Project Title

Predicting Alzheimer's Disease Diagnosis Using Machine Learning Models

Team Members

Mike Xia N18489155

Ken Li N14232508

---

1. Background and Motivation

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that gradually impairs memory, cognition, and daily functioning. Despite decades of research, the underlying causes of AD remain unclear, and no single biological mechanism has been definitively identified. This scientific uncertainty makes early detection especially challenging, yet early diagnosis is crucial for effective intervention and care planning.

Given this context, our project seeks to explore whether machine learning techniques can help identify which patient characteristics—demographic, behavioral, cognitive, or biomedical—exhibit strong associations with Alzheimer's Disease. By examining these relationships, we hope to gain insights into potential risk factors that may contribute to AD.

In addition, we apply classification models to predict whether a patient is likely to have Alzheimer's based on their observed features. Through this dual approach—feature correlation analysis and supervised classification modeling—we aim to evaluate both the interpretability and predictive potential of machine learning for supporting Alzheimer's-related research.

---

2. Dataset Description

https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data

The dataset contains 2149 patient records and 35 variables, including:

- Demographics: Age, Gender, Ethnicity, Education
- Health Behaviors: Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality
- Clinical Metrics: BMI, Blood Pressure, Cholesterol
- Cognitive & Behavioral Measures: MMSE score, MemoryComplaints, FunctionalAssessment, Disorientation, Forgetfulness, ADL, BehavioralProblems
- Target Variable: Diagnosis (1 = Alzheimer's, 0 = No Alzheimer's)

---

3. Research Question

This project investigates two central questions:

1. Prediction:

   Can machine learning classification models accurately predict Alzheimer's Disease (AD) diagnosis using demographic, behavioral, cognitive, and clinical features?

2. Interpretability / Risk Factors:

   Which features show the strongest association with Alzheimer's Disease, and to what extent do certain demographic, lifestyle, or biomedical variables contribute to the likelihood of an AD diagnosis?

Together, these questions allow us to assess both the predictive performance of machine learning models and the relative importance of key features, offering insights into potential factors linked to Alzheimer's Disease.

---

4. Methods

4.1 Data Preprocessing

- Encode categorical variables (e.g., Gender, Ethnicity)
- Standardize numerical features
- Train-test split (80/20)

4.2 Models to be Trained

1. Logistic Regression
    - Baseline interpretable linear model
2. Support Vector Machine (SVM)
    - Nonlinear boundary with RBF kernel
3. Random Forest Classifier
    - Ensemble model + feature importance

4.3 Evaluation Metrics

- Accuracy
- Confusion Matrix
- ROC-AUC (primary metric)

- Precision / Recall (important for medical prediction

---

5. Expected Outcomes

- Comparative performance of the three models
- Identification of key features associated with Alzheimer's risk
- Visualizations including:
    - Correlation heatmap
    - Feature importance (Random Forest)
    - ROC curves for all models
- Final recommendation for best-performing model

---

6. Feasibility

Given the quality and structure of the dataset, this project is highly feasible. The Alzheimer's dataset contains **2149 observations and 35 well-defined variables**, providing sufficient sample size and feature richness for supervised learning. Because the dataset is clean and complete, we can perform preprocessing steps such as encoding categorical variables and standardizing numerical features with minimal difficulty. Exploratory analysis—such as correlation heatmaps and feature-importance metrics—can be conducted early to understand the relationships among variables and to prepare the data for model development.

Since our primary goal is to build a **classification model** that predicts Alzheimer's diagnosis, we will implement several widely used and computationally efficient algorithms, including **Logistic Regression**, **Support Vector Machine (SVM)**, and **Random Forest**. To further

compare ensemble performance, we may also explore **bagging** and **boosting** methods. All of these models can be trained using standard machine learning libraries within the project timeline.

Model performance will be evaluated using established metrics such as **accuracy**, **ROC-AUC**, **precision**, **recall**, and **confusion matrices**, ensuring a comprehensive assessment of predictive utility. Taken together, the dataset quality, methodological suitability, and manageable computational requirements make the project fully achievable within the course timeframe.