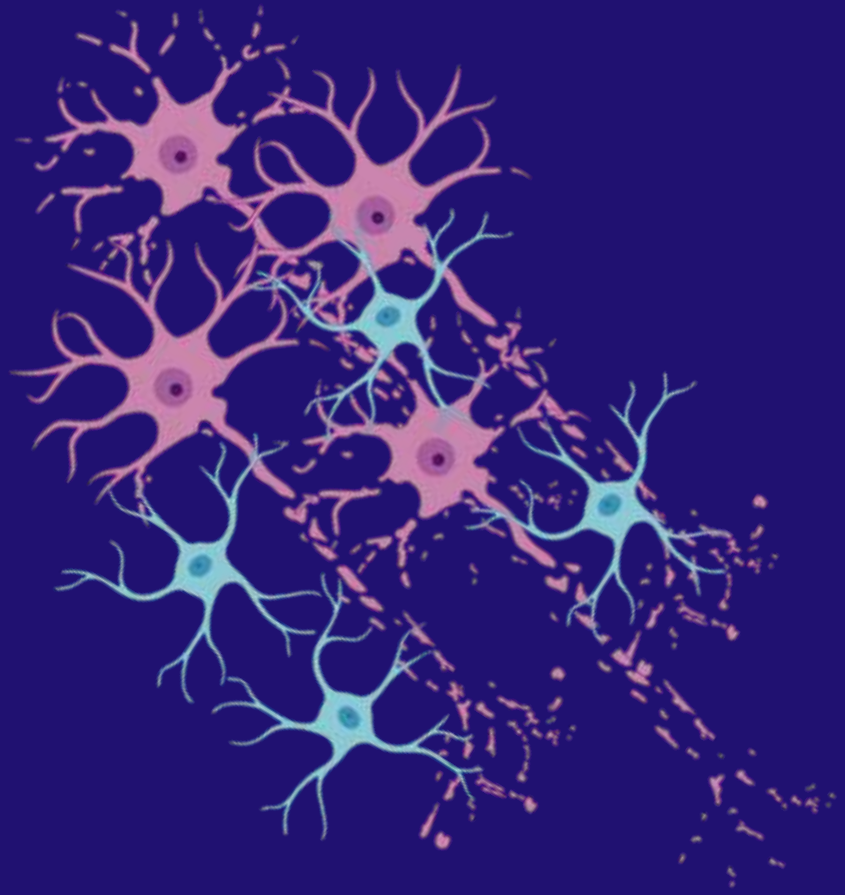
University of Pisa
Department of Information Engineering



CogniPredictAD

Francesco Panattoni

Project for Data Mining and Machine Learning

Table of Contents

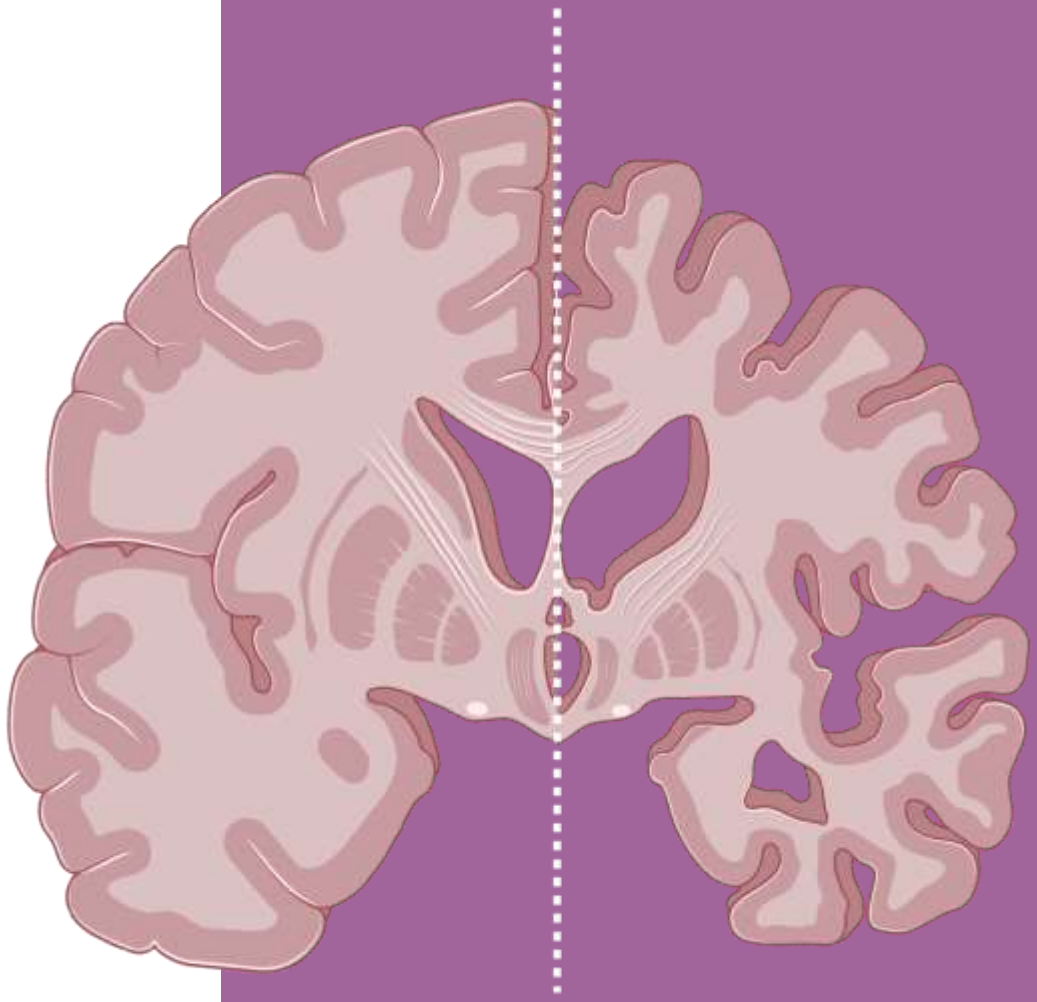
Motivations and Clinical Background

Alzheimer’s Disease

- **Alzheimer’s Disease** is a progressive neurodegenerative disease that affects memory, cognitive function, and daily living skills. It is the most common form of dementia and has no definitive cure.

- It primarily affects the **elderly** and its incidence is increasing as the population ages. It has a **significant social, family, and economic impact**, requiring **long-term care**.

- **Early diagnosis is difficult but crucial** to slowing the progression of the disease and improving quality of life.

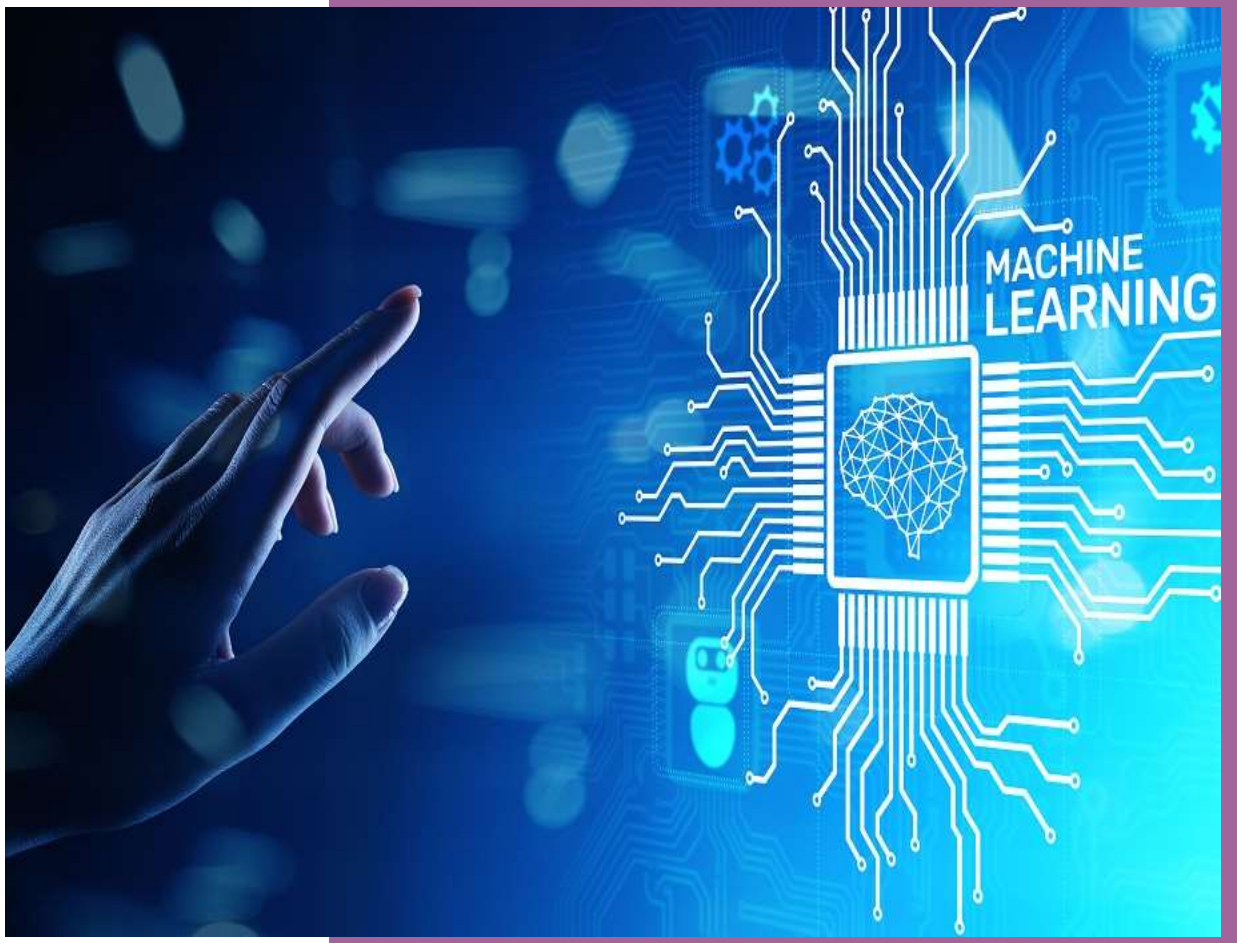


So why Machine Learning?

- **Machine Learning models** could support doctors in diagnosis by **analyzing large amounts of data**, **identifying hidden patterns**, and **improving accuracy and speed**.

- Alzheimer's disease is **multifactorial**. Machine learning helps integrate **complex and nonlinear information**.

- They help **personalize clinical pathways** and **identify at-risk patients** before the most serious symptoms appear.



Dataset ADNI

ADNIMERGE.csv

- The **Alzheimer's Disease Neuroimaging Initiative (ADNI)** is a longitudinal, multicenter, observational study involving over 60 clinical sites in the United States and Canada.
- Launched in **2004** and divided into the following phases: **ADNI1** (2004–2009), **ADNIGO** (2009–2010), **ADNI2** (2011–2016), and **ADNI3** (2016–2022).
- **ADNI4** is the most recent phase of the ADNI study and was initiated in **2022**.
- **16,421 rows x 116 columns.** Its columns are divided into current visit columns and baseline visit columns (with "_bl" suffix) to aid in quick comparison.
- Obtained from the **fusion of clinical data** collected during phases ADNI1, ADNIGO, ADNI2 and ADNI3. Unfortunately, the ADNI4 data has not yet been merged.
- The dataset contains **numerous visits** from different patients, with **associated diagnoses**.

Features of ADNIMERGE.csv

- **Diagnosis (target)**: DX, DX_bl
- **Administrative**: RID, COLPROT, ORIGPROT, PTID, SITE, VISCODE, update_stamp
- **Timestamps**: EXAMDATE
- **Demographics**: AGE, PTGENDER, PTEDUCAT, PTETHCAT, PTRACCAT, PTMARRY, APOE4
- **PET Imaging**: FDG, PIB, AV45, FBB
- **CSF Biomarkers**: ABETA, TAU, PTAU
- **Clinical Scores**: CDRSB, ADAS11, ADAS13, ADASQ4, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, RAVLT_perc_forgetting, LDELTOTAL, DIGITSCOR, TRABSCOR, FAQ, MOCA
- **ECog (self-report)**: EcogPtMem, EcogPtLang, EcogPtVisspat, EcogPtPlan, EcogPtOrgan, EcogPtDivatt, EcogPtTotal
- **ECog (informant-report)**: EcogSPMem, EcogSPLang, EcogSPVisspat, EcogSPPlan, EcogSPOrgan, EcogSPDivatt, EcogSPTotal
- **MRI Imaging**: FLDSTRENG, FSVERSION, IMAGEUID, Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, MidTemp, ICV
- **Composite Scores**: mPACCdigit, mPACCtrailsB
- **Baseline Values**: all variables with the suffix _bl
- **Time Measures**: Years_bl, Month_bl, Month, M

Strengths and Weaknesses of ADNIMERGE.csv

Strengths

- It groups demographic, cognitive, imaging (MRI, PET) and CSF biomarker data into a **single large dataset**, **useful for integrated analyses**.
- **Rigorous post-data acquisition** correction procedures, which reduce technical variability and increase the statistical reliability of the features;
- One of the most **widely used datasets** in Alzheimer's disease research, with well-documented protocols and support for harmonization and comparative studies.

Weaknesses

- Many variables have **numerous missing values**, and the missingness varies depending on the diagnosis or phase of the visit, so it is **not Missing Completely at Random**.
- Participants are predominantly **white**, **highly educated**, **motivated**, and **married**, **reducing their representativeness** of the general population.
- Many features are **highly correlated or duplicated**, increasing computational complexity and the risk of **overfitting** in ML models.

Preprocessing

Preprocessing Pipeline

I **Data Preparation:**
This phase involves building the dataset from the original ADNI files. All the preprocessing operations that prevent data leakage from the train dataset to the test dataset are implemented. In fact, it is precisely here that we split the ADNIMERGE.csv dataset into train and test datasets.

II **Data Preprocessing:**
This phase involves transforming the dataset to make it suitable for machine learning. These operations would risk data leakage if evaluated on the entire dataset. Therefore, they are performed on the train dataset, and the test dataset is modified accordingly to make it consistent, before evaluating the models built on the train. It is divided into *Data Cleaning*, *Data Transformation*, *Outlier Detection*, and *Data Reduction*.

Data Preparation

1

Selection of baseline visits only

Only baseline visits were considered (VISCODE == 'bl'), so as to predict the diagnosis based on the information from the first visit.

2

SMC diagnosis management

SMC (Subjective Memory Concern) subjects were realigned with DX and finally reclassified as CN, since subjective perception is not predictable from objective data.

3

Consolidating baseline columns

Clean up duplicates and merge ”baseline” values into visit columns.

4

Error Handling

Deleted the row with RAVLT perc forgetting = -316.667 and RAVLT forgetting = -19.

5

Text Category Cleaning

Features representing ethnicity and social status have been standardized for better readability and to avoid inconsistencies.

6

Encoding of categorical variables

One-hot encoding (PTMARRY), binary mapping (PTGENDER: Male=1, Female=0) and ordinal encoding for DX (CN=0, EMCI=1, LMCI=2, AD=3).

7

Preliminary Feature Reduction

Removal of columns not relevant to the diagnosis.

8

Splitting train/test

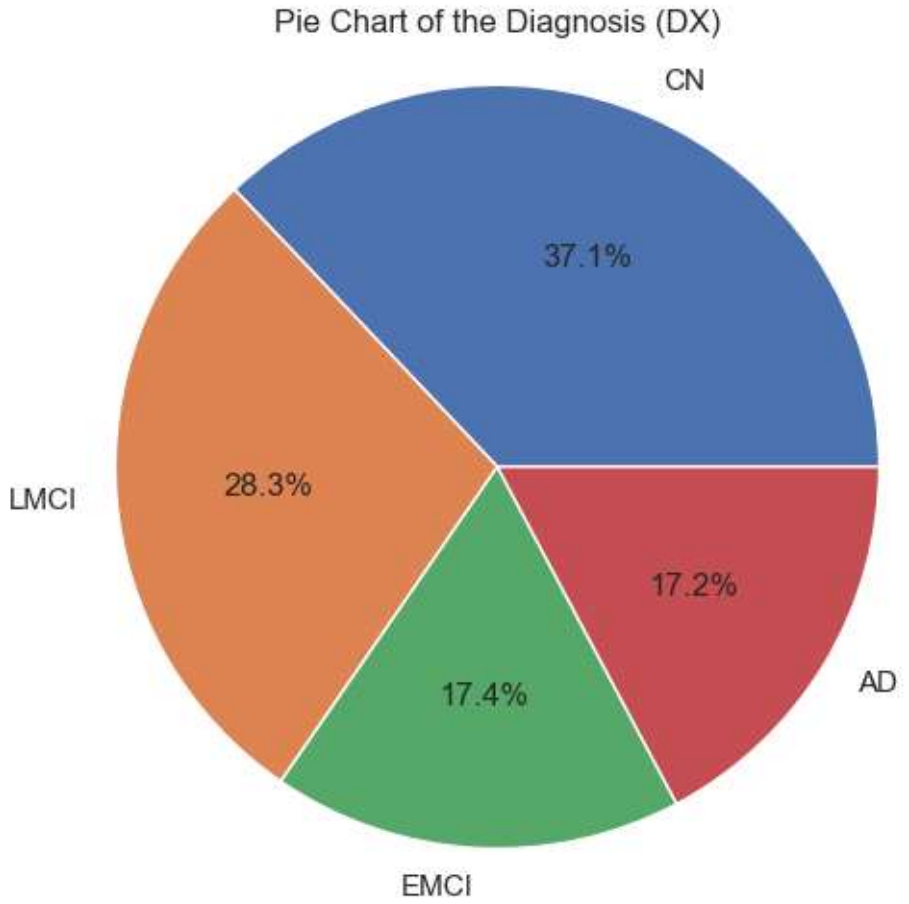
Separation into training and test datasets while avoiding leakage.

Multiclass Problem: DX and DX_bl

- DX_bl has 5 possible values:
  - CN: Cognitively Normal
  - SMC: Subjective Memory Concern
  - EMCI: Early Mild Cognitive Impairment
  - LMCI: Late Mild Cognitive Impairment
  - AD: Alzheimer's Disease

- DX has 3 possible values:
  - CN: Cognitively Normal
  - MCI: Mild Cognitive Impairment
  - Dementia: Alzheimer's Disease

- **We create a new DX as our target with this 4 classes:**
  - CN: Cognitively Normal
  - EMCI: Early Mild Cognitive Impairment
  - LMCI: Late Mild Cognitive Impairment
  - AD: Alzheimer's Disease



Data Cleaning

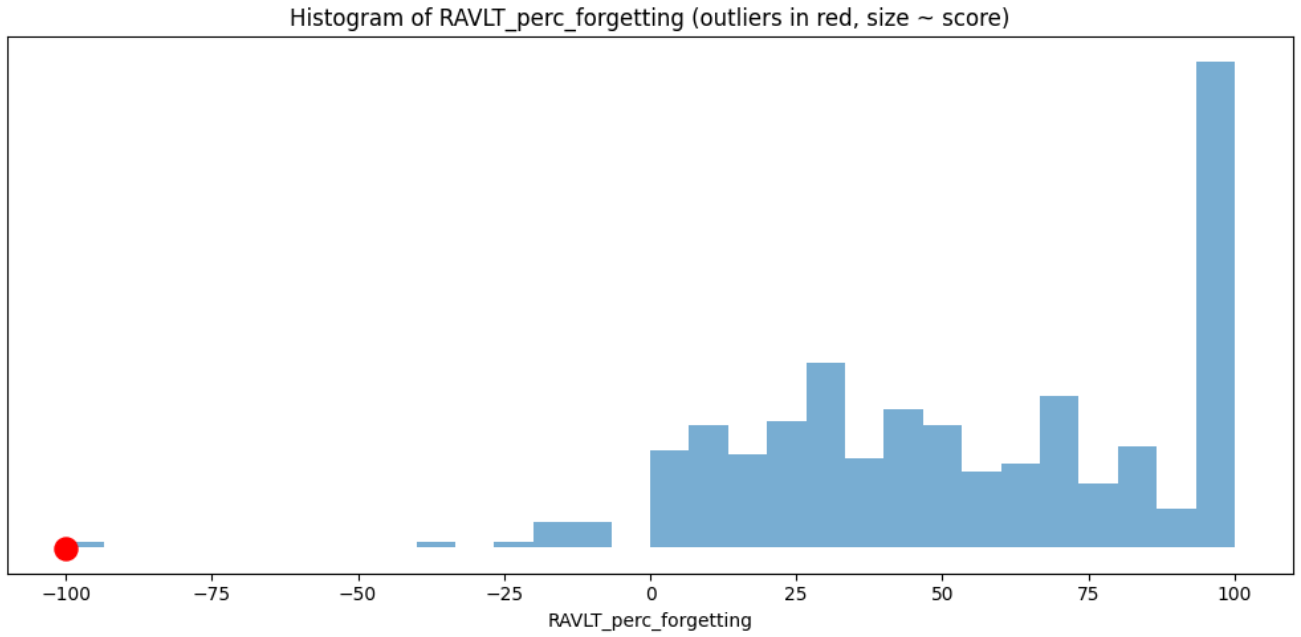
- **Handling missing values:** Identifying percentages of missing values and using KNN Imputer for continuous variables.
- **Numeric Value Conversion:** Convert almost all cognitive scales and age from float to int, correcting for approximations due to imputation or format errors.

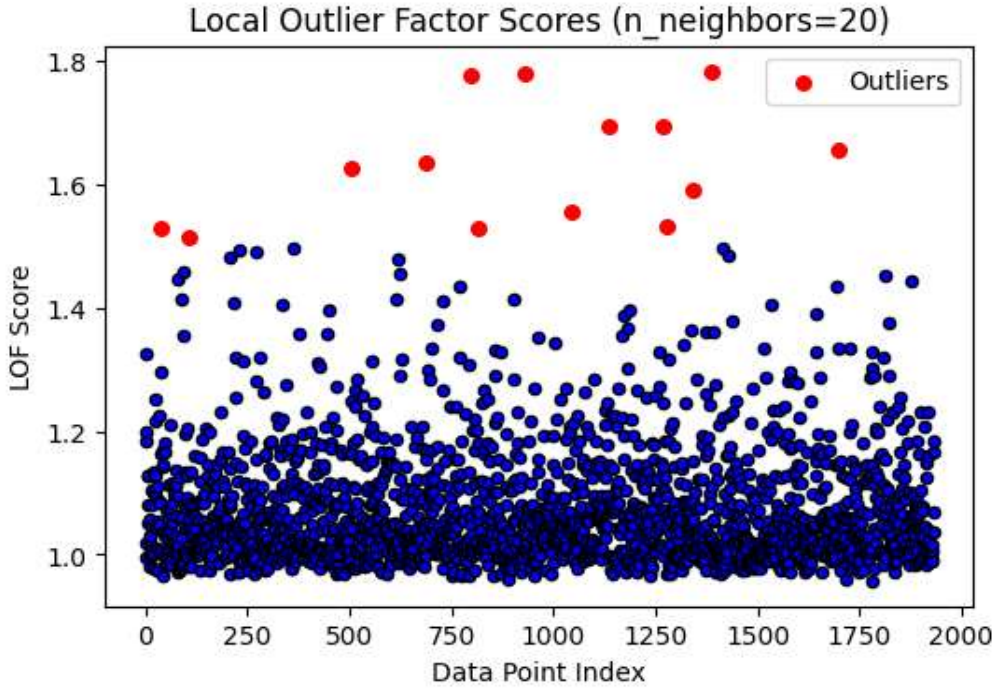
Data Transformation

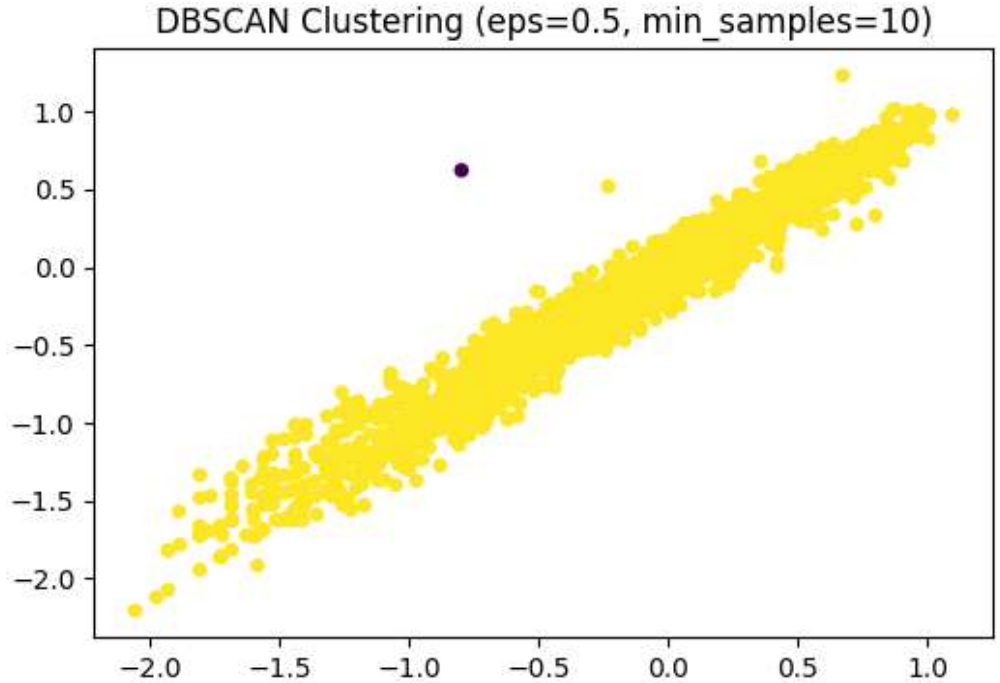
- **Creation of new CSF metrics:** TAU/ABETA and PTAU/ABETA ratios more predictive than single measures according to the literature.
- **MRI normalization to ICV:** Necessary to correct for differences due to gender and cranial size.

Outlier Detection

- **Univariate Analysis:** use *IQR* and *Z-score* for each column to find outliers.
- **Multivariate Analysis:** create groups of variables (EcogPt, EcogSP, Neuropsych, MRI, MRI/ICV, CSF, CSF/ABETA, mPACC), apply LOF and DBSCAN on the normalized data (RobustScaler).
- **Problematic outlier removal:** high unlikely values are replaced with the mean by class.







Univariate Analysis

- **Univariate Analysis** takes a **single feature** from the dataset and uses the parametric **IQR** and **Z-score threshold**. It combines the indices reported by both and builds a summary table for each index with a value, IQR flag, Z-score flag, and a score that is the sum of the two flags (0, 1, or 2). **Points with a score of 2 are considered "robust" univariate outliers**. **Those with a score of 1 are considered weaker signals**. **All are evaluated on a case-by-case basis in the notebook**.

Multivariate Analysis

- **Multivariate analysis** works on a group of features. The data is first scaled using a **RobustScaler**. The function calls **LOF** (saving the **LOF scores**, where a higher score indicates greater local anomaly) and **DBSCAN** (points **labeled −1** are considered noise). It aggregates the indices returned by LOF and DBSCAN, calculates a LOF/DBSCAN flag and a **score for each index** (always 0, 1, or 2, given by the sum of the flags), **stores the maximum LOF_score** among the tested configurations, and builds a summary. **Only points reported by both methods and with a LOF_score above 2 are considered "extreme".**

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:** Four complementary methods were used: *Pearson correlation* ($|r| \geq 0.6$), *Mutual Information* (top 25), *SelectKBest* (Kruskal–Wallis, k=25), and *RFE* (with Random Forest) to capture linear, nonlinear, univariate, and model-based relationships. For each method, a list of top features was obtained and the number of occurrences of each was counted. Variables selected by at least three methods were retained as "core" variables because they were more robust and less dependent on a single criterion. Additionally, some clinical variables deemed important for interpretability were deliberately preserved.

Classification & Results

Model Selection

We selected this classification models:

- **Decision Tree**
- **Random Forest**
- **Extra Trees**
- **XGBoost**
- **LightGBM**
- **CatBoost**
- **Multinomial Logistic Regression**
- **Bagging**

Hybrid Sampling

To overcome the class imbalance we use in combination:

- **Random Under-Sampling** (RUS) to reduce the number of instances in the majority classes (CN and LMCI), preventing the dataset from becoming excessively biased toward synthetic examples;

- **Synthetic Minority Over-sampling Technique for Nominal and Continuous features** (SMOTENC) to generate new synthetic examples of the minority classes (EMCI and AD).

The problem with CDRSB, LDELTOTAL, and mPACCdigit

- The **CDRSB**, **LDELTOTAL**, and **mPACCdigit** cognitive scores show significantly higher predictive power than other variables.
- This **can improve model accuracy**, but creates the **risk of feature dominance**, where a few variables excessively influence predictions.
- This imbalance can cause **local overfitting**: excellent performance on ADNI but possible loss of accuracy on external or more heterogeneous populations.
- It is not possible to definitively determine whether these variables **are simply very strong predictors of Alzheimer's disease diagnosis**.
- We divided the pipeline in dataset **with** CDRSB, LDELTOTAL and mPACCdigit and **without** CDRSB, LDELTOTAL and mPACCdigit.

Classification Pipeline

All subsequent operations will be the same, but will be executed separately within 4 distinct pipelines, corresponding to the 4 types of datasets generated.

DS1

Dataset without Sampling and with CDRSB, LDELTOTAL, and mPACCdigit

DS2

Dataset with Sampling and with CDRSB, LDELTOTAL, and mPACCdigit

DS3

Dataset without Sampling and without CDRSB, LDELTOTAL, and mPACCdigit

DS4

Dataset with Sampling and without CDRSB, LDELTOTAL, and mPACCdigit

Model Construction

1

Hyperparameter Tuning

To optimize the performance of the classifiers, a Grid Search procedure with 5-fold-cross validation was adopted.

2

Models Building

We built the models on the training dataset.

3

Explainability

Used SHAP Summary Plots on all models to see which features were most important overall and by class. Built trees and rules for Decision Trees.

4

Models Final Evaluation

We evaluated the models on the train dataset, using 5-fold cross-validation, and on the test dataset, as well as using the explainability results. The evaluations primarily considered the statistics on the test dataset.

IV Classification & Results

MODEL	Accuracy	Balanced Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	F1 Score (macro)	ROC AUC (macro)
Random_Forest1	0.925620	0.919814	0.927101	0.925620	0.925820	0.916918	0.986475
Extra_Trees1	0.923554	0.918812	0.924969	0.923554	0.924003	0.914303	0.986232
XGBoost0	0.927686	0.916805	0.928391	0.927686	0.927514	0.918030	0.987596
Random_Forest0	0.929752	0.916327	0.930994	0.929752	0.929415	0.920545	0.983876
XGBoost1	0.923554	0.915250	0.924380	0.923554	0.923706	0.913798	0.986799
Extra_Trees0	0.923554	0.913191	0.924021	0.923554	0.923567	0.913565	0.988441
CatBoost1	0.919421	0.912752	0.920484	0.919421	0.919728	0.910155	0.987497
LightGBM1	0.919421	0.911565	0.920235	0.919421	0.919545	0.909524	0.987098
CatBoost0	0.921488	0.908992	0.921941	0.921488	0.921237	0.910847	0.988699
LightGBM0	0.919421	0.904757	0.920713	0.919421	0.918885	0.907530	0.984300
Bagging0	0.919421	0.904077	0.922416	0.919421	0.919240	0.907923	0.984449
Bagging1	0.904959	0.895301	0.906196	0.904959	0.905261	0.892969	0.983476
Decision_Tree1	0.898760	0.893037	0.901509	0.898760	0.899541	0.886243	0.980261
Multinomial_Logistic_Regression1	0.878099	0.873117	0.883112	0.878099	0.878508	0.862857	0.980601
Decision_Tree0	0.876033	0.872226	0.881332	0.876033	0.877111	0.861493	0.974606
Multinomial_Logistic_Regression0	0.873967	0.871962	0.881084	0.873967	0.874781	0.859787	0.981506

IV Classification & Results

MODEL	Accuracy	Balanced Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	F1 Score (macro)	ROC AUC (macro)
XGBoost1	0.735537	0.720959	0.745818	0.735537	0.739242	0.717172	0.907103
Extra_Trees0	0.737603	0.717186	0.738250	0.737603	0.736785	0.714046	0.909280
LightGBM1	0.725207	0.716317	0.740019	0.725207	0.730068	0.709832	0.908058
Extra_Trees1	0.719008	0.706432	0.728482	0.719008	0.721057	0.698828	0.906806
Bagging1	0.725207	0.705921	0.728153	0.725207	0.725832	0.704301	0.903807
Multinomial_Logistic_Regression0	0.719008	0.699875	0.718814	0.719008	0.717156	0.697048	0.907081
XGBoost0	0.739669	0.699017	0.728746	0.739669	0.731404	0.703253	0.912593
Random_Forest0	0.716942	0.697227	0.717294	0.716942	0.715909	0.691878	0.905293
Multinomial_Logistic_Regression1	0.708678	0.695913	0.715986	0.708678	0.709280	0.689989	0.903669
CatBoost1	0.706612	0.692950	0.717981	0.706612	0.710555	0.689426	0.909166
Random_Forest1	0.702479	0.686877	0.711632	0.702479	0.704450	0.680898	0.904030
LightGBM0	0.729339	0.682113	0.712688	0.729339	0.715176	0.682723	0.912346
Bagging0	0.723140	0.673841	0.713842	0.723140	0.713053	0.682405	0.903966
CatBoost0	0.723140	0.673356	0.708919	0.723140	0.710123	0.678616	0.909338
Decision_Tree1	0.640496	0.652118	0.688104	0.640496	0.652207	0.635709	0.856708
Decision_Tree0	0.661157	0.646413	0.690373	0.661157	0.672628	0.654678	0.837981

DS3

"...0"

"...1"

DS4

Model Choosing

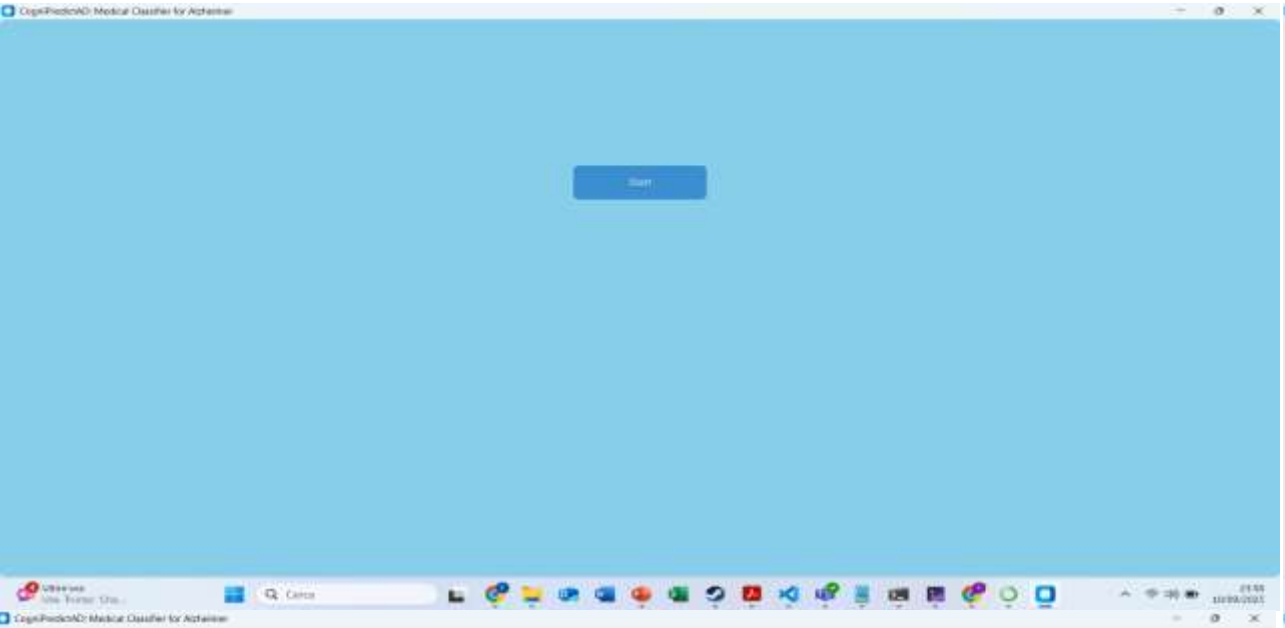
- **With CDRSB, LDELTOTAL, and mPACCdigit: (DS1, DS2)**
  Random Forest1 was chosen as the main model and Decision Tree1 as the XAI, due to the best metrics (balanced accuracy, F1, ROC-AUC).

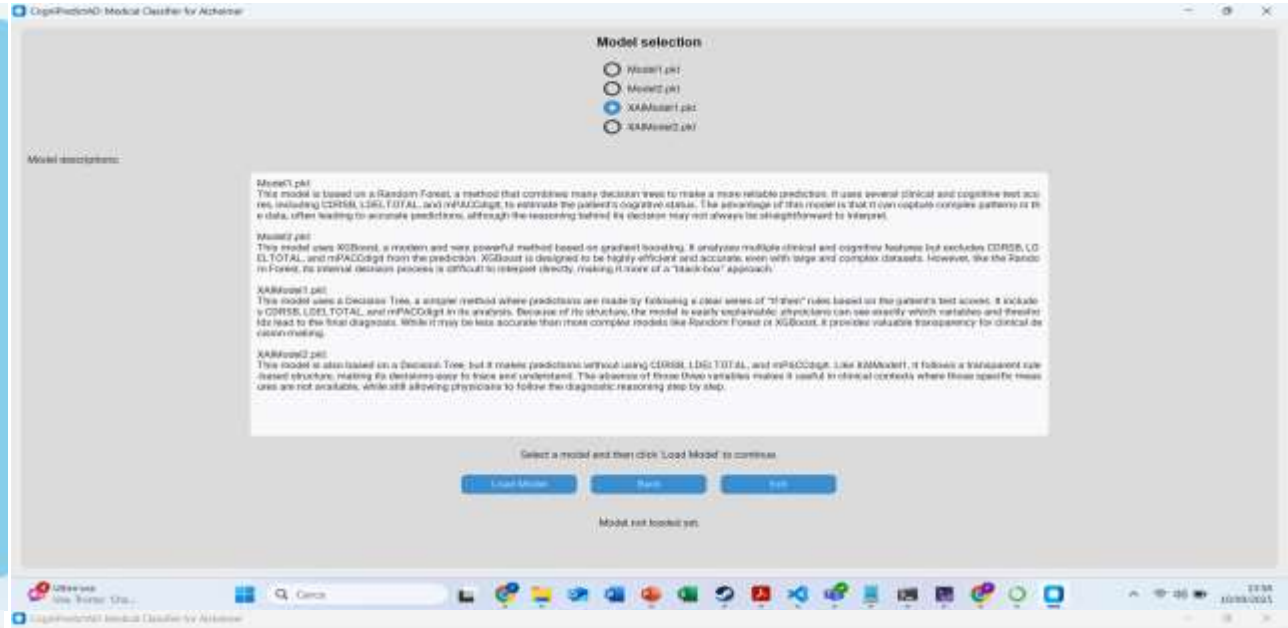
- **Without CDRSB, LDELTOTAL, and mPACCdigit: (DS3, DS4)**
  XGBoost1 was chosen as the main model and Decision Tree1 as the XAI, based on testing performance.

- The saved models are **Model1.pkl** (1/RandomForest1), **XAIModel1.pkl** (1/Decision Tree1), **Model2.pkl** (2/XGBoost1), and **XAIModel2.pkl** (2/Decision Tree1).

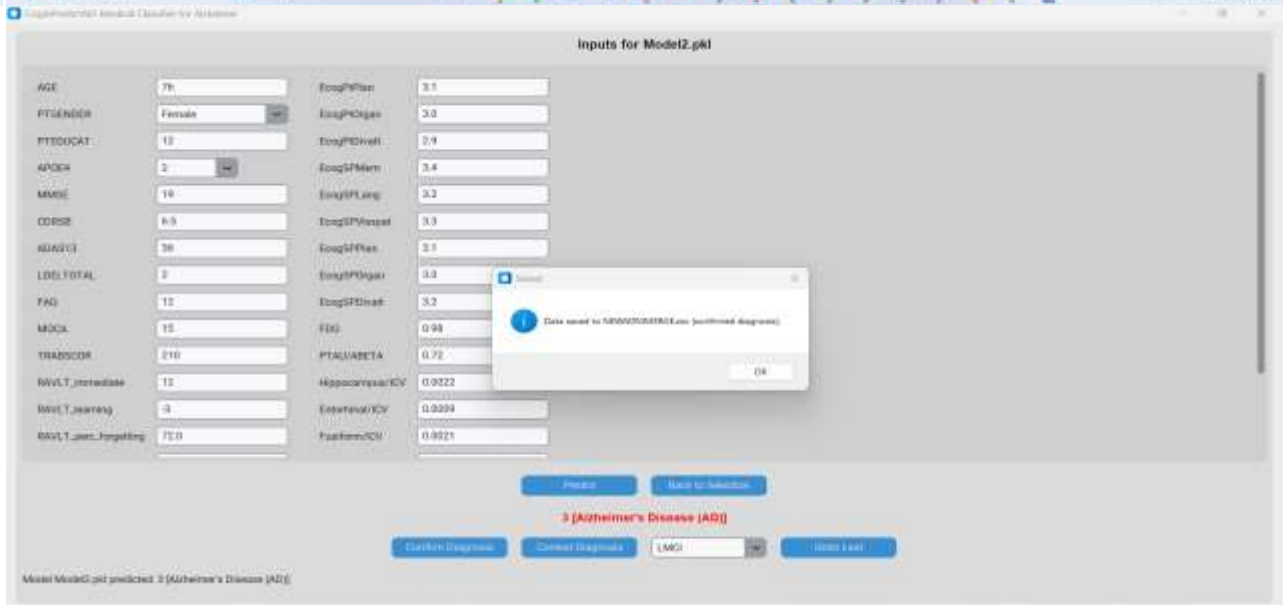
Applications & Conclusions

App




Model selection
Select a model and then click 'Load Model' to continue.


Inputs for Model1.pkl
0 (Cognitively Normal (CN))


Inputs for Model2.pkl
3 (Alzheimer's Disease (AD))

Conclusions

- **Dataset limitations:** only 2,419 patients, many missing values (CSF, PET), strong dependence on 3 cognitive endpoints. Risk of local overfitting, dataset bias, and imputations increasing noise. External validation required.
- **Model Performance:** The models perform well overall, especially Model1. Furthermore, XAIModel1 and XAIModel2 are easily interpretable. If the three features appear unpredictable externally, there are Model2 and XAIModel2.
- **Application value:** useful as a support (screening, risk stratification), but obviously does not replace clinical evaluation.
- **Future developments:** expand cohorts (ADNI4, external), integrate with similar datasets, and include geographic area as a feature.

Thanks for your attention!

Brain
(sagittal cut)

Brain
(coronal cut)

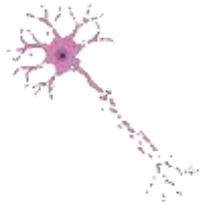
Brain
(lateral)

Brain with
Alzheimer's

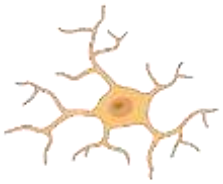
Brain with regions
(coronal cut)

Amyloid-beta plaque

Myelinated
motor neuron



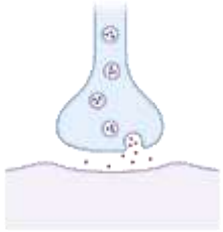
Degenerating
motor neuron



Microglia

Astrocyte

Dynamic line
neurons



Synaptic cleft