Predicting Alzheimer's Diagnosis Using Lifestyle and Health Data

Authors: Christos Symeou, Leontios Ioannou

Course: EPL448 – Data mining on the web

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 analyse 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 behaviour of the target variable. This helped us identify any correlations, and feature types that would influence our modelling approach.

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 the hyperparameters and try to improve the performance even further.

2. Exploratory Data Analysis

Before starting the preprocessing and modelling, 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 modelling.

```
Country                              object
Age                                   int64
Gender                               object
Education Level                       int64
BMI                                 float64
Physical Activity Level              object
Smoking Status                       object
Alcohol Consumption                  object
Diabetes                             object
Hypertension                         object
Cholesterol Level                    object
Family History of Alzheimer's        object
Cognitive Test Score                  int64
Depression Level                     object
Sleep Quality                        object
Dietary Habits                       object
Air Pollution Exposure               object
Employment Status                    object
Marital Status                       object
Genetic Risk Factor (APOE-ε4 allele) object
Social Engagement Level              object
Income Level                         object
Stress Levels                        object
Urban vs Rural Living                object
Alzheimer's Diagnosis                object
dtype: object
```

Figure 1. Data types of each feature in the Alzheimer's prediction dataset.

2.2 Target Variable: Alzheimer's Diagnosis

The target variable in this dataset is **Alzheimer's Diagnosis**, which indicates whether 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 analysing 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 labelled **"No"**
- The remaining **41.3%** are labelled **"Yes"**

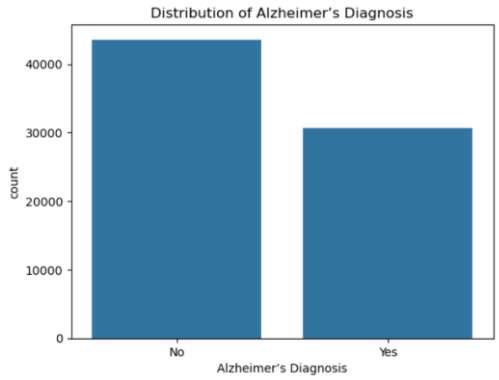

Figure 2. Distribution of the target variable, showing the number of individuals with and without an Alzheimer's diagnosis.

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.

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 analysed 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 most of the cases clustering around **75 to 90+** years.

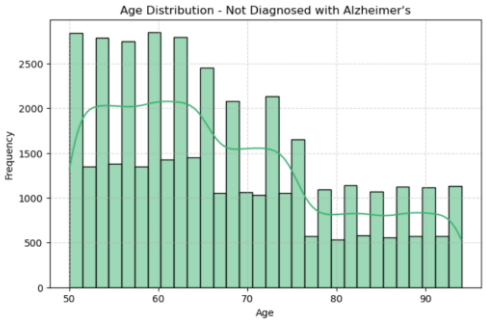

Figure 3. Age distribution of individuals not diagnosed with Alzheimer's.

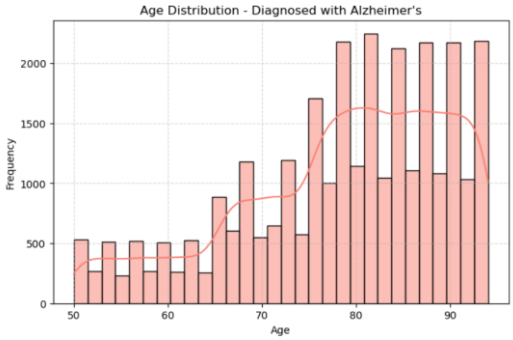

Figure 4. Age distribution of individuals diagnosed with Alzheimer's

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.

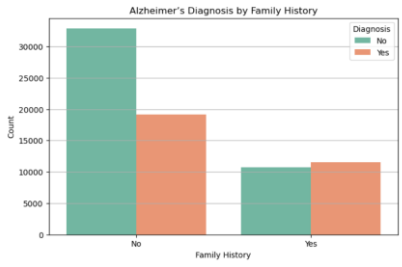

Figure 5. Distribution of Alzheimer's diagnosis based on family history.

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.

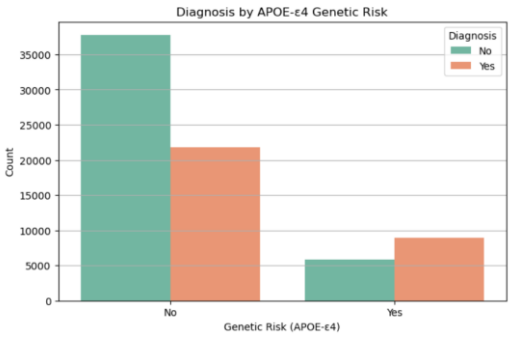

Figure 6. Alzheimer's diagnosis by APOE-ε4 genetic risk.

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

2.4. Secondary Features Analysis

In this section, we explore additional features that may be hypothesized to influence Alzheimer's diagnosis, categorized by demographics, lifestyle, socioeconomic status, medical history, and environmental conditions. However, based on our correlation analysis, we found that most of these secondary features do not exhibit strong linear relationships with the diagnosis outcome.

2.4.1 Demographic Features: Personal Identity

We looked at Gender and Marital Status to see if they had any effect on Alzheimer's diagnosis. The results were similar across both genders, and there were only small differences between marital status groups. So, these features don't seem to have a strong connection to the diagnosis on their own.

2.4.2 Demographic Features: Socioeconomic Status

We looked at Employment Status, Income Level, and Urban vs Rural Living to see if they affect Alzheimer's diagnosis. The differences between the groups were small, and the diagnosis rates looked similar across all categories. This suggests that these socioeconomic features don't have a strong effect on the likelihood of being diagnosed.

2.4.3 Education Level by Alzheimer's Diagnosis

We also looked at Education Level to see if it was related to Alzheimer's diagnosis. The number of diagnosed and non-diagnosed people stays similar across all education levels, with no clear trend. This suggests that in our dataset, education level doesn't seem to have a big impact on the chances of being diagnosed.

2.4.4 Socioeconomic & Lifestyle Factors: Social & Behavioural Elements

We checked Social Engagement Level, Stress Levels, and Sleep Quality to see if they affect Alzheimer's diagnosis. All three features showed very similar results across the different categories, with no

major differences between diagnosed and non-diagnosed people. So, these features don't seem to have a strong effect on the outcome in this dataset.

2.4.5 Socioeconomic & Lifestyle Factors: Lifestyle & Behavioural Habits

We looked at habits like Physical Activity, Smoking, Alcohol Consumption, and Dietary Habits. The differences between groups were small across all these features. This suggests that, at least in our dataset, these lifestyle choices don't seem to have a strong direct impact on Alzheimer's diagnosis.

2.4.6 Health History & Conditions: Medical Conditions & Biometrics

We looked at health-related features like Diabetes, Hypertension, Cholesterol Level, and Depression Level. The diagnosis rates across different categories for each of these features are quite similar, with no strong trends. This suggests that, in our dataset, these medical and biometric conditions don't have a clear link to Alzheimer's diagnosis.

2.4.7 Air Pollution Exposure by Alzheimer's Diagnosis

We also checked Air Pollution Exposure to see if it might be linked to Alzheimer's diagnosis. The results were very similar across the high, medium, and low exposure groups. So, in our dataset, air pollution doesn't seem to have a strong effect on the chances of being diagnosed.

2.4.8 Alzheimer's Diagnosis Proportion by BMI

We explored whether BMI levels were linked to Alzheimer's diagnosis. As shown in Figure 7, the proportions of diagnosed and non-diagnosed individuals are consistent across all BMI values. This suggests that BMI doesn't have a strong relationship with Alzheimer's diagnosis in our dataset.

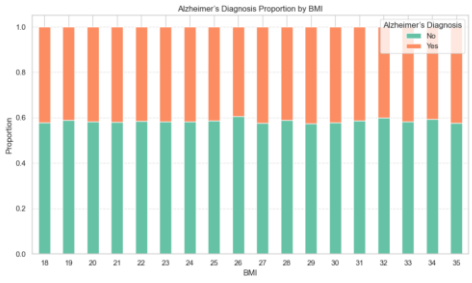

Figure 7. Proportion of Alzheimer's diagnosis across different BMI levels.

2.4.9 Alzheimer's Diagnosis Proportion by Country

We analysed the proportion of Alzheimer's diagnoses across different countries to explore any possible geographic relation. According to the chart while there are small differences—for example, slightly higher proportions in Japan and Mexico—the overall patterns are similar. This suggests that country alone doesn't strongly affect Alzheimer's diagnosis in our dataset.

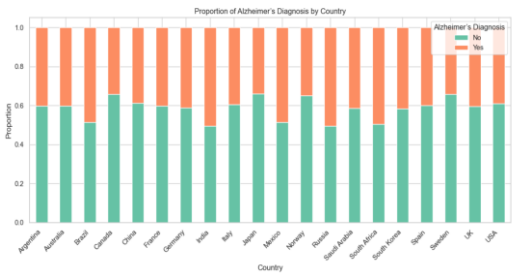

Figure 8. Proportion of Alzheimer's diagnosis by country.

2.5 Correlation Heatmap

We created a correlation heatmap to see how strongly each feature is related to Alzheimer's diagnosis. Most features show **very low correlation**, meaning they don't have a clear linear relationship with the target.

Age stands out with a **moderate positive correlation (~0.42)**, confirming that it's the most important feature. **Family History of Alzheimer's** and **Genetic Risk Factor (APOE-ε4)** also show small positive correlations, while the rest of the features have values close to zero.

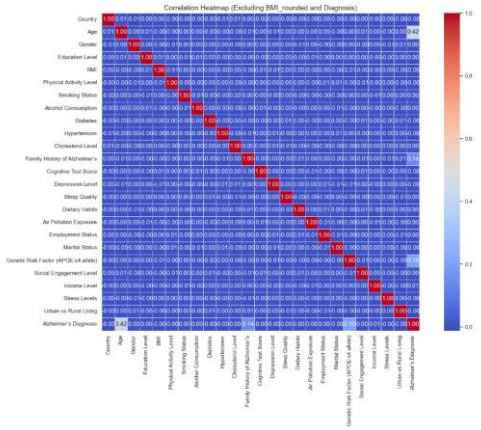

Figure 9. Correlation heatmap of all features and Alzheimer's diagnosis

3. Initial Modelling Approach

Before applying any preprocessing, we split the original dataset into training and test sets using an 80/20 split (test size=0.2) and a fixed random state=42 to ensure reproducibility (training set size: 59426 samples, test set size: 14857 samples). This split was performed prior to scaling and encoding to avoid data leakage.

Then, we performed scaling, encoding, and transformations on the training set to create multiple pre-processed versions. Each version applied different combinations of preprocessing techniques to better prepare the data for modelling. We then applied various classifiers using default hyperparameters to each version of the dataset. This allowed us to observe how different models reacted to different types of input preprocessing.

To evaluate the performance of each model-version combination, we used cross-validation. This helped us compare results more fairly and identify the best-performing models and preprocessing pipelines. These findings guided the selection of the top models we would later fine-tune in Approach B.

3.1 Pre-processed Dataset Versions

To explore how different preprocessing techniques affect model performance, we created four versions of the dataset. Each version applies a unique combination of transformations to prepare the data for training.

3.1.1 Dataset V1

Dataset V1 applies RobustScaler to the numeric features to reduce the influence of outliers, and uses OrdinalEncoder to convert categorical variables into integer labels. This version represents a simple preprocessing strategy and serves as a baseline for evaluating model performance under standard conditions.

3.1.2 Dataset V2

In Dataset V2, the numeric features are transformed using the Yeo-Johnson method, which helps normalize skewed distributions. Categorical features are still encoded using OrdinalEncoder. This

version is designed to evaluate whether addressing skewness in the data improves model accuracy compared to the baseline.

3.1.3 Dataset V3

Dataset V3 extends the preprocessing of V2 by replacing the OrdinalEncoder with OneHotEncoder for categorical features. This approach treats each category independently and is useful for models that perform better with non-ordinal categorical data. The numeric features continue to use the Yeo-Johnson transformation for normalization.

3.1.4 Dataset V4

Dataset V4 includes all the preprocessing steps from V3 and adds dimensionality reduction using PCA. The PCA transformation reduces the number of features while retaining approximately 95% of the original variance. Based on the PCA analysis, around 38 principal components are enough to preserve most of the information, making this version more efficient for training while minimizing information loss. Below is the PCA explained variance chart, which visually confirms that the first few components contribute the most to the total variance.

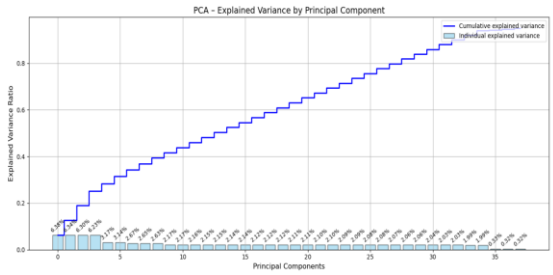

Figure 10. Explained variance by principal component using PCA.

3.2 Model Selection – Evaluation of Classifiers on Datasets V1–V4

In this section, we evaluate how different classifiers perform across the four pre-processed versions of the dataset (V1 to V4). We tested a wide range of models to see which ones work best. These include:

- Random Forest
- AdaBoost
- XGBoost
- CatBoost
- Support Vector Machine (SVC)
- K-Nearest Neighbors (KNN)
- Logistic Regression
- Decision Tree
- Gaussian Naive Bayes.

Evaluation Strategy

Each classifier was trained on every version of the dataset. To keep things fair and consistent, we used **10-fold cross-validation** and focused on the **weighted F1-score** as the main metric. All results were recorded and visualized so we could easily compare performance. The goal here is to spot which models and preprocessing versions perform best, so we can use them later for testing and possibly fine-tuning.

3.3 Model Selection Based on Evaluation Results

After evaluating the performance of 9 different classifiers on the four pre-processed dataset versions (V1–V4), we identified the top-performing models and the most effective preprocessing strategy. The results (shown below) indicate that the two best-performing classifiers on average were: **AdaBoostClassifier** and **CatBoostClassifier**

classifier	mean
AdaBoost	0.721706
CatBoost	0.721223
RandomForest	0.713851
SVC	0.712432
XGBoost	0.711542
LogisticRegression	0.709385
GaussianNB	0.677600
KNeighbors	0.639637
DecisionTree	0.634846

Figure 11. Mean F1-scores for each classifier across all dataset versions

Additionally, when comparing the performance across dataset versions, **Dataset V2** had the highest average F1-score, followed closely by **Dataset V1**. Based on these results, we chose to move forward with both **V1 and V2** for the final evaluation phase.

featureset	mean
v2	0.697919
v1	0.695175
v3	0.693518
v4	0.687708

Figure 12 . Mean F1-scores for each dataset version (V1–V4)

Based on these findings, we proceed with Approach B, where we aim to improve performance through hyperparameter tuning. Specifically, we use pipelines combined with GridSearchCV to fine-tune the top two best-performing models: AdaBoost and CatBoost. Each pipeline integrates the model with its respective best-performing dataset version (V1 or V2), allowing us to optimize the entire process — from preprocessing to prediction — in a structured and reproducible way.

3.4 Alternative Approach: Feature Reduction

As an alternative approach, we experimented with a minimal feature set by retaining only the three most important features — **Age**, **Family History of Alzheimer's**, and **Genetic Risk Factor (APOE-ε4 allele)** — along with **Country**. All other features were dropped from the dataset for this experiment. Despite the significant reduction in dimensionality, the results remained consistent: **AdaBoost (V2)** and **CatBoost (V2)** continued to be the top-performing models. This indicates that the selected key features captured most of the predictive signal in the data, and that the classifiers were robust even when trained on a highly simplified feature set. However, for the main modeling pipeline, we chose to continue using the full feature set without dropping any columns to ensure that any weaker but potentially useful secondary features remained available to the models during training and optimization.

4. Fine-Tuning Top Models Using Pipelines and GridSearchCV (Approach B)

In this section, we use pipelines and GridSearchCV to fine-tune the hyperparameters of the top two best-performing models identified in Approach A. The pipelines incorporate the best-performing feature set versions (Datasets V1 and V2) to ensure optimal preprocessing is consistently applied during model evaluation. This approach helps streamline the workflow and avoid data leakage, while GridSearchCV allows for an exhaustive search over parameter combinations using cross-validation.

4.1 Defining Feature Groups

Before building the pipelines, we first categorized the dataset's features based on their types and preprocessing needs:

num_features: Includes all numerical columns in the dataset, such as Age, Education Level, BMI, and Cognitive Test Score.

features_to_scale: A subset of the numerical features that require scaling or transformation, depending on the chosen preprocessing strategy (e.g., Age and Cognitive Test Score for V1/V2).

cat_features: Categorical features, selected based on the data type (object) for encoding.

This step ensures the correct preprocessing methods are applied to each feature group during pipeline construction.

4.2 Preprocessing Pipelines

To automate and streamline data transformations, we created separate preprocessing pipelines for numerical and categorical features. These pipelines ensure consistency and avoid data leakage during model training and evaluation.

We implemented:

Numerical Pipelines: num_pipeline_v1 scales selected features using RobustScaler, which is resistant to outliers. num_pipeline_v2 uses the PowerTransformer (Yeo-Johnson) method to normalize skewed distributions.

Categorical Encoding: cat_pipeline encodes categorical variables using OrdinalEncoder, handling unknown categories gracefully.

Column Transformers: preprocessor1 combines num_pipeline_v1 for selected scaled features and passes categorical features through cat_pipeline. preprocessor2 does the same but applies num_pipeline_v2 for unskewing instead of scaling.

4.3 Classification Pipelines

We constructed four different pipelines to combine preprocessing steps with classifier training for our top-performing models, **AdaBoostClassifier** and **CatBoostClassifier**. Each classifier is evaluated with both top-performing preprocessing strategies (from Dataset V1 and V2).

AdaBoostPipelines: pipeline_ada_v1 uses preprocessor1 (based on RobustScaler) followed by AdaBoostClassifier.pipeline_ada_v2 uses preprocessor2 (based on Yeo-Johnson transformation) followed by AdaBoostClassifier.

CatBoostPipelines: pipeline_cat_v1 mirrors pipeline_ada_v1 but uses CatBoostClassifier with silent=True to suppress training logs. pipeline_cat_v2 mirrors pipeline_ada_v2 and uses CatBoostClassifier.

These pipelines encapsulate both preprocessing and modelling steps, allowing efficient experimentation with hyperparameter tuning while preserving the structure and integrity of the data.

4.4 Hyperparameter Grids for GridSearchCV

To optimize model performance, we defined separate hyperparameter grids for **AdaBoost** and **CatBoost** classifiers. These grids are used in conjunction with **GridSearchCV** to systematically explore combinations of hyperparameters and identify the best configuration for each pipeline.

For the **AdaBoostClassifier**, we tune the number of estimators (n_estimators), learning rate (learning_rate), and the algorithm type (algorithm), although only 'SAMME' is used in this context.

For the **CatBoostClassifier**, we explore variations in tree depth (depth), number of boosting iterations (iterations), and learning rate (learning_rate), which are critical for controlling model complexity and convergence.

These parameter grids serve as the foundation for the hyperparameter search conducted in Approach B.

4.5 Running GridSearchCV for Hyperparameter Tuning

In this stage, we use GridSearchCV to fine-tune the top-performing classifiers — AdaBoost and CatBoost — using the preprocessing pipelines defined earlier. Each pipeline is evaluated using 5-fold cross-validation to ensure robust performance estimation. We focus on the weighted F1 score to better handle the class imbalance in our dataset.

For each classifier, the grid search identifies:

- The best combination of hyperparameters.
- The highest cross-validation F1 score.
- The final performance on the test set.

This process helps ensure that only the most optimized versions of each model move forward for final evaluation.

4.5.1 Results Discussion

The **GridSearchCV** results for the four pipelines (AdaBoost and CatBoost with both preprocessing approaches) provide valuable insights into model performance and generalization.

1. AdaBoost – V1 and V2:

Best Parameters:

- algorithm: SAMME (default, works well for AdaBoost)
- learning_rate: 1.0 (suggesting a moderate learning rate is best)
- n_estimators: 100 (indicating a sufficient number of weak learners)

Cross-validation (CV) F1 score: 0.7291
The model performs well during 5-fold validation.

Test F1 score: 0.7337
Shows minimal overfitting and strong generalization to unseen data.

2. CatBoost – V1 and V2:

Best Parameters:

- depth: 6 (optimal tree depth)
- iterations: 100 (standard number of boosting rounds)
- learning_rate: 0.01 (slower learning, more stable performance)

CV F1 score: 0.7274
Slightly behind AdaBoost in validation performance.

Test F1 score: 0.7304
Nearly matches AdaBoost in generalization performance.

Key Observations:

- Both **AdaBoost** and **CatBoost** show **very similar test performance**, around 0.73. This suggests either model could be a strong choice.
- **AdaBoost** shows a bit more flexibility due to its higher learning rate and the use of the SAMME algorithm.
- **CatBoost** maintains stability with a low learning rate and reduced tree depth, which aligns with its design goals.

Next Steps:

1. **Further Fine-Tuning:** Explore advanced techniques (e.g., more granular hyperparameter search or ensemble combinations) to push performance further.
2. **Model Selection:** Since performance is very close, more evaluation (e.g., feature importance or larger CV folds) might support a final model decision.

4.6 Final Model Evaluation: AdaBoost with Preprocessor V2

We selected **AdaBoost combined with Preprocessor V2** as the final model based on its strong performance, balanced metrics, and consistent generalization across datasets.

The model achieved a **weighted F1 score of 0.7298 on the training set** and **0.7337 on the test set**, showing minimal overfitting and stable performance. These scores indicate that the model generalized well and was not overly tuned to the training data.

The **classification report on the test set** reveals solid performance across both classes:

- **Class 0 (No Alzheimer's):** Precision = 0.79, Recall = 0.75, F1-score = 0.77
- **Class 1 (Alzheimer's):** Precision = 0.66, Recall = 0.71, F1-score = 0.69
- **Overall accuracy:** 73%
- **Macro & Weighted averages** of all scores ≈ 0.73

```
F1 score on training dataset: 0.7297688233699589
F1 score on test dataset: 0.7337417138619833

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.79      0.75      0.77      8714
           1       0.66      0.71      0.69      6143

    accuracy                           0.73     14857
   macro avg       0.73      0.73      0.73     14857
weighted avg       0.74      0.73      0.73     14857
```

Figure 13. Classification report showing F1 scores, precision, and recall for the test set.

This configuration proved to be the most effective in balancing simplicity, model performance, and generalizability, making it well-suited for future deployment or further enhancements such as ensemble stacking or post-processing for clinical support systems.

4.6.1 Confusion Matrix – AdaBoostClassifier (V2)

The following confusion matrix presents the performance of the final **AdaBoostClassifier (V2)** on the test set after hyperparameter tuning via GridSearchCV.

Interpretation

- *True Positives (4376):* Alzheimer's cases correctly predicted.
- *True Negatives (6508):* Non-Alzheimer's cases correctly predicted.
- *False Positives (2206):* Non-Alzheimer's cases incorrectly predicted as Alzheimer's.
- *False Negatives (1767):* Alzheimer's cases missed by the model.

Key Insights

- The model shows a solid balance between **sensitivity and specificity**, maintaining a **recall for Alzheimer's around 71%**, which is crucial in healthcare applications.
- While the number of **false positives (2206)** is notable, it is acceptable in a medical setting where early detection is prioritized.
- The confusion matrix supports the conclusion that **AdaBoost V2** offers reliable classification performance across both classes and maintains generalization.

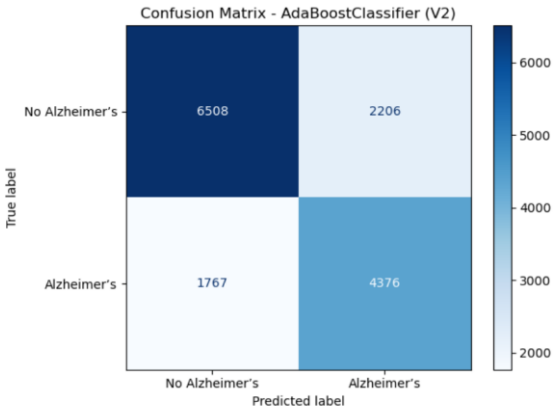

Figure 14. Confusion matrix for the CatBoostClassifier

5. Conclusion

In this project, we explored various preprocessing techniques and classification models to predict Alzheimer's diagnosis using a global dataset. We began with exploratory data analysis to understand the dataset's structure and identify key predictive features. This analysis guided the creation of four distinct pre-processed dataset versions (V1–V4), each incorporating different combinations of scaling, transformation, and encoding methods.

Nine classifiers were evaluated on each version using 10-fold cross-validation, and their performances were compared based on the **weighted F1-score**. The best-performing classifiers were **AdaBoost** and **CatBoost**, with **Dataset V2** and **Dataset V1** emerging as the most effective preprocessing strategies.

We then proceeded to **Approach B**, where we applied **machine learning pipelines** and **GridSearchCV** to fine-tune the top models with their respective optimal preprocessing pipelines. The tuning process revealed that both models performed well and generalized effectively, with **AdaBoost showing slightly better overall stability and balance**.

The final selected model, **AdaBoost with Preprocessor V2**, achieved:

- **F1-score on training set:** 0.7298
- **F1-score on test set:** 0.7337
- **Accuracy:** 0.73
- **Balanced performance** across both classes, as confirmed by the confusion matrix and classification report.

These results demonstrate that classification models can be a valuable tool in predicting Alzheimer's disease risk, particularly when combined with robust preprocessing and model optimization techniques.