

# Negation and Uncertainty Detection using Classical and Machine Learning Techniques

Judit Félez Guerrero<sup>a,1</sup> and Èlia Campos Villaró<sup>b,2</sup>

<sup>a</sup>1704833  
<sup>b</sup>1703842

**Abstract**—An abstract is a brief summary that outlines the key aspects of a work. An example of a famous abstract is reproduced verbatim here for illustration purposes [vaswani\_attention\_2017]: The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results.

**Keywords**—a, b, c, d

## Contents

1	Introduction	1
2	EDA, Exploratory Data Analysis	1
2.1	Descripció general del dataset	1
2.2	Anàlisi	1
3	Preprocessing	2
4	Metric selection	2
5	Model selection	2
6	Final analysis	2
	References	2

## 1. Introduction

L'objectiu d'aquesta pràctica és desenvolupar un model capaç de classificar pacients segons la presència o absència de la malaltia d'Alzheimer. La detecció precoç d'aquesta malaltia és clau per millorar la qualitat de vida dels pacients i optimitzar el tractament, i els models de classificació automàtica poden ser una eina per ajudar al professionals de la salut.

Per dur a terme aquest estudi, hem utilitzat una base de dades disponible a la plataforma *Kaggle*, <https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset>, que conté informació clínica de 2149 pacients. Cada pacient està descrit amb 35 variables, que inclouen mesures mèdiques, resultats de proves i dades demogràfiques. La selecció d'aquesta base de dades es va basar en la seva mida suficient i la varietat d'informació disponible, que permet explorar diferents enfocaments de modelització

Aquest informe documenta tot el procés seguit, des de l'exploració inicial de les dades i el preprocessament d'aquestes, fins a la selecció de mètriques i models, per finalitzar amb un anàlisi dels diferents resultats obtinguts. L'objectiu final és identificar quin model ofereix la millor precisió i robustesa per a la classificació de pacients amb Alzheimer.

## 2. EDA, Exploratory Data Analysis

### 2.1. Descripció general del dataset

El dataset utilitzat en aquest estudi conté informació clínica i demogràfica de 2149 pacients, identificats mitjançant un codi únic, *PatientID*, que va des de 4751 fins a 6900. Per a cada pacient disposem d'un conjunt ampli de variables que inclouen dades demogràfiques,

factors d'estil de vida, historial mèdic, mesures clíniques, avaluacions cognitives i funcionals, simptomatologia i informació diagnòstica.

Les variables demogràfiques inclouen l'edat (*Age*, de 60 a 90 anys), el gènere (*Gender*, on 0 representa home i 1 dona), l'etnicitat (*Ethnicity*, codificada com 0: Caucasian, 1: African American, 2: Asian, 3: Other) i el nivell educatiu (*EducationLevel*, amb valors des de 0: None fins a 3: Higher),

Pel que fa als factors d'estil de vida,, el dataset recull l'índex de massa corporal (*BMI*, entre 15 i 40), consum de taba (*Smoking*), consum setmanal d'alcohol (*AlcoholConsumption*, de 0 a 20 unitats), el nivell d'activitat física setmanal (*PhysicalActivity*, de 0 a 10 hores), així com puntuacions de qualitat de la dieta (*DietQuality*, de 0 a 10) i de la son (*SleepQuality*, de 4 a 10).

L'historial mèdic dels pacients està representat per variables binàries que indiquen la presència de diferents condicions: antecedents familiars d'Alzheimer (*FamilyHistoryAlzheimers*), malaltia cardiovascular (*CardiovascularDisease*), diabetis, depressió, lesions cranials (*HeadInjury*) i hipertensió.

Les mesures clíniques inclouen la pressió arterial (*SystolicBP*, 90–180 mmHg; *DiastolicBP*, 60–120 mmHg) i diversos paràmetres de perfil lipídic, com el colesterol total, LDL, HDL i els triglicèrids.

Pel que fa a les avaluacions cognitives i funcionals, el dataset incorpora la puntuació MMSE (*Mini-Mental State Examination*, entre 0 i 30), l'avaluació funcional (*FunctionalAssessment*, 0–10), la presència de queixes de memòria, problemes de comportament i la puntuació d'activitats de la vida diària (*ADL*).

Els símptomes associats al deteriorament cognitiu també es registren com a variables binàries, incloent la confusió, desorientació, canvis de personalitat, dificultats per completar tasques i oblitats freqüents.

Finalment, la variable objectiu del nostre estudi és la columna *Diagnosis*, que indica la presència (1) o absència (0) de diagnòstic d'Alzheimer. La qual està desbalancejada, ja que aproximadament el 65% dels pacients no presenten Alzheimer a diferència del 35% que sí que el presenten.

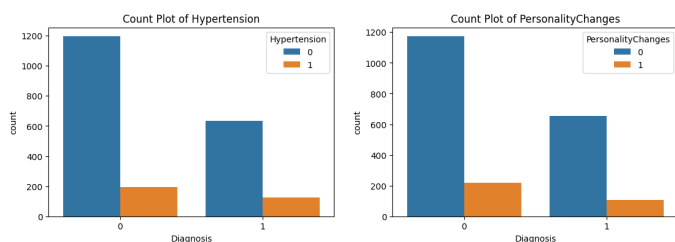
El dataset també inclou una columna denominada *DoctorInCharge*, que conté informació confidencial i presenta el mateix valor per a tots els pacients, de manera que no aporta variabilitat i quedarà exclosa de l'anàlisi.

### 2.2. Anàlisi

De les primeres coses que vam fer després de descarregar les dades va ser classificar-les en variables numèriques i categòriques, ja que el tipus de gràfic a generar depèn de la naturalesa de la variable.

Vam considerar les variables categòriques aquelles amb un màxim de deu valors únics. Per a aquestes, es van generar gràfics de barres agrupades pel diagnòstic, amb l'objectiu d'identificar variables que mostressin diferències evidents entre pacients amb i sense Alzheimer. La majoria de variables no mostraven gran canvis en les proporcions segons el diagnòstic, com es pot observar a la *Figura 1*:

Algunes variables, però, presentaven diferències significatives. Per exemple, la variable *MemoryComplaints* mostrava una distribució clarament diferenciada segons el diagnòstic, *Figura 2*, fet que era esperable.



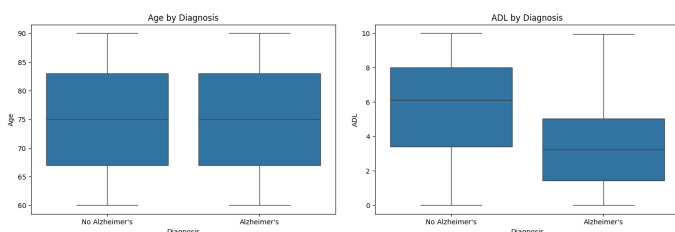
(a) Distribució de la hipertensió segons el diagnòstic

(b) Distribució de queixes de memòria segons el diagnòstic

**Figure 1.** Comparativa de variables categòriques segons el diagnòstic

Per a les variables numèriques, es van utilitzar boxplots, ja que els count plots no són informatius per a aquests tipus de dades. En general, moltes variables numèriques no mostren diferències evidents entre grups, com és el cas de l'edat, *Figura 3*.

Algunes variables, com *ADL*, sí mostren diferències clares entre pacients amb i sense Alzheimer, *Figura 4*.

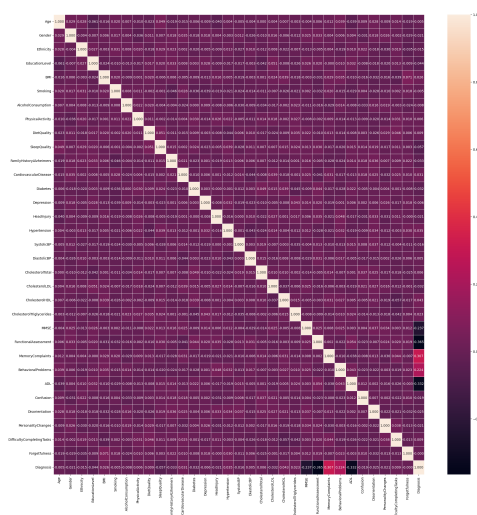


(a) Distribució de l'edat segons el diagnòstic

(b) Distribució de l'ADL segons el diagnòstic

**Figure 2.** Comparativa de variables numèriques segons el diagnòstic

Per obtenir una anàlisi més sòlida de la rellevància de cada variable, vam calcular la matriu de correlació entre totes les variables numèriques, *Figura 5*. Aquesta matriu va indicar que les correlacions entre variables eren molt baixes, cosa que ens va portar a la decisió de no eliminar cap variable i considerar-les totes en el model.

**Figure 3.** Matriu de correlació entre variables numèriques

A partir d'aquests resultats, vam concloure que certs models com KNN podrien no ser els més indicats per a aquesta base de dades, ja que les variables tenen poca correlació i la diferenciació entre classes no és immediata a simple vista.

### 3. Preprocessing

El tractament de dades realitzat va començar analitzant la informació proporcionada per la funció `df.info()`. L'output indicava que totes les variables contienien el mateix nombre de valors no nuls (2149, corresponents a tots els pacients), excepte *DoctorInCharge*, que com ja s'havia descrit, tenia el mateix valor per a tots els individus. Atès que no aporta informació rellevant, es va eliminar del dataset.

Com que les variables categòriques estaven codificades numèricament i no com a text, no va ser necessari realitzar cap procés d'*encoding* addicional.

A continuació, es va fer una detecció d'outliers per a cada variable numèrica utilitzant el mètode del rang interquartílic (IQR). Després d'aplicar aquest criteri, es va observar que les dades no contienien valors anòmals destacables, de manera que no va ser necessari aplicar cap tractament addicional.

FALTA PARLAR DEL TRAIN-TEST SPLIT PERCENTATGE UTILITZAT

### 4. Metric selection

La selecció de mètriques és un aspecte fonamental en l'avaluació de models de classificació, especialment en problemes on les classes estan desbalancejades. En el nostre cas, la variable objectiu, *Diagnosis*, presenta una distribució desigual: aproximament el 65% dels pacients no tenen Alzheimer, mentre que el 35% sí en presenten. Aquest desequilibri implica que una mètrica basada únicament en el percentatge d'encerts, com l'*accuracy*, podria donar resultats enganyosos. Per exemple, un model que classifiqués tots els pacients com a sans ja obtindria una precisió del 65%, tot i ser completament inútil per a la detecció de la malaltia.

Per aquest motiu, hem optat per utilitzar l'*F1-score* com a mètrica principal d'avaluació. L'*F1-score* combina la *precision* i el *recall* en una única mesura, mitjançant la seva mitjana harmònica. Aquesta característica la fa especialment adequada en escenaris on és igualment important reduir els falsos positius, que podrien generar diagnòstics erronis i preocupació innecessària, i minimitzar els falsos negatius, encara més crític en el context mèdic, ja que un pacient amb Alzheimer no detectat podria no rebre el tractament o seguiment adequat.

A diferència de l'*accuracy*, l'*F1-score* penalitza de manera equilibrada els errors en qualsevol de les dues direccions i reflecteix millor el rendiment real del model en situacions de desbalanceig com la nostra.

Tot i centrar-nos en aquesta mètrica, també hem tingut en compte el *precision* i el *recall* per separat, ja que poden oferir informació complementària sobre el comportament dels models. No obstant això, és l'*F1-score* la mètrica que hem utilitzat per comparar els diferents models i seleccionar-ne el millor.

### 5. Model selection

### 6. Final analysis

Tau  $\LaTeX$  template built by Guillermo Jimenez.