Informe Final: Predicción de Accidentes Cerebrovasculares

Víctor De Marco **Alejandro Diez**

Índice

1. Objetivo y contexto del proyecto

El presente proyecto tiene como objetivo desarrollar un modelo de aprendizaje automático que permita predecir si una persona tiene riesgo de sufrir un accidente cerebrovascular (ACV), utilizando variables médicas y demográficas. Este problema es relevante por su impacto en la salud pública y por la posibilidad de aplicar intervenciones preventivas si se detecta riesgo a tiempo. El trabajo se apoya en el conjunto de datos 'Stroke Prediction Dataset' disponible en Kaggle.

2. Metodología seguida en cada fase

Fase 1: Análisis Exploratorio de Datos (EDA)

- Se cargaron los datos con pandas y se inspeccionaron valores nulos, tipos de variables y distribución.

- Visualizaciones con matplotlib y seaborn: histogramas, boxplots, countplots y scatterplots, para así dar un soporte visual a toda esa informacion y que se entienda mejor lo que se está haciendo en cada momento.

- Se analizó la correlación entre variables numéricas y la variable objetivo, mostrando la matriz de correlación correspondiente.

Fase 2: Preprocesamiento de Datos

- Transformación de las variables nulas por la media de la variable.

- Codificación One-Hot para variables categóricas.

- Normalización con StandardScaler.

- Creación de un pipeline para el procesamiento de los datos.

Fase 3: Modelado y Evaluación

- Los modelos elegidos para evaluar y entrenar fueron: Regresión Logística, Árboles, Random Forest y SVM.

- Se evaluaron con validación cruzada y las métricas elegidas para evaluar su desempeño fueron: precisión, recall, F1-score, AUC-ROC.

Fase 4: Ajuste de Hiperparámetros

-Se establecen los modelos y parámetros a buscar en ellos.

- Se aplicó GridSearchCV para identificar los parámetros óptimos de cada modelo.

- Se comparan los modelos antes y después de ajustar.

-Se llega a la conclusión de que el mejor modelo según las métricas obtenidas es el Random Forest después de ajustar.

-Se muestran unos gráficos con información, la curva Roc y la matriz de confusión, de cada modelo con y sin los hiperparámetros.

Fase 5: Técnicas Avanzadas

-Se utiliza Smote en los modelos anteriores.

-Se comparan las métricas de los modelos antes y después de aplicarles Smote.

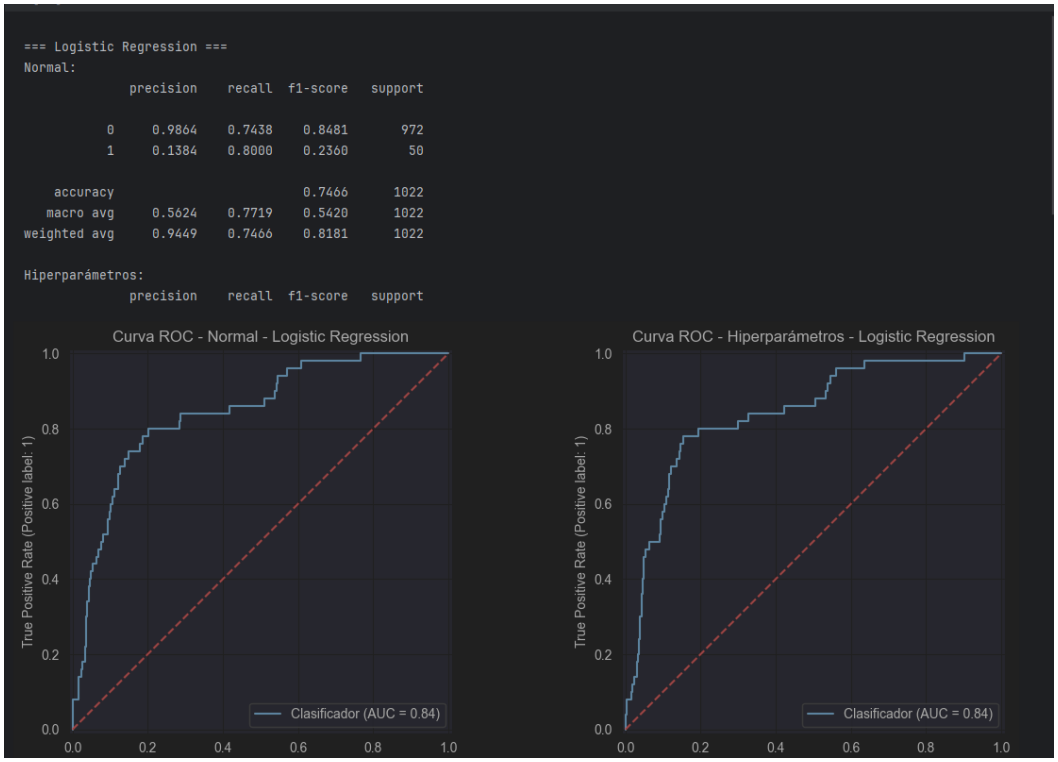
-Se prueba a utilizar de forma conjunta el clasificador Gradiente Boosting y Smote.

3. Principales hallazgos y resultados

- Edad, nivel de glucosa y BMI son variables predictivas claves para determinar si el paciente tiene riego de sufrir un ACV.

- El mejor modelo obtenido es el Random Forest despues de ajustar.

-Para observar mejor los resultados estadísticos y gráficos referentes a los modelos entrenados se recomienda visitar el apartado Graficos en el archivo Jupyter.



-Se muestra una captura de lo que puedes encontrar en el apartado anteriormente mencionado, datos estadísticos de los modelos como la precisión, el recall y su f1 score o graficos como su matriz de confusión/curva Roc.

4. Comparación de modelos y justificación final

Comparación de modelos:

	Accuracy (Normal)	Accuracy (Ajustado)	Precision (Normal)	Precision (Ajustado)
Random Forest	0.951076	0.849070	0.400000	0.173618
Logistic Regression	0.739239	0.730432	0.133078	0.132925
Decision Tree	0.917316	0.805519	0.126850	0.138023
SVM (Support Vector Machine)	0.748042	0.725785	0.121322	0.128400

Recall (Normal)	Recall (Ajustado)	F1 (Normal)	F1 (Ajustado)	Roc_auc (Normal)	Roc_auc (Ajustado)
0.015000	0.562692	0.028804	0.264953	0.787603	0.839222
0.788974	0.819231	0.227713	0.228669	0.838741	0.839859
0.115256	0.562179	0.120540	0.221455	0.536801	0.702821
0.668333	0.799103	0.205288	0.221229	0.795066	0.838249

- Aquí se puede observar las métricas obtenidas de los modelos, destacar el aumento del f1-score del modelo Random Forest después de ser ajustado o el aumento general del Recall en todos los modelos después del ajuste.

Random Forest fue elegido el mejor modelo debido a su buen desempeño antes y después de ser ajustado, su alto valor de precisión y accuracy y su buen f1-score tras ser ajustado.

5. Dificultades y soluciones realizadas

- Durante el desarrollo del trabajo se encontraron distintas dificultades tales como el tratamiento de los valores faltantes o el correcto ajuste de hiperparámetros para obtener el mejor modelo posible de todos los candidatos disponibles. Aun así, dichas dificultades se solucionaron sin demasiados problemas y así lo demuestran los resultados obtenidos en este trabajo.

6. Posibles mejoras o líneas de trabajo futuras

- Usar modelos avanzados como XGBoost o LightGBM.

- Implementar pipelines con validación incluida.

- Desplegar un sistema interactivo basado en el modelo.

- Incluir más variables clínicas si están disponibles.