Clustering Analysis on the National Health and Nutrition Examination Survey(NHANES) Dataset

Saleha Muzammil
Haiqa Abdul Rauf

May 8, 2024

1 Introduction

This project aims to apply clustering analysis techniques to the National Health and Nutrition Examination Survey (NHANES) dataset to identify patterns in dietary habits, physical activity, and their correlations with various health outcomes. Our goal is to provide insights that can aid in the formulation of targeted nutritional guidelines and public health interventions.
The importance of this project lies in its potential to improve public health outcomes. By identifying patterns and correlations between dietary habits, physical activities, and diseases, we can develop targeted interventions and personalized health recommendations. This can lead to a reduction in the prevalence of various illnesses and an overall improvement in population health.

2 Mining of Massive Data Techniques with PySpark

We employ PySpark, the Apache Spark's Python API, to manage and analyze the large-scale NHANES dataset. PySpark is ideal for processing large datasets efficiently across distributed systems with its in-memory computing and extensive machine learning library, MLlib. Here's how we will use PySpark:

- **K-means Clustering with PySpark:** PySpark's MLlib offers a scalable implementation of the K-means algorithm, crucial for efficiently clustering large datasets. We will segment the NHANES dataset into groups with similar health and nutrition characteristics, specifying the number of clusters and initialization mode to enhance meaningful segmentation.
- **Bisecting K-means Clustering with PySpark:** Using the BisectingKMeans algorithm, an alternative to traditional hierarchical clustering, we will further analyze and validate clusters formed by K-means, providing a dendrogram for visual examination.
- **Principal Component Analysis (PCA) with PySpark:** Employ PCA for dimensionality reduction before clustering to reduce the dataset into principal components, facilitated by PySpark's MLlib on large datasets.

Data Preprocessing and Management: Prior to clustering, we will implement significant preprocessing steps using PySpark's DataFrame API, including handling missing values, normalizing data, and encoding categorical variables.

Visualizations and Insights Extraction: Generate visualizations directly from Spark DataFrames using integrations like Plotly and Databricks notebooks to facilitate immediate insights and interpretation of clustering results.

3 Project Objective

- Identify common dietary and physical activity patterns among U.S. populations.
- Discover correlations between these patterns and specific health outcomes.
- Provide a data-driven basis for targeted nutritional guidelines and public health interventions.

4 Related Work

While our aim is to work on the latest NHANES dataset, previous work has been done on mining of NHANES data of 1999- 2008. This paper [3] mines the NHANES data of 1999-2008 to extract health and nutrition patterns. The paper adapts and extends data mining techniques, including association rule mining and clustering algorithms, to extract knowledge about diabetes and high blood pressure from NHANES data. The methodological approaches in the paper can provide insights or frameworks that we might adopt for our clustering analysis. Another study [2] explores the association between dietary patterns and the risk of developing type 2 diabetes among adults in the U.S. using a clustering approach. It identifies specific dietary patterns that are significantly associated with an increased risk of diabetes. This study [1] provides a precedent for using clustering to analyze dietary patterns in relation to health outcomes. Our project aims to extend this approach by including a broader range of health outcomes and incorporating additional variables like physical activity levels. This study employed clustering analysis to identify patterns of unhealthy behaviors among participants in the Aerobics Center Longitudinal Study. The behaviors analyzed included smoking, physical inactivity, poor diet, and excessive alcohol consumption. Similar to this research, our project intends to use clustering techniques on the NHANES dataset to uncover patterns of dietary and physical activity behaviors and their relationships with various health metrics. This study offers a methodological framework that we can adapt to analyze how combined lifestyle factors influence health outcomes in a diverse, national sample. This paper's approach to clustering and its focus on comprehensive lifestyle assessment can provide valuable insights into designing our analysis and interpreting complex behavioral data within NHANES.

5 Dataset Details

Dataset: National Health and Nutrition Examination Survey (NHANES)

- **Source:** Centers for Disease Control and Prevention (CDC)
- **Size:** Approximately 5,000 participants annually
- **Processing with PySpark:** The dataset will be loaded into PySpark DataFrames, allowing for distributed data processing across the Spark cluster. This includes handling demographic, dietary, physical examination, and laboratory data.

6 Methodology

Our methodology encompasses:

- Initializing a **Spark** session and creating **Spark DataFrames**.
- Visualizing data to understand key attributes.
- **Pre-processing data in Spark**: handling missing values, encoding categorical variables, and performing data normalization.

- Using **inner joins in Spark** to amalgamate multiple CSV files into a comprehensive DataFrame.
- Employing **broadcasting** for efficient data merging of smaller files.
- Analyzing column correlations to identify meaningful relationships.
- Applying **PCA** for dimensionality reduction and **clustering using K-Means and Bisecting K-Means from PySpark's MLlib**. Metrics like **silhouette coefficient and sum of squared errors** are calculated to determine optimal clustering.
- Comparing clustering outcomes and visualizing in 3D for enhanced data interpretation.
- **Deploying** our work on **Databricks** platform and creating **Spark clusters and jobs**.

7 Results

To compare and determine the best clustering method between KMeans and Bisecting KMeans, we can analyze the Silhouette scores for different values of K and PCA components.

7.1 KMeans Clustering

- Best Silhouette Score: 0.1849 (2 PCA components, K = 2)
- Average Silhouette Score: -0.2191 (average of all scores)

7.2 Bisecting KMeans Clustering

- Best Silhouette Score: 0.3988 (K = 2)
- Average Silhouette Score: 0.1898 (average of all scores)

Based on the average Silhouette scores, the Bisecting KMeans method out- performs the KMeans method. However, it's important to note that the best performing KMeans configuration (2 PCA components, K = 2) achieves a higher Silhouette score than any individual Bisecting KMeans configuration. Therefore, if the goal is to achieve the best possible clustering for this specific dataset, the KMeans method with 2 PCA components and K = 2 should be chosen. If a more generalized approach is needed, where overall performance is more important than finding the absolute best configuration, the Bisecting KMeans method may be preferred due to its higher average Silhouette score

8 Conclusion

Utilizing PySpark's capabilities, this project effectively manages the expansive NHANES dataset to uncover pivotal health and nutrition patterns. These findings contribute to understanding the relationship between lifestyle factors and health outcomes, guiding public health strategies.

References

[1] Mariane Héroux, Ian Janssen, Duck-chul Lee, Xuemei Sui, James R Hebert, and Steven N Blair. Clustering of unhealthy behaviors in the aerobics center longitudinal study. *Prevention Science*, 13:183–195, 2012.

[2] Rob M van Dam, Eric B Rimm, Walter C Willett, Meir J Stampfer, and Frank B Hu. Dietary patterns and risk for type 2 diabetes mellitus in us men. *Annals of internal medicine*, 136(3):201–209, 2002.

[3] Jun won Lee and Christophe Giraud-Carrier. Results on mining nhanes data: A case study in evidence-based medicine. *Computers in biology and medicine*, 43(5):493–503, 2013.