Informe del Modelo de Clasificación – Detección de Stroke

1. Introducción

Este informe describe el desarrollo de un modelo de clasificación orientado a predecir el riesgo de sufrir un stroke (accidente cerebrovascular) a partir de datos de salud. A su vez este proyecto puede dividirse en dos partes: un primer algoritmo basado en machine learning “clásico”, que a través de variables como BMI y Age puede predecir el riesgo de Stroke; y un modelo de computer visión que nos permite inferir el riesgo en función de tomografías. Para aumentar la robustez del modelo de machine learning, se fusionaron dos datasets distintos con información complementaria. Esta estrategia permitió incrementar el número de muestras y enriquecer la diversidad de los registros utilizados para el entrenamiento y prueba.

3. Algoritmo de Machine Learning

3.1. Descripción de los Datasets

Dataset 1

Variable	Descripción
HeartDiseaseorAttack	Indicador de si la persona ha tenido enfermedad cardíaca o ataque al corazón. (0 = No, 1 = Sí)
HighBP	Hipertensión diagnosticada. (0 = No, 1 = Sí)
HighChol	Colesterol alto diagnosticado. (0 = No, 1 = Sí)
CholCheck	Ha tenido un chequeo de colesterol en los últimos 5 años. (0 = No, 1 = Sí)
BMI	Índice de Masa Corporal (peso en relación a la altura).
Smoker	Ha fumado al menos 100 cigarrillos en su vida. (0 = No, 1 = Sí)
Stroke	Ha sufrido un accidente cerebrovascular. (0 = No, 1 = Sí)
Diabetes	Diagnóstico de diabetes. Puede estar codificado como 0 (no), 1 (sí), 2 (prediabetes) u otro.
PhysActivity	Participa en actividad física en el último mes, fuera del trabajo. (0 = No, 1 = Sí)
Fruits	Consume fruta al menos una vez al día. (0 = No, 1 = Sí)
Veggies	Consume vegetales al menos una vez al día. (0 = No, 1 = Sí)

HvyAlcoholConsump	Consumo excesivo de alcohol (más de 14 bebidas por semana para hombres o más de 7 para mujeres).
AnyHealthcare	Tiene acceso a algún tipo de cobertura de salud. (0 = No, 1 = Sí)
NoDocbcCost	No pudo ver al médico en los últimos 12 meses debido al costo. (0 = No, 1 = Sí)
GenHlth	Salud general autoinformada (1 = Excelente, 5 = Mala).
MentHlth	Número de días con mala salud mental en los últimos 30 días.
PhysHlth	Número de días con mala salud física en los últimos 30 días
DiffWalk	Dificultad para caminar o subir escaleras. (0 = No, 1 = Sí)
Sex	Sexo biológico. (0 = Mujer, 1 = Hombre)
Age	Grupo de edad codificado (p.ej. 1 = 18–24, 2 = 25–29, ..., 13 = 80+)
Education	Nivel educativo. (1 = Menos de secundaria, 6 = Universidad completa)
Income	Nivel de ingresos. (1 = <$10k, ..., 8 = >$75k)

Dataset 2

Variable	Descripción
gender	Género de la persona (Male, Female, Other)
age	Edad numérica en años
hypertension	Hipertensión diagnosticada. (0 = No, 1 = Sí)
heart_disease	Presencia de enfermedad cardíaca . (0 = No, 1 = Sí)
ever_married	Estado civil. (Yes = Casado/a alguna vez, No = Nunca casado/a)
work_type	Tipo de ocupación (Private, Self-employed, Govt job, children, Never worked).
Residence_type	Zona de residencia (Urban o Rural)
avg_glucose_level	Nivel promedio de glucosa en sangre.
bmi	Índice de Masa Corporal (peso en relación a la altura).
smoking_status	Estado de fumador (never smoked, formerly smoked, smokes, unknown).
stroke	Ha sufrido un accidente cerebrovascular. (0 = No, 1 = Sí)

Dataset unido

Variable	Descripción
Sex	Género de la persona (Male, Female, Other)
Age	Edad numérica en años
HighBP	Hipertensión diagnosticada. (0 = No, 1 = Sí)
HeartDiseaseorAttack	Presencia de enfermedad cardíaca . (0 = No, 1 = Sí)
BMI	Índice de Masa Corporal (peso en relación a la altura).
Smoker	Estado de fumador (0 = No, 1=Sí).
Stroke	Ha sufrido un accidente cerebrovascular. (0 = No, 1 = Sí)

3.2. Selección y Optimización del Modelo

Se probaron diversos clasificadores como:

- Random Forest
- XGBoost
- Gradient Boosting
- Naive Bayes
- K-Nearest Neighbors (KNN)
- LGBMClassifier
- RidgeClassifier

El mejor rendimiento se obtuvo utilizando Gradient Boosting, que mostró un balance sólido entre precisión y capacidad de generalización.

3.3. Hiperparámetros Óptimos del Modelo

-colsample_bytree=1
-eval_metric='logloss'
-gamma=0
-learning_rate=1
-max_depth=16
-n_estimators=125
-random_state=42
-reg_alpha=0.1
-reg_lambda=1
-subsample=1
-use_label_encoder=False

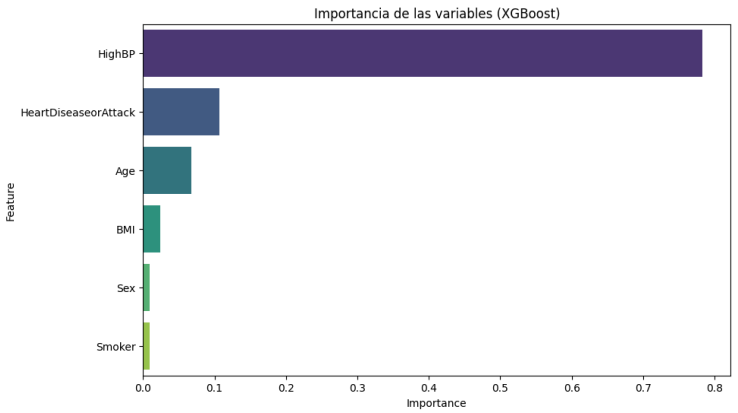
3.4 . Métricas del modelo

Matriz de confusión

	precision	recall	f1-score	support
0.0	0.94	0.91	0.92	947
1.0	0.96	0.97	0.97	2108
accuracy			0.95	3055
macro avg	0.95	0.94	0.94	3055
weighted avg	0.95	0.95	0.95	3055

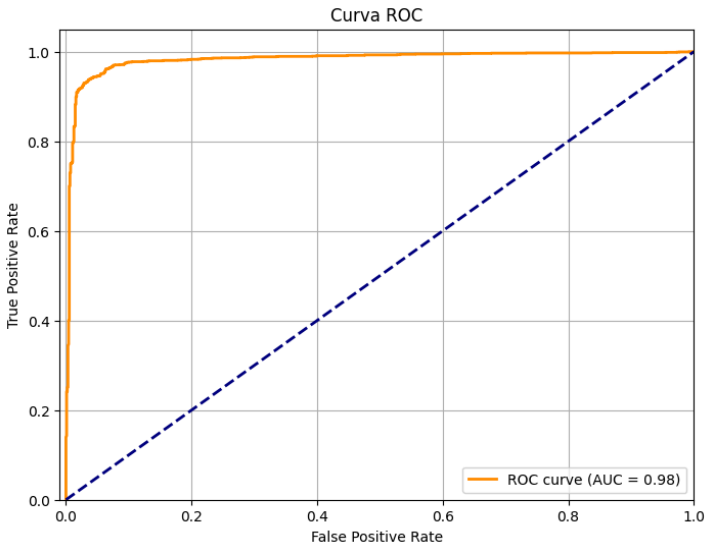
El modelo presenta un rendimiento general sólido, con una **exactitud del 95%**. La clase positiva (stroke) obtiene una **precisión del 96%** y un **recall del 97%**, lo que indica una alta capacidad para detectar correctamente casos reales de stroke. El **F1-score ponderado es 0.95**, lo que refleja un buen equilibrio entre precisión y sensibilidad en un conjunto de datos ligeramente desbalanceado. Sin embargo, se observa una leve disminución en el recall para la clase negativa (0.91), lo que sugiere la presencia de algunos falsos positivos.

Importancia de las variables



La variable más influyente en la predicción de *stroke* es HighBP (hipertensión), con una importancia de 0.78, lo que indica que tiene un peso significativamente mayor que el resto. Le siguen HeartDiseaseorAttack (antecedentes cardíacos) con 0.11 y Age (edad) con 0.067. Variables como BMI, Sex y Smoker tienen una contribución marginal al modelo, con valores por debajo de 0.03. Esto sugiere que la hipertensión es el factor clave en las predicciones realizadas por el modelo.

Curva ROC



El modelo presenta un excelente poder discriminativo, con un AUC de 0.98 en la curva ROC. Esto refleja una alta capacidad del modelo para diferenciar entre casos positivos y negativos de *stroke*, lo cual es fundamental en aplicaciones clínicas donde se busca minimizar errores en el diagnóstico.

4. Algoritmo de Redes Neuronales

4.1 Selección y Optimización del Modelo

Para esta tarea de clasificación de imágenes se utilizó una red neuronal convolucional basada en la arquitectura ResNet50, un modelo profundo ampliamente validado en el ámbito de la visión por computador. Se optó por utilizar una versión preentrenada en ImageNet, aplicando transfer learning para aprovechar el conocimiento previamente adquirido por la red en tareas de clasificación general. Posteriormente, se adaptó la capa de salida para ajustarse al número de clases del problema actual. Esta elección permitió obtener un modelo robusto, capaz de aprender características discriminativas relevantes incluso con un conjunto de datos limitado.

4.2 Hiperparámetros Óptimos del Modelo

Durante el entrenamiento del modelo se seleccionaron los siguientes hiperparámetros, ajustados para maximizar el rendimiento en el conjunto de validación:

-Tamaño de batch: 32
-Número de épocas: 20 (con parada anticipada si no hay mejora del F1-score)
-Tasa de aprendizaje inicial: 0.0001
-Optimizador: Adam
-Scheduler de tasa de aprendizaje: Reducción escalonada cada 5 épocas
-Función de pérdida: Cross Entropy Loss con pesos balanceados según la distribución de clases

-Técnicas de regularización: Dropout y data augmentation (rotaciones, cambios de color, flips)

Estos valores fueron determinados empíricamente tras varias iteraciones de experimentación, evaluando métricas como la precisión, recall y F1-score para asegurar un rendimiento equilibrado entre clases.

4.3. Métricas del modelo

Matriz de confusión

	precision	recall	f1-score	support
Bleeding	1	0.97	0.98	947
Ischemia	0.97	0.95	0.96	2108
Normal	0.96		0.97	3055
accuracy			0.97	1775
macro avg	0.98	0.97	0.97	1775
weight	0.97	0.97	0.97	1775

El modelo tiene un desempeño **excelente** para clasificar entre sangrado, isquemia y casos normales. Las precisiones y recalls cercanos o superiores al 95% indican que tiene muy pocos falsos positivos y falsos negativos en cada clase, lo que es crucial para aplicaciones médicas donde errores pueden ser costosos. La alta exactitud global (97.2%) refuerza esta conclusión.