Alzheimer's Disease Risk Factors — Capstone Report

1. Project Overview

Alzheimer's disease is a progressive neurological disorder that affects memory, cognitive function, and behavior. It is the most common form of dementia and a major global health challenge. Early diagnosis and prevention are critical, yet difficult due to the complex nature of its risk factors.

This capstone project investigates potential indicators of Alzheimer's disease using a dataset sourced from Kaggle. The dataset contains 2,149 records and 35 variables spanning demographics, lifestyle habits, medical history, cognitive assessments, and a binary diagnosis indicator. The primary goal is to analyze patterns in these variables and identify relationships that can support early detection or future predictive modeling.

2. Dataset Summary

The dataset includes a wide range of attributes:
- Demographics: Age, Gender, Ethnicity, Education Level
- Lifestyle: Smoking, Alcohol Consumption, Diet Quality, Physical Activity, Sleep Quality
- Medical History: Diabetes, Depression, Cardiovascular Disease, Head Injury, Hypertension
- Biometrics: Blood Pressure, Cholesterol (Total, HDL, LDL, Triglycerides)
- Cognitive and Functional Metrics: MMSE, Functional Assessment, Confusion, Forgetfulness, Behavioral Changes

The target variable is `Diagnosis` (0 = not diagnosed, 1 = diagnosed with Alzheimer’s disease). The dataset was clean, with no missing values or structural issues.

3. Data Cleaning

To prepare the dataset for analysis:
- Duplicate records were removed to ensure integrity.
- The `DoctorInCharge` column was dropped as it contained a single non-informative value.
- Categorical variables were mapped for clarity:
  - Gender: 0 → Female, 1 → Male
  - Ethnicity: 0 → White, 1 → Black, 2 → Asian, 3 → Hispanic
  - Education Level: 0 → High School or Less, 1 → Some College, 2 → Bachelor's Degree, 3 → Graduate Degree

No imputation was necessary due to the absence of missing values. The cleaned dataset was ready for exploration and modeling.

4. Exploratory Data Analysis (EDA)

We visualized distributions of key numeric variables using histograms and boxplots. Age was skewed toward older adults, as expected for an Alzheimer's-focused dataset. Alcohol consumption showed a right-skewed distribution, while physical activity and sleep quality appeared more normally distributed.

A correlation heatmap revealed strong relationships among cardiovascular metrics but relatively weak correlations between cognitive scores and biometric data. These visualizations provided foundational insights for subsequent analysis.

5. In-Depth Analysis by Diagnosis

We grouped patients by `Diagnosis` and calculated mean differences across all numeric features. Patients with Alzheimer's showed higher levels of confusion, forgetfulness, and difficulty completing tasks. These symptoms were consistently associated with the diagnosis group.

Non-diagnosed patients tended to have better physical activity levels and sleep quality, possibly indicating protective factors. This step helped prioritize which

features to investigate more closely.

6. Cognitive Assessment: MMSE Score Analysis

The MMSE (Mini-Mental State Examination) is a widely used cognitive test scored out of 30. A score below 24 typically suggests cognitive impairment.

In our analysis:
- Diagnosed patients had a mean MMSE score of 11.99 (std dev 7.23)
- Non-diagnosed patients had a mean of 16.27 (std dev 8.93)

A boxplot visualized this difference. These results confirm the MMSE's role as a meaningful indicator of cognitive health and Alzheimer's diagnosis.

7. Statistical Testing: Chi-Square Analysis

We conducted Chi-Square tests on three categorical variables: Gender, Ethnicity, and Education Level. None of them showed statistically significant association with Alzheimer's diagnosis at the $p < 0.05$ level.

Variable	p-Value	Significant?
Gender	0.354	No
Ethnicity	0.098	No
Education Level	0.217	No

While these variables are important demographically, they did not differ significantly in diagnosed vs. non-diagnosed groups in this dataset.

8. Predictive Modeling: Logistic Regression

We trained a logistic regression model using a subset of features:
- Age, BMI, Physical Activity, Sleep Quality
- Confusion, Forgetfulness, Difficulty Completing Tasks

The model yielded:
- **Accuracy:** 62.2%
- **Confusion Matrix:** Predicted only the "not diagnosed" class
- **Top contributing features:** Sleep Quality, BMI, Age

This result demonstrates class imbalance and highlights the need for advanced techniques (resampling, regularization, alternative models). It still provided insight into which features carry the most predictive weight.

9. Conclusion and Final Insights

This project explored a rich dataset of Alzheimer's risk factors and produced several key findings:
- Cognitive symptoms like forgetfulness and confusion strongly differentiate diagnosed patients.
- MMSE scores validated as a critical metric in diagnosis prediction.
- Lifestyle factors such as sleep quality and physical activity show protective potential.
- Categorical demographic variables were not significantly associated with diagnosis in this dataset.
- Logistic regression showed predictive potential but needs refinement due to imbalance.

Future work should focus on predictive model improvement and apply the same analysis techniques to longitudinal or multi-modal datasets (e.g., imaging, genetics). This study highlights the power of data analytics to uncover meaningful patterns in clinical data.