Alzheimer's Disease Prediction

Tessa Clary
Applied Computer Science
University of Colorado
Boulder, CO
sccl8634@colorado.edu

Problem Statement and Motivation:

Alzheimer's Disease is a complex and devastating condition that impacts millions of people around the globe. Early detection is paramount for patient care as well as advancing our knowledge of the disease. The motivation for this project is to develop a predictive model that can help identify those who are at risk of developing early-stage Alzheimer's. This will be done using the techniques and knowledge that we have learned about data mining and analysis in this course. The goal of this project is to discover key risk factors and patterns that can help lead to earlier diagnosis and better treatments.

Literature Survey:

"A smart Alzheimer's patient monitoring system with IoT-assisted technology through enhanced deep learning approach"
This study utilized IoT devices and deep learning techniques to help predict Alzheimer's. Given an accuracy and precision rate of 98% and 97% respectively, there is obviously a lot of value in using IoT to accurately monitor patients in real time. This sets a high benchmark for this project as similar rates of accuracy and precision should still be potentially achievable.

"Population measures of subjective cognitive decline: A means of advancing public health policy to address cognitive health"
This study focuses on SCD, or subjective cognitive decline. It is an early cognitive indicator that complements this project given that it includes an early-stage assessment parameter. Given the issues with studying Alzheimer's Disease via population-based measures, this can potentially help alleviate the difficulty.

Proposed Work:

The data set will be collected and cleaned, with inconsistencies, missing or incorrect values, and outliers being identified and handled. Relevant data from the literature survey will also be examined for pertinence; integrating SCD is likely.

After this, data preprocessing will begin. Feature engineering will be used to extract relevant variables, and the data will be normalized for consistency. Data visualization here can also help identify anomalies or patterns. Data transformation will then be carried out at this point in the form of finalizing the solution to issues such as missing values or converting data types.

Given that this project is focused on finding early-stage Alzheimer's Disease, the model design will focus on demographic data, health factors, and relevant correlations in an attempt to predict patient outcomes. To assess the model's performance, a number of evaluation methods will be used including accuracy, precision, and recall.

The difference between this project and prior works is that the given prior works had a far narrower scope of predictive factors. This will be a more inclusive and comprehensive analysis of overall demographic and health factors, which may dilute the model's predication capability,

but should also do a better job of showing which variables if any impact the development of the disease.

Data Set:

This project focuses on a published data set from the CDC called: “Alzheimer's Disease and Healthy Aging Data”. It is 250,000 rows with 39 column labels. It includes variables for:

- Year Start / End
- Location
- Disease Indicator Category:
    - Caregiving
    - Cognitive Decline
    - Mental Health
    - Nutrition/Physical Activity/Obesity
    - Overall Health
    - Screenings and Vaccines
    - Smoking and Alcohol Use
- Age Group
- Gender
- Race

Data.gov Dataset: https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data

Same Dataset hosted by the CDC: https://data.cdc.gov/Healthy-Aging/Alzheimer-s-Disease-and-Healthy-Aging-Data/hfr9-rurv

CDC’s page that uses the same dataset for visualization purposes: https://www.cdc.gov/aging/agingdata/index.html

Evaluation Methods:

This dataset will be evaluated primarily through the use of a regression model- accuracy, precision, recall, F1 score, and ROC-AUC will be considered as metrics. A confusion matrix will also reveal useful insights into the model’s performance as well. In the event that the initial model is too generalized to return quality performance metrics, additional data transformation may be carried out alongside other scope changes in order to ensure a final model with a real capacity to predict. Additional models may be developed beyond the initial regression as needed to further explore the data set.

Tools:

The majority of the code for this project will be written in Python for its flexibility with data manipulation, analysis, and modeling. Numpy is also going to be potentially used given its ability to handle numerical operations and arrays. Pandas is irreplaceable when it comes to data manipulation, cleaning and structuring so it will be relied on heavily. SciPy expands the scientific and technical computing functions available to the project, so it may come into play if the initial analysis does not have quality results. Scikit-Learn provides additional ML algorithms and tools for pertinent use as well. SQL may be used depending on how easy the data set is to load and manipulate. For data visualization, Matplotlib will be used with the potential for either Seaborn or Plotly as needed. Github will be used for version control and project management.

Milestones:

1. Data Preprocessing (1 Week)
    a. The data set will be cleaned, with any issues or gaps in it addressed.
    b. Exploratory data analysis will be performed to gain early insights.
    c. Preprocessing will take place during this time, including

feature engineering, normalization, and data transformation.

2. Model Design (1 Week)
    a. During this stage, the scope of the initial model will be defined and developed, with its objectives and target variables specified.
    b. Model architecture will be chosen and a baseline model will be developed.
3. Model Training (1 Week)
    a. The model will be trained using the preprocessed data.
4. Evaluation (1 Week)
    a. Its performance will then be benchmarked against other relevant solutions to help gauge its relative accuracy alongside other evaluation metrics.
    b. Accuracy, precision, recall, F1 score, and ROC-AUC along with other measures as needed.
5. Model Optimization (1 Week, Project Part 3 Goal on 12/4)
    a. Based on the results of the evaluation, additional tuning and refinement of the model may be necessary to a variable degree at this stage.
6. Write-up (1 Week)
    a. Documentation will be finalized, and project conclusion will have been reached.

Milestone 1 is currently being carried out, with milestones 2 and 3 to be performed over the next few weeks. By fall break, milestones 3 and 4 will be in progress. At this stage, initial analysis will likely be complete and the feasibility of the project goal can then be reexamined if the scope likely needs to be updated. Ideally this allows for focus on model optimization for the rest of the course so long as the model does not require extensive rework. Once the optimizations are sufficient, the remainder of the available time will be spent preparing the project presentation.