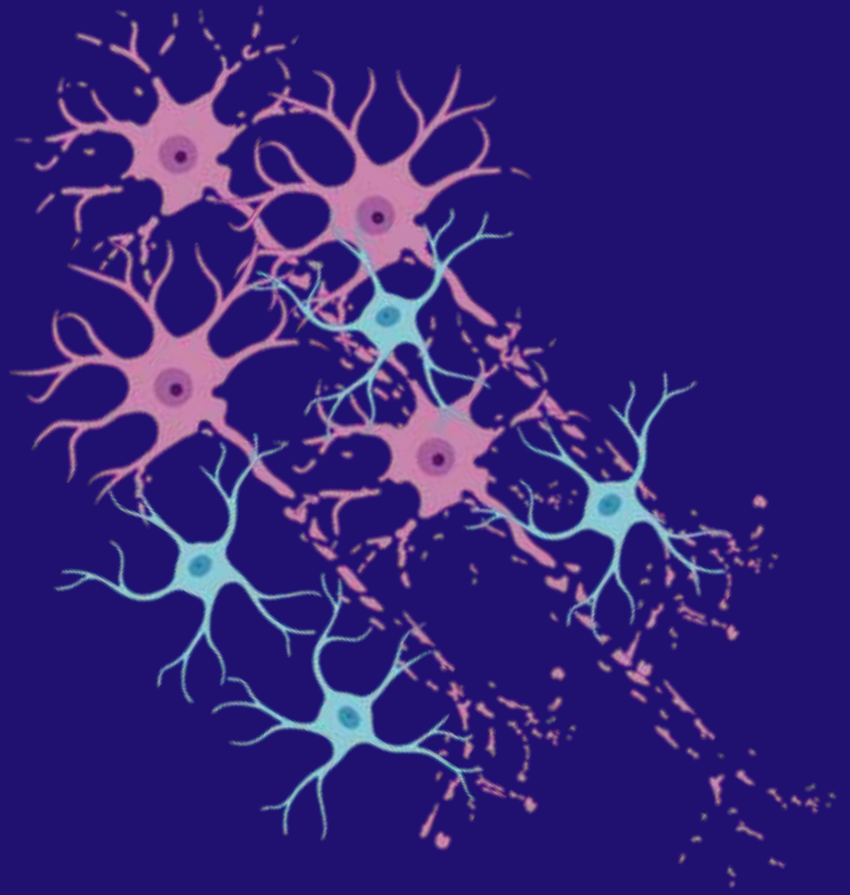
University of Pisa
Department of Information Engineering
Artificial Intelligence and Data Engineering



CogniPredictAD

Francesco Panattoni

Project for Data Mining and Machine Learning

Table of Contents

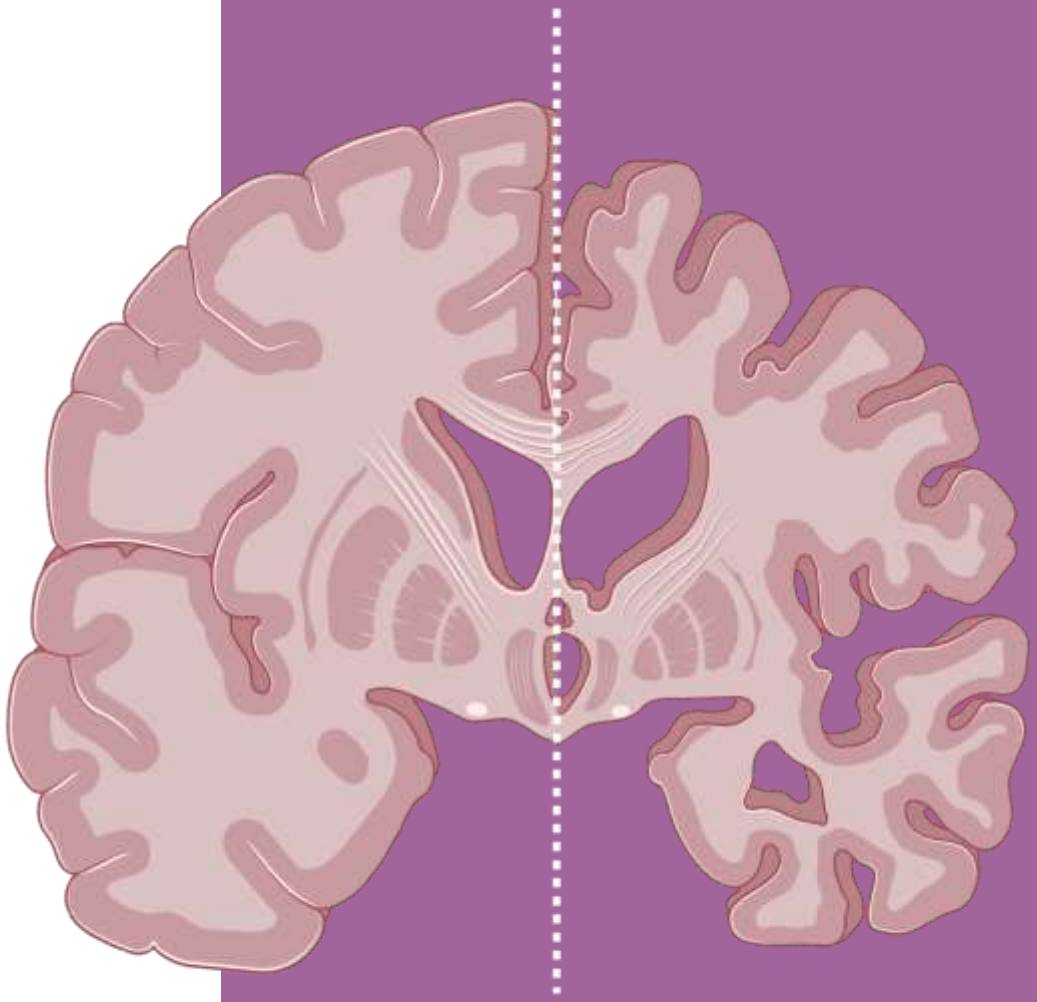
Motivations and Clinical Background

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** for slowing the progression of the disease and improving quality of life.

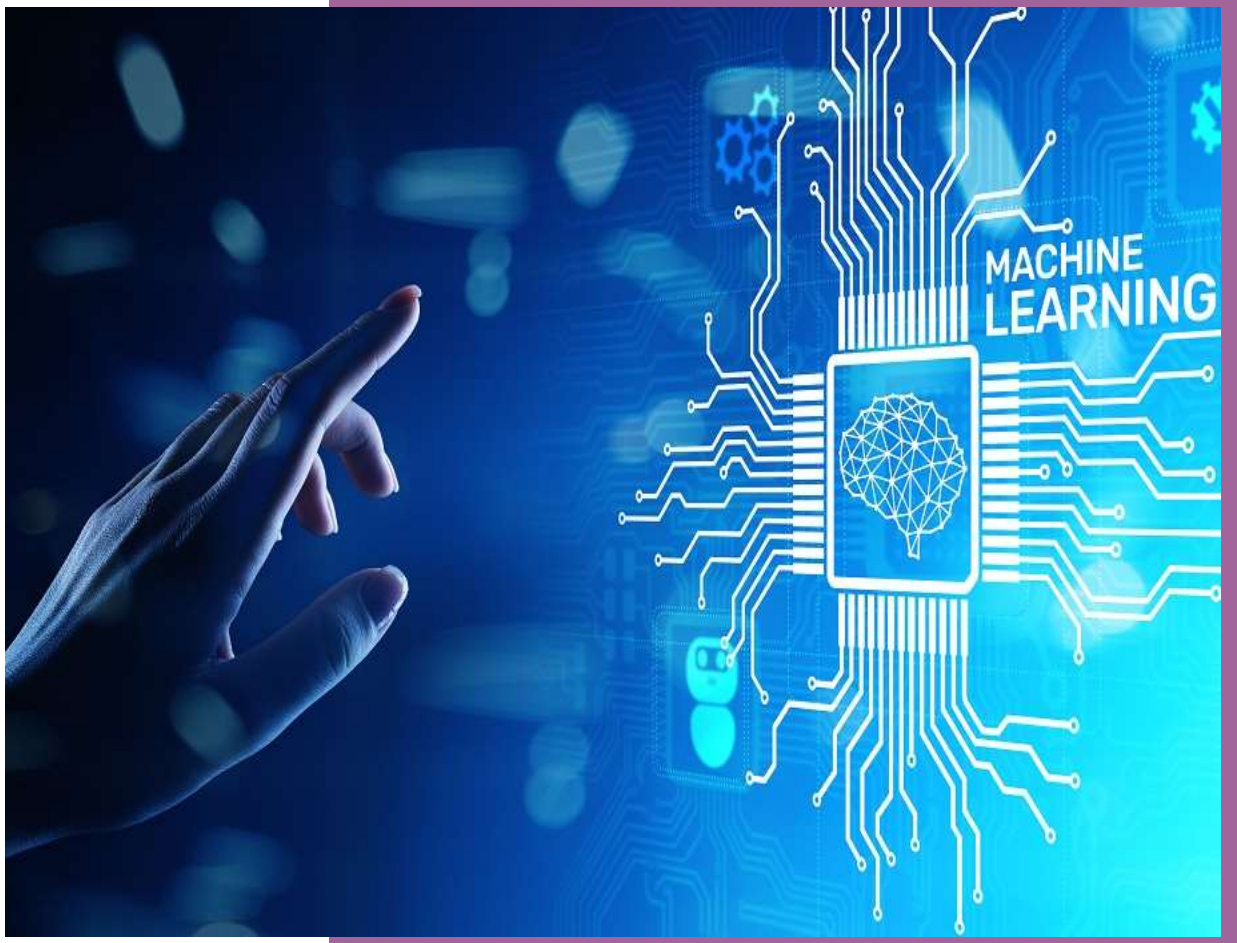


So why Machine Learning?

- **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**, the most recent phase of the study, was initiated in **2022.**
- **16,421 rows x 116 columns.** The columns are divided into current visit columns and baseline visit columns (with the "_bl" suffix) to aid 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.

Data Preparation

The **Data Preparation** involves building the **learning set** from the original ADNIMERGE table file. This phase focuses on transforming the visit-centric CSV into a single-row-per-subject baseline cohort and generating compact, multimodal feature tables suitable for modeling.

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, ensuring consistency with objectively defined diagnostic categories.

3

Consolidating baseline columns

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

4

Text Category Cleaning

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

5

Encoding of categorical variables

One-hot encoding (PTMARRY) and binary mapping (PTGENDER: Male=1, Female=0).

6

Preliminary Feature Reduction

Removal of columns not relevant to the diagnosis.

7

Error Correction

Clinically implausible values are detected using IQR Method and either clamped, nullified, or set aside for model-aware imputation rather than being corrected.

8

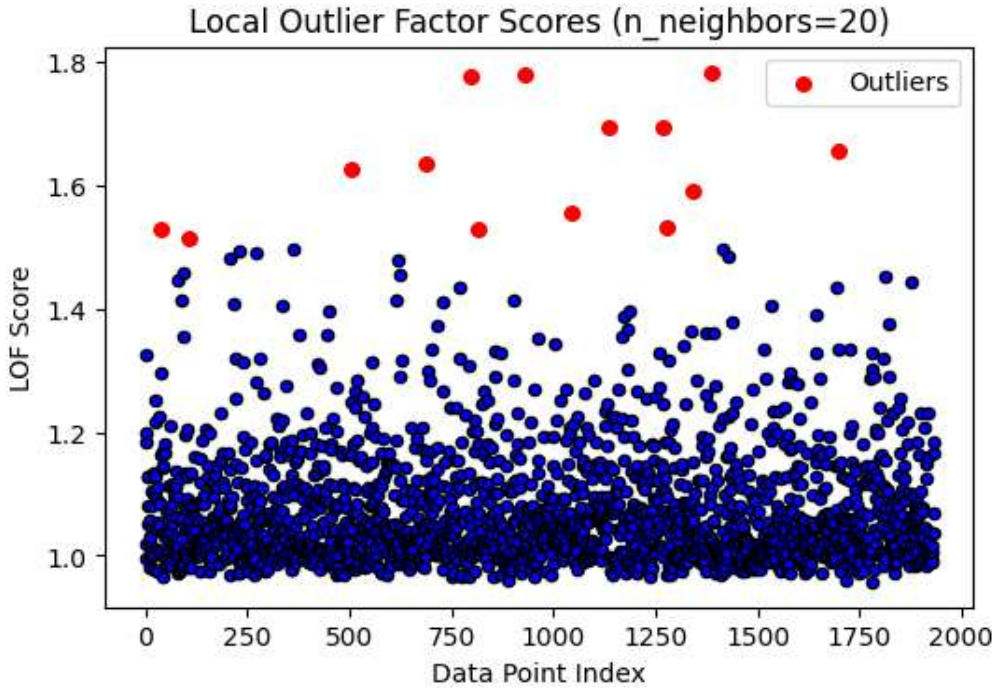
Holdout Method

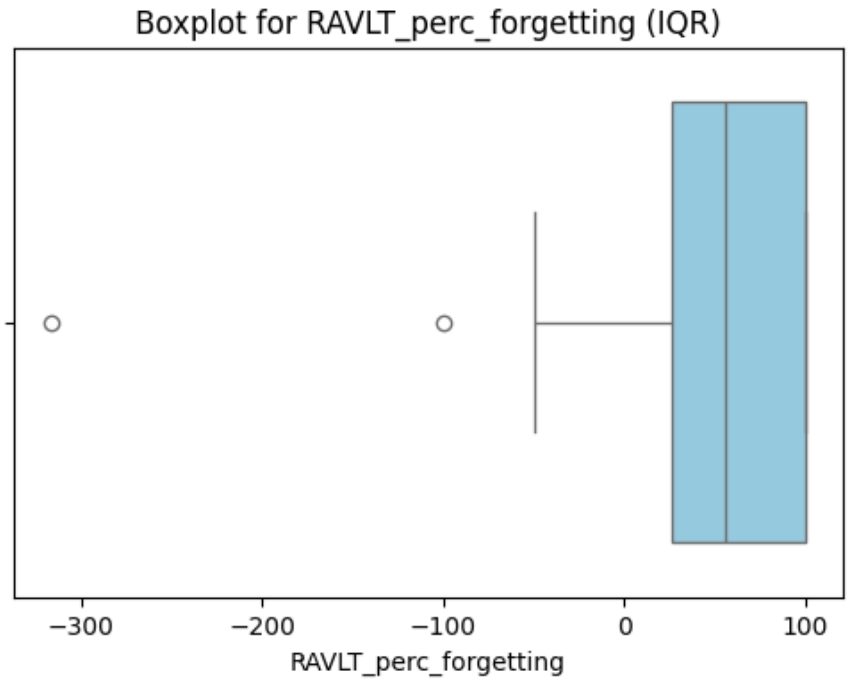
Separation of the learning set into training and testing datasets while avoiding leakage.

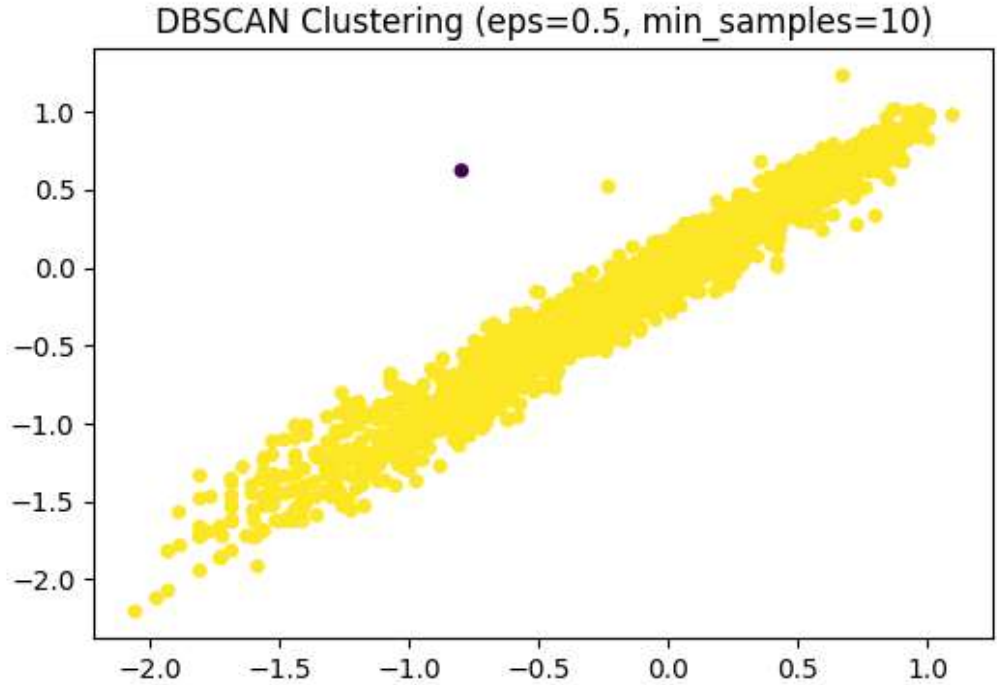
II Dataset ADNI

Outlier Detection

- **Univariate Analysis:** use *IQR* for each column to find outliers and *to remove problematic outliers*.
- **Multivariate Analysis:** create groups of columns (Cognitive Score, MRI/ICV, CSF/ABETA), apply *LOF* on the normalized data (RobustScaler) to find outliers and *to analyze them*.





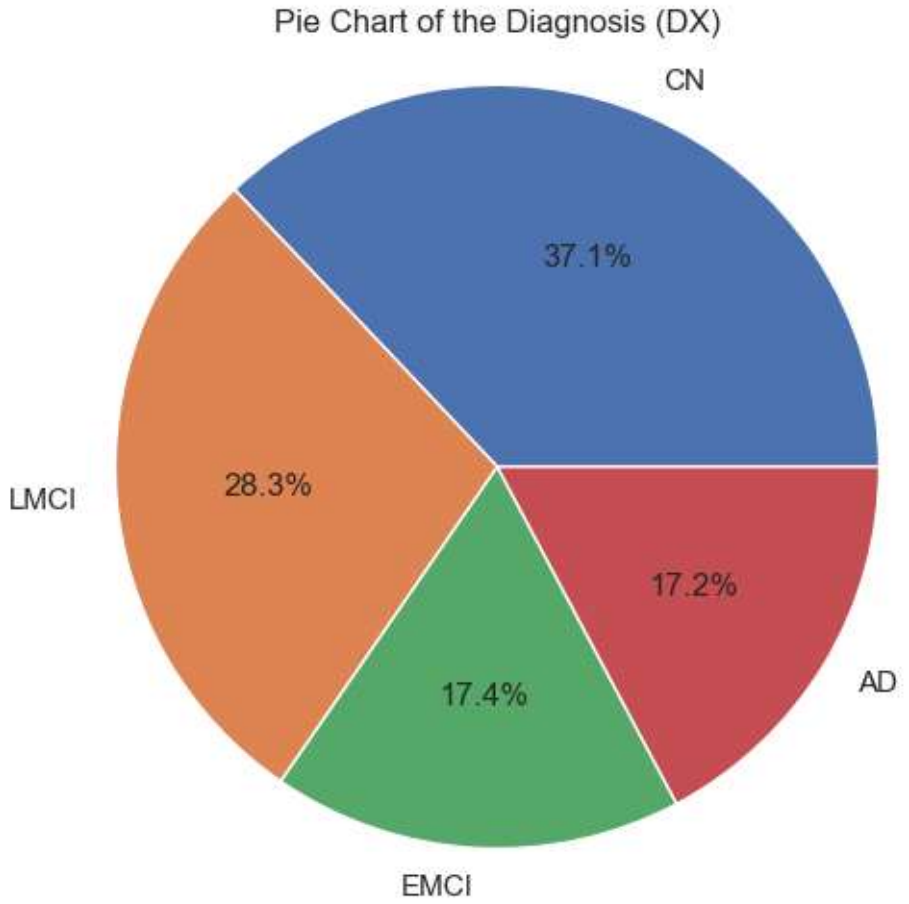


Multiclass Problem: DX and DX_bl

- 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



Preprocessing

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 set*, and the *test set* is modified accordingly to make it consistent, before evaluating the models built on the training dataset. Preprocessing is divided into *Data Cleaning*, *Data Transformation, Data Reduction and eventually Hybrid Sampling*.

Data Cleaning

- **Handling missing values:** Identifying percentages of missing values and using KNN Imputer for continuous variables.
- **Numeric Value Conversion:** Converted nearly all cognitive scales and age from float to int, correcting approximations 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.

Data Reduction

- **Biomarker raw values replaced by ratios:** *TAU*, *PTAU*, *ABETA* are replaced by *TAU/ABETA* and *PTAU/ABETA*;
- **MRI are normalized by ICV:** *Ventricles*, *Hippocampus*, *Entorhinal*, *Fusiform*, *MidTemp*, *WholeBrain*, *ICV* are replaced by *Ventricles/ICV*, *Hippocampus/ICV*, *Entorhinal/ICV*, *Fusiform/ICV*, *MidTemp/ICV*, and *WholeBrain/ICV*;
- **Removal of redundant features:** *ADAS11*, *ADASQ4*, *EcogPtTotal*, *EcogSPTotal*, and *mPACCtrailsB* were removed because they had a high correlation with other features and their informative value was low compared to correlated features.
- **Outcome of reduction:** The dataset is streamlined to key demographic, cognitive, CSF, and MRI features, minimizing noise and redundancy while preserving diagnostic information. This focused baseline table enhances model interpretability, prevents data leakage, and provides a robust foundation for multiclass classification.

Hybrid Sampling

To overcome class imbalance, we used a combination of:

- **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).

We use another pipeline to evaluate performance on the dataset with and without sampling.

Classification & Results

Model Selection

We selected the following classification models:

- **Decision Tree**
- **Random Forest**
- **Extra Trees**
- **Adaptive Boosting**
- **Multinomial Logistic Regression**

Model Construction

1

Hyperparameter Tuning

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

2

Model Building

Models evaluated with *Repeated Stratified 5×5 Cross-Validation* and compared using *Wilcoxon signed-rank tests* on outer-fold F1-macro scores.

3

Explainability

We used *SHAP Summary Plots* on all models to identify 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 *Repeated Stratified 5×5 Cross-Validation*, and on the *test dataset*, as well as using the explainability results. *The evaluations primarily considered the statistics on the test dataset.*

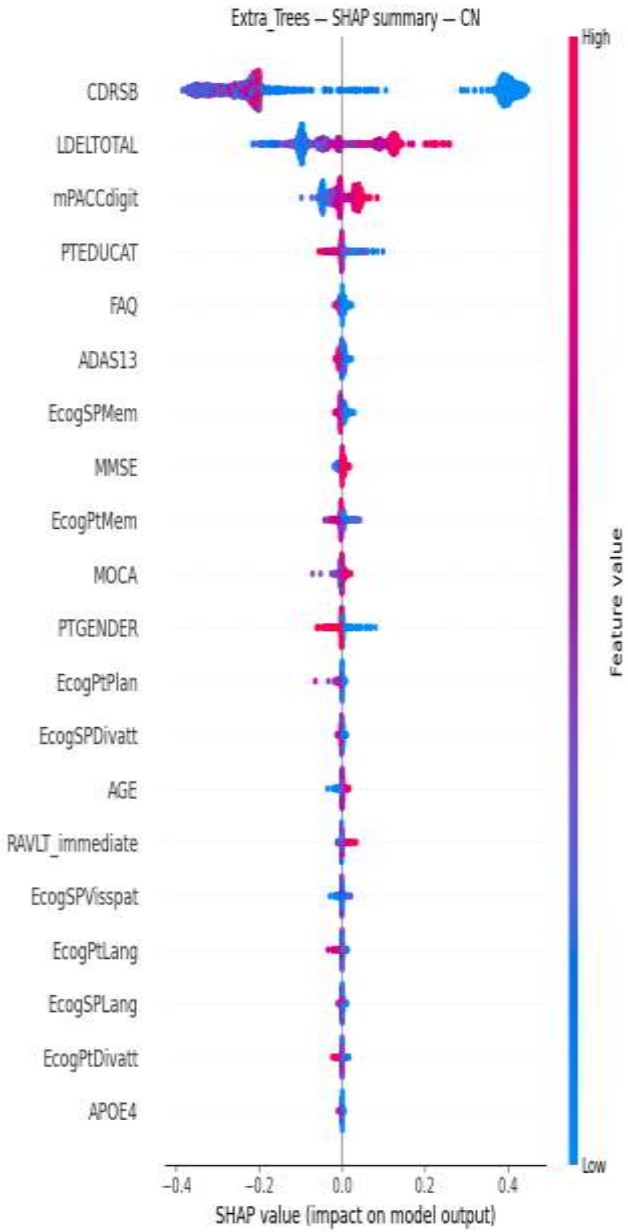
Results (on Test set)

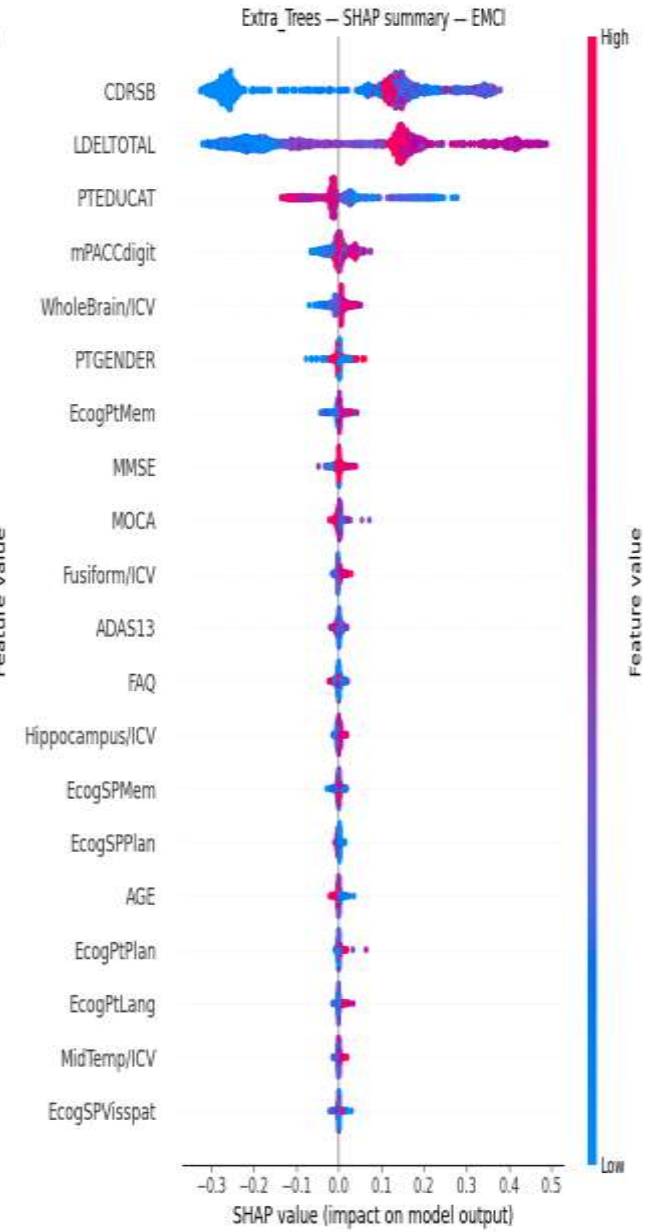
Model	F1 Score (macro)	Accuracy	Balanced Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	ROC AUC (macro)
Extra_Trees	0.9376	0.9442	0.9408	0.9448	0.9442	0.9443	0.9867
Extra_Trees_Sampled	0.9359	0.9421	0.9411	0.9435	0.9421	0.9425	0.9890
Random_Forest	0.9301	0.9380	0.9341	0.9387	0.9380	0.9381	0.9886
Adaptive_Boosting	0.9285	0.9360	0.9347	0.9378	0.9360	0.9363	0.9878
Random_Forest_Sampled	0.9271	0.9339	0.9358	0.9367	0.9339	0.9344	0.9863
Adaptive_Boosting_Sampled	0.9262	0.9339	0.9329	0.9361	0.9339	0.9343	0.9890
Decision_Tree_Sampled	0.9131	0.9236	0.9178	0.9244	0.9236	0.9235	0.9804
Decision_Tree	0.8934	0.9050	0.9026	0.9096	0.9050	0.9057	0.9824
Multinomial_Logistic_Regression	0.8700	0.8843	0.8816	0.8893	0.8843	0.8843	0.9825
Multinomial_Logistic_Regression_Sampled	0.8677	0.8822	0.8754	0.8851	0.8822	0.8826	0.9827

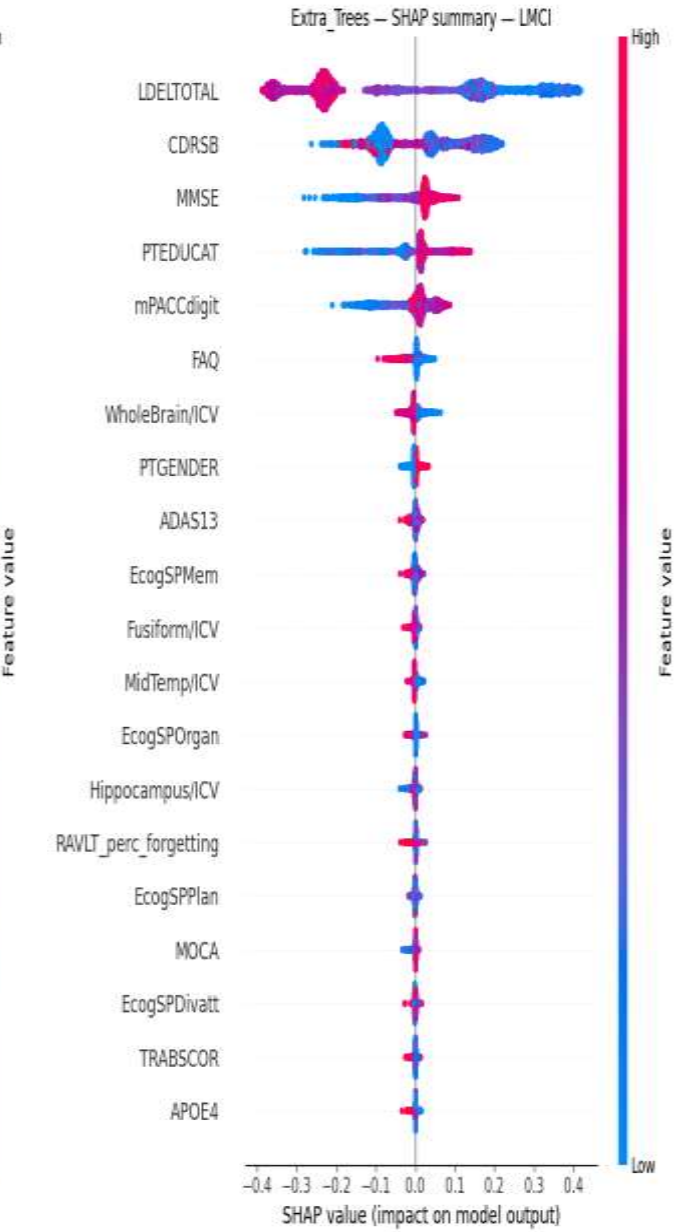
The problem with CDRSB, LDELTOTAL, and mPACCdigit

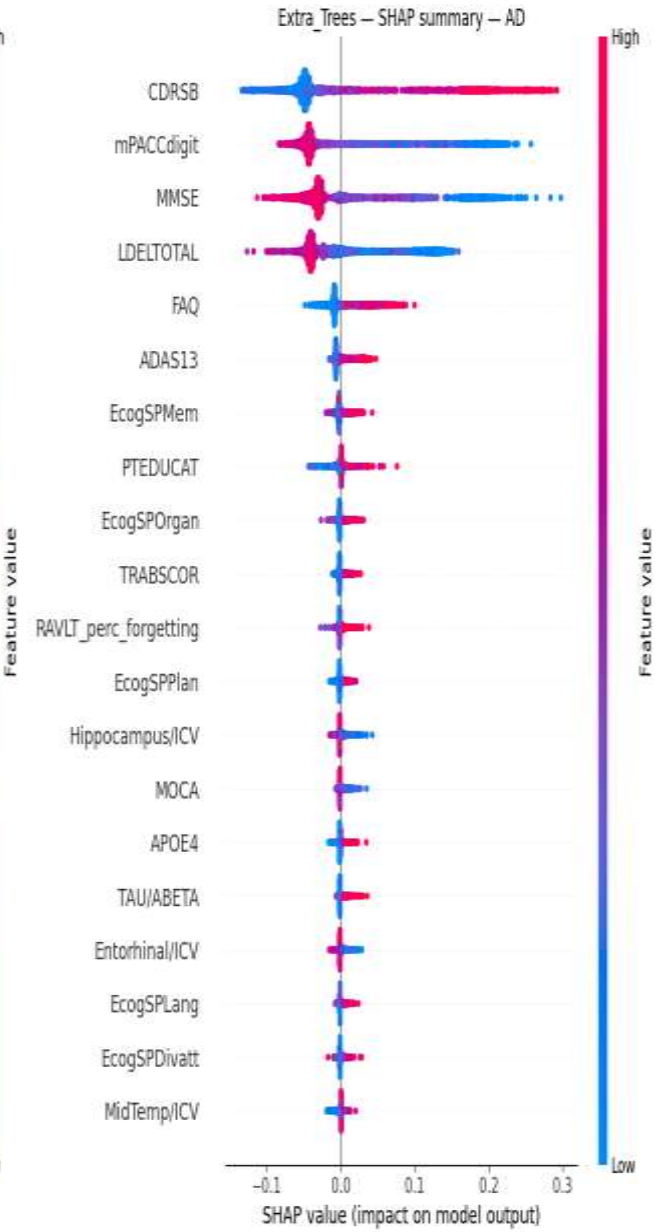
- The **CDRSB**, **LDELTOTAL**, and **mPACCdigit** cognitive scores show significantly higher predictive power than other variables.
- The **Kruskal–Wallis** test confirms this pattern. The features show the largest group differences, indicating that they naturally dominate the separation between diagnostic classes.
- 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 into datasets **with** CDRSB, LDELTOTAL and mPACCdigit and **without** CDRSB, LDELTOTAL and mPACCdigit.

IV Classification & Results









Without CDRSB, LDELTOTAL and mPACCdigit

Model	F1 Score (macro)	Accuracy	Balanced Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	ROC AUC (macro)
Adaptive_Boosting	0.7303	0.7459	0.7327	0.7456	0.7459	0.7437	0.9037
Adaptive_Boosting_Sampled	0.7112	0.7293	0.7157	0.7316	0.7293	0.7277	0.9001
Random_Forest	0.7061	0.7252	0.7052	0.7256	0.7252	0.7245	0.9022
Random_Forest_Sampled	0.7035	0.7190	0.7050	0.7251	0.7190	0.7208	0.9041
Extra_Trees_Sampled	0.7011	0.7211	0.7075	0.7258	0.7211	0.7202	0.9003
Extra_Trees	0.6998	0.7293	0.6952	0.7215	0.7293	0.7238	0.9105
Multinomial_Logistic_Regression _Sampled	0.6829	0.7066	0.6933	0.7135	0.7066	0.7063	0.8921
Multinomial_Logistic_Regression	0.6768	0.7004	0.6882	0.7055	0.7004	0.6996	0.8952
Decision_Tree	0.6603	0.6632	0.6622	0.6960	0.6632	0.6736	0.8467
Decision_Tree_Sampled	0.6482	0.6508	0.6543	0.6990	0.6508	0.6626	0.8445

Model Choosing

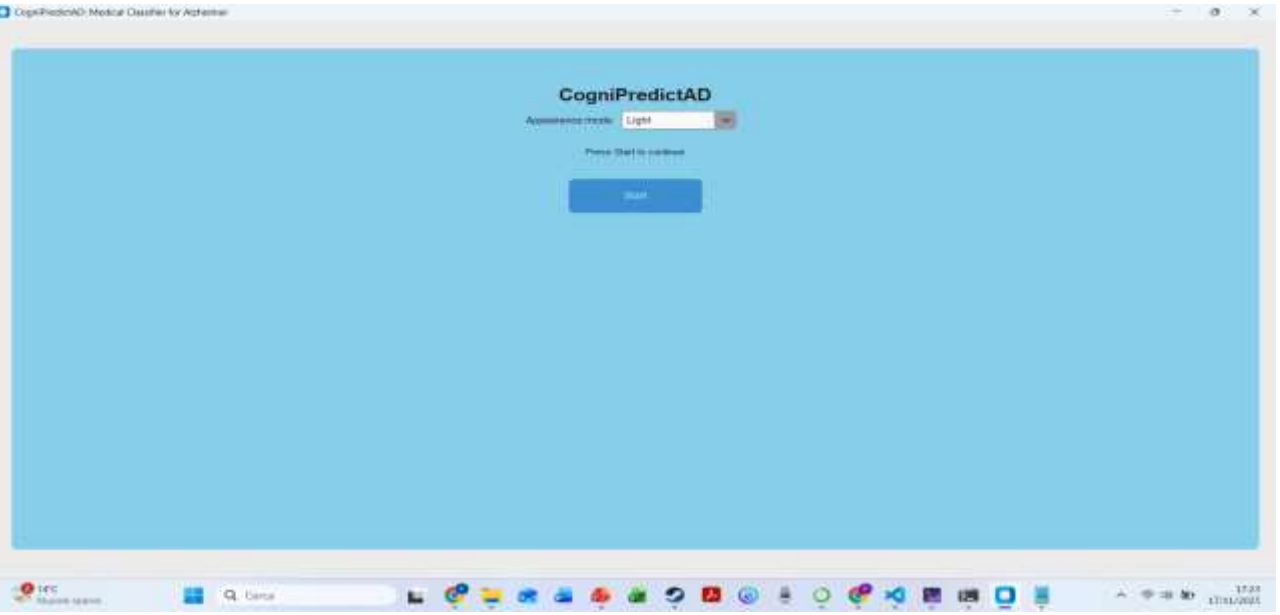
- **With CDRSB, LDELTOTAL, and mPACCdigit:**
  Extra_Trees was chosen as the main model and Decision_Tree_Sampled as the XAI model, due to the best metrics (Balanced Accuracy, F1, ROC-AUC).

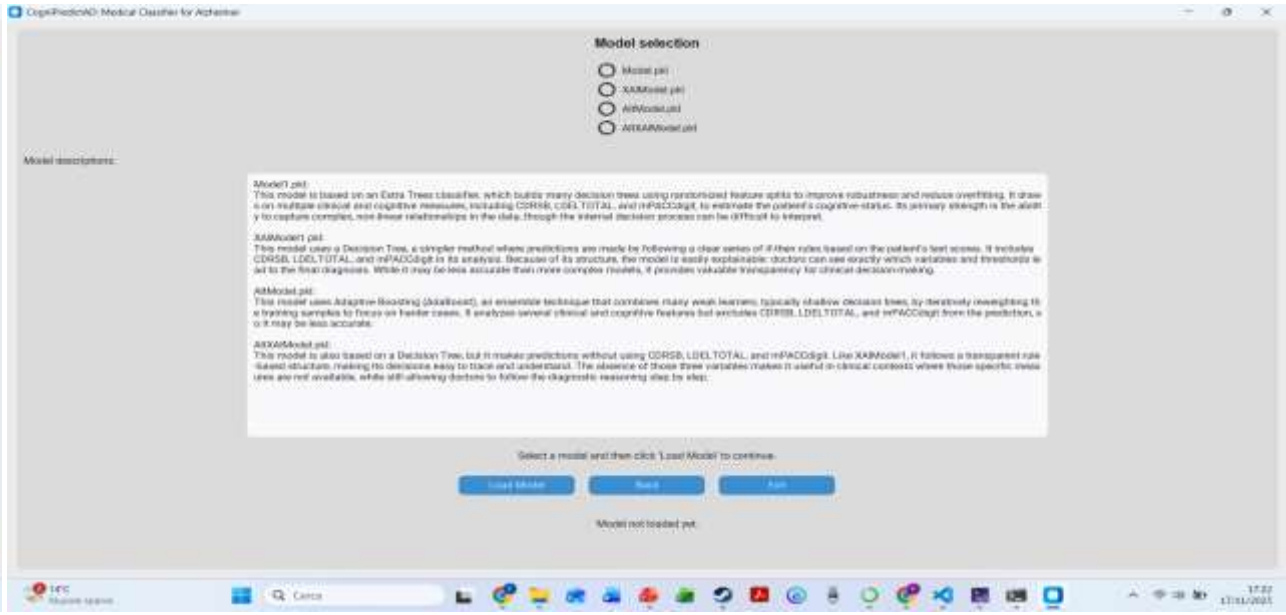
- **Without CDRSB, LDELTOTAL, and mPACCdigit:**
  Adaptive_Boosting was chosen as the main model and Decision_Tree as the XAI, based on testing performance.

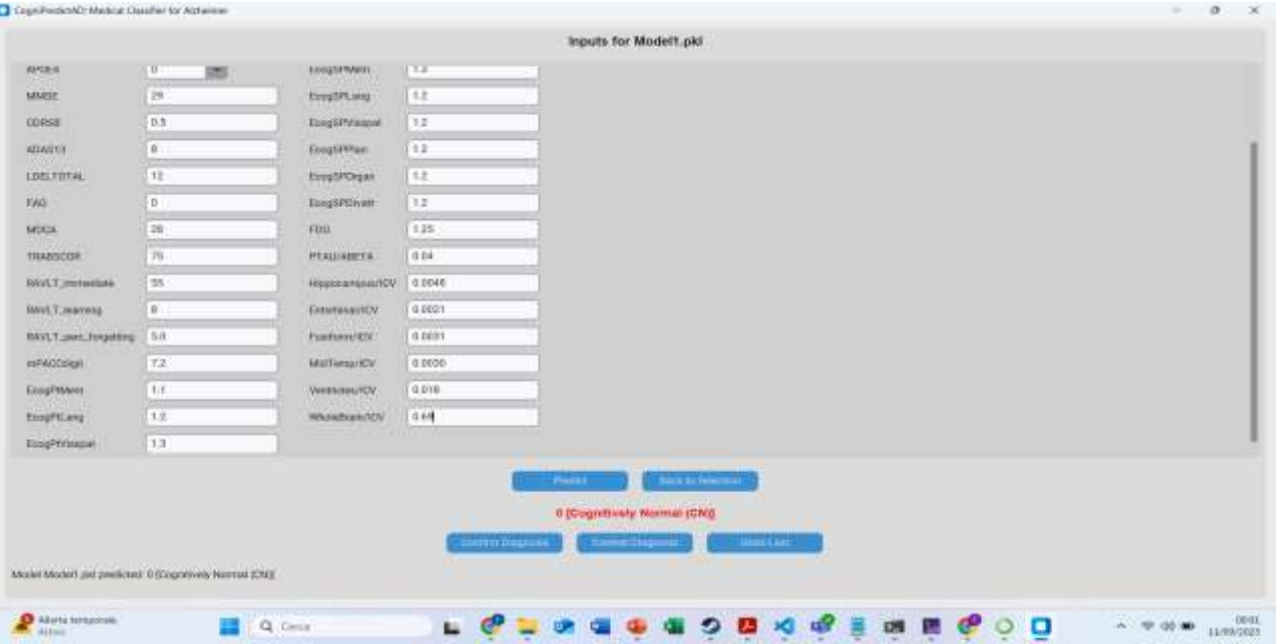
- The saved models are **Model.pkl** (Extra_Trees), **XAIModel.pkl** (Decision_Tree_Sampled), **AltModel.pkl** (alternative/Adaptive_Boosting), and **AltXAIModel.pkl** (alternative/Decision_Tree).

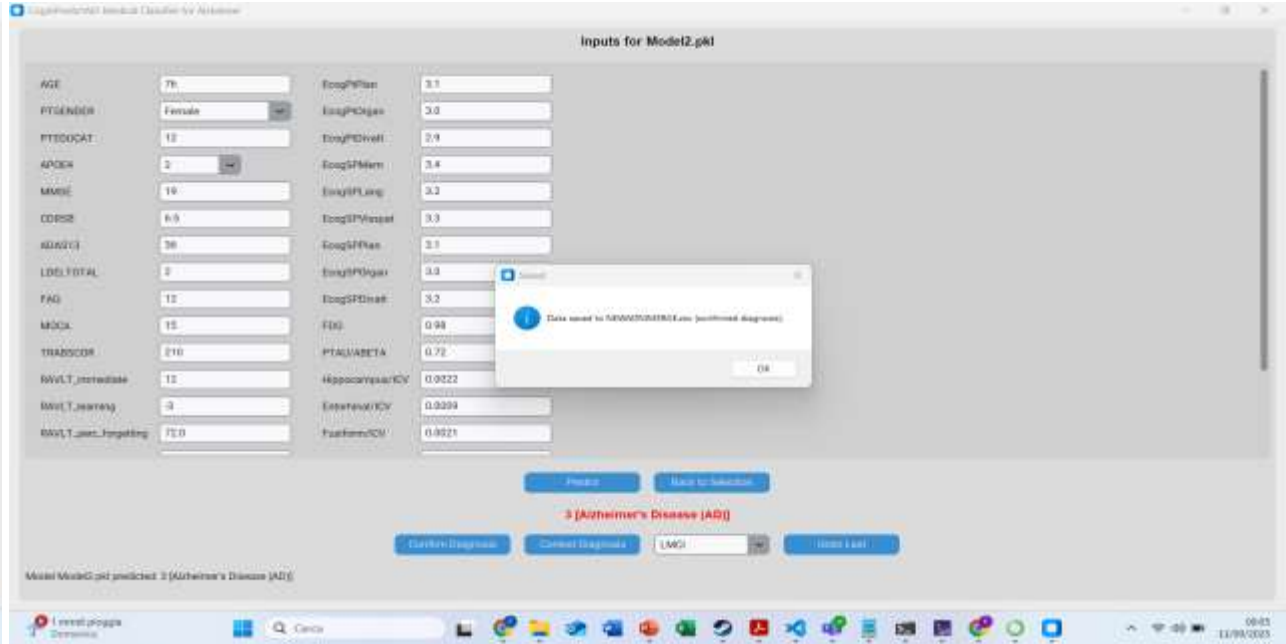
Applications & Conclusions

APP


CogniPredictAD
Start


Model selection
Model descriptions:
Select a model and then click 'Load Model' to continue.
Model not loaded yet.


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 three cognitive scores. Risk of local overfitting, dataset bias, and imputations increasing noise. External validation required.
- **Model Performance:** The models perform well overall, especially *Model.pkl*. Furthermore, *XAIModel.pkl* and *AltXAIModel.pkl* are easily interpretable. If the three features prove unpredictable in external validation, *AltModel.pkl* and *AltXAIModel.pkl* can be used instead.
- **State of the Art:** Although it was not possible to make a precise state of the art due to the lack of similar studies, the statistics make it a solid project.
- **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)

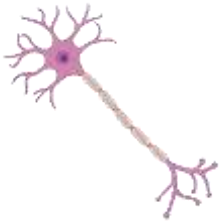
Brain
(coronal cut)

Brain
(lateral)

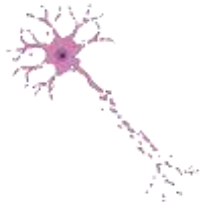
Brain with
Alzheimer's

Brain with regions
(coronal cut)

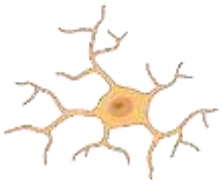
Amyloid-beta plaque



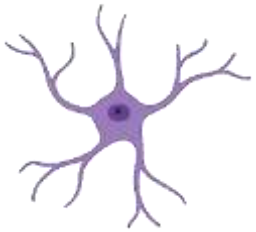
Myelinated
motor neuron



Degenerating
motor neuron

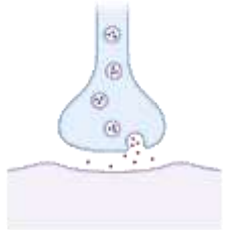


Microglia



Astrocyte

Dynamic line
neurons



Synaptic cleft