**Predicting Alzheimer’s Disease Through Classification Models**

Authors: Christos Symeou, Leontios Ioannou

Subject: EPL448 – Data mining

Submission Date: 27/4/2025

**1.Introduction**

Alzheimer’s disease is a serious brain disorder that slowly affects memory and thinking skills, and eventually makes it difficult for a person to carry out simple tasks. It’s the most common cause of dementia and mainly affects older people. As life expectancy increases around the world, the number of people living with Alzheimer’s is also expected to rise, which makes early detection more important than ever.

Being able to predict Alzheimer’s early can make a big difference. It gives people a chance to start treatment sooner, make lifestyle changes, and plan ahead while they still can. With the help of machine learning, we can now analyze large datasets to find patterns that might help identify who is at risk.

For this project, we worked with the Alzheimer’s Prediction Dataset (Global). The main goal of the project is to build classification models that can predict whether someone is at risk of developing Alzheimer’s disease. We began by performing exploratory data analysis (EDA) to understand the structure of the dataset, the distribution of features, and the behavior of the target variable. This helped us identify any imbalances, correlations, and feature types that would influence our modeling approach.

A screenshot of a computer program

AI-generated content may be incorrect.After EDA, we conducted initial experiments with different preprocessing strategies and a wide range of machine learning models using their default hyperparameters. This helped us understand which combinations performed best on our dataset. Based on these results, we selected the top two models along with their best-performing preprocessing techniques. For these selected models, we then used pipelines with GridSearchCV to fine-tune thehyperparameters and try to improve the performance even further.

**2. Exploratory Data Analysis**

Before starting the preprocessing and modeling, we performed exploratory data analysis (EDA) to better understand the structure of our dataset and identify any issues or patterns that might influence our approach.

**2.1 Dataset Structure**

The dataset used in this project is called the Alzheimer’s Prediction Dataset (Global) and contains detailed records of individuals from various backgrounds. It includes a total of **74,283 entries** and **25 columns**, each representing either personal, lifestyle, medical, or cognitive data. One of the key advantages of this dataset is that it is **complete**, with **no missing values**, allowing us to move directly into analysis without needing to impute or drop data.

As shown in the figure below, most of the features are stored as object types, meaning they are **categorical** variables such as gender, country, smoking status, and employment. A smaller number of features, like Age, Education Level, BMI, and Cognitive Test Score, are stored as numeric types (int64 or float64). These features are suitable for statistical analysis and machine learning without much conversion. Additionally, the dataset contains no irrelevant or identifier fields like patient names or IDs, which makes it clean and ready for modeling. *(Refer to the image showing the column names and data types)*

**2.2 Target Variable: Alzheimer’s Diagnosis**

The target variable in this dataset is **Alzheimer’s Diagnosis**, which indicates whether or not an individual has been diagnosed with Alzheimer’s disease. This is a **binary classification problem**, where:

* **"Yes"** means the individual has been diagnosed
* **"No"** means the individual has not been diagnosed

After analyzing the distribution of the target variable, we observed that the dataset is balanced, since the proportion is very close to 50/50.

* Around **58.7%** of the individuals are labeled **"No"**
* The remaining **41.3%** are labeled **"Yes"**

Relying only on accuracy can be misleading, so we used additional metrics like **F1-score** to get a better picture of each model’s performance. *(We also included a bar chart showing the counts for each class).*

A graph of a number of blue squares

AI-generated content may be incorrect.

**2.3 Key Feature Analysis**

In this section, we explore a few of the most informative features in relation to Alzheimer’s Diagnosis. We aim to understand how these features differ between individuals diagnosed with Alzheimer’s and those not diagnosed. We also visualize the distribution and importance of each feature to better interpret its role in prediction.

The selected features for deeper analysis are:

1. Age
2. Family History of Alzheimer’s
3. Genetic Risk Factor (APOE-ε4)

**2.3.1 Age vs Alzheimer’s Diagnosis**

Age is one of the most well-known and studied risk factors for Alzheimer’s disease. To explore this further, we analyzed how age differs between individuals diagnosed with Alzheimer’s and those who are not. The plots below show **separate histograms with KDE curves** for each group:

* The **non-diagnosed group** has a higher concentration of individuals between ages **50 to 70**.
* The **diagnosed group** tends to be older, with the majority of cases clustering around **75 to 90+** years.

**A graph showing the age of a person

AI-generated content may be incorrect.**

**A graph of age distribution

AI-generated content may be incorrect.**These distributions support the established fact that the likelihood of developing Alzheimer’s increases with age.

**2.3.2 Family History vs Alzheimer’s Diagnosis**

Family history is a strong indicator of genetic or environmental predisposition. Individuals with a family history of Alzheimer’s are generally at higher risk. The count plot below shows:

* Individuals with **no family history** are more likely to be undiagnosed.
* Among those with a **positive family history**, the number of diagnosed individuals is slightly **higher than undiagnosed**, indicating a potential correlation.

A graph with green and orange squares

AI-generated content may be incorrect.This suggests that family history is a relevant predictor, although not definitive on its own.

**2.3.3 Genetic Risk Factor (APOE-ε4) vs Diagnosis**

The APOE-ε4 allele is one of the most recognized genetic risk factors associated with Alzheimer’s disease.

The plot below illustrates the relationship:

* Most individuals without the APOE-ε4 allele are not diagnosed.
* A notable proportion of individuals who have the APOE-ε4 allele are diagnosed, highlighting its importance.

**A graph with green and orange bars

AI-generated content may be incorrect.**

This further reinforces the genetic aspect of Alzheimer’s risk, making APOE-ε4 a **critical variable in classification models**.