McGill University Desautels Faculty of Management

MGSC 661 Multivariate Statistical Analysis

**Final Report:**

# Predicting Alzheimer's Disease in Hospital Patients Using Machine Learning


Emmanuel Okerein, 261267987

Arantzazu Arregui Gonzalez, 261284753

Chloee Liew, 261111208

Nicholas Stanfield, 261068709


December 6th, 2025

## 1. Introduction and Business Context

Alzheimer's disease (AD) stands as the most prevalent neurodegenerative disorder and the primary cause of dementia, characterised by a progressive decline in memory, cognitive function, and the ability to perform daily tasks (Arjaria et al., 2024; L. Wang et al., 2021). However, the disease remains incurable, making early detection critical for effective disease management (Antor et al., 2021; Rani et al., 2024). Conventional diagnostic techniques often rely heavily on complex clinical judgments, neuroimaging, and neuropsychological assessments, which can be time-consuming, expensive, and subject to variability in predictive accuracy (Kavitha et al., 2022; X. Wang et al., 2024).

To address these diagnostic challenges, machine learning (ML) and data science approaches have emerged as powerful tools for developing automated and objective risk prediction models (Ezzati et al., 2019). Previous studies have demonstrated the efficacy of algorithms such as Support Vector Machines (SVM), Random Forests, and Logistic Regression in distinguishing between cognitively normal individuals and those with AD using demographic, clinical, and neuroimaging data (Neelaveni & Devasana, 2020; Antor et al., 2021). Building on this body of research, this project aims to construct a predictive model designed to diagnose Alzheimer's disease.

## 2. Methodology

### *2.1. Dataset*

Our dataset was found on Kaggle and provides comprehensive information on over 2,000 hospital patients. Each row represents a unique patient. The features can be split into six main categories: demographics, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, and symptoms. The target variable was binary where one indicated a diagnosis of AD and zero not diagnosed with AD. The dataset is very clean and detailed likely owing to its synthetic nature. While we could not find further information on how the dataset was generated, one possibility is that the data is real but anonymized and altered using a deep learning model. All the data is numerical or, if categorical, encoded already. In addition, there is no missing values meaning the dataset is almost ready for modelling.

### *2.2. Exploratory Data Analysis*

The distributions of the features were explored to gain a richer understanding of the dataset (see Appendix). The age range is between 60 and 90, there is an even gender split and the majority of the patients are Caucasian indicating this dataset is likely based off a Western country. Next, we explored the distribution of the target variable (see Appendix). It is slightly imbalanced, with 35% of the patients having a positive diagnosis. The imbalance is not so extreme as to require adjusting the data and/or model, but it is also not representative of real patients. We expect that in real hospitals the diagnosis rate (even among older patients) to be much lower and thus slightly unrealistic in our data.

Finally, from our progress report, we discovered that only a few variables have an absolute (linear) correlation with the target variable above 0.1 (see Appendix). We examined the distribution of these important variables to gain a better understanding of their meaning (see Appendix). Functional assessment and Activity of Daily Living (ADL) were both uniformly distributed from 0 to 10 where a lower score indicated a worse result for the patient. Mini-Mental State Examination (MMSE) was similarly uniformly distributed where a lower score indicates a worse patient life but scaled from 0 to 30. Memory complaints and behavioural problems are both binary variables where 1 indicates complaints/problems, and they occurred in roughly one fifth of all patients.

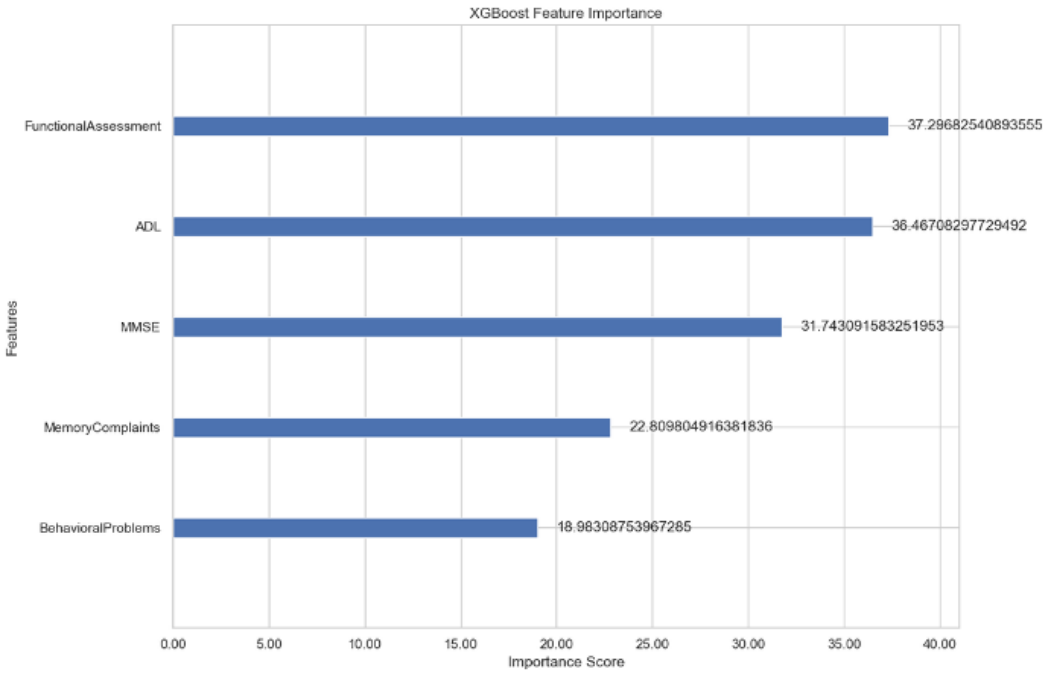
## 3. Analysis and Results

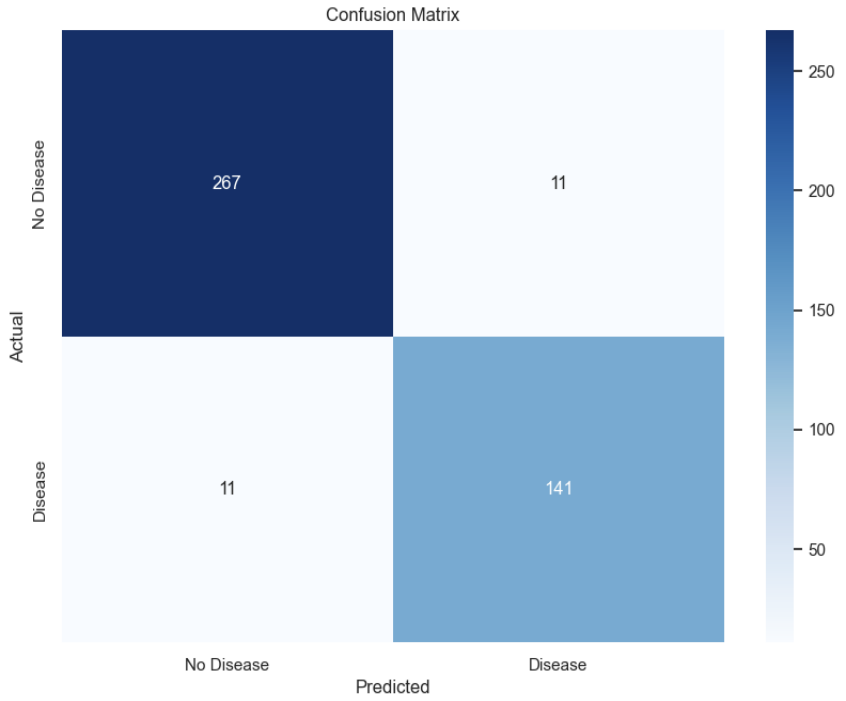
### *3.1. XGBoost*

XGBoost (Extreme Gradient Boosting) is a high-performance machine learning algorithm based on gradient boosted decision trees. This model was selected for the AD prediction task because it outperforms linear models in capturing the complex and non-linear relationships inherent in cognitive and behavioral assessments. It also handles interaction effects that are present in AD indicators, for example, functional with cognitive decline, while reducing the risk of overfitting, which is crucial with medical data where noise and variability are common.

We used XGBoost's feature importance scores to determine which variables contribute most strongly to classification. Top drivers include Functional Assessment, showing that reduced functional ability is highly indicative of AD. Activities of Daily Living (ADL) declines and low scores signal early impairment and correlate strongly with disease beginning. MMSE (Cognitive Performance) plays a major role but is slightly less influential than physical/functional metrics. Report of memory complaints and presence of behavioural problems are important predictors too. These insights show that functional capability and daily living performance are more predictive of AD than subjective memory concerns alone. This aligns with clinical patterns where functional decline often precedes or accompanies cognitive deterioration. The feature importance is shown below.

To evaluate the model's predictive ability, we examined the confusion matrix. The model correctly classifies 95% of all patients. Among patients predicted to have AD, 93% truly have the disease (high precision). The model successfully identifies 93% of all AD cases, missing only 7% (high recall). These results confirm that the model not only performs well overall but also maintains strong performance for both identifying AD and avoiding false alarms.




*3.2. Logistic Regression*

We developed a Logistic Regression baseline model for the features that showed importance in the XGBoost model, i.e., the results of cognitive and functional assessments. These predictors have a significance level $p < 0.001$, supporting the interpretation that these are the most reliable indicators of AD presence in a patient, as shown in the following table.

| Variable | Coefficient | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|
| Intercept | 3.8610 | 0.000 | 3.407 | 4.315 |
| MMSE | -0.1072 | 0.000 | -0.123 | -0.091 |
| FunctionalAssessment | -0.4441 | 0.000 | -0.495 | -0.393 |
| MemoryComplaints | 2.5896 | 0.000 | 2.266 | 2.913 |
| BehavioralProblems | 2.4733 | 0.000 | 2.119 | 2.828 |
| ADL | -0.4139 | 0.000 | -0.464 | -0.364 |

The baseline model demonstrated good discriminative ability with an area under the curve (AUC) of 0.907. The following table presents the key results of the baseline model:

| Metric | Value | Interpretation |
|---|---|---|
| Accuracy | 0.846 | Model correctly classifies 84.5% of the patients based on their cognitive and functional assessments. |
| Precision | 0.806 | Of all the patients predicted to have AD, 74% actually have it. |
| Recall | 0.743 | Model identifies 80.5% of actual AD cases. |
| Specificity | 0.902 | The proportion of actual negative cases that are correctly predicted as negative. |
| ROC-AUC | 0.907 | In 90.7% of randomly paired cases, the model correctly assigns a higher probability of diagnosis to the positive case. |

Although the accuracy of this model is lower than the accuracy of the XGBoost, it is more interpretable. Therefore, we can use it to make inferences and test the hypotheses we outlined for this research.

Based on the literature, we expected positive coefficients for both Age and Family History, indicating higher log-odds of AD, and a negative coefficient for Education. Furthermore, we hypothesize a negative coefficient for the interaction term Age × Education, suggesting that as education levels rise, the effect of age on the probability of developing AD decreases. Contrary to our hypothesis, neither age nor family history of AD showed significant associations with AD diagnosis. Education level and the age × education interaction term are non-significant, providing no evidence for educational moderation of age-related AD risk.

The second hypothesis is that lower MMSE scores and worse functional status are strongly associated with higher odds of Alzheimer's. Similarly, hypertension, diabetes, cardiovascular disease, higher LDL and triglycerides, and lower HDL are associated with higher odds of Alzheimer's, independent of age and sex. As seen in the baseline and XGBoost models, cognitive and functional measures showed strong

associations with AD diagnosis. The MMSE score demonstrated a strong negative effect, as did functional assessment scores and ADL scores. Memory complaints and behavioral problems were positively associated with AD diagnosis. In contrast, vascular risk factors showed no significant associations, as determined by their p-values shown in the table below.

Finally, we evaluate the impact of lifestyle factors and demographic interactions. We hypothesize that adverse behaviors, such as physical inactivity, poor diet, sleep disturbances, depression, and substance use, are associated with increased log-odds of AD. The model also tests the Age × Gender interaction to assess if the age-related risk is steeper for women, and the Ethnicity variable to determine if diagnostic probabilities differ across groups when holding cognitive deterioration constant. None of these factors reached statistical significance, also summarized in the table below. This suggests that there is no evidence to support these hypotheses.

Our analysis reveals a disconnect between established AD risk factors and diagnostic predictors. Traditional risk factors, such as age, family history, lifestyle, likely influence whether someone develops AD over their lifetime. However, once cognitive decline is measurable, these historical factors become diagnostically irrelevant, and the current functional status becomes more determining. The results are summarized in the following table.

| Hypothesis | Feature | Coefficient | P-Value |
|---|---|---|---|
| H1: Demographics & Education | Age | -0.0205 | 0.0897 |
| H1: Demographics & Education | FamilyHistoryAlzheimers | -0.0751 | 0.6054 |
| H1: Demographics & Education | EducationLevel | -0.6790 | 0.2463 |
| H1: Demographics & Education | Age:EducationLevel | 0.0078 | 0.3121 |
| H2: Clinical Indicators | Hypertension | 0.1844 | 0.2921 |
| H2: Clinical Indicators | Diabetes | 0.0256 | 0.8877 |
| H2: Clinical Indicators | CardiovascularDisease | 0.1528 | 0.3799 |
| H2: Clinical Indicators | CholesterolLDL | -0.0031 | 0.0354 |
| H2: Clinical Indicators | CholesterolTriglycerides | 0.0008 | 0.2089 |
| H2: Clinical Indicators | CholesterolHDL | 0.0046 | 0.0942 |
| H3: Lifestyle & Behavioral | PhysicalActivity | -0.0051 | 0.8133 |
| H3: Lifestyle & Behavioral | DietQuality | 0.0122 | 0.5761 |
| H3: Lifestyle & Behavioral | SleepQuality | -0.0541 | 0.1297 |
| H3: Lifestyle & Behavioral | Depression | 0.0547 | 0.7235 |
| H3: Lifestyle & Behavioral | AlcoholConsumption | -0.0077 | 0.4780 |
| H3: Lifestyle & Behavioral | Smoking | -0.2105 | 0.1313 |

| | | | |
|---|---|---|---|
| H4: Age-Gender Interaction | Age | -0.0147 | 0.1312 |
| H4: Age-Gender Interaction | Gender | -0.7368 | 0.4863 |
| H4: Age-Gender Interaction | Age:Gender | 0.0091 | 0.5170 |
| H5: Ethnicity | Ethnicity | -0.0447 | 0.4866 |

## 4. Discussion and Recommendations

Our analysis highlights that although age, lifestyle habits and vascular conditions are widely cited in academic literature as long-term contributors to Alzheimer's development, these factors become diagnostically irrelevant when functional deterioration is measurable. Neither demographic attributes nor medical history variables were statistically significant in predicting diagnosis, suggesting that screening patients based on background characteristics does not meaningfully support clinical decision making once symptoms emerge.

Instead, diagnosis relies mainly on current functional and cognitive performance, with MMSE, Functional Assessment, ADL scores and patient-reported memory and behavioural complains consistently ranking as the strongest predictors across both logistic regression and XGBoost models. This indicates that Alzheimer's should not be approached as a demographic risk-profiling exercise but rather, as a functional impairment evaluation. Therefore, hospitals should shift resources toward strengthening structured cognitive and functional assessment protocols, integrating patient-reported symptoms into intake procedure and also, reducing reliance on demographic or lifestyle screening variables that do not improve diagnostic accuracy. In practice, this means prioritising tools and workflows that measure present cognitive status rather than attempting to infer diagnosis from historical exposure or risk, as doing so is both more predictive and clinically meaningful.

## 5. Conclusion

Our finding strongly supports prioritising functional and cognitive assessment in Alzheimer's disease diagnosis. MMSE, Functional Assessment and ADL scores consistently emerged as the most influential predictors, which indicates that present cognitive performance rather than demographic or historical risk factors drives diagnostic accuracy. The continued significance of subjective indicators such as memory complaints and behavioural problems further highlights that patient-reported symptoms provide meaningful diagnostic value when combined with objective testing. Overall, these results suggests that Alzheimer's detection should focus less on profiling long-term risks and more on evaluating current cognitive impairment. While future work using real clinical data is needed to confirm the generalizability of our findings, the evidence clearly points toward functional decline, and not background characteristics as the most reliable foundation for Alzheimer's evaluation.

**References**


Antor, M. B., Jamil, A. H. M. S., Mamtaz, M., Khan, M. M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict Alzheimer's disease. *Journal of Healthcare Engineering*, *2021*, Article 9917919. https://doi.org/10.1155/2021/9917919

Arjaria, S. K., Rathore, A. S., Bisen, D., & Bhattacharyya, S. (2024). Performances of machine learning models for diagnosis of Alzheimer's disease. *Annals of Data Science*, *11*(1), 307–335. https://doi.org/10.1007/s40745-022-00452-2

Ezzati, A., Zammit, A. R., Harvey, D. J., Habeck, C., Hall, C. B., & Lipton, R. B. (2019). Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *Journal of Alzheimer's Disease*, *71*(4), 1027–1036. https://doi.org/10.3233/JAD-190262

Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). Early-stage Alzheimer's disease prediction using machine learning models. *Frontiers in Public Health*, *10*, Article 853294. https://doi.org/10.3389/fpubh.2022.853294

Neelaveni, J., & Devasana, M. S. G. (2020). Alzheimer disease prediction using machine learning algorithms. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 101–104). IEEE. https://doi.org/10.1109/ICACCS48705.2020.9074248

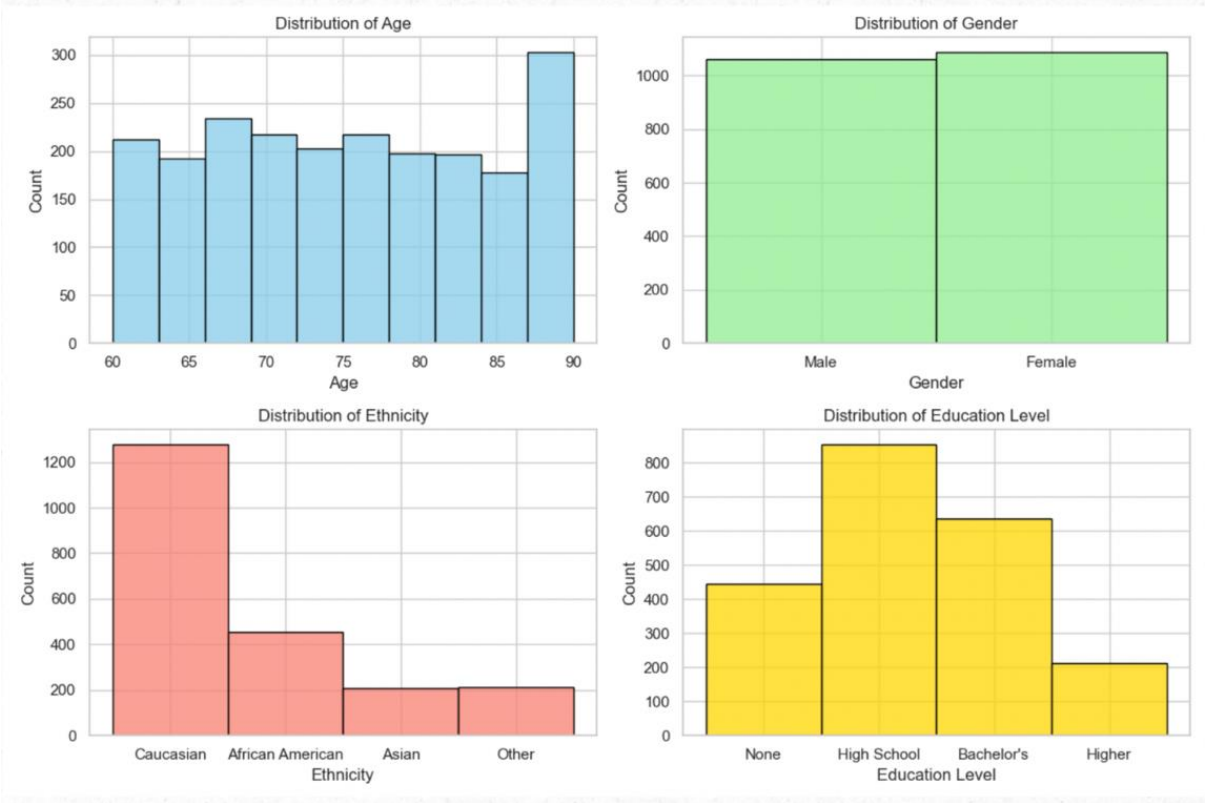
Rani, P., Iwendi, C., Lamba, R., Sachdeva, R. K., & Kumar, K. (2024). A machine learning model for Alzheimer's disease prediction. *IET Cyber-Physical Systems: Theory & Applications*, *9*(2), 125–134. https://doi.org/10.1049/cps2.12090

Wang, L., Li, P., Hou, M., Zhang, X., Cao, X., & Li, H. (2021). Construction of a risk prediction model for Alzheimer's disease in the elderly population. *BMC Neurology*, *21*, Article 271. https://doi.org/10.1186/s12883-021-02276-8

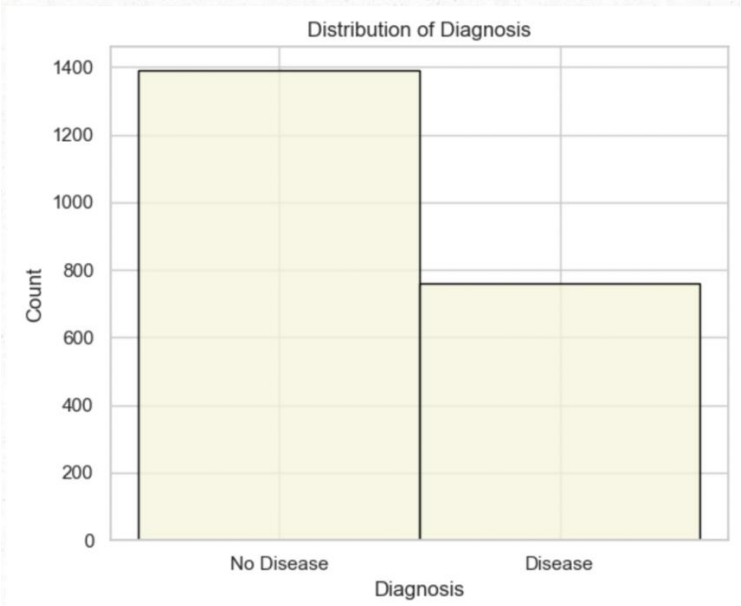
Wang, X., Zhou, S., Ye, N., Li, Y., Zhou, P., Chen, G., & Hu, H. (2024). Predictive models of Alzheimer's disease dementia risk in older adults with mild cognitive impairment: A systematic review and critical appraisal. *BMC Geriatrics*, *24*, Article 531. https://doi.org/10.1186/s12877-024-05044-8

## Appendix

### Distribution of Demographic Variables



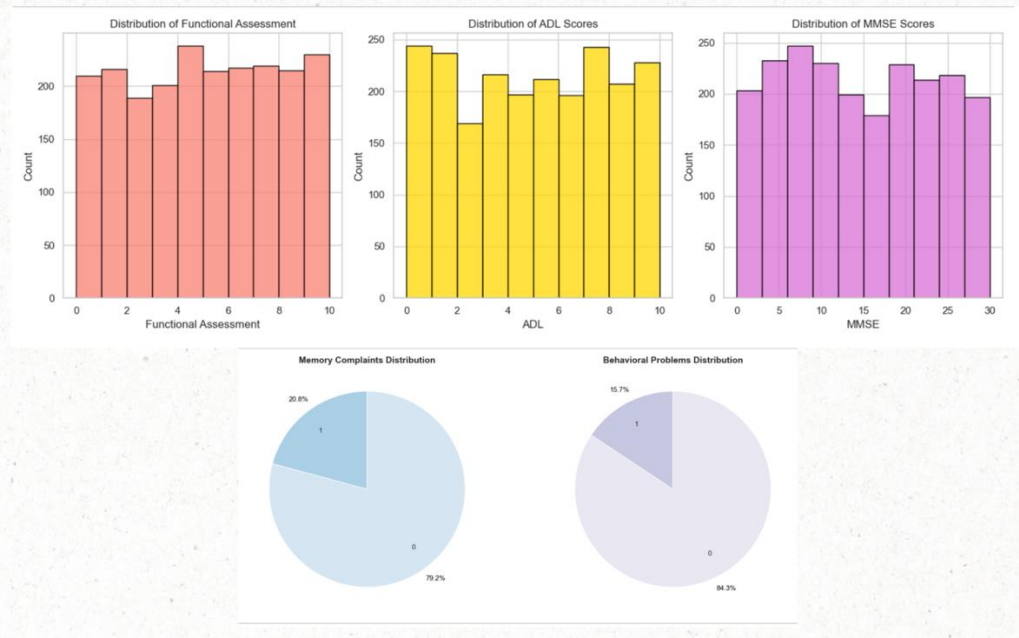

### Distribution of Target Variable



**Strong Correlation with Target Variable**

| | Diagnosis |
|---|---|
| **Diagnosis** | 1.00 |
| **MemoryComplaints** | 0.31 |
| **BehavioralProblems** | 0.22 |
| **MMSE** | -0.24 |
| **ADL** | -0.33 |
| **FunctionalAssessment** | -0.36 |

**Distribution of Key Target Variables**



## Features Results for Logistic Regression models

Multivariate Analysis (All Features)

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | 5.2976 | 0.0000 |
| MemoryComplaints | 2.5967 | 0.0000 |
| BehavioralProblems | 2.5053 | 0.0000 |
| ADL | -0.4216 | 0.0000 |
| FunctionalAssessment | -0.4482 | 0.0000 |
| MMSE | -0.1071 | 0.0000 |
| CholesterolLDL | -0.0029 | 0.0501 |
| CholesterolHDL | 0.0047 | 0.0857 |
| SleepQuality | -0.0554 | 0.1270 |
| Smoking | -0.2138 | 0.1316 |
| HeadInjury | -0.3272 | 0.1411 |
| Age | -0.0103 | 0.1429 |
| EducationLevel | -0.0880 | 0.2144 |
| CholesterolTriglycerides | 0.0007 | 0.2401 |
| Hypertension | 0.1979 | 0.2650 |
| Confusion | -0.1542 | 0.3337 |
| CardiovascularDisease | 0.1676 | 0.3412 |
| AlcoholConsumption | -0.0087 | 0.4283 |
| FamilyHistoryAlzheimers | -0.1106 | 0.4563 |
| Disorientation | -0.1206 | 0.4923 |
| DifficultyCompletingTasks | 0.1008 | 0.5625 |
| Ethnicity | -0.0374 | 0.5650 |
| DiastolicBP | 0.0019 | 0.6047 |
| BMI | -0.0044 | 0.6221 |
| DietQuality | 0.0100 | 0.6513 |
| Depression | 0.0661 | 0.6721 |
| PersonalityChanges | -0.0711 | 0.6965 |
| Gender | -0.0479 | 0.7074 |
| PhysicalActivity | -0.0073 | 0.7399 |
| SystolicBP | -0.0007 | 0.7726 |
| CholesterolTotal | 0.0002 | 0.8745 |
| Diabetes | 0.0153 | 0.9331 |
| Forgetfulness | 0.0036 | 0.9795 |

Demographics Analysis

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | -0.2814 | 0.4716 |
| Age | -0.0017 | 0.7379 |
| Gender | -0.0886 | 0.3272 |
| Ethnicity | -0.0286 | 0.5313 |
| EducationLevel | -0.1027 | 0.0410 |

Lifestyle Analysis

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | -0.3607 | 0.2067 |
| BMI | 0.0075 | 0.2309 |
| Smoking | -0.0251 | 0.8020 |
| AlcoholConsumption | -0.0029 | 0.7117 |
| PhysicalActivity | 0.0044 | 0.7829 |

DietQuality 0.0079 0.6104
SleepQuality -0.0677 0.0084

Medical History Analysis

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | -0.5697 | 0.0000 |
| FamilyHistoryAlzheimers | -0.1723 | 0.1027 |
| CardiovascularDisease | 0.1906 | 0.1322 |
| Diabetes | -0.1915 | 0.1382 |
| Depression | -0.0382 | 0.7359 |
| HeadInjury | -0.1606 | 0.3137 |
| Hypertension | 0.2071 | 0.0971 |

Clinical Measurements Analysis

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | -0.7002 | 0.1235 |
| SystolicBP | -0.0012 | 0.4774 |
| DiastolicBP | 0.0005 | 0.8317 |
| CholesterolTotal | 0.0003 | 0.7596 |
| CholesterolLDL | -0.0015 | 0.1579 |
| CholesterolHDL | 0.0037 | 0.0570 |
| CholesterolTriglycerides | 0.0004 | 0.3200 |

Cognitive & Functional Analysis

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | 3.8610 | 0.0 |
| MMSE | -0.1072 | 0.0 |
| FunctionalAssessment | -0.4441 | 0.0 |
| MemoryComplaints | 2.5896 | 0.0 |
| BehavioralProblems | 2.4733 | 0.0 |
| ADL | -0.4139 | 0.0 |

Symptoms Analysis

| Feature | Coefficient | P-Value |
|---|---|---|
| intercept | -0.5482 | 0.0000 |
| Confusion | -0.0986 | 0.3825 |
| Disorientation | -0.1447 | 0.2500 |
| PersonalityChanges | -0.1278 | 0.3182 |
| DifficultyCompletingTasks | 0.0511 | 0.6782 |
| Forgetfulness | -0.0052 | 0.9578 |

========== H1: Baseline + Age/Edu Interaction ==========
Generalized Linear Model Regression Results
==========

| | | | |
|---|---|---|---|
| Dep. Variable: | Diagnosis | No. Observations: | 2149 |
| Model: | GLM | Df Residuals: | 2139 |
| Model Family: | Binomial | Df Model: | 9 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -801.41 |

```
Date:               Tue, 02 Dec 2025   Deviance:                    1602.8
Time:                       21:16:27   Pearson chi2:                5.22e+03
No. Iterations:                    6   Pseudo R-squ. (CS):           0.4251
Covariance Type:           nonrobust
==============================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
intercept                5.5366      0.950      5.830      0.000       3.675       7.398
MMSE                    -0.1068      0.008    -13.192      0.000      -0.123      -0.091
FunctionalAssessment    -0.4435      0.026    -17.052      0.000      -0.494      -0.393
MemoryComplaints         2.6013      0.166     15.712      0.000       2.277       2.926
BehavioralProblems       2.4942      0.182     13.738      0.000       2.138       2.850
ADL                     -0.4149      0.026    -16.172      0.000      -0.465      -0.365
Age                     -0.0205      0.012     -1.697      0.090      -0.044       0.003
FamilyHistoryAlzheimers -0.0751      0.145     -0.517      0.605      -0.360       0.210
EducationLevel          -0.6790      0.586     -1.159      0.246      -1.827       0.469
Age:EducationLevel       0.0078      0.008      1.011      0.312      -0.007       0.023
==============================================================================================

==================== H2: Baseline + Clinical Indicators ====================
            Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Diagnosis   No. Observations:                 2149
Model:                            GLM   Df Residuals:                     2135
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -797.37
Date:               Tue, 02 Dec 2025   Deviance:                       1594.7
Time:                       21:16:28   Pearson chi2:                 5.92e+03
No. Iterations:                    6   Pseudo R-squ. (CS):           0.4272
Covariance Type:           nonrobust
==============================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
intercept                4.5301      0.650      6.972      0.000       3.257       5.803
MMSE                    -0.1074      0.008    -13.218      0.000      -0.123      -0.091
FunctionalAssessment    -0.4449      0.026    -16.988      0.000      -0.496      -0.394
MemoryComplaints         2.5974      0.166     15.674      0.000       2.273       2.922
BehavioralProblems       2.4929      0.182     13.670      0.000       2.135       2.850
ADL                     -0.4187      0.026    -16.245      0.000      -0.469      -0.368
Hypertension             0.1844      0.175      1.053      0.292      -0.159       0.527
Diabetes                 0.0256      0.181      0.141      0.888      -0.330       0.381
CardiovascularDisease    0.1528      0.174      0.878      0.380      -0.188       0.494
CholesterolLDL          -0.0031      0.001     -2.104      0.035      -0.006      -0.000
CholesterolTriglycerides 0.0008      0.001      1.256      0.209      -0.000       0.002
CholesterolHDL           0.0046      0.003      1.674      0.094      -0.001       0.010
Age                     -0.0099      0.007     -1.416      0.157      -0.024       0.004
Gender                  -0.0561      0.126     -0.445      0.656      -0.303       0.191
==============================================================================================

==================== H3: Baseline + Lifestyle ====================
            Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Diagnosis   No. Observations:                 2149
Model:                            GLM   Df Residuals:                     2136
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -801.01
Date:               Tue, 02 Dec 2025   Deviance:                       1602.0
Time:                       21:16:28   Pearson chi2:                 4.69e+03
```

No. Iterations: 6 Pseudo R-squ. (CS): 0.4253
Covariance Type: nonrobust

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 4.4409 | 0.462 | 9.615 | 0.000 | 3.536 | 5.346 |
| MMSE | -0.1074 | 0.008 | -13.278 | 0.000 | -0.123 | -0.092 |
| FunctionalAssessment | -0.4457 | 0.026 | -17.078 | 0.000 | -0.497 | -0.395 |
| MemoryComplaints | 2.5928 | 0.166 | 15.663 | 0.000 | 2.268 | 2.917 |
| BehavioralProblems | 2.4712 | 0.182 | 13.610 | 0.000 | 2.115 | 2.827 |
| ADL | -0.4151 | 0.026 | -16.219 | 0.000 | -0.465 | -0.365 |
| PhysicalActivity | -0.0051 | 0.022 | -0.236 | 0.813 | -0.048 | 0.037 |
| DietQuality | 0.0122 | 0.022 | 0.559 | 0.576 | -0.031 | 0.055 |
| SleepQuality | -0.0541 | 0.036 | -1.515 | 0.130 | -0.124 | 0.016 |
| Depression | 0.0547 | 0.155 | 0.354 | 0.723 | -0.248 | 0.358 |
| AlcoholConsumption | -0.0077 | 0.011 | -0.710 | 0.478 | -0.029 | 0.014 |
| Smoking | -0.2105 | 0.140 | -1.509 | 0.131 | -0.484 | 0.063 |
| BMI | -0.0035 | 0.009 | -0.398 | 0.690 | -0.021 | 0.014 |

========== H4: Baseline + Age-Gender Interaction ==========

Generalized Linear Model Regression Results

Dep. Variable: Diagnosis No. Observations: 2149
Model: GLM Df Residuals: 2133
Model Family: Binomial Df Model: 15
Link Function: Logit Scale: 1.0000
Method: IRLS Log-Likelihood: -796.45
Date: Tue, 02 Dec 2025 Deviance: 1592.9
Time: 21:16:28 Pearson chi2: 6.16e+03
No. Iterations: 6 Pseudo R-squ. (CS): 0.4277
Covariance Type: nonrobust

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 4.9877 | 0.837 | 5.957 | 0.000 | 3.347 | 6.629 |
| MMSE | -0.1072 | 0.008 | -13.185 | 0.000 | -0.123 | -0.091 |
| FunctionalAssessment | -0.4450 | 0.026 | -16.965 | 0.000 | -0.496 | -0.394 |
| MemoryComplaints | 2.6010 | 0.166 | 15.681 | 0.000 | 2.276 | 2.926 |
| BehavioralProblems | 2.5036 | 0.183 | 13.688 | 0.000 | 2.145 | 2.862 |
| ADL | -0.4191 | 0.026 | -16.233 | 0.000 | -0.470 | -0.368 |
| Age | -0.0147 | 0.010 | -1.509 | 0.131 | -0.034 | 0.004 |
| Gender | -0.7368 | 1.058 | -0.696 | 0.486 | -2.811 | 1.337 |
| Age:Gender | 0.0091 | 0.014 | 0.648 | 0.517 | -0.018 | 0.036 |
| EducationLevel | -0.0846 | 0.070 | -1.210 | 0.226 | -0.222 | 0.052 |
| Hypertension | 0.1804 | 0.175 | 1.030 | 0.303 | -0.163 | 0.524 |
| Diabetes | 0.0230 | 0.182 | 0.126 | 0.899 | -0.333 | 0.379 |
| CardiovascularDisease | 0.1612 | 0.174 | 0.925 | 0.355 | -0.180 | 0.503 |
| CholesterolLDL | -0.0030 | 0.001 | -2.054 | 0.040 | -0.006 | -0.000 |
| CholesterolTriglycerides | 0.0008 | 0.001 | 1.237 | 0.216 | -0.000 | 0.002 |
| CholesterolHDL | 0.0046 | 0.003 | 1.692 | 0.091 | -0.001 | 0.010 |

```
==================== H5: Baseline + Ethnicity ====================
            Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Diagnosis   No. Observations:                 2149
Model:                            GLM   Df Residuals:                     2142
Model Family:                Binomial   Df Model:                            6
Link Function:                 Logit   Scale:                          1.0000
Method:                         IRLS   Log-Likelihood:                -803.75
Date:               Tue, 02 Dec 2025   Deviance:                       1607.5
Time:                       21:16:28   Pearson chi2:                 5.04e+03
No. Iterations:                    6   Pseudo R-squ. (CS):             0.4238
Covariance Type:           nonrobust
==============================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
intercept              3.8923      0.236     16.488      0.000       3.430       4.355
MMSE                  -0.1072      0.008    -13.283      0.000      -0.123      -0.091
FunctionalAssessment  -0.4443      0.026    -17.097      0.000      -0.495      -0.393
MemoryComplaints       2.5890      0.165     15.711      0.000       2.266       2.912
BehavioralProblems     2.4703      0.181     13.654      0.000       2.116       2.825
ADL                   -0.4138      0.026    -16.221      0.000      -0.464      -0.364
Ethnicity             -0.0447      0.064     -0.696      0.487      -0.171       0.081
==============================================================================================
```