**Comparación de Modelos de Machine Learning y Deep Learning para la Predicción de Accidentes Cerebrovasculares: Un Estudio Experimental**

Michael Rafael Ventura Bautista

Fernielis Vicioso Rodriguez

Julian Alejandro Núñez Toribio

Universidad Autonoma de Santo Domingo

Ciencia de Datos 2 | INF-8239

Dr. Silverio del Orbe

26 de octubre de 2025

**Resumen**

Este estudio presenta una comparación de diferentes técnicas de machine learning (ML) y deep learning (DL) para la predicción de accidentes cerebrovasculares (ACV), utilizando un conjunto de datos clínicos. Se preprocesaron los datos mediante la imputación de valores faltantes, balanceo de clases con SMOTE y la normalización de las variables. Se entrenaron múltiples modelos: Regresión Logística, Random Forest, Máquinas de Soporte Vectorial (SVM) y redes neuronales (una capa y múltiples capas). La evaluación de los modelos se realizó mediante métricas clave como accuracy, precision, recall, F1-score, y AUC, además de las curvas ROC y matrices de confusión. Los resultados mostraron que Random Forest y las Redes Neuronales de Múltiples Capas fueron los modelos con mejor rendimiento, especialmente en términos de AUC y F1-Score, indicando una capacidad superior para capturar la complejidad del problema. En contraste, la Regresión Logística mostró un rendimiento inferior, mientras que SVM presentó un desempeño intermedio. La mejora del modelo de redes neuronales de múltiples capas frente al de una sola capa fue notable, destacándose en su capacidad para representar patrones complejos. Estos hallazgos sugieren que las técnicas de Deep Learning ofrecen ventajas significativas en la predicción de ACV, aunque los modelos clásicos como Random Forest siguen siendo altamente efectivos.

Palabras clave: Predicción de ACV, Machine Learning, Deep Learning, Random Forest, Redes Neuronales, SVM, AUC, F1-Score, SMOTE, Regresión Logística.

**Contenido**

[Resumen 2](#_Toc212298590)

[Marco de Referencia 4](#_Toc212298591)

[Método 5](#_Toc212298592)

[Evaluación de los Modelos 7](#_Toc212298593)

[Resultados y Discusión 8](#_Toc212298594)

[Conclusiones 10](#_Toc212298595)

[Referencias 11](#_Toc212298596)

[Anexos 12](#_Toc212298597)

**Marco de Referencia**

Los accidentes cerebrovasculares (ACV) representan una de las principales causas de muerte y discapacidad a nivel mundial, lo que resalta la importancia de su predicción temprana para mejorar los tratamientos preventivos y médicos. Diversos enfoques de aprendizaje automático (machine learning) y aprendizaje profundo (deep learning) han sido aplicados para este tipo de problemas de clasificación, con el fin de identificar factores predictivos en los datos clínicos (Zhou et al., 2020). Modelos clásicos como Regresión Logística, Random Forest y SVM se han utilizado con éxito en clasificación binaria, mientras que las redes neuronales, especialmente las de múltiples capas, han ganado popularidad en la última década debido a su capacidad para capturar relaciones no lineales complejas (Goodfellow et al., 2016).

En cuanto a la técnica de preprocesamiento, el balanceo de clases es una etapa crucial en el análisis de datos desbalanceados, como ocurre en el caso de los ACV, donde los eventos de interés (ACV) son menos frecuentes que los de no-evento (Kubat & Matwin, 1997). El uso de SMOTE (Synthetic Minority Over-sampling Technique) ha demostrado ser eficaz para mejorar la representación de las clases minoritarias sin perder información relevante.

A pesar de los avances en las redes neuronales, los modelos clásicos siguen siendo útiles debido a su simplicidad y facilidad de interpretación, lo que los hace adecuados para tareas donde la explicabilidad es importante (Breiman, 2001). Sin embargo, el poder predictivo de las redes neuronales profundas sigue superando a los modelos tradicionales, como se ha observado en estudios recientes de clasificación médica (Ribeiro et al., 2016).

**Método**

**Datos**

El conjunto de datos utilizado en este estudio proviene de registros clínicos de pacientes con información sobre factores de riesgo para accidentes cerebrovasculares (<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset> ). Las variables incluyen edad, presión arterial, nivel de azúcar en sangre, índice de masa corporal (IMC), entre otros. El objetivo es predecir la presencia o ausencia de un accidente cerebrovascular (ACV).

**Pre procesamiento de los Datos**

El pre procesamiento incluyó:

**Imputación de valores faltantes**: Las variables con valores faltantes fueron imputadas utilizando la media (en el caso de variables continuas).

**Balanceo de clases**: Dado el desbalance en las clases (la clase de ACV es menos frecuente), se utilizó el algoritmo **SMOTE** para generar ejemplos sintéticos de la clase minoritaria.

**Codificación y normalización**: Las variables categóricas fueron convertidas en variables numéricas mediante técnicas de codificación y todas las variables numéricas fueron escaladas para mejorar la convergencia de los modelos.

**Modelos Evaluados**

**Regresión Logística:** Un modelo estadístico basado en una función sigmoide que modela la probabilidad de ocurrencia de un evento binario.

**Random Forest:** Un modelo basado en un ensamblaje de árboles de decisión que maneja bien relaciones no lineales.

**Máquinas de Soporte Vectorial (SVM):** Un modelo que busca el margen óptimo entre las clases para maximizar la separación.

**Redes Neuronales:**

* **1 capa:** Una red con una sola capa oculta.
* **Múltiples capas:** Una red con varias capas ocultas, diseñada para capturar patrones más complejos.

**Evaluación de los Modelos**

Los modelos fueron evaluados utilizando las siguientes métricas:

**Accuracy**: Proporción de predicciones correctas.

**Precision**: Proporción de predicciones positivas correctas respecto al total de predicciones positivas.

**Recall**: Proporción de predicciones positivas correctas respecto al total de casos positivos reales.

**F1-Score**: Promedio armónico de precision y recall, usado para balancear ambos en problemas desbalanceados.

**AUC (Área bajo la curva ROC)**: Mide la capacidad del modelo para discriminar entre las clases.

**Curvas ROC**: Representación gráfica de la tasa de falsos positivos frente a la tasa de verdaderos positivos.

**Resultados y Discusión**

**Rendimiento de los Modelos**

| **Modelo** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| --- | --- | --- | --- | --- | --- |
| Regresión Logística | 0.8730 | 0.8731 | 0.8722 | 0.8726 | 0.9469 |
| Random Forest | 0.9625 | 0.9648 | 0.9598 | 0.9623 | 0.9932 |
| SVM | 0.8992 | 0.8933 | 0.9062 | 0.8997 | 0.9590 |
| NN - 1 capa | 0.8895 | 0.9055 | 0.8691 | 0.8869 | 0.9635 |
| NN - Múltiples capas | 0.9260 | 0.9089 | 0.9464 | 0.9273 | 0.9801 |

Los resultados indican que Random Forest y Redes Neuronales de múltiples capas fueron los modelos con mejor desempeño general. Ambos modelos mostraron altos valores de AUC (0.9932 y 0.9801, respectivamente) y F1-Score (0.9623 y 0.9273, respectivamente), lo que sugiere que estos modelos capturan muy bien la complejidad del problema. En particular, el modelo de Random Forest alcanzó un AUC cercano al 1, lo que indica una capacidad excepcional para distinguir entre los casos de ACV y no-ACV.

El modelo de Redes Neuronales de múltiples capas también mostró una mejora significativa con respecto al modelo de 1 capa, con un AUC de 0.9801 frente a 0.9635, lo que refleja una mayor capacidad para modelar relaciones no lineales y patrones complejos en los datos. Esto resalta la importancia de utilizar redes neuronales profundas cuando se trata de problemas complejos como la predicción de enfermedades.

En comparación, la Regresión Logística mostró un desempeño inferior, con un AUC de 0.9469 y un F1-Score de 0.8726. Aunque este modelo es interpretativo y fácil de implementar, su capacidad predictiva es limitada en comparación con modelos más complejos. SVM, por otro lado, tuvo un desempeño intermedio, mostrando un AUC de 0.9590 y un F1-Score de 0.8997.

**Discusión**

Los modelos de Deep Learning (especialmente el de múltiples capas) demostraron una mayor capacidad para capturar patrones complejos, lo que es crucial en el ámbito médico, donde las relaciones entre las variables no siempre son lineales. Sin embargo, esto viene a costa de una mayor complejidad computacional, que podría ser un desafío en entornos con recursos limitados. Por otro lado, Random Forest sigue siendo una opción robusta y altamente efectiva, con la ventaja adicional de ser más fácil de interpretar que las redes neuronales.

**Limitaciones**

Una de las limitaciones de este estudio es la falta de un análisis más profundo de la interpretabilidad de los modelos, especialmente en las redes neuronales, donde la "caja negra" es un desafío. Además, el tiempo de entrenamiento para los modelos de Deep Learning fue considerablemente más largo que para los modelos clásicos.

**Conclusiones**

Los resultados de este estudio sugieren que los modelos de Deep Learning, específicamente las redes neuronales de múltiples capas, superan a los modelos clásicos en términos de rendimiento predictivo, especialmente en métricas clave como AUC y F1-Score. Sin embargo, modelos clásicos como Random Forest ofrecen un rendimiento competitivo y son más fáciles de interpretar, lo que los convierte en una opción adecuada en escenarios donde la interpretabilidad es crucial. Se recomienda el uso de Redes Neuronales de Múltiples Capas para maximizar la precisión en la predicción de ACV, especialmente en entornos con recursos computacionales adecuados.

# **Referencias**

Breiman, L. (2001). *Random forests. Machine Learning.*

Goodfellow, I. B. (2016). *Deep learning.* MIT Press.

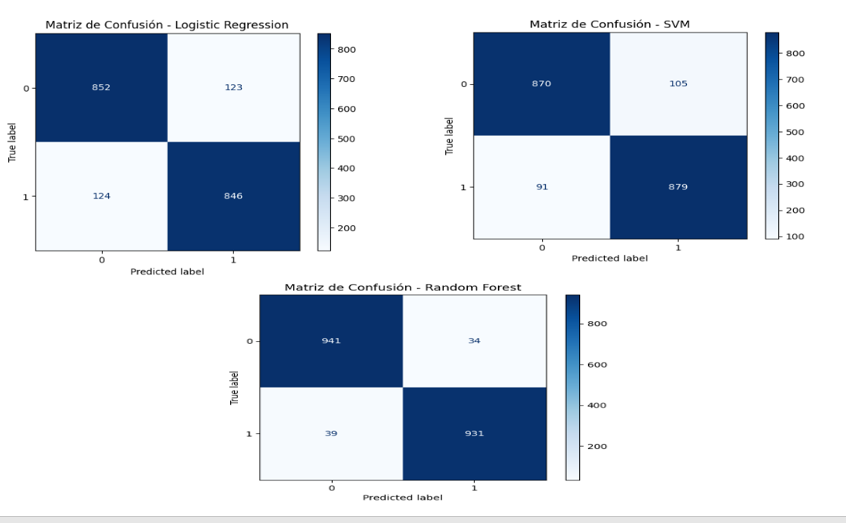
Kubat, M. &. (1997). *Addressing the curse of imbalanced training sets: One-sided selection. ICML.*

Ribeiro, M. T. (2016). *"Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

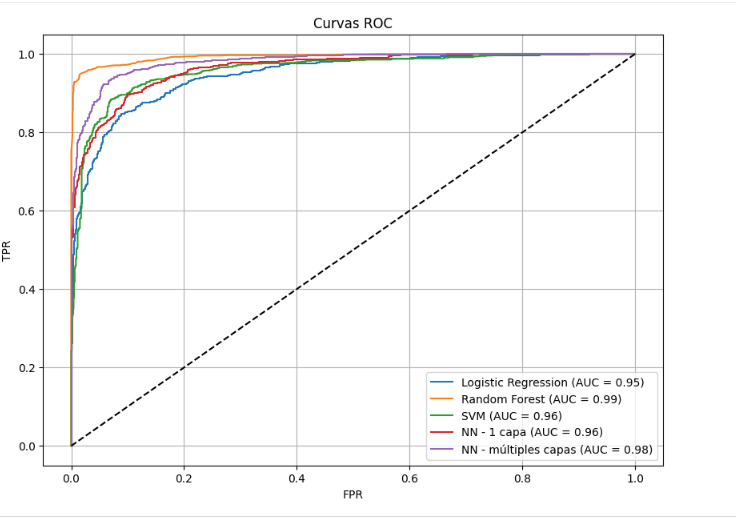
Zhou, J. e. (2020). *Predicting stroke risk using machine learning techniques: A review. Journal of Stroke and Cerebrovascular Diseases.*

**Anexos**

**Anexo A:** Matriz Confusion



**Anexo B:** Grafica las curvas ROC (Receiver Operating Characteristic)



**Anexo C:** Informe detallado metricas obtenidas

--- Logistic Regression ---

precision recall f1-score support

0 0.87 0.87 0.87 975

1 0.87 0.87 0.87 970

accuracy 0.87 1945

macro avg 0.87 0.87 0.87 1945

weighted avg 0.87 0.87 0.87 1945

--- Random Forest ---

precision recall f1-score support

0 0.96 0.97 0.96 975

1 0.96 0.96 0.96 970

accuracy 0.96 1945

macro avg 0.96 0.96 0.96 1945

weighted avg 0.96 0.96 0.96 1945

--- SVM ---

precision recall f1-score support

0 0.91 0.89 0.90 975

1 0.89 0.91 0.90 970

accuracy 0.90 1945

macro avg 0.90 0.90 0.90 1945

weighted avg 0.90 0.90 0.90 1945