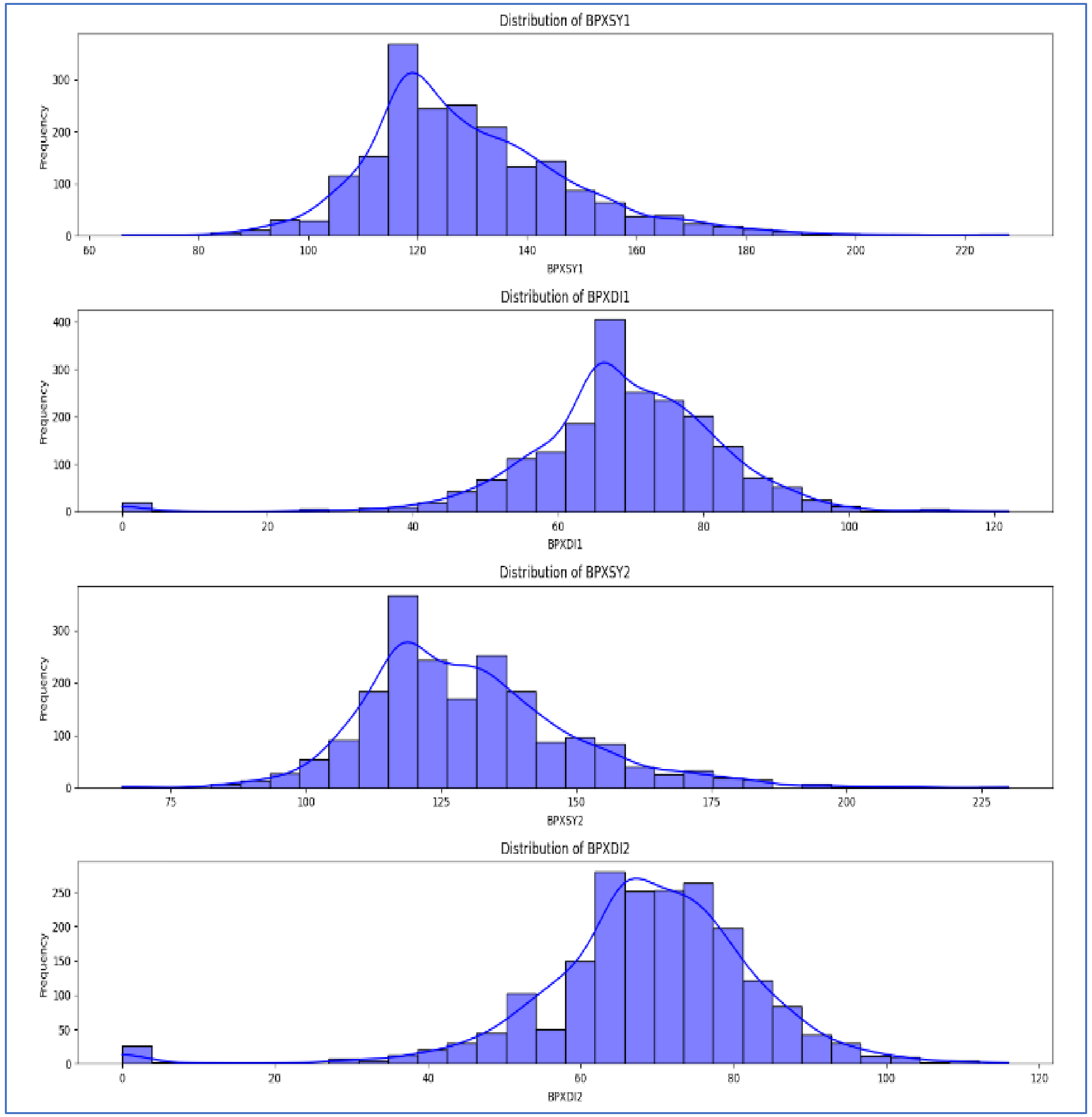


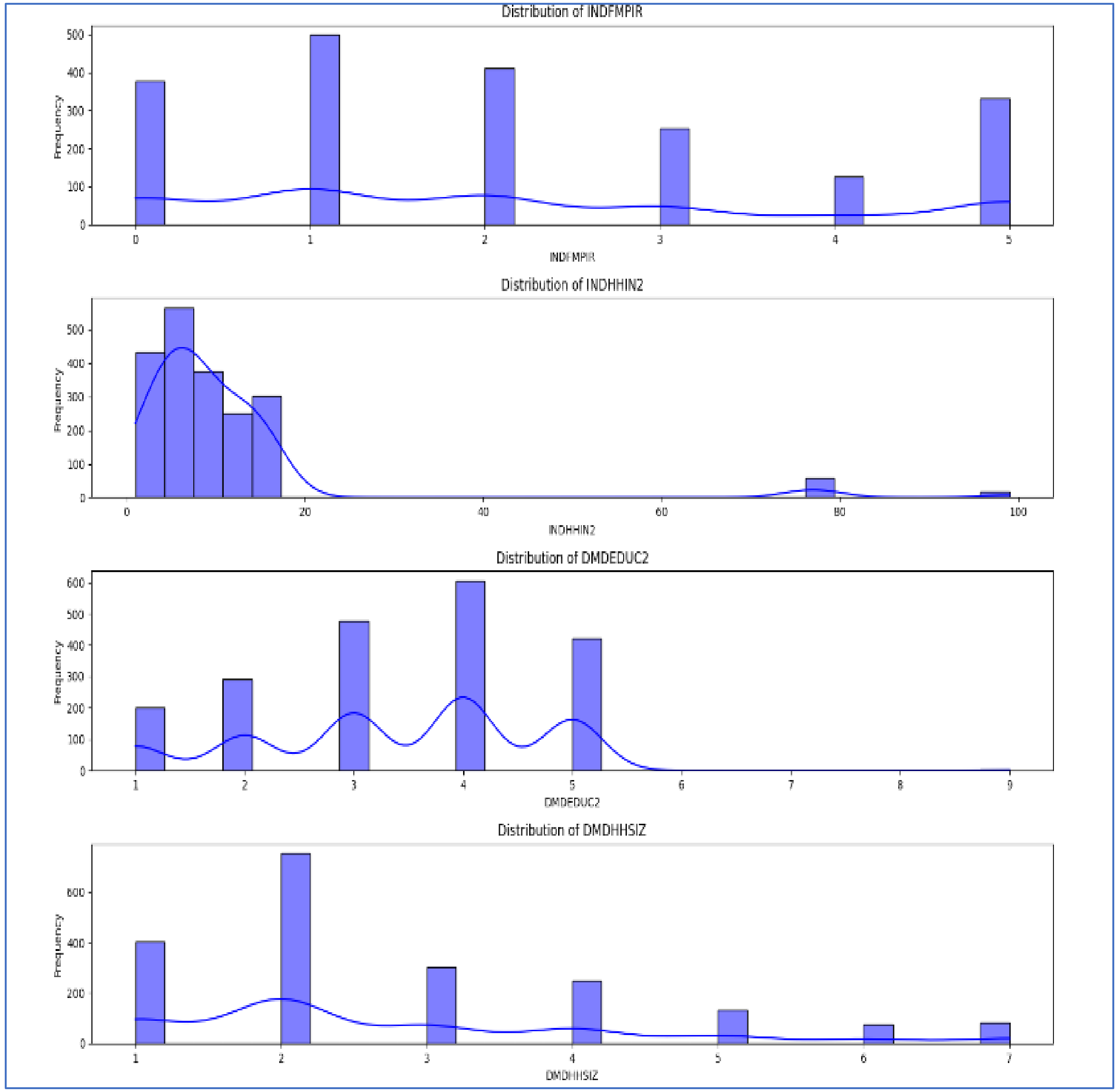
```
Frequent Itemsets:
    support                          itemsets
0  0.400782                   (RIDAGEYR_High)
1  0.241424               (RIDAGEYR_VeryHigh)
2  0.214069                 (INDFMPIR_Medium)
3  0.244030                   (WTINT2YR_High)
4  0.239687                   (WTMEC2YR_High)
5  0.558402                   (RXDRSC1_1_I10)
6  0.213200                  (RXDRSC1_1_E780)
7  0.268780                   (RXDRSC1_2_I10)
8  0.250977  (RIDAGEYR_High, RXDRSC1_1_I10)
9  0.232306  (WTMEC2YR_High, WTINT2YR_High)

Generated Association Rules:
       antecedents      consequents  antecedent support  consequent support
0  (RIDAGEYR_High)  (RXDRSC1_1_I10)            0.400782            0.558402
1  (RXDRSC1_1_I10)  (RIDAGEYR_High)            0.558402            0.400782
2  (WTMEC2YR_High)  (WTINT2YR_High)            0.239687            0.244030
3  (WTINT2YR_High)  (WTMEC2YR_High)            0.244030            0.239687
```

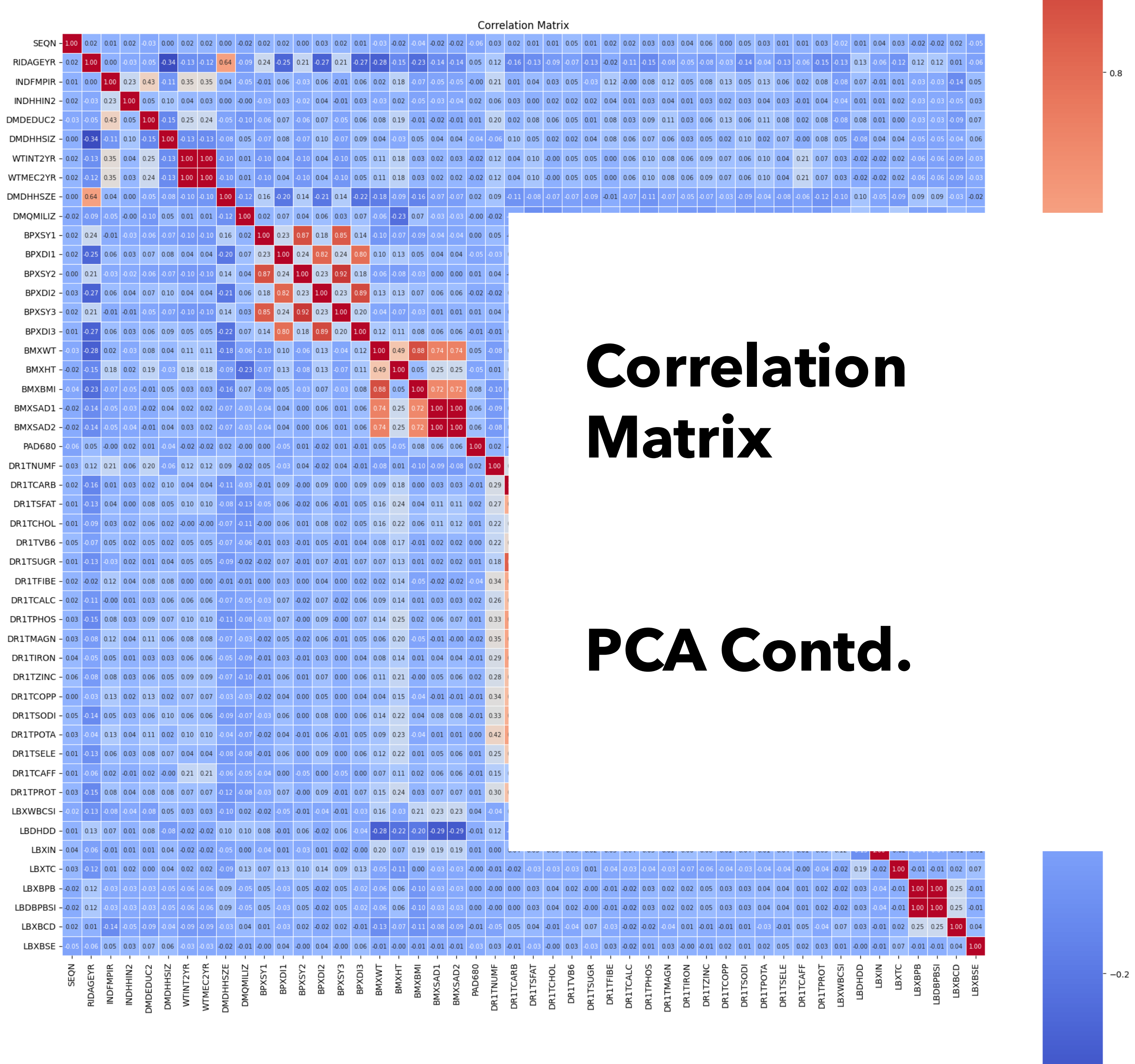
WTMEC2YR = Weightage, WTINT2YR = Interview weightage
RXDRSC1= Medication disease code for having diabetes
RIDAGEYR = Age

Distribution Analysis for PCA





- Standardization is performed for PCA
- Box-cox was performed for skewed



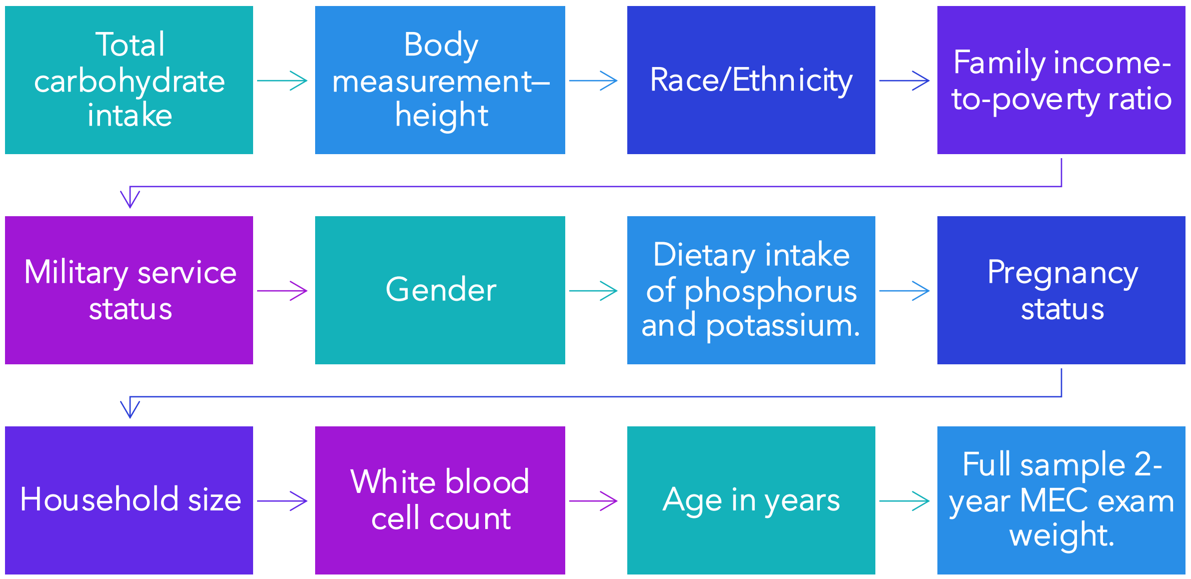
Correlation Matrix

PCA Contd.

Dropping Features using Association analysis & Correlation matrix -

'WTINT2YR', 'RIDEXPRG_3.0', 'BPXSY2', 'BPXSY3', 'BPXDI2', 'BPXDI3', 'BMXBMI', 'BMXSAD1', 'BMXSAD2', 'DR1TSUGR', 'DR1TPROT', 'LBDBPBSI'

PCA Output



Total carbohydrate intake

Body measurement–height

Race/Ethnicity

Family income-to-poverty ratio

Military service status

Gender

Dietary intake of phosphorus and potassium.

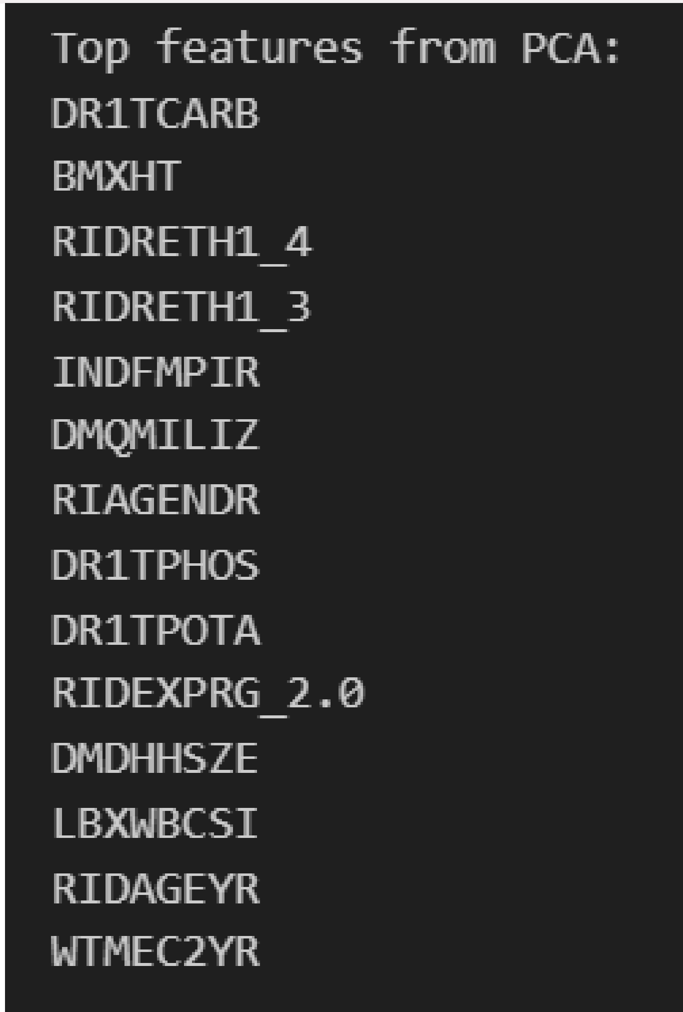
Pregnancy status

Household size

White blood cell count

Age in years

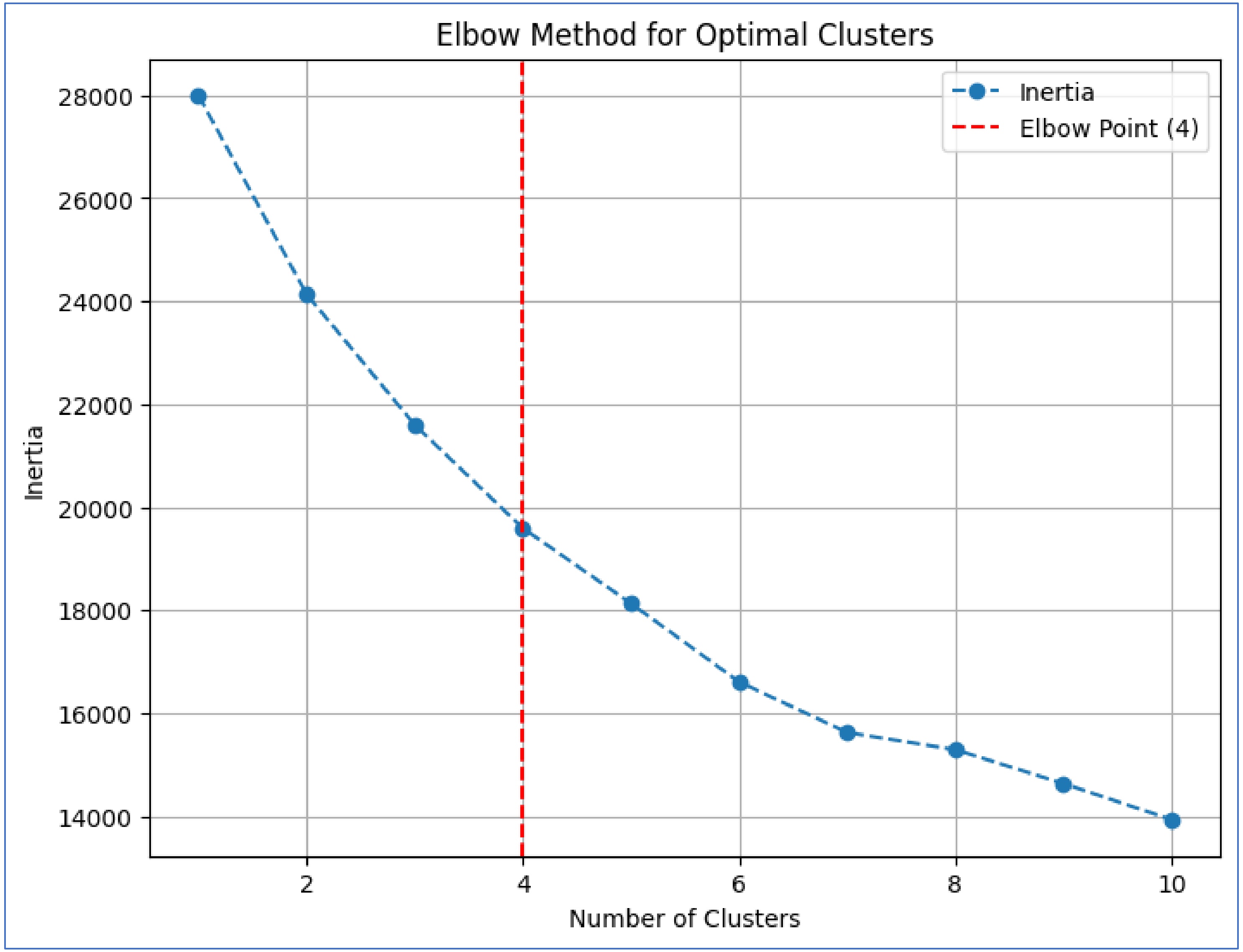
Full sample 2-year MEC exam weight.



```
Top features from PCA:
DR1TCARB
BMXHT
RIDRETH1_4
RIDRETH1_3
INDFMPIR
DMQMILIZ
RIAGENDR
DR1TPHOS
DR1TPOTA
RIDEXPRG_2.0
DMDHHSZE
LBXWBCSI
RIDAGEYR
WTMEC2YR
```

Cluster Optimization

PCA Contd



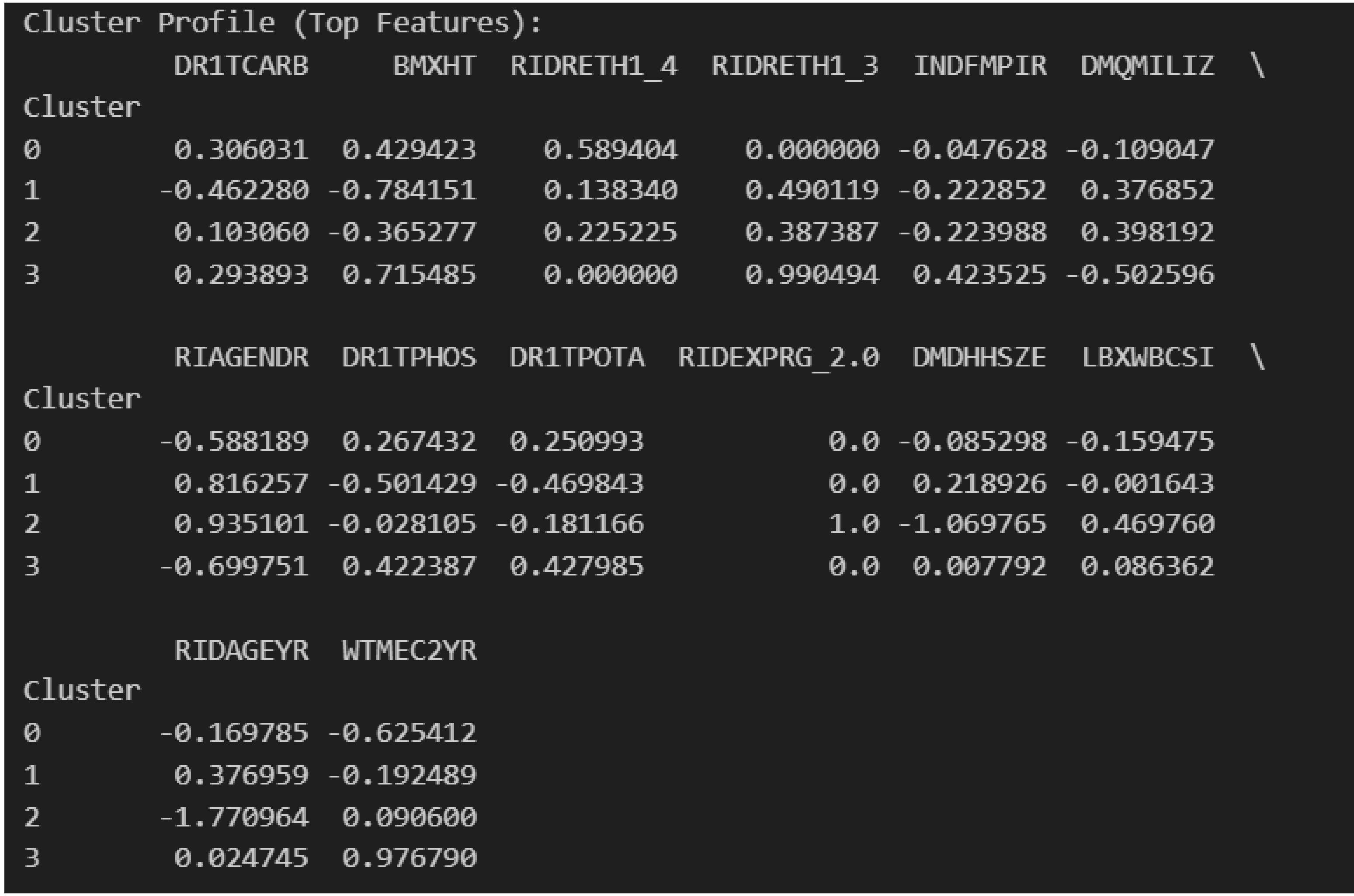
Cluster Profile (PCA Contd)

Cluster 0: higher DR1TCARB, high BMXHT, RIDRETH1_4, below-avg INDFMPIR, negative DMQMILIZ, RIDAGEYR below mean

Cluster 1: Lower DR1TCARB, low BMXHT, Below-average INDFMPIR, Positive DQMILIZ, Majority RIAGENDR_0, Higher RIDAGEYR

Cluster 2: Moderate DR1TCARB, Below-average BMXHT, Predominantly RIDEXPRG_2.0, negative DMDHHSZE, High LBXWBCSI, negative RIDAGEYR

Cluster 3: Higher DR1TCARB, high BMXHT, Predominantly RIDRETH1_3. Above-average INDFMPIR, Negative DQMILIZ, Majority RIAGENDR_1, Average RIDAGEYR



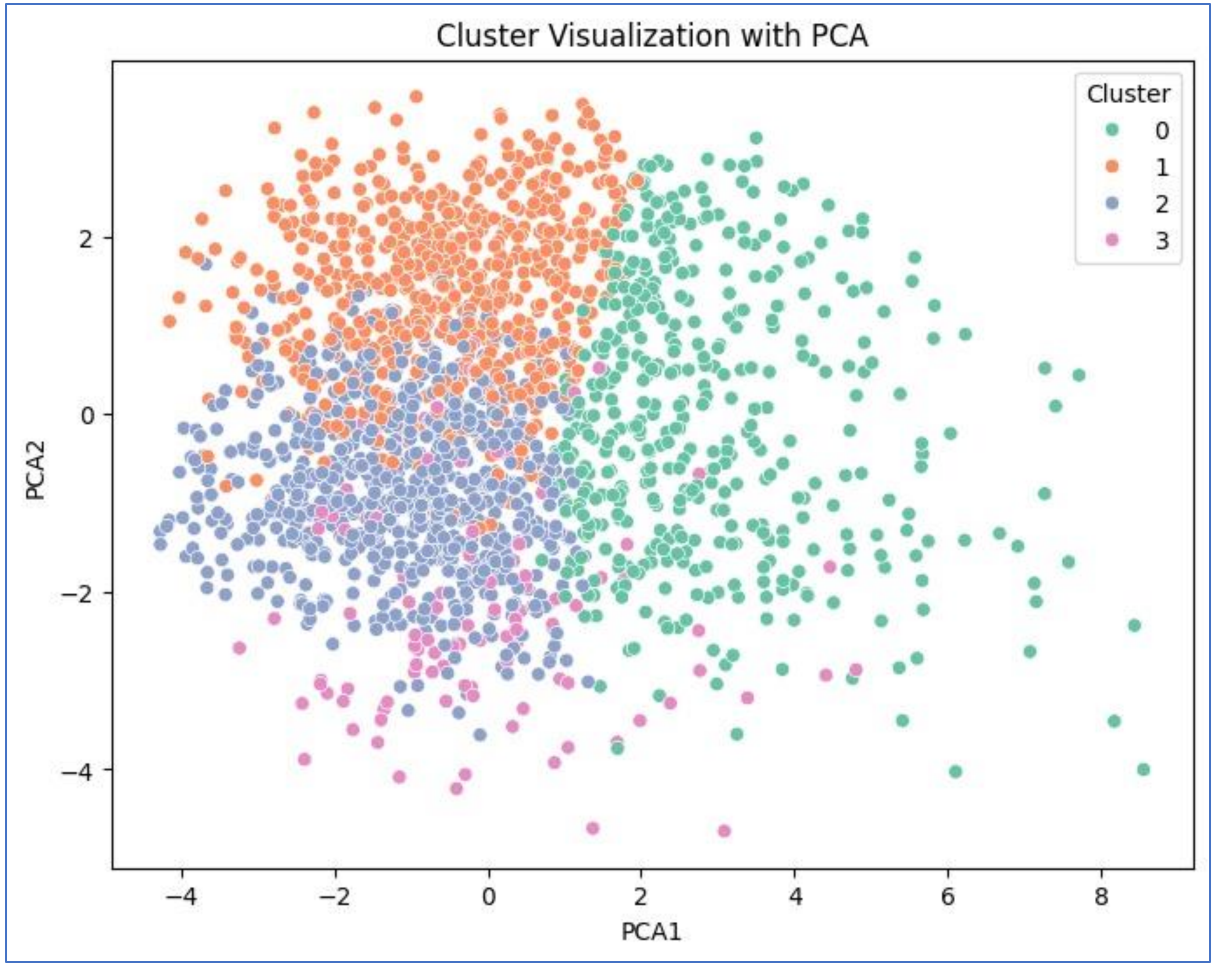
```
Cluster Profile (Top Features):
         DR1TCARB     BMXHT  RIDRETH1_4  RIDRETH1_3  INDFMPIR  DMQMILIZ  \
Cluster
0        0.306031  0.429423    0.589404    0.000000 -0.047628 -0.109047
1       -0.462280 -0.784151    0.138340    0.490119 -0.222852  0.376852
2        0.103060 -0.365277    0.225225    0.387387 -0.223988  0.398192
3        0.293893  0.715485    0.000000    0.990494  0.423525 -0.502596

         RIAGENDR  DR1TPHOS  DR1TPOTA  RIDEXPRG_2.0  DMDHHSZE  LBXWBCSI  \
Cluster
0       -0.588189  0.267432  0.250993           0.0 -0.085298 -0.159475
1        0.816257 -0.501429 -0.469843           0.0  0.218926 -0.001643
2        0.935101 -0.028105 -0.181166           1.0 -1.069765  0.469760
3       -0.699751  0.422387  0.427985           0.0  0.007792  0.086362

         RIDAGEYR  WTMEC2YR
Cluster
0       -0.169785 -0.625412
1        0.376959 -0.192489
2       -1.770964  0.090600
3        0.024745  0.976790
```

PCA Cluster

- Cluster 0: This cluster likely represents younger Non-Hispanic Black females who are taller than average.
- Cluster 1: This cluster likely consists of older males from diverse racial and ethnic backgrounds.
- Cluster 2: This cluster primarily represents young pregnant females with smaller household sizes.
- Cluster 3: This cluster likely represents Non-Hispanic White females of average age with higher income levels.



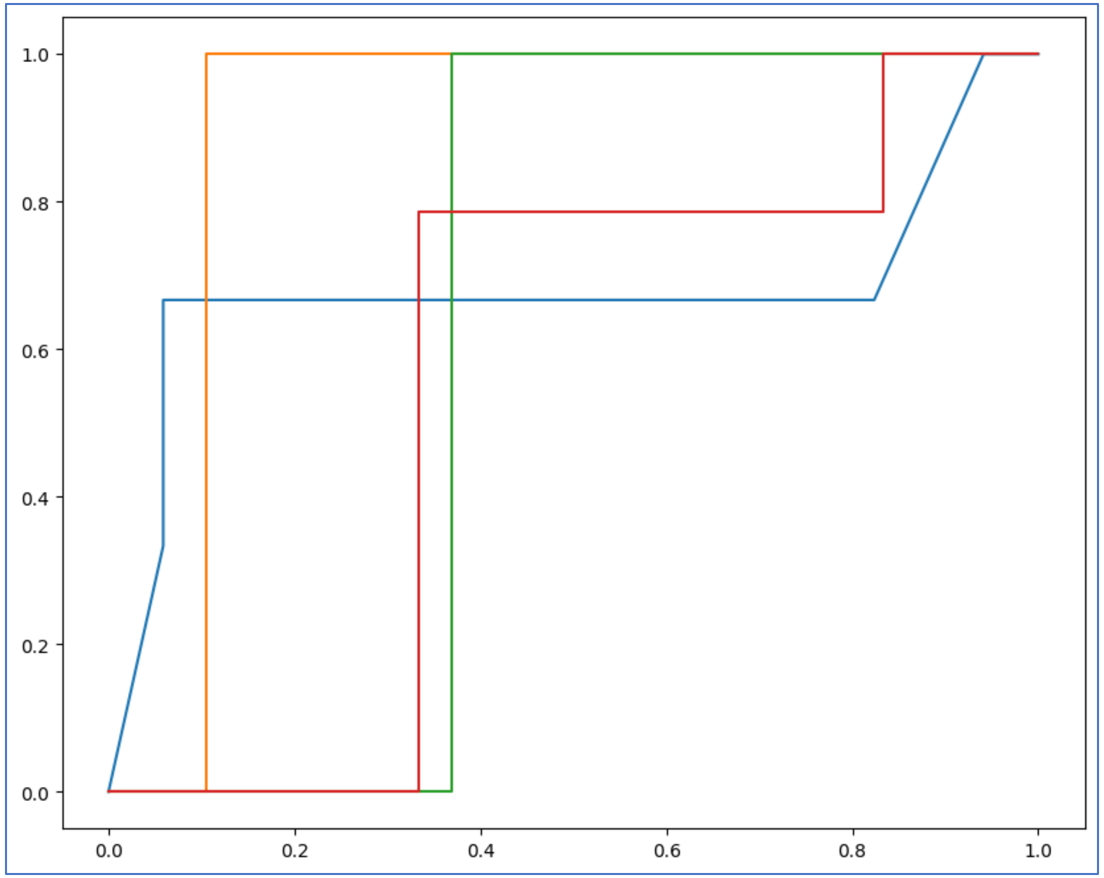
Classification (process)

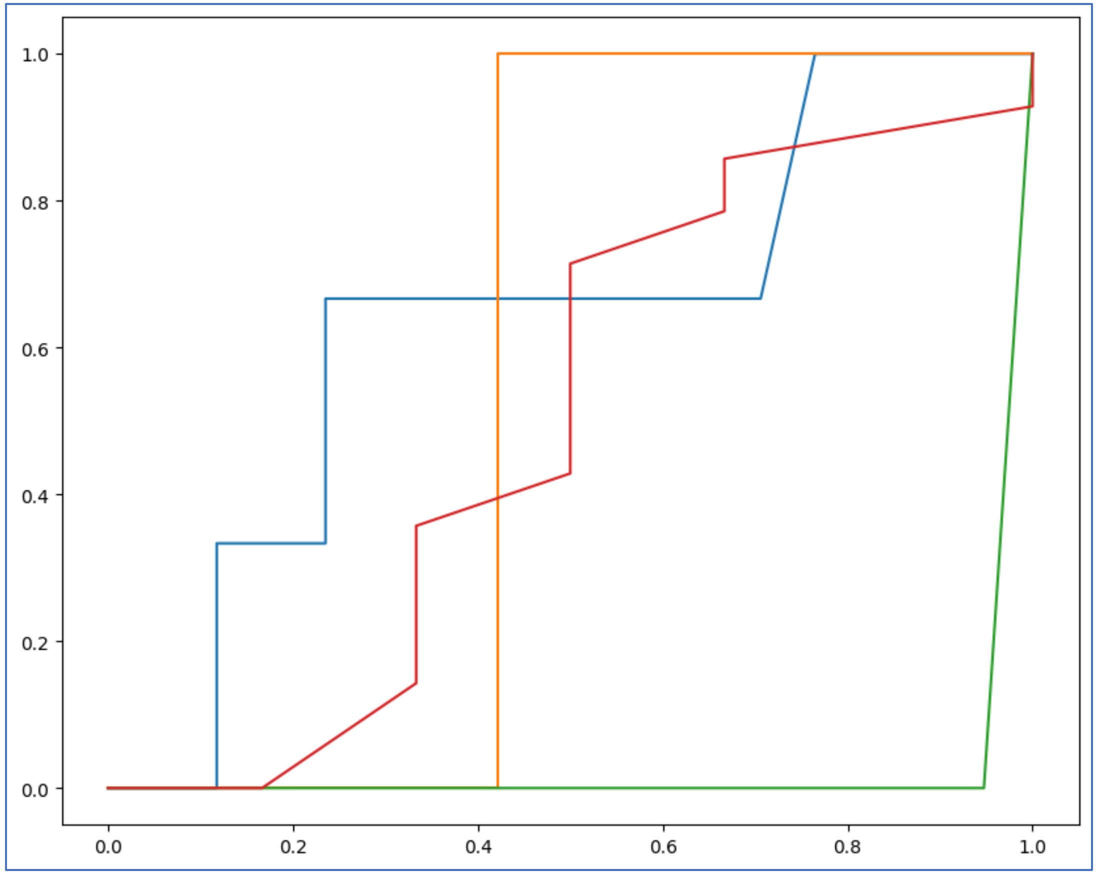
- Objective: Classify the target variable RXDRSC1 using a Random Forest classifier.
- Dataset: Mention key features, missing values handled by median imputation, and target variable encoding.
- Challenge: Address class imbalance and evaluate multi-class performance.

- Model 1 – Baseline Random Forest:
- A Random Forest classifier was trained with n_estimators=100 and random_state=42.
- Key Metrics:
    - Accuracy: 65%.
    - Classification Report: Poor recall and F1-scores for minority classes.
    - Confusion Matrix: Predictions were biased toward the majority class (class 3).
- Feature Importance: Top predictors included *DR1TSUGR* (sugar intake), *LBXTC* (total cholesterol), and *DR1TCALC* (calcium intake).

- Model 2 – Random Forest with Class Balancing:
- Addressed class imbalance using class_weight='balanced'.
- Key Metrics:
    - Accuracy: Remained at 65%.
    - Classification Report and Confusion Matrix: Slight improvements in recall and F1-scores for minority classes.
- Feature Importance: Top predictors shifted slightly, with *DR1TVB6* (vitamin B6 intake), *DR1TSFAT* (saturated fat intake), and *LBDHDD* (HDL cholesterol) emerging as significant.

ROC Curves





(Model 1) ROC Curve Observations:

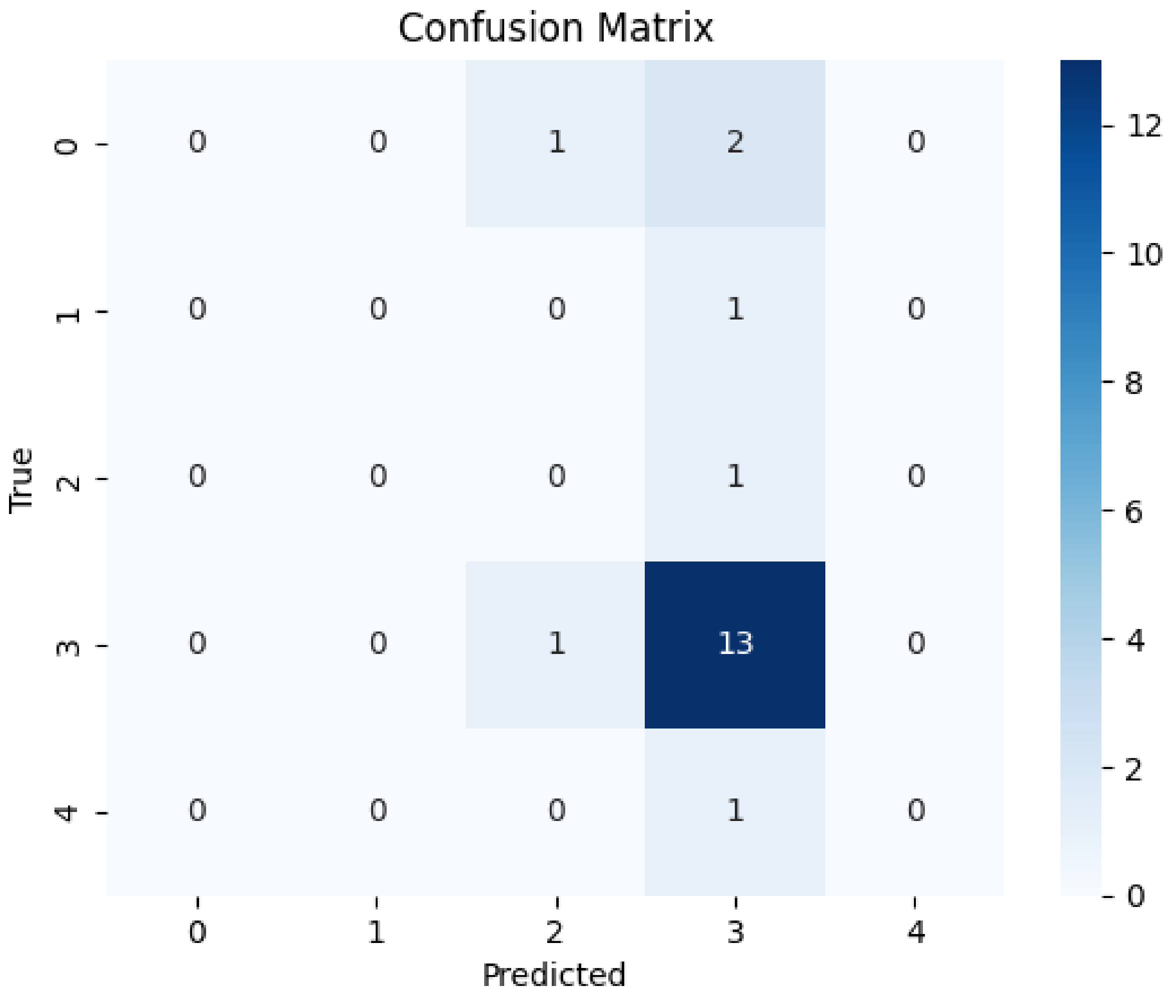
- Abrupt and irregular transitions in the ROC curves indicate inconsistent performance for some classes.
- AUC values suggest low predictive power for minority classes.

(Model 2) ROC Curve Observations:

- Smoother and more consistent curves compared to Model 1.
- Higher AUC values for minority classes, indicating improved performance..

Classification Results

CONFUSION MATRIX:



Confusion Matrix

True \ Predicted	0	1	2	3	4
0	0	0	1	2	0
1	0	0	0	1	0
2	0	0	0	1	0
3	0	0	1	13	0
4	0	0	0	1	0

References

1. Centers for Disease Control and Prevention (CDC). National Health and NutritionExamination Survey (NHANES), 2021-2023;https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023

2. Ambika Satija, Edward Yu, Walter C Willett, Frank B Hu Understanding NutritionalEpidemiology and Its Role in Policy 2015; 10.3945/an.114.007492

3. Sareen S Gropper The Role of Nutrition in Chronic Disease 2023; 10.3390/nu15030664

4. Kimokoti R.W., Millen B.E. Nutrition for the Prevention of Chronic Diseases. Med. Clin.N. Am. 2016;100:1185–1198. doi: 10.1016/j.mcna.2016.06.003

5. Colten HR, Altevogt BM, editors. Sleep Disorders and Sleep Deprivation: An UnmetPublic Health Problem. Washington (DC): National Academies Press (US); 2006

6. Cátia Reis , Sara Dias , Ana Maria Rodrigues , Rute Dinis Sousa , Maria João Gregório ,Jaime Branco , Helena Canhão , Teresa Paiva Sleep duration, lifestyles and chronicdiseases: a cross-sectional population-based study 2018; 10.5935/1984-0063.20180036