ALZAWARE: Predictive Model for Early Detection of Alzheimer's Disease Using Social Determinants of Health

Business Understanding.

The AlzAware Project seeks to harness the power of predictive modeling to identify early signs of Alzheimer's Disease (AD) and Alzheimer's Disease-Related Dementias (AD/ADRD) by analyzing social determinants of health. Using data from the Mexican Health and Aging Study (MHAS), this initiative investigates how factors like socioeconomic status, education, and access to healthcare influence cognitive decline. The project aspires to empower early interventions, reduce health disparities, and improve care accessibility for underserved populations.

Join us as we explore some guides that will help us unlock early detection of AD and ADRD.

Research Question.

Primary Research Question

- How can predictive modeling using social determinants of health enable early detection of Alzheimer's Disease (AD) and Alzheimer's Disease-Related Dementias (AD/ADRD)?

Secondary Research Questions

1. Understanding Risk Factors

- What are the key social determinants of health (e.g., education, income, healthcare access) that significantly influence the risk of cognitive decline and AD/ADRD?
- How do socioeconomic status and education levels impact the progression of cognitive decline?

2. Early Detection

- Can predictive models accurately identify individuals at risk of AD/ADRD using non-clinical data?

- What combination of social determinants provides the most reliable indicators for early cognitive impairment?

3. Health Disparities and Bias

- How do disparities in healthcare access and socioeconomic conditions affect the predictive accuracy of models for marginalized communities?
- What strategies can be implemented to ensure the model minimizes bias across diverse demographic groups?

4. Model Accessibility and Scalability

- How can we design a predictive model that is accessible to regions with limited clinical and diagnostic resources?
- To what extent can the developed model be generalized to predict AD/ADRD risk in populations outside of the MHAS dataset?

5. Outcomes and Interventions

- How can early identification of at-risk individuals through the model facilitate targeted interventions?
- What are the potential public health and economic benefits of implementing such a predictive model in underserved populations?

6. Technical Questions

- Which machine learning techniques are most effective for analyzing social determinants of health in relation to AD/ADRD?
- What metrics should be prioritized to evaluate the model's success in early detection and bias mitigation?

Broader Impact Questions

- How can predictive modeling initiatives like AlzAware reduce global health disparities in diagnosing and managing Alzheimer’s Disease?

- What role can such models play in shaping public health policies aimed at addressing cognitive health in aging populations?

- These research questions aim to guide the AlzAware Project in achieving its objectives while addressing critical challenges in early detection, healthcare equity, and model applicability.

Problem Statement.

Alzheimer's Disease affects millions globally, with prevalence expected to rise significantly as populations age. Current diagnostic approaches often fail to detect early cognitive impairment, particularly in marginalized communities where access to healthcare is limited. Social determinants of health, such as education, income, and healthcare access, play a crucial yet underutilized role in understanding and predicting cognitive decline. This project addresses the gap by developing a predictive model that integrates these non-clinical factors, enabling early intervention and reducing disparities in healthcare outcomes.

Objectives.

Main Objective

- To develop a predictive model for the early detection of Alzheimer's Disease (AD) and Alzheimer's Disease-Related Dementias (AD/ADRD) by leveraging social determinants of health

Specific Objectives

- Improved Early Detection: Identify individuals at risk of AD/ADRD based on non-clinical factors, enabling timely intervention.
- Bias Mitigation: Ensure the model provides accurate predictions across diverse demographics, minimizing disparities.
- Enhanced Accessibility: Develop a model that can be applied broadly, requiring only widely available social health data.
- Potential for Generalization: Provide a framework that can be adapted for AD/ADRD prediction in other populations and regions.

Data Understanding.

Data Sources:

The dataset used in this project was derived from the Mexican Health and Aging Study (MHAS), a publicly available longitudinal survey focusing on adults aged 50 and above in Mexico. This comprehensive dataset contains detailed information on demographic, socioeconomic, health, and lifestyle factors, making it ideal for exploring the impact of social determinants of health (SDOH) on cognitive outcomes.

Dataset Overview:

The dataset includes information collected over multiple years, specifically 2003, 2012, 2016, and 2021. These years were selected to provide historical data (2003 and 2012) for training the predictive model and target outcomes (2016 and 2021) for evaluating cognitive health.

- The dataset consists of several key variables used in the analysis:
    - Demographics: Age, gender, marital status, and place of residence.
    - Socioeconomic Factors: Education level, income, and employment status.
    - Health Metrics: Self-reported health, chronic conditions, and body mass index (BMI).
    - Lifestyle Behaviors: Physical activity, smoking status, and alcohol consumption.
    - Cognitive Scores: Assessment of cognitive health over time, used as the primary outcome variable.

Data Preparation.

In this section, we will focus on data cleaning to prepare the dataset for analysis. The following methods will be applied:

- Renaming columns for clarity and consistency
- Handling missing data appropriately

- Identifying and removing duplicate records
- Merging multiple datasets
- Grouping the data for better structure

Additionally, we will perform feature engineering, which includes:

- Selecting relevant columns for the analysis
- Dropping irrelevant columns
- Filtering the dataset to include only the relevant rows

Feature Engineering.

Next, we move on to feature engineering.

We will design custom transformers to capture various aspects of the data, such as temporal changes, education progression, marital transitions, chronic illnesses, and more. This approach is likely to enhance the predictive power of our model by incorporating domain knowledge.We will also be implementing each feature engineering step as a separate transformer class, which promotes code modularity and reusability. We will be creating CustomFeatureEngineer classes and integrating them into a preprocessing pipeline, so as to ensure that all transformations are consistently applied to both training and test data.

Creating Temporal Features:

Since we have data from 2003 and 2012, we can enhance predictive modeling by creating temporal features that capture changes in individuals' characteristics over time and the duration since the last measurement.

-Change Over Time: For individuals with data from both 2003 and 2012, calculate the change or rate of change in features over time.

- Duration Since Last Measurement: Include the time gap between the last available feature data and the target year.

We'll focus on numerical and ordinal variables suitable for calculating changes.

EDA

The Exploratory Data Analysis (EDA) focuses on understanding key features within the dataset, including demographic, health, lifestyle, and composite health scores

- **Demographics**:
  - Analyzed key demographic variables such as age, marital status, locality size, education level, number of children, and spouse's gender.
  - These distributions help identify the diversity within the dataset.
- **Health and Lifestyle Variables**:
  - Explored health perceptions, limitations in daily living, depressive symptoms, health coverage, vaccination status, exercise frequency, and tobacco use.
  - These variables are important in understanding the overall health and wellbeing of individuals in the dataset.
- **Composite Score Analysis**:
  - The composite score aggregates various health and lifestyle domains, and its distribution is generally bell-shaped with most values around the middle range.
  - Limitation variables (e.g., daily living activities, mobility, depression) show skewed distributions, with most individuals having fewer limitations.
- **Visual Analysis**:
  - **Histograms**:
    - Each variable’s distribution was visualized using histograms.
    - Composite score: Bell-shaped distribution.
    - Limitation variables: Heavily skewed, indicating fewer individuals report high limitations.

  - **Scatter Plot Relationships**:
    - **Composite Score vs. Limitations**: Slight downward trend suggesting a negative correlation—more limitations tend to correlate with lower composite scores.
    - **Limitations Co-occurrence**: Scatter plots show that individuals with limitations in one area often have limitations in other areas (e.g., someone with mobility issues may also have limitations in daily living activities).

- **Notable Observations**:
  - **Low Composite Scores and Higher Limitations**: Individuals with more limitations tend to have lower composite scores, suggesting that physical or mental limitations affect performance.
  - **Clustering at Low Limitation Values**: Most data points are concentrated around low values (0, 1, or 2) for each limitation variable.
- **Implications of the Analysis**:
  - **Impact of Limitations on Performance**: Negative correlation suggests physical and mental limitations could hinder cognitive health assessments and composite scores.
  - **Co-Occurrence of Limitations**: Interrelationships among limitations indicate a need for holistic support for individuals facing multiple challenges, as addressing one limitation may improve others.

This EDA offers valuable insights into the dataset, identifying key factors that may influence cognitive health and early detection of Alzheimer’s Disease and related dementias.

Modeling

The modeling approach will focus on predictive machine learning techniques tailored to the structure and goals of the dataset:

1. Techniques: Models like Linear Regression and ensemble Learning.
2. Target Variable: The primary outcome is the cognitive health score, which is assessed over time and categorized into early risk stages of AD/ADRD.
3. Feature Engineering: Temporal changes, education progression, and health metrics will be key predictors.
4. Model Selection: A baseline logistic regression will evaluate the dataset's predictability. Subsequent models will include feature selection and hyperparameter tuning.
5. Validation: K-fold cross-validation ensures robust results. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC will measure success.

Conclusions

The AlzAware project demonstrates the potential of social determinants of health in predicting early signs of Alzheimer’s Disease. Key findings from data exploration reveal that demographic and socioeconomic factors significantly influence cognitive health. By implementing machine learning models, the project addresses gaps in early detection and provides a scalable solution for underserved populations.

Recommendations

1. Health Policy: Incorporate insights from the model into public health strategies for targeted early interventions.

2. Community Outreach: Develop awareness programs in regions with limited access to healthcare.

3. Integration: Collaborate with healthcare providers to use the model in clinical workflows.

4. Bias Reduction: Continuously refine the model to ensure fairness across diverse populations.

Future Improvement Ideas

1. Data Expansion: Incorporate additional datasets for broader demographic coverage.

2. Model Refinement: Explore advanced neural networks for improved prediction accuracy.

3. Personalization: Tailor interventions based on individual risk profiles predicted by the model.

4. Longitudinal Studies: Use future survey data to validate and enhance model performance.

5. Explainability: Develop interpretable models to gain insights into the key predictors driving cognitive health outcomes.

Deployment

1. Method: Deploy the model as a web application for healthcare professionals and policymakers.
2. Functionality:
   - Input data on social determinants to predict AD/ADRD risk.
   - Provide risk scores and feature importance explanations.
   - Generate tailored intervention recommendations.
3. Platform: Cloud-based deployment ensures accessibility and scalability.
4. Integration: API endpoints allow seamless integration with electronic health records (EHRs) and other systems.