Predicción de Riesgo de Accidente Cerebrovascular

Con equidad.

2025-04-28

Práctica Profesional II

Profesor: Carlos Ignacio Charletti

Equipo: Dante Javier Pagano, Eugenia Barozzi, Federico Gurrea, Juan Marcelo Molina, Julieta Battauz y Laura Peralta

Github: https://github.com/Data-Dinasty/Data-Dinasty_TSDCIA_ISPC-PPII/tree/master

colab del proyecto: pp2-entrega2.ipynb

Data Summary:
https://github.com/Data-Dinasty/Data-Dinasty_TSDCIA_ISPC-PPII/blob/master/docs/data/data_summary.md

Informe de hallazgos

El dataset contiene información clínica y demográfica de 5110 personas con el objetivo de predecir la ocurrencia de un accidente cerebrovascular. Incluye variables como edad, nivel de glucosa en sangre, índice de masa corporal, antecedentes de hipertensión, enfermedades cardíacas, tipo de trabajo, estado civil, entre otras. La variable objetivo es binaria (stroke: 0 o 1) y el conjunto de datos presenta un desbalance significativo entre clases.

Sección técnica

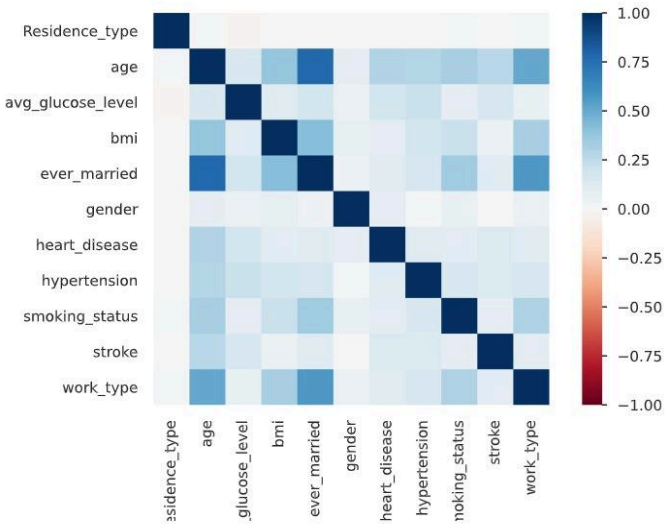
1. Valores faltantes:

- La variable “bmi” tiene 201 valores faltantes (3.93% del total).
- Distribución de valores faltantes en “bmi” por género:
    - 51.7% de los casos corresponden a personas identificadas como hombres
    - 48.3% de los casos corresponden a personas identificadas como mujeres.
    - 0% para la categoría "Other".

2. Correlaciones destacadas:

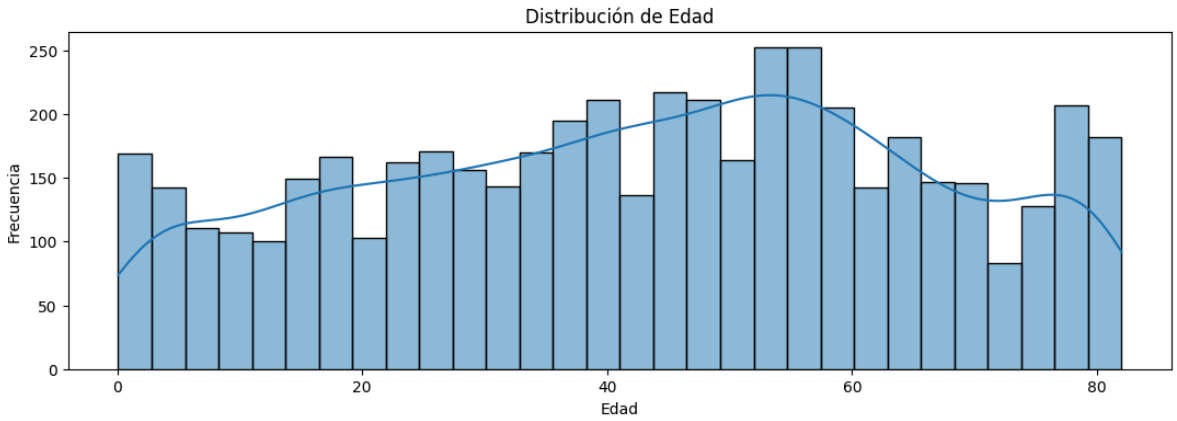
- “stroke” muestra correlaciones moderadas con:
    - age: 0.25
    - hypertension: 0.13
    - heart_disease: 0.13
    - avg_glucose_level: 0.13

- “bmi” presenta una correlación más débil con “stroke”: 0.04



3. <u>Distribuciones generales</u>**:**

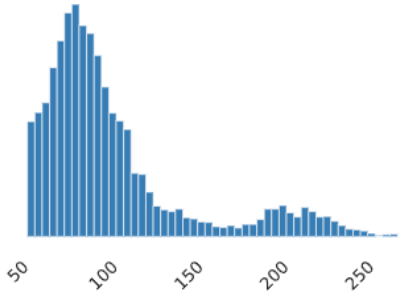
- Edad media: **43.2 años** (máx. 82)



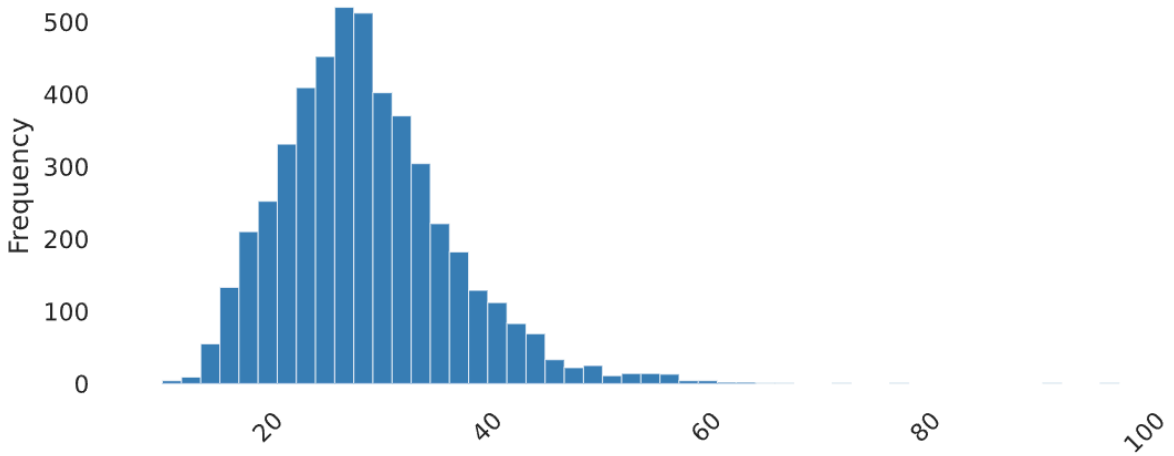
- Glucosa media: **106.1 mg/dL**

Distinct	3852
Distinct (%)	78.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	105.30515

Minimum	55.12
Maximum	271.74
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	76.7 KiB



- BMI medio: **28.9** (mín. 10.3, máx. 97.6)



- Tasa general de ACV (stroke): **4.87%**

stroke
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	316.4 KiB

Informe ético

Mitigación de Sesgos en los Datos y Proceso de Limpieza

Introducción

La integridad ética en proyectos de ciencia de datos no solo depende de la precisión estadística, sino también de la equidad en la representación de grupos poblacionales y en la interpretación de resultados. Este informe aborda cómo se identificaron y mitigaron posibles sesgos en el dataset de predicción de ACV, y qué acciones se tomaron en el proceso de limpieza para minimizar impactos éticamente problemáticos.

1. Identificación de Posibles Sesgos

a. Representación Demográfica

- La categoría gender = Other presenta **solo un registro**, lo que impide cualquier análisis estadístico fiable. Esta **subrepresentación** implica un sesgo en la recolección de datos que debe ser reconocido al comunicar resultados.

- Los datos muestran una **distribución desigual por tipo de trabajo**, donde grupos como children y Never_worked están muy poco representados, lo que podría distorsionar las conclusiones sobre la relación entre ocupación y riesgo de ACV.

b. Variables Proxy de Riesgo Socioeconómico

- **Residence_type** y **work_type** están asociadas con la tasa de ACV. Estas variables pueden estar actuando como *proxies* de condiciones estructurales (acceso a salud, nivel económico, nivel educativo), lo que puede introducir sesgos al entrenar modelos predictivos si no se tratan con cautela.

c. Disparidades Observadas

- Existen diferencias en la tasa de ACV entre hombres (5.1%) y mujeres (4.7%), y entre residentes urbanos (5.2%) y rurales (4.5%). Estas diferencias pueden estar

influenciadas por factores sociales o de acceso, y no necesariamente por causas médicas.

2. Limpieza de Datos Éticamente Responsable

a. Manejo de Valores Faltantes

- La variable bmi contiene un 3.93% de valores faltantes. Se evaluaron distintas estrategias de imputación:
  - Imputación por media segmentada por género y grupo etario.
  - Imputación usando regresión o algoritmos como KNN.

- Se evitó eliminar registros con valores faltantes para no reducir la diversidad de la muestra ni distorsionar la distribución de variables clave.

b. Normalización y Codificación

- Se aplicó una codificación *one-hot* para variables categóricas como gender, work_type y Residence_type para evitar que se interpreten como ordinales por los modelos.

- Se evitó reescalar variables que no lo requerían médicamente (como age) para preservar su interpretación clínica directa.

c. Exclusión de Casos No Representativos

- Dado que gender = Other representa solo un caso, se excluyó de los análisis estadísticos y visualizaciones, pero se reportó esta exclusión explícitamente para no invisibilizar al grupo.

3. Recomendaciones para la Mitigación de Sesgos

1. **Auditorías de equidad** especialmente respecto a género y residencia.

2. **Validación estratificada** por subgrupos (género, ocupación, tipo de residencia) para asegurar que el modelo no favorezca a un grupo en lugar de otro.

3. **Documentar explícitamente las exclusiones** o transformaciones realizadas para mantener la transparencia.

4. **Considerar métricas de fairness** como igualdad de oportunidades o precisión balanceada, si se desarrollan modelos de riesgo clínico.

5. **Mejorar el muestreo** en futuras recolecciones de datos para garantizar representatividad equitativa entre todos los grupos sociales y de género.

Conclusión

El análisis ético de un dataset va más allá de su estructura implica entender las consecuencias que los modelos y análisis pueden tener en la vida de las personas. En este trabajo se aplicaron buenas prácticas de limpieza de datos y se identificaron fuentes de sesgo que podrían condicionar los resultados. Recomendamos continuar este enfoque ético en todas las etapas del ciclo de vida del proyecto.