University of Pisa

Master's degree in Artificial Intelligence and Data Engineering

Data Mining and Machine Learning project presentation

CogniPredictAD

Student:

Francesco Panattoni

GitHub Repository: https://github.com/Tenshin000/CogniPredictAD/tree/main

ACADEMIC YEAR 2024/2025

Contents

Abstract

This report introduces CogniPredictAD, a data mining and machine learning project designed to analyze clinical and biological parameters from the ADNI dataset, with the goal of predicting final clinical diagnoses (CN, EMCI, LMCI, AD) based on baseline data.

The analysis includes a detailed examination of preprocessing techniques: advanced missing data management, normalizations, and feature engineering and feature vectorization using TF-IDF. We assess the models based on their accuracy and present the results through various evaluations.

Due to the ambiguity of the high predictivity of three values (CDRSB, LDELTOTAL, mPACCdigit), I build two families of models: with and without these features to reduce the risk of overfitting on ADNI.

The modeling phase includes Decision Tree, Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, Multinomial Logistic Regression, and Bagging; hybrid sampling techniques, Grid Search for hyperparameter optimization, and cross-validation are applied. With the full set, the models achieve very high metrics (e.g., accuracy ≈ 0.93, ROC-AUC ≈ 0.989), while by removing the three dominant features, the LightGBM with hybrid sampling retains the best balanced performance (balanced accuracy ≈ 0.7299, ROC-AUC ≈ 0.913). For transparency, an optimized Decision Tree was also selected, and the trees and explanatory rules were saved.

The conclusions highlight good predictive performance on the ADNIMERGE dataset, but caution against possible sample bias and the need for external validation (by adding additional patients to the dataset or through data integration with similar datasets) before any clinical use.

1 Introduction

Early and accurate diagnosis of **Alzheimer's disease** (AD) is a clinical and social priority: intervening before cognitive impairment becomes severe allows for the planning of therapies, treatments, and support strategies, and the testing of interventions that slow decline. However, the disease is complex and multifactorial: clinical signs, cognitive tests, Cerebrospinal Fluid (CSF) biomarkers, genetics (e.g., APOE4), and neuroimaging measures interact in a nontrivial way. For this reason, **Machine Learning** (ML) techniques are particularly well-suited: they can integrate multimodal information, model nonlinear relationships, and identify combinations of features that improve the discrimination between **cognitively normal** (CN), **mild cognitive impairment** (MCI), and full-blown **Alzheimer's subjects** (AD).

A dataset widely used in the literature for these purposes is **ADNI**[1] (**Alzheimer's Disease Neuroimaging Initiative**), a multicenter longitudinal study that collects clinical, cognitive, genetic, CSF, and imaging data from USA and Canada. In this project, I worked with the **ADNIMERGE.csv** tabular file, which is the merged version of the ADNI data and contains repeated visits over time, many clinical variables/biomarkers, and metadata; the notebook shows the direct import of this file as a starting point.

In this project notebooks, baseline visits were selected, extensive cleaning and imputation of missing features was performed, MRI volumes were normalized for ICV, and derived features (biological ratios and cognitive scores) were constructed. Variable selection methods were then applied to create two distinct sets (with and without the three dominant cognitive features).

The modeling compared trees and ensembles (including LightGBM/XGBoost/CatBoost), using hyperparameter optimization and sampling strategies. The CatBoost1 model was chosen as Model1. The LightGBM model, chosen as Model2, maintained good performance even when excluding the dominant features.

[1] The Alzheimer's Disease Neuroimaging Initiative is a longitudinal, multicenter, observational study involving over 60 clinical sites in the United States and Canada. Launched in 2004 by the National Institute on Aging (NIA) in collaboration with the pharmaceutical industry, the initiative aims to develop and validate biomarkers to improve the diagnosis and monitoring of Alzheimer's disease.

Furthermore, XAIModel1 and XAIModel2 represent the explainable decision trees for the dataset with and without the dominant features.

2 Dataset

ADNIMERGE.csv is the **ADNI** merged table used as the main input in the notebook: the copy used by the project contains 16,421 rows (representing visits) and 116 columns before any cleaning and selection, and incorporates repeat visits for each subject (VISCODE, EXAMDATE), identifiers (RID, PTID), and both the initial screening diagnosis (DX_bl) and the more complete diagnosis assigned at the baseline visit (DX).

The structure is mixed but rich: there are demographics (AGE, PTGENDER, PTEDUCAT, PTETHCAT, PTRACCAT, PTMARRY), genetics (APOE4), numerous cognitive and clinical scores (MMSE, CDRSB, ADAS11/13, LDELTOTAL, FAQ, MOCA, TRABSCOR, RAVLT_..., mPACC*), CSF and PET biomarkers (ABETA, TAU, PTAU, FDG, also columns such as PIB and AV45), and MRI volumetric measures (Ventricles, Hippocampus, Entorhinal, Fusiform, MidTemp, WholeBrain, ICV).

ADNIMERGE.csv, however, isn't simply a concatenation: many variables are derived from source files. For example, the variable Hippocampus is derived from the sum of the left/right components (ST29SV + ST88SV) taken from the original *FreeSurfer*[2] files. Therefore, to understand exactly how a measure was calculated, it's good practice to examine the merge script and the data dictionaries of the source tables.

3 Multiclass Problem

As we've seen, we have a different class distribution between DX_bl and DX.

- **DX_bl** can be "CN", "SMC", "EMCI", "LMCI", and "AD". It indicates the screening diagnosis, i.e. the preliminary clinical judgment assigned during the first evaluation visit. It is a Screening diagnosis.

- **DX** can be "CN", "MCI", and "Dementia". It is instead the diagnosis assigned during the baseline visit (denoted by *VISCODE* equal to "bl"), after a more in-depth clinical evaluation.

"AD" and "Dementia" are the same thing despite the different names.

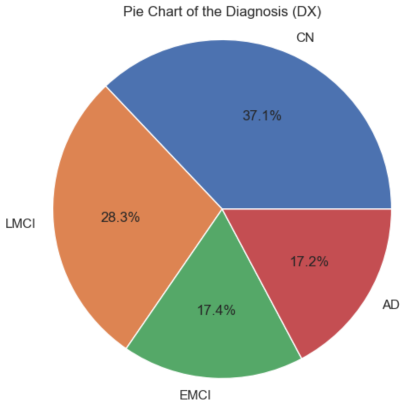
The acronym **SMC** refers to *Subjective Memory Concern*, i.e., cognitively normal (CN) subjects reporting perceived memory issues. Since predicting a subjective perception from objective data is not meaningful. So we reassign it based on the value it has in DX.

Furthermore, we divide MCI into EMCI (Early MCI) and LMCI (Late MCI), assuming that DX_bl values accurately distinguish EMCI and LMCI when DX equals MCI. This is because the division into EMCI and LMCI reflects the degree of cognitive impairment and the risk of progression to dementia.

- **EMCI:** mild cognitive deficits, often detectable only with more sensitive tests. Lower or slower risk of progression to dementia.

- **LMCI:** more marked and evident impairment, greater impact on daily life. Higher risk of progression to Alzheimer's or other dementias.

Therefore, we decide to keep them in the diagnostic prediction.



[2] FreeSurfer is an open-source software designed for the analysis and visualization of neuroimaging data, in particular structural (but also functional or diffusion) MRI scans, and provides a complete processing workflow.

Ultimately, our target variable will be a modified version of the *DX* column, which can now take on four values: "CN", "EMCI", "LMCI", and "AD." This makes our problem a multiclass problem.

4 Data Exploration

Explorations reveal that the dataset has 16,421 rows and 116 columns. However, these records represent the various visits, and we are only interested in the baseline ones. The dataset contains 2,419 useful patients (using "useful" means those who did not have a NULL baseline diagnosis) for the proposed problem.

Many columns contain significant percentages of missing cases. The diagnostic classes of the baseline sample are unbalanced, but not extremely unbalanced.

Demographic and risk analyses show bias in the ADNI sample. Ethnicity is heavily skewed toward white subjects, with high average levels of education (peaks at 16–18 years of schooling), and many married individuals. There are more men than women, but overall the number is not disproportionate. This, however, implies that models may perform worse on more heterogeneous clinical populations.

The *Data Exploration* was then divided into three parts:

1. the preliminary data exploration of the raw dataset;
2. the data exploration after splitting and preparing the actual dataset;
3. the data exploration after Preprocessing to select the classification models.

5 Data Preprocessing

We divided the *Preprocessing* into two phases.

In the first phase, which involved preparing the dataset, we performed all the transformations and cleaning operations that did not involve the risk of *data leakage*, applying them to the entire dataset before splitting it into training and testing.

In the second phase, however, we applied the transformations that could introduce data leakage exclusively to the training set, with the sole exception of imputing missing values: in this case, the *KNN*[3] Imputer was trained on the training set and then used for both training and testing, to ensure consistency and avoid leakage. Preprocessing has in turn been divided into: *Data Cleaning*, *Data Transformation*, *Outlier Detection* and *Data Reduction.*

5.1 Data Preparation

- **Selection of baseline visits only:** The first clinical-operational choice was to work only on baseline visits (VISCODE == "bl"), because the goal is to predict the diagnosis based on the information collected at the first visit;
- **SMC diagnosis management:** Records with *DX_bl* = SMC were realigned using the DX variable and ultimately classified as "CN", for the reasons stated in the chapter "Multiclass Problem";
- **Consolidating bl columns:** Clean up duplicates and merge "baseline" values into main columns;
- **Dirty value handling:** Deleting row with RAVLT_perc_forgetting = -316.667 and RAVLT_forgetting = -19;
- **Text category cleaning (ethnicity, race, marriage):** String values standardized to improve readability and avoid inconsistencies;
- **Encoding of categorical variables:** One-hot encoding (PTETHCAT, PTMARRY), binary mapping (PTGENDER: Male=1, Female=0) and ordinal encoding for DX (CN=0, EMCI=1, LMCI=2, AD=3).
- **Preliminary Feature Reduction:** Removal of columns not relevant to the diagnosis;
- **Splitting train/test:** *Separation into training and test sets while avoiding leakage.*

[3] K-Nearest Neighbors

5.2 Data Cleaning

- **Handling missing values:** Identifying percentages of missing values and using KNN Imputer for continuous variables ;

- **Numeric Value Conversion:** Convert cognitive scales and age from float to int, correcting for approximations due to imputation or format errors.

5.3 Data Transformation

- **Creation of new CSF metrics:** *TAU/ABETA* and *PTAU/ABETA* ratios more predictive than single measures according to the literature;

- **MRI normalization to ICV (Intracranial Volume):** Necessary to correct for differences due to gender and cranial size.

5.4 Outlier Detection

- **Univariate Analysis:** use IQR and Z-score for each column to find outliers;

- **Multivariate Analysis:** Constructs groups (EcogPt, EcogSP, Neuropsych, MRI, MRI/ICV, CSF, CSF/ABETA, mPACC), applies *LOF* and *DBSCAN* to normalized data (*RobustScaler*[4]). Only data points reported by both techniques and with a *LOF_score* greater than 2 are kept to flag them as "extreme";

- **Cleaning up problematic outliers:** Outliers with values that were clearly out of range and therefore deemed highly unlikely were replaced with the mean by class.

5.5 Data Reduction

- **Removal of redundant features:** *ADAS11*, *ADASQ4*, *EcogPtTotal*, *EcogSPTotal*, *mPACCtrailsB*, and *TAU* were removed because they had a high correlation with other features and their informative value was low;

- **Attribute Subset Selection:** I apply four selection methods to the train: *Pearson correlation* ($|r| \geq 0.6$), *mutual information* (top 25), *SelectKBest f_classif* ($k = 25$), and *Recursive Feature Elimination* (RFE with Random Forest). It combines the results by counting how many times each feature appears and retains those selected at least three times, plus other features deemed useful even though they were counted less frequently.

5.6 Some Considerations

5.6.1 Correlation

The dataset contains groups of highly correlated variables (e.g., different neuropsychological scores, ECG components, and MRI volumetric measurements). Rather than eliminating them through aggressive reduction, we decided to retain them and rely on models that are intrinsically robust to correlation. This choice was motivated by two main reasons:

1. **Clinical interpretability:** Correlated variables can describe different facets of the same function or biomarker. Removing them would impoverish medical interpretation;

2. **Complementary predictive value:** Even correlated measures may contain specific variance useful for distinguishing clinical subgroups.

5.6.2 Normalization

During the data preparation process, no global normalization or standardization was applied to all variables.

This choice was driven by one reason: we wanted to **preserve the clinical interpretability**. Maintaining variables in their original units facilitates the medical interpretation of the results and comparability with clinical reference values. Normalization would have made it more difficult to attribute direct clinical significance to the transformed values.

[4] $x_i' = \frac{x_i - \text{median}(X)}{Q_3(X) - Q_1(X)}$

Table 1: Description of the final dataset attributes

Attribute	**Description**	**Category**
DX	Clinical diagnosis at the time of visit: CN, SMC, EMCI, LMCI, AD	Diagnosis
AGE	Participant's age at time of visit	Demographics
PTGENDER	Participant's gender (Male/Female)	Demographics
PTEDUCAT	Years of formal education completed	Demographics
APOE4	Number of APOE $\varepsilon 4$ alleles (0, 1, or 2), a genetic risk factor for Alzheimer's	Demographics
MMSE	Mini-Mental State Examination score (0–30, higher = better)	Clinical Scores
CDRSB	Clinical Dementia Rating - Sum of Boxes (0–18, higher = worse)	Clinical Scores
ADAS13	ADAS-Cog 13-item total score (higher = worse)	Clinical Scores
LDELTOTAL	Logical Memory II delayed recall total score	Clinical Scores
FAQ	Functional Activities Questionnaire – functional impairment score	Clinical Scores
MOCA	Montreal Cognitive Assessment – global cognitive function (0–30)	Clinical Scores
TRABSCOR	Trail Making Test Part B – time in seconds (higher = worse)	Clinical Scores
RAVLT_immediate	RAVLT total immediate recall score (sum over 5 trials)	Clinical Scores
RAVLT_learning	Learning score (Trial 5 minus Trial 1 of RAVLT)	Clinical Scores
RAVLT_perc_forgetting	Percent forgetting from RAVLT (higher = worse)	Clinical Scores
mPACCdigit	Modified Preclinical Alzheimer's Cognitive Composite – Digit Symbol test	Composite Scores
EcogPtMem	Subject self-reported memory complaints (ECog)	ECogPT
EcogPtLang	Subject self-reported language difficulties (ECog)	ECogPT
EcogPtVisspat	Subject self-reported visuospatial difficulties (ECog)	ECogPT
EcogPtPlan	Subject self-reported planning difficulties (ECog)	ECogPT
EcogPtOrgan	Subject self-reported organizational issues (ECog)	ECogPT
EcogPtDivatt	Subject self-reported divided attention issues (ECog)	ECogPT
EcogSPMem	Informant-reported memory complaints (ECog)	ECogSP
EcogSPLang	Informant-reported language issues (ECog)	ECogSP
EcogSPVisspat	Informant-reported visuospatial issues (ECog)	ECogSP
EcogSPPlan	Informant-reported planning problems (ECog)	ECogSP
EcogSPOrgan	Informant-reported organization issues (ECog)	ECogSP
EcogSPDivatt	Informant-reported divided attention issues (ECog)	ECogSP
FDG	FDG PET SUVR – brain glucose metabolism	Biomarkers
PTAU/ABETA	CSF phosphorylated tau protein/Aβ42 ratio	Biomarkers
Hippocampus/ICV	Volume of hippocampus/Intracranial volume ratio from MRI	MRI
Entorhinal/ICV	Volume of the entorhinal cortex/Intracranial volume ratio from MRI	MRI
Fusiform/ICV	Fusiform gyrus volume/Intracranial volume ratio from MRI	MRI
MidTemp/ICV	Middle temporal gyrus volume/Intracranial volume ratio from MRI	MRI
Ventricles/ICV	Volume of ventricles/Intracranial volume ratio from MRI	MRI
WholeBrain/ICV	Whole brain volume/Intracranial volume ratio from MRI	MRI

6 Model Selection

The following classification algorithms were chosen:

1. **Decision Tree:** This model constructs a series of "if → then" rules (split on individual features) to separate classes using a binary tree. Each leaf of the tree corresponds to a prediction. It was chosen because it is immediately interpretable (*XAI*[5]);

2. **Random Forest:** Builds many different decision trees on subsamples of the data and averages their predictions. This reduces variance compared to a single tree and improves robustness to noise, outliers, and collinearity;

3. **Extra Trees[6]:** Similar to Random Forest but chooses more random splits, increasing diversity among trees and often reducing overfitting on noisy features. It was tested to compare with Random Forest and evaluate whether increased randomness improved generalization across the dataset;

4. **XGBoost:** A boosting algorithm that builds trees sequentially, each improving the errors of the previous one. It is highly efficient, regularized, and capable of capturing nonlinear interactions between variables, while also controlling overfitting;

5. **LightGBM:** A gradient boosting implementation designed to be very fast and scalable. It uses techniques (leaf-wise splitting, binning) that make it particularly efficient on heterogeneous datasets;

6. **CatBoost:** A boosting variant that natively handles categorical variables and has robust default hyperparameters to reduce overfitting. It is suitable for working with clinical data;

7. **Multinomial Logistic Regression:** A linear model that estimates the probabilities of membership in each class using a linear combination of features. It requires feature standardization to function properly, which was ensured with StandardScaler in the pipeline. It was included as a simple and interpretable statistical baseline, useful for comparing whether the gain from complex models is consistent with a linear solution;

8. **Bagging[7]:** An ensemble approach that trains several models on different bootstrap samples drawn from the dataset, and then combines their outputs by averaging. The main effect is a reduction in model variance, leading to more stable predictions. It was selected as a baseline method because, in clinical datasets where variability and noise are substantial, bagging provides a straightforward way to assess how much predictive stability can be gained simply by aggregating multiple weak learners.

7 Hyperparameter Selection and Hybrid Sampling

7.1 Grid Search

To optimize the performance of the classifiers, a **Grid Search** procedure with layered cross-validation was adopted. Grid Search was chosen because it allows for a systematic and controlled exploration of the most relevant hyperparameters for each model, ensuring reproducibility and the ability to transparently compare the tested configurations.

7.2 Hybrid Sampling

The dataset also presented a slight imbalance in diagnostic classes, as already discussed in the "Multiclass Problem" chapter.

To address this problem, a **Hybrid Sampling strategy** was applied, combining:

[5] Explainable Artificial Intelligence

[6] Extremely Randomized Trees

[7] Bootstrap Aggregating

[8] Random Under-Sampling

1. **RUS**[8] to reduce the number of instances in the majority classes, preventing the dataset from becoming excessively biased toward synthetic examples;
2. **SMOTENC**[9] to generate new synthetic examples of the minority classes, taking into account the mixed nature of the variables (continuous and categorical).

We kept the ”old” dataset (the one obtained with Preprocessing) and the ”new” one (the resampled one) and tried Hyperparameter Tuning on both.

7.3 The problem with CDRSB, LDELTOTAL, and mPACCdigit

The cognitive scores *CDRSB*[10], *LDELTOTAL*[11] and *mPACCdigit*[12] show exceptionally high predictive power compared to the rest. While this may be advantageous in terms of model accuracy, it also raises the concern of feature dominance: **a small number of variables may disproportionately drive the predictions, while many others contribute minimally**. This imbalance can lead to a form of local overfitting, where **models appear highly effective on the ADNI dataset but lose performance when applied to more heterogeneous clinical populations or external data.**

However, this assumption cannot be verified, as it is equally possible that these three variables are genuine strong predictors of Alzheimer’s diagnosis.

So the issue does not reflect a weakness of the cognitive measures themselves, but rather the possibility of dataset bias: the strength of these predictors may be tied to the specific characteristics of ADNI rather than to generalizable diagnostic patterns.

To address this, the modeling strategy should consider two complementary approaches:

1. building a predictive model that leverages these dominant variables;
2. building an alternative model that excludes them.

Hybrid Sampling will be applied to both datasets and then we’ll compare whether the standard model or the resampled model performs better on the test set.

7.4 Hyperparameter Optimization

In practice, we’ll end up with two models: one that can also use CDRSB, LDELTOTAL, and mPACCdigit, chosen from the standard and resampled models, and one that doesn’t use CDRSB, LDELTOTAL, and mPACCdigit, chosen from the standard and resampled models.

Therefore, we need to run four Grid Searches.

...

8 Classification

8.1 Building Models

...

8.2 Results

...

8.3 Final Decision

...

9 Conclusions

9.1 Real World Applications

...

9.2 Improvements for Future Works

...

9.3 Final Considerations

...

[9] Synthetic Minority Over-sampling Technique for Nominal and Continuous features

[10] Clinical Dementia Rating - Sum of Boxes

[11] Logical Memory II delayed recall total score

[12] Modified Preclinical Alzheimer’s Cognitive Composite – Digit Symbol test