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

Predicción de ACV

Predicción de Riesgo de Accidente Cerebrovascular con Equidad

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

Objetivo del Proyecto

El objetivo de este proyecto es desarrollar un modelo de *machine learning* para predecir la probabilidad de que una persona sufra un accidente cerebrovascular (ACV) utilizando variables clínicas y socioeconómicas. Además buscar un buen desempeño técnico, el modelo será evaluado en términos de equidad, con el fin de garantizar que no reproduzca sesgos en la predicción, especialmente en relación con el género y la condición laboral. Este enfoque contribuirá a mejorar la prevención en salud pública y reducir las desigualdades.

Alcance del Proyecto

Incluye:

- Uso del dataset stroke-prediction .
- Análisis exploratorio de datos clínicos y demográficos.
- Desarrollo de modelos de clasificación supervisada.
- Evaluación del rendimiento técnico del modelo (precisión, AUC, recall, etc.).
- Evaluación de la equidad del modelo mediante métricas de *fairness*.
- Documentación en GitHub y seguimiento del proceso usando TDSP y Jira/Trello.

Excluye:

- Implementación clínica del modelo en entornos hospitalarios reales.
- Recolección de nuevos datos o integración con sistemas sanitarios locales.

- Diagnóstico médico o validación con datos de pacientes en tiempo real.

Metodología

Se sigue la metodología TDSP (Team Data Science Process), que estructura el trabajo en las siguientes fases: entendimiento del negocio, exploración de datos, modelado, validación, despliegue y evaluación. Se utilizarán herramientas como Python, Scikit-learn, MLflow y Fairlearn para asegurar un desarrollo técnico robusto y una revisión de sesgos en el modelo.

1. Justificación de la elección del dataset

Dataset elegido: Stroke Prediction Dataset

Impacto social: El ACV es una de las principales causas de muerte y discapacidad a nivel mundial. Predecir la probabilidad de sufrir un ACV a partir de características clínicas y socioeconómicas permite realizar intervenciones tempranas, mejorar la calidad de vida y reducir los costos del sistema de salud.

Datos sensibles: El dataset incluye variables relacionadas con el género, la edad, la hipertensión, enfermedades cardíacas, IMC, tipo de trabajo y nivel de glucosa. Estas variables permiten explorar no solo modelos predictivos, sino también analizar posibles sesgos algorítmicos, lo cual es crucial para evaluar la equidad y justicia del modelo.

Aplicabilidad local: Este tipo de análisis puede replicarse en contextos locales con bases de datos clínicas, por lo que es un modelo de referencia útil para proyectos de salud pública local.

2. Pregunta de negocio (Business Question)

Pregunta: ¿El modelo de predicción de ACV tiene un desempeño equitativo entre géneros y grupos socioeconómicos, además de ser clínicamente útil?

Esta pregunta busca:

- Construir un modelo predictivo (clasificación binaria: ¿tendrá ACV o no?).
- Analizar la equidad del modelo usando métricas de *Fairlearn*.

- Generar valor para los stakeholders del sistema de salud y la comunidad.

3. Stakeholders (Partes Interesadas)

Rol	Descripción
Autoridades Sanitarias	Interesadas en políticas preventivas basadas en datos para reducir el riesgo de ACV en la población.
Profesionales de la Salud	Podrán utilizar el modelo para identificar pacientes en riesgo y actuar preventivamente.
Comunidad General	Beneficiarios del proyecto al recibir diagnósticos más justos y prevención temprana.
Equipo de Ciencia de Datos	Desarrolladores y responsables técnicos y éticos del modelo. Incluye:
- Project Manager	Coordina y supervisa el desarrollo del proyecto.
- Data Engineer	Se encarga del procesamiento y limpieza de los datos.
- Data Scientist	Desarrolla y evalúa los modelos predictivos.
- Ethical Reviewer	Evalúa el modelo en términos de justicia, equidad y sesgos.

Métricas de éxito (técnicas y de equidad)

Métricas Técnicas:

- **Accuracy:** % de predicciones correctas.
  Requerido: Accuracy > 90%.

- **Precision:** Qué tan confiables son los positivos que predijo.
  Requerido: Precision > 85%.

- **Recall:** Cuántos positivos verdaderos logra detectar.
  Requerido: Recall > 80%. (Crucial en salud).

- **F1-Score:** Balance entre precisión y recall.
  Requerido: F1-Score > 82%.

- **AUC-ROC:** Capacidad de separar 0 vs 1 en general.
  Requerido: AUC > 0.85.

Métricas de Equidad

El dataset incluye atributos sensibles como **Sexo**, **Residence_type** (rural o urbano), y **work_type** (tipo de trabajo), por lo que el modelo debe garantizar que no discrimine en base a:

- **Género**

- **Ubicación (rural vs urbana)**

Métricas de Equidad comunes:

- **Demographic Parity:** Que la tasa de predicción positiva sea parecida entre grupos.
  Éxito: Diferencia < 5%.

- **Equal Opportunity:** Que la tasa de verdaderos positivos sea parecida entre grupos.
  Éxito: Diferencia < 5%.

Estas métricas se evaluarán usando *Fairlearn*, con especial foco en el género y tipo de ocupación como atributos sensibles.

Modelos de ML para este set de datos

1. **Logistic Regression:** Comenzaremos con regresión logística, debido a su rapidez y facilidad de interpretación.

2. **Random Forest:** Para mejorar los resultados, utilizaremos Random Forest como un modelo más robusto.

Herramientas Utilizadas

- **Python, Scikit-learn** para modelado.
- **Fairlearn** para evaluación de equidad.
- **MLflow** para gestión de experimentos.
- **GitHub** para control de versiones.
- **TDSP** como metodología de desarrollo.

Este enfoque asegurará que el modelo no solo sea técnico y predictivamente efectivo, sino también ético y justo para todos los grupos representados en los datos.