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**.

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 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.

4 Data Preprocessing

The first clinical-operational choice for *Data Preprocessing* 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; this filter reduced the dataset from 16,421 to 2,419 rows.

I discover that the acronym **SMC** stands for "*Subjective Memory Concern*", and refers to *cognitively normal* (CN) subjects who perceive memory difficulties. Introduced in ADNI2 to increase variability in the CN and MCI groups, this category does not correspond to an objective deficit. Therefore, it would be inappropriate to train machine learning models to predict a phenomenon based solely on the patient's subjective perception from objective data. Therefore, observations with DX_bl = "SMC" were realigned using the DX variable, but ultimately all were classified as CN.

...

[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.