Yamanki Santander Cruz

Detección de Indicadores Cognitivos de Demencia Temprana mediante la aplicación de Algoritmos de Inteligencia Artificial y Procesamiento de Lenguaje Natural.

2022

Universidad Autónoma de Querétaro

Facultad de Ingeniería

Detección de Indicadores Cognitivos de Demencia Temprana mediante la aplicación de Algoritmos de Inteligencia Artificial y Procesamiento de Lenguaje Natural.

Tesis

Que como parte de los requisitos para obtener el grado Maestro en Ciencias en Inteligencia Artificial

Presenta

Yamanki Santander Cruz

Dirigido por:

Dr. Saúl Tovar Arriaga

Co-Dirigido por:

Dr. Sebastián Salazar Colores

Universidad Autónoma de Querétaro

Facultad de Ingeniería

Maestría en Ciencias en Inteligencia Artificial

Detección de indicadores cognitivos de demencia temprana mediante la aplicación de algoritmos de inteligencia artificial y procesamiento de lenguaje natural.

TESIS

Que como parte de los requisitos para obtener el grado de Maestro en Ciencias en Inteligencia Artificial

Presenta:

Ing. Yamanki Santander Cruz

Dirigido por:

Dr. Saúl Tovar Arriaga

Co-Dirigido por:

Dr. Sebastián Salazar Colores

SINODALES

Dr. Saúl Tovar Arriaga
Presidente

Dr. Sebastián Salazar Colores
Secretario

Dr. Wilfrido Jacobo Paredes García
Vocal

Dr. Juan Manuel Ramos Arreguin
Suplente

Dr. Humberto Guendulain Arenas
Suplente

Centro Universitario, Querétaro, QRO. México.
Mayo 2022

A mis seres queridos, quienes me han acompañado a lo largo del camino, hasta aquí.

Agradecimientos

Agradezco primeramente al Consejo Nacional de Ciencia y Tecnología (CONACYT) por proveer los fondos para la realización de este trabajo de investigación y mi formación académica. De igual modo, a la Universidad Autónoma de Querétaro y a la dirección de Investigación y Posgrado de la Facultad de Ingeniería por proveer las herramientas y medios necesarios para poder finalizar mis estudios de posgrado, así como a mis profesores, guías en mi desarrollo académico, especialmente al Dr. Saul Tovar Arriaga quien dirigió mi proyecto de tesis, al Dr. Wilfrido Jacobo Paredes García por abrirme las puertas al mundo del Procesamiento de Lenguaje Natural, y al Dr. Sebastián Salazar Colores por su trabajo y apoyo incansable. Por último, pero no menos importante agradezco a mi Madre y amigos, sin su apoyo no hubiera sido posible concluir este objetivo.

Abstract

Dementia is a neurodegenerative disease that leads to the development of cognitive deficits like aphasia, apraxia, and agnosia. It is currently considered one of the major medical problems worldwide, affecting the elderly. This disease presents symptoms gradually, with complications varying throughout its stages. As patients' cognition deteriorates, they are unable to perform daily tasks without assistance, resulting in additional medical expenses. As it is an incurable disease, tools and methods are needed to take care of patients in their early stages. State-of-the-art methods have shown that the use of syntactic-type linguistic features provides a sensitive and non-invasive tool for detecting dementia in its early stage. However, these methods lack relevant semantic information. For the aforementioned, in this work, we propose a novel methodology based on semantic features approach by using sentence embeddings computed by Siamese BERT-Networks (SBERT) along with Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and Artificial Neural Network (ANN) as classifiers. Our methodology involves 17 demographic, lexical, syntactic and semantic features extracted from the Corpus Pit database provided by Dementiabank. The mutual information score demonstrates a dependence between our features and the MMSE score, proving that they are relevant for the dementia classification task. Experimental tests performance based on accuracy, precision, recall, F1 score (77%, 80%, 80%, 80%) have validated that our methodology performs better than syntax-based methods and the BERT approach when only linguistic features are used.

Resumen

La demencia es una enfermedad neurodegenerativa que conduce al desarrollo de déficits cognitivos como la afasia, la apraxia y la agnosia. Actualmente se considera uno de los principales problemas médicos a nivel mundial, afectando a las personas mayores. Esta enfermedad presenta los síntomas de forma gradual, con complicaciones que varían a lo largo de sus etapas. A medida que la cognición de los pacientes se deteriora, son incapaces de realizar las tareas cotidianas sin ayuda, lo que supone gastos médicos adicionales. Al tratarse de una enfermedad incurable, se necesitan herramientas y métodos para atender a los pacientes en sus primeras fases. Los métodos más avanzados han demostrado que el uso de características lingüísticas de tipo sintáctico proporciona una herramienta sensible y no invasiva para detectar la demencia en su fase inicial. Sin embargo, estos métodos carecen de información semántica relevante. Por lo anterior, en este trabajo proponemos una metodología novedosa basada en el enfoque de características semánticas mediante el uso de incrustaciones de oraciones computadas por redes BERT siamesas (SBERT) junto con máquina de soporte vectorial (SVM), vecinos más cercanos (KNN), bosques aleatorios y redes neuronales artificiales (RNA) como clasificadores. Nuestra metodología incluye 17 características demográficas, léxicas, sintácticas y semánticas extraídas de la base de datos Corpus Pit proporcionada por Dementiabank. La puntuación de información mutua demuestra una dependencia entre nuestras características y la puntuación MMSE, demostrando que son relevantes para la tarea de clasificación de la demencia. Los resultados de las pruebas experimentales basados en la exactitud, la precisión, la sensibilidad y la puntuación F1 (77%, 80%, 80%, 80%) han validado que nuestra metodología se comporta mejor que los métodos basados en la sintaxis y el enfoque BERT cuando sólo se utilizan características lingüísticas.

Abreviaturas y Siglas

EA – Enfermedad de Alzheimer

DCL – Deterioro Cognitivo Leve

DL – Demencia Leve

DM – Demencia Moderada

DS – Demencia Severa

MOCA – Montreal Cognitive Assessment – Evaluación cognitiva de Montreal

MMSE – Mini-Mental State Examination – Mini-Examen del Estado Mental

IA – Inteligencia Artificial

RNA – Red neuronal artificial

ML – Machine Learning – Aprendizaje automático

PLN – Procesamiento de Lenguaje Natural

BERT – Bidirectional Encoder Representations from Transformers – Representaciones de Codificadores Bidireccionales a partir de Transformadores

SBERT – Siamese BERT-Networks – Redes Siamesas BERT

SimCos – Similitud de Coseno

SVM – Support Vector Machine – Máquina de Soporte Vectorial

KNN – K – Nearest Neighbors – K- vecinos más cercanos

IM – Coeficiente de información mutua

Índice General

AGRADECIMIENTOS ..... **IV**

ABSTRACT ..... **V**

RESUMEN ..... **VI**

ABREVIATURAS Y SIGLAS ..... **VII**

ÍNDICE GENERAL ..... **VIII**

ÍNDICE TABLAS ..... **X**

ÍNDICE FIGURAS ..... **XI**

CAPÍTULO 1 ..... **1**

INTRODUCCIÓN **1**

1.1 DEMENCIA 1

1.2 DESCRIPCIÓN DEL PROBLEMA 3

1.3 JUSTIFICACIÓN 4

HIPÓTESIS **5**

OBJETIVOS **6**

1.4.1 OBJETIVO GENERAL 6

1.4.2 OBJETIVOS ESPECÍFICOS 6

CAPÍTULO 2 ..... **7**

ANTECEDENTES **7**

2.1 ESTADO DEL ARTE 7

2.2 CARACTERÍSTICAS LINGÜÍSTICAS DE PACIENTES EA 11

MARCO TEÓRICO **13**

2.3 PROCESAMIENTO DE LENGUAJE NATURAL 13

2.3.1 Análisis léxico *14*

2.3.2 Análisis sintáctico *15*

2.3.3 Análisis semántico *16*
2.4 Similitud de textos 17
2.5 Aprendizaje profundo + PLN 17
2.6 Aprendizaje Automático 20
2.6.1 K-vecinos más cercanos *20*
2.6.2 Bosques Aleatorios *23*
2.6.3 Máquina de Soporte Vectorial *25*
2.6.4 Red Neuronal Artificial *28*
2.7 Selección de las características 30
CAPÍTULO 3 ..... **32**
METODOLOGÍA **32**
3.1 Conjunto de datos 32
3.2 Metodología propuesta 33
3.3 Preprocesamiento 35
3.4 Extracción de Características 37
3.4.1 Extracción de características léxicas *37*
3.4.2 Extracción de características sintácticas *39*
3.4.3 Extracción de características semánticas *40*
3.5 Selección de características 45
3.6 Clasificadores automáticos 46
3.6.1 Selección de hiper parámetros *47*
CAPÍTULO 4 ..... **49**
RESULTADOS **49**
4.1 Métricas de evaluación 49
4.2 Metodología de evaluación 50
4.3 Discusión 55
CAPÍTULO 5 ..... **57**
CONCLUSIONES **57**
5.1 Trabajos Futuros 58
REFERENCIAS **59**

ANEXOS **69**

ANEXO 1 ..... **69**

MUESTRA DE TRANSCRIPCIÓN MANUAL DE EVALUACIÓN "COOKIE THIEF" 69

ANEXO 2 ..... **71**

DESCRIPCIÓN "GROUND-TRUTH" 71

ANEXO 3 ..... **72**

CERTIFICADO "FORMACIÓN EN LÍNEA SOBRE LA PROTECCIÓN DE LOS PARTICIPANTES EN LA INVESTIGACIÓN EN SERES HUMANOS" 72

CERTIFICADO NECESARIO PARA OBTENER ACCESO A LA BASE DE DATOS. 72

Índice Tablas

Tabla 1.Principales tipologías de demencia [9]. 2

Tabla 2. Prevalencia (%) de demencia estratificada por grupos de edad y zona [17]. 4

Tabla 3. Estado del Arte. 10

Tabla 4. Descripción de características lingüísticas de pacientes EA en sus etapas progresivas [43]. 12

Tabla 5. Funciones de Kernel, ... 27

Tabla 6. Preprocesamiento de Corpus. 36

Tabla 7. Palabras clave a mapear en las transcripciones [46]. 41

Tabla 8. Ideas principales extraídas a partir del sistema de clasificación de la coherencia temática basado en [45]. 42

Tabla 9. Modelos y sus hiper parámetros. 47

Tabla 10. Media de los resultados obtenidos durante la etapa de prueba utilizando únicamente rasgos léxico-sintácticos. 50

Tabla 11. Media de los resultados medios obtenidos durante la etapa de prueba utilizando características léxicas, sintácticas y semánticas. 51
Tabla 12 . Transcripción extraída del conjunto de Datos DementiaBank [84]. 70

Índice Figuras

Figura 1. Ejemplo de Árbol Sintáctico. 15
Figura 2. Adaptación propia del modelo BERT [59]. Los textos en bruto se introducen en el modelo para predecir la etiqueta binaria. Toma el token CLS como entrada y luego pasa al bloque Transformer, dando como resultados vectores incrustados. 18
Figura 3. Adaptación propia del modelo SBERT [60]. Teniendo la frase A y la frase B como entrada, las incrustaciones u y v se producen después de la agrupación BERT. La similitud de estas incrustaciones se calcula utilizando la similitud del coseno. 19
Figura 4. Modelo KNN 21
Figura 5. Método de codo. 22
Figura 6. Modelo Bosques Aleatorios. 24
Figura 7. Modelo SVM. 25
Figura 8. Kernelización 27
Figura 9. Arquitectura Red Neuronal Artificial. 28
Figura 10. Imagen utilizada en la prueba "Cookie Thief "del Examen diagnóstico de afasia de Boston, fuente [44]. 33
Figura 11. Metodología propuesta. 34
Figura 12. Proceso de extracción de características léxicas. 38

Figura 13. Comparación de observaciones durante la descripción a) Paciente EA en b) Paciente CN, tomado de [89]. 41
Figura 14. Proceso de extracción de características semánticas. 44
Figura 15. Histograma de la puntuación obtenida por cada una de las características relativas a la puntuación MMSE en el análisis de información mutua..... 46
Figura 16. Modelo Red Neuronal Artificial implementado. 48
Figura 17. Estadísticas de clasificadores métrica Exactitud. 52
Figura 18. Estadísticas de clasificadores métrica Precisión. 52
Figura 19. Estadísticas de clasificadores métrica Sensibilidad. 53
Figura 20. Estadísticas de clasificadores métrica Puntuación F1 53
Figura 21. Comparación estadística modelos implementados vs BERT 55

CAPÍTULO 1

Introducción

1.1 Demencia

La demencia es denominada como una enfermedad neurodegenerativa mayor [1], condición con subtipos etiológicos (véase Tabla 1), en los que se reconocen diferentes niveles de disfunción generalizados en Deterioro Cognitivo Leve (DCL), Demencia leve (DL), Demencia Moderada (DM) y Demencia Severa (DS) [2]. Afecta a los adultos mayores, mostrando sus primeros signos a la edad de 60 años. Los síntomas característicos son la pérdida de memoria y la presencia de múltiples déficits cognitivos como es el caso de la Afasia, que refiere a un trastorno de lenguaje, donde se ven afectadas la expresión oral, escrita y la comprensión lectora [3], la Apraxia que se presenta como una alteración en el sistema nervioso que dificulta la planeación motora [4], y la Agnosia que es la disminución de la percepción sensorial y dificultad para reconocer objetos [5]. Síndrome que se refleja en aspectos de las funciones y procesos intelectuales, como la atención, la memoria, la cognición, la toma de decisiones, la planificación, el razonamiento, el juicio, la comprensión perceptiva, el lenguaje y la función visoespacial. Dentro de sus subtipos se destaca la Enfermedad de Alzheimer (EA) siendo la más prevalente, del cual se registra 50 millones de personas diagnosticadas [6], seguido por la Demencia Vascular (DC) a la que se atribuye 20% de los casos registrados de demencia [7].

Debido a que es una enfermedad incurable el tratamiento se centra en retrasar el deterioro cognitivo. Lo que convierte a la detección temprana en la medida más eficaz para detener la progresión del deterioro cognitivo [8].

Tabla 1.Principales tipologías de demencia *[9]*.

Demencias	Etiología
Enfermedad de Alzheimer	Casos esporádicos, con un inicio tardío (≥ 65 años) y una etiología poco clara. El mejor predictor en este caso es la edad. Sin embargo, alrededor del 5 al 15% de los casos son familiares; el 50% de estos casos tiene un inicio temprano (presenil) (< 65 años) y habitualmente se relacionan con mutaciones genéticas específicas.
Demencia vascular	La demencia vascular típicamente ocurre cuando múltiples infartos cerebrales pequeños (o a veces hemorragias) producen suficiente pérdida neuronal o axonal como para deteriorar la función encefálica.
Enfermedad de Pick	Cambios patológicos en la demencia frontotemporal, como atrofia, pérdida neuronal, gliosis y por la presencia de neuronas anormales (células de Pick) que contienen inclusiones (cuerpos de Pick).
Enfermedad de Huntington	Resultante de una mutación en el gen huntingtin (HTT), que produce una repetición anormal de la secuencia CAG del DNA que codifica el aminoácido glutamina, produciendo una proteína grande denominada huntingtina, que se acumulan dentro de las neuronas.
Enfermedad de Parkinson	Es asume una predisposición genética. Alrededor del 10% de los pacientes tienen antecedentes familiares de enfermedad de Parkinson. Se han identificado varios genes anormales. Una mutación en la repetición rica en leucina cinasa 2 (LRRK2; también conocida como PARK8) es un gen que codifica la proteína dardarina.
Enfermedad de Creutzfeldt-Jakob	La ECJ se sospecha en los pacientes más jóvenes sintomáticos que se han expuesto a carne contaminada por priones.
Demencia en la infección por Sida (VIH)	Es causada por el daño neuronal del virus VIH. Sin embargo, puede ser el resultado de otros trastornos, como una infección secundaria por virus JC que produce un leuco encefalopatía multifocal progresiva y linfoma del sistema nervioso central.

El procedimiento para el diagnóstico consiste en corroborar déficits en dos o más áreas cognitivas con pruebas neuropsicológicas, exámenes físicos y neurológicos. Seguido por la aplicación de evaluaciones cognitivas como la Evaluación cognitiva de Montreal (MOCA) y mini examen del estado mental (MMSE). Por último, se realizan pruebas médicas para descartar otras causas de síntomas similares a la demencia [1].

1.2 Descripción del Problema

El diagnóstico de la demencia puede retrasarse debido a la naturaleza de sus síntomas. Es normal que los pacientes no tengan noción de lo que les acontece, es necesario contar con el apoyo de familiares y cuidadores, quienes al notar cambios relativos al padecimiento acudan con un especialista. Atender enfermedades relacionadas a una condición mental es aún un estigma para la sociedad, que a su vez implica atención médica especializada en servicios de salud mental a la cual no toda la población tiene acceso debido a situaciones de carácter socio-económico [10], lo que dificulta el diagnóstico y denominación. Esta situación se da principalmente en regiones en desarrollo como América Latina, donde la prevalencia de la demencia se estima en un 11% en la población adulta [11].

Por otro lado, el estudio de la demencia en México y el trato que se le ha dado resulta escaso comparado con el crecimiento y alcance que presenta. Al igual que la mayoría de los países en desarrollo, hay un retraso en el estudio de esta. Estos factores se suman a la problemática de la detección y tratamiento de este padecimiento que va al alza [10]. Es imperativo establecer un método de diagnóstico temprano para controlar el proceso degenerativo del paciente y garantizar un estilo de vida digno [12].

Con el objeto de brindar herramientas útiles para el diagnóstico temprano se han desarrollado investigaciones y algunos proyectos tecnológicos basados en software e Inteligencia Artificial (IA). La implementación de IA ha explorado el uso de Redes Neuronales Artificiales (RNA), algoritmos de Aprendizaje automático (ML), y técnicas de Procesamiento de Lenguaje Natural (PLN) para la clasificación de Demencia [13]. Sin

embargo, hay algunos retos computacionales que necesitan ser resueltos para minimizar errores en dichos trabajos. En general se necesita solucionar el método de procesamiento de datos, explorar características y patrones de demencia, así como buscar soluciones ante desventajas como la escasez de datos clínicos [14].

1.3 Justificación

La demencia afecta a más de 50 millones de personas mundialmente, cifra que aumenta cada 3 segundos. Se prevén 135 millones de personas afectadas para 2050 [15]. Considerándose un problema mayor de salud mundial. Actualmente más de 10 millones de personas viven con este padecimiento sólo en la región de América [16]. En la Tabla 2 se presentan porcentajes de prevalencia de demencia estratificada por grupos de edad y zona geográfica, publicados como una estimación de la prevalencia global de demencia en el mundo [17].

Tabla 2. Prevalencia (%) de demencia estratificada por grupos de edad y zona *[17].*

	60-64	**65-69**	**70-74**	**75-79**	**80-84**	**≥85**
Europa	0.9	1.4	3.3	5.9	12.1	24.7
América	0.8	1.6	3.2	6.8	12.9	30.5
África	0.4	0.8	1.6	3.0	5.7	12.3
Asia	0.8	1.7	3.4	6.2	13.1	24.3
Oceanía	0.7	1.3	2.6	4.7	9.0	16.0

Ya que la demencia predomina en adultos mayores, el incremento de longevidad implica un aumento en la tasa epidemiológica, pronosticando que enfermedades degenerativas como la demencia en sus diferentes etiologías aumente a nivel mundial [18].

Se estima que para el año 2050 la esperanza de vida se incremente a 81.29 años, por lo que la sociedad mexicana estará constituida en una buena parte por personas adultas mayores [19]. En México, en 2004 se registraba un total de 800 mil personas con demencia. Para 2030

se espera que el número de personas con demencia en el país aumente a más de 1.5 millones [20].

Siendo un padecimiento que repercute gravemente la cognición, a medida en que incrementa la severidad del deterioro el paciente se vuelve incapaz de valerse por sí mismo, volviéndose dependiente de familiares o cuidadores. Por consecuencia la calidad de vida del enfermo, así como la de aquellos que lo rodean se ve comprometida.

En los últimos años, el uso de algoritmos basados en IA, han mostrado avances significativos en la detección de patrones de demencia. Las técnicas de PLN han obtenido resultados prometedores en la búsqueda de patrones, combinados con algoritmos de ML logran aportar una herramienta fiable para la clasificación de demencia contribuyendo a la detección temprana [21].

Hipótesis

Una metodología basada en técnicas de Inteligencia Artificial y Procesamiento de Lenguaje Natural permitirá la búsqueda de patrones de demencia que contribuirá en la generación de un algoritmo de clasificación eficaz para el diagnóstico temprano de dicho padecimiento.

Objetivos

1.4.1 Objetivo general

Diseñar y desarrollar una metodología que permita la detección de patrones asociados a la demencia, basada en la aplicación de técnicas de Procesamiento de Lenguaje Natural y algoritmos de Inteligencia Artificial. Para efectuar evaluaciones clínicas que contribuyan a el diagnóstico temprano de Demencia.

1.4.2 Objetivos específicos

Objetivo específico 1.1: Inquirir indicadores de demencia para seleccionar las posibles técnicas de búsqueda de patrones, con base en el Estado del Arte.

Objetivo 2.1: Obtener el acceso a una base de datos adecuada para el desarrollo del proyecto.

Objetivo específico 3.1: Desarrollar una metodología que sirva como herramienta para la clasificación de Demencia utilizando algoritmos de Inteligencia Artificial y técnicas de Procesamiento de Lenguaje Natural.

Objetivo específico 3.1: Realizar las pruebas pertinentes al algoritmo para validar los resultados.

CAPÍTULO 2

Antecedentes

2.1 Estado del Arte

Durante los últimos años la tecnología ha permitido el diseño de ensayos clínicos preventivos, desarrollo de modelos y la elaboración de algoritmos dirigidos a la prevención de los déficits asociados con la demencia.

Los algoritmos de PLN basados en el aprendizaje profundo muestran una amplia área de oportunidad para hacer incursiones en el dominio de la salud debido a su capacidad para analizar grandes cantidades de datos multimodales mediante el procesamiento computacional [22, 21]. Sin embargo, estos métodos de PLN requieren una gran cantidad de datos para tener un buen rendimiento. Las bases de datos clínicas en el estudio de la demencia presentan escasez y acceso limitado a su información. Ante este inconveniente, Masrani et al. [23] construyeron un conjunto de datos formado por varios miles de entradas de blog, algunas de personas con demencia y otras de personas de control. Utilizan este conjunto de datos para diseñar un clasificador de EA basado en Bosques Aleatorios, KNN, y en RNA logrando una precisión del 84%. Las características lingüísticas de los pacientes con EA se han utilizado para la clasificación de esta enfermedad.

Roak et al. [24] han explorado el recuento de pausas y el análisis de la complejidad sintáctica extraída de las transcripciones de audio construidas manualmente a partir de las grabaciones realizadas durante los exámenes neuropsicológicos, mediante los cuales realizaron un modelo de clasificación basado en Máquina de soporte vectorial con Kernel polinomial de segundo orden, obteniendo una tasa de precisión del 86% resultante de su clasificación. Otros estudios proponen reforzar los modelos de clasificación añadiendo información demográfica

[25, 26], registros médicos electrónicos (EMR) [27] y medidas médicas como la puntuación MMSE [28]. Karlekar et al. [29] plantean un método de clasificación de tres modelos neuronales artificiales basados en redes neuronales convolucionales (CNNs), módulo de memoria (LSTM-RNNs), y su combinación para distinguir entre las muestras de lenguaje de los pacientes con EA y los pacientes de control logrando un 84,9 % de precisión en la base de datos Corpus Pit proporcionada por Dementiabank. Solís-Rosas et al. [30] realizaron un análisis sintáctico en profundidad, incluyendo pausas, palabras de relleno, palabras formuladas, reinicios, repeticiones, enunciados incompletos y habla difusa. Realizaron dos métodos de clasificación automática, uno utilizando una RNA de 3 capas y otro utilizando una SVM con un kernel polinomial. Su tasa de clasificación máxima alcanzó el 86,42% de precisión en la Colección de Conversaciones de Carolina [31]. En otro estudio, Eyigoz et al. [32] utilizan variables lingüísticas con variables clínicas y demográficas en sus modelos de predicción. Extraído de la base de datos del estudio cardíaco de Framingham [33], su estudio alcanzó una precisión del 70%.

En los últimos años se han realizado algunas investigaciones en las que se propone el uso de características acústicas extraídas de los registros de audio de las pruebas neuropsicológicas además de las características lingüísticas [34, 35, 36, 37]. Estos estudios basados en el conjunto de datos de habla ADReSS Challenge [38] consisten en dos tareas principales: la primera, predecir cuál sería la puntuación alcanzada por un paciente en el test MMSE, considerando su rendimiento durante una evaluación neuropsicológica, y la segunda, clasificar la demencia. Principalmente, un modelo SVM, un modelo Bosques Aleatorios y algunas arquitecturas de RNA han sido realizadas tratando de resolver estas tareas, logrando una tasa de precisión del 77%, una tasa de exactitud del 77%, una sensibilidad del 76% y una puntuación F1 del 77%.

Balagopalan et al. [39] han recurrido por primera vez el uso de características semánticas como la distancia media del coseno entre los enunciados y la distancia media del coseno entre los enunciados de 300 dimensiones WORD2VEC [40] y las unidades de contenido de las imágenes, además de las características acústicas y lingüísticas; así como, la clasificación del texto por representaciones de codificadores bidireccionales a partir de transformadores

(BERT). Tras la etapa de clasificación, obtuvieron una exactitud del 81%, una precisión del 83%, con sensibilidad del 79% y una puntuación f1 del 81%. En el caso de la clasificación por BERT, las métricas conseguidas fueron las siguientes, 83% de exactitud, 86% de precisión, 79% de sensibilidad y 83% de puntuación F1.

La Tabla 3 muestra un resumen de los trabajos antes mencionados donde se contiene la base de datos utilizada, las características extraídas, modelos de clasificación desempeñados y los resultados obtenidos en cada uno de los trabajos con la finalidad de tener un panorama claro de las propuestas del estado del arte.

A través de esta breve reseña de algunos de los proyectos realizados sobre la búsqueda de patrones de demencia mediante técnicas de PLN, podemos observar que se utilizaron principalmente características lingüísticas léxicas y sintácticas. Sin embargo, queda por explorar cómo los elementos semánticos aportan información útil a la hora de realizar una detección automatizada de la demencia.

Tabla 3. Estado del Arte.

Título	Autores	Base de Datos	Características	Modelo de clasificación	Resultados
Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. [24]	Roark, B., Mitchell,M., Hosom,J., Hollingshead,K. & Kaye, J.(2011)	Grupo de cohorte de Layton Aging & Alzheimer's Disease Center.	Puntajes de evaluaciones. Complejidad léxica. Medidas de duración de discurso.	SVM utilizando kernel polinomial	86% exactitud
Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia. [23]	Masrani, V., Murray, G., Field, T. & Carenini, G., (2017)	Texto recolectado de 2805 publicaciones de blogs públicos.	Complejidad sintáctica. Psicolingüística.	Clasificador bayesiano, RNA, Bosques Aleatorios, Árbol de decisión, KNN, Log. Reg, Bosques Aleatorios	84% exactitud
Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. [29]	Karlekar, S., Niu, T., & Bansal, M., (2018)	DementiaBank, Corpus Pitt.	Texto codificado con el método POS-Tag.	CNN, LSTM-RNN, CNN-LSTM	84.9% exactitud
Augmenting word2vec with latent Dirichlet allocation within a clinical Application. [41]	Budhk,A., & Rudzicz,F., (2018)	The Wisconsin Longitudinal Study (WLS), Corpus Pitt	Vector de temas. Temas inductores.	SVM usando kernel lineal, Regresión logística, Bosques Aleatorios, Potenciación del gradiente	77.5% exactitud
Search for Dementia Patterns in Transcribed Conversations using Natural Language Processing. [30]	Solís Rosas, D., Tovar Arriaga, S., & Aceves Fernandez, M.A., (2019)	The Carolina Corpus Conversation database	Complejidad sintáctica. Conteo de pausas.	RNA, SVM	86.9% exactitud
An Automatic System for Dementia Detection using Acoustic and Linguistic Features. [35]	Gonzalez-Atienza,M., Gonzalez-Lopez, J.A., &. Peinado, A.M (2020)	The ADReSS dataset	Características acústicas. Complejidad sintáctica.	LDA, SVM, Bosques Aleatorios, AdaBoost	87% exactitud
To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer. [39]	Balagopalan, A.,Eyre, B.,Rudzicz,F.,& Novikova, J.(2020)	The ADReSS dataset, Corpus Pitt	Características léxico-sintácticas, acústicas y semánticas.	SVM, RNA, Bosques Aleatorios, Clasificador bayesiano, BERT, LOSO	87% exactitud

2.2 Características lingüísticas de pacientes EA

El avance en el conocimiento de la demencia ha hecho posible realizar su diagnóstico en etapas tempranas, incluso antes de la aparición de síntomas clínicos, motivando a la revisión de los criterios de diagnóstico y la búsqueda de alternativas para atender la demencia en sus estados preclínicos. En la actualidad, los marcadores cognitivos tempranos de la demencia se centran en la memoria episódica y en la orientación [42], sin embargo, investigaciones recientes han descubierto que la memoria semántica y las habilidades lingüísticas pueden ser herramientas sensibles para el diagnóstico temprano [43].

La coherencia global del habla espontánea de los pacientes con EA muestra alteraciones de la comprensión semántica y de la pérdida de memoria. Las alteraciones de la comprensión semántica incluyen errores al nombrar objetos o acciones que dan lugar a una denominación categórica incorrecta de las entidades, un fallo en la organización de las estructuras de conocimiento semántico o una desconexión entre la organización jerárquica del conocimiento y la producción verbal [44]. Las alteraciones de la pérdida de memoria se reflejan en el vocabulario restringido de los pacientes con EA y en sus dificultades para encontrar palabras adecuadas [45]. El deterioro lingüístico aporta al diagnóstico diferencial y permite predecir el grado de severidad de la demencia [43]. La Tabla 4 muestra las principales características lingüísticas de pacientes EA en sus etapas progresivas.

Las evaluaciones neuropsicológicas se realizan para detectar dichas alteraciones y déficits. Una de las evaluaciones neuropsicológicas más utilizadas es el test “Cookie thief” [46], que consiste en la descripción de una imagen que proporciona información valiosa sobre el discurso oral del paciente. Este test permite el análisis lingüístico desde varias perspectivas, en las que se puede explorar la gramática, la expresión lingüística, la calidad de la información transmitida, el dominio del tema y la organización narrativa, identificadas como áreas de deterioro en las etapas preclínicas de la EA.

Tabla 4. Descripción de características lingüísticas de pacientes EA en sus etapas progresivas *[43].*

	Estado inicial o leve	Estado moderado	Estado severo.
Semántica	Dificultad para comprender oraciones de contenido complejo. Comprensión de ideas simples. Dificultad para recuperar palabras en conversación espontánea. Lenguaje fluido, pero poco concreto. Sustitución de palabras (parafasias semánticas, uso de circunloquios).	Dificultad moderada para recuperar palabras en conversación espontánea y en tareas específicas. Comprensión escasa de frases de contenido complejo. Dificultad para tareas de nominación y categorización. Aumento y predominio de parafasias semánticas. Reducción del vocabulario expresivo.	Comprensión únicamente de elementos significativos. Limitaciones importantes para nominar y recuperar las palabras. Comprensión solamente de palabras u oraciones muy cortas o con elementos familiares. Reducción del vocabulario. Utilización únicamente de elementos significativos. Predominio de parafasias semánticas.
Sintáctica	Dificultad leve para comprender oraciones de organización extensa y/o compleja. Estructura sintáctica conservada Repetición afectada para oraciones largas.	Comprensión de oraciones sintácticamente simples, Disminución en la compresión de secuencias y series. Producción verbal con formas sintácticas reconocibles. Se mantienen morfemas sintácticos. Repetición afectada para oraciones simples. Omisión de conectores y palabras funcionales de la oración.	Comprensión severamente alterada para frases o limitada a palabras cortas y familiares. Limitación en el uso de lenguaje automático. Repetición casi extinta, incluso para palabras monosílabas. Omisión frecuente de palabras funcionales.
Fonológica	Sistema fonológico conservado.	Escaso recobro de la representación auditiva de la palabra. Dificultades en el procesamiento del lenguaje oral. Confusión ocasional de patrones de pronunciación	Imprecisiones en la conversión fonológica. Ecolalia (repetición parcial o total de forma incontrolada de frases expresadas por el interlocutor). Palilalia (repetición de una palabra de forma incontrolada). Logoclonias (repetición de sílabas) de forma incontrolada.
Pragmática	Pérdida de preguntas y referencias del narrador. Disminución en tiempo y contenido del discurso. Divagación y tópico difuso Olvido, reiteración e ideas incompletas en la conversación.	Frases inacabadas Repetición de ideas en la conversación. Pérdida del tópico y abandono de la conversación. Limitación en la toma de turnos en la conversación.	Pérdida de ideas claves del discurso. No reconoce intencionalidad. Conversación casi ausente, limitada o imprecisa. Ausencia de automonitoreo y autocorrección. Mutismo. Imposibilidad para mantener el tópico.
Lectoescritura	Comprensión lectora conservada para contenidos simples. Escritura alterada en la forma (ej., perdida de acentuaciones, disortografía),	Comprensión lectora alterada para mensajes de alta complejidad gramatical. Escritura de palabras y frases cortas Reducción de la variedad de elementos en la redacción. Dificultad para iniciar de forma espontánea la escritura.	Afectación severa para la comprensión lectora. Afectación casi total de la escritura por componente apráxico. Puede conservarse la escritura de letras y/o palabras monosílabas.

Los métodos de Procesamiento del Lenguaje Natural (PLN) han tenido un buen desempeño en tareas de extracción de rasgos, y búsqueda de patrones en el contexto lingüístico para tareas relacionadas con el problema que se aborda en este trabajo [47, 48, 49].

Marco Teórico

2.3 Procesamiento de lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN), es una rama de Inteligencia Artificial, la cual trata computacionalmente el lenguaje en cada una de sus variantes. Este proceso es posible mediante la aplicación de modelos matemáticos determinados por lingüistas computacionales e ingenieros especialistas en el área, basándose en patrones estructurales y codificaciones que permiten traducir el lenguaje humano a lenguaje máquina, basados en teorías de N. Chomsky [50]. Entendiendo el lenguaje natural como la forma del humano para comunicarse y el lenguaje máquina como la información codificada para ser procesada por un ordenador.

El PLN da lugar al diseño de sistemas que realicen tareas lingüísticas complejas como lo son traducción de textos, resúmenes, recuperación de información, transcripción de audio, análisis de sentimientos etc. Un sistema PLN se sustenta en los niveles de análisis que conforman al lenguaje natural, definidos como fonología, morfología, sintaxis, semántica y pragmática [51].

- Nivel Fonológico: Refiere a la fonética de las palabras que involucra los sonidos (fonemas), resultantes de la producción lingüística en general.
- Nivel Morfológico: Construcción de significado de una palabra a partir de morfemas, siendo esta la unidad mínima analizable gramaticalmente.

- Nivel Sintáctico: Define la estructura por la cual se hace una conjunción de palabras para generar una oración, bajo las reglas gramaticales específicas de la lengua.
- Nivel Semántico: Atribuye significado a las palabras de forma aislada o a la conjunción de una oración.
- Nivel Pragmático: Contextualiza el uso de las oraciones reconociendo un subnivel recursivo. Donde se interpreta el discurso respecto al contenido e intencionalidad.

De esta manera las técnicas PLN se aproximan a la comprensión de los mecanismos humanos de comunicación y la manipulación del mismo.

2.3.1 Análisis léxico

El análisis léxico se refiere a la validación de las palabras desde un punto de vista lingüístico, derivado de los lexemas de la lengua en la que se ha escrito el texto [52]. Para ejemplificar esta tarea podemos identificar la palabra *CANTO*, y producir el análisis:

CANTAR+VERB+PRESIND+1P+SING

Donde CANTAR es su forma infinitiva, +VERB refiere al elemento gramatical con el que se identifica a la palabra, en este caso ‘verbo’; +PRESIND describe el tiempo presente indicativo, +1P indica la primera persona, y +SING indica el número singular. Las técnicas básicas empleadas en este análisis comienzan con ‘Tokenizar” el texto, donde se mapean las palabras empleadas aplicando una separación entre ellas. Tras la tokenización, se procederá a buscar las palabras en un diccionario y extraer su significado. En este nivel, tenemos pocos indicios de la estructura sintáctica y de la intencionalidad de las palabras. Sin embargo, se puede realizar un análisis relativamente superficial del texto.

2.3.2 Análisis sintáctico

El análisis sintáctico se refiere al uso de las palabras dentro de una frase [53]. Es necesario tener conocimiento concreto de las reglas gramaticales puesto que se basa en etiquetar cada uno de los componentes que aparecen en la oración y analizar cómo las palabras se combinan para formar construcciones gramaticalmente correctas. La metodología de este proceso consiste en descomponer la frase, generando un árbol sintáctico (Figura 1) que permite movernos por dicha estructura.

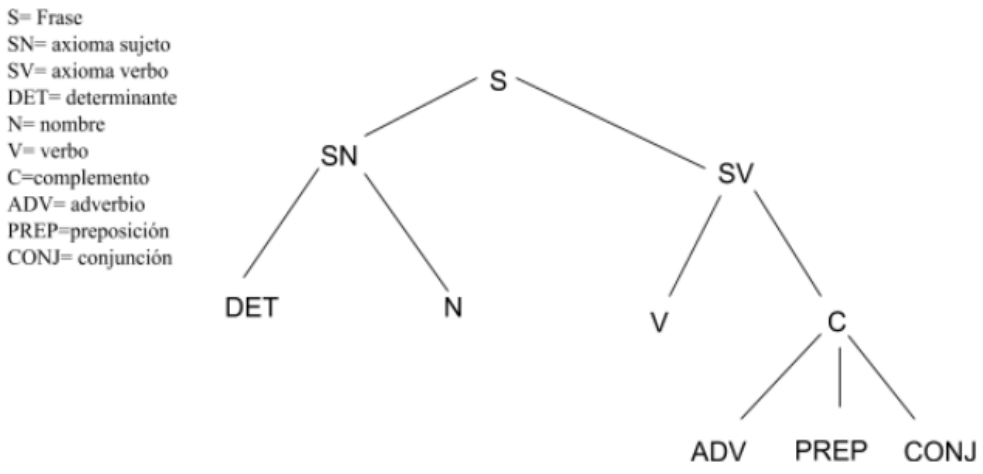

Figura 1. Ejemplo de Árbol Sintáctico.

Los árboles sintácticos son formas gráficas utilizadas para expresar la estructura de la oración, consistente en nodos conectados por ramas [54], construidos a partir de identificadores de elementos gramaticales. Con ello es posible realizar un análisis de la complejidad sintáctica y denotar errores gramaticales.

Las técnicas sintácticas utilizadas con frecuencia son la lematización, que se encarga de reducir la complejidad de las palabras a su forma simple, eliminado sufijos y conjugaciones. El análisis de dependencia, que evalúa las relaciones entre las palabras de una frase. Otra

etapa del análisis del texto consiste en asociar cada token a una categoría gramatical o parte del discurso (POS).

2.3.3 Análisis semántico

Hay dos enfoques de la semántica. El primero, se refiere al significado que tienen las palabras por sí mismas. El segundo, considera el significado que adquieren por su uso en una circunstancia determinada. El análisis semántico necesita trabajar sobre la estructura sintáctica a nivel computacional porque establece una relación que da sentido a una frase a través de las palabras que la preceden o proceden [54]. Por desgracia, esta es una de las tareas más difíciles referente al PLN ya que los procesos de la tecnología semántica ignoran la influencia del contexto o las intenciones del hablante; sin embargo, es posible contextualizar el texto matemáticamente.

Las técnicas de análisis semántico incluyen la extracción de entidades, consistente en identificar y extraer entidades categóricas como personas, lugares, o cosas. Es esencial para simplificar el análisis contextual del lenguaje natural.

Generación de lenguaje natural - Es el proceso de convertir la información de la intención semántica del ordenador en lenguaje humano legible. Los chatbots lo utilizan para responder de forma eficaz y realista a los usuarios.

Comprensión del lenguaje natural: consiste en convertir fragmentos de texto en representaciones estructuradas de forma lógica para que los programas informáticos puedan manipularlos fácilmente.

2.4 Similitud de textos

Las palabras que aparecen en contextos similares tienden a tener significados similares. La similitud se refiere a las características comunes entre dos instancias, una magnitud que se calcula por la distancia entre ellas, y se establece un umbral adecuado para decidir si son similares o no [55]. Esta métrica es muy útil para la minería de datos, la recuperación de información y los procesos de comparación de textos. En este caso, se ha aplicado para encontrar información que coincida con los puntos relevantes de la descripción. Esta investigación utiliza la similitud del coseno (SimCos) para comparar documentos extrapolando el texto plano a un espacio vectorial n-dimensional, calculando el coseno del ángulo formado entre dos vectores. Se calcula con la ecuación (1), donde *P* es el peso del documento, *n* representa el número de términos, *d* es el documento y *q* es la consulta [56].

$$Sim\cos(d,q) = \frac{\Sigma(p_{(n,d)} \times p_{(n,q)})}{\Sigma(p_{(n,d)})^2 x \Sigma(p_{(n,q)})^2} \quad (1)$$

2.5 Aprendizaje profundo + PLN

Manejar el lenguaje natural a nivel computacional es imposible sin antes realizar un embebido de este. Este proceso consiste en realizar una representación vectorial de la oración de tal manera que persevere su significado [57]. Mediante la embebido de frases, podemos extraer una representación numérica de una frase en la que es posible encapsular su contenido semántico. Los métodos de incrustación de frases suelen basarse en modelos lingüísticos neuronales profundos previamente entrenados, como GPT [58], BERT [59] y SBERT [60].

BERT [59] es un modelo transformador para PNL, desarrollado por Google. La arquitectura de su modelo es un codificador transformador bidireccional multicapa, diseñado para preentrenar representaciones bidireccionales profundas a partir de un texto plano mediante el condicionamiento conjunto de un contexto izquierda-derecha. Se entrena en un gran corpus de texto sin etiquetar que le permite aprender información a nivel de palabra y de frase, con el fin de realizar un enmascaramiento lingüístico eficiente. Es útil para tareas como el reconocimiento de entidades, el etiquetado del habla, la respuesta a preguntas, la predicción de palabras y la clasificación de textos. La arquitectura de su modelo se compone de capas de atención multicabezales. Existen variaciones de la arquitectura del modelo, sin embargo, su modelado básico consta de 12 bloques transformadores, 768 capas ocultas, 12 cabezas de autoatención, y el número total de parámetros para el modelo preentrenado es de 110M. La Figura 2 muestra la arquitectura del BERT.

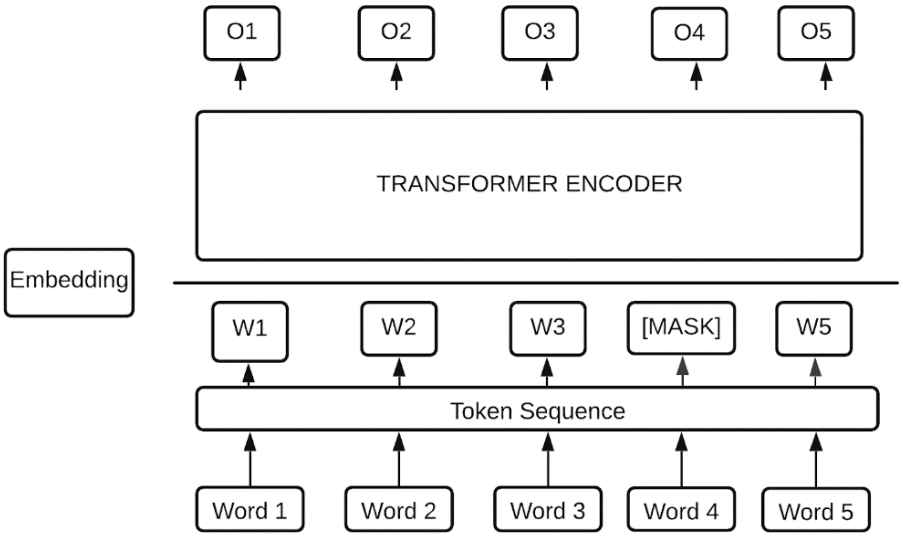

Figura 2. Adaptación propia del modelo BERT *[59]*. Los textos en bruto se introducen en el modelo para predecir la etiqueta binaria. Toma el token CLS como entrada y luego pasa al bloque Transformer, dando como resultados vectores incrustados.

Sin embargo, el modelo BERT no es capaz de realizar la búsqueda de similitudes semánticas ni la agrupación. Esta observación ha inspirado la realización de un nuevo modelo de

incrustación de frases. SBERT [60] es un modelo modificado basado en BERT que utiliza estructuras de red para derivar incrustaciones de frases semánticamente significativas realizando la similitud del coseno y otras tareas como la similitud textual semántica, la búsqueda semántica o la minería de paráfrasis. Integra la red siamesa con un modelo BERT preentrenado. SBERT añade una operación de pooling a la salida de BERT. Esta capa de agrupación nos permite crear una representación de tamaño fijo para frases de entrada de longitud variable. La Figura 3 muestra la arquitectura de SBERT.

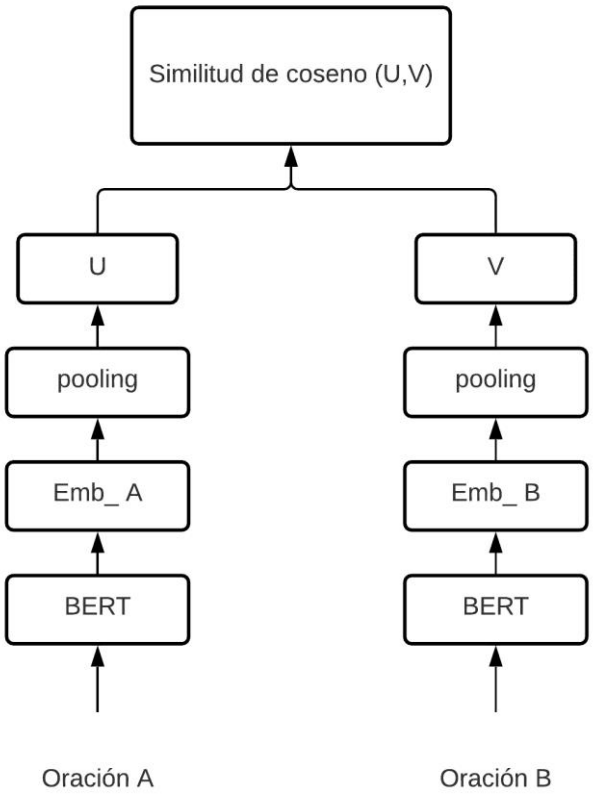

Figura 3. Adaptación propia del modelo SBERT *[60]*. Teniendo la frase A y la frase B como entrada, las incrustaciones u y v se producen después de la agrupación BERT. La similitud de estas incrustaciones se calcula utilizando la similitud del coseno.

2.6 Aprendizaje Automático

En la actualidad gracias a la tecnología y el acceso masivo a medios digitales, se hace plausible la generación y disposición de grandes cantidades de datos. Una de las ramas involucradas en el manejo, manipulación y análisis de datos es el aprendizaje automático, también conocido por su nomenclatura en el lenguaje inglés como *Machine Learning*. El cual podemos definir como un conjunto de métodos capaces de detectar patrones de manera automática para después utilizarlos en tareas de predicción, clasificación, etc. [61].

Debido a la amplitud de esta rama y a su aplicación multidisciplinaria hay una gran variedad de algoritmos y métodos que la conforman. Para este trabajo en específico los algoritmos de interés son aquellos capaces de resolver tareas de clasificación.

Entendemos a la clasificación como aprendizaje supervisado, de enfoque discriminante. Consiste en predecir salidas categóricas a partir de un conjunto de variables de entrada [62]. Existe una variedad de modelos de clasificación como el Clasificador Bayesiano, Árboles de decisión, Máquina de soporte vectorial, etc.

2.6.1 K-vecinos más cercanos

El algoritmo de clasificación supervisada K-vecinos más cercanos, (K-Nearest Neighbors, KNN), basa su lógica en el cálculo de distancias entre puntos. Todo tipo de dato puede ser proyectado como un punto en un espacio n-dimensional, donde n es el número de atributos de entrada. Los cuales se busca etiquetar por clase [63]. Visualizar una nube de puntos de dimensión $n > 2$ resulta complejo para la abstracción humana, sin embargo, matemáticamente es posible realizar cálculos que nos permiten conocer su relación. La distancia euclidiana es la magnitud que separa a dos puntos existentes en un espacio ($\mathbb{R}$) de n dimensiones ($\mathbb{R}^n$) dada por la norma de sus diferencias, es decir, como la magnitud del

vector que lleva a uno en otro. En $\mathbb{R}^2$ la Ecuación 2 define el cálculo de la distancia euclidiana [64].

$$d\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2}$$

El algoritmo hace una distinción entre las observaciones similares, consideradas los puntos más cercanos entre sí, a los que se denomina pertenecientes a una misma clase. De esta forma, a partir del uso de un conjunto de entrenamiento y tras definir un número K de puntos vecinos cercanos, el algoritmo aprende a realizar una clasificación, Figura 4.

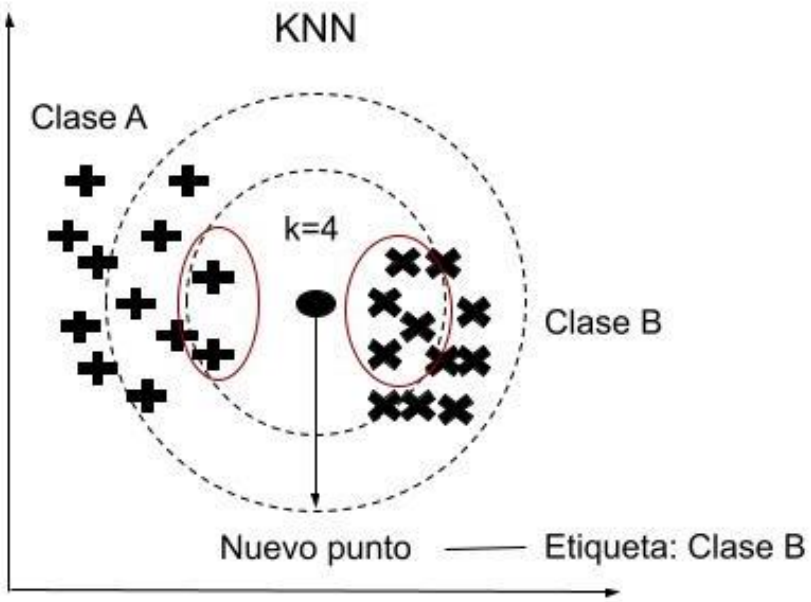

Figura 4. Modelo KNN

Definir un número k vecinos, es crucial en la efectividad de clasificación, así como en el desempeño computacional del algoritmo. Para poder hacer una correcta estimación del número k, se implementa el método de codo, consistente en la ejecución del modelo de forma iterativa, incrementando el número k [65]. El modelo es evaluado durante las ejecuciones, usando el cálculo del error. La puntuación obtenida es trazada dando como resultado una representación gráfica de la relación existente entre la exactitud con respecto al número k. La gráfica tiende a generar una curva donde la tasa de error disminuye, siendo el número k ideal aquel en el que se forma el "codo" de la curva [66], como se muestra en la Figura 5.

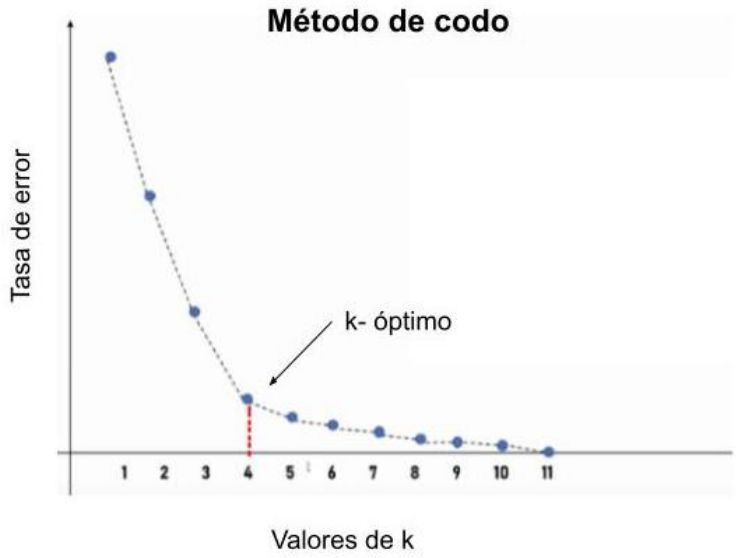

Figura 5. Método de codo.

El razonamiento del método pudiese parecer a simple vista empírico. Sin embargo, los conceptos de sesgo y varianza juegan un rol elemental en el cálculo del número k. Definimos el sesgo como la diferencia entre el valor objetivo y la predicción del modelo [67], y a la varianza como la medida de dispersión de los datos [68]. Los altos niveles de sesgo y la varianza implican un escaso aprendizaje de patrones debido a la simplificación de las características, así como sobreajuste de modelo, errores reflejados directamente en la

clasificación. El método de codo realiza una compensación de sesgo-varianza tras el aumento en el número de K.

El clasificador KNN presenta dificultades. Para determinar el vecino más cercano de un nuevo punto se debe calcular la distancia entre todas las instancias de entrenamiento. Por lo que el rendimiento en tiempo de ejecución es lento. Sin embargo, el modelo es sencillo ya que no es necesario ajustar parámetros complejos para su construcción.

Para obtener resultados de buena calidad, el modelo K-NN requiere una normalización para evitar cualquier sesgo y un número significativo de registros de entrenamiento con las máximas permutaciones posibles de los atributos de entrada. Aunque el modelo no es bueno para generalizar la relación entrada-salida, sigue siendo un modelo bastante eficaz para aprovechar las relaciones existentes en los registros de entrenamiento [63].

2.6.2 Bosques Aleatorios

El método de aprendizaje Bosques Aleatorios (Random Forest) es implementado para resolver tareas de clasificación y regresión. Su metodología de ensamble construye una colección de árboles de decisión a partir de selección aleatoria de características (Figura 6) [69]. Cada árbol de decisión se genera a partir de un indicador de selección de atributos, el coeficiente de ganancia de información, el ratio de ganancia y el índice de Gini para cada atributo [70]. Y se entrena utilizando una muestra independiente de los datos de entrenamiento. Cada árbol del conjunto actúa como clasificador base para determinar la etiqueta de clase de una instancia, siendo electa la de mayor aparición. En el caso de la regresión, la media de todas las salidas de los árboles se considera el resultado final.

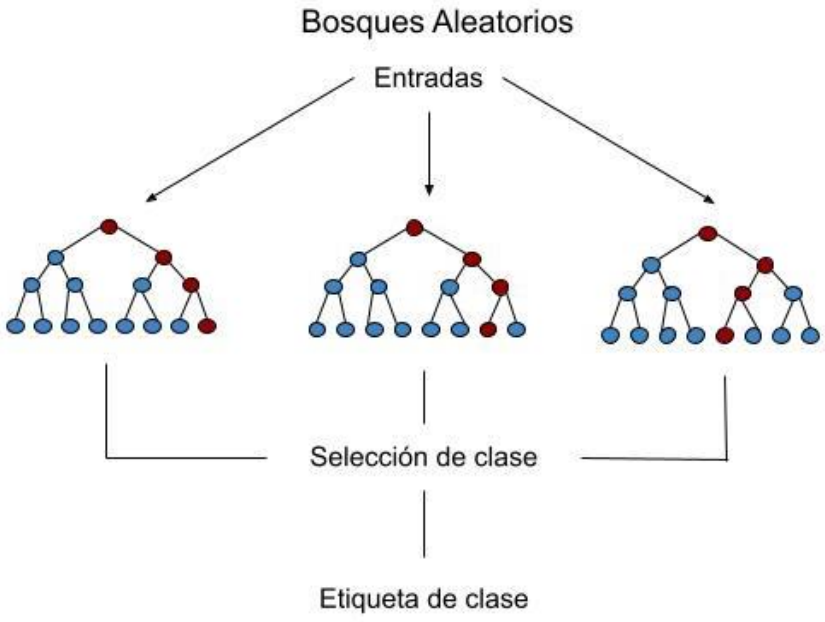

Figura 6. Modelo Bosques Aleatorios.

Los parámetros necesarios para definir el modelo son el número de árboles en el conjunto, el criterio de parada para cada árbol, el número de características a seleccionar aleatoriamente y el criterio de división [71]. Se considera que cuantos más árboles conformen el algoritmo, mejor será su rendimiento. Sin embargo, utilizar más árboles de los necesarios supone un desperdicio de recursos. Para elegir el número ideal de árboles, se evalúa el modelo iniciando con un número n propuesto, acrecentando n en cada prueba hasta el punto en el que el rendimiento deja de mejorar.

Los Bosques Aleatorios tienen un buen rendimiento, a costa de un gran uso de memoria. Una desventaja de este algoritmo es la pérdida de inteligibilidad del modelo. No obstante, el modelo no se enfrenta al problema del sobreajuste porque toma la media de todas las predicciones, anulando los sesgos.

2.6.3 Máquina de Soporte Vectorial

La Máquina de soporte vectorial, por sus siglas en inglés (SVM). Es un método de clasificación supervisado cuyo principio busca delimitar semi espacios para la separación de datos, partiendo de la suposición de que los datos pueden ser separados linealmente, es decir, existe un hiperplano que puede separar perfectamente las clases. Tras la definición de un conjunto básico de puntos que pueden ayudar a identificar y fijar el límite, el cual definimos como margen a partir de los vectores de soporte [72]. Si hablamos de un espacio $\mathbb{R}^2$, el límite puede ser una línea recta o una curva (Figura 7). En $\mathbb{R}^3$ puede ser un plano o una superficie compleja irregular. El modelo se construye al definir el hiperplano, el cual se consigue al dar solución a un problema de optimización. Dicho problema busca el margen máximo, es decir la distancia máxima entre los vectores de soporte de ambas clases.

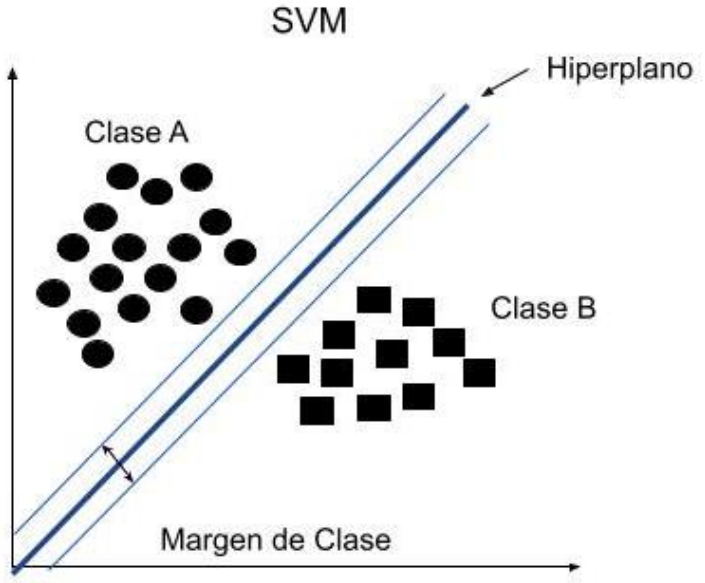

Figura 7. Modelo SVM.

Se requiere encontrar el par (w, b) que clasifique correctamente los vectores x_i en dos clases y_i [73], por lo que el hiperplano se expresa como:

$$H = w^T \cdot x_i + b \tag{3}$$

Donde *w* es el vector ortogonal al hiperplano, x_i los puntos que se encuentran sobre el hiperplano y b es el sesgo [72]. La distancia positiva y negativa del punto más cercano a H está dado por d^+ y d^- respectivamente, por lo que el margen de separación entre clases se define como:

$$\rho = d^+ + d^- = 2\frac{1}{\|w\|} \tag{4}$$

El problema de maximización del margen se formula:

$$\min_{(w,b)} \quad \frac{1}{2}\|w\|^2 \tag{5}$$

Sujeta a:

$$y_i(w^T x_i + b) \geq 1 \tag{6}$$

Se resuelve como problema de optimización cuadrática, mediante multiplicadores de Lagrange [74]. El Lagrangiano L del problema SVM es:

$$L(w_1 b, \alpha) = \frac{1}{2} w^T w - \sum \alpha_i [y_i(w^T x_i + b) - 1] \tag{7}$$

Una vez que se define un límite, se procede a comprobar si los datos se encuentran dentro de la frontera.

Suponemos que los datos son linealmente separables, sin embargo, esto regularmente no es una realidad. Por lo que el método de SVM emplea funciones matemáticas que denominamos como kernel, para proyectar los puntos generados por los datos, en un espacio dimensional superior, de modo que sea posible realizar una separación lineal [75]. La Kernelización consiste en el mapeo de las covariables de X en un espacio de mayor dimensión Z y aplicar

el clasificador en el espacio mayor Z [76]. Dando lugar a un clasificador de menor simplicidad computacional, Figura 8.

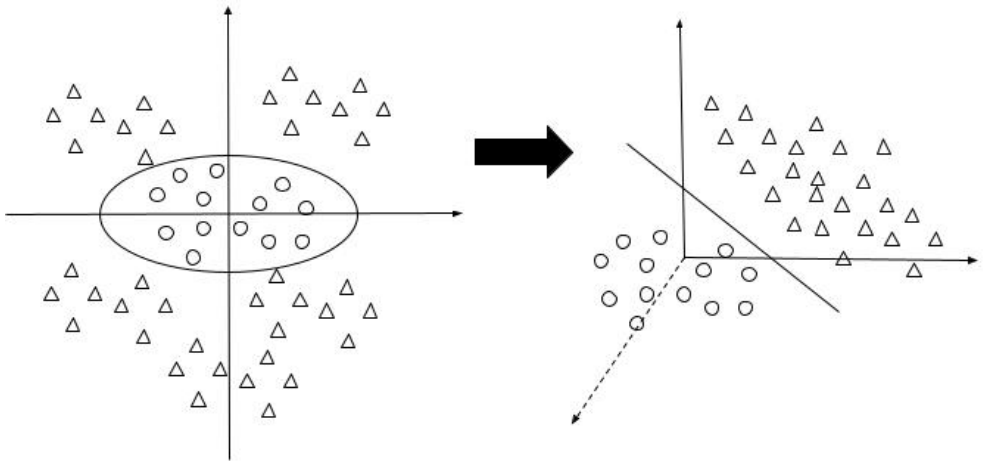

Figura 8. Kernelización.

Algunas de las funciones de kernel más comunes son polinomio, base radial y función sigmoidal. La Tabla 5 contiene algunos de los kernels de interés para este trabajo.

Tabla 5. Funciones de Kernel,

Kernel	Función
Lineal	$\langle x, x' \rangle$
Polinomial	$(\gamma \langle x, x' \rangle + r)^d$
RBF	$\exp(-\gamma \|x - x'\|^2)$

+

La etiqueta de clase se determina mediante la posición de la observación respecto al hiperplano. La ventaja de una SVM es que, una vez establecida la frontera, la mayor parte de los datos de entrenamiento son redundantes.

2.6.4 Red Neuronal Artificial

El modelo de red neuronal artificial busca emular el proceso biológico de una neurona. Su arquitectura básica puede describirse como un grafo cuyos nodos son el equivalente a las neuronas y las aristas a los enlaces entre ellas Figura 9 [77]. Existen variantes de esta arquitectura, sin embargo, suelen estar constituidas una agrupación de nodos cercana a la entrada denominada capa de entrada o nodos de entrada, a esta dependiendo de la arquitectura es seguida por un número n de capas, para concluir con una última capa de nodos o un único nodo de salida.

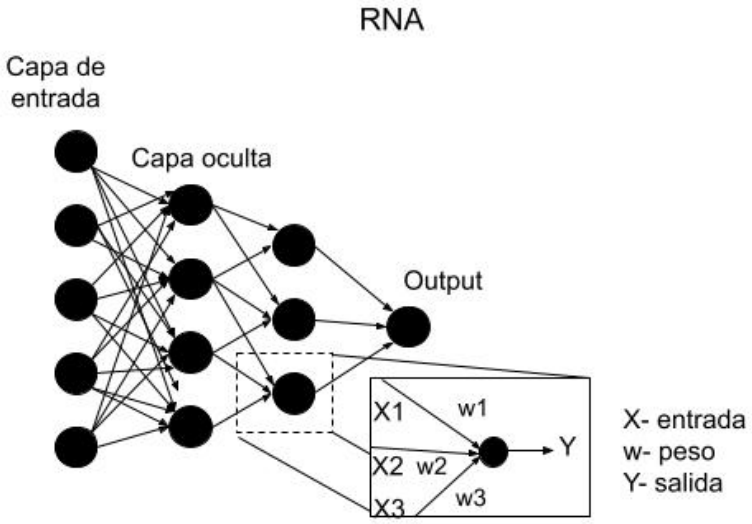

Figura 9. Arquitectura Red Neuronal Artificial.

Este algoritmo de autoaprendizaje comienza con un mapeo inicial aleatorio y, a partir de este se da paso a un autoajuste que ocurre durante una serie de iteraciones del modelo denominadas épocas.

La configuración sobre la que se construye el modelo, emplea parámetros conocidos como pesos, relacionados para ajustar con precisión la salida deseada. Una función de agregación,

la cual realiza la suma de todas las entradas ponderadas por sus pesos, una función de activación cuyo objetivo es acotar los valores de salida. Y una función de transferencia, la cual determina la salida total [78].

El método de aprendizaje de una red neuronal relaciona los atributos de entrada y la etiqueta de clase de salida a través de una técnica llamada retro propagación. La clave del entrenamiento es encontrar los pesos adecuados para una buena estimación. El modelo utiliza cada observación de entrenamiento para estimar el error de la salida predicha en comparación con la salida esperada [63]. El cálculo del error es usado para ajustar los pesos con el fin de minimizar el margen de error en el siguiente registro de entrenamiento, repetidas veces hasta que el error esté dentro del rango aceptable.

Como se ha mencionado existen variaciones en la estructura del modelo, propuestas tras la experimentación y búsqueda de solución de diversas tareas. La topología dos capas, como se muestra en la Figura 8, se denomina perceptrón, la forma más simple de red neuronal artificial. La utilización de múltiples capas brinda una capacidad de aprendizaje profundo para poder extraer características de nivel superior de los datos sin procesar.

Un aparente inconveniente de la implementación de un modelo RNA es el tiempo que conlleva optimizar los parámetros. No existen directrices consistentes sobre el número de capas ocultas y de nodos dentro de cada capa oculta. Por lo que habría que probar muchos parámetros para optimizar la selección de los mismos. Sin embargo, una vez construido el modelo, es fácil de implementar.

Un método que aborda los problemas anteriores de optimización de parámetros es el denominado aprendizaje por transferencia (Transfer Learnig), el cual, transfiere los conocimientos adquiridos en un modelo preentrenado, como punto de partida para un modelo en una nueva tarea relacionada. La práctica más común es el “Fine Tuning”, el cual consiste en entrenar previamente un modelo de red neuronal artificial a partir de un conjunto de datos de origen, para después generar un nuevo modelo denominado “Target Model” al cual se agrega una capa de salida [79]. La capa de salida se entrenará desde cero, mientras que los parámetros de todas las demás capas se ajustarán en función de los parámetros del modelo

de origen para beneficiarse del uso de bases de datos ricas en ejemplos que eviten el problema del exceso de sobreajuste.

2.7 Selección de las características

El interés por interpretar los resultados de los modelos de aprendizaje automático e implementar acciones que representen una mejoría en estos, ha sido abordado desde diferentes enfoques. La construcción de los modelos y la definición de sus parámetros no son las únicas variables que se anteponen en el resultado final [80]. Los datos de entrada y la manera en que se preprocesan son tan importantes como el modelo mismo.

Es importante entender que la utilización de atributos para representar información es por sí mismo un conocimiento sobre el problema. Estas instancias ocasionalmente contienen columnas que no están relacionadas con el objetivo. Del mismo modo, las características pueden no necesariamente no estar relacionadas, sino que son redundantes. Las características redundantes aumentan el espacio de búsqueda de parámetros del optimizador del modelo innecesariamente [81]. La eliminación de estas características de la selección de datos de entrada, aporta un gran valor a la disminución de ruido generado en el modelo. La selección de características puede entenderse como la reducción del conjunto de datos a un subconjunto de características que aporten información relevante.

Existen diferentes métricas que operan sobre las características, su distribución de valores, y exploran su relación con la clase objetivo. Su aplicación depende del tipo de característica en cuestión y a los objetivos. Las diferentes métricas se basan en supuestos sobre los datos.

Una forma fácil de seleccionar características es calcular la covarianza de cada característica con respecto a la variable dependiente seleccionando aquellas cuyos valores de correlación sean altos. Sin embargo, para este trabajo nos referiremos específicamente a la métrica de información ya que aventaja el método de la covarianza, ya que mide la dependencia general

de las variables aleatorias sin hacer ninguna suposición sobre la naturaleza de sus relaciones subyacentes.

La definición de información mutua (IM) depende de la distribución de dos variables aleatorias en espacios de probabilidad diferentes. IM es la cantidad de información conocida sobre una segunda variable aleatoria si conocemos la primera variable. Su principio se relaciona con la teoría de entropía de la información de una variable aleatoria. Mide la relación entre dos variables aleatorias, incluso si la relación no es lineal. Se basa en el concepto de probabilidad de relevancia de cada atributo medido empleando una prueba de permutación, lo que permite descartar las variables irrelevantes, así como ordenar por importancia las que son relevantes [82]. Se calcula mediante la ecuación (8), basada en la entropía de Shannon [83] del par "(X, Y)", donde p es la función de distribución de probabilidad conjunta de *X* e *Y*, y *P(x)* y *P(y)* son las funciones de distribución de probabilidad marginales de *X* e *Y* respectivamente.

$$I(x,y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{8}$$

La métrica IM generaliza bien a la clasificación multiclase, puede emplearse como métrica de múltiples características. Es importante señalar que la medida de puntuación de información mutua se define entre [0 - 1] [49]. Son posibles dos casos:

1. I(x,y) = 0, las variables son independientes, es decir, no comparten información.
2. I(x,y) > 0, existe una asociación entre X e Y, por lo que comparten información.

CAPÍTULO 3

Metodología

3.1 Conjunto de datos

Este estudio utilizó la base de datos Pitt Corpus [84, 85], integrada por audios, transcripciones, datos demográficos y los resultados de la prueba Mini-Examen del Estado Mental (MMSE) recogidos como parte del protocolo administrado por el Estudio de Alzheimer y Demencias Relacionadas de la Facultad de Medicina de la Universidad de Pittsburgh. Entre los participantes había 242 adultos mayores de control, 308 con enfermedad de Alzheimer diagnostica-probable y otros diagnósticos de demencia.

El Corpus Pitt contiene transcripciones resultantes de las tareas denominadas “Cookie Thief”, “Fluency” y “Recall”. Para este estudio, nos centramos en la prueba del “Cookie thief” ya que proporciona una prueba estandarizada que ha sido utilizada en varios estudios en el pasado y posee características relevantes para estudios de esta naturaleza [46]. El cuadro consiste en una escena doméstica familiar, mostrada en la Figura 10, cuya descripción requiere el uso de vocabulario básico esencial aprendido en la infancia y el contraste de personajes y lugares.

El objetivo de aplicar la evaluación “Cookie Thief” es reunir información sobre el déficit de comunicación y el déficit de atención existentes en los pacientes. Se considera la forma más eficaz de obtener una de discurso que pueda estandarizarse en muchos sujetos [86].

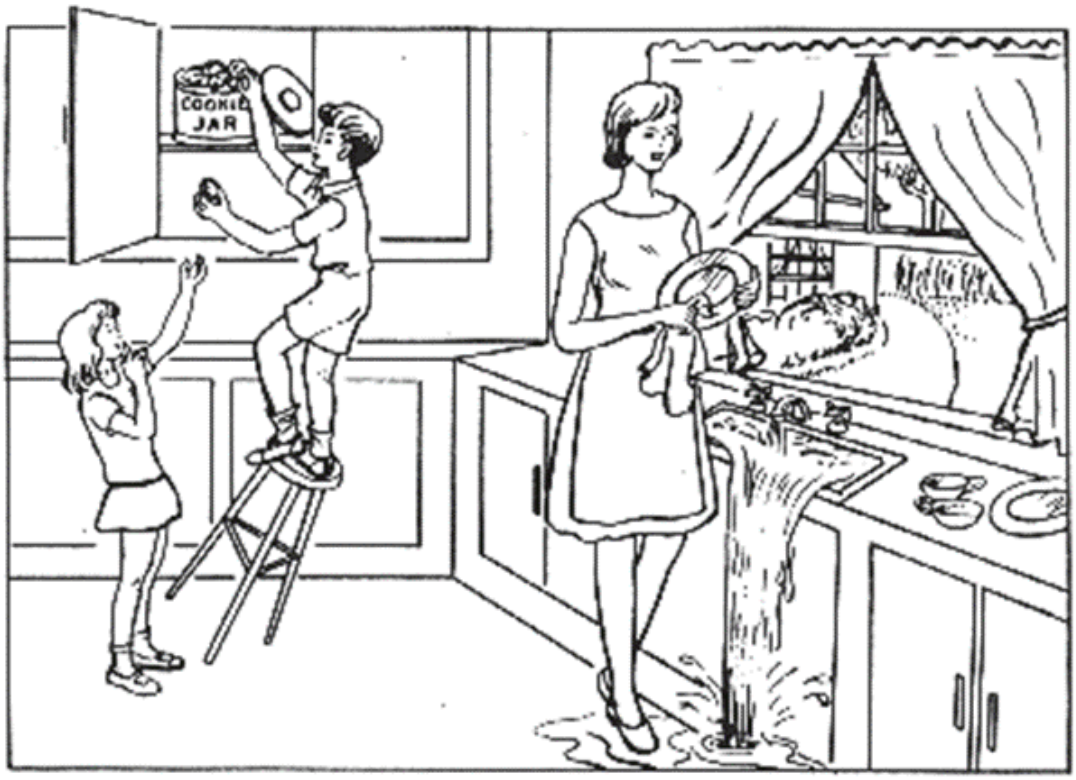

Figura 10. Imagen utilizada en la prueba "Cookie Thief "del Examen diagnóstico de afasia de Boston, fuente [44].

Las ventajas de este instrumento son su enfoque pictórico claro que reduce la ambigüedad sobre el tema, baja demanda de memoria porque el estímulo sigue estando disponible para el sujeto en el momento de la evaluación, minimiza los factores de confusión en el análisis debido a la naturaleza controlada del contenido del discurso; y cuando se utiliza para reevaluar, controla la progresión [87].

3.2 Metodología propuesta

Los métodos más avanzados en la extracción de patrones de demencia han demostrado que el uso de características lingüísticas de tipo sintáctico proporciona una herramienta sensible

y no invasiva para detectar la demencia en su fase inicial [24, 29, 30]. Sin embargo, estos métodos carecen de información semántica relevante. Por lo anterior, este trabajo se enfoca en desarrollar una metodología con la incorporación de características semánticas mediante el uso de incrustaciones de oraciones computadas por redes BERT siamesas (SBERT) junto con Máquina de soporte vectorial (SVM), K-vecinos más cercanos (KNN), Bosques aleatorios y Red neuronal artificial (RNA) como modelos de clasificación. El proceso metodológico propuesto se muestra en la Figura 11.

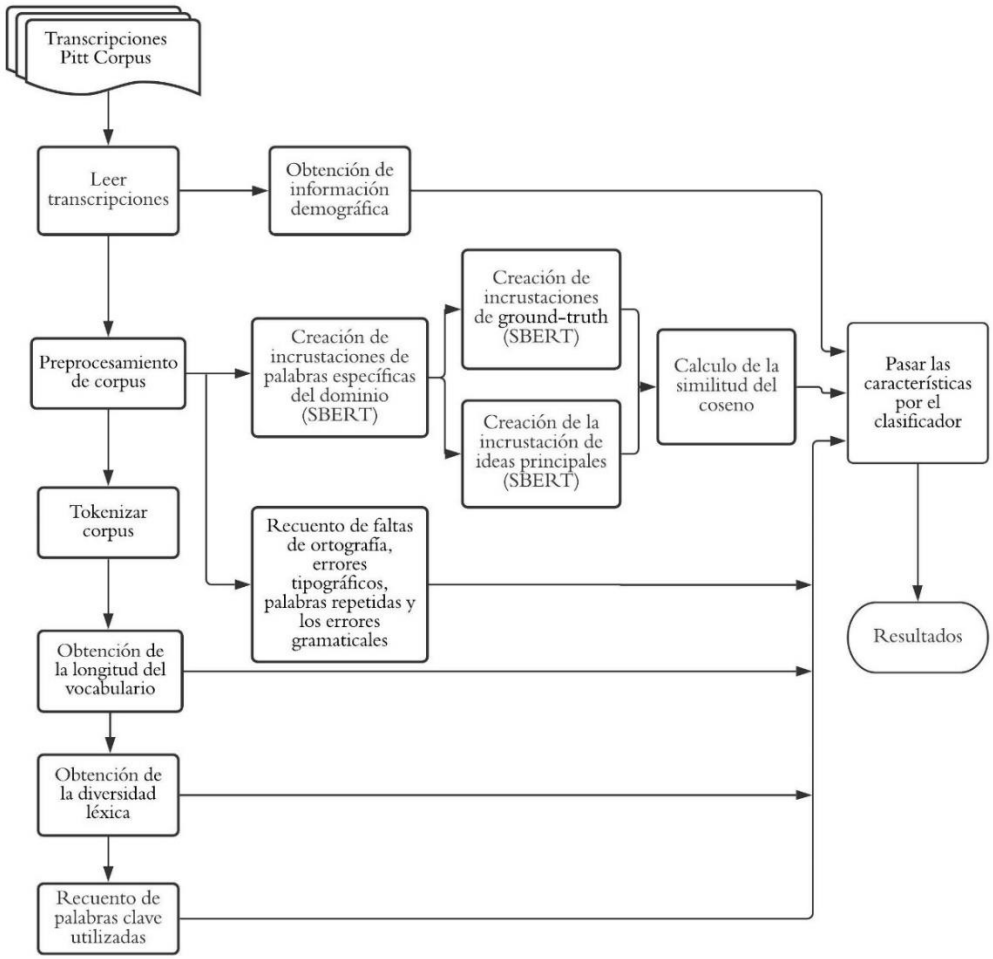

Figura 11. Metodología propuesta.

La metodología propuesta puede subdividirse en tres etapas, preprocesamiento, extracción de características y etapa de clasificación.

El preprocesamiento del texto consiste en extraer el discurso del paciente y convertirlo en texto plano. La extracción de características se realiza en varios niveles, inicia con la captura del número de faltas de ortografía, errores gramaticales, longitud del vocabulario, diversidad léxica, el número de palabras clave utilizadas. De forma complementaria, se incorporan características adicionales de distinta naturaleza, como información demográfica como la edad y los años de educación, además de características obtenidas a partir de Redes Siamesas BERT (SBERT) como parte de la búsqueda semántica de ideas principales, similitud semántica con respecto a ground-truth y el recuento de ideas principales empleadas.

En la etapa de clasificación, se incorporan a la metodología la máquina de soporte de SVM, modelo KNN, Bosques Aleatorios y RNA en una etapa de clasificación de EA. Para medir el rendimiento de nuestro método, aplicamos métricas comúnmente encontradas en la literatura, como la exactitud, la precisión, sensibilidad y la puntuación F1.

3.3 Preprocesamiento

Como se menciona en el apartado *3.1*, la base de datos proporciona un conjunto de muestras de discurso oral, clasificadas en sujetos Control y EA. La grabación de cada discurso oral se transcribió manualmente a nivel de palabra, siguiendo el protocolo de códigos para el análisis de transcripciones humanas por sus siglas en inglés (CHAT) de TalkBank [88]. Estas transcripciones contienen una breve información sobre el sujeto de prueba, donde se alberga su identificador, edad, género, puntaje alcanzado en la prueba MMSE y su diagnóstico. Se transcriben las indicaciones e intervenciones del investigador, así como la descripción proporcionada durante la prueba. En el Anexo 1 se puede encontrar una muestra de estas transcripciones.

El corpus extraído de la transcripción se genera mediante un filtrado, que consiste en extraer únicamente la respuesta del participante. El preprocesamiento aplicado al corpus consiste en eliminar de este texto la simbología CHAT agregada por los investigadores, los signos de puntuación, caracteres especiales y convertir el texto a minúsculas. La Tabla 6 muestra una comparativa del corpus extraído antes y después de su procesamiento.

Tabla 6. Preprocesamiento de Corpus.

Fragmento de Corpus	
Corpus sin preprocesar	['*', 'par', ':', 'mhm', '.', '\x15', '3609_4282', '\x15'] ['*', 'par', ':', "there's", 'a', 'young', 'boy', '&', 'uh', 'going', 'in', 'a', 'cookie', 'jar', '.', '\x15', '5096_10600', '\x15'] ['*', 'par', ':', 'and', "there's", 'a', '[', '/', ']', '&', 'lit', 'a', 'girl', '[', '/', '/', ']', 'young', 'girl', '.', '\x15', '10600_13143', '\x15'] ['*', 'par', ':', 'and', "i'm", 'sayin', '(', 'g', ')', "he's", 'a', 'boy', '(', 'be', ')', 'cause', '<', 'you', 'can', '>', '[', '/', '/', ']', '&', 'hard', "it's"] ['hardly', '[', '/', '/', ']', 'hard', 'to', 'tell', 'anymore', '.', '\x15', '13143_18888', '\x15'] ['*', 'par', ':', '&', 'uh', 'and', "he's", '[', '/', ']', "he's", 'in', 'the', '&', 'c', '&', 't', 'cookie', 'jar', '.', '\x15', '19885_23542', '\x15'] …..
Corpus preprocesado	mhm there s a young boy uh going in a cookie jar and there s a lit a girl young girl and i m sayin g he s a boy be cause you can hard it's hardly hard to tell anymore uh and he s he s in the c t cookie jar and there s a s stool that he is on and it already is starting to fall over and so is the water in the sink uh is ev overflowing in the sink hm i i don't know about the this hickey here i whether that's more than what i said es uh like it uh the wife or g i_mean uh the the mother is near the girl and she s uh w uh h she has uh has oh uh i i can t think of the she has uh the she s tryin g to wipe uh wipe dishes oh a and stop the water from going out

3.4 Extracción de Características

Una vez que el texto plano está listo para ser procesado, se extraen las características lingüísticas descritas anteriormente. Dicha extracción se realiza en los niveles lingüísticos, léxico, sintáctico y semántico. Mediante las técnicas PLN descritas en la sección *2.3*. Al final de la extracción de características, se ha generado un Marco de Datos donde se almacenan todas las propiedades lingüísticas obtenidas en el texto.

3.4.1 Extracción de características léxicas

Como fue mencionado anteriormente un elemento importante de la tarea de descripción de imágenes es la capacidad de recuperar elementos léxicos [46]. En este nivel de análisis se recopilan unidades de información (Figura 12), que son palabras que transmiten con precisión información relevante para el estímulo que se solicita, también se busca cuantificar el discurso, para ello se han calculado cuatro métricas que captan la complejidad léxica del texto:

- Ortografía correcta de la palabra. Cada palabra se compara con un diccionario para corroborar que las palabras están escritas correctamente.
- Errores tipográficos. La búsqueda de errores tipológicos se realizó recurriendo de nuevo al uso de diccionarios.
- Longitud del vocabulario. Se ha realizado la tokenización del texto, operación en la que se realiza una separación de las palabras utilizadas dentro del corpus. A partir de la tokenización de las palabras, es posible realizar un modelo de Bolsa de Palabras en el que se identifican todas las palabras utilizadas a lo largo del texto. Una vez identificadas estas palabras, se cuenta la frecuencia con la que aparecen en el texto. A continuación, es posible medir la longitud del vocabulario (VL).

- Diversidad léxica. Dado que el vocabulario de las personas con demencia en los estadios medio y tardío tiende a ser muy pobre [10], la diversidad léxica (LD) es una métrica importante en este estudio, ya que indica la cantidad de vocabulario utilizado durante la descripción. Es posible calcular la diversidad léxica a partir de la siguiente fórmula, en la que la longitud del texto (LT) se divide por el número de palabras del vocabulario utilizado Longitud del vocabulario (LV):

$$DL = LT/LV \tag{9}$$

Es importante aclarar que el cálculo de los errores ortográficos y mecanográficos no es perfecto, ya que uno de estos errores puede no estar identificado en el diccionario.

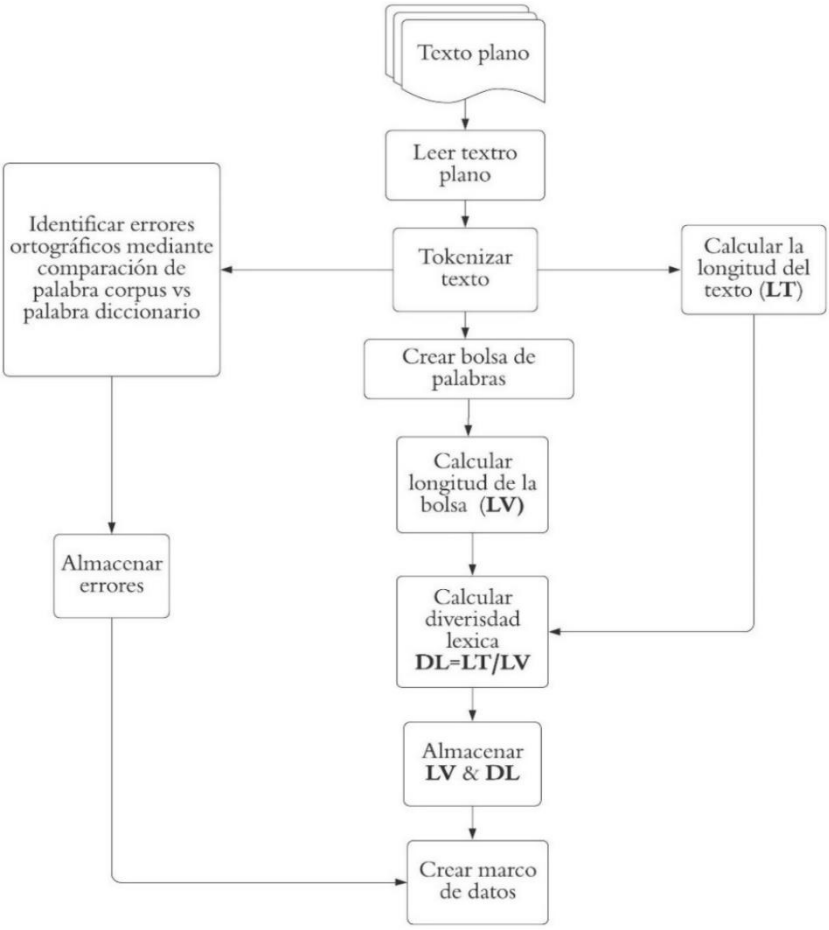

Figura 12. Proceso de extracción de características léxicas.

3.4.2 Extracción de características sintácticas

Es habitual que a los pacientes diagnosticados con EA les resulte muy difícil mantener una conversación debido a sus constantes dificultades para encontrar la palabra adecuada o sustituirla por otra incorrecta e incluso olvidando que ya han dicho esa palabra [34].

Un patrón indicativo de la demencia es la mala estructuración de las oraciones y la tendencia a repetir las palabras mencionadas en un periodo corto, ya que la memoria episódica está en proceso de deterioro. El análisis sintáctico de la producción narrativa de los pacientes es necesario para obtener los errores gramaticales y de estructuración de las frases. Para ello, hacemos uso de la herramienta API Language Tool, basada en el etiquetado de partes del habla (POS-Tagger), donde se asignan partes del habla a cada palabra (y otros tokens), como sustantivo, verbo y adjetivo [37]. Se recogen los errores gramaticales cometidos durante la descripción, así como el número de palabras repetidas en la misma frase.

Algoritmo 3.4.2: Extracción de características sintácticas.

Entada: Texto plano (plain text)
Salida: Núm. De errores gramaticales, núm. de palabras repetidas
tool =LanguageTool('en-US')
Semantic Features (gram_error, w_rep):

```
1.  mstks = [ ]
2.  for i← 1 to length (plain text) by 1 do:
3.      matches = tool.check(text[i])
4.      nu_mis= length (matches)
5.      for j←1 to length (matches) by 1 do:
6.          if nu_mis>0 then:
7.            mstks.append(matches[j])
8.          end
9.      end
10. end
11. grammar=0
12. word_r=0
13. for f←1 to length (mstks) by 1 do:
14.         if mstks[f]==’spelling’:
15.                 grammar += grammar
```

```
16.         elif mstks[f]=='word repetition':
17.                 word_r += word_r
18.          end
19. end
20. return gram_error, w_rep
```

3.4.3 Extracción de características semánticas

Como se ha mencionado anteriormente, investigaciones del estado del arte han trabajado mucho en la identificación de rasgos para denotar afasia en la etapa temprana de EA, centrándose en rasgos como la frecuencia de las palabras, velocidad del habla y complejidad sintáctica. Por otro lado, se ha demostrado que la disfunción en habilidades semánticas puede encontrarse en la etapa temprana de EA caracterizadas por una falta de coherencia semántica y un deterioro de las capacidades de comprensión de información escrita y auditiva [42, 86]. Por lo tanto, para captar los deterioros de la comprensión en el discurso, suponemos que los errores de comprensión suelen dar lugar a respuestas inadecuadas.

Como ya se ha mencionado en la Sección *3.1* la evaluación "Cookie Thief" evalúa los niveles de atención además de las habilidades lingüísticas, requiere que el sujeto de prueba proporcione información importante sobre la imagen estímulo que está siendo mostrada. Hallazgos sobre esta evaluación han concluido que, en comparación con los sujetos de control, los diagnosticados con EA producen descripciones más reducidas y son incapaces de identificar algunas áreas de interés sobre la imagen [18, 86, 46]. La Figura 13 muestra la comparativa de las descripciones realizadas por un individuo con EA y uno participante control (CN) denotando las palabras importantes y el flujo en el que han sido mencionadas durante la descripción [89].

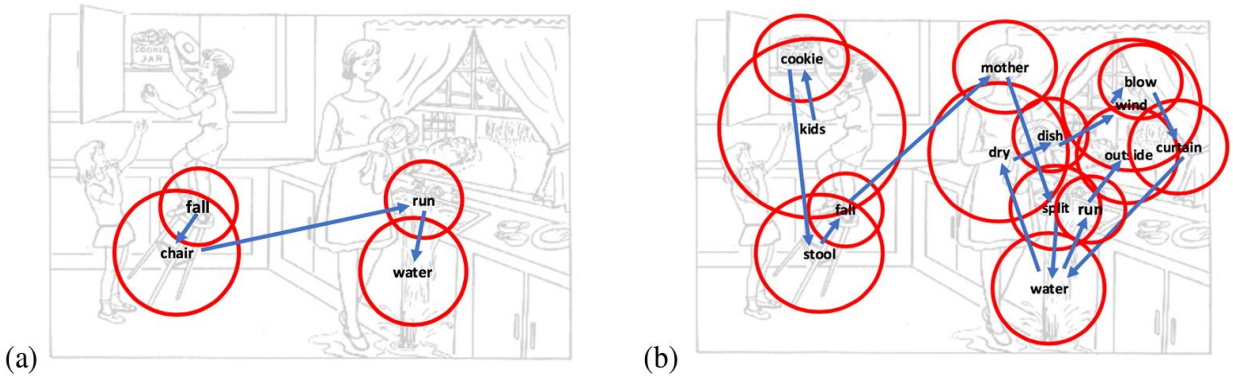

Figura 13. Comparación de observaciones durante la descripción a) Paciente EA en b) Paciente CN, tomado de *[89]*.

En estas evaluaciones también se extrae información puntual que debería formar parte del discurso de los participantes durante la tarea de descripción. Al encontrar estas palabras en las descripciones, suponemos que el paciente se está refiriendo correctamente a la escena de la imagen y no está divagando en otro tema, ya que se ha demostrado una tendencia de los pacientes con EA a presentan cambios de información en la descripción de la imagen y en los conceptos [87]. Las Tablas 7 y 8 describen las palabras clave e ideas principales de la secuencia de acción de la imagen respectivamente.

Tabla 7. Palabras clave a mapear en las transcripciones *[46]*.

PALABRAS CLAVE		
boy	jar	overflowing
girl	plate	spilling
woman	sink	asking
kitchen	stool	unconcerned
window	water	indifferent
cabinet	taking	stealing
cookies	cookie	wobbling
counter	falling	handing
curtain	drying	mother
dishes	washing	sister
faucet	doing	brother
floor		

Tabla 8. Ideas principales extraídas a partir del sistema de clasificación de la coherencia temática basado en *[45]*.

TEMA	ANÁLISIS	IDEA PRINCIPAL
Escena de robo de galletas	Cláusula o frase nominal que identifica una descripción global del contenido de una secuencia de enunciados.	• Children stealing cookies
Actividad de los personajes	Esas secuencias de subtemas que abarcan las actividades realizadas por cada uno de los personajes	• Woman doing dishes • Girl reaching for a cookie • Woman not noticing
Información adicional	Un nivel adicional en la jerarquía de la descripción	• Boy on stool • Sink overflowing • Stool falling

Como parte de la extracción de características semánticas se ha realizado un proceso de embebido de texto, con el objetivo de tener un umbral de búsqueda amplio en las respuestas de los sujetos de prueba y evitar sesgo de la información. Mediante la implementación de SBERT. Al convertir de texto plano en un vector incrustado se añade carga semántica a la representación vectorial del discurso de los sujetos de prueba. Con ello es posible realizar cálculos de similitud semántica, lo que amplía el umbral de búsqueda de los elementos de importancia para la evaluación, mencionados con anterioridad, ventaja sobre la búsqueda textual de estos elementos dentro del discurso.

Para extraer la información a nivel semántico se han implementado 4 técnicas (Figura 13), con las que se recuperan las siguientes características semánticas:

- Búsqueda de palabras clave: Cantidad total de palabras clave (Tabla 7) mapeadas dentro del corpus de la transcripción.
- Cálculo de similitud con respecto a ideas centrales del texto: Para este análisis se hace uso de las ideas principales definidas en la Tabla 8. El método de extracción se efectúa

a partir de segmentar el corpus en frases cortas. Seguido por el proceso de embebido tanto de la idea principal en cuestión como de las oraciones resultantes para finalmente realizar el cálculo de similitud de coseno entre estas. Esto no sólo facilita el cálculo de la similitud, sino que también beneficia al minimizar el sesgo de información dentro del discurso. Como resultado se obtienen 7 características, las cuales contienen el grado de similitud semántica de cada una de las ideas principales mapeadas en la descripción del participante.

- Número de ideas centrales abordadas en la descripción: Tras haber obtenido el grado de similitud semántica entre las ideas principales y las oraciones que componen la descripción, es posible observar que en ocasiones la similitud alcanzada está por debajo del 50%, debido a que la oración toma algunos elementos de la idea principal sin embargo no refiere sustancialmente a esta. Por tanto, se considera que la idea principal no ha sido realmente abordada. Para el recuento de ideas principales abordadas en el texto, se considera sólo aquellas que superan el 50% de grado de similitud.
- Cálculo de similitud respecto a la descripción determinada como ground-truth: Para tener una visión general de lo descrito, también se ha realizado un cálculo de similitud de coseno entre el corpus de la descripción y el gorund-truth (Anexo 2). La magnitud de la similitud alcanzada nos da una aproximación del rendimiento del sujeto de prueba en la tarea de descripción.

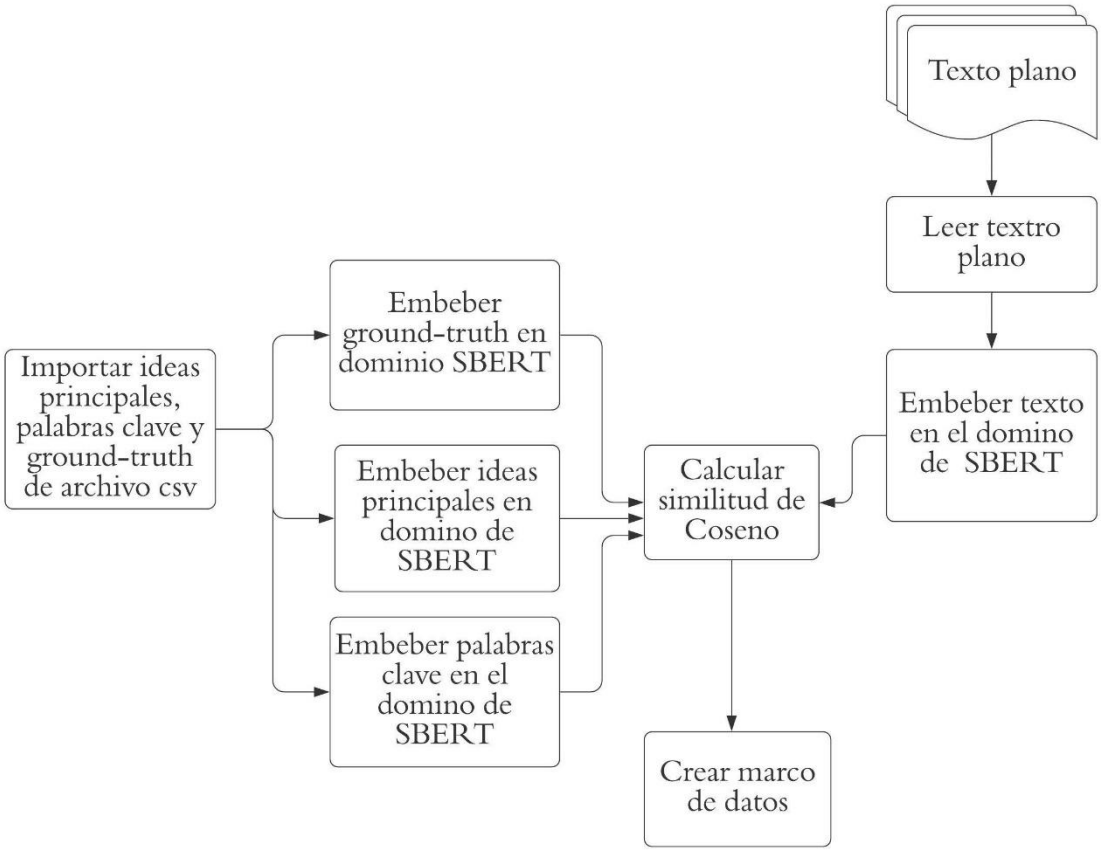

Figura 14. Proceso de extracción de características semánticas.

3.5 Selección de características

Uno de los principales objetivos de esta investigación, además de extraer rasgos lingüísticos de las evaluaciones neuropsicológicas es cuantificar la utilidad de estos rasgos en la tarea de clasificación de la demencia. Para ello, se ha aplicado el método de correlación para evaluar la asociación lineal bidireccional entre dos variables. Se ha centrado en la aplicación del método de correlación de Spearman [90]. Debido al tipo de variables utilizadas, en este caso, la correlación buscada es la existente entre la puntuación del test MMSE y el diagnóstico de demencia. La puntuación del MMSE proporciona una escala ampliamente aceptada para estimar la gravedad de la demencia de 0 a 30, basada en una serie de preguntas relacionadas con la orientación, el registro, la atención, la falta de memoria y el lenguaje [1]. Una puntuación de 20 a 24 sugiere demencia leve, de 13 a 20 sugiere demencia moderada y menos de 12 indica demencia grave [91]. El cálculo del coeficiente de correlación se ha obtenido utilizando el paquete estadístico Pingouin [92], los valores resultantes son un coeficiente de correlación de -0,789 una significación de 2,26561e-118, y una potencia estadística igual a 1.

La puntuación del MMSE y la denominación de demencia presentan una fuerte correlación negativa, con alta potencia estadística. Esto significa que cuanto mayor sea la puntuación obtenida en la prueba, el paciente puede ser descartado como alguien con deterioro cognitivo o diagnosticado de demencia. En ese sentido, podemos inferir que la puntuación del MMSE es indicativa de demencia. Esto nos ayudará a determinar la influencia de las características lingüísticas extraídas del texto en la clasificación de la demencia. Para la parametrización de esta influencia, se ha aplicado el método del Coeficiente de Información Mutua [82].

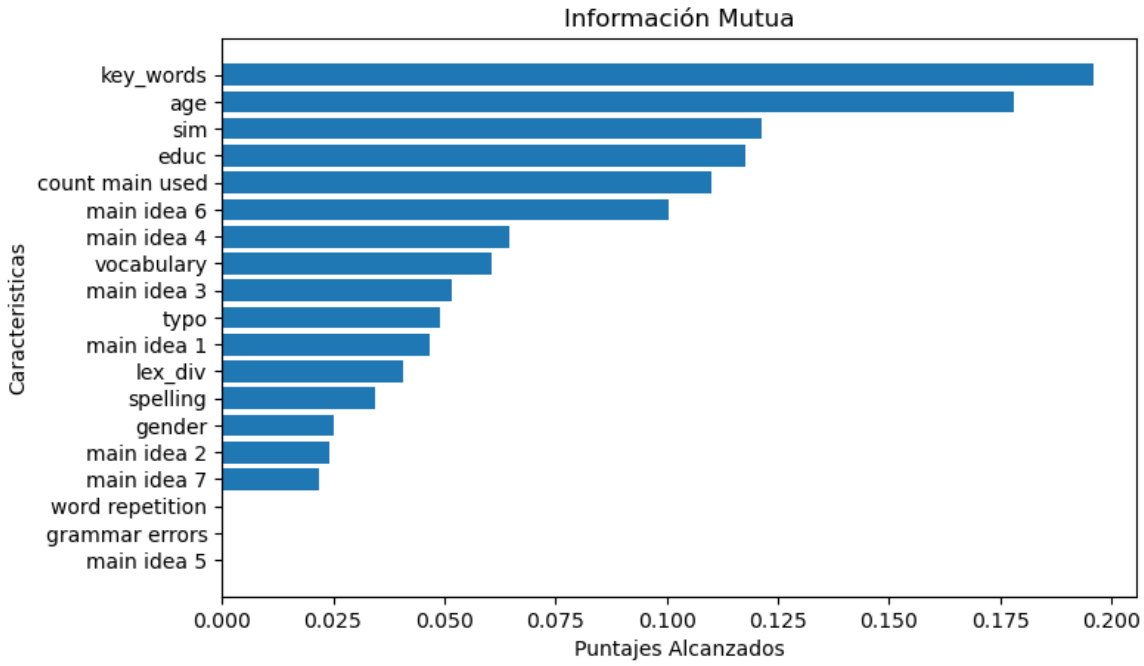

Figura 15. Histograma de la puntuación obtenida por cada una de las características relativas a la puntuación MMSE en el análisis de información mutua.

Las puntuaciones obtenidas en el método de información mutua muestran que la característica más relacionada es la denominada palabras clave extraída de las características semánticas (Figura 15). Los datos con mayor relevancia para la denominación de la demencia son la edad y la escolaridad, así como, algunas de las características semánticas. Esto reafirma la idea de utilizar este tipo de características aplicadas al modelo de clasificación.

3.6 Clasificadores automáticos

La tarea de clasificación automática de personas con y sin demencia se ha realizado aplicando algoritmos de aprendizaje automático supervisado. Los modelos de implementación han sido descritos en la sección *2.4*. Como se ve en la Tabla 9, se han utilizado seis modelos del paquete SKlearn Python [93]. Además, de los modelos de ML mencionados anteriormente,

se realizó una clasificación por medio de un modelo de Red Neuronal Artificial de la librería tensorflow [94].

3.6.1 Selección de hiper parámetros

Para definir los hiper parámetros del clasificador, se tuvieron en cuenta diferentes aspectos. En el caso de K-vecinos más cercanos, para definir el número de K vecinos, se implementó el método del codo [95]. Mientras que, para el Bosques Aleatorios, se realizó una validación cruzada. Por otro lado, para la Máquina de soporte vectorial, se utilizaron kernels estándar para mapear las observaciones. Para encontrar la mejor configuración para la Red Neuronal Artificial (RNA) se ha utilizado la librería Autokeras [96].

Tabla 9. Modelos y sus hiper parámetros.

Modelo	**Hiper parámetros**
KNN	K = 30
Bosques aleatorios	Árboles = 100, profundidad máxima = 6
SVM Lineal	kernel = 'lineal'
SVM Polinomio	núcleo = 'poly', grado=2
SVM precalculado	kernel = 'precomputed'
SVM RBF	kernel = 'rbf'
Red neuronal artificial	Véase la figura 15

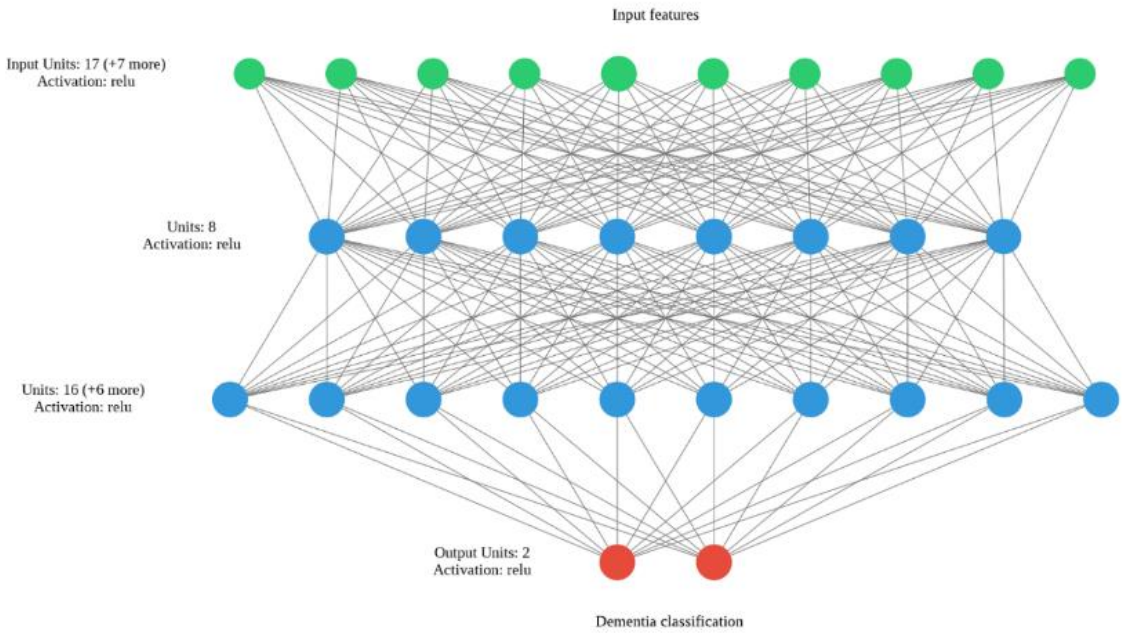

Figura 16. Modelo Red Neuronal Artificial implementado.

El modelo de red neuronal implementado (Figura 16), se compone por una capa de entrada de 17 nodos seguido de una capa densa oculta de 8 nodos, una capa densa de 16 nodos, y finalmente una capa de salida de 2 nodos. Cada capa hace uso de una función de activación Rectificada Lineal (ReLU) [97].

$$f(x) = \max(0, x) = \begin{cases} 0 \rightarrow x < 0 \\ 1 \rightarrow x \geq 0 \end{cases} \qquad (10)$$

Y como salida no retribuye la clasificación binaria.

CAPÍTULO 4

Resultados

4.1 Métricas de evaluación

Para obtener una visión precisa del rendimiento de nuestra metodología propuesta, se siguieron dos estrategias. La primera, fue el uso de métricas comúnmente utilizadas en estudios similares: métricas de exactitud, precisión, recuerdo y puntuación F1 [98]. Estas métricas se definen en términos de los valores estimados correctamente, denominados verdaderos positivos (TP) y verdaderos negativos (TN). Así como, las predicciones incorrectas, denominadas falsos positivos (FP) y falsos negativos (FN) [99].

- Exactitud: Representa el porcentaje de predicciones correctas respecto al total. En nuestro contexto, la Exactitud se refiere a la relación entre las predicciones correctas de pacientes EA y todos los casos analizados.

$$Exactitud = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \quad (11)$$

- Precisión: También llamada tasa de verdaderos positivos. En este caso, la precisión nos da la proporción entre los pacientes clasificados como pacientes con demencia según nuestro método y los pacientes que realmente tienen demencia.

$$Precisión = \frac{TP}{TP+FP} \times 100 \quad (12)$$

- Sensibilidad: El porcentaje de casos de demencia positivos que se identificaron correctamente.

$$Sensibilidad = \frac{TP}{TP+FN} \times 100 \quad (10)$$

- Puntuación F1: Es la media armónica entre sensibilidad y precisión. Esta métrica elimina el sesgo causado por tener datos desequilibrados, como en nuestro conjunto de datos (Tabla 1).

$$Puntuación\ F1 = \frac{2}{\frac{1}{Presicion} \times \frac{1}{Sensibilidad}} \times 100 \quad (13)$$

4.2 Metodología de evaluación

Dado que la cardinalidad del conjunto de datos es reducida (550). La segunda para disminuir el posible sesgo en la medición del rendimiento debido es el uso de la validación cruzada k-fold. Para cada modelo de clasificación, los datos se dividieron en 10 pliegues, dividiendo los datos en entrenamiento y prueba. El método cruzado de 10 pliegues se iteró 10 veces. Las tablas 10 y 11 muestran las estadísticas de cada métrica por algoritmo de clasificación.

La Tabla 10 es el resultado de la clasificación realizada exclusivamente con características léxicas y sintácticas, compuesto por la cantidad de vocabulario utilizado, la diversidad léxica, los errores ortográficos, número de palabras repetidas, añadiendo la edad y años de escolaridad. Estas características se utilizan en [20 - 29]. En esta experimentación, se alcanzó una precisión media del 71% utilizando KNN, siendo el valor medio de clasificación más bajo y el más alto del 75% cuando se utilizó SVM polinomial y pre computada. El valor medio de precisión alcanzado fue del 68% -80%, siendo KNN el de mayor porcentaje.

Tabla 10. Media de los resultados obtenidos durante la etapa de prueba utilizando únicamente rasgos léxico-sintácticos.

Modelo	Exactitud	Precisión	Sensibilidad	Puntuación F1	Tiempo (seg.)
Bosque aleatorio	0.74	0.78	0.74	0.76	**0.01**
KNN	0.71	**0.80**	0.63	0.71	0.04

SVM Lineal	0.74	0.78	**0.76**	**0.77**	**0.01**
SVM Polinomio	**0.75**	0.79	**0.76**	**0.77**	**0.01**
SVM precalculado	0.74	0.78	**0.76**	**0.77**	**0.01**
SVM RBF	0.75	0.79	0.75	0.76	**0.01**
Red neuronal artificial	0.72	0.68	0.72	0.70	5.35

Nuestro objetivo es añadir indicadores semánticos para reforzar el modelo. La Tabla 11 muestra el rendimiento de los algoritmos de clasificación cuando se añaden al modelo las características semánticas extraídas, que en este caso son la búsqueda de palabras clave en el texto, el cálculo de la similitud del coseno con las ideas principales propuestas en la Tabla 8, y el cálculo de la similitud con respecto a la descripción considerada nuestra ground-truth. Los porcentajes alcanzados en las métricas de evaluación aumentaron, llegando a una precisión media del 78% cuando se utilizó la RNA, sin embargo, la SVM polinómica alcanzó mayores porcentajes en el resto de las métricas.

Tabla 11. Media de los resultados medios obtenidos durante la etapa de prueba utilizando características léxicas, sintácticas y semánticas.

Modelo	Exactitud	Precisión	Sensibilidad	Puntuación de la F1	Tiempo (seg.)
Bosque aleatorio	0.76	0.79	0.78	0.78	**0.01**
KNN	0.74	**0.81**	0.70	0.75	0.04
SVM Lineal	0.77	0.80	0.79	0.79	**0.01**
SVM Polinomio	0.77	0.80	**0.80**	**0.80**	**0.01**
SVM precalculado	0.77	0.80	0.79	0.79	**0.01**
SVM RBF	0.77	**0.81**	0.78	0.79	**0.01**
Red neuronal	**0.78**	0.73	0.79	0.76	8.43

Las Figuras 17-20 muestran el comportamiento estadístico de los resultados obtenidos por cada uno de los modelos implementados, correspondientes a cada métrica considerada en este estudio.

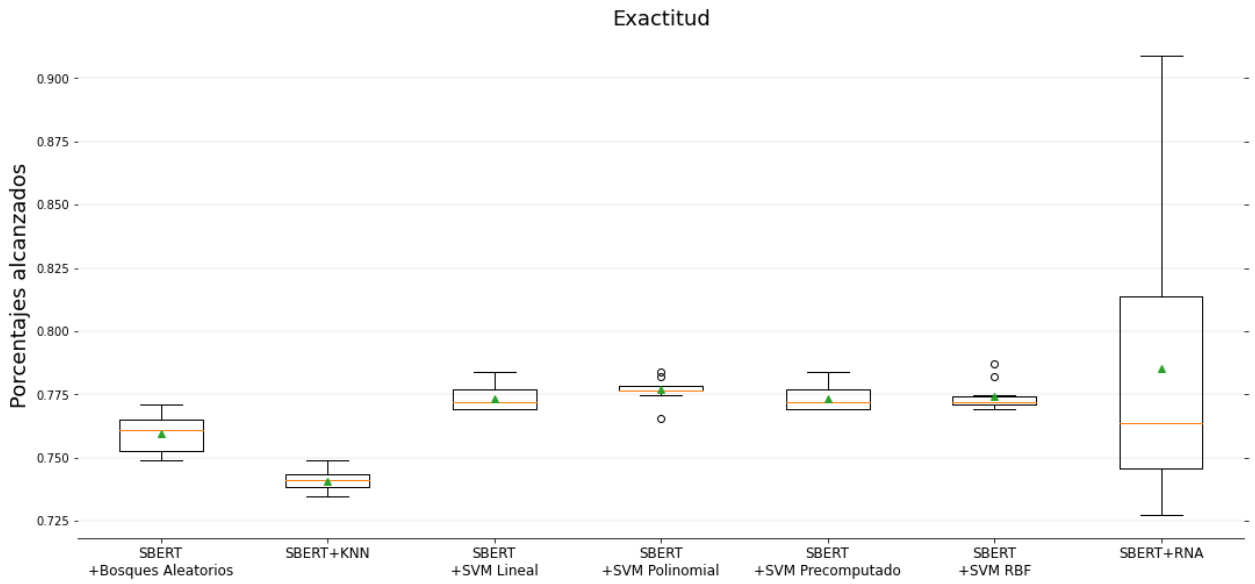

Figura 17. Estadísticas de clasificadores métrica Exactitud.

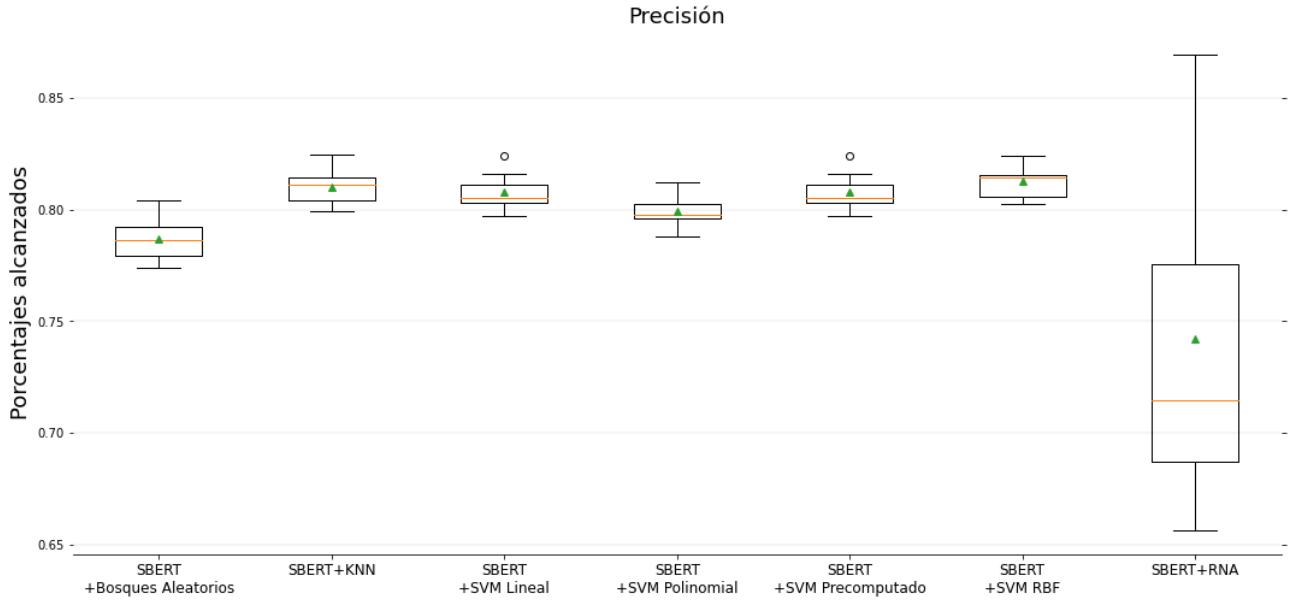

Figura 18. Estadísticas de clasificadores métrica Precisión.

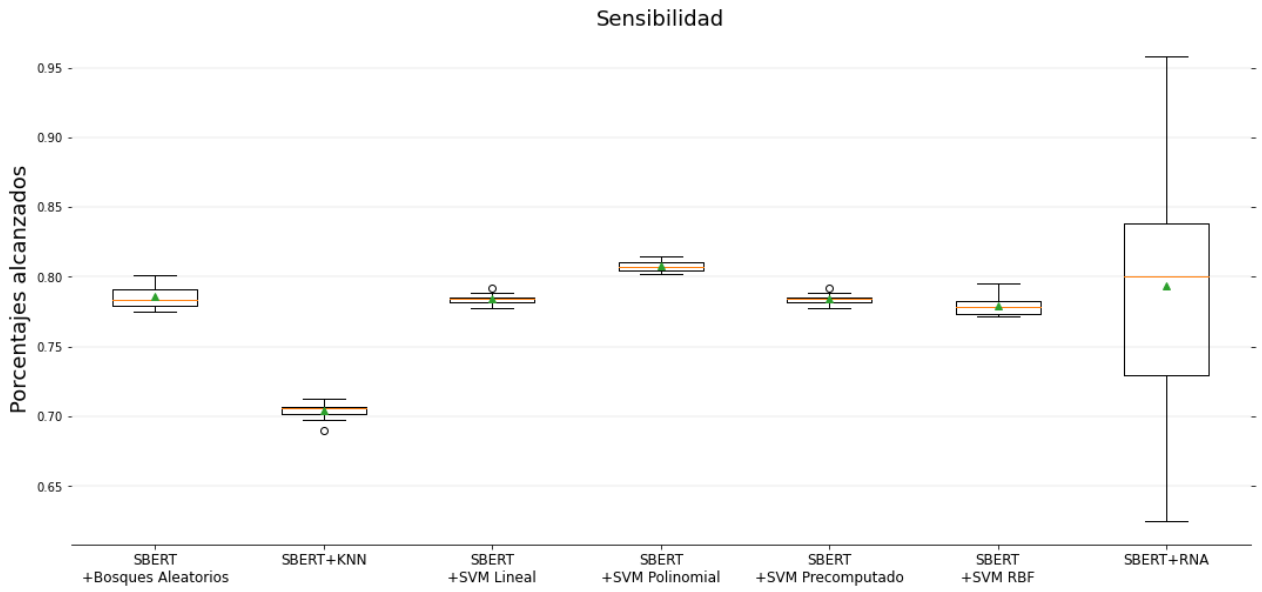

Figura 19. Estadísticas de clasificadores métrica Sensibilidad.

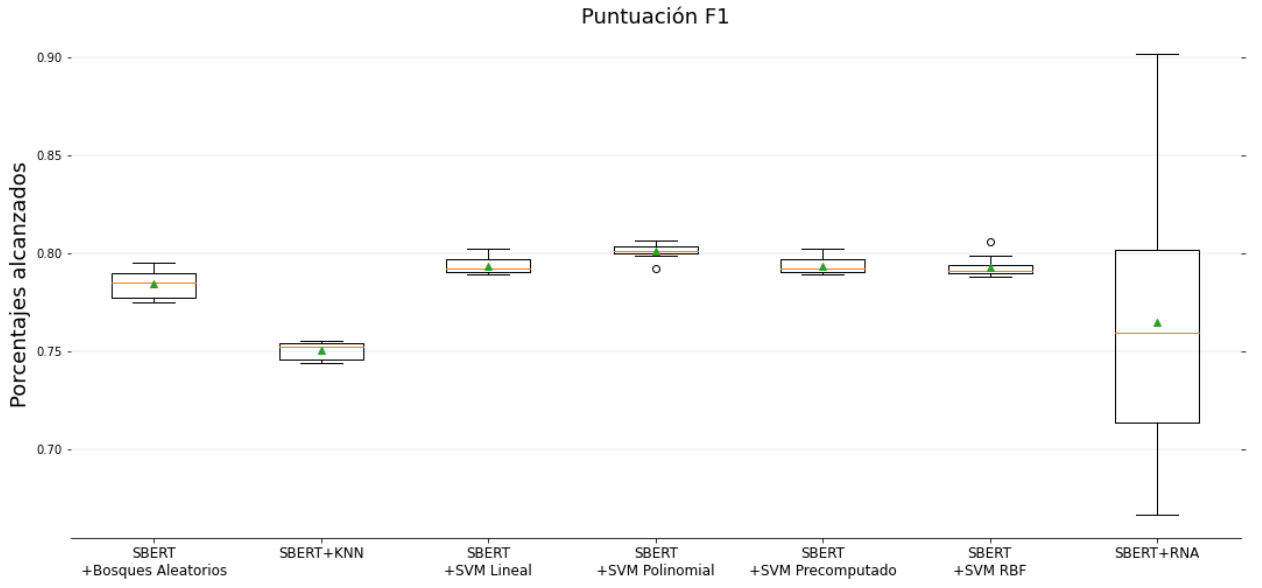

Figura 20. Estadísticas de clasificadores métrica Puntuación F1

A lo largo de la experimentación es posible observar que los modelos de clasificación que mantienen un comportamiento normal son aquellos basados en el método SVM con sus diferentes kernels. Manteniendo un valor medio de exactitud 77%, y un valor máximo de 79% en el caso específico del modelo con kernel polinomial. El modelo con menor desempeño con base a esta métrica es KNN alcanzando un valor medio de 74% y un valor máximo de 75%. Por otro lado, el modelo basado en RNA destaca por obtener un valor mayor al 90% de exactitud en algunas de las pruebas K-fold, sin embargo, los valores alcanzados a lo largo de las pruebas son inconstantes, fenómeno que se repite en las métricas.

Respecto a la métrica de precisión los modelos de clasificación exceptuando RNA mantienen valores de 80% en sus valores medios y como máximo 82%, siendo el caso de Bosques Aleatorios. El desempeño de los clasificadores dentro de la métrica de sensibilidad varía entre ellos, sin embargo, sus distribuciones tienden a ser normales a excepción del modelo RNA como ya se ha mencionado antes.

El puntaje F1 mantiene valores medios cercanos al 80% en la mayoría de los casos, a diferencia de los modelos KNN y RNA. Debido a las fluctuaciones que presentan durante la experimentación en las métricas de sensibilidad y precisión.

Para comparar el rendimiento de nuestra metodología propuesta con un método del estado del arte [35] en términos de tasa de precisión, describimos a través de las 10 iteraciones, el rendimiento de la clasificación en la Figura 21. Es importante mencionar que la prueba BERT se realizó en las mismas condiciones de datos, excepto por el uso de transcripciones manuales y por disponer únicamente del corpus incrustado.

La RNA muestra el mejor rendimiento en muchos k-fold. Sin embargo, la variabilidad de su tasa de clasificación advierte sobre su fiabilidad. Mientras que los clasificadores SVM mantienen una tasa de clasificación normal, con una tasa de clasificación media del 77%, lo que indica que en términos generales es el clasificador más viable para esta tarea.

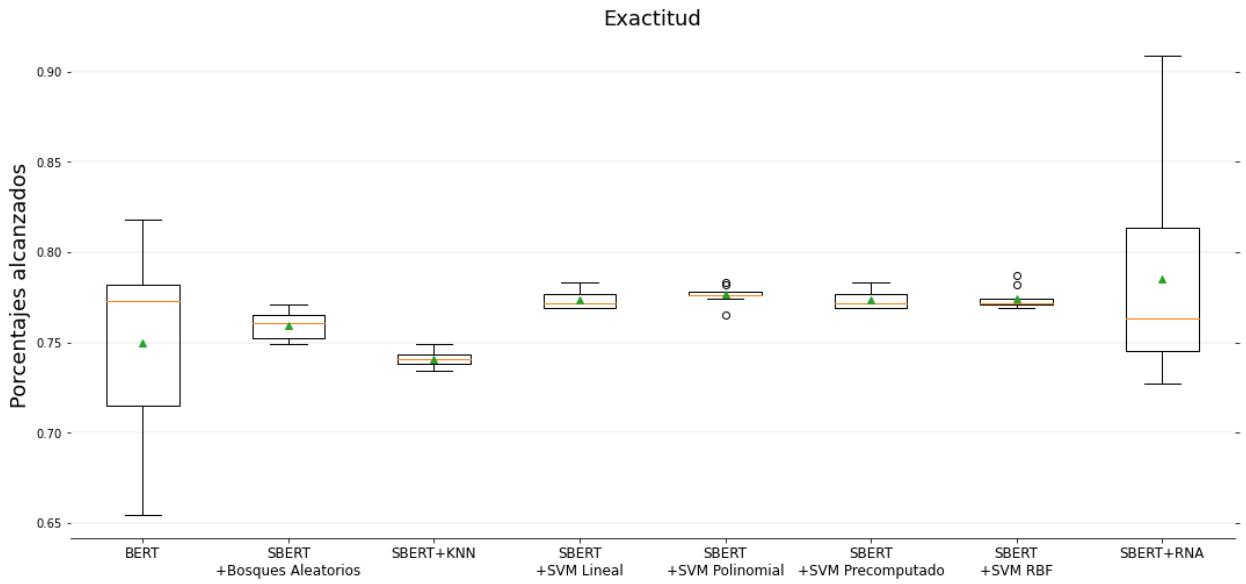

Figura 21. Comparación estadística modelos implementados vs BERT

4.3 Discusión

El enfoque dado a este trabajo de investigación se inspiró en las investigaciones lingüísticas realizadas por especialistas en esta área clínica [32, 46, 87]. A lo largo de la realización de este trabajo, nos encontramos con desafíos en la obtención de las características propuestas debido a la prevalencia de datos dispersos y ruidosos. Las técnicas de preprocesamiento redujeron el ruido; sin embargo, se requiere una combinación de técnicas para eliminarlo aún más. La implementación de técnicas de PNL en la base de datos aumenta el espectro de búsqueda de patrones y características. Sin embargo, sigue existiendo un factor de error difícil de determinar.

Por otro lado, el uso de modelos preentrenados minimiza la carga computacional necesaria para realizar el proceso de incrustación del texto, proporcionando una interpretación semántica del mismo debido a la arquitectura implementada en estos modelos y a la riqueza de los datos con los que está preentrenado.

En cuanto a la extracción de rasgos semánticos, que es nuestro objeto de interés, se ha calculado el grado de utilidad de los rasgos. La aplicación del método de información mutua revela un grado de dependencia entre las características propuestas y la puntuación MMSE. Siendo los rasgos correspondientes al uso de palabras clave, la edad, la similitud de la descripción con respecto al ground-truth y el grado académico, los que tienen una mayor puntuación en la evaluación. Con esto, podemos inferir que la información semántica y demográfica tiene un mayor peso en el modelo de clasificación. En este punto, es importante señalar que, al experimentar con una selección diferente de características, se ha demostrado que el uso de características lingüísticas léxicas y sintácticas tiene un rendimiento menor que el realizado al añadir características semánticas. A partir de los experimentos, el método de clasificación con mejores puntuaciones en las métricas fue el modelo SVM polinomial de segundo orden y el modelo RNA, alcanzando una precisión media del 77% y del 78%.

Un diagnóstico erróneo se traduce directamente en una pérdida de tiempo para detener la progresión de esta enfermedad. Es entonces importante mantener un bajo número de falsos positivos sobre los falsos negativos para la validación. Por ello, se ha utilizado el resultado obtenido en la puntuación F1, se ha obtenido una media del 80% para esta métrica.

CAPÍTULO 5

Conclusiones

Los rasgos lingüísticos, como se postula en las investigaciones aquí mencionadas, son un elemento que refleja un déficit cognitivo, específicamente la afasia. Es evidente que no es posible utilizar este aspecto como determinante en la clasificación clínica de la demencia. Sin embargo, puede utilizarse como prueba de apoyo para descartar una probable demencia, especialmente en sus primeras etapas. Podemos argumentar que un modelo básico para la denominación de la demencia podría prescindir de la puntuación del MMSE, si se utilizan características léxicas, sintácticas, semánticas y demográficas. Los resultados obtenidos de nuestra prueba experimental, han demostrado que el enfoque SBERT a través de la similitud del coseno es adecuado para generar características semánticas. Además, se ha demostrado que la configuración mediante el uso de SBERT+SVM se comportó mejor en cada métrica que el resto de los modelos, a excepción de la configuración SBERT+RNA cuya tasa de precisión es mejor en algunos de los k-fold. A través de la comparación entre el modelo léxico-sintáctico y el modelo semántico-sintáctico, se ha comprobado que la implementación de características semánticas al modelo aumenta la tasa de métrica en aproximadamente un 3%.

Si bien la metodología propuesta ha tenido un rendimiento competente para la discriminación de la demencia, seguir buscando características lingüísticas o clínicas más relevantes podría reflejarse en la mejora del rendimiento de la tarea. Se ha demostrado que los rasgos semánticos mejoran el rendimiento de los modelos de clasificación debido al enfoque con el que se desarrolla la evaluación diagnóstica.

5.1 Trabajos Futuros

La Demencia es una enfermedad con un alto impacto en nuestra sociedad, no solo desde un enfoque clínico, sino también, económico y social. Se continuará con la búsqueda de herramientas que puedan contribuir al diagnóstico temprano de este padecimiento. En este sentido, investigaciones emergentes sugieren emplear indicadores como biomarcadores a modelos de clasificación de demencia. Se buscarán otros rasgos sintácticos y semánticos que nos proporcionen información sobre la memoria semántica de los sujetos de prueba y la implementación de un modelo de ensamble.

Referencias

[1] T. Widiger, “Diagnostic and statistical manual of mental disorders (DSM),” in *Psychology*, Oxford University Press, 2011. Accessed: Apr. 25, 2022. [Online]. Available:http://dx.doi.org/10.1093/obo/9780199828340-0022

[2] J. González-Hernández and T. Ramos F., “RELACIÓN MÉDICO-PACIENTE EN EL CONTEXTO DE LA DEMENCIA,” *Revista Médica Clínica Las Condes*, vol. 27, no. 3, pp. 357–362, May 2016, doi: 10.1016/j.rmclc.2016.06.009.

[3] R. A. G. Victoriano and A. Hornauer-Hughes, “Afasia: Una perspectiva clínica,” *unknown*,Dec.01,2014.https://www.researchgate.net/publication/318659697_Afasia_una_perspectiva_clinica

[4] J. Peña-Casanova, *Neurología de la conducta y neuropsicología*. Ed. Médica Panamericana,2007.

[5] E. B. Inc, *Britannica Enciclopedia Moderna*. Encyclopaedia Britannica, Inc., 2011.

[6] Alzheimer’s Association, “500,” *Alzheimer’s Disease and Dementia*, 2021. https://www.alz.org/alzheimer-demencia. (accessed Apr. 25, 2022).

[7] A. C. Gracia-Rebled *et al.*, “El efecto de la ocupación laboral en la incidencia de demencia vascular: un estudio de cohortes de 12 años de seguimiento,” *Revista de Psiquiatría y Salud Mental*, Jul. 2020, doi: 10.1016/j.rpsm.2020.05.002.

[8] P. A. Roa Rojas, A. Martínez Ruiz, and M. del C. García Peña, *La Enfermedad de Alzheimer y otras demencias como problema nacional de salud*. Intersistemas S.A de C.V, 2017, pp. 1–33.

[9] J. Huang, “Demencia,” *Manuales MSD*, Dec. 04, 2019. https://www.msdmanuals.com/es/professional/trastornosneurol%C3%B3gicos/delirio-y-demencia/demencia#v1036599_es.

[10] F. M. Tapia, W. L. Castro, C. M. Poblete, C. M. Soza, and Array, “Stigma towards mental disorders: characteristics and interventions,” *Salud Mental*, vol. 38, no. 1, pp. 53–58, Feb. 2015, doi: 10.17711/SM.0185-3325.2015.007.

[11] C. Zurique Sánchez *et al.*, “Prevalencia de demencia en adultos mayores de América Latina: revisión sistemática,” *Revista Española de Geriatría y Gerontología*, vol. 54, no. 6, pp. 346–355, Nov. 2019, doi: 10.1016/j.regg.2018.12.007.

[12] J. L. Barranco Quintana, M. F. Allam, A. Serrano del Castillo, and R. Fernández-Crehuet Navajas, “Factores de riesgo de la enfermedad de Alzheimer,” *Revista de Neurología*, vol. 40, no. 10, p. 613, 2005, doi: 10.33588/rn.4010.2004360.

[13] M. Dashwood, G. Churchhouse, M. Young, and T. Kuruvilla, “Artificial intelligence as an aid to diagnosing dementia: an overview,” *Progress in Neurology and Psychiatry*, vol. 25, no. 3, pp. 42–47, Jul. 2021, doi: 10.1002/pnp.721.

[14] P. Sawyer, A. Sutcliffe, P. Rayson, and C. Bull, “Dementia and Social Sustainability: Challenges for Software Engineering,” May 2015. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1109/icse.2015.188

[15] A. D. International, “World Alzheimer Report 2015, The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends”.

[16] M. Arthur, “Institute for Health Metrics and Evaluation,” *Nursing Standard*, vol. 28, no. 42, pp. 32–32, Jun. 2014, doi: 10.7748/ns.28.42.32.s33.

[17] S. Lpez-Pousa and J. Garre-Olmo, “Demencia. Concepto. Clasificación. Epidemiología. Aspectos socioeconómicos,” *Medicine - Programa de Formación Médica Continuada Acreditado*, vol. 9, no. 77, pp. 4921–4927, Apr. 2007, doi: 10.1016/s0211-3449(07)75473-5.

[18] W. W. Hung, J. S. Ross, K. S. Boockvar, and A. L. Siu, “Recent trends in chronic disease, impairment and disability among older adults in the United States,” *BMC Geriatrics*, vol. 11, no. 1, Aug. 2011, doi: 10.1186/1471-2318-11-47.

[19] T. Álvarez Cisneros, S. Torres Castro, B. Mena Montes, and N. M. Torres Carrillo, “Genero y Salud en Cifras,” vol. 15, no. 3, Feb. 2020.

[20] “Demencias, una Visión panorámica : temas derivados del V Simposio de Medicina Geriátrica realizado el 5 y 6 de septiembre del 2014 San Luis Potosí, S.L.P.,” 2015.

[21] M. Dashwood, G. Churchhouse, M. Young, and T. Kuruvilla, “Artificial intelligence as an aid to diagnosing dementia: an overview,” *Progress in Neurology and Psychiatry*, vol. 25, no. 3, pp. 42–47, Jul. 2021, doi: 10.1002/pnp.721.

[22] C. Brierley, “AI could detect dementia years before symptoms appear,” *University of Cambridge*, Aug. 11, 2021. https://www.cam.ac.uk/stories/AIdementia

[23] V. Masrani, G. Murray, T. Field, and G. Carenini, “Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia,” 2017. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.18653/v1/w17-2329

[24] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011, doi:10.1109/tasl.2011.2112351.

[25] L. Cleret de Langavant, E. Bayen, and K. Yaffe, "Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study," *Journal of Medical Internet Research*, vol. 20, no. 7, p. e10493, Jul. 2018, doi:10.2196/10493.

[26] K.-S. Na, "Prediction of future cognitive impairment among the community elderly: A machine-learning based approach," *Scientific Reports*, vol. 9, no. 1, Mar. 2019, doi:10.1038/s41598-019-39478-7.

[27] L. A. Jennings, A. Hackbarth, N. Wenger, Z. Tan, and D. B. Reuben, "AN AUTOMATED APPROACH TO IDENTIFYING PATIENTS WITH DEMENTIA USING ELECTRONIC MEDICAL RECORDS," *Innovation in Aging*, vol. 1, no. suppl_1, pp. 1381–1382, Jun. 2017, doi: 10.1093/geroni/igx004.5084.

[28] V. S. Nori, C. A. Hane, D. C. Martin, A. D. Kravetz, and D. M. Sanghavi, "Identifying incident dementia by applying machine learning to a very large administrative claims dataset," *PLOS ONE*, vol. 14, no. 7, p. e0203246, Jul. 2019, doi:10.1371/journal.pone.0203246.

[29] S. Karlekar, T. Niu, and M. Bansal, "Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models," 2018. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.18653/v1/n18-2110

[30] D. S. Rosas, S. T. Arriaga, and M. A. A. Fernandez, "Search for Dementia Patterns in Transcribed Conversations using Natural Language Processing," Sep. 2019. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1109/iceee.2019.8884572

[31] Madical University of South Carolina, "Carolinas Conversations Collection ," *Mission*, 2009. https://carolinaconversations.musc.edu/ccc/about/ (accessed Apr. 25, 2022).

[32] E. Eyigoz, S. Mathur, M. Santamaria, G. Cecchi, and M. Naylor, "Linguistic markers predict onset of Alzheimer's disease," *EClinicalMedicine*, vol. 28, p. 100583, Oct. 2020, doi:10.1016/j.eclinm.2020.100583.

[33] T. R. Dawber, G. F. Meadors, and F. E. Moore Jr., "Epidemiological Approaches to Heart Disease: The Framingham Study," *American Journal of Public Health and the Nations Health*, vol. 41, no. 3, pp. 279–286, Mar. 1951, doi: 10.2105/ajph.41.3.279.

[34] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, “Multimodal Inductive Transfer Learning for Detection of Alzheimer’s Dementia and its Severity,” Oct. 2020. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2020-3137

[35] M. Gonzalez-Atienza, A. M. Peinado, and J. A. Gonzalez-Lopez, “An Automatic System for Dementia Detection using Acoustic and Linguistic Features,” Mar. 2021. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.21437/iberspeech.2021-56

[36] D. Mukherji, M. Mukherji, and N. Mukherji, “Early Detection of Alzheimer’s Disease with Low-Cost Neuropsychological Tests: A Novel Predict-Diagnose Approach using Recurrent Neural Networks,” Cold Spring Harbor Laboratory, Jan. 2021. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1101/2021.01.17.21249822

[37] N. Linz, J. Troger, J. Alexandersson, M. Wolters, A. Konig, and P. Robert, “Predicting Dementia Screening and Staging Scores from Semantic Verbal Fluency Performance,” Nov. 2017. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1109/icdmw.2017.100

[38] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge,” Oct. 2020. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2020-2571

[39] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, “To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer’s Disease Detection,” Oct. 2020. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2020-2557

[40] J. Bhatta, D. Shrestha, S. Nepal, S. Pandey, and S. Koirala, “Efficient Estimation of Nepali Word Representations in Vector Space,” *Journal of Innovations in Engineering Education*, vol. 3, no. 1, pp. 71–77, Mar. 2020, doi: 10.3126/jiee.v3i1.34327.

[41] A. Budhkar and F. Rudzicz, “Augmenting word2vec with latent Dirichlet allocation within a clinical application,” *ACL Anthology*. https://aclanthology.org/N19-1414/

[42] F. Cuetos, J. Rodríguez-Ferreiro, and M. Menéndez, “Semantic Markers in the Diagnosis of Neurodegenerative Dementias,” *Dementia and Geriatric Cognitive Disorders*, vol. 28, no. 3, pp. 267–274, 2009, doi: 10.1159/000242438.

[43] J. Hernandez, “Demencias: los problemas de lenguaje como hallazgos tempranos,” *Acta Neurológica Colombiana*.

[44] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, “Speaking in Alzheimer’s Disease, is That an Early Sign? Importance of Changes in Language Abilities in Alzheimer’s Disease,” *Frontiers in Aging Neuroscience*, vol. 7, Oct. 2015,doi:10.3389/fnagi.2015.00195.

[45] J. O. de Lira, T. S. C. Minett, P. H. F. Bertolucci, and K. Z. Ortiz, “Analysis of word number and content in discourse of patients with mild to moderate Alzheimer’s disease,” *Dementia & Neuropsychologia*, vol. 8, no. 3, pp. 260–265,Sep.2014,doi:10.1590/s1980-57642014dn83000010.

[46] C. Mackenzie, M. Brady, J. Norrie, and N. Poedjianto, “Picture description in neurologically normal adults: Concepts and topic coherence,” *Aphasiology*, vol. 21, no. 3–4, pp. 340–354, Mar. 2007, doi:10.1080/02687030600911419.

[47] M. Ibrahim and R. Ahmad, “Class Diagram Extraction from Textual Requirements Using Natural Language Processing (NLP) Techniques,” 2010. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1109/iccrd.2010.71

[48] R. Ismail, Z. Abu Bakar, and N. Abd. Rahman, “EXTRACTING KNOWLEDGE FROM ENGLISH TRANSLATED QURAN USING NLP PATTERN,” *Jurnal Teknologi*, vol. 77, no. 19, Nov. 2015, doi: 10.11113/jt.v77.6515.

[49] J. Zhang and N. M. El-Gohary, “Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking,” *Journal of Computing in Civil Engineering*, vol. 30, no. 2, p. 04015014, Mar. 2016, doi:10.1061/(asce)cp.1943-5487.0000346.

[50] J. M. Guerrero, “chomsky y la gramatica generativa.pdf,” *Scribd*. https://es.scribd.com/document/363357772/chomsky-y-la-gramatica-generativa-pdf

[51] A. C. Vásquez, H. V. huerta, J. P. Quispe, and A. M. Huayna, “Procesamiento de lenguaje natural,” *Revista de investigación de Sistemas e Informática*, vol. 6, no. 2, pp. 45–54, Dec. 2009, doi: 5923.

[52] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*. CRC Press,2010.

[53] Dr. (smt). M. V. Reddy, “Semantical and syntactical analysis for NLP,” *Dr. (smt). Mallamma V Reddy - Academia.edu*, May 21, 2014. https://www.academia.edu/7117928/Semantical_and_syntactical_analysis_for_NLP

[54] Eduardo Sosa, "Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones (Parte I)," *El profesional de la información*. http://profesionaldelainformacion.com/contenidos/1997/enero/procesamiento_del_lenguaje_natural_revisin_del_estado_actual_bases_tericas_y_aplicaciones_parte_i.html

[55] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Information Sciences*, vol. 307, pp. 39–52, Jun. 2015, doi: 10.1016/j.ins.2015.02.024.

[56] J. Han, M. Kamber, and J. Pei, "Getting to Know Your Data," in *Data Mining*, Elsevier, 2012, pp. 39–82. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1016/b978-0-12-381479-1.00002-2

[57] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. PrenticeHall,2009.

[58] S. Adilov, "Generative Pre-Training from Molecules," *ChemRxiv*. https://chemrxiv.org/engage/chemrxiv/article-details/6142f60742198e8c31782e9e

[59] U. Kamath, K. L. Graham, and W. Emara, "Bidirectional Encoder Representations from Transformers (BERT)," in *Transformers for Machine Learning*, Boca Raton: Chapman and Hall/CRC, 2022, pp. 43–70. Accessed: Apr. 25, 2022. [Online]. Available:http://dx.doi.org/10.1201/9781003170082-3

[60] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," 2019. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.18653/v1/d19-1410

[61] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[62] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press, 2019.

[63] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann, 2014.

[64] L. J. Krajewski and L. P. Ritzman, *Administración de operaciones: estrategia y análisis ; incluye CD*. Pearson Educación, 2000.

[65] R. Gove, "Using the elbow method to determine the optimal number of clusters for k-means clustering," *bl.ocks.org*. https://bl.ocks.org/rpgove/0060ff3b656618e9136b (accessed Apr. 25, 2022).

[66] R. Manuel José, R. Tinguaro J., C. Rafael, and G. Daniel, *Big data para científicos sociales. Una introducción*. CIS, 2020.

[67] Data sience team, "DATA SCIENCE," *Data Science*. https://datascience.eu/es/. (accessed Apr. 25, 2022).

[68] S. F. Fugazzi, *ABC Economipedia*. Lulu.com, 2015.

[69] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering*, vol. 2, no. 1, pp. 602–609, Oct. 2014, doi: 10.1080/21642583.2014.956265.

[70] M. Koning and C. Smith, *Decision Trees and Random Forests: A Visual Introduction for Beginners*. Independently Published, 2017.

[71] C. Schönbach, K. Nakai, S. Ranganathan, and M. A. Khan, *Encyclopedia of Bioinformatics and Computational Biology: Applications / Mohammad Asif Khan (Centre for Bioinformatics, Perdana University, Selangor, Malaysia)*. 2018.

[72] N. Cristianini, J. Shawe-Taylor, and D. of C. S. R. H. J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[73] S. Maldonado and R. Weber, "MODELOS DE SELECCIÓN DE ATRIBUTOS PARA SVMS," *Congreso Latino-Iberoamericano de Investigación Operativa*, Sep. 2012.

[74] E. Valveny, J. G. Sabaté, and R. B. Caselles, "Support Vector Machines (SVM): Desarrollo matemático - Bag of Words (BoW)," *Coursera*. https://www.coursera.org/lecture/clasificacion-imagenes/support-vector-machines-svm-desarrollo-matematico-ztrcv (accessed Apr. 25, 2022).

[75] Á. F. Godoy Viera, "Técnicas de aprendizaje de máquina utilizadas para la minería de texto," *Investigación Bibliotecológica. Archivonomía, Bibliotecología e Información*, vol. 31, no. 71, p. 103, Mar. 2017, doi: 10.22201/iibi.0187358xp.2017.71.57812.

[76] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media, 2013.

[77] S. Shalev-Shwartz and S. Ben-David, "Neural Networks," in *Understanding Machine Learning*, Cambridge: Cambridge University Press, pp. 228–242. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1017/cbo9781107298019.021

[78] J. P. Lévy Mangin, “Las redes neuronales artificiales: Aspectos generales,” in *Las redes neuronales artificiales. Fundamentos teoricos y aplicaciones practicas*, Netbiblo, pp. 16–45. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.4272/978-84-9745-246-5.ch2

[79] Univerity of Arizona , “13.2. Fine-Tuning — Dive into Deep Learning 0.17.5 documentation.” https://d2l.ai/chapter_computer-vision/fine-tuning.html (accessed Apr. 25, 2022).

[80] M. Chapman-Rounds, U. Bhatt, E. Pazos, M.-A. Schulz, and K. Georgatzis, “FIMAP: Feature Importance by Minimal Adversarial Perturbation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, pp. 11433–11441, May 2021, doi: 17362.

[81] P. Duboue, “Features, Reduced: Feature Selection, Dimensionality Reduction and Embeddings,” in *The Art of Feature Engineering*, Cambridge University Press, 2020, pp. 79–111. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1017/9781108671682.006

[82] B. E. Boyle, “Feature Selection Using Mutual Information,” in *Computer Oriented Learning Processes*, Dordrecht: Springer Netherlands, 1976, pp. 287–297. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-94-010-1545-5_14

[83] “Teoría de la información de Claude E. Shannon,” *DIA*. http://dia.austral.edu.ar/Teor%C3%ADa_de_la_informaci%C3%B3n_de_Claude_E._Shannon

[84] DementiaBank consortium group, “DementiaBank.” https://dementia.talkbank.org/. (accessed Apr. 25, 2022).

[85] J. T. Becker, “The Natural History of Alzheimer’s Disease,” *Archives of Neurology*, vol. 51, no. 6, p. 585, Jun. 1994, doi: 10.1001/archneur.1994.00540180063015.

[86] E. Giles, K. Patterson, and J. R. Hodges, “Performance on the Boston Cookie theft picture description task in patients with early dementia of the Alzheimer’s type: Missing information,” *Aphasiology*, vol. 10, no. 4, pp. 395–408, May 1996, doi: 10.1080/02687039608248419.

[87] J. O. de Lira, T. S. C. Minett, P. H. F. Bertolucci, and K. Z. Ortiz, “Analysis of word number and content in discourse of patients with mild to moderate Alzheimer’s disease,” *Dementia & Neuropsychologia*, vol. 8, no. 3, pp. 260–265, Sep. 2014, doi: 10.1590/s1980-57642014dn83000010.

[88] B. MacWhinney, “References,” in *The Childes Project*, Psychology Press, 2014, pp. 207–210. Accessed: Apr. 25, 2022. [Online]. Available:

http://dx.doi.org/10.4324/9781315805672-25

[89] B. Mirheidari, Y. Pan, T. S. Walker, and H. Christensen, “Detecting Alzheimer’s Disease by estimating attention and elicitation path through the alignment of spoken...,” *unknown*, Oct. 01, 2019. https://www.researchgate.net/publication/336208488_Detecting_Alzheimer’s_Disease_by_estimating_attention_and_elicitation_path_through_the_alignment_of_spoken_picture_descriptions_with_the_picture_prompt

[90] Y. Dodge, “Spearman Rank Correlation Coefficient,” in *SpringerReference*, Berlin/Heidelberg: Springer-Verlag. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1007/springerreference_221490

[91] A. J. Mitchell, “The Mini-Mental State Examination (MMSE): Update on Its Diagnostic Accuracy and Clinical Utility for Cognitive Disorders,” in *Cognitive Screening Instruments*, Cham: Springer International Publishing, 2017, pp. 37–48. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-44775-9_3

[92] R. Vallat, “Pingouin: statistics in Python,” *Journal of Open Source Software*, vol. 3, no. 31, p. 1026, Nov. 2018, doi: 10.21105/joss.01026.

[93] A. Pajankar and A. Joshi, “Introduction to Machine Learning with Scikit-learn,” in *Hands-on Machine Learning with Python*, Berkeley, CA: Apress, 2022, pp. 65–77. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-1-4842-7921-2_5

[94] Tensorflow company, “API Documentation,” *TensorFlow*. https://www.tensorflow.org/api_docs (accessed Apr. 25, 2022).

[95] H. Humaira and R. Rasyidah, “Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm,” 2020. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.4108/eai.24-1-2018.2292388

[96] H. Jin, Q. Song, and X. Hu, “Auto-Keras: An Efficient Neural Architecture Search System,” Jul. 2019. Accessed: Apr. 26, 2022. [Online]. Available: http://dx.doi.org/10.1145/3292500.3330648

[97] R. Valbuena, *Inteligencia Artificial: Investigación Científica Avanzada Centrada en Datos*. ROIMAN VALBUENA, 2021.

[98] M. A. A. Fernández, *Inteligencia artificial para programadores con prisa*. Universo de Letras, 2022.

[99] J. Zou, Y. Han, and S.-S. So, “Overview of Artificial Neural Networks,” in *Methods in Molecular BiologyTM*, Totowa, NJ: Humana Press, 2008, pp. 14–22. Accessed: Apr. 25, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-1-60327-101-1_2

[100] T. A. CISNEROS, S. T. CASTRO, B. M. MONTES, and N. M. T. CARRILLO, “Alzheimer: Diferencias por género entre América Latina y otras regiones del mundo,” Jan. 01, 2017. http://repositorio.inger.gob.mx/jspui/handle/20.500.12100/17213 (accessed Apr. 25, 2022).

ANEXOS

Anexo 1

Muestra de Transcripción Manual de Evaluación “Cookie Thief”

NOTAS SOBRE EL ARCHIVO DE CHAT

Formato de identificación del material de transcripción

eng|Pitt|PAR|edad|género|Dx|Participante|MMSE|

eng- Lenguaje en el que se desarrolla la entrevista

Pitt- Titulo de la investigación en la que se produjo la base de datos

PAR- Código de identificación

Dx- Diagnostico (CX= control, DX = Demencia)

MMSE- puntaje obtenido en la evaluación MMSE

Simbología

Identificador del orador: *PAR – Participante, * INV – Investigador

%mor = descomposición morfológica de la frase

%gram = descomposición gramatical de la frase

“@” = adition letters

“word(word)word” = palabras incompletas

“[/]” = pausas

“[: text]” = palabra remplazada

“[x N]” = silabas repetidas

“·0_1073·” = tiempo de alineación

Tabla 12 . Transcripción extraída del conjunto de Datos DementiaBank [84].

```
@UTF8
@PID:    11312/t-00002426-1
@Begin
@Languages:    eng
@Participants:    PAR Participant, INV Investigator
@ID:    eng|Pitt|PAR|75;|female|ProbableAD||Participant|15||
@ID:    eng|Pitt|INV|||||Investigator|||
@Media: 007-3, audio
*INV:    on in the picture . 0_2865
%mor:    adv|on prep|in det:art|the n|picture .
%gra:    1|0|INCROOT 2|1|JCT 3|4|DET 4|2|POBJ 5|1|PUNCT
*PAR:    well the girl is telling the boy to get the cookies down but don't tell your mother . 2865_8887
%mor:    co|well det:art|the n|girl aux|be&3S part|tell-PRESP det:art|the n|boy inf|to v|get det:art|the n|cookie-PL adv|down conj|but
         mod|do~neg|not v|tell det:poss|your n|mother .
%gra:    1|5|COM 2|3|DET 3|5|SUBJ 4|5|AUX 5|0|ROOT 6|7|DET 7|5|OBJ 8|9|INF
         9|7|XMOD 10|11|DET 11|9|OBJ 12|9|JCT 13|5|CONJ 14|16|AUX 15|14|NEG
         16|13|COORD 17|18|DET 18|16|OBJ 19|5|PUNCT
*PAR:    and the boy is also falling over off the stool . 8887_13642
%mor:    coord|and det:art|the n|boy aux|be&3S adv|also part|fall-PRESP adv|over prep|off det:art|the n|stool .
%gra:    1|6|LINK 2|3|DET 3|6|SUBJ 4|6|AUX 5|6|JCT 6|0|ROOT 7|6|JCT 8|6|JCT
         9|10|DET 10|8|POBJ 11|6|PUNCT
*PAR:    and the mother is letting the water run out_of the sink . 13642_17948
%mor:    coord|and det:art|the n|mother aux|be&3S part|let-PRESP det:art|the n|water v|run prep|out_of det:art|the n|sink .
%gra:    1|5|LINK 2|3|DET 3|5|SUBJ 4|5|AUX 5|0|ROOT 6|7|DET 7|8|SUBJ 8|5|COMP
         9|8|JCT 10|11|DET 11|9|POBJ 12|5|PUNCT
*PAR:    and she's dryin(g) dishes . 17948_20100
%mor:    coord|and pro:sub|she~aux|be&3S part|dry-PRESP n|dish-PL .
%gra:    1|4|LINK 2|4|SUBJ 3|4|AUX 4|0|ROOT 5|4|OBJ 6|4|PUNCT
*PAR:    I don't quite get that but then &=laughs +... [+ exc] 20100_23860
%mor:    pro:sub|I mod|do~neg|not adv|quite v|get adv|that conj|but adv:tem|then +...
%gra:    1|5|SUBJ 2|5|AUX 3|2|NEG 4|5|JCT 5|0|ROOT 6|5|JCT 7|5|CONJ 8|7|COORD
         9|5|PUNCT
*PAR:    &uh she has water on the floor and [/] and basically it's kind_of &uh a distressing scene . 23860_35314
%mor:    pro:sub|she aux|have&3S n|water prep|on det:art|the n|floor
         coord|and adv|basic&dadj-AL-LY pro:per|it~cop|be&3S adv|kind_of
         det:art|a n:gerund|distress-PRESP n|scene .
%gra:    1|3|SUBJ 2|3|AUX 3|0|INCROOT 4|3|NJCT 5|6|DET 6|4|POBJ 7|10|LINK
         8|10|JCT 9|10|SUBJ 10|3|CJCT 11|10|JCT 12|14|DET 13|14|MOD 14|11|POBJ
         15|3|PUNCT
*PAR:    everything's goin(g) haywire . 35314_37790
%mor:    pro:indef|everything~aux|be&3S part|go-PRESP adj|haywire .
%gra:    1|3|SUBJ 2|3|AUX 3|0|ROOT 4|3|JCT 5|3|PUNCT
*INV:    okay . 37790_39719
%mor:    co|okay .
%gra:    1|0|INCROOT 2|1|PUNCT
*PAR:    she needs to turn off the water . 39719_44091
%mor:    pro:sub|she v|need-3S inf|to v|turn prep|off det:art|the n|water .
%gra:    1|2|SUBJ 2|0|ROOT 3|4|INF 4|2|COMP 5|4|JCT 6|7|DET 7|5|POBJ 8|2|PUNCT
*PAR:    if she turned off the water she'd be a hundred percent better off .44091_47843
%mor:    conj|if pro:sub|she v|turn-PAST prep|off det:art|the n|water
         pro:sub|she~mod|genmod cop|be det:art|a det:num|hundred n|percent
         adv|good&CP prep|off .
%gra:    1|3|LINK 2|3|SUBJ 3|9|CJCT 4|3|JCT 5|6|DET 6|4|POBJ 7|9|SUBJ 8|9|AUX
         9|0|ROOT 10|12|DET 11|12|QUANT 12|9|PRED 13|14|JCT 14|12|NJCT 15|9|PUNCT
*INV:    okay . 47843_50116
%mor:    co|okay .
%gra:    1|0|INCROOT 2|1|PUNCT
*INV:    okay . 50116_53247
%mor:    co|okay .
%gra:    1|0|INCROOT 2|1|PUNCT
@End
```

Anexo 2

Descripción “Ground-truth”

Formato: Texto Plano

“a mother who is drying dishes next to the sink in the kitchen she is not paying attention and has left the tap on as a result water is overflowing from the sink meanwhile two children are attempting to take cookies from a jar when their mother is not looking one of the children a boy has climbed onto a stool to get up to the cupboard where the cookie jar is stored the stool is rocking precariously the other child a girl is standing next to the stool and has her hand outstretched ready to be given cookies”

Anexo 3

Certificado “Formación en línea sobre la protección de los participantes en la investigación en seres humanos”

Certificado necesario para obtener acceso a la base de datos.

CERTIFICATE

OF COMPLETION

PHRP Online Training, Inc. certifies that

Santander Yamanki

has successfully completed the web-based course "Protecting Human Research Participants Online Training."

Date Completed: **2020-10-03** Certification Number: **2857764**

PHRP
Protecting Human
Research Participants