Early Disease Prediction

Section A1 - Team 1

Rebecca Bubis, Sanjal Desai, Yashna Meher, Mishil Trivedi

5 Jan 2025

**1. Motivation**The increasing prevalence of chronic diseases such as obesity, diabetes, and cardiovascular conditions has created a public health challenge. Understanding the demographic, dietary, and health-related factors influencing these conditions is crucial for policymakers, healthcare professionals, and researchers. The NHANES post pandemic dataset provides a unique opportunity to analyze these factors through a combination of interview data and physical examination results. Late-stage diagnoses of chronic or critical illnesses (e.g. Stroke and liver disease) lead to higher treatment costs, poorer patient outcomes, and increased mortality rates. Early detection can significantly improve patient prognosis and reduce healthcare costs.

**2. Business Relevance**

* Proactive Care
  + Clustering techniques can group patients into high and low risk categories based on different factors including but not limited to demographics, lifestyle, labs, medical history, etc. Patient risk segmentation can minimize hospital readmission rates and assist with targeted screening programs by allowing for early detection of diabetes, liver disease, and cardiovascular conditions.
* Cost Savings
  + Late diagnosis of disease results in higher treatment costs. Early detection through proactive care can minimize these costs and reduce additional costs incurred through readmission rates.
* Improved Outcomes
  + The proactive care achieved from these clustering techniques can enhance patient quality of life and survival rates due to early detection in targeted screening programs that stem from it.

**3. Problem Statement**Early warning signs often go unnoticed due to the lack of structured screening processes, leaving healthcare professionals to rely on subjective assessments. Additionally, traditional diagnostic methods often rely on manually intensive processes that may miss underlying patterns in health data. Machine learning techniques offer an opportunity to uncover hidden trends, detect correlations between demographic and lifestyle factors, and improve the identification of high-risk individuals. Our project aims to apply unsupervised machine learning to analyze symptom-based medical records, identifying patient clusters that exhibit early-stage disease patterns. This will enable healthcare providers to intervene proactively and create a preventative environment. By doing so, we can enhance diagnostic accuracy and improve early intervention strategies.

**4. Dataset and Data Source**

The National Health and Nutrition Examination Survey ([NHANES](https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023)) provides comprehensive datasets essential for analyzing early-stage disease patterns. Below is an overview of key datasets, including descriptions of pertinent variables:

* Demographics Data

This dataset includes key demographics like age (RIDAGEYR), gender (RIAGENDR), race/ethnicity (RIDRETH1), education, and income. These variables help segment patients and analyze how social determinants influence early disease patterns and health disparities.

* Examination Data

Comprising results from physical examinations, this dataset offers measurements such as blood pressure, body mass index (BMI), and waist circumference. These metrics are vital for detecting early warning signs of diseases. Key variables include BPXSY1 (systolic blood pressure), BPXDI1 (diastolic blood pressure), and BMXBMI (BMI), which help in assessing cardiovascular and metabolic health risks.

* Laboratory Data

This dataset provides laboratory test results, including blood glucose levels, cholesterol levels, and other biomarkers indicative of early-stage diseases. Variables such as LBXGLU (fasting blood glucose) and LBXTC (total cholesterol) are instrumental in identifying risks for conditions like diabetes and cardiovascular diseases.

* Questionnaire Data

Containing self-reported information, this dataset covers medical history, lifestyle factors, and symptomatology. Variables like SMQ020 (smoking status) and PAQ605 (physical activity frequency) provide insights into behaviors that may contribute to health risks, aiding in clustering patients based on lifestyle and health history.

* Dietary Data

This dataset analyzes dietary habits linked to disease risk, with key variables like DR1TKCAL (caloric intake) and DR1TCARB (carbohydrate intake). It supports identifying early-stage disease patterns using unsupervised machine learning to enhance proactive healthcare.

**5. Proposed Methodologies**

**Potential Clustering Algorithms:**

1. **K-Means:** Simple and efficient for large datasets, but requires specifying the number of clusters (k).
2. **Hierarchical Clustering:** Useful for understanding relationships between clusters but computationally expensive for large datasets.
3. **Principal Component Analysis (PCA)**: To reduce dimensionality and uncover key health indicators that contribute most significantly to health outcomes.
4. **DBSCAN:** Effective for identifying outliers and handling noise in data.
5. **Gaussian Mixture Models (GMM):** Probabilistic approach that can handle overlapping clusters. (link attached appendix)
6. **Self-Organizing Maps (SOM):** Neural network-based clustering for visualizing high-dimensional data. (link attached appendix)

**Feature Engineering:**

* Normalize or standardize features (e.g., lab results, vitals).
* Handle missing data (imputation or removal).
* Reduce dimensionality using PCA or t-SNE to improve clustering performance.

**APPENDIX:**

### **Process Flow: NHANES Post-Pandemic Data Analysis**

1. Data Acquisition & Understanding
   * Retrieve the NHANES dataset and review its structure, including key variables related to demographics, laboratory results, and lifestyle factors.
   * Assess data completeness, identify missing values, and determine initial suitability for analysis.
2. Data Preprocessing & Cleaning
   * Handle missing and inconsistent values through imputation techniques.
   * Normalize numerical variables, encode categorical data, and perform exploratory data analysis (EDA) to detect trends and distributions.
3. Feature Engineering & Selection
   * Identify key health indicators and reduce dimensionality using Principal Component Analysis (PCA) if necessary.
   * Construct composite features that enhance predictive insights and analytical robustness.
4. Unsupervised Machine Learning Analysis
   * Apply clustering techniques (K-Means, DBSCAN) to segment patients into risk groups.
   * Use association rule mining to uncover relationships between health behaviors and chronic disease prevalence.
5. Insights Extraction & Interpretation
   * Translate machine learning outputs into actionable business and healthcare insights.
   * Generate data visualizations (heatmaps, cluster plots, association graphs) to communicate findings effectively.
6. Evaluation & Business Impact Assessment
   * Compare results with existing public health research to validate findings.
   * Assess potential cost savings, improved patient outcomes, and operational efficiencies for healthcare stakeholders.
7. Implementation & Strategic Recommendations
   * Provide data-driven recommendations for early disease detection programs and targeted healthcare interventions.
   * Suggest areas for further research and potential integration of findings into predictive healthcare models.

This structured approach ensures a rigorous and business-relevant analysis, delivering insights that support both public health decision-making and strategic healthcare initiatives.

**References:**

<https://onlinelibrary.wiley.com/doi/10.1155/2022/6913043>

<https://pmc.ncbi.nlm.nih.gov/articles/PMC8392842/>

<https://www.researchgate.net/publication/376951863_Unsupervised_machine_learning_for_disease_prediction_a_comparative_performance_analysis_using_multiple_datasets>

<https://pmc.ncbi.nlm.nih.gov/articles/PMC6568064/>

<https://pubmed.ncbi.nlm.nih.gov/32744577/#:~:text=Objectives:%20To%20explore%20the%20risk,a%20higher%20risk%20of%20stroke>

**AI usage Disclosure:**

**For this particular project we have used one specific AI tool - ChatGPT;**

1. **Content Generation:** Assisting in drafting and refining project documentation, including sections related to methodology, business relevance, and technical descriptions.

* Chat Prompts : Draft an introduction that highlights the importance of early disease prediction using data analytics
* Summarize the methodologies used for clustering techniques in simple, clear language [link](https://chatgpt.com/share/67a4047c-eb54-8005-b53d-dadfa5c67185)

1. **Data Analysis Guidance:** Providing recommendations on clustering techniques, feature engineering strategies, and potential methodologies for data segmentation and disease prediction.

* Chat Prompts : Explain the difference between K-Means and DBSCAN clustering algorithms for a healthcare data project in as simple terms as possible. [Link](https://chatgpt.com/share/67a4047c-eb54-8005-b53d-dadfa5c67185)

1. **Research Assistance:** Supporting literature review efforts by summarizing relevant research articles and suggesting credible sources.

* Chat Prompts : For the given paper give us key takeaways which would be directly in association with the project that we are working on. [Link](https://chatgpt.com/share/67a4047c-eb54-8005-b53d-dadfa5c67185)

1. **Proofreading and Formatting:** Enhancing the clarity, coherence, and structure of the written content.

* Chat Prompts : Review this section for clarity and conciseness [Link](https://chatgpt.com/share/67a4047c-eb54-8005-b53d-dadfa5c67185)