Modeling For Alzheimer's Diagnosis: Predictive Analysis*

STA314 Group28

Yuguang Duan 1007824348 Pingxuan Ren 1007613708
Yiyi Feng 1007763984 Nida Li 1008507411

December 3, 2024

This study developed a model to predict Alzheimer's disease diagnoses based on patients' demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, and symptoms. Additionally, we analyzed all predictive factors included in the model to identify their influence on Alzheimer's diagnosis outcomes, aiming to uncover factors that are more likely to contribute to the onset of Alzheimer's disease. Finally, we analyzed patients' health factors to explore whether health-related behaviors (such as smoking and drinking) play a decisive role in Alzheimer's disease diagnosis.

Table of contents

*Code and data are available at: https://github.com/william03duan/Alzheimer-Disease-prediction-314.git.

1 Introduction

Alzheimer's disease is a complex condition with no known definitive cause or cure. Its impact on individuals and families is profound, affecting both cognitive function and quality of life. As the global population ages, Alzheimer's prevalence is expected to rise, emphasizing the need for early diagnosis and research. Identifying factors contributing to the onset of this disease is crucial for developing effective interventions. This study aims to predict Alzheimer's diagnosis using demographic, medical, and lifestyle factors, and to build a predictive model for early detection.

Our analysis focuses on predicting Alzheimer's diagnosis based on variables such as age, gender, medical history (e.g., hypertension, diabetes), cognitive performance, and lifestyle factors (e.g., smoking, alcohol consumption). We employed a Random Forest model to estimate diagnosis outcomes, incorporating predictors like age, ethnicity, cardiovascular health, memory complaints, and physical activity levels. The goal is to identify which factors most significantly influence the risk of Alzheimer's.

Our model predicts a higher likelihood of Alzheimer's diagnosis in patients with cardiovascular disease, depression, and low cognitive performance, with health behaviors such as smoking and physical inactivity also playing key roles. This study aims to inform future research and clinical practices by identifying key predictors for early diagnosis and intervention strategies.

The remainder of this paper is structured as follows: Section 2 describe the detail of research questions addressed in our study. Section 3 provides an overview of the dataset, including data processing and data analyzation. Section 4 explains the modeling approach, and the algorithms chosen in this study to assess the model. Section 5 presents the findings, including a summary of the predicted diagnose result for patients, and conclusions drawn for our research questions. In Section 6, we discuss the implications of these results, the limitations of our analysis, and what can we do next to improve our model. Additional methodological details and diagnostics are included in the appendix.

2 Problem statement

This study, while constructing a predictive model for Alzheimer's diagnosis, also investigates two research questions:

- Among the 32 predictors included in the model, covering patients' demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, and symptoms, which predictors are most critical for accurate diagnosis?
- Within lifestyle factors, including BMI, smoking, alcohol consumption, physical activity, diet quality, and sleep quality, which factors have a statistically significant influence on Alzheimer's diagnosis?

Answering these questions will deepen our understanding of the key contributors to Alzheimer's diagnosis and provide guidance for patients to adopt healthier lifestyles, potentially reducing the risk of developing the disease.

3 Statistical Analysis

3.1 Data Processing

Data preprocessing or feature engineering: steps for‘ data cleaning, transformation, feature selection, and feature engineering techniques applied

3.2 Data Analyzation

To gain a comprehensive understanding of the dataset before modeling, we created various visualizations to quickly grasp its structure. Five types of charts were produced:

- Boxplots and Histograms for quantitative data, illustrating their distributions and highlighting outliers or patterns.
- Pie Charts and Bar Charts for categorical data, providing insights into the proportions and frequencies of each category.
- A Correlation Matrix for all predictors and diagnose data, showing the relationships and dependencies between variables.

Boxplots are available at . The boxplot analysis was conducted for each variable. The analysis of 15 variables revealed no outliers.

Histograms are available at . The histograms highlight distinct patterns in the dataset. Participant ages show a sharp peak at 90, indicating a larger proportion of older individuals.

Triglyceride levels cluster around 300, suggesting elevated levels in a significant number of participants. Sleep quality scores are concentrated near 8, reflecting generally good sleep among participants, with fewer reporting poor sleep. MMSE scores exhibit notable variability, with a dip between 10 and 20, pointing to differences in cognitive abilities within the sample. These trends provide meaningful insights and warrant further exploration.

Piecharts are available at . From the perspective of patients' demographic details, the gender distribution is nearly balanced, with Caucasians significantly outnumbering other ethnic groups, and the majority of patients having a High School education level. From a health standpoint, most individuals are non-smokers and have no history of Alzheimer's in their family, cardiovascular disease, diabetes, depression, head injuries, or hypertension. Similarly, the majority do not report symptoms such as memory impairment, confusion, disorientation, personality changes, difficulty completing tasks, or forgetfulness.

Barcharts are available at . The bar plot reveals the number of Alzheimer's diagnoses across different categories for each factor. For most factors, there are no significant differences in diagnosis rates among categories. For instance, in the gender factor, males and females have similar diagnosis rates. However, notable differences are observed in certain cases. Caucasians have a slightly higher diagnosis rate compared to other ethnic groups. Among health-related factors, patients with hypertension, memory complaints, behavioral problems, and forgetfulness exhibit significantly higher probabilities of an Alzheimer's diagnosis. Interestingly, patients with a family history of Alzheimer's do not show an elevated diagnosis rate, suggesting that the disease may not have strong genetic predisposition.

Figure for correlation matrix is available at . The correlation matrix shows most variables have weak correlations with diagnosis (between -0.2 and 0.2). However, MMSE, Functional Assessment, and ADL show stronger negative correlations (below -0.2), indicating higher scores are linked to a lower likelihood of Alzheimer's. In contrast, memory complaints and behavioral problems have stronger positive correlations (above 0.2), suggesting that more severe symptoms increase the likelihood of a diagnosis. These findings align with the bar chart analysis, highlighting these variables' predictive value.

4 Model

4.1 Model Justification

For our study, we utilized the Random Forest algorithm to predict Alzheimer's disease diagnosis. Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their outputs for predictions. This algorithm was chosen for its robustness, ability to handle datasets with mixed variable types, and effectiveness in identifying important features. It also mitigates the risk of overfitting, making it a reliable choice for complex datasets such as ours.

The Kaggle dataset provided pre-separated training and test datasets, eliminating the need for manual data partitioning. The model was trained on the training dataset, which included 32 predictors covering demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, and symptoms. The dependent variable was Diagnosis, indicating whether a patient was diagnosed with Alzheimer's disease. We implemented the Random Forest model using the randomForest package in R, setting the number of variables randomly sampled at each split (mtry) to 6 and enabling the calculation of feature importance. The trained model was then evaluated on the test dataset, and predictions were generated using the predict function.

While training the model, we explored other machine learning algorithms, including k-Nearest Neighbors (kNN) and Naïve Bayes. However, these models yielded significantly lower prediction accuracies compared to Random Forest. For instance, kNN struggled with higher-dimensional data, and Naïve Bayes, with its assumption of feature independence, proved less effective in capturing complex interactions between predictors. Random Forest outperformed these alternatives in both accuracy and interpretability, supporting its suitability for our analysis.

In this context, Random Forest is reasonable as it aligns with the objectives of the study. Beyond predicting Alzheimer's diagnosis, it identifies the most influential predictors among the 32 included variables. This feature importance analysis provides valuable insights into the primary factors contributing to diagnosis, such as cognitive and functional assessments, which showed strong correlations with Alzheimer's likelihood.

( )

4.2 Model Evaluation

Specify the metrics chosen to evaluate model performance (e.g., accuracy, precision, recall, F1 score) and explain their appropriateness.

5 Result and Conclusion

Table 1: Feature Importance for Alzheimer's Diagnosis Model

	Mean Decrease Accuracy	Mean Decrease Gini
PatientID	-2.0281580	17.501944
Age	1.1136721	13.700540
Gender	-1.6312724	2.348315
Ethnicity	0.5180901	4.517400
EducationLevel	-0.5483328	5.514173

	Mean Decrease Accuracy	Mean Decrease Gini
BMI	-0.1658121	18.367270
Smoking	-2.1245737	2.006530
AlcoholConsumption	-0.6144742	17.018871
PhysicalActivity	-0.0890730	18.745054
DietQuality	-0.2474637	18.662884
SleepQuality	0.2960033	19.039733
FamilyHistoryAlzheimers	-0.8710437	2.394133
CardiovascularDisease	-1.3477476	1.987291
Diabetes	1.2916761	2.125278
Depression	0.2202775	1.896398
HeadInjury	-1.4943062	1.519876
Hypertension	0.5819497	1.841131
SystolicBP	-1.9517112	15.447448
DiastolicBP	-0.6072691	14.103701
CholesterolTotal	-0.5999197	17.489800
CholesterolLDL	-0.6797411	16.639124
CholesterolHDL	0.6010065	17.105686
CholesterolTriglycerides	0.6188522	18.941319
MMSE	57.4804854	90.847339
FunctionalAssessment	70.2112687	133.791213
MemoryComplaints	53.0998132	55.192959
BehavioralProblems	41.5222619	37.220173
ADL	65.9871204	112.234999
Confusion	0.2445177	1.773776
Disorientation	-0.4093992	1.911363
PersonalityChanges	-1.9090836	1.758660
DifficultyCompletingTasks	-0.0693538	1.569454
Forgetfulness	-1.0955541	2.598917
DoctorInCharge	0.0000000	0.000000

Table 1 presents the feature importance scores for two metrics—Mean Decrease Accuracy and Mean Decrease Gini—derived from a random forest model predicting Alzheimer's diagnosis. These scores provide insight into how each feature contributes to the model's predictive performance.

Mean Decrease Accuracy measures how much the accuracy of the model decreases when the feature is excluded from the model. A higher value indicates that the feature is more critical to accurate predictions. For example, MMSE (Mini-Mental State Examination), Functional Assessment, Memory Complaints, and Behavioral Problems all show very high values, suggesting these features significantly impact the model's accuracy. These findings align with

our understanding that cognitive and behavioral factors are critical in diagnosing Alzheimer's disease.

Mean Decrease Gini assesses how much a feature improves the homogeneity of the nodes and leaves in the random forest. A higher value indicates that the feature plays a larger role in reducing impurity and improving the model's decisions. Features like MMSE, Functional Assessment, Memory Complaints, and Behavioral Problems also show strong values here, which further suggests that these variables have strong discriminatory power for classifying Alzheimer's diagnoses.

Some features, such as PatientID, Gender, and Family History of Alzheimer's, show negative values for both Mean Decrease Accuracy and Mean Decrease Gini, indicating that their contribution to the model is less significant or even potentially detrimental. Specifically, PatientID does not contribute to the model, with both scores at 0, suggesting it has no predictive power.

Interestingly, lifestyle factors such as BMI, Smoking, Alcohol Consumption, Physical Activity, and Diet Quality show moderate values for Mean Decrease Gini, indicating they provide some predictive power but are less influential than cognitive and behavioral factors.

In conclusion, the most important predictors of Alzheimer's diagnosis in this model are cognitive and functional assessments, such as MMSE, Functional Assessment, Memory Complaints, and Behavioral Problems, which have the highest Mean Decrease Accuracy and Mean Decrease Gini scores. These results suggest that improving cognitive and behavioral evaluations could enhance Alzheimer's diagnosis prediction.

( )

6 Discussion

6.1 Summery of Findings

6.2 Limitation

6.3 Future Study

Appendix a.

6.1 Data Retrieval

7 References