

# INFORME DE RECERCA: PREDICCIÓ DE MALALTIA D'ALZHEIMER MITJANÇANT AVALUACIONS COGNITIVES I FUNCIONALS

## RESUM EXECUTIU

Aquest estudi desenvolupa un model predictiu per al diagnòstic de la malaltia d'Alzheimer utilitzant dades clíniques estàndard. El model XGBoost calibrat amb funció sigmoide aconsegueix un rendiment excepcional, amb una àrea sota la corba ROC de 0.948 i una precisió del 93.8%. L'anàlisi identifica les activitats de la vida diària (ADL), l'avaluació funcional i el test MMSE com les variables més predictives. El model demostra consistència entre diferents subgrups demogràfics, amb un rendiment particularment notable en pacients majors de 85 anys (AUC: 0.982). Aquests resultats suggereix que les avaluacions cognitives i funcionals estàndard poden constituir una eina efectiva per al suport al diagnòstic d'Alzheimer en entorns clínics.

**Paraules clau:** Alzheimer, predicció, aprenentatge automàtic, avaluació cognitiva, diagnòstic precoç

## 1. INTRODUCCIÓ I CONTEXT

### 1.1 Fons del Problema

La malaltia d'Alzheimer representa una de les principals causes de discapacitat i dependència entre les persones grans a nivell mundial. El diagnòstic precís i oportú és essencial per a una gestió adequada de la malaltia, però continua representant un repte en la pràctica clínica diària.

### 1.2 Objectius de la Recerca

Aquest projecte busca respondre a la pregunta: "Fins a quin punt les avaluacions cognitives i funcionals estàndard permeten predir el diagnòstic d'Alzheimer?" Mitjançant tècniques d'aprenentatge automàtic, s'ha desenvolupat un model predictiu que pot servir com a eina de suport als professionals sanitaris.

### 1.3 Context del Dataset

Les dades utilitzades en aquest estudi provenen d'un dataset públic disponible a Kaggle, que conté informació de 2.149 pacients i inclou múltiples variables clíniques, demogràfiques i d'avaluació cognitiva.

## 2. DADES I METODOLOGIA

### 2.1 Descripció del Dataset

- Font:** Kaggle - Alzheimer's Disease Dataset
- Població:** 2.149 pacients
- Període:** Dades recollides de manera transversal
- Prevalença:** 35.3% (758 casos positius de 2.149)
- Variables inicials:** 35 variables incloent dades demogràfiques, clíniques i d'avaluació

### 2.2 Preprocessat de Dades

El procés de preprocessat ha inclòs:

- Eliminació de variables no predictives (PatientID, DoctorInCharge)
- Imputació de valors missing mitjançant mediana per variables numèriques i moda per variables categòriques
- Conversió de variables binàries a format categòric
- Creació de grups d'edat (<65, 65-74, 75-84, 85+)
- Estandarització de variables numèriques
- Codificació one-hot de variables categòriques

### 2.3 Models Implementats

S'han comparat quatre enfocaments de modelatge:

- Regressió Logística:** Model lineal amb regularització
- Random Forest:** Ensemblatge amb 200 arbres de decisió
- XGBoost:** Algoritme de gradient boosting
- XGBoost Calibrat:** Versions amb calibració sigmoide i isotònica

### 2.4 Estratègia d'Avaluació

- Divisió de dades:** 60% entrenament (1.289 pacients), 20% validació (430 pacients), 20% test (430 pacients)
- Validació:** Hold-out estratificat per mantenir la distribució de la variable objectiu (Diagnosis)
- Mètriques principals:** ROC-AUC, PR-AUC, Brier Score, precisió, sensibilitat, especificitat
- Tècniques de calibració:** Platt scaling i calibració isotònica

## 3. RESULTATS

### 3.1 Rendiment Comparatiu dels Models

El model XGBoost calibrat amb funció sigmoide ha demostrat el millor rendiment global, amb una àrea sota la corba ROC de 0.948 i una àrea sota la corba Precision-Recall de 0.939. La Taula 1 resumeix el rendiment de tots els models avaluats en el conjunt de test.

Taula 1. Comparativa de Rendiment dels Models en el Conjunt de Test

Model	ROC-AUC	PR-AUC	Brier Score
Regressió Logística	0.882	0.820	0.139
Random Forest	0.936	0.904	0.098
XGBoost	0.939	0.922	0.052
XGBoost + Sigmoid	0.948	0.939	0.053
XGBoost + Isotonic	0.946	0.947	0.052

### 3.2 Variables Predictives Més Importants

L'anàlisi d'importància de variables ha revelat patrons consistents entre els diferents models. El Random Forest i XGBoost han identificat variables similars com a més predictives, encara que amb diferents ponderacions.

Taula 2. Variables Més Importants per Tipus de Model

Posició	Random Forest	Importància	XGBoost	Importància
1	ADL	0.150	Memory Complaints	0.151
2	Functional Assessment	0.145	Behavioral Problems	0.131
3	MMSE	0.105	MMSE	0.108
4	Memory Complaints	0.098	Functional Assessment	0.096
5	Physical Activity	0.029	ADL	0.094

### 3.3 Anàlisi d'Equitat per Subgrups

El model ha demostrat un rendiment consistent a través de diferents subgrups demogràfics, amb variacions mínimes en les mètriques principals.

Taula 3. Rendiment per Grups d'Edat

Grup d'Edat	n	Sensibilitat	Precisió	ROC-AUC
<65 anys	63	0.870	0.909	0.896
65-74 anys	151	0.906	0.923	0.936
75-84 anys	135	0.927	0.950	0.957
85+ anys	81	0.886	0.969	0.982

Taula 4. Rendiment per Gènere

Gènere	n	Sensibilitat	Precisió	ROC-AUC
Dones	220	0.903	0.929	0.955
Homes	210	0.900	0.947	0.937

### 3.4 Anàlisi de Calibració

La calibració de probabilitats ha mostrat que el model XGBoost original ja produïa probabilitats ben calibrades, amb un Brier Score de 0.052. Les tècniques de calibració aplicades no han millorat significativament aquesta mètrica, indicant que les probabilitats inicials ja eren properes a les freqüències observades.

### 3.5 Matrius de Confusió i Llidars Operatius

La matriu de confusió del model amb llindar 0.5 ha mostrat un bon equilibri entre sensibilitat i especificitat:

#### Matriu de Confusió - Llindar 0.5

- Veritables Negatius: 269 (62.6%)
- Falsos Positius: 9 (2.1%)
- Falsos Negatius: 15 (3.5%)
- Veritables Positius: 137 (31.9%)

L'anàlisi de diferents llindars operatius ha revelat que el model manté una sensibilitat consistent del 90.1% across diferents llindars, mentre que l'especificitat augmenta lleugerament amb llindars més alts.

## 4. DISCUSSIÓ

### 4.1 Interpretació dels Resultats Principals

Els resultats d'aquest estudi demostren que les avaluacions cognitives i funcionals estàndards poden constituir una base sòlida per a la predicció de la malaltia d'Alzheimer. L'ROC-AUC de 0.948 obtingut pel model XGBoost calibrat indica una capacitat discriminativa excel·lent.

La consistència en la identificació de variables clau entre diferents models (ADL, MMSE, Functional Assessment) reforça la validesa d'aquests constructes com a indicadors rellevants en el procés diagnòstic. És particularment notable que les avaluacions funcionals (ADL) emergeixin com a variables tan predictives, ja que això reflecteix la importància de l'impacte de la malaltia en la vida diària del pacient.

### 4.2 Implicacions Pràctiques

El model desenvolupat ofereix diverses aplicacions potencials en entorns clínics:

**Per a cribatge poblacional** (llindar 0.3):

- Alta sensibilitat (90.1%) assegura la detecció de la majoria dels casos
- Taxa acceptable de falsos positius (2.3%)
- Ideal per a programes de detecció precoç en atenció primària

**Per a diagnòstic en consulta** (llindar 0.5):

- Equilibri òptim entre sensibilitat (90.1%) i especificitat (96.8%)
- Baixa taxa d'errors tant falsos positius com falsos negatius
- Apropiada per a l'ús rutinari en especialitats relacionades

**Per a confirmació diagnòstica** (llindar 0.7):

- Alta precisió (93.8%) assegura la fiabilitat dels casos identificats
- Especialment útil en contextos d'alta especialització
- Pot suportar decisions terapèutiques importants

### 4.3 Equitat i Validesa Externa

L'anàlisi per subgrups revela que el model manté un rendiment consistent a través de diferents grups d'edat i gènere. La lleugera variació en sensibilitat observada en el grup de menys de 65 anys (87.0%) pot reflectir diferències en la presentació clínica de la malaltia en aquest grup d'edat, però no representa una diferència clínicament significativa.

El rendiment excepcional en el grup de majors de 85 anys (AUC: 0.982) és particularment rellevant, ja que aquest grup representa una proporció important dels pacients amb Alzheimer i sovint presenta desafiaments diagnòstics addicionals.

### 4.4 Limitacions i Consideracions

Tot i els resultats prometedors, cal considerar algunes limitacions:

- Naturalesa del dataset:** Encara que el dataset de Kaggle proporciona una base sòlida per a l'anàlisi, la falta d'informació detallada sobre els criteris de diagnòstic i el context de recollida de dades pot limitar la generalització dels resultats.
- Variables disponibles:** L'estudi no inclou marcadors biològics o d'imatge, que són components importants en el diagnòstic contemporani d'Alzheimer.
- Validació externa:** Es necessitaria validació en cohorts independents per confirmar la generalització dels resultats.

## 5. LIMITACIONS I FUTURES LINIES DE RECERCA

### 5.1 Limitacions Methodològiques

La principal limitació d'aquest estudi rau en l'ús d'un dataset públic sense informació detallada sobre els procediments de recollida de dades i els criteris diagnòstics específics utilitzats. Això pot afectar la validesa externa dels resultats.

Adicionalment, la naturalesa transversal de les dades no permet l'anàlisi de la progressió de la malaltia al llarg del temps, que és un aspecte crucial en l'Alzheimer.

### 5.2 Futures Direccions

Les següents línies de recerca podrien expandir les troballes d'aquest estudi:

- Integració amb biomarcadors:** Combinar les avaluacions clíniques amb marcadors biològics (biomarcadors en líquid cefalorraquidi, PET amiloide) per millorar l'exactitud diagnòstica.

2. **Models longitudinals:** Desenvolupar models que capturin canvis en les avaluacions al llarg del temps per a la detecció més precoç de la malaltia.
3. **Validació multicèntrica:** Validar el model en cohorts independents de diferents centres i poblacions.
4. **Implementació en entorns reals:** Provar la utilitat clínica del model en entorns assistencials reals mitjançant estudis d'implementació.

## 6. CONCLUSIONS

---

Aquest estudi demostra que les avaluacions cognitives i funcionals estàndard, quan s'analitzen mitjançant tècniques avançades d'aprenentatge automàtic, poden proporcionar una predicció altament precisa de la malaltia d'Alzheimer. El model desenvolupat aconsegueix un rendiment excepcional (ROC-AUC: 0.948) i manté aquest rendiment a través de diferents subgrups demogràfics.

Les implicacions pràctiques d'aquests resultats són significatives, ja que suggereixen que les dades clíniques rutinàriament disponibles podrien ser utilitzades més efectivament per al suport al diagnòstic d'Alzheimer. El model ofereix flexibilitat operativa mitjançant diferents llindars, adaptant-se a diversos contextos clínics des del cribatge fins a la confirmació diagnòstica.

Si bé es requereix validació addicional abans de la implementació clínica generalitzada, aquest enfocament representa un pas prometedori cap a l'optimització del procés diagnòstic en la malaltia d'Alzheimer.

## 7. REFERÈNCIES

---

1. Dataset: Alzheimer's Disease Dataset. Kaggle. Disponible a: <https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data>
2. McKhann, G. M., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.
3. OpenAI. (2024). ChatGPT [Model de llenguatge gran]. <https://chatgpt.com>

\*Nota: Aquesta eina d'IA es va utilitzar com a assistent en el desenvolupament d'aquest projecte per a: • Optimització de codi Python i implementació de models • Anàlisi de resultats i validació de mètriques • Revisió d'estratègies de preprocessament de dades • Assistència en la redacció tècnica de l'informe

Tot el disseny del projecte, les decisions metodològiques i la interpretació dels resultats van ser realitzades pel/la estudiant, amb l'IA funcionant com a eina de suport tècnic.