B-565 Data Mining | Fall 2024

A Project Report on

Prediction Health Outcomes using Nutritional intake, socioeconomic factors and sleep data

Hanush Kumar *Huy Duc Pham* *Neha Anil Chede*

Abstract

Chronic diseases such as hypertension, type 2 diabetes, asthma, and depression are major public health concerns, with their prevalence increasing worldwide. Traditional healthcare models often focus on treatment rather than prevention, and many individuals remain unaware of their risk for developing such conditions. This study seeks to address this gap by developing a predictive model for health outcomes, using data from NHANES that includes nutritional intake, socioeconomic factors, and sleep patterns. Through machine learning and association rule mining, we aim to uncover hidden relationships between these factors and the onset of chronic diseases. By identifying patterns and classifying individuals into risk groups, the model will provide actionable insights to guide public health interventions and personalized healthcare strategies. The model's performance will be evaluated using metrics like ROC curves and F1 scores to assess its potential in predicting at-risk populations for targeted prevention efforts.

Keywords: Health outcome prediction, nutrition, chronic diseases, socioeconomic factors, sleep patterns, association rules, diabetes, depression, principal component analysis, insomnia, classification, clustering

Table of Contents

1. Introduction

Understanding the relationship between factors such as demographic factors, dietary habits, and biomarkers is critical for the development of preventative medicine and personalized health care strategies. As large-scale health datasets become more accessible, there is a unique opportunity to use this data to better understand how these factors contribute to the prevalence of chronic diseases and health risks. In this study, we aim to investigate the association between demographic variables, dietary habits, and biomarkers using data provided by the Centers for Disease Control and Prevention[1]. Through statistical analysis, association rule mining and predictive modeling, we aim to identify patterns that can help classify the likelihood of individuals contracting specific diseases like essential hypertension, type 2 diabetes mellitus, asthma, major depressive disorder, and insomnia. By leveraging association rule mining, we will uncover strong associations between lifestyle factors and the development of chronic health conditions. By examining high-dimensional data structures and identifying these meaningful patterns, our goal is to provide insights into disease risk factors that could inform public health interventions and strategies for disease prevention.

2. Dataset

The data for this study is sourced from the National Health and Nutrition Examination Survey (NHANES) 2013-2014[1], a nationally representative survey conducted by the Centers for Disease Control and Prevention (CDC). NHANES is designed to assess the health and nutritional status of the U.S. population, combining interviews, physical examinations, and laboratory tests to gather a wide range of health-related data. The survey provides valuable insights into chronic conditions, dietary habits, physical activity, environmental exposures, and socio-economic factors that influence health outcomes.

For the purposes of this study, the NHANES 2013-2014 dataset is used to analyze the effects of **nutritional intake, socio-economic factors**, and **sleep** data on the prediction of health outcomes. The dataset is divided into six primary categories, each containing unique variables that are critical to understanding the relationships between lifestyle factors and the risk of developing chronic diseases. The key categories and variables relevant to this study are outlined below:

Dataset	Variable	Description	Type

Demographics	Age	The age of the participant, in years	Continuous
	Gender	The sex of the participant (e.g., Male, Female)	Categorical
	Race	The racial or ethnic background of the participant (e.g., White, Black, Hispanic, Asian)	Categorical
	Education	The highest level of education completed by the participant (e.g., High school, College)	Categorical
	Household Structure	The type of household (e.g., living alone, with spouse, with children)	Categorical
Dietary	Nutrient Intake	The amount of specific nutrients consumed, such as sodium, vitamin B12, and other vitamins/minerals	Continuous
	Meal Patterns	Information on the number of meals consumed daily and meal timings	Categorical
	Food frequency	The frequency with which certain foods (e.g., fruits, vegetables, processed foods) are consumed	Categorical
Examination	Blood pressure	Blood pressure measurements, including systolic and diastolic values	Categorical
	Weight	Participant's body weight (in kg)	Continuous
	Height	Participant's height (in cm)	Continuous
	BMI	A ratio of weight to height (kg/m^2)	Continuous
	Body Dimensions	Includes leg length, upper arm length, arm circumference, waist circumference, and sagittal abdominal diameters (1 and 2), all measured in centimeters	Continuous
Laboratory	Cholesterol	Blood HDL-cholesterol levels	Continuous
	Insulin	Concentration of insulin in the blood	Continuous
	Lead	Lead concentration in the blood	Continuous
	Cadmium	Cadmium concentration in the blood	Continuous

	Selenium	Selenium levels in the blood	Continuous
Medications	Prescription	The use of prescription medications and the types of medications used	Categorical
	Primary Diagnosis	The primary health condition diagnosed (e.g., hypertension, diabetes)	Categorical
	Secondary Diagnosis	Additional medical conditions diagnosed (e.g., asthma, depression)	Categorical
Questionnaire	Alcohol consumption and frequency	Questions about alcohol use, including if the participant had 12+ drinks in the past year, drinking frequency, and average daily intake over the past 12 months	Categorical/ Continuous
	Medication usage	Whether the participant is taking a prescription for hypertension	Categorical
	Fast food consumption	Frequency of eating at fast food restaurants	Categorical
	Family income	Monthly family income (1 ($0 - $399), 2 ($400 - $799 and so on)	Categorical
	History of Diagnosis	Whether the participant has ever been told they have asthma or arthritis.	Categorical
	Need for Equipment	Whether the participant needs special equipment to walk.	Categorical
	Confusion	Whether the participant experiences confusion or memory problems	Categorical
	Physical activity	Minutes moderate recreational activities	Continuous
	Smoking behaviour	Current smoking status, age of regular smoking onset, and frequency	Categorical/ Continuous

2.1 Data Loading and Filtering

The first step of the analysis involved loading and filtering data from six primary datasets. This process was crucial to ensure the data used for subsequent analysis was both relevant and meaningful.

- **Medications Data**:
    - The medications dataset was loaded from medications.csv. This file contains information about prescriptions associated with individuals, identified by the unique identifier SEQN.
    - To focus on health conditions of interest, we filtered the RXDRSC1 column for specific prefixes:
        - E11 and E11.: Indicating Type 2 Diabetes.
        - E78 and E78.: Indicating disorders of lipoprotein metabolism (e.g., high cholesterol).
        - I10 and I10.: Indicating essential hypertension.
- **Demographic Data**:
    - The demographic dataset was loaded from demographic.csv. Only columns directly relevant to our analysis were selected, including:
        - SEQN: Unique identifier.
        - RIAGENDR: Gender of the individual.
        - RIDAGEYR: Age in years.
        - RIDRETH1: Ethnicity.
        - INDFMPIR: Family income-to-poverty ratio.
        - INDHHIN2: Total household income.
        - DMDEDUC2: Education level.
        - DMDHHSIZ: Household size.
        - RIDEXPRG: Pregnancy status (for females).
    - For RIDEXPRG, missing values for females were replaced with a placeholder value of 3, while all males were assigned 0. This ensured the column was complete and interpretable.
    - Columns with missing values, such as INDFMPIR, INDHHIN2, and DMDEDUC2, were imputed with their respective column means.
- **Dietary Data**:

- The dataset dietary.csv was loaded to analyze nutritional intake. Key variables like total caloric intake and macronutrient consumption were extracted to study their impact on health outcomes.

- **Examination Data**:
  - The examination dataset was loaded from examination.csv. Only columns directly relevant to our analysis were selected, including:
    - BPXSY1: Blood Pressure Systolic 1
    - BPXDI1: Blood Pressure Diastolic 1
    - BPXSY2: Blood Pressure Systolic 2
    - BPXDI2: Blood Pressure Diastolic 2
    - BPXSY3: Blood Pressure Systolic 3
    - BPXDI3: Blood Pressure Diastolic 3
    - BMXWT: Weight
    - BMXHT: Height
    - BMXBMI: Body Mass Index
    - BMDBMIC: BMI Class (Children)
    - BMXLEG: Leg Length
    - BMXARML: Upper Arm Length
    - BMXARMC: Arm Circumference
    - BMXWAIST: Waist Circumference
    - BMXSAD1: Sagittal Abdominal Diameter 1
    - BMXSAD2: Sagittal Abdominal Diameter 2
  - For columns with missing values, the following imputation strategies were applied:
    - BPXSY1, BPXDI1, BPXSY2, BPXDI2, BPXSY3, BPXDI3, BMXWT, BMXHT, BMXBMI, BMXLEG, BMXARML, BMXARMC, BMXWAIST, BMXSAD1, and BMXSAD2: Missing values were imputed using the column mean to maintain the continuity of the data.
    - BMDBMIC: Missing categorical data were imputed with the most frequent category (mode) to preserve the categorical nature of this variable.

- **Laboratory Data**:
  - Laboratory test results were sourced from laboratory.csv, which included biomarkers such as cholesterol levels, fasting glucose, and other critical health

indicators. These values helped in identifying patterns associated with chronic conditions.

- **Questionnaire Data**:
    - The questionnaire dataset was loaded from questionnaire.csv. Only columns directly relevant to our analysis were selected, including:
        - ALQ101: Had at least 12 alcohol drinks/1 yr?
        - ALQ120Q: How often drink alcohol over past 12 months
        - ALQ130: Average number of alcoholic drinks/day over past 12 months
        - BPQ040A: Taking prescription for hypertension
        - DID040: Age when first told you had diabetes
        - CBQ505: Eat at fast food/pizza places
        - CBQ550: Eat at restaurants with waiter
        - DUQ217: How often would you use marijuana?
        - HEQ010: Ever told you have Hepatitis B?
        - IND235: Monthly family income
        - MCQ010: Ever been told you have asthma
        - MCQ160C: Doctor ever said you had arthritis
        - PFQ054: Need special equipment to walk
        - PFQ057: Experience confusion/memory problems
        - PAD675: Minutes of moderate recreational activities
        - PAD680: Minutes of sedentary activity
        - SMD030: Age started smoking cigarettes regularly
        - SMQ040: Do you now smoke cigarettes?
    - For columns with missing values, the following imputation strategies were applied:
        - ALQ120Q, ALQ130, DID040, PAD675, PAD680, and SMD030: Missing values were imputed using the column mean to maintain the continuity of the data.
        - ALQ101, ALQ120U, BPQ040A, CBQ505, CBQ550, DUQ217, HEQ010, IND235, MCQ010, MCQ160C, PFQ054, PFQ057, and SMQ040: Missing categorical data were imputed with the most frequent category (mode) to preserve the categorical nature of these variables.

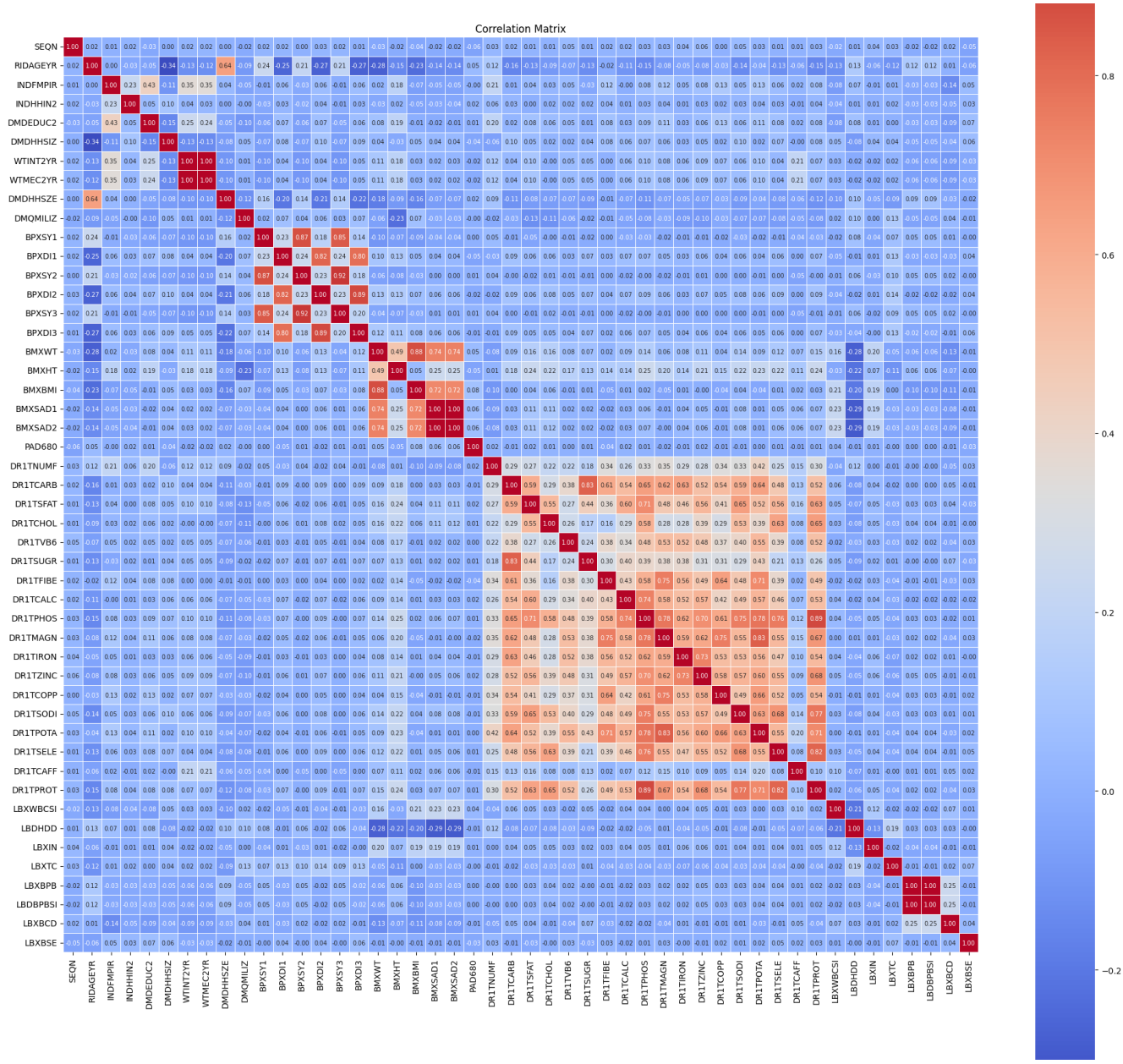
Each dataset was linked using the SEQN identifier to create a unified dataset. This approach ensured that all relevant aspects of an individual's health, behavior, and demographics were captured comprehensively.

2.2 Data Pre-processing:

The StandardScaler from sklearn was used to standardize all numerical features. The motivation behind this step was to address differences in scale among variables. For example, features like RIDAGEYR (age, typically ranging from 0 to 80) and INDFMPIR (income-to-poverty ratio, ranging from 0 to values over 5) were brought to the same scale, allowing clustering and dimensionality reduction techniques to function optimally.

The skewness of several numerical features was examined using the scipy.stats.skew function. Highly skewed variables can distort statistical analyses and clustering algorithms. Variables with excessive skewness, such as INDFMPIR (family income-to-poverty ratio) and INDHHIN2 (total household income), were transformed using the Box-Cox transformation. This method ensures that variables approximate a normal distribution, which enhances the performance of downstream statistical methods. For instance, the skewness of INDFMPIR dropped significantly after the transformation.

We analyzed the correlation matrix to identify and address multicollinearity among variables. Variables with high correlation values were removed to avoid redundancy and improve model performance. Specifically, we dropped variables such as 'WTINT2YR' (Weighting Factor), 'RIDEXPRG_3.0' (Pregnancy Status), 'BPXSY2', 'BPXSY3', 'BPXDI2', 'BPXDI3' (Blood Pressure Measurements), 'BMXBMI' (Body Mass Index), 'BMXSAD1', 'BMXSAD2' (Sagittal Abdominal Diameters), 'DR1TSUGR' (Sugar Intake), 'DR1TPROT' (Protein Intake), and 'LBDBPBSI' (Blood Pressure Biomarker).



3. Methodology and Results

The primary objective is to develop a predictive model that can classify individuals into risk groups based on these factors, enabling early identification of at-risk populations and informing public health interventions. The analysis proceeds through the following key steps: Association Analysis, Principal Component Analysis (PCA), Clustering, and Classification.

3.1 Association analysis

To uncover relationships between key health and biometric attributes, association rule mining was conducted using the Apriori algorithm. The analysis focused on data from health questionnaires, demographic information, and examination results, incorporating factors like lifestyle habits, chronic conditions, and biometric measurements. Using a minimum support threshold of 0.6 and a confidence threshold of 0.8, frequent itemsets and association rules were identified, shedding light on potential predictors of health risks.

The methodology involved preprocessing the data to ensure consistency and accuracy. Variables such as DIQ010, HEQ010, and MCQ010 were recoded to reflect binary choices, where "2" indicates "Yes" (presence of a condition) and "1" indicates "No" (absence of a condition). Continuous demographic variables, including RIDAGEYR (age), INDFMPIR (poverty-income ratio), and WTINT2YR (household income), were categorized into labeled bins (e.g., Low, Medium, High) to allow for discrete analysis. Categorical medication data, such as RXDRSC1_1, was one-hot encoded to create binary columns for each unique category. Data was then subjected to the Apriori algorithm to identify frequently co-occurring conditions and their interrelationships. This process enabled the discovery of association rules, which highlight how the presence of one health condition increases the likelihood of another. By focusing on the antecedents and consequents of these rules, the analysis reveals key insights into the relationships among diabetes, hepatitis, and asthma.

The most significant rules with high confidence and lift are listed below. These rules reveal how the presence of one condition increases the likelihood of another.

Antecedents	Consequents	Support	Confidence	Lift
(DIQ010_2.0) — Diabetes	(HEQ010_2.0) — Hepatitis	0.7441	0.8564	1.0276
(HEQ010_2.0) — Hepatitis	(DIQ010_2.0) — Diabetes	0.7441	0.8929	1.0276
(MCQ010_2.0) — Asthma	(DIQ010_2.0) — Diabetes	0.7341	0.9084	1.0455
(DIQ010_2.0) — Diabetes	(MCQ010_2.0) — Asthma	0.7341	0.8448	1.0455
(MCQ010_2.0) — Asthma	(HEQ010_2.0) — Hepatitis	0.6936	0.8583	1.0299
(HEQ010_2.0) — Hepatitis	(MCQ010_2.0) — Asthma	0.6936	0.8323	1.0299
(DIQ010_2.0, HEQ010_2.0) —	(MCQ010_2.0) — Asthma	0.6210	0.8346	1.0329

Diabetes + Hepatitis				
(MCQ010_2.0, HEQ010_2.0) — Asthma + Hepatitis	(DIQ010_2.0) — Diabetes	0.6210	0.8954	1.0305
(MCQ010_2.0, DIQ010_2.0) — Asthma + Diabetes	(HEQ010_2.0) — Hepatitis	0.6210	0.8460	1.0153

The association rules reveal meaningful relationships between chronic health conditions, providing insights into how co-occurring conditions may predict health risks. The strong link between diabetes (DIQ010_2.0) and hepatitis (HEQ010_2.0), supported at 74.41% with a confidence of 85.64%, suggests a non-random, meaningful relationship between these conditions. Similar patterns are seen with asthma (MCQ010_2.0), which is associated with both diabetes and hepatitis. These findings indicate that individuals with diabetes are more likely to exhibit signs of hepatitis or asthma, underscoring the potential need for comprehensive health monitoring for individuals with these chronic conditions. The rules involving combinations of antecedents—such as (DIQ010_2.0, HEQ010_2.0) leading to (MCQ010_2.0)—highlight the compounded influence of multiple health conditions on the development of asthma. This observation implies that co-occurring chronic conditions have a synergistic effect, raising the likelihood of developing a third health condition.

The association rules highlight key relationships between diabetes, hepatitis, and asthma, with shared co-occurrence patterns that suggest possible links in underlying health risk factors. These relationships will be further explored using Principal Component Analysis (PCA), which reduces the dimensionality of the dataset and identifies key components that explain the most variance. By visualizing how variables like DIQ010, HEQ010, and MCQ010 cluster within principal components, PCA will provide a deeper understanding of the shared dimensions that underlie these chronic health conditions.

3.2 Principal Component Analysis

Given the high dimensionality of the dataset, PCA was performed to reduce the number of variables while retaining the maximum variance in the data. This dimensionality reduction

technique helps eliminate noise and multicollinearity, making the dataset more manageable and suitable for subsequent analysis. PCA was applied to reduce the dataset's dimensionality while capturing 95% variance.

PCA Output -

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

Observations:

Dominant Variables: The top features identified by PCA include BMXHT (Height), INDFMPIR (Income-to-Poverty Ratio), RIDAGEYR (Age), and WTMEC2YR (Weighting Factor). These variables contribute significantly across the principal components:

BMXHT: Height emerges as a strong driver of variance, likely reflecting its association with demographic and health-related patterns.

INDFMPIR: Continues to stand out as a dominant feature, highlighting its socioeconomic importance in explaining variance.

RIDAGEYR: Age remains critical, reflecting its influence on age-dependent health and lifestyle factors.

WTMEC2YR: The weighting factor is influential, likely due to its importance in survey design and population representation.

Several categorical features play a significant role in explaining variance in the dataset. For instance, RIDRETH1_3 and RIDRETH1_4, likely representing ethnic or racial groups, highlight the importance of demographic diversity as a key component of variability. DMQMILIZ (Military Status) also contributes significantly, reflecting its association with specific demographic or lifestyle factors. Additionally, RIDEXPRG_2.0 (Pregnancy Status) emerges as an influential variable, possibly tied to household dynamics and health-related patterns.

Health and nutritional variables further stand out as critical contributors to variance. For example, DR1TPOTA (Potassium Intake) and DR1TPHOS (Phosphorus Intake) underscore the pivotal role of nutrition, while DR1TCARB (Carbohydrate Intake) highlights dietary habits as a key factor. Moreover, LBXWBCSI (White Blood Cell Count) emphasizes the significance of biological markers, reinforcing the importance of health-related data in explaining variability across the dataset.

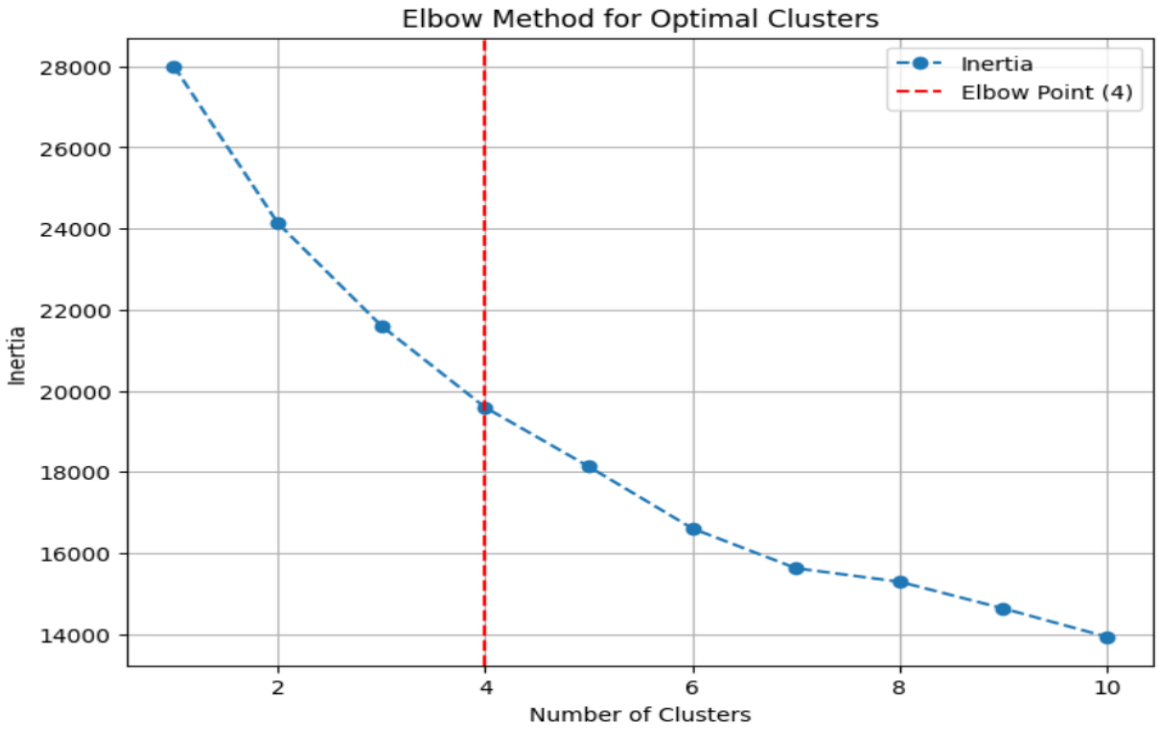
The output of the association analysis directly informed the PCA process by guiding variable prioritization and interpretation. The strong relationships identified between chronic conditions like Diabetes (DIQ010_2.0), Hepatitis (HEQ010_2.0), and Asthma (MCQ010_2.0)—evident from their high support, confidence, and lift values—highlighted these as key features in the dataset. This ensured that health-related variables were retained and explored further during PCA. The observed co-occurrence of these conditions suggested shared variability, which PCA captured by grouping related variables within principal components.

Additionally, while the association rules focused on direct relationships, PCA revealed broader patterns involving influential features like DR1TPOTA (Potassium Intake), LBXWBCSI (White Blood Cell Count), and RIDAGEYR (Age), which may indirectly influence the chronic conditions identified earlier. The lift values greater than 1.0 in the association analysis validated the non-random nature of these connections, allowing for a more meaningful interpretation of PCA outputs. Together, the association analysis and PCA provided complementary insights, with association rules identifying specific co-occurrences and PCA expanding on these findings to uncover shared variance among key demographic, health, and nutritional variables.

3.3 Cluster Analysis

After performing PCA, K-means clustering was applied to group individuals based on the patterns identified in the data. Clustering helps identify natural groupings within the dataset, revealing how demographic, socio-economic, and lifestyle factors relate to each other. This step also provides an exploratory view of the data and can be used to better understand how different groups of individuals are at varying levels of risk for specific health outcomes.

```
Optimal number of clusters: 4
```



The Elbow Method plot clearly identifies 4 clusters as the optimal number for this dataset. Inertia, which measures the sum of squared distances between data points and their respective cluster centers, decreases as the number of clusters increases. However, the rate of decrease slows significantly after 4 clusters, where the "elbow point" is observed. This point, marked by the red dashed line, indicates the balance between reducing inertia and avoiding overfitting. Beyond this, adding more clusters yields diminishing returns with minimal improvements in inertia.

The Cluster Profile analysis highlights distinct patterns across the top features, helping to characterize the four identified clusters:

Cluster	BMXHT	RIDRETH1_3	DR1TPOTA	DMQMILIZ	RIDRETH1_4	WTMEC2YR	RIAGENDR	DR1TPHOS	RIDEXPRG_2.0	INDFMPIR	RIDAGEYR	DR1TCARB	DMDHHSZE	LBXWBCSI
0	0.429423	0.000000	0.250993	-0.109047	0.589404	-0.625412	-0.588189	0.267432	0.0	-0.047628	-0.169785	0.306031	-0.085298	-0.159475
1	-0.784151	0.490119	-0.469843	0.376852	0.138340	-0.192489	0.816257	-0.501429	0.0	-0.222852	0.376959	-0.462280	0.218926	-0.001643
2	-0.365277	0.387387	-0.181166	0.398192	0.225225	0.090600	0.935101	-0.028105	1.0	-0.223988	-1.770964	0.103060	-1.069765	0.469760
3	0.715485	0.990494	0.427985	-0.502596	0.000000	0.976790	-0.699751	0.422387	0.0	0.423525	0.024745	0.293893	0.007792	0.086362

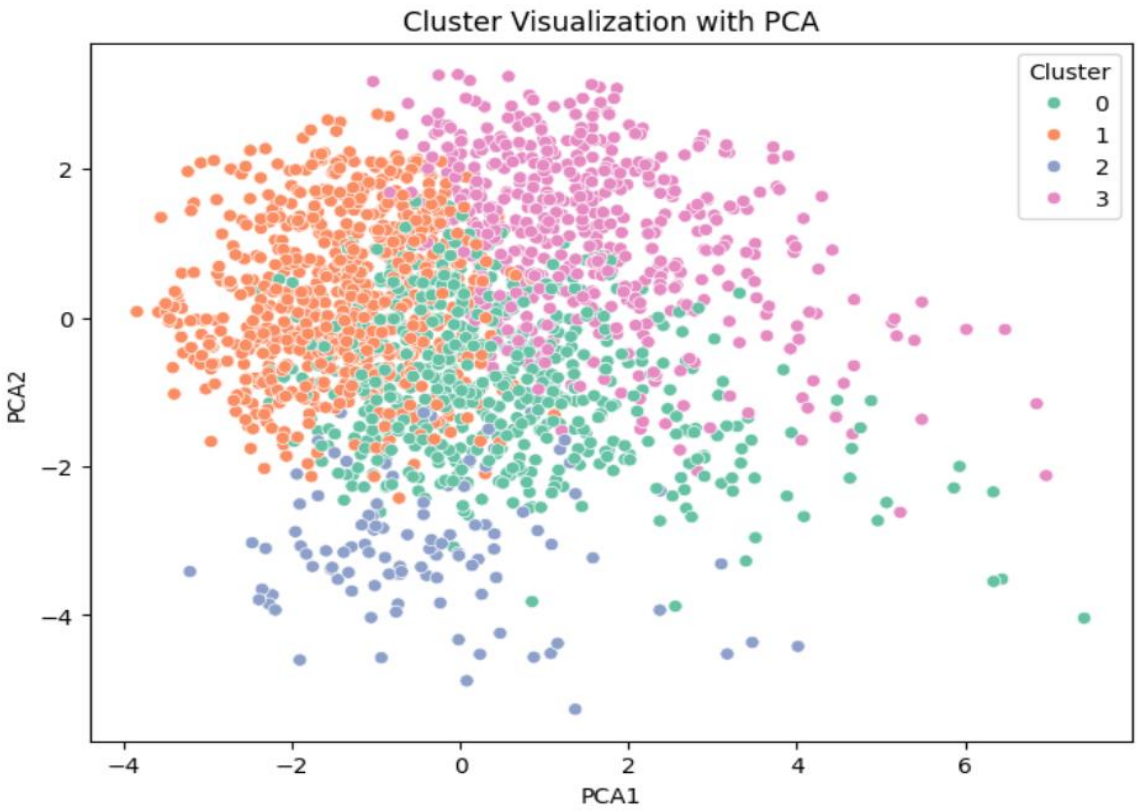
Cluster 0: This cluster is characterized by higher BMXHT (Height) and DR1TPOTA (Potassium Intake) values. It also shows positive contributions for RIDRETH1_4 (Ethnic Group) and DR1TPHOS (Phosphorus Intake). However, it has low WTMEC2YR (Weighting Factor) and

negative RIAGENDR (Gender), suggesting that individuals in this group may belong to a specific demographic with moderate nutritional attributes and underrepresentation in weighting factors.

Cluster 1: Cluster 1 stands out with lower BMXHT (Height) and DR1TPOTA (Potassium Intake), alongside higher RIAGENDR (Gender) and DMQMILIZ (Military Status). It also exhibits a negative DR1TPHOS (Phosphorus Intake) and DR1TCARB (Carbohydrate Intake). This profile suggests a group with shorter height, low potassium intake, and a notable military demographic presence.

Cluster 2: This cluster is defined by high RIAGENDR (Gender) and RIDEXPRG_2.0 (Pregnancy Status), which has a value of 1.0, indicating pregnancy is a defining attribute. It also shows low RIDAGEYR (Age) and DMDHHSZE (Household Size). The positive LBXWBCSI (White Blood Cell Count) highlights a health-related component, suggesting younger individuals with pregnancy status and moderate white blood cell counts.

Cluster 3: Cluster 3 is dominated by high BMXHT (Height), RIDRETH1_3 (Ethnic Group), and WTMEC2YR (Weighting Factor). It also has positive contributions for DR1TPHOS (Phosphorus Intake) and DR1TCARB (Carbohydrate Intake). This indicates a group with taller individuals, high ethnic diversity, and strong dietary attributes, particularly phosphorus and carbohydrate intake.



The PCA plot provides a clear visualization of how the clusters are distributed across the two most significant principal components. While Cluster 1 and Cluster 2 are more compact and distinct, Clusters 0 and 3 show broader spread and overlap, suggesting a more complex relationship between their underlying features and we can adopt tSNE to do deeper cluster analysis on a local level.

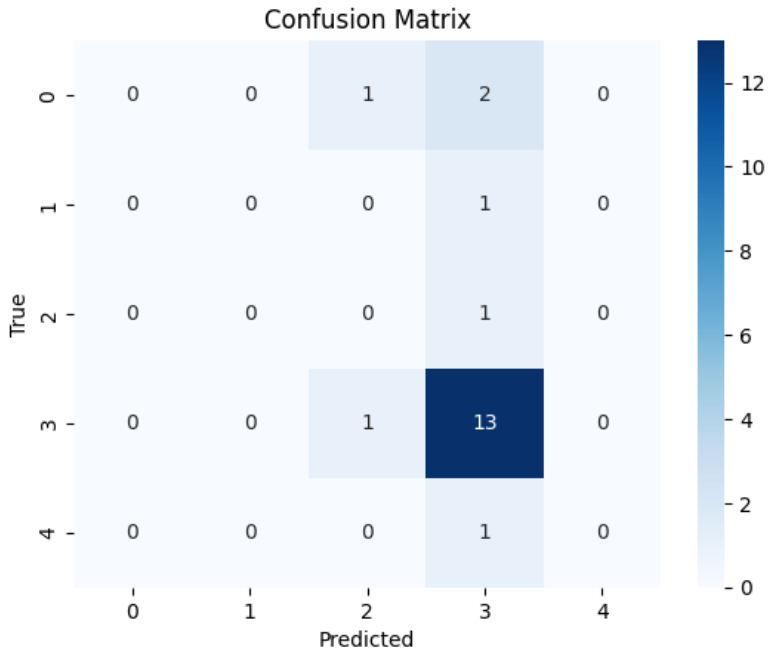
3.4 Classification

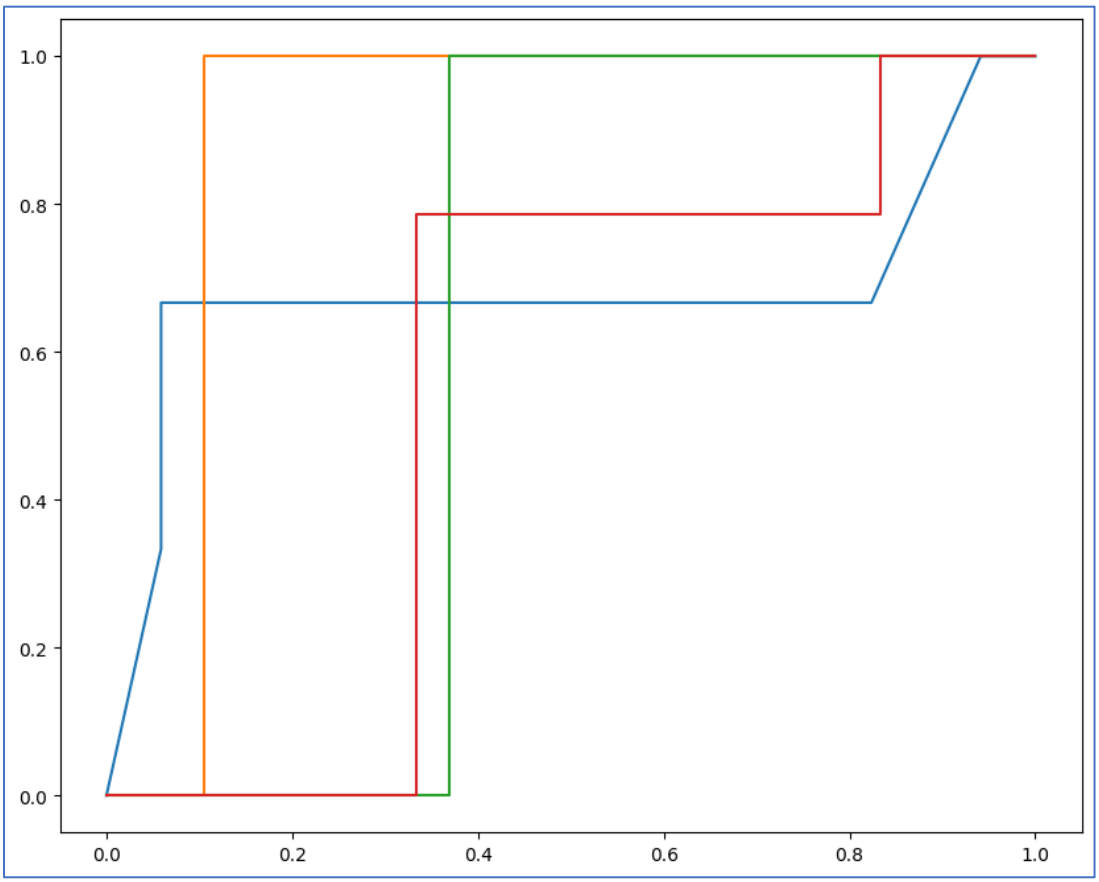
To classify individuals into risk categories (e.g., low, high, very-high risk), we utilized Random Forest, an ensemble learning method known for handling both categorical and continuous variables and capturing complex relationships between predictors. The model aimed to predict the likelihood of chronic conditions such as hypertension, diabetes, or asthma based on various features derived from socio-economic, dietary, and behavioral factors. Cross-validation was used to tune model parameters, with performance assessed using key metrics such as ROC curve, Precision, Recall, and F1-score to ensure reliable predictions.

The classification process involved two main stages: first, a baseline Random Forest model without addressing class imbalance, followed by an improved model incorporating class balancing. The target variable (RXDRSC1) represents the presence or absence of specific chronic diseases, categorized into multiple risk levels (e.g., low, medium, high risk). The dataset was appropriately encoded to allow the model to process both categorical and continuous features effectively, optimizing for multi-class classification.

Model 1: Baseline Random Forest

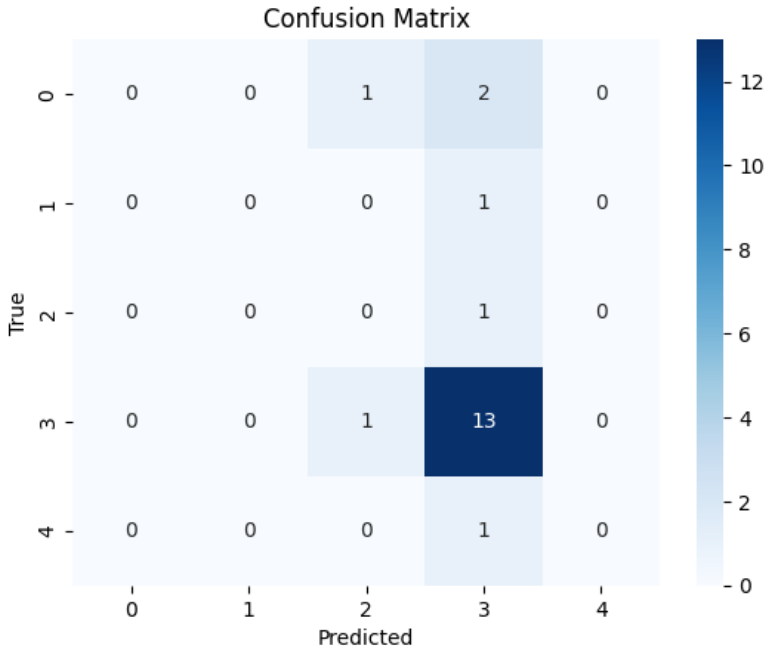
In the first stage, a baseline Random Forest model was trained with 100 trees (n_estimators=100) and a fixed random seed (random_state=42) for reproducibility. This model achieved 65% accuracy but struggled with minority classes. The classification report highlighted poor recall and F1-scores for these underrepresented classes, indicating a bias toward the majority class (class 3). The confusion matrix confirmed that most predictions were clustered in the majority class, leading to significant misclassifications for the minority categories.

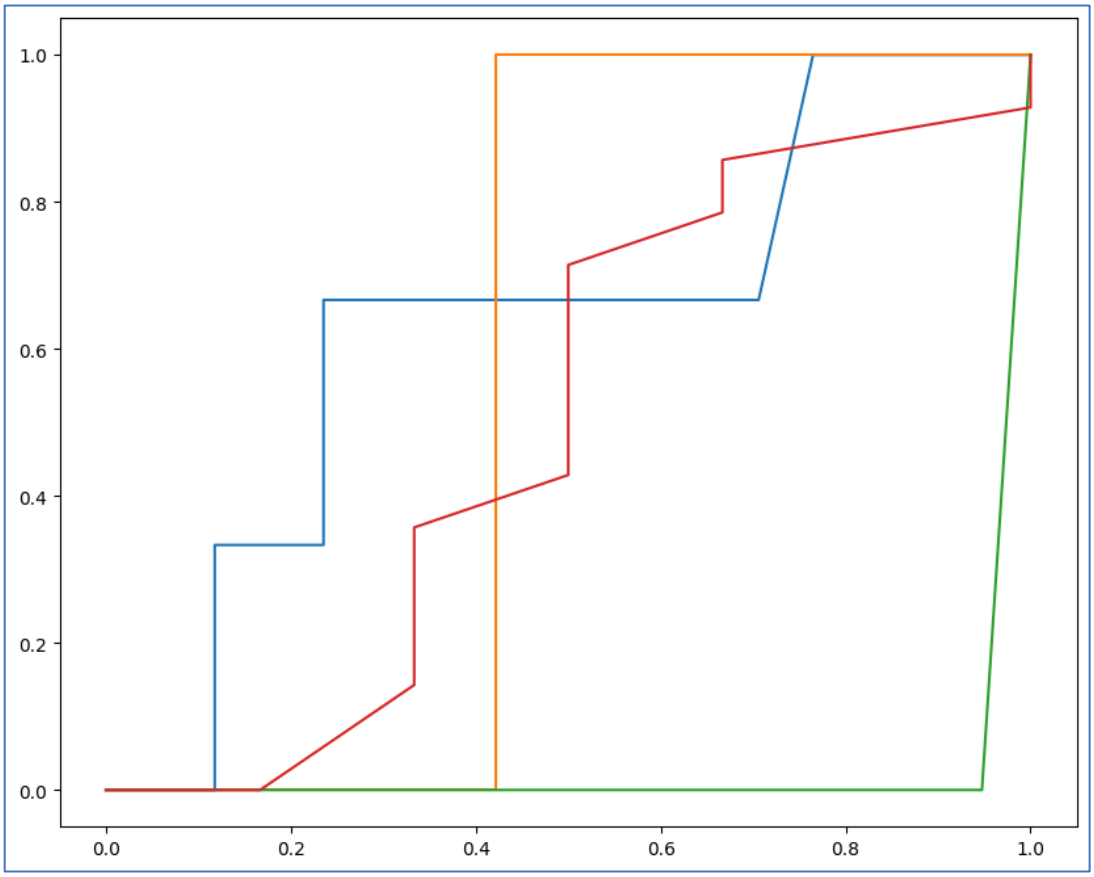




Model 2: Random Forest with Class Balancing

To address the class imbalance issue, we introduced class balancing in the second model using the class_weight='balanced' parameter in the Random Forest algorithm. Although the overall accuracy remained at 65%, this model showed slight improvements in performance for the minority classes. The classification report indicated better recall and F1-scores for these underrepresented groups, while the confusion matrix revealed a reduction in misclassifications.





Both models were evaluated using Receiver Operating Characteristic (ROC) curves. For Model 1, the ROC curves were irregular, especially for minority classes, and the AUC values were low,

indicating poor predictive power for these underrepresented categories. In contrast, Model 2 showed smoother and more consistent curves, with higher AUC values for the minority classes, reflecting the improvement in performance after applying class balancing.

4. Conclusion

In this study, we utilized multiple analytical techniques to explore the relationship between socio-economic factors, dietary habits, and chronic disease risk. The association analysis revealed significant patterns between lifestyle factors and disease risk, highlighting key associations like the links between diabetes (DIQ010_2.0), hepatitis (HEQ010_2.0), and asthma (MCQ010_2.0).

Additionally, we successfully applied Principal Component Analysis (PCA), and clustering techniques to explore the relationships between socio-economic factors, nutritional intake, and chronic disease risks. PCA further reduced data dimensionality while retaining 95% of the variance, identifying dominant contributors such as BMXHT (Height), RIDAGEYR (Age), and DR1TPOTA (Potassium Intake).

The clustering analysis uncovered four distinct groups with unique demographic and nutritional profiles, offering actionable insights for public health strategies. For example, Cluster 2 identified younger individuals with pregnancy-related attributes, suggesting the need for targeted healthcare monitoring for maternal health. Similarly, the dietary patterns observed in Cluster 3 highlight the importance of nutritional interventions for individuals with strong dietary influences.

For classification, the Random Forest classifier provided useful insights into chronic disease risk, though challenges like class imbalance persisted. While the model achieved 65% accuracy, it struggled with underrepresented classes. Class balancing slightly improved performance for these minority groups, but further optimization is needed. The feature importance analysis identified key predictors such as sugar intake, cholesterol levels, and socio-economic factors, which align with existing literature on chronic disease risk.

5. Discussion

The association analysis revealed key patterns in the co-occurrence of chronic health conditions, providing insight into how certain conditions are linked. Using the Apriori algorithm, frequent

itemsets and association rules were identified, highlighting strong relationships between diabetes (DIQ010), hepatitis (HEQ010), and asthma (MCQ010). This categorization made it possible to identify patterns such as higher asthma prevalence among certain age groups when combined with the presence of diabetes and hepatitis. By recoding categorical data into binary indicators and categorizing continuous demographic variables into meaningful ranges, the association analysis uncovered complex interdependencies among chronic conditions, offering valuable perspectives on how multiple health risks may be interconnected.

The Principal Component Analysis (PCA) and clustering revealed significant patterns within the dataset that provide valuable insights into the interplay between demographic, nutritional, and health-related factors. For example, the identification of younger individuals with pregnancy status in Cluster 2 highlights the importance of tracking health indicators for specific subgroups that may require targeted interventions. Similarly, the presence of strong dietary attributes like phosphorus and carbohydrate intake in Cluster 3 aligns with existing research that associates these nutrients with metabolic health and chronic disease risks.

The findings also underscore the importance of ethnic diversity (features such as RIDRETH1_3 and RIDRETH1_4) and its correlation with distinct health and socioeconomic patterns. For instance, Cluster 1's lower height and potassium intake, coupled with a notable military status (DMQMILIZ), may reflect occupational or lifestyle-driven factors affecting health outcomes. These results emphasize how PCA and clustering complement each other to uncover hidden structures in high-dimensional data and suggest meaningful groupings that may not be obvious through traditional statistical methods. By interpreting these clusters within a broader health context, we gain a deeper understanding of risk factors that can inform public health policies and strategies for targeted healthcare interventions.

The classification analysis revealed that dietary factors such as sugar consumption (DR1TSUGR), total cholesterol (LBXTC), and calcium intake (DR1TCALC) were the most important predictors of chronic disease risk. These findings support the well-established role of diet in the development of conditions like hypertension, diabetes, and cardiovascular diseases. After addressing class imbalance in Model 2, Vitamin B6 intake (DR1TVB6), saturated fat intake (DR1TSFAT), and HDL cholesterol (LBDHDD) emerged as significant predictors. These results highlight the importance of both dietary habits and metabolic health markers in chronic disease risk.

6. Future Work

Future research could improve the classification model's performance by exploring advanced techniques such as hyperparameter tuning and experimenting with alternative machine learning algorithms like Gradient Boosting or Neural Networks. Incorporating longitudinal or time-series data would allow for a more dynamic understanding of how lifestyle factors evolve and influence chronic disease risk over time. Expanding the feature set to include genetic information or more detailed health data could further refine predictions. Finally, addressing class imbalance through more advanced methods like SMOTE could improve the model's ability to accurately classify underrepresented risk categories.

7. References

- Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey (NHANES), 2021-2023; https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023
- Ambika Satija, Edward Yu, Walter C Willett, Frank B Hu Understanding Nutritional Epidemiology and Its Role in Policy 2015; 10.3945/an.114.007492
- Sareen S Gropper The Role of Nutrition in Chronic Disease 2023; 10.3390/nu15030664
- 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
- Colten HR, Altevogt BM, editors. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. Washington (DC): National Academies Press (US); 2006
- Cátia Reis , Sara Dias , Ana Maria Rodrigues , Rute Dinis Sousa , Maria João Gregório , Jaime Branco , Helena Canhão , Teresa Paiva Sleep duration, lifestyles and chronic diseases: a cross-sectional population-based study 2018; 10.5935/1984-0063.20180036