Predicción de accidentes cerebrovasculares utilizando redes neuronales y máquinas de soporte vectorial

Huaccharaqui Fernando
Facultad de Ingeniería de Sistemas
Universidad ESAN
Lima, Perú
18100209@ue.edu.pe

***Abstract*— El presente trabajo de investigación se centra en utilizar redes neuronales y máquinas de soporte vectorial para predecir la ocurrencia de accidentes cerebrovasculares utilizando el conjunto de datos de Cerebral Stroke Prediction Dataset de Kaggle. El estudio tiene como objetivo comparar el rendimiento de estos dos modelos de aprendizaje automático supervisado para determinar el enfoque óptimo en la predicción de accidentes cerebrovasculares. Al aprovechar el poder de estos modelos, se pueden obtener ideas valiosas que permitan la detección temprana e intervención para mitigar el riesgo de accidentes cerebrovasculares. Los hallazgos de este estudio tienen el potencial de contribuir al campo de la salud al proporcionar una herramienta confiable y eficiente para la predicción de accidentes cerebrovasculares, lo que finalmente conduce a una mejora en la atención y los resultados para los pacientes.**

Keywords*— *accidentes cerebrovasculares, predicción, redes neuronales, máquinas de soporte vectorial, dataset de* Cerebral Stroke *en Kaggle.

I. INTRODUCCIÓN

Los accidentes cerebrovasculares (ACV) representan una de las principales causas de muerte y discapacidad en todo el mundo. Según la Organización Mundial de la Salud, cada año ocurren 15 millones de ACV a nivel mundial, de los cuales 5 millones resultan en muertes y otros 5 millones en discapacidad permanente [1]. En el Perú, según el Instituto Nacional de Salud (INS), el ACV es la segunda causa de muerte a nivel nacional, con una tasa de mortalidad de 39,9 por cada 100.000 habitantes [2].

Los ACV pueden ser prevenibles a través de la modificación de factores de riesgo como la hipertensión, el tabaquismo, la diabetes y la obesidad. Sin embargo, la identificación temprana de estos factores de riesgo y la implementación de intervenciones de prevención efectivas a menudo se ven obstaculizadas por la falta de acceso a la atención médica, la falta de conciencia sobre los síntomas del ACV y la falta de comprensión de los factores de riesgo.

En este contexto, las técnicas de aprendizaje automático pueden desempeñar un papel crucial en la identificación temprana de personas en riesgo de sufrir un ACV. Diversos estudios han demostrado la eficacia de las técnicas de aprendizaje automático en la predicción de enfermedades cardiovasculares. Por ejemplo, Deo [3] discutió cómo el aprendizaje automático puede ser utilizado para predecir el riesgo de enfermedad cardiovascular, destacando la importancia de las técnicas de aprendizaje automático en el campo de la medicina predictiva.

En este trabajo, nuestro objetivo es construir un modelo de aprendizaje automático que pueda predecir con precisión la ocurrencia de un ACV. Para ello, utilizaremos un conjunto de datos de pacientes que incluye información demográfica, historial médico y factores de estilo de vida. A través de un análisis exhaustivo de estos datos y la aplicación de técnicas avanzadas de aprendizaje automático, esperamos desarrollar un modelo que pueda identificar a las personas en riesgo de sufrir un ACV, permitiendo así intervenciones tempranas y potencialmente salvadoras.

II. MARCO TEÓRICO

A. Algoritmos de predicción utilizados

1) Redes neuronales artificiales

Las Redes Neuronales Artificiales (RNA) son sistemas de cálculo inspirados en el funcionamiento del cerebro humano y han demostrado ser herramientas muy útiles en la clasificación y predicción de eventos complejos y no lineales. Los principales componentes de una RNA incluyen neuronas (o nodos), capas y conexiones ponderadas. Las neuronas se agrupan en capas y las conexiones entre las neuronas de diferentes capas se ponderan según su relevancia para la predicción

2) Máquinas de Vectores de Soporte (SVM)

Las Máquinas de Vectores de Soporte (SVM) son un conjunto de algoritmos de aprendizaje supervisado que se utilizan para clasificación y regresión. Dada un conjunto de ejemplos de entrenamiento, cada uno marcado como perteneciente a una de dos categorías, un algoritmo de entrenamiento SVM construye un modelo que asigna nuevos ejemplos a una categoría u otra. Un modelo SVM es una representación de los ejemplos como puntos en el espacio, mapeados de manera que los ejemplos de las categorías separadas están divididos por un espacio claro que es tan amplio como sea posible.

3) Técnica de Sobremuestreo de Minorías Sintéticas (SMOTE)

La Técnica de Sobre muestreo de Minorías Sintéticas (SMOTE) es un enfoque para tratar los conjuntos de datos desequilibrados mediante la generación de instancias sintéticas de la clase minoritaria. Este enfoque puede mejorar el rendimiento de los modelos de aprendizaje automático en conjuntos de datos desequilibrados al proporcionar más ejemplos de la clase minoritaria para el entrenamiento [4].

B. Principales indicadores de rendimiento

1) Exactitud (Accuracy)

La precisión es la relación entre las predicciones correctas y el total de predicciones. Es un indicador de rendimiento comúnmente utilizado en la clasificación binaria.

2) Sensibilidad (Recall)

La sensibilidad, también conocida como tasa de verdaderos positivos o recall, es la proporción de positivos reales que se identifican correctamente. En el contexto de la predicción de ACV, esto sería la proporción de personas que realmente tienen un ACV y que el modelo identifica correctamente.

3) Precisión (Precision)

La precisión, en el contexto de los indicadores de rendimiento, es la proporción de identificaciones positivas que fueron realmente correctas. En el caso de la predicción de ACV, sería la proporción de personas a las que el modelo predijo que tendrían un ACV y que realmente lo tuvieron.

4) Área bajo la curva ROC (AUC-ROC)

El AUC-ROC es una medida de rendimiento para los problemas de clasificación en diferentes umbrales de clasificación. AUC representa el grado o medida de separabilidad, es decir, cuánto el modelo es capaz de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo en la predicción de 0s como 0s y 1s como 1s [5].

III. MATERIALES Y MÉTODOS

A. Base de datos

La base de datos utilizada para predecir la incidencia de accidentes cerebrovasculares se obtuvo de Kaggle. Los datos contienen información de salud y estilo de vida de individuos, con el objetivo de predecir la probabilidad de que ocurran accidentes cerebrovasculares. Los datos contienen las siguientes columnas:

- Género: Género del paciente (Masculino y Femenino).
- Edad: Edad del paciente.
- Hipertensión: Si el paciente tiene o no hipertensión.
- Enfermedad del corazón: Si el paciente tiene o no una enfermedad del corazón.
- Casado: Si el paciente está casado o no.
- Tipo de trabajo: Tipo de trabajo que realiza el paciente.
- Tipo de residencia: Tipo de residencia del paciente.
- Nivel medio de glucosa en sangre: Nivel medio de glucosa en la sangre del paciente.
- IMC: Índice de masa corporal del paciente.
- Estado de fumador: Estado de fumador del paciente.
- Accidente cerebrovascular: Si el paciente ha tenido o no un accidente cerebrovascular.

B. Visualización de datos

Se generaron varias visualizaciones para entender mejor los datos. Estas visualizaciones incluyeron:

- Histogramas que muestran la distribución de las variables 'age', 'avg_glucose_level', y 'bmi'.
- Una matriz de correlación para visualizar la correlación entre las diferentes características.

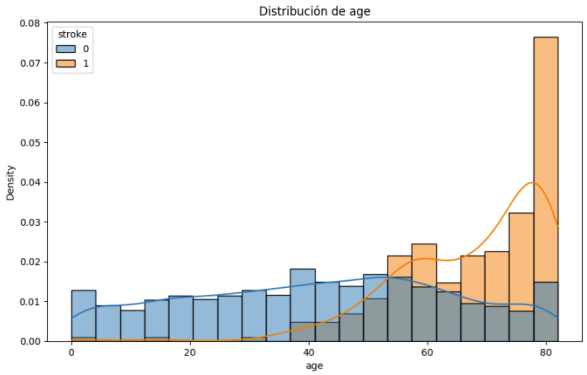

Figura 1. Distribución de la Edad

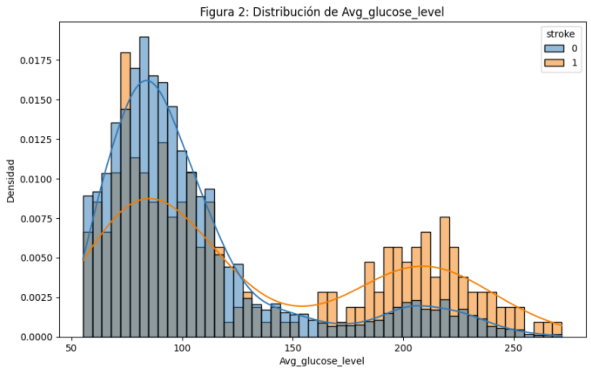

Figura 2. Distribución del Nivel Promedio de Glucosa

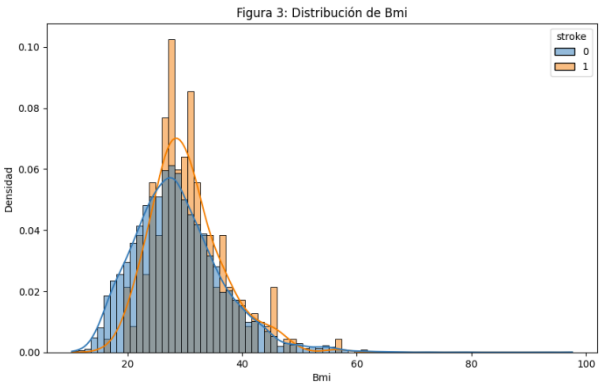

Figura 3. Distribución del Índice de Masa Corporal (IMC)

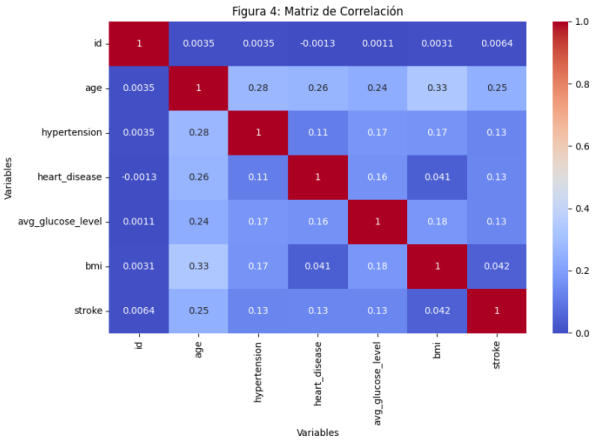

Figura 4. Matriz de Correlación

C. Metodología

1) Aplicación del modelo Support Vector Machines (SVM)

Para aplicar el modelo SVM a los datos de la investigación, se realizaron los siguientes pasos:

- Se codificaron las variables categóricas en columnas numéricas, asignando un número a cada variable de 'género', 'ever_married', 'work_type', 'Residence_type', y 'smoking_status'.
- Se llenaron los valores faltantes en la columna 'bmi' con la media de esa columna.
- Se dividieron los datos en conjuntos de entrenamiento y prueba, asignando el 20% de los datos al conjunto de prueba.
- Se aplicó la técnica de sobremuestreo SMOTE solo en los datos de entrenamiento para manejar el desequilibrio de clases en la variable objetivo.
- Se entrenó el modelo SVM con los datos de entrenamiento y se realizaron predicciones en el conjunto de prueba.
- Se calculó el área bajo la curva ROC (AUC) y se trazó la curva ROC.

2) Aplicación del modelo de Redes Neuronales Artificiales (RNA)

El modelo de RNA también se aplicó a los datos de la investigación, siguiendo un proceso similar al del modelo SVM. Sin embargo, antes de entrenar el modelo, se realizó una búsqueda en cuadrícula para optimizar los parámetros del modelo. Los parámetros considerados incluyeron el número de capas ocultas, la función de activación, el solucionador, el coeficiente de regularización y la tasa de aprendizaje.

Después de entrenar el modelo con los parámetros óptimos, se realizaron predicciones en el conjunto de prueba y se calculó el AUC y se trazó la curva ROC, al igual que con el modelo SVM.

Finalmente, se compararon los resultados de los dos modelos y se visualizó la comparación mediante una tabla y una gráfica de la curva ROC.

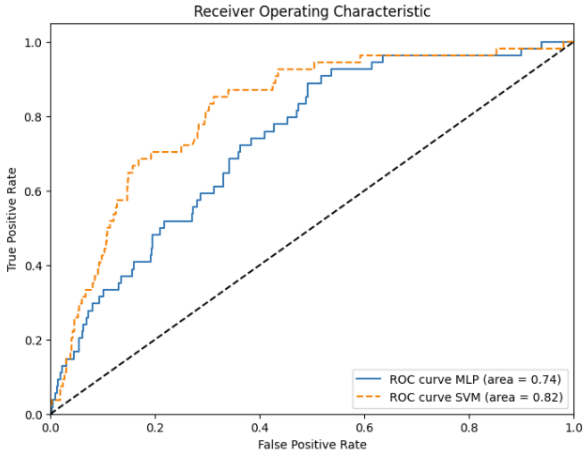

Figura 5. Curva ROC.

IV. RESULTADOS Y DEBATE

A. Resultados de predicción y de pruebas de rendimiento del modelo Support Vector Machines (SVM)

El modelo SVM fue entrenado y probado para predecir la probabilidad de que un individuo experimente un accidente cerebrovascular. Las métricas de rendimiento del modelo incluyen la exactitud (accuracy), la sensibilidad (recall) y la precisión (precision). Los resultados se presentan en la Tabla 1.

TABLA 1
RESULTADOS DE PREDICCIÓN UTILIZANDO EL MODELO SVM

Accuracy	Recall	Precisión
0.711	0.778	0.129

Además, el área bajo la curva ROC (AUC) para el modelo SVM fue de 0.82, lo que indica un buen rendimiento del modelo en la clasificación de los individuos que han sufrido un accidente cerebrovascular.

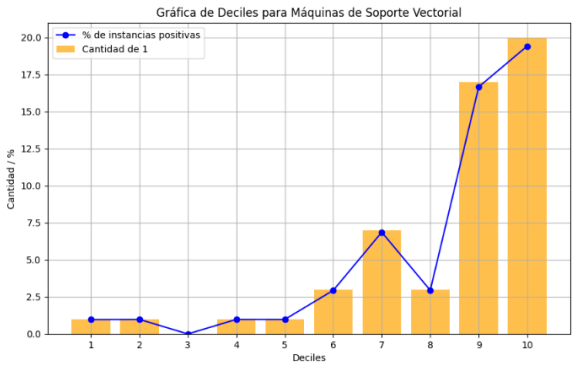

Figura6. Gráfico de deciles SVM

B. Resultados de predicción y de pruebas de rendimiento del modelo Redes Neuronales Artificiales (RNA)

Similarmente, el modelo de RNA fue entrenado y probado. Las métricas de rendimiento del modelo se presentan en la Tabla 2.

TABLA 2
RESULTADOS DE PREDICCIÓN UTILIZANDO EL MODELO RNA

Accuracy	Recall	Precisión
0.752	0.559	0.110

El área bajo la curva ROC (AUC) para el modelo RNA fue de 0.74, lo que indica un buen rendimiento del modelo, aunque ligeramente inferior al del modelo SVM.

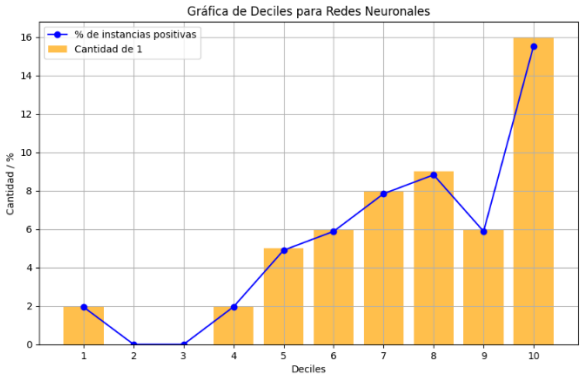

Figura6. Gráfico de deciles RNA

C. Decisión sobre el mejor modelo de predicción

Comparando las métricas de rendimiento y el AUC de ambos modelos, se puede concluir que, aunque el modelo RNA tiene una accuracy ligeramente mayor, el modelo SVM tiene una mejor sensibilidad y un AUC mayor. Por lo tanto, el modelo SVM podría ser más adecuado para este problema particular, ya que es crucial identificar correctamente a los individuos que tienen un alto riesgo de sufrir un accidente cerebrovascular.

V. CONCLUSIONES

En este estudio, se compararon dos modelos de aprendizaje automático, SVM y RNA, para predecir la probabilidad de que un individuo sufra un accidente cerebrovascular basándose en una serie de características de salud y estilo de vida.

Los resultados indican que el modelo SVM podría ser más adecuado para este problema, ya que muestra un mejor rendimiento en términos de sensibilidad y AUC. Aunque el modelo RNA tiene una accuracy ligeramente superior, la capacidad de identificar correctamente a los individuos que tienen un alto riesgo de sufrir un accidente cerebrovascular es crucial, y en este aspecto el modelo SVM supera al modelo RNA.

Estos hallazgos demuestran el potencial de los modelos de aprendizaje automático en la identificación de individuos en riesgo de sufrir accidentes cerebrovasculares, lo que podría permitir intervenciones tempranas y personalizadas para prevenir estos eventos. Sin embargo, se necesitan más investigaciones para optimizar estos modelos y validar sus predicciones en diferentes poblaciones y contextos.

VI. REFERENCIA BIBLIOGRÁFICAS

[1]Organización Mundial de la Salud. (2021). Las 10 principales causas de muerte. [En línea] Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[2] Instituto Nacional de Salud. (2021). Mortalidad en el Perú 2020. [En línea] Disponible en: http://www.dge.gob.pe/portal/docs/tools/boletin/2021/02.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.