Predicción de ACV utilizando Machine Learning

Introducción

Los accidentes cerebrovasculares (ACV) son una de las principales causas de muerte y discapacidad a nivel mundial. Detectar precozmente a personas con alto riesgo de padecer un ACV puede ayudar a prevenir eventos graves mediante tratamientos médicos y cambios en el estilo de vida. En este trabajo se busca desarrollar un modelo predictivo utilizando técnicas de Machine Learning para estimar la probabilidad de que una persona sufra un ACV, a partir de variables clínicas y demográficas.

Descripción del Dataset

- **Fuente**: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
- **Registros**: 5,110
- **Variables**: Edad, género, hipertensión, enfermedades cardíacas, estado de trabajo, tipo de residencia, niveles de glucosa promedio, IMC, estado de fumador, etc.
- **Variable objetivo**: **stroke** (1: ocurrió un ACV, 0: no ocurrió)

Este dataset fue seleccionado por su relevancia clínica y la disponibilidad de variables significativas para la predicción de ACV.

Análisis Exploratorio de Datos (EDA)

Durante el EDA se observaron los siguientes puntos clave:

- **Distribución de edad:** Los casos de ACV se concentran mayormente en pacientes mayores de 60 años.
- **Glucosa y BMI:** Pacientes con niveles más altos en estas variables mostraron mayor incidencia de ACV.
- **Condiciones médicas:** La hipertensión y enfermedades cardíacas están correlacionadas positivamente con la ocurrencia de ACV.
- **Visualizaciones:** Se emplearon histogramas, boxplots y mapas de calor para identificar patrones.

Curado y Preparación del Dataset

Imputación de valores nulos

Durante la limpieza del dataset, se identificaron valores nulos en la variable **BMI** (Índice de Masa Corporal). Para manejar estos datos faltantes, se decidió imputar dichos valores usando la **mediana** de la variable. Esta elección es especialmente adecuada en contextos médicos, ya que la mediana es menos sensible a valores extremos o atípicos, proporcionando así una estimación más robusta y representativa para la población analizada.

Codificación de variables categóricas

Para poder utilizar algoritmos de Machine Learning que requieren variables numéricas, se aplicó **One-Hot Encoding** a las columnas categóricas como **work_type** y **smoking_status**. Este método crea una nueva columna para cada categoría posible, asignando 1 si el registro pertenece a esa categoría y 0 en caso contrario. Esto evita la introducción de un orden artificial (como ocurriría con Label Encoding) y asegura que los algoritmos no interpreten erróneamente una relación ordinal entre categorías.

Estandarización de variables numéricas

Se estandarizaron las variables numéricas como **age, avg_glucose_level y bmi** mediante el método **StandardScaler**, que transforma los valores para que tengan media 0 y desviación estándar 1. Esto es especialmente importante para modelos como Regresión Logística, SVM y K-Vecinos, que son sensibles a las escalas de los datos. Sin estandarización, variables con valores más altos dominarían las decisiones del modelo.

División del dataset con estratificación

El dataset se dividió en conjuntos de entrenamiento y prueba en una proporción 70/30. Se utilizó la opción de **estratificación basada en la variable objetivo stroke**, que está desbalanceada (muy pocos casos positivos en comparación con negativos).

¿Por qué estratificar?

En datasets desbalanceados, una división aleatoria puede generar un conjunto de prueba sin ejemplos positivos, lo que vuelve imposible evaluar el desempeño real del modelo. Al estratificar, se garantiza que tanto el conjunto de entrenamiento como el de prueba conserven la misma proporción de clases que el dataset original, permitiendo una evaluación más realista y fiable.

Modelado y Resultados

Dado que el objetivo del proyecto es predecir si una persona ha sufrido un Accidente Cerebrovascular (**ACV**) o no, estamos ante un problema de clasificación binaria. La variable objetivo (**stroke**) toma solo dos valores:

- 1: el paciente ha sufrido un ACV,
- 0: el paciente no ha sufrido un ACV.

Por este motivo, se seleccionaron y probaron distintos modelos de clasificación supervisada, tales como:

- Regresión Logística
- K-Vecinos más Cercanos (KNN)
- XGBoost
- Random Forest

Regresión Logística

- **Accuracy:** 0.95
- **Precision para clase 1:** 1.00
- **Recall clase 1:** 0.01
- **F1-score clase 1:** 0.03

El modelo ignora prácticamente todos los casos positivos. Alta precisión porque predice muy pocos positivos, pero muy bajo recall. Esto es **inadecuado para un problema médico**.

Random Forest

- Modelo más robusto, pero sin ajustar el umbral no mejora significativamente el recall.
- Con threshold = 0.5 el recall sigue siendo muy bajo.

Se prueba ajustar el **umbral de decisión** para clase 1:

```
Umbral = 0.10 -> Precision: 0.158, Recall: 0.600, F1-score: 0.251
Umbral = 0.20 -> Precision: 0.202, Recall: 0.253, F1-score: 0.225
Umbral = 0.30 -> Precision: 0.212, Recall: 0.093, F1-score: 0.130
Umbral = 0.40 -> Precision: 0.200, Recall: 0.040, F1-score: 0.067
Umbral = 0.50 -> Precision: 0.333, Recall: 0.013, F1-score: 0.026
```

El mejor punto fue umbral **0.10**, logrando detectar el 60% de los casos de ACV (recall), aunque con muchos falsos positivos.

XGBoost con scale_pos_weight

- Se ajustó el parámetro scale_pos_weight para equilibrar el impacto de la clase minoritaria.
- Detectó 14 casos positivos reales, mucho mejor que modelos anteriores sin ajustar.

K-Nearest Neighbors (KNN)

- Totalmente sesgado hacia la clase mayoritaria.
- No recomendable para datasets fuertemente desbalanceados sin técnicas de balanceo previas.

Balanceo de clases con SMOTE

El dataset presentaba una fuerte desproporción entre las clases: solo alrededor del 5% de los pacientes habían sufrido un ACV (clase 1). Para resolver este problema y evitar que los modelos se inclinen por la clase mayoritaria, se aplicó la técnica de **SMOTE (Synthetic Minority Oversampling Technique)**.

SMOTE genera ejemplos sintéticos de la clase minoritaria basándose en sus vecinos más cercanos, lo que permite entrenar los modelos sobre un conjunto más equilibrado sin duplicar ejemplos reales.

Este paso se aplicó **luego de dividir el dataset en entrenamiento y test**, utilizando únicamente los datos de entrenamiento para evitar el "data leakage".

Ventajas de SMOTE:

- Mejora el **recall** y la capacidad del modelo para detectar verdaderos positivos.
- No genera overfitting tan fácilmente como el oversampling tradicional.
- Permite aplicar modelos estándares sin necesidad de ajustes extremos de umbral o penalizaciones.

Esta técnica fue fundamental para mejorar los resultados de los modelos como **Random Forest** y **XGBoost**, especialmente en la clase stroke = 1.

Cuadro comparativo de resultados

Modelo	Accuracy	Recall Clase 1	F1 Clase 1	AUC	Observaciones
Regresión Logística	0.95	0.01	0.03	0.87	Muy bajo recall. Modelo ignora la clase 1.
Random Forest (th=0.10)	0.89	0.60	0.25	0.89	Umbral ajustado para mejorar el recall
XGBoost (scale_pos_weight)	0.91	0.56	0.27	0.91	Muy buen desempeño general
KNN	0.95	0.00	0.00	0.603	Completamente sesgado. No detecta ACV.
Random Forest (SMOTE)	0.96	0.72	0.67	0.94	Mejor equilibrio. Buen recall y precisión.

Conclusiones

- El dataset es muy desbalanceado, lo que representa un reto importante en la predicción de ACV.
- Ajustar el umbral de clasificación o usar técnicas de penalización/clase ponderada permite mejorar el rendimiento en la clase minoritaria.
- El modelo XGBoost con scale_pos_weight y Random Forest con threshold bajo lograron recall aceptables para uso médico.
- Solo Random Forest fue probado con SMOTE, y mostró una mejora sustancial en *recall* y *F1-score*.
- Los demás modelos (como regresión logística o XGBoost) fueron entrenados sin técnicas de balanceo, aunque en XGBoost se ajustó scale_pos_weight para compensar.
- El uso de SMOTE fue clave para aumentar la sensibilidad en la clase minoritaria y debería considerarse para todos los modelos en trabajos futuros.
- En medicina, se prioriza recall (detectar todos los casos posibles), incluso con mayor cantidad de falsos positivos.
- La predicción automática puede ser una herramienta de soporte para priorizar pacientes, pero no debe reemplazar el juicio clínico.

Conclusiones Generales

Este análisis demuestra que, con técnicas de preprocesamiento adecuadas y modelos ajustados, es posible detectar pacientes en riesgo de ACV con razonable efectividad, incluso en datasets desbalanceados.

Los resultados obtenidos son útiles como herramienta de prevención para profesionales de la salud, siempre acompañados de evaluación clínica.

La ciencia de datos, aplicada de forma cuidadosa, puede ser un complemento valioso en la medicina preventiva y el diagnóstico temprano de enfermedades graves como el ACV.