Reporte Final: Proyecto de Ciencia de Datos I - Introducción a fundamentos

Análisis de Alzheimer y Factores de Riesgo

Introducción

El presente trabajo tiene como objetivo explorar un conjunto de datos relacionado con la enfermedad de Alzheimer (fuente Kaggle), analizando factores de riesgo (como síndrome metabólico, enfermedades cardíacas, diabetes y tabaquismo), factores protectores (educación y actividad física), evaluación cognitiva y de actividades de la vida diaria y finalmente síntomas relacionados a trastornos cognitivos.

La intención principal es identificar patrones que contribuyan al desarrollo de la enfermedad y construir un modelo predictivo enfocado en minimizar los falsos negativos (error tipo II), priorizando la sensibilidad del modelo.

Descripción de la Base de Datos

La base de datos utilizada contiene 2,149 registros con 33 columnas que incluyen variables demográficas, clínicas, de estilo de vida y de evaluación cognitiva:

- **Variables Numéricas**: Edad, IMC, consumo de alcohol, actividad física, calidad de dieta, presión arterial, colesterol (total, LDL, HDL, triglicéridos), puntajes de MMSE (mini mental state examination), evaluación funcional y puntaje de actividades de la vida diaria (ADL).
- **Variables Categóricas**: Género, Tabaquismo, Antecedentes familiares, Educación, entre otros.

El objetivo principal del modelo a desarrollar es predecir la variable objetivo “Diagnóstico” (0 = Sin Alzheimer, 1 = Con Alzheimer).

Limpieza de Datos y Feature Engineering

Procesos de limpieza e ingeniería de variables

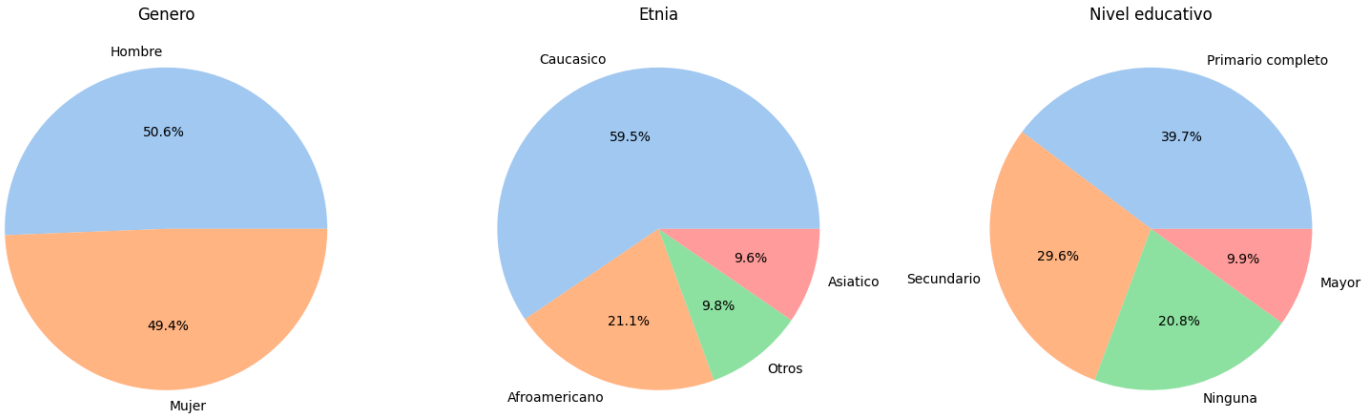
1. **Valores nulos**: No se encontraron valores nulos en las columnas.
2. **Se mapearon variables categóricas para realizar one hot encoding y aumentar la sensibilidad de la selección de variables y modelos.**
    - **Etnia**: caucasico, afroamericano, asiatico, otro
    - **Nivel Educativo**: Ninguna, primario completo, secundario, y mayor.
3. **En base a la variable numérica IMC (BMI), se crearon tres variables categóricas peso regular, sobrepeso y obesidad.**
4. **Se utilizó StandardScaler para utilizar la misma escala en todas las variables numéricas.**

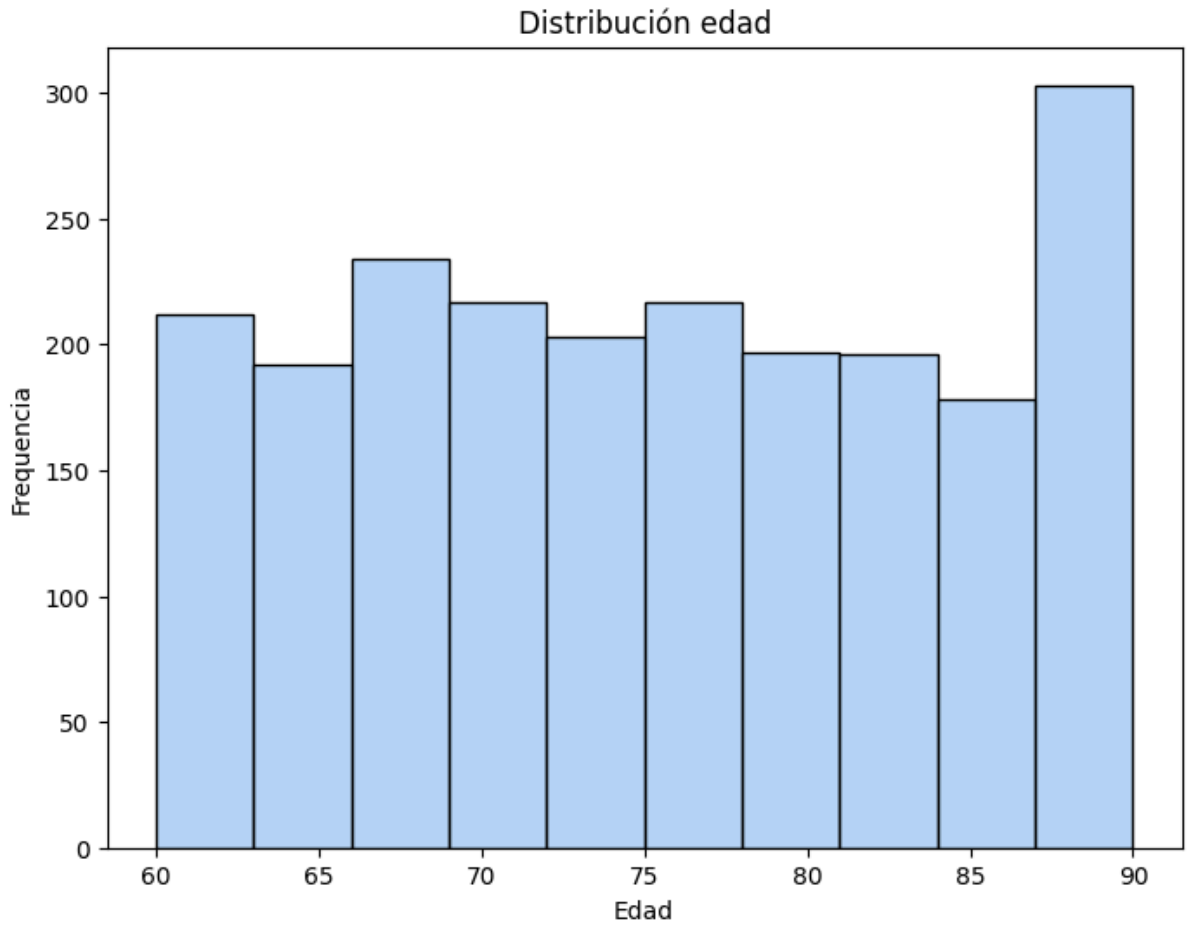
Selección de Variables

Se eliminaron variables con baja correlación (< 0.001) respecto al diagnóstico. A modo de simplificar los modelos predictivos y disminuir los tiempos de cómputos.

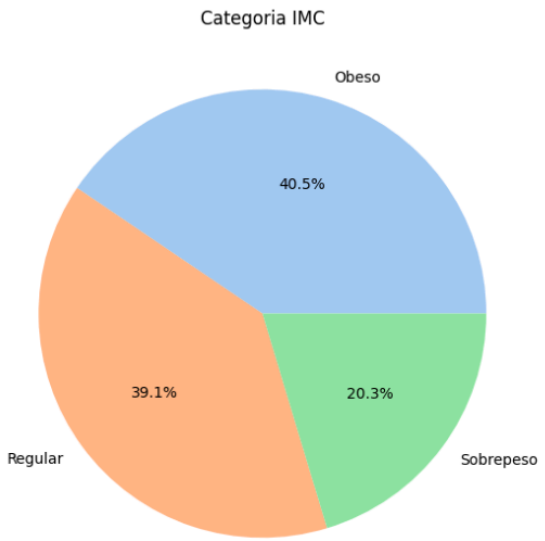
Análisis Exploratorio de Datos (EDA)

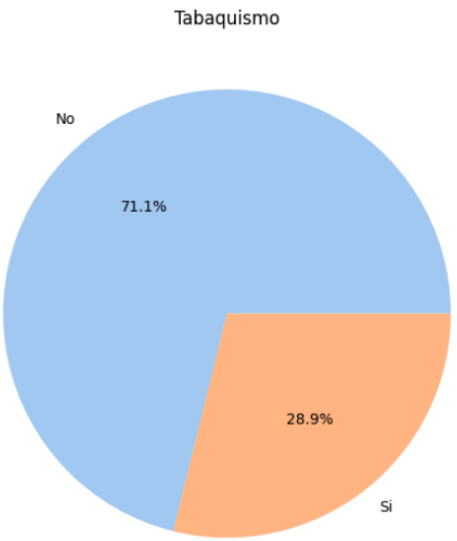
Distribuciones y Gráficos Representativos



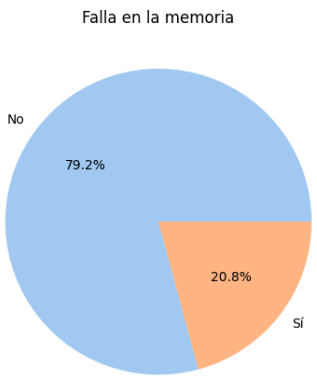


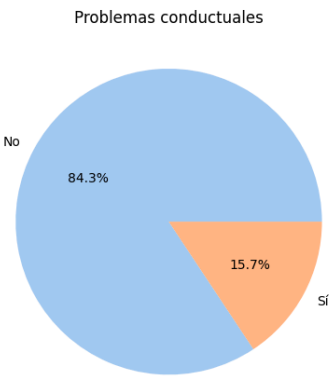
- El grupo etario se distribuyó principalmente entre los 60 y 90 años.

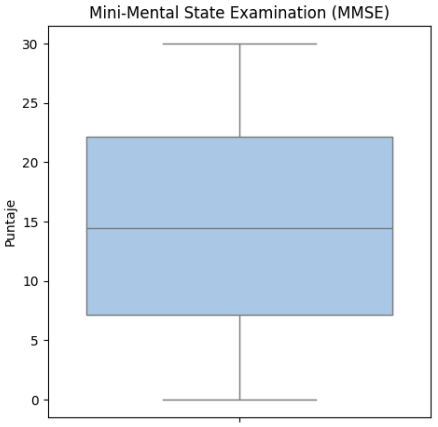


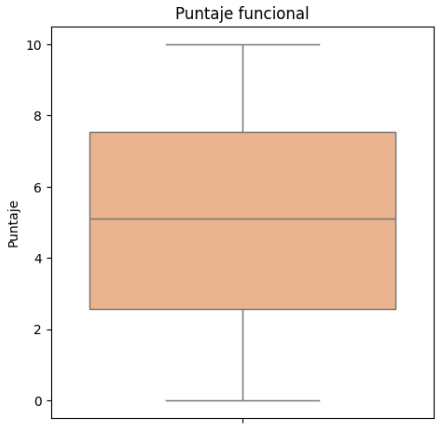


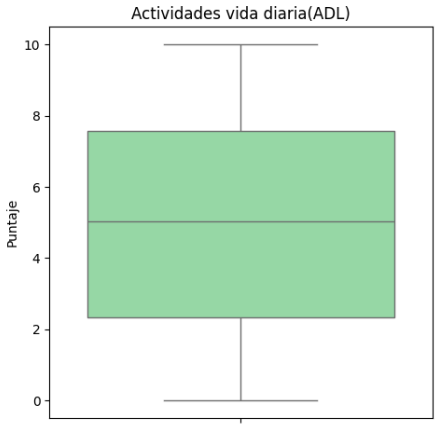
Evaluación cognitiva y funcional



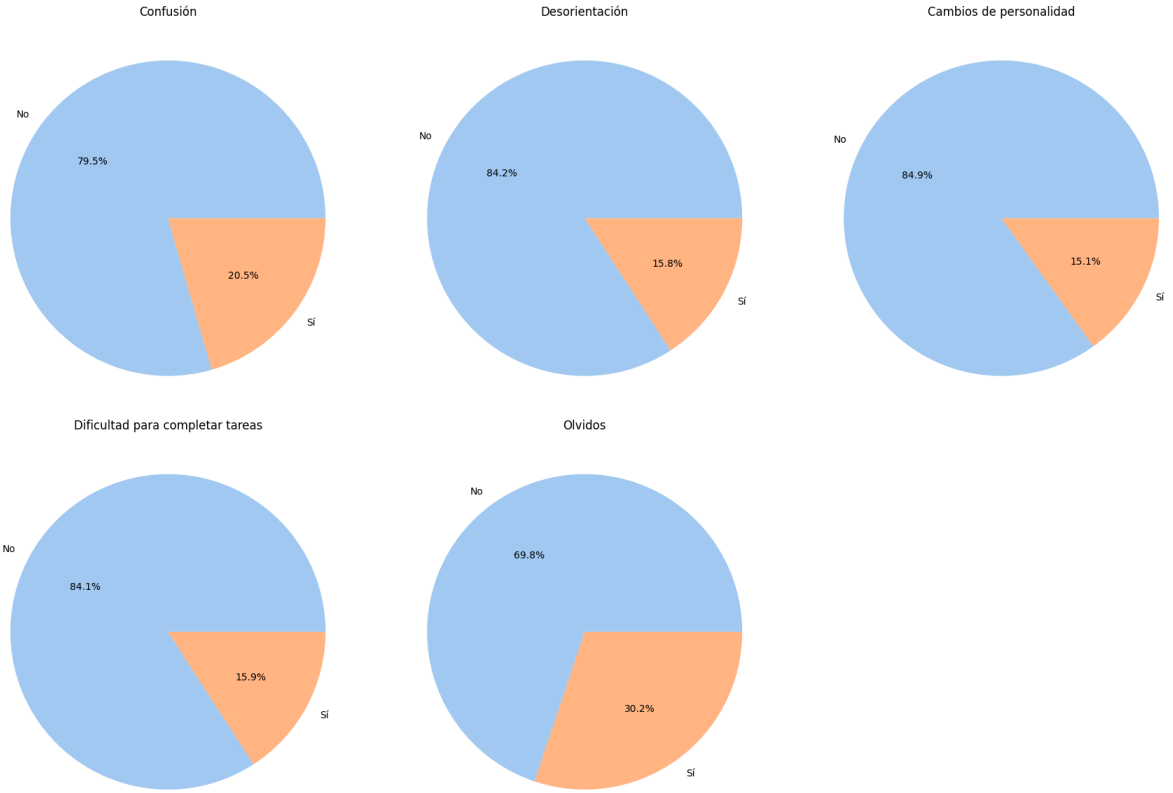


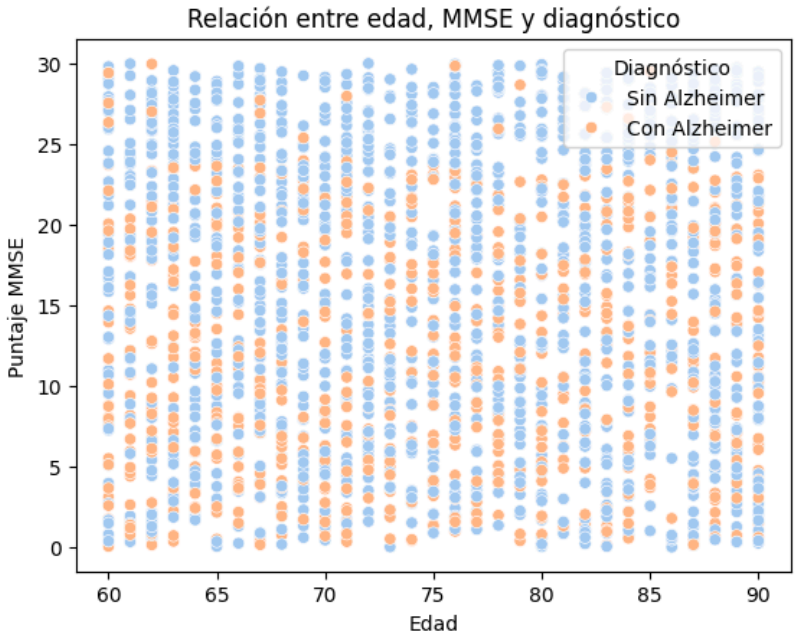






Sintomatología asociada





- No se observa una clara relación lineal entre el puntaje de MMSE y la edad en relación al diagnóstico de Alzheimer. Probablemente debido a relaciones no lineales complejas, y que el deterioro cognitivo, evaluado por MMSE, puede tener causas diferentes como por ejemplo depresión, déficit de atención u otros tipos de demencia.

Creación de Modelos Predictivos

Se desarrollaron cuatro modelos predictivos:

1. **Regresión Logística**
2. **K-Nearest Neighbors (KNN)**
3. **Árbol de Decisión**
4. **Random Forest**

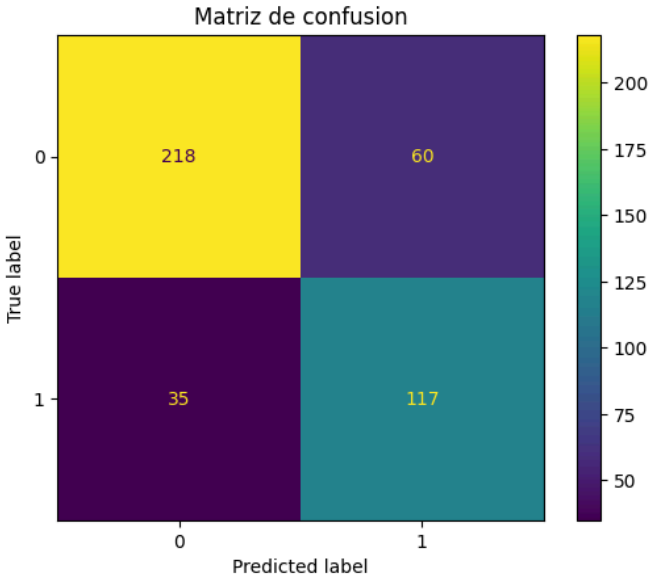
- En general para los cuatro modelos se realizó:
    - Se utilizó stratify en base a la variable dependiente para mantener balance (aproximadamente 65 -35 %) en los muestreos de prueba y testeo.
    - A su vez para la validación cruzada se utilizó StratifiedKFold
    - Utilizacion de GridSearchCV para encontrar los mejores hiperparametros, se busco el mejor valor de recall para disminuir el error tipo II. Manteniendo de todas formas un nivel elevado de ROC AUC (Receiver Operating Characteristic - Area Under the Curve).
- En específico para el modelo de regresión logística además se realizó:
    - Utilización de PolynomialFeatures.
    - Evaluación de hiperparametros con penalidad L1, L2 y elastic net.(se creo una funcion con python para filtrar parámetros compatibles en base a la penalidad y a los solvers)
- Se creó una función de python para extraer de igual manera para cada modelo:
    - Accuracy, precision, recall, F1-score, ROC AUC
    - Gráfico de matriz de confusión
    - Gráfico de ROC

Árbol de Decisión (Modelo Destacado)

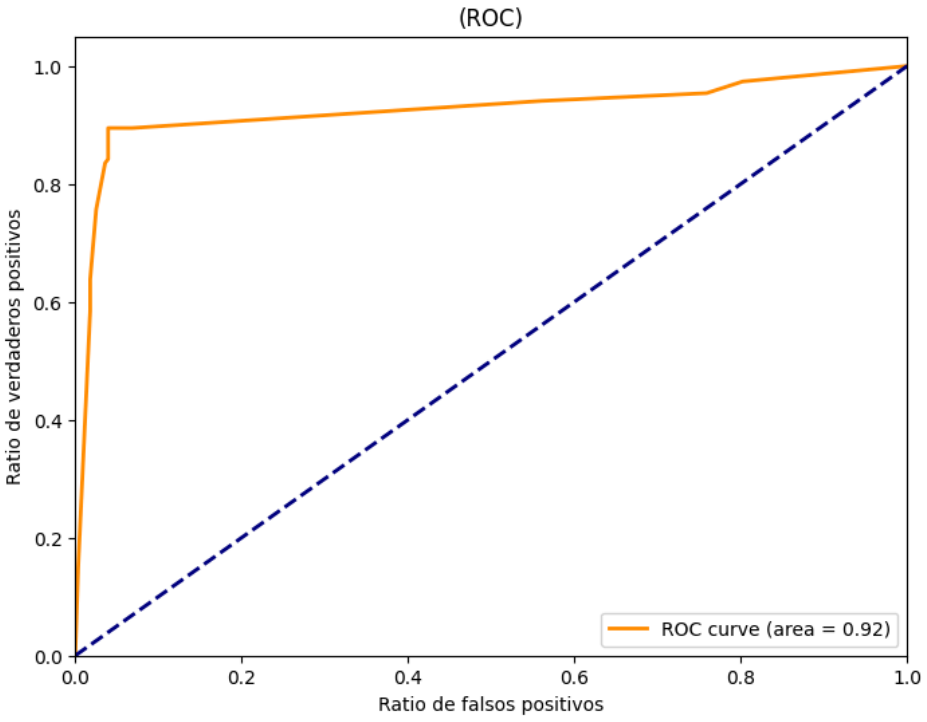
- **Fue el modelo predictivo con mejor recall (objetivo principal) con casi un 90%. Manteniendo una alta precisión con ROC AUC de 92.45%**
- **Hiperparámetros Optimizados**:
  {'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 30, 'splitter': 'best'}
- **Métricas de Evaluación:**

Métrica	Valor
Precisión	91.89%
Recall	89.47%
F1-Score	90.67%
ROC-AUC	92.45%

- **Matriz de confusión:**



- **Gráfico de ROC AUC:**



Visualización del Árbol de Decisión

- Se utilizó graphviz para visualizar el mejor modelo de árbol de decisiones.
- En el mismo se observa que el modelo prioriza variables como **evaluación funcional**, **MMSE** y **problemas conductuales**.

Resumen y Conclusiones

1. **Insights Principales**:
    - La edad avanzada, la obesidad y las condiciones crónicas como hipertensión y diabetes contribuyen significativamente al riesgo de Alzheimer.
    - Factores cognitivos y funcionales como **MMSE** y **ADL** tienen una fuerte asociación con el diagnóstico.
2. **Desempeño del Modelo**:
    - El árbol de decisión alcanzó un buen equilibrio entre sensibilidad y precisión, lo que lo hace adecuado para minimizar falsos negativos.
    - Otro beneficio del árbol de decisión en comparación con otros modelos fue su rápida computación y búsqueda de hiperparametros, además de su mayor interpretabilidad, permitiendo valorar cuáles fueron las variables con mayor correlación con mayor jerarquía en el árbol de decisiones.
3. **Limitaciones**:
    - La base de datos es transversal y no incluye datos longitudinales para evaluar la progresión del Alzheimer.
    - La base de datos no especifica si se excluyen otros diagnósticos de demencia.
    - Algunas variables importantes como imágenes cerebrales, datos genéticos (e.g., APOE4) o variables funcionales como test de reloj o CDR, no están incluidas.
4. **Mejoras Futuras**:
    - Ampliar el tamaño de la muestra y recolectar datos longitudinales.
    - Incorporar variables adicionales como biomarcadores, historial detallado de medicamentos y patrones de actividad cerebral.
    - Probar modelos avanzados como Boost o redes neuronales.
