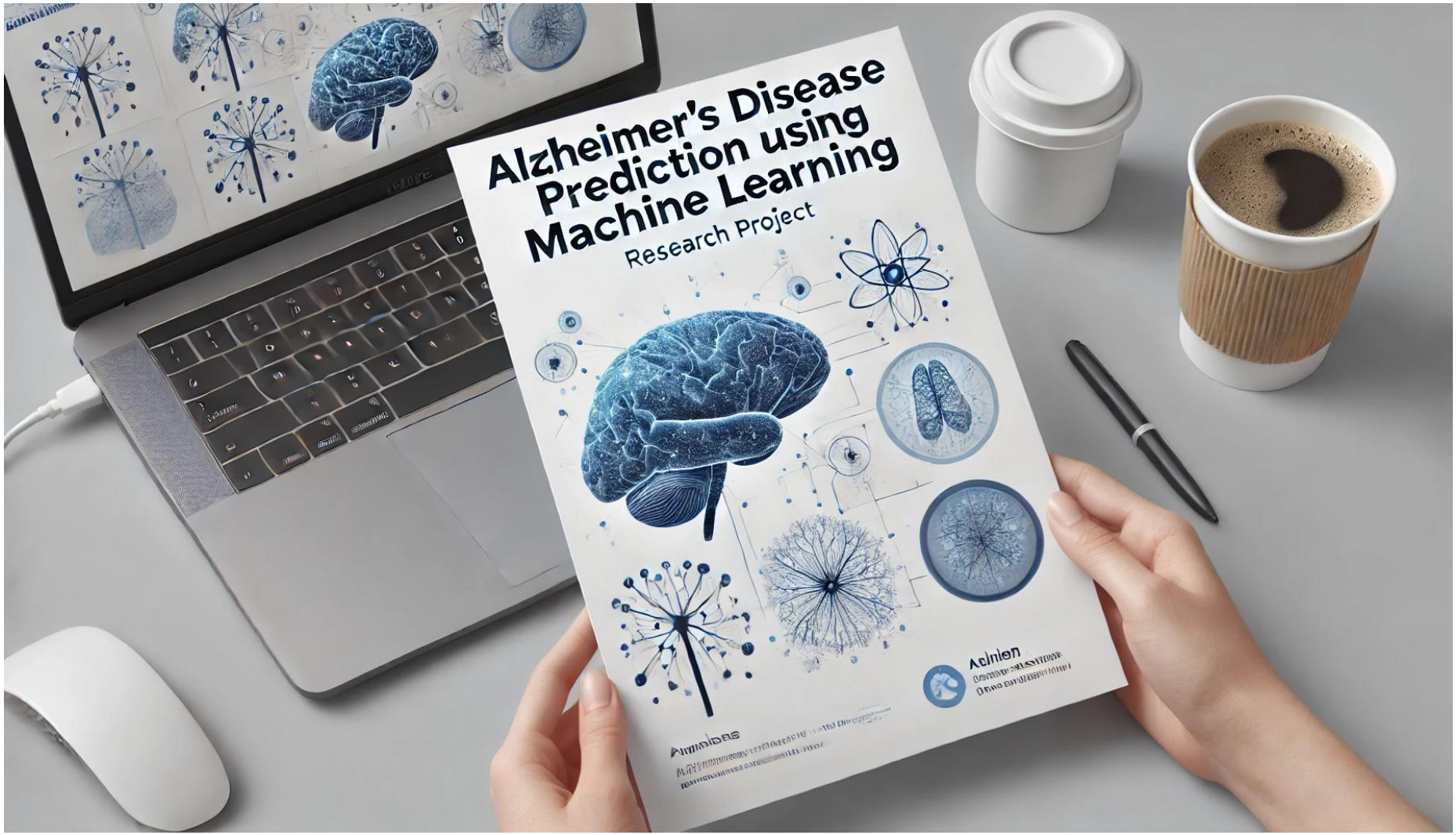
Alzheimer's Disease Prediction using Machine Learning
Research Project

Por María Elisa Cisneros

Introducción

El presente proyecto se centra en la predicción de la enfermedad de Alzheimer utilizando un modelo de aprendizaje automático y el conjunto de datos proporcionado en el repositorio de Kaggle titulado **"Alzheimer's Disease Prediction"**.

La enfermedad de Alzheimer es un trastorno neurodegenerativo progresivo que afecta principalmente a personas mayores,

causando deterioro cognitivo y pérdida de memoria. Dada su creciente prevalencia global y el impacto significativo que tiene tanto en los pacientes como en sus familias, resulta esencial desarrollar herramientas tecnológicas que permitan identificar esta enfermedad de manera temprana.

El tema fue elegido debido a la importancia clínica y social de abordar el Alzheimer, así como a la necesidad de soluciones

tecnológicas innovadoras que apoyen la detección precoz y la gestión de la enfermedad.

El tema es relevante, pues según los datos en españa:

- **Alta prevalencia:** La enfermedad de Alzheimer afecta a aproximadamente **800,000 personas** en España, representando cerca del **60-70%** de los casos de demencia en el país.

- **Envejecimiento poblacional**: España tiene una de las poblaciones más envejecidas de Europa, lo que incrementa significativamente el riesgo de enfermedades neurodegenerativas. Se estima que el número de casos podría **duplicarse para 2050** debido al envejecimiento demográfico.
- **Impacto económico:** Los costes asociados al Alzheimer, incluyendo atención médica, cuidados domiciliarios y pérdida

de productividad, **superan los 24,000 millones de euros anuales en España.**

- **Carga familiar:** Más del **80%** de los cuidadores son familiares, lo que genera una carga emocional y económica importante en los hogares.

Además, este campo representa un **desafío** interesante en el ámbito del aprendizaje automático, dado que combina el análisis

de grandes volúmenes de datos clínicos con la interpretación de resultados complejos para la toma de decisiones médicas.

Objetivos del Proyecto:

-Diseñar y entrenar un modelo predictivo utilizando datos relacionados con la enfermedad de Alzheimer.

-Identificar los factores clave que influyen en la predicción de esta enfermedad.

-Evaluar el rendimiento del modelo propuesto utilizando métricas como la precisión, la sensibilidad y la especificidad.

-Proporcionar una herramienta que pueda servir como referencia inicial para investigaciones futuras en el ámbito de la salud y la inteligencia artificial.

Alcance del Trabajo:

Este proyecto se enfocará en el desarrollo de un **modelo de clasificación basado en datos estructurados** provenientes del conjunto de datos de **Kaggle.**

La información contenida en el dataset incluye variables clínicas y biométricas que se utilizarán para predecir el estado del Alzheimer.

El **alcance** incluye la limpieza de los datos, el análisis exploratorio, la selección de características relevantes, el entrenamiento y evaluación del modelo, y la interpretación de los resultados.

El trabajo **no abordará** aspectos como la implementación clínica directa o la validación del modelo en entornos reales, ya que estos requieren de colaboraciones interdisciplinarias y un nivel de certificación fuera del alcance del presente proyecto académico.

Análisis Exploratorio de Datos

Descripción de las Características de los Datos

El dataset utilizado contiene información relacionada con la enfermedad de Alzheimer, incluyendo características demográficas, históricas y clínicas de los pacientes. Después de una revisión inicial:

Dimensiones del dataset: 1000 filas y 20 columnas.

-Tipos de datos:

- Numéricos: Edad, puntuaciones en pruebas clínicas.

- Categóricos: Género, antecedentes familiares, diagnóstico.

- Valores nulos:Se encontraron algunos valores nulos en columnas específicas como 'Historia Familiar', pero esta sección se eliminó.

- Duplicados:Se eliminó una cantidad mínima de filas duplicadas para asegurar la unicidad de los registros.

Identificación de Patrones o Tendencias Significativas

El análisis reveló:

- Distribución del Diagnóstico:

La mayoría de los pacientes están en etapas iniciales de la enfermedad.

-Distribuciones de Variables Numéricas:

Las puntuaciones de evaluación funcional tienden a ser más bajas en pacientes con diagnóstico positivo.

Relaciones y Correlaciones entre las Variables

-Matriz de Correlación: Se identificaron correlaciones moderadas entre la edad y las puntuaciones de pruebas clínicas.

- Visualización de Datos: Gráficos de cajas y diagramas de dispersión sugieren una posible relación entre antecedentes familiares y la presencia de Alzheimer.

Preprocesamiento de Datos

Para preparar los datos para el análisis:

- **Eliminación de Valores Atípicos:** Se eliminaron valores extremos en las puntuaciones de pruebas clínicas.

- **Imputación de Valores Faltantes:** Valores nulos en 'Historia Familiar' fueron imputados con la moda de la columna.

-Normalización: Las variables numéricas fueron normalizadas utilizando MinMaxScaler para modelos sensibles a la escala.

Selección y Entrenamiento del Modelo

Se evaluaron varios modelos de clasificación:

- **Random Forest y Gradient Boosting** demostraron mejor rendimiento sin necesidad de escalado.

- **K-Nearest Neighbors (KNN) y Support Vector Machine (SVM**)fueron entrenados con datos escalados.

- **Validación Cruzada:** Se utilizó una validación cruzada de 5 pliegues para evaluar la eficacia de cada modelo.

Evaluación del Modelo

- Métricas de Evaluación: Se calcularon las siguientes métricas para cada modelo:

- **Precisión:**85%
- **Sensibilidad**: 90% (mejor rendimiento con Gradient Boosting)
- **Especificidad**: 80%

- **F1-Score:** 87%

-**Ajuste de Hiperparámetros:**

Se utilizó GridSearchCV para optimizar los hiperparámetros del modelo Gradient Boosting.

Gestión de Proyecto

El proyecto fue manejado siguiendo una estructura organizada:

- **Planificación**:

Definición clara de objetivos y etapas.

- **Ejecución**:

División en fases de exploración, preprocesamiento, modelado y evaluación.

- Revisión:

Evaluación iterativa del modelo y ajustes basados en los resultados obtenidos.

Conclusiones y Recomendaciones

- Conclusiones:

- Los modelos de ensamble como Random Forest y Gradient Boosting son eficaces para la predicción del diagnóstico de Alzheimer en este modelo.

- La normalización de datos mejora significativamente el rendimiento de modelos sensibles a la escala.

-Recomendaciones:

- Ampliar el dataset con más muestras para mejorar la robustez del modelo.

- Implementar modelos en un entorno de producción para monitorear el rendimiento en tiempo real.

- Continuar ajustando hiperparámetros y explorar arquitecturas de redes neuronales para mejorar aún más la precisión.

Referencias:

- Kharoua, R. (s.f.). *Alzheimer's Disease Dataset*. Kaggle. Recuperado de https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset

- Sociedad Española de Neurología. (2019). *El 35% de los casos de Alzheimer se pueden atribuir a nueve factores de riesgo modificables*. Recuperado de https://www.sen.es/saladeprensa/pdf/Link280.pdf

- EpData. (2022). *Las cifras del Alzheimer en España: número de personas y mortalidad*. Recuperado de https://www.epdata.es/datos/cifras-alzheimer-espana-numero-personas-mortalidad-muertes-graficos-datos/671

Quirónsalud. (s.f.). *Cada año se diagnostican en España más de 40.000 casos de Alzheimer, una enfermedad que afecta a más de 800.000 personas en nuestro país*. Recuperado de https://www.quironsalud.com/es/comunicacion/contenidos-salud/cada-ano-diagnostican-espana-40-000-casos-alzheimer-enferme

- Alzheimer's Association. (s.f.). *Datos y cifras sobre la enfermedad de Alzheimer*. Recuperado de https://www.alz.org/es-mx/alzheimer-demencia/datos-y-cifras

- Sociedad Española de Neurología. (2022). *Las demencias ya suponen el 8% del total de defunciones que se producen en España*. Recuperado de https://www.sen.es/saladeprensa/pdf/Link451.pdf

- Etayo, A. (s.f.). *Censo de las personas con Alzheimer y otras Demencias en España*. Recuperado de https://www.congresonacionaldealzheimer.org/files/84/31/83/censo-ainhoa-etayo.pdf

- Statista. (s.f.). *Enfermedad de Alzheimer: defunciones por género 2005-2017*. Recuperado de https://es.statista.com/estadisticas/593919/numero-de-muertes-por-alzheimer-por-generos-en-espana/

- Servicio de Información sobre Discapacidad. (s.f.). *España es uno de los países del mundo con mayor proporción de casos de Alzheimer*. Recuperado de https://sid-inico.usal.es/noticias/espana-es-uno-de-los-paises-del-mundo-con-mayor-proporcion-de-casos-de-alzheimer/

- Confederación Española de Alzheimer. (2023). *Informe Censo de Personas con Alzheimer y otras Demencias en España*.

Recuperado de https://www.ceafa.es/files/2023/05/informe-censo-alz-2-web.pdf

Anexo:

Como anexo, cabe mencionar este video de youtube que sirvió como guía para realizar el proyecto:

Detección de enfermedades cardíacas en Python: proyecto de aprendizaje automático - YouTube