ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

MODELO PREDICTIVO PARA LA PREVENCIÓN DE ACCIDENTES CEREBROVASCULARES

Profesor
Manuel Eugenio Morocho

Autores
Carlos Calán Juna
Selena Pozo Barahona

2024

RESUMEN

El objetivo de este estudio fue aplicar y evaluar diferentes algoritmos de Machine Learning para la predicción de un ACV, con algunas métricas de predicción como la accuracy, ROC AUC (área bajo la curva ROC), precisión, recall (sensibilidad) y F1 Score. El diseño de la investigación tiene un corte transversal, correlacional y predictivo. Este enfoque permite recopilar datos en un único punto temporal, analizar las relaciones entre diversas variables, y desarrollar modelos que predicen la probabilidad de ocurrencia de un ACV. El conjunto de datos se obtuvo del repositorio digital Kaggle, esta consta de 5,110 registros y 12 variables.

El estudio utiliza una metodología de análisis predictivo de datos que comprende varias fases clave para el desarrollo del proyecto, entre ellas la recolección de los datos, la limpieza, el análisis EDA y la selección y preparación de variables. Posteriormente, se aplican modelos de Machine Learning, como regresión logística, random forest y redes neuronales, para construir y evaluar estos modelos, utilizando Python y algunas librerías como Pandas, Seaborn, Matplotlib y NumPy. Finalmente, se realizó una evaluación de los resultados de cada modelo, donde se determinó que el que más se ajusta a nuestro objetivo es la regresión logística.

ABSTRACT

The objective of this study was to apply and evaluate different Machine Learning algorithms for the prediction of a stroke, with some prediction metrics such as accuracy, ROC AUC (area under the ROC curve), precision, recall (sensitivity) and F1 Score. The research design has a cross-sectional, correlational and predictive design. This approach allows you to collect data at a single time point, analyze the relationships between various variables, and develop models that predict the probability of stroke occurrence. The data set was obtained from the Kaggle digital repository, it consists of 5,110 records and 12 variables.

The study uses a predictive data analysis methodology that includes several key phases for the development of the project, including data collection, cleaning, EDA analysis, and the selection and preparation of variables. Subsequently, Machine Learning models, such as logistic regression, random forest and neural networks, are applied to build and evaluate these models, using Python and some libraries such as Pandas, Seaborn, Matplotlib and NumPy. Finally, an evaluation of the results of each model was carried out, where it was determined that the one that best fits our objective is logistic regression.

ÍNDICE DE CONTENIDO

RESUMEN 2
ABSTRACT 3
INTRODUCCIÓN 1
REVISIÓN DE LITERATURA 2
ANTECEDENTES DE LA INVESTIGACIÓN 2
IDENTIFICACIÓN DEL OBJETO DE ESTUDIO 8
PLANTEAMIENTO DEL PROBLEMA 11
OBJETIVO GENERAL 12
OBJETIVOS ESPECÍFICOS 12
JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA 13
SELECCIÓN DE LA BASE DE DATOS 13
LIMPIEZA, PREPROCESAMIENTO Y/O TRANSFORMACIÓN DE DATOS 14
IDENTIFICACIÓN DE VARIABLES 17
DEFINICIÓN DE VARIABLES 19
DESCRIPCIÓN DE VARIABLES 21
VISUALIZACIÓN DE VARIABLES 24
SELECCIÓN DEL MODELO ESTADÍSTICO 25
RESULTADOS 28
ANÁLISIS DE LOS MODELOS ESTADÍSTICOS E INTERPRETACIÓN DE RESULTADOS 28
DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN 35
CONCLUSIONES 40
RECOMENDACIONES 41
REFERENCIAS 42

ÍNDICE DE TABLAS

Tabla 1 Matriz de investigaciones similares 6
Tabla 2 Diccionario de variables..... 18
Tabla 3 Intervalos del índice de masa corporal (IMC) 19
Tabla 4 Niveles normales de azúcar en sangre antes de las comidas 20
Tabla 5 Niveles normales de azúcar en sangre después de las comidas 20
Tabla 6 Niveles normales de azúcar al acostarse 21
Tabla 7 Matriz de resumen estadístico de variables numéricas 29
Tabla 8 Resultados de los modelos de aprendizaje supervisado 29
Tabla 9 Matriz de confusión – Regresión Logística 30
Tabla 10 Matriz de confusión – Random Forest 30
Tabla 11 Matriz de confusión – Redes Neuronales 30
Tabla 12 Resultado de los modelos con ajuste de hiperparámetros 31
Tabla 13 Matriz de confusión con hiperparámetros – Regresión Logística 31
Tabla 14 Matriz de confusión con hiperparámetros – Random Forest 32
Tabla 15 Matriz de confusión con hiperparámetros – Redes Neuronales 32

ÍNDICE DE FIGURAS

Figura 1 Proceso de importación y manejo de las librerías de Python 15
Figura 2 Descripción general de las variables del conjunto de datos 15
Figura 3 Transformación del tipo de datos de las variables 16
Figura 4 Imputación de datos con la media..... 17
Figura 6 Diagrama de cajas de las variables numéricas..... 22
Figura 7 Distribución de variables continuas..... 23
Figura 8 Matriz de correlación de variables..... 23
Figura 9 Distribución de las variables cualitativas..... 24
Figura 10 Técnica SMOTE para el balanceo de clases 29
Figura 11 Matriz de confusión por modelo..... 30
Figura 12 Matriz de confusión por modelo con hiperparámetros 32
Figura 13 Evaluación de las variables predictoras..... 33

INTRODUCCIÓN

Los accidentes cerebrovasculares (ACV) son la primera causa de discapacidad y la tercera causa de mortalidad a nivel mundial (Olascoaga & Ascue, 2020), este se produce cuando el flujo sanguíneo se reduce o no llega al cerebro, impidiendo el paso de oxígeno y nutrientes, afectando directamente a las arterias que van hacia y dentro del cerebro (Ramos et al., 2024).

Según un estudio realizado por Avellán et al. (2022), los ACV se asocian al estilo de vida de las personas, al estrés personal y laboral, falta de actividad física, así como a trastornos de alimentación, siendo estos, factores modificables y/o mejorables en las personas. Su impacto no está ligado únicamente a quien lo padece, sino también a los familiares y personas cercanas al afectado, ya que la repercusión de un ACV puede ser difícil de sobrellevar debido al proceso de rehabilitación e inclusive de discapacidad que resulta de esta enfermedad. Además, de convertirse en una problemática de salud pública a nivel general, debido a sus altos porcentajes de padecimiento.

Es importante mencionar que existen factores que están ligados a una mayor posibilidad de que un paciente sufra o no un ACV, varios de estos factores se analizarán a lo largo de este estudio para identificar su nivel de relación a la hora de predecir un ACV. Entre ellos: la edad, el índice de masa corporal (IMC), el nivel de azúcar en sangre y el hábito de fumador de una persona.

En este contexto, la presente investigación hará una comparación de los algoritmos de Machine Learning más utilizados en la predicción de enfermedades, específicamente: regresión logística, random forest y redes neuronales. Para luego realizar un análisis del modelo que más se adapte con el objetivo del estudio, de acuerdo a distintas métricas estadísticas.

REVISIÓN DE LITERATURA

ANTECEDENTES DE LA INVESTIGACIÓN

El avance de la tecnología y la aplicación de técnicas de aprendizaje automático estadístico se ha diversificado a muchas áreas, en el sector médico específicamente, se han evidenciado estudios predictivos que buscan identificar métodos de prevención para las enfermedades más comunes en la población a nivel mundial, obteniendo casos de éxito gracias al alto porcentaje de precisión en las predicciones.

Campos et al. (2019), indican que el *Machine Learning* se ha convertido en un aliado muy importante en el control de enfermedades crónicas no transmisibles, pues ayuda a agilizar procesos y/o tareas que en el pasado requerían mucho tiempo y que por lo general estaban susceptibles a errores.

Dentro de las enfermedades crónicas no transmisibles (ENT), las enfermedades y/o accidentes cerebrovasculares constituyen una de las principales causas de mortalidad y discapacidad global. Siendo esta, la primera causa de discapacidad en el adulto y la tercera en la tasa de mortalidad en los países desarrollados (Yurieski Pérez et al., 2023).

Varios de estos aportes a la investigación se listan a continuación:

Según Olascoaga y Ascue (2020), los accidentes cerebrovasculares (ACV) afectan alrededor de 15 millones de personas, de estos más de 5 millones de personas fallecen y otros pocos sobreviven con secuelas irreversibles. Así lo mencionan en su obra, titulada: "*Desarrollo de un algoritmo con redes neuronales para la predicción de ACV en pacientes diabéticos"*, en esta obra los autores mediante la inteligencia artificial predictiva aplican el uso de redes neuronales, con el objetivo de predecir con el mínimo error, la probabilidad de que una persona pueda o no, sufrir un ACV, centrándose en la sensibilidad y la especificidad del modelo. La base de datos utilizada en este caso de estudio se obtuvo del repositorio Kaggle, el dataset cuenta con más de 17,372 registros y alrededor de 9 variables. Los resultados del análisis mostraron que el modelo incrementó la sensibilidad al 93% y la especificidad al 94%, ambos superando el rango mínimo estimado del 88% para estas variables. Los hallazgos en este

estudio confirman la efectividad del enfoque basado en redes neuronales para la predicción de ACV en pacientes diabéticos.

Por otra parte, es importante considerar todos los factores que pueden influir en un accidente cerebrovascular y discernir los que no sean útiles para el análisis, esta evaluación previa es crucial para identificar las variables más relevantes que podrían influir en los resultados del modelo.

Los factores que permiten evaluar la posibilidad de sufrir un ACV, son muy diversos, estos pueden estar ligados a problemas de salud, malos hábitos e inclusive pueden estar ligados a factores emocionales y psicológicos. Jerez y Madero-Cabib (2021), estiman que alrededor del 90% de los casos de ACV se atribuyen a factores modificables que pueden ser corregidos de acuerdo con el estilo de vida del paciente, entre ellos la presión arterial, la obesidad, el colesterol, los patrones dietéticos inadecuados, entre otros. Sin embargo, se ha evidenciado la afectación por parte de factores de tipo psicosocial, del entorno y del comportamiento, principalmente el estrés.

Bajo este contexto, los autores realizaron una recopilación de datos cara a cara con 802 personas entre 65 a 75 años en circunstancias familiares y laborales completamente distintas. En este caso, los autores aplican un modelo de regresión logística posterior a un análisis de secuencia multicanal, para evaluar si el estrés producido en un entorno familiar o laboral debe ser considerado como un factor de riesgo en distintas etapas y dominios de la vida del paciente. Los resultados indican que el escenario donde existe la presencia de estrés en ambientes familiares y con ausencia de estrés laboral, refiere un riesgo 10 veces mayor a estar en un ambiente libre de estrés (Jerez & Madero-Cabib, 2021).

Como estos estudios, existen otras investigaciones que han aplicado modelos predictivos para medir el riesgo de padecer una enfermedad crónica no transmisible (ENT), que como se había descrito anteriormente, son la principal causa de muerte y discapacidad permanente.

Entre las enfermedades crónicas no transmisibles se encuentran las siguientes: enfermedades cardiovasculares, diabetes mellitus, enfermedades cerebrovasculares e hipertensión arterial, esto de acuerdo al artículo titulado:

"*Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo*" por Avellán et al. (2022), donde se aplican modelos de Machine Learning para predecir estas enfermedades. El objetivo de este estudio, es proporcionar un recurso médico que facilite la toma de decisiones por parte de los profesionales de la salud y la detección temprana de estas enfermedades. Esta investigación utilizó una base de datos de libre acceso del Instituto Nacional de Estadísticas y Censos (INEC), y con el apoyo de un profesional de la salud se identificó los factores riesgos necesarios para el análisis, en el cual se aplicaron tres algoritmos: Bayes Nayves, Regresión Logística y Random Forest (Avellán et al., 2022).

Según Avellán et al. (2022), indican que se aplicaron los tres modelos para la predicción de estas enfermedades, resultando algunos modelos más efectivos que otros en el análisis de cada enfermedad, esto debido a la naturaleza de los datos. Los resultados de este estudio indican que Random Forest tiene un mayor rendimiento en el caso de las enfermedades cerebrovasculares, con un valor de F1_Score de 0.927; en el caso de la diabetes mellitus, el modelo de Regresión logística es el que más se ajusta con un F1_Score de 0.781 y por su parte, Bayes Nayves tiene un mejor desempeño en las enfermedades cardiovasculares con un F1_Score de 0.836. Por último, para la hipertensión arterial los resultados arrojados por los modelos son muy bajos, lo que indica que la data no es suficiente para realizar un buen estudio.

En un estudio similar realizado por Blanc-Pihuave et al. (2020), se enfatiza en la efectividad al usar Machine Learning para este tipo de análisis, debido a su capacidad para aprender a partir de grandes volúmenes de datos y mejorar procesos de clasificación y/o predicción. En el área de la salud, esto permite facilitar y hacer más preciso el diagnóstico médico gracias a la capacidad de procesamiento, comparación y síntesis de la información. Los autores usaron algoritmos de aprendizaje supervisado, como redes bayesianas y redes neuronales para poder identificar la presencia o ausencia de una enfermedad cardiovascular, o a su vez el nivel de riesgo para adquirir una enfermedad.

En otra investigación realizada por Ramos et al. (2024), indica que la clave para prevenir o reducir de manera efectiva el impacto de un ACV es identificar los factores de riesgo más relevantes que están asociados a este

padecimiento. Para este análisis, los autores realizaron una evaluación de los métodos estadísticos más utilizados en el sector salud, que se han usado para realizar investigaciones que modelan los datos para el diagnóstico patológico de otros padecimientos, como el cáncer de mama con posibilidad de metástasis a otros órganos y la predicción de muerte por COVID-19, en los cuales se aplicó regresión logística (LR) para su análisis (Bustan & Poerwanto, 2021; Yupari et al., 2021).

En el caso anterior los autores utilizaron una base de datos de ensayos clínicos abiertos, la cual contiene información de 219 personas con un rango de edad de 20 a 89 años. Este, contiene variables importantes a considerarse para la predicción de un ACV, como la hipertensión y la diabetes en los pacientes, los autores determinaron que el modelo con mayor precisión es la regresión logística, donde se determinó que la edad y el índice de masa corporal son las variables predictoras que más afectan en un ACV (Ramos et al., 2024).

Si bien, para este tipo de casos médicos la mayoría de los autores han decidido buscar una solución mediante Machine Learning, hay que tener en cuenta que no es una solución aplicable a todo tipo de análisis, por lo cual es importante identificar en que situaciones es correcto y necesario utilizar el aprendizaje automático (Campos et al., 2019).

Tabla 1

Matriz de investigaciones similares

Título de la investigación	Problema	Objetivos	Fuente de datos	Metodología	Resultado	Modelo utilizado
Desarrollo de un algoritmo con redes neuronales para la predicción de ACV en pacientes diabéticos.	No hay un estudio que permita evaluar que tan propenso está un paciente diabético de sufrir un ACV.	Determinar en qué medida un algoritmo basado en redes neuronales, puede ayudar a predecir un ACV.	Base de datos de Kaggle, conformado por 17372 registros y 9 variables.	Para el desarrollo se utilizó la metodología SEMMA (sample, explore, modify, model, assess), además del uso de la librería H2O para generar el modelo de Deep Learning.	Los resultados superaron el mínimo aceptable: Precisión – 91% Accuracy – 94% Sensibildiad – 93% Especificidad – 94%	Redes Neuronales
Trayectorias de estrés familiar y laboral y su asociación con accidentes cerebrovasculares.	Necesidad de identificar si el enfrentar cotidianamente situaciones de estrés, puede considerarse un factor de riesgo para desarrollar un ACV.	Observar las trayectorias de estrés laboral y/o familiar, y determinar su asociación con ACV.	La información fue recolectada mediante encuestas a 802 personas chilenas en diferentes circunstancias.	Se utilizó análisis de secuencias multicanal para reconstruir trayectorias de estrés familiar y/o laboral, para luego utilizar modelos de LR para determinar la asociación de estas trayectorias con un ACV.	Las personas con un nivel de estrés laboral y/o familiar constante o con etapas de inactividad prolongada son más propensos a sufrir un ACV.	Regresión Logística
Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo.	Alto porcentaje y en ascenso, de personas con hipertensión arterial, principal factor de riesgo para desarrollar un ECV.	Predicción temprana de las principales enfermedades que atacan la salud del país, para elaborar un insumo médico que apoye la toma de decisiones médicas.	Base de datos con 5235 registros y 847 atributos de personas mayores de 60 años, obtenida del Instituto Nacional de Estadísticas y Censos (INEC).	Se implementó la metodología Crisp-DM (Cross – Industry Standard Process for Data Mining), para el manejo de fases y el proceso KDD (Knowledge Discovery Databases). Estas metodologías no son de secuencia rígida, por lo cual permiten regresar a	Se utilizaron 3 modelos diferentes para con un nivel de precisión diferente para cada enfermedad (enfermedades cardiovasculares, accidente cerebrovascular, diabetes mellitus, hipertensión arterial).	Random Forest, Bayes Naives y Regresión Logística.

				etapas anteriores para la mejora de procesos.		
Modelo computacional de clasificación de aprendizaje de máquina supervisado, para el análisis de datos cardiovasculares y pronóstico médico.	Falta de material de apoyo para que los especialistas puedan determinar, predecir y/o detectar una enfermedad, de acuerdo a diferentes factores de riego.	Diseño de un modelo de clasificación con el apoyo de métodos probabilísticos (redes bayesianas y redes neuronales) que permitan modelar los principales factores de riesgo de un ECV.	El estudio toma una base de datos de la Universidad de Ryerson, que tiene dos pilares, la recopilación de información de exámenes médicos e información del estilo de vida de los pacientes.	Aplicación de técnicas de Machine Learning mediante los algoritmos de Naive Bayes y Elvira, para la predicción de un ECV de acuerdo a los factores de riesgo.	La investigación aplica métodos de Machine Learning para modelar este tipo de enfermedades con un porcentaje de error menor al 20% en su entrenamiento. Haciendo más preciso el modelo de predicción y/o de diagnóstico.	Redes Bayesianas y Redes Neuronales
Predicción de accidente cerebrovascular utilizando regresión logística.	Alto grado de personas que padecen de uno o más enfermedades, consideradas como factores de riesgo, lo cual provoca un incremento en la aparición de un ACV.	Identificar los factores de riesgo más significativos de un accidente cerebrovascular o ictus, además de tener la capacidad de predecir la aparición de la enfermedad.	Base de datos de un ensayo clínico abierto del Hospital Popular de Guillin, con datos de 219 personas, en un rango de edad de 20 a 89 años. Contiene algunas variables consideradas como factor de riesgo, como: edad, índice de masa corporal, azúcar en sangre, presión arterial, sexo y frecuencia cardíaca.	Uso de la curva de ROC para determinar el mejor punto de corte para el diagnóstico, evaluando la variación de la sensibilidad y la especificidad.	La curva ROC está cerca de uno y la AUC es de 0.859, por lo tanto, se considera que tiene una buena capacidad predictiva. Cuatro predictores son los más significativos, como son: la edad, el índice de masa corporal, la presión arterial y la diabetes. Donde el aumento de la edad y del índice de masa corporal aumentan el riesgo de sufrir un ACV y también de la probabilidad de mortalidad en un paciente.	Regresión Logística

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

Para comprender la gravedad y la prevalencia de los ACV, es crucial distinguir entre sus dos principales tipos: el isquémico y el hemorrágico. Cada uno presenta diferentes mecanismos de origen, pero ambos representan amenazas significativas para la salud. Estos eventos no solo afectan de manera directa la vida de los pacientes, sino que también imponen una carga considerable en los sistemas de salud. A continuación, se describen en detalle las características y causas de ambos tipos de ACV, con el fin de proporcionar un contexto sólido para el análisis posterior.

El Accidente Cerebrovascular Isquémico es el tipo más común, con cerca del 85 % de los casos de ACV. Sucede cuando una arteria que lleva sangre al cerebro se obstruye o se reduce, usualmente a causa de un coágulo sanguíneo. La formación de estos coágulos puede deberse a la aterosclerosis, donde las placas de grasa se acumulan en las arterias, o a coágulos que se desplazan desde cualquier parte del cuerpo hasta una arteria cerebral.

El Accidente Cerebrovascular Hemorrágico, constituye aproximadamente el 15% de los casos de ACV y se produce cuando una arteria cerebral se rompe, causando una hemorragia. Entre las causas más comunes se destaca la hipertensión no controlada, la aneurisma y las malformaciones arteriovenosas.

Existen varios factores de riesgo para los accidentes cerebrovasculares, entre ellos como: la hipertensión arterial (factor de riesgo más significativo), enfermedades cardiacas, diabetes, colesterol alto, tabaquismo, obesidad y sedentarismo, por mencionar las más importantes.

Según estudios realizados a nivel mundial, por Feigin et al. (2021), se estima que en el año 2019 hubo aproximadamente 12,2 millones de nuevos casos de ACV (casos incidentes) y 101 millones de casos existentes (casos prevalentes). El ACV también contribuyó a 143 millones de AVAD (años perdidos por discapacidad o muerte) y 6,55 millones de muertes.

Los países de ingresos bajos y medios soportaron una carga significativamente mayor de muertes por ACV en comparación con los países de ingresos altos.

Las tasas de muerte y discapacidad a causa de ACV variaron considerablemente entre diferentes países y regiones del mundo.

Estos datos resaltan la considerable carga global del ACV y la necesidad de crear, mejorar y aplicar nuevas estrategias de prevención y tratamiento de la enfermedad, particularmente para individuos más jóvenes y en países con ingresos bajos y medios.

Según Daniel Moreno-Zambrano et al. (2016), en Ecuador el panorama es muy similar, este desorden es una de las primeras causas de mortalidad desde 1975, año en el cual alcanzó el noveno lugar y 25 años después en 1990, se posicionó como primera causade muerte en el país.

Actualmente, los sistemas de salud enfrentan múltiples desafíos en la prevención de ACV, como la identificación temprana de individuos en riesgo, la implementación de estrategias preventivas eficaces y la gestión de recursos para tratar a los pacientes afectados. A menudo, la detección de factores de riesgo se realiza de forma reactiva, o sea, después de mostrar síntomas o sufrir un ACV. Este enfoque no solo es ineficaz, sino también costoso en términos de atención médica y recursos humanos.

Los métodos tradicionales de evaluación del riesgo suelen basarse en datos limitados y en modelos estadísticos que no capturan la complejidad y la interacción de múltiples factores de riesgo. Esto resulta en una prevención subóptima y una asignación ineficiente de recursos de salud pública. Por lo tanto, es imperativo desarrollar modelos predictivos avanzados que puedan identificar de manera proactiva a las personas con mayor riesgo de sufrir un ACV, permitiendo intervenciones preventivas oportunas y personalizadas.

Un modelo predictivo para la prevención de ACV utiliza técnicas de aprendizaje automático para evaluar de manera precisa y anticipada el riesgo de ACV en individuos. Este tipo de modelo tiene varias ventajas significativas:

Los modelos predictivos pueden analizar grandes volúmenes de datos de salud, incluidos historiales médicos, hábitos de vida, antecedentes familiares, y resultados de pruebas diagnósticas, permitiendo identificar patrones complejos

y sutiles que pueden no ser evidentes a través de métodos tradicionales, facilitando la detección temprana de individuos en riesgo.

Al proporcionar una evaluación del riesgo individualizada, un modelo predictivo puede permitir diseñar intervenciones preventivas específicas para cada paciente. Esto puede abarcar modificaciones en el estilo de vida, medicación preventiva y un seguimiento médico más regular. La personalización de la atención mejora la eficacia de las intervenciones y maximiza los resultados positivos para los pacientes.

La prevención efectiva de ACV no solo reduce la incidencia de estos eventos, sino que también disminuye la carga de discapacidad a largo plazo. Esto mejora notablemente la calidad de vida de los pacientes y sus también de sus familiares, además esto puedo reducir los costos asociados a la atención médica de estos pacientes a largo plazo.

PLANTEAMIENTO DEL PROBLEMA

Los métodos tradicionales de evaluación del riesgo suelen basarse en datos limitados y en modelos estadísticos que no capturan la complejidad y la interacción de múltiples factores de riesgo. Esto resulta en una prevención subóptima y una asignación ineficiente de recursos de salud pública. Por lo tanto, es imperativo desarrollar modelos predictivos avanzados que puedan identificar de manera proactiva a las personas con mayor riesgo de sufrir un ACV, permitiendo intervenciones preventivas oportunas y personalizadas.

La adopción de un enfoque analítico puede ser la apropiada para abordar este problema ya que el análisis de datos puede ser útil para descubrir patrones y factores de riesgo específicos relacionados con la aparición de enfermedades crónicas no transmisibles entre las personas.

Un modelo predictivo para la prevención de ACV utiliza técnicas de aprendizaje automático para evaluar de manera precisa y anticipada el riesgo de ACV en individuos. Este tipo de modelo tiene varias ventajas significativas:

Los modelos predictivos pueden analizar grandes volúmenes de datos de salud, incluidos historiales médicos, hábitos de vida, antecedentes familiares, y resultados de pruebas diagnósticas, permitiendo identificar patrones complejos y sutiles que pueden no ser evidentes a través de métodos tradicionales, facilitando la detección temprana de individuos en riesgo.

Al proporcionar una evaluación del riesgo individualizada, un modelo predictivo puede permitir diseñar intervenciones preventivas específicas para cada paciente, desde cambios en el estilo de vida, hasta seguimiento médico más frecuente y dirigido. La personalización de la atención mejora la eficacia de las intervenciones y maximiza los resultados positivos para los pacientes.

La prevención efectiva de ACV no solo reduce la incidencia de estos eventos, sino que también disminuye las cifras de muerte y discapacidad a largo plazo, reduciendo los costos asociados a la atención y mejora de la calidad de vida de los pacientes. La prevención de ACV puede generar ahorros sustanciales en la atención médica, la rehabilitación y el ausentismo laboral, además que puede contribuir a mejorar la salud y el bienestar general de los pacientes.

OBJETIVO GENERAL

Evaluar y comparar el rendimiento de diversos modelos predictivos basados en técnicas de análisis de datos con la finalidad de identificar las personas con alto riesgo de sufrir un accidente cerebrovascular.

OBJETIVOS ESPECÍFICOS

Identificar y seleccionar diferentes algoritmos de aprendizaje automático y técnicas estadísticas más adecuados para predecir el riesgo que tiene una persona de sufrir un accidente cerebrovascular.

Comparar el desempeño de los modelos aplicados para identificar la precisión y robustez de cada uno, para la predicción de ACV mediante los factores de riesgo asociados a esta enfermedad.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Los accidentes cerebrovasculares (ACV) se han convertido en una problemática de salud a nivel mundial, está es una de las enfermedades crónicas no transmisibles más estudiadas debido a las cifras tan altas de casos registrados, el nivel de incidencia anual de un ACV es de 200 casos por 100,000 habitantes y una prevalencia de 600 casos en la misma muestra poblacional, dónde las mujeres representan un porcentaje significativamente mayor que los hombres en torno a pacientes afectados por esta enfermedad (Bender del Busto, 2019).

Según Piloto et al. (2020), los ACV no solo constituyen un problema médico potencialmente alto, sino también un problema social y económico significativo; más allá de los desafíos médicos que enfrentan los pacientes y sus familias, como la rehabilitación y el tratamiento continuo, los ACV influyen también en el entorno familiar y psicológico de los afectados, debido a las preocupaciones económicas en torno a los costos asociados a la atención médica prolongada y la pérdida de productividad laboral.

En este contexto, la presente investigación tiene como objetivo evaluar los factores de riesgo más significativos a la hora de predecir un ACV, esto con la finalidad de identificar aquellos factores que pueden modificarse para mejorar la prevención o diagnóstico temprano de esta enfermedad. Para lograrlo, se utilizarán técnicas de Machine Learning y análisis de datos. Para lo cual, se evaluarán diferentes modelos, como la regresión logística, redes neuronales y random forest, para sustentar la precisión y confiabilidad de las predicciones del análisis.

SELECCIÓN DE LA BASE DE DATOS

Para este proyecto se ha tomado una base de datos de acceso libre disponible en la plataforma digital Kaggle, la cual alberga la comunidad de Data Science más grande del mundo. Kaggle brinda herramientas y recursos de libre acceso que facilitan los trabajos de investigación, permitiendo a los investigadores contribuir al avance científico mediante casos de estudio.

La base de datos utilizada en esta investigación, contiene información de varios parámetros considerados como factores de riesgo de padecer un ACV, entre ellos: el sexo, la edad, distintas enfermedades y el hábito fumador de un paciente (Fede, 2021). El conjunto de datos se encuentra en formato csv, está conformado de 5,110 registros y 12 variables, las cuales se utilizarán para predecir la probabilidad de que un paciente sufra un ACV en base a las características que contiene la base de datos.

LIMPIEZA, PREPROCESAMIENTO Y/O TRANSFORMACIÓN DE DATOS

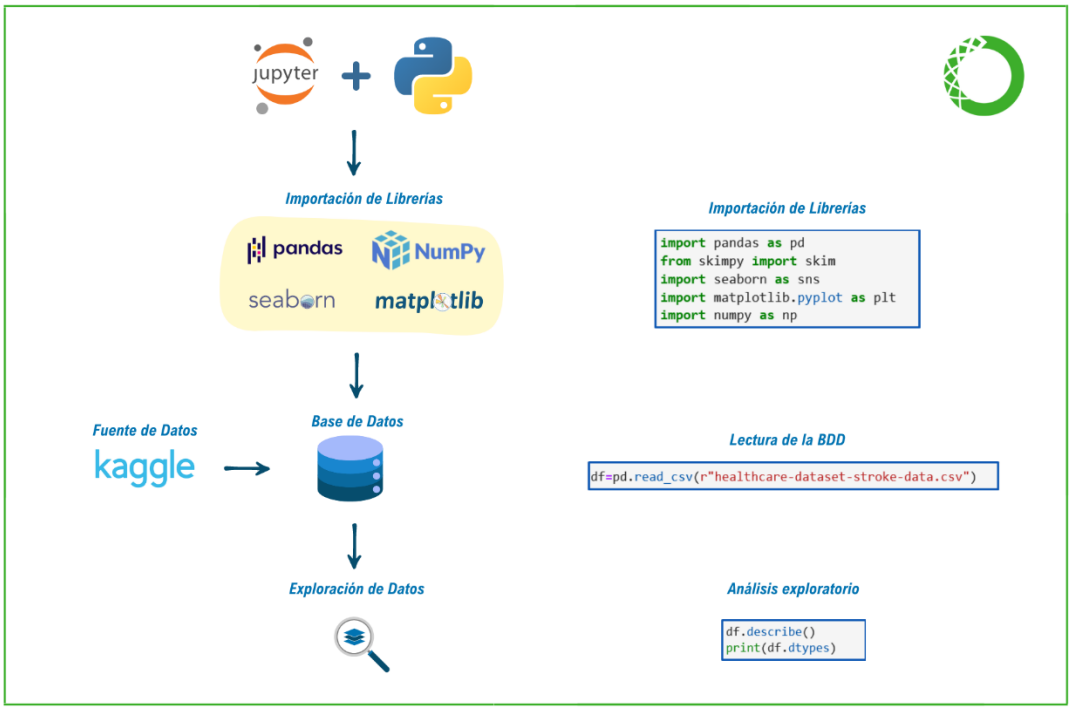
Para el preprocesamiento y análisis de datos se hará uso del lenguaje de programación Python, el cual está más orientado al Data Science, Machine Learning y al procesamiento de grandes cantidades de datos. Este, gracias a herramientas como Anaconda y Jupyter Notebooks, permiten crear un entorno de desarrollo muy versátil, robusto y unificado que facilita el desarrollo e implementación de proyectos de esta índole, permitiendo la integración no solo de código, sino de texto, imágenes, visualizaciones y ejecución de otros lenguajes de programación a través de un navegador.

Además, para el análisis exploratorio, limpieza y análisis de datos se utilizarán librerías propias de Python que simplifican el trabajo y aceleran el proceso de desarrollo, entre ellas: Pandas, Seaborn, Matplotlib y Numpy. Las cuales se usan para la manipulación y análisis de los datos, además de facilitar la creación de visualizaciones estadísticas.

El proceso de carga, preprocesamiento y transformación de datos se realiza con el uso de las librerías descritas anteriormente. El proceso inicia con la importación de estas librerías al entorno de desarrollo de Jupyter Notebook, para luego leer el archivo csv del caso de estudio mediante Pandas (Figura 1). Posteriormente, se inicia con el análisis exploratorio de los datos, donde se aplican herramientas estadísticas descriptivas y gráficas que permiten comprender la estructura original de los datos, además permiten identificar posibles valores atípicos que podrían sesgar el análisis. Para el análisis exploratorio se puede usar las funciones *describe* para obtener una descripción estadística de las variables numéricas, como: el valor mínimo, el valor máximo, porcentiles, etc. y la función *type* para identificar el tipo de datos con el que se está trabajando (Castillo et al., 2023).

Figura 1

Proceso de importación y manejo de las librerías de Python



Nota: Elaboración propia

La base de datos utilizada en este proyecto se obtuvo del repositorio digital Kaggle, como se mencionó previamente. Este conjunto de datos incluye distintas variables que pueden considerarse como factores de riesgo para sufrir un accidente cerebrovascular (ACV). A continuación, se presenta una descripción detallada de la composición de esta base de datos:

Figura 2

Descripción general de las variables del conjunto de datos

```
df.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

```
print(df.shape)
```

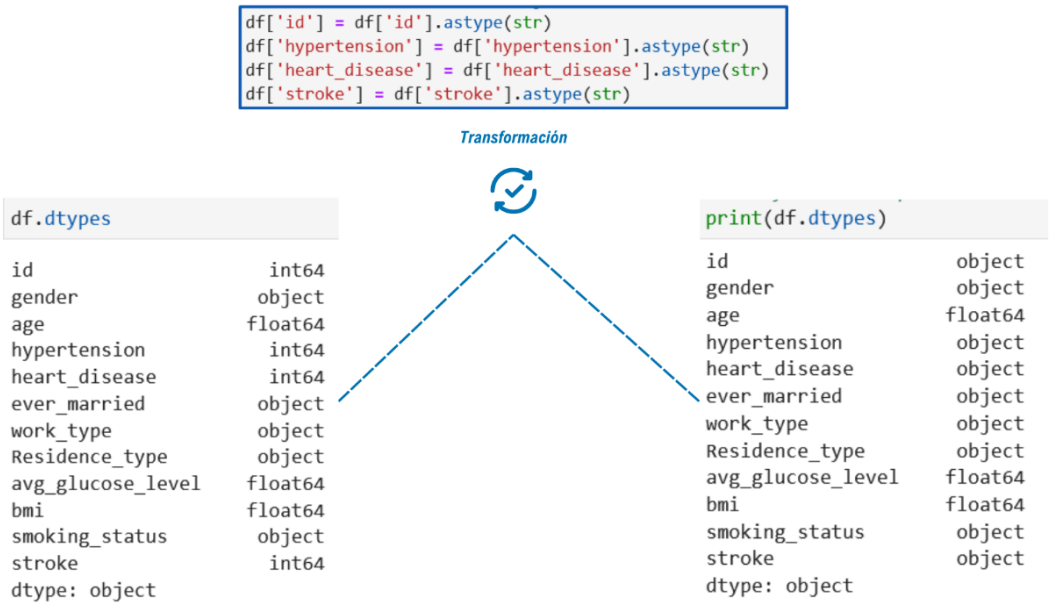
```
(5110, 12)
```

Nota: La base de datos posee 5,110 registros de diferentes pacientes, cada uno conformado por 12 columnas que representan diferentes características del paciente.

Como parte del preprocesamiento, se evaluaron los tipos de datos iniciales del dataset y se realizaron las transformaciones necesarias para cada variable.

Figura 3

Transformación del tipo de datos de las variables

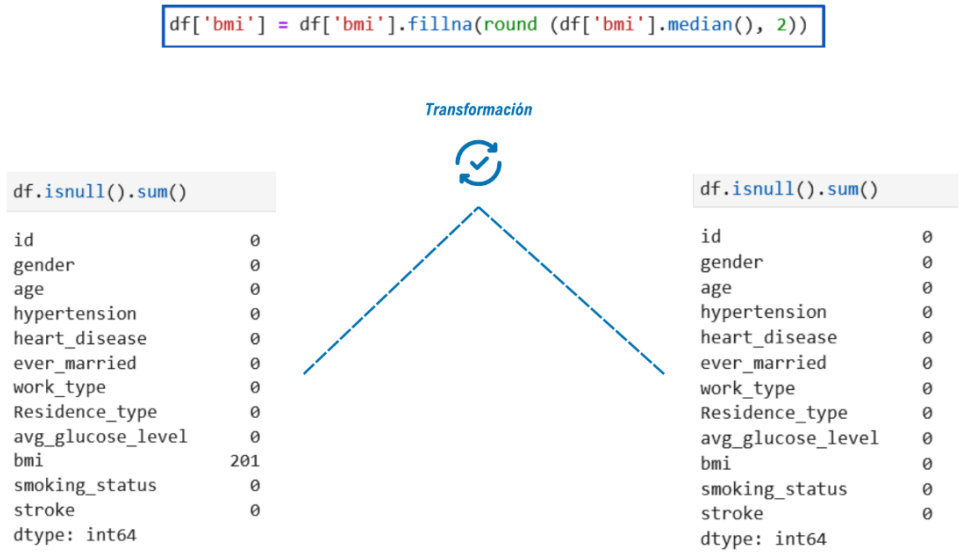


Nota: Elaboración propia.

Además, se verificó si la base de datos tiene valores nulos, valores perdidos o no válidos. Con lo cual se pudo observar que la columna bmi (índice de masa corporal), tiene 201 valores nulos equivalente al 3% del total de registros de la base, pudiendo eliminarlos sin ningún problema (Figura 4). Sin embargo, para el caso de estudio se opta por realizar una imputación con la media, esto con la finalidad de no perder información.

Figura 4

Imputación de datos con la media



Nota: Elaboración propia.

Posterior a ello se verifica que no existan valores duplicados en el conjunto de datos, utilizando el ID único de cada paciente como referencia. El análisis confirmó que el dataset contiene un total de 5,110 registros únicos, lo que indica que no hay presencia de valores duplicados en el dataset. Garantizando así, la integridad y la calidad de los datos, por ende, un análisis más preciso y confiable en las etapas posteriores de este proyecto.

IDENTIFICACIÓN DE VARIABLES

Luego de un ligero análisis de la base de datos, se pueden identificar distintas variables que pueden contribuir en la predicción de que un paciente pueda o no sufrir un ACV en función de estas variables, mismas que representan factores de riesgo potenciales, que, al ser evaluadas en conjunto, puede proporcionar una predicción más precisa de la probabilidad de ocurrencia de un ACV, a continuación, se describen el contexto de cada una de ellas:

Tabla 2

Diccionario de variables

Columna	Tipo de Dato	Descripción
id	Cualitativo / Categórica	Identificador único del paciente.
gender	Cualitativo / Categórica	Género del paciente: "Male", "Female" u "Other".
age	Cuantitativo / Continua	Edad del paciente.
hypertension	Cualitativo / Categórica	Indica si el paciente tiene o no hipertensión: 1 “SI”, 0 “NO”.
heart_disease	Cualitativo / Categórica	Indica si el paciente tiene o no enfermedades del corazón: 1 “SI”, 0 “NO”
ever_married	Cualitativo / Categórica	Indica si el paciente alguna vez se ha casado o no: "True" o "False"
work_type	Cualitativo / Categórica	El estatus del paciente en torno a su trabajo: "children", "Govt_jov", "Never_worked", "Private" o "Self-employed".
Residence_type	Cualitativo / Categórica	El estatus del paciente en torno al tipo de su residencia, existen valores de: "Rural" o "Urban".
avg_glucose_level	Cuantitativo / Continua	Nivel del azúcar promedio en la sangre.
bmi	Cuantitativo / Continua	Índice de masa corporal.
smoking_ status	Cualitativo / Categórica	El estatus del paciente en sus hábitos de fumar: "formerly smoked", "never smoked", "smokes" or "unknown".
Stroke	Cualitativo / Categórica	Indica si el paciente ha tenido un ACV o no: 1 “SI”, 0 “NO”.

DEFINICIÓN DE VARIABLES

En el área médica, tanto el índice de masa corporal (IMC) como el nivel de azúcar en sangre, son medidas que indican si un paciente está dentro de los rangos saludables de estos indicadores. El primero evalúa la relación del peso y la estatura para identificar los niveles de peso que puedan provocar problemas de salud, debido al déficit o al exceso de peso de un individuo (Hernández & Orlandis, 2021). Por otra parte, los niveles de azúcar en sangre o glucosa en sangre representan también un riesgo para la salud, altos niveles de azúcar se denomina hiperglucemia y provoca que el organismo no pueda usar la insulina correctamente, y se denomina hipoglucemia cuando estos niveles han bajado demasiado que necesitan volver a un rango normal, mediante recomendaciones alimenticias y de estilo de vida del paciente (Eske, 2021).

De acuerdo con lo publicado en el sitio oficial de la (Organización Mundial de la Salud / OMS, 2024), en el año 2019 se registraron cinco millones de muertes por enfermedades crónicas no transmisibles (ENT) como los accidentes cerebrovasculares (ACV). En este artículo, se menciona que los pacientes con sobrepeso, independientemente de la edad, tienen una mayor probabilidad de sufrir una ENT. Según la OMS, el diagnóstico de sobrepeso y obesidad se determina por el cálculo del índice de masa corporal (IMC), mediante la siguiente fórmula: $peso(kg)/estatura^2\ (m^2)$.

Las categorías de IMC se detallan a continuación:

Tabla 3

Intervalos del índice de masa corporal (IMC)

Rango	IMC (kg/m^2)
Peso Insuficiente	< 18.5
Peso normal o saludable	> 18.5 < 24.9
Sobrepeso	> 25.0 < 29.9
Obesidad	> 30.0

Nota: Recuperado de Enciclopedia Médica A.D.A.M. (2024)

Otro factor importante a tomarse en cuenta en este estudio, son los niveles de azúcar en sangre o glucosa en sangre. Este indicador evalúa si el nivel de azúcar está por encima o bajo los valores normales en los seres humanos, esta se mide en miligramos de azúcar por decilitro (mg/dL) o en milimoles de azúcar por litro $(mmol/L)$, en los casos donde se presenta un elevado porcentaje de azúcar en sangre, los profesionales de salud lo denominan hiperglucemia, esta con el tiempo puede provocar daños irreparables en algunos órganos y sistemas del organismo, por otra parte, si un paciente tiene niveles bajos de azúcar, se denomina hipoglucemia. Los niveles normales de azúcar en sangre, varían de acuerdo al estado de salud de los pacientes, si estos están sanos o tienen algún grado de diabetes como antecedente (Organización Mundial de la Salud / OMS, 2023).

Tabla 4

Niveles normales de azúcar en sangre antes de las comidas

Rango de edad	**Glucosa (mg/dL^2)**
Adultos	90 a 130
Niños de 13 a 19 años	90 a 130
Niños de 6 a 12 años	90 a 180
Niños menores de 6 años	100 a 180

Nota: Recuperado de Enciclopedia Médica A.D.A.M. (2024)

Tabla 5

Niveles normales de azúcar en sangre después de las comidas

Rango de edad	**Glucosa (mg/dL^2)**
Adultos	< 180

Nota: Los rangos de azúcar después de las comidas deben medirse una a dos horas después de la comida. Recuperado de la Enciclopedia Médica A.D.A.M. (2024)

Tabla 6

Niveles normales de azúcar al acostarse

Rango de edad	Glucosa (mg/dL^2)
Adultos	90 a 150
Niños de 13 a 19 años	90 a 150
Niños de 6 a 12 años	100 a 180
Niños menores de 6 años	110 a 200

Nota: Recuperado de Enciclopedia Médica A.D.A.M. (2024)

DESCRIPCIÓN DE VARIABLES

Una vez realizada la descripción de las variables, se realiza un resumen estadístico de las variables, la cual incluye estadísticas descriptivas de las variables numéricas como: la media, la desviación estándar, los cuartiles y los valores máximos y mínimos de cada variable.

Tabla 7

Matriz de resumen estadístico de variables numéricas

	age	avg_glucose_level	bmi
count	5110.000000	5110.000000	5110.000000
mean	43.226614	106.147677	28.862035
std	22.612647	45.283560	7.699562
min	0.080000	55.120000	10.300000
25%	25.000000	77.245000	23.800000
50%	45.000000	91.885000	28.100000
75%	61.000000	114.090000	32.800000
max	82.000000	271.740000	97.600000

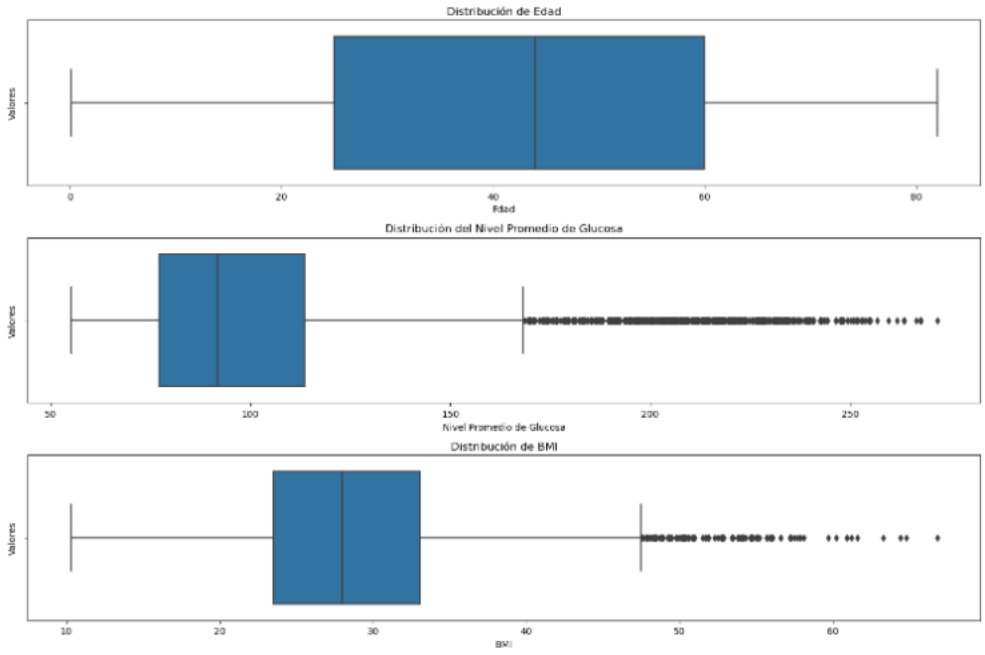
Se observa que el valor máximo del índice de masa corporal (IMC) es 97.6.

Según la Organización Mundial de la Salud (OMS), un índice de masa corporal ya en situaciones de obesidad mórbida el límite es de 70, dado que estos valores exceden significativamente este umbral, se atribuye que se trata de errores en la recolección de los datos. Por lo tanto, se procede a filtrar y eliminar estos valores para asegurar la integridad y precisión del análisis.

Para tener una descripción gráfica más clara de los datos, se realiza un diagrama de cajas que permite identificar los valores atípicos de estas variables. Este gráfico es bastante eficiente a la hora de visualizar la dispersión y distribución de los datos, así como para detectar posibles anomalías que podrían influir en el análisis (Figura 5).

Figura 5

Diagrama de cajas de las variables numéricas



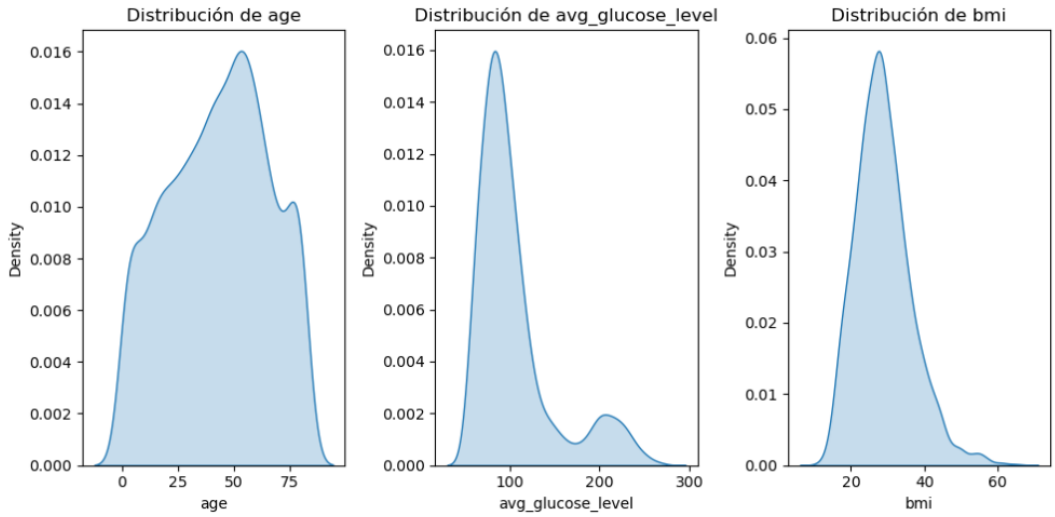
Nota: Elaboración propia.

Se puede observar que las variables del IMC y del nivel de azúcar en sangre presentan valores atípicos. Si bien estos valores son atípicos en el conjunto de datos, es importante aclarar que siguen siendo posibles en los seres humanos. Por lo tanto, estos valores no se eliminan ya que podrían representar casos extremos pero válidos.

A continuación, se realiza la distribución de las variables continuas para analizar su comportamiento en el conjunto de datos, como patrones y/o tendencias (Figura 6). Seguidamente, se revisa la correlación de estas variables, mediante un mapa de calor de la matriz de correlación:

Figura 6

Distribución de variables continuas



Nota: Elaboración propia.

Figura 7

Matriz de correlación de variables

Nota: Elaboración propia.

La correlación de datos se refiere a la relación de los campos o variables de un conjunto de datos, esta varía de -1 a 1. Una correlación positiva indica que, si el valor de una variable aumenta, la otra también tiende a aumentar. Por el contrario, una correlación negativa ocurre cuando el valor de una variable

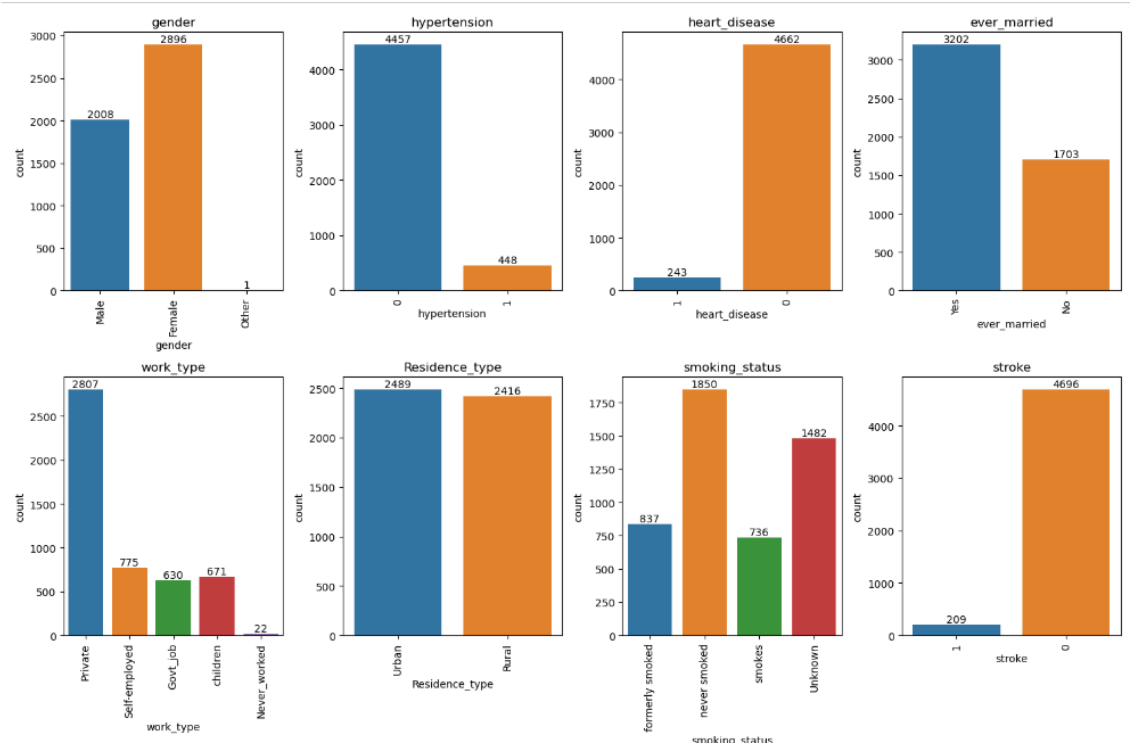
aumenta y la otra tiende a disminuir. En este caso, se puede observar que no existe una fuerte correlación entre las variables continuas del dataset, los valores entre las variables categóricas son muy bajos para determinar una alta correlación.

VISUALIZACIÓN DE VARIABLES

En este apartado, se muestra la distribución de los datos en torno a las variables cuantitativas para una exploración más amplia de los datos.

Figura 8

Distribución de las variables cualitativas



Nota: Elaboración propia.

Se puede observar que en la variable del género existe un paciente con valor “Otro”, y que la mayoría de pacientes no padecen de hipertensión ni enfermedades del corazón. Además, en el estado del fumador hay presencia de valores desconocidos, por lo cual se debe analizar la posibilidad y el método de imputación adecuado para dichos valores. Por otra parte, la variable objetivo del estudio tiene en su mayoría valores de la case 0, lo que indica que la mayor cantidad de pacientes no ha sufrido un ACV, lo que indica que es un problema

de clases desbalanceadas, lo cual debe tenerse en cuenta en el entrenamiento de los modelos.

SELECCIÓN DEL MODELO ESTADÍSTICO

La selección de métodos estadísticos en una investigación debe estar alineada estrechamente con los objetivos del mismo. Los métodos deben ajustarse a la naturaleza de los datos y las relaciones de las variables más significativas del análisis. En el presente estudio, cuyo objetivo es la predicción de los accidentes cerebrovasculares (ACV), se han seleccionado tres algoritmos de aprendizaje supervisado:

- **Regresión Logística**

La regresión logística se trata de un conjunto de técnicas estadísticas que buscan comprobar una hipótesis o analizar la relación de las variables dependientes categóricas con las variables independientes, sean estas categóricas y/o cuantitativas. Este modelo busca estudiar la probabilidad de que un evento ocurra en función de las variables significativas del estudio, en este caso, se busca predecir la posibilidad de que una persona sufra un accidente cerebrovascular (Martínez & Pérez, 2024).

Según Rodríguez (2024), cuando una variable dependiente consta únicamente de dos categorías se utiliza la regresión logística dicotómica o binaria, en el caso que la variable a predecir conste de más categorías se recurre a la regresión logística politómica, el objetivo en cualquier caso es modelar la probabilidad de que un paciente sufra un ACV o no.

En este caso de estudio, se utiliza una regresión logística binaria, donde la variable dependiente puede tomar únicamente dos valores; 1 para el caso positivo de la probabilidad de sufrir un ACV y 0 como la probabilidad de no sufrirlo.

Modelo Matemático de la Regresión Logística Binaria Simple

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Dónde P es la probabilidad de que un evento ocurra y $1 - P$ de que no ocurra, X por su parte es la variable independiente asociada a P, e es la base del

logaritmo neperiano (2.71828) y β_0 y β_1 son coeficientes del modelo (Rodríguez, 2024).

- **Random Forest**

Random forest es un algoritmo de Machine Learning que trabaja con varios árboles de decisión, donde cada uno de ellos tiene una participación en la predicción final del modelo (Aracena et al., 2022), este algoritmo ha demostrado ser bastante eficiente en la categorización de enfermedades, su manera de ensamblar varios árboles de decisión, simula la toma de decisiones en los seres humanos, por lo cual, los resultados resultan ser bastante competitivos (López et al., 2023).

Modelo Matemático del modelo Random Forest

Según (Yangón & Reyes, 2023) el Random Forest se puede expresar como una ecuación matemática, donde cada árbol de decisión se considera como una función individual, y la predicción resultante se define mediante la combinación de las salidas de todos los árboles. Obteniendo la siguiente ecuación matemática:

$$\hat{Y} = \frac{1}{n} \sum_{i}^{n} = 1\, Yi$$

Dónde $\hat{Y}$ hace referencia a la predicción final del modelo, Yi son las predicciones individuales de cada árbol de decisión y la variable *n* indica el número total de árboles en el Random Forest.

- **Redes Neuronales**

Según Sarmiento (2020), las redes neuronales se basan en las redes neuronales biológicas, estas son capaces de aprender y generalizar, permitiendo reconocer patrones, identificar y predecir comportamientos. El autor menciona que, en el campo de la biomedicina las redes neuronales son un gran aporte a la ciencia, con su aplicación se han mejorado los análisis de grandes cantidades de datos del sector salud, mejorando así los tiempos de diagnóstico y pronóstico de enfermedades, que con el tiempo se han vuelto más rápidos y precisos.

Las redes neuronales utilizan tareas de clasificación y regresión para poder aprender, generalizar y procesar datos automáticamente, la clasificación organiza los datos de entrada y la regresión predice un valor de salida. De estas aplicaciones se concluye que las redes neuronales son precisas y eficientes para el reconocimiento de patrones y predicción de comportamientos (Sarmiento, 2020).

Modelo Matemático del modelo Redes Neuronales

En la investigación realizada por Arias (2022), se enfatiza en la composición de una red neuronal, las cuales están compuestas por un número determinado de elementos denominados neuronas, las cuales se encuentra agrupadas por capas, de tal forma que una neurona en determinada capa se trata de una combinación de la capa anterior. De acuerdo a esta investigación se define la siguiente función:

$$Y = f\left(\sum_{j} WijXj + b\right)$$

Dónde, Xj corresponden a los datos de entrada para el modelo, Wij se refiere a los pesos de las entradas, b es un término aditivo o sesgo y f es una función de activación, donde se obtienen las predicciones de que un determinado caso pertenezca o no a una clase (Arias, 2022).

RESULTADOS

ANÁLISIS DE LOS MODELOS ESTADÍSTICOS E INTERPRETACIÓN DE RESULTADOS

Cómo parte inicial del entrenamiento de los modelos estadísticos, se debe dividir el dataset en un conjunto de datos de entrenamiento y otro de prueba, siguiendo la convención más común de 80% para el entrenamiento y 20% para la prueba. En este caso se identificó un problema de clases desbalanceadas, por lo cual, es crucial tomar medidas para equilibrar los datos en el conjunto de entrenamiento, esto evitará que el modelo favorezca siempre a la clase predominante, lo cual no es nuestro objetivo principal. Balancear las clases garantiza que el modelo aprenda a reconocer y predecir correctamente todas las clases, mejorando así su rendimiento.

Para lo cual, se realizó previamente la limpieza y preprocesamiento de la data, eliminando columnas innecesarias para el análisis, verificando la duplicidad de los datos, removiendo registros imposibles en los seres humanos dentro de las variables medidas de IMC y glucosa en sangre, y finalmente, realizando una transformación de tipos de variables numéricas y categóricas. Obteniendo así, un conjunto de datos desbalanceado, conformado de:

- 3 758 muestras de la clase 0 (NO padeció un ACV).
- 166 muestras de la clase 1 (SI padeció un ACV).

Para el balanceo se probaron distintas técnicas de balanceo, como: Random Oversampling, NearMiss y SMOTE, obteniendo mejores resultados con SMOTE (Synthetic Minority Over-sampling Technique). Esta técnica genera artificialmente nuevos registros para la clase minoritaria al interpolar entre las instancias existentes de esa clase. De esta manera, logramos un conjunto de datos de entrenamiento más equilibrado, lo cual ayuda a evitar que el modelo se sesgue hacia la clase predominante y mejora su capacidad para aprender a predecir ambas clases de manera efectiva. Después de la aplicación de esta técnica, el conjunto de datos queda de la siguiente forma:

- 3 758 muestras de la clase 0 (NO padeció un ACV).
- 3 758 muestras de la clase 1 (SI padeció un ACV).

Figura 9

Técnica SMOTE para el balanceo de clases

```
sm = SMOTE(random_state=1)
X_train_res, Y_train_res = sm.fit_resample(X_train, Y_train.ravel())

Antes OverSampling, numero de registros con '1': 166
Antes OverSampling, numero de registros con '0': 3758

Después del OverSampling, el tamaño de train_X: (7516, 10)
Después del OverSampling, el tamaño de train_y: (7516,)

Después del OverSampling, numero de registros con '1': 3758
Después del OverSampling, numero de registros con '0': 3758
```

Nota: Elaboración propia.

Una vez separado el dataset en entrenamiento y prueba, se procede a entrenar los modelos que se describieron en el capítulo anterior: regresión logística, random forest y redes neuronales.

El entrenamiento de los dos primeros se realiza sin ajuste de hiperparámetros para evaluar su desempeño inicial y para las redes neuronales se implementa el perceptrón como la unidad básica de estas, con el fin de capturar patrones básicos en los datos. Además de cada modelo, se obtiene las métricas de evaluación en cada uno para analizar y comparar su rendimiento, entre ellas: exactitud, área bajo la curva ROC (ROC AUC), precisión, recall y F1-score. Estas métricas proporcionan una visión comprensiva de la forma en que cada modelo clasifica y discrimina entre clases, permitiendo comparar su efectividad y facilitar la elección de un enfoque más adecuado para nuestro problema de clasificación. Después del entrenamiento de los modelos, se obtienen los siguientes resultados para las métricas de evaluación:

Tabla 8

Resultados de los modelos de aprendizaje supervisado

Modelo	Exactitud	ROC AUC	Precisión	Recall	F1
Regresión Logística	0.73	0.84	0.11	0.77	0.20
Random Forest	0.95	0.80	0.20	0.02	0.04
NN	0.48	0.71	0.08	0.95	0.14

Tabla 9

Matriz de confusión – Regresión Logística

Regresión logística	Predicted no stoke	Predicted stroke
No Stroke	683	255
Stroke	10	33

Tabla 10

Matriz de confusión – Random Forest

Random forest	Predicted no stoke	Predicted stroke
No Stroke	934	4
Stroke	42	1

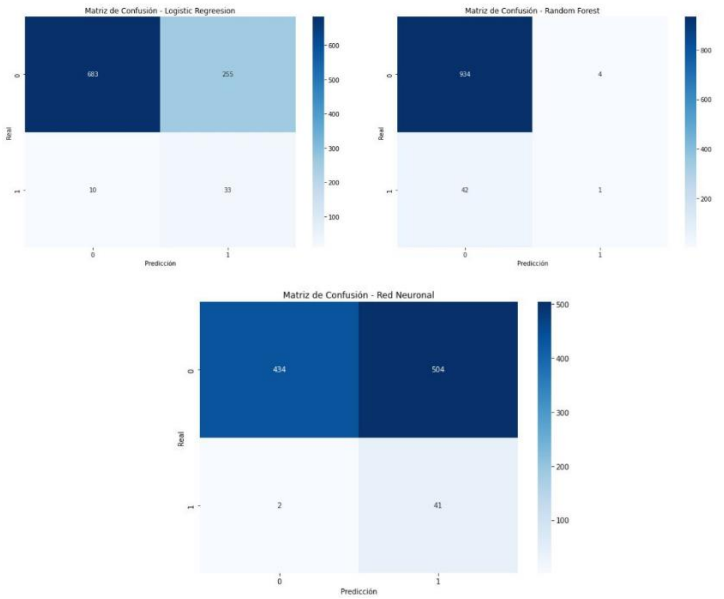
Tabla 11

Matriz de confusión – Redes Neuronales

NN	Predicted no stoke	Predicted stroke
No Stroke	434	504
Stroke	2	41

Figura 10

Matriz de confusión por modelo



Nota: Elaboración propia.

De acuerdo con los resultados obtenidos en cada modelo, se puede evidenciar un mejor desempeño con la regresión logística. Sin embargo, aunque las medidas de ROC AUC y el Recall, tienen valores relativamente altos, lo que indica un buen rendimiento en la discriminación de clases y en la identificación de la mayor cantidad de casos positivos posibles. Sin embargo, la precisión y el f1 son muy bajos, lo que sugiere un alto número de falsos positivos.

Por lo cual, se realiza un ajuste de hiperparámetros para optimizar el rendimiento de los modelos, en este caso para la búsqueda de los mejores valores se ocupa Grid Search, que es una técnica comúnmente usada en aprendizaje automático para encontrar los mejores hiperparámetros para un modelo.

Los hiperparámetros son configuraciones que no se aprenden directamente del proceso de entrenamiento del modelo, sino que se establecen antes de que comience el entrenamiento y afectan significativamente el rendimiento del modelo. Además, para la red neuronal se agregan dos capas adicionales. Con estos cambios se obtienen los siguientes resultados:

Tabla 12

Resultado de los modelos con hiperparámetros

Modelo	Exactitud	ROC AUC	Precisión	Recall	F1
Regresión Logística	0.730	0.845	0.120	0.814	0.209
Random Forest	0.953	0.815	0.200	0.023	0.042
NN	0.759	0.844	0.133	0.465	0.207

Tabla 13

Matriz de confusión con hiperparámetros – Regresión Logística

Regresión logística	Predicted no stoke	Predicted stroke
No Stroke	681	257
Stroke	8	35

Tabla 14

Matriz de confusión con hiperparámetros – Random Forest

Random forest	Predicted no stoke	Predicted stroke
No Stroke	934	4
Stroke	42	1

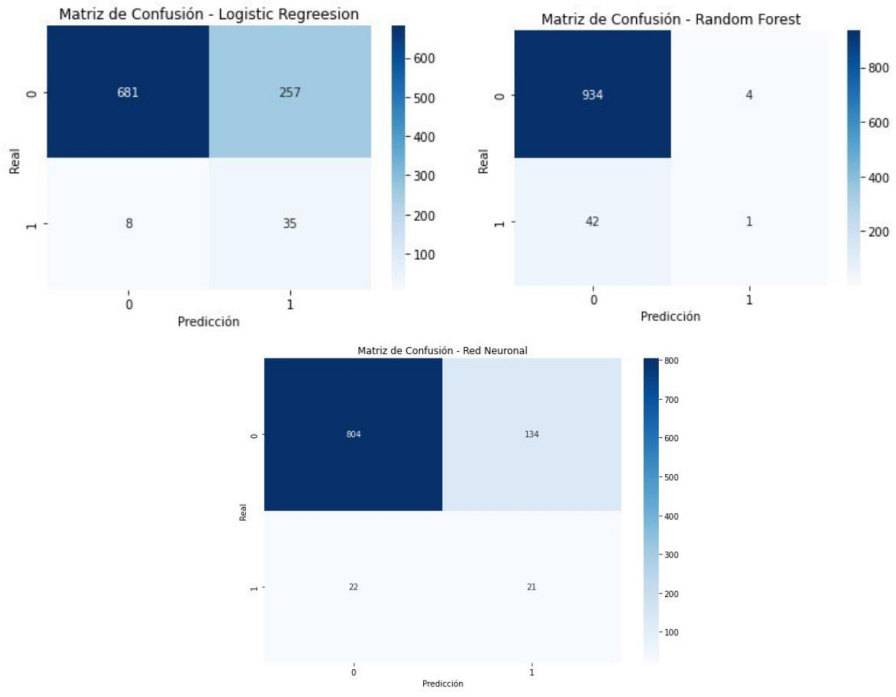
Tabla 15

Matriz de confusión con hiperparámetros – Redes Neuronales

NN	Predicted no stoke	Predicted stroke
No Stroke	808	130
Stroke	23	20

Figura 11

Matriz de confusión por modelo con hiperparámetros



Nota: Elaboración propia.

Después del ajuste de hiperparámetros, se puede observar que la regresión logística es la que mejor se ajusta al análisis, el modelo tiene un ROC AUC y un Recall relativamente alto, lo que determina que puede identificar la mayoría de casos donde los pacientes tienen riesgo de sufrir un ACV. Sin embargo, su

precisión y F1 son bastante bajo, lo que indica que el modelo tiene muchos falsos positivos.

Este estudio tiene como objetivo identificar los pacientes que están propensos a sufrir un ACV, el Recall en este caso es una métrica crucial para minimizar los falsos negativos, es decir, no incluir a las personas que si tienen riesgo de padecer la enfermedad. Por lo tanto, el modelo de regresión logística parece ser el más óptimo ya que tiene el Recall más alto entre los modelos, con un valor de 0.814, esto a pesar que el valor de su precisión sea baja. Lo que significa que, aunque se produzcan muchos falsos positivos, la mayoría de los verdaderos casos de riesgo se identificarán.

Evaluación de variables significativas

Ahora, utilizando el modelo de regresión logística, se procede a evaluar cuáles son las variables más significativas para el estudio. Para lo cual, se procede a analizar los coeficientes del modelo, que indican la fuerza y la dirección de la relación entre las variables predictoras y la probabilidad de que un paciente tenga un ACV. La identificación de estas variables proporcionan información valiosa para la prevención y gestión de los factores de riesgo asociados a un ACV, permitiendo mejorar la problemática mediante la implementación de programas de control de las variables con mayor peso en la predicción.

Figura 12

Evaluación de las variables predictoras

Feature_0	gender
Feature_1	age
Feature_2	hypertension
Feature_3	heart_disease
Feature_4	ever_married
Feature_5	work_type
Feature_6	Residence_type
Feature_7	avg_glucose_level
Feature_8	bmi
Feature_9	smoking_status

```
==============================================================================
Dep. Variable:                      y   No. Observations:                 7516
Model:                          Logit   Df Residuals:                     7505
Method:                           MLE   Df Model:                           10
Date:                Sat, 13 Jul 2024   Pseudo R-squ.:                0.3406
Time:                        19:58:22   Log-Likelihood:                -3435.0
converged:                       True   LL-Null:                       -5209.7
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.6910      0.229    -24.897      0.000      -6.139      -5.243
Feature_0      0.0612      0.068      0.900      0.368      -0.072       0.195
Feature_1      0.0794      0.002     34.745      0.000       0.075       0.084
Feature_2      0.6914      0.090      7.670      0.000       0.515       0.868
Feature_3      0.5635      0.119      4.717      0.000       0.329       0.798
Feature_4     -0.2030      0.102     -1.981      0.048      -0.404      -0.002
Feature_5      0.0254      0.036      0.704      0.481      -0.045       0.096
Feature_6      0.1813      0.067      2.720      0.007       0.051       0.312
Feature_7      0.0024      0.001      3.974      0.000       0.001       0.004
Feature_8      0.0126      0.005      2.575      0.010       0.003       0.022
Feature_9      0.1748      0.034      5.158      0.000       0.108       0.241
==============================================================================
```

Nota: Elaboración propia.

Se puede observar que las variables más significativas a la hora de predecir un ACV, son la hipertensión y las enfermedades del corazón, estas variables tienen un impacto positivo, lo que significa que un incremento de estas variables está asociado con un mayor riesgo de sufrir un ACV. Estas tienen una fuerte correlación con la probabilidad de sufrir un ACV y son esenciales para entrenar el modelo predictivo.

Otras variables la como: la edad, el nivel de glucosa en sangre y el historial de que, si un paciente es fumador o no, tienen coeficientes más pequeños, lo que sugiere un impacto menor en la predicción de un ACV. Sin embargo, de acuerdo con varios estudios, estas variables también influyen en la ocurrencia de un ACV.

Esto se puede corroborar con la literatura estudiada en esta investigación, donde se documenta que la edad es un factor de riesgo casi inminente, debido al estrechamiento y endurecimiento de las arterias. Sin embargo, existen otros factores de riesgo bastante relevantes que pueden ser modificables en cada paciente, como la hipertensión, las enfermedades cardíacas, el nivel de azúcar en sangre y el estado del fumador, para lo cual se pueden elaborar estrategias preventivas y tratamientos personalizados.

Por otra parte, en términos de investigación, estos hallazgos pueden guiar estudios futuros para explorar más a fondo la relación de estas variables con la probabilidad de sufrir un ACV, permitiendo a los profesionales de la salud realizar intervenciones más eficaces ante esta problemática.

DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

El análisis que se llevó a cabo determina, que variables son consideradas como factores de riesgo a la hora de predecir un accidente cerebrovascular. La comparación de los modelos de aprendizaje supervisado como la regresión logística, random forest y redes neuronales, han demostrado ser efectivos para la predicción de muchas enfermedades. Por lo cual, al realizar un análisis minucioso con estos tres modelos, se pudo evaluar cuál de ellos es el más óptimo para este caso de estudio.

Con base en los resultados obtenidos de los modelos predictivos desarrollados durante el proyecto, se recomienda la implementación de un modelo de regresión logística optimizado para la predicción de accidentes cerebrovasculares (ACV). Esta recomendación se fundamenta en los altos valores de ROC AUC y recall obtenidos por este modelo en comparación con los otros evaluados. El valor elevado de ROC AUC indica una excelente capacidad discriminativa del modelo, mientras que el alto recall garantiza una mayor sensibilidad en la detección de casos de ACV, minimizando la probabilidad de falsos negativos. Esta combinación de métricas sugiere que la regresión logística optimizada es la opción más eficaz para identificar individuos en riesgo y, por ende, es la estrategia recomendada para su implementación en entornos clínicos y de salud pública.

Factores como la edad, estado del fumador, y problemas de salud como la hipertensión, la azúcar en sangre y las enfermedades cardiovasculares son las que mayor riesgo aportan en este caso. Siendo estos, indicadores que pueden ser modificables a lo largo del tiempo, excepto la edad. Por tal razón, es importante reconocer que factores son los más significativos y cómo los profesionales de salud pueden diagnosticar, controlar y tratar estas dolencias para reducir el porcentaje de ACV en los seres humanos, y por ende reducir los porcentajes de discapacidad y muerte que este produce diariamente.

Las estrategias que un individuo o una organización puede generar a partir de este estudio son diversas y altamente beneficiosas. Una de ellas se enfoca en el monitoreo de la hipertensión y los niveles de glucosa en sangre, este monitoreo no solo puede prevenir los ACV, sino también facilitar el diagnóstico

temprano e inclusive promover un manejo efectivo de estos riesgos. La adopción de estas prácticas permite una intervención oportuna que puede salvar vidas y mejorar la calidad de vida de los pacientes.

Además, el desarrollo de programas de salud enfocados en el control de peso y la nutrición para pacientes con un índice de masa corporal (IMC) elevado, representa una estrategia sencilla, pero efectiva.

Desde la perspectiva técnica, el modelo predictivo se puede fortalecer mediante la inclusión de variables que se asocien a los ACV, como los niveles de estrés o preocupación de los pacientes (Jerez & Madero-Cabib, 2021). Esta ampliación permitirá obtener predicciones más precisas. Implementar estos programas en empresas públicas y/o privadas puede crear un ambiente laboral con menor nivel de estrés, además de reducir el riesgo de un ACV en los empleados. Este enfoque no solo mejora la salud de los trabajadores, sino que también puede conducir a la creación de estrategias empresariales innovadoras en el manejo del personal, incrementando la productividad y el bienestar general en el entorno laboral.

IMPLICACIONES EN LA ORGANIZACIÓN

La implementación de modelos de machine Learning para predecir la probabilidad de que una persona sufra un ACV, tiene repercusiones significativas en la organización, tanto del punto de vista estratégico como operativo. En este contexto, se han evaluado previamente varios modelos como regresión logística, random forest y redes neuronales, destacando la Regresión Logística como el más óptimo para este caso de estudio. Si bien el recall nos indica que la sensibilidad para la identificación de estos casos es bastante buena, los niveles de precisión son bajos, estos miden la proporción de verdaderos positivos entre todas las predicciones positivas.

Lo que sugiere que el modelo identifica un bajo porcentaje de verdaderos ACV entre todos los casos que predice como positivos, lo cual implica un alto número de falsos positivos. Esto es bastante relevante, ya que, si la organización depende únicamente de esta predicción para la implementación de medidas preventivas o de diagnóstico, podría estar invirtiendo recursos en personas que no requieren atención inmediata.

Por otra parte, el recall (0.814) es bastante alto, lo que indica que el modelo es eficaz en identificar a la mayoría de los pacientes que realmente sufrirán un ACV. En el área de la salud, un recall es muy importante porque la prioridad es detectar la mayor cantidad posible de casos positivos para poder intervenir a tiempo. Sin embargo, este desbalanceo de la precisión y el recall debe ser gestionado cuidadosamente, para evitar sobrecarga de recursos humanos y económicos de la organización, debido a la cantidad de falsos positivos.

A nivel estratégico, las organizaciones podrían utilizar este modelo como una herramienta predictiva que permita segmentar en grupos de riesgo a los pacientes y enfocar las medidas preventivas para aquellos que tenga una mayor probabilidad de sufrir un ACV, esto con la evaluación de los factores de riesgos más significativos de esta enfermedad. Dentro de los cuales se encuentran la hipertensión y las enfermedades cardíacas, seguidos por el índice de masa corporal (IMC), los niveles de azúcar en sangre, el tabaquismo y la edad. Esto no solo optimizaría el uso de recursos, sino que también podría

mejorar la calidad de vida de los pacientes mediante la detección temprana, la prevención y un seguimiento personalizado.

Un punto a considerar, es que la mayoría de factores de riesgo que más se asocian a un ACV son factores modificables, que quiere decir que pueden mejorar a través del tiempo mediante medidas preventivas y mejoras en el estilo de vida del paciente. Para una segmentación de pacientes a nivel general se podría tomar en cuenta, las siguientes:

- Programas y/o campañas de educación y concientización de la salud.
- Promoción de hábitos saludables.
- Guías prácticas para fomentar el autocuidado.
- Políticas de estilo de vida en el trabajo.

Ahora bien, esta segmentación se podría llevar de la mano con algún reglamento o protocolo clínico con validaciones médicas adicionales. Esto permitiría un discernimiento más preciso antes de proceder con diagnósticos o intervenciones preventivas personalizadas. Dado que estas personas presentan una mayor propensión a sufrir un ACV y las medidas preventivas para estos pacientes deben ser más drásticas, especializadas y con un mejor control y seguimiento. Para este grupo menos numeroso, la organización puede incluir estrategias distintas como:

- Monitoreo médico regular.
- Cambios de estilo de vida dedicados y dirigidos.
- Intervenciones psicológicas y apoyo social, personal y/o familiar.

En resumen, las implicaciones organizacionales de implementar un modelo predictivo para identificar el riesgo de ACV son profundas. No solo se potencia la capacidad de la organización para intervenir de manera temprana y efectiva, sino que también se optimizan los recursos y se mejora la calidad de la atención. A través de un enfoque basado en datos, se logra una mayor precisión en la identificación de pacientes en riesgo, lo que permite personalizar las intervenciones y, en última instancia, mejorar los resultados de salud. Esta estrategia no solo beneficia a los pacientes, sino que también

fortalece la posición de la organización como líder en la prevención y tratamiento de enfermedades críticas.

CONCLUSIONES

Si bien varios modelos muestran buen desempeño en diferentes métricas, la elección del más adecuado depende de la naturaleza específica del problema que enfrentamos. En nuestro caso, buscamos identificar personas con mayor riesgo de sufrir un accidente cerebrovascular. Es crucial que el modelo pueda detectar correctamente la categoría de alto riesgo. Por este motivo, hemos optado por evaluar los modelos utilizando la métrica de recall. Aunque esto podría resultar en una disminución en precisión y exactitud, la regresión logística ha demostrado ser efectiva en la identificación de esta clase menos común. Esto garantiza una captura más completa de las personas con mayor probabilidad de desarrollar esta condición médica, lo cual es fundamental para nuestros objetivos específicos.

Al examinar las variables que tienen mayor incidencia en nuestro modelo, observamos que están consistentemente respaldadas por la literatura y la teoría revisada. Por ejemplo, factores como el sobrepeso, problemas cardíacos y el hábito de fumar regularmente han demostrado estar significativamente asociados con un mayor riesgo de accidente cerebrovascular. Estos hallazgos no solo validan nuestra selección de variables, sino que también refuerzan la importancia de considerar estos factores de riesgo en la evaluación y predicción de la probabilidad de sufrir un evento cerebrovascular.

La interpretación de modelos de Machine Learning, como la regresión logística aplicada en nuestro trabajo, desempeña un papel crucial al encontrar qué variables y características tienen un mayor impacto en la predicción del riesgo de accidente cerebrovascular. Esta capacidad no solo permite identificar factores de riesgo significativos, sino que también facilita la generación de insights clínicos profundos. Con una comprensión más clara de estos modelos, y con diversos estudios adicionales, los profesionales de la salud pueden tomar decisiones más informadas y estratégicas para mejorar la atención y el manejo de pacientes con riesgo de accidente cerebrovascular, promoviendo así mejores resultados y una atención más personalizada.

RECOMENDACIONES

El machine learning es una herramienta potente para la toma de decisiones, pero su efectividad depende en gran medida de comprender todo el contexto empresarial circundante. En este caso, hemos priorizado un modelo con alto recall, que es crucial para identificar correctamente los verdaderos positivos. Sin embargo, es importante reconocer que en otros escenarios el costo de falsos positivos puede superar ampliamente los beneficios de una alta sensibilidad. Por esta razón, se recomienda siempre tener un conocimiento profundo del entorno completo del problema y considerar todas las implicaciones y variaciones posibles que puedan surgir. Esto garantiza una implementación efectiva y ajustada a las necesidades específicas de cada situación.

Se recomienda desarrollar herramientas de apoyo a la decisión basadas en Machine Learning (ML) para la identificación temprana y el manejo proactivo del riesgo de accidente cerebrovascular. Se espera que la implementación de estas herramientas genere beneficios como la reducción de la incidencia de accidentes cerebrovasculares, la mejora de los resultados clínicos, la optimización de los recursos sanitarios y el empoderamiento de los profesionales de la salud.

REFERENCIAS

Aracena, C., Villena, F., Arias, F., & Dunstan, J. (2022). Aplicaciones de aprendizaje automático en salud. *Revista Médica Clínica Las Condes*, *33*(6), 568–575.

Arias, A. (2022). *Aplicación de técnicas estadísticas para modelos de clasificación supervisada con muestras de datos desbalanceadas.* [Máster en Técnicas Estadísticas]. Universidade da Coruña.

Avellán, S., Holguín, C., & Cruz, M. del R. (2022). Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo. *Serie Científica de La Universidad de Las Ciencias Informáticas*, *15*(8), 37–50.

Bender del Busto, J. (2019). Las enfermedades cerebrovasculares como problema de salud. *Revista Cubana de Neurología y Neurocirugía*, *9*(2), 1–7.

Blanc-Pihuave, G., Cevallos-Torres, L., & Arteaga-Vera, J. (2020). Modelo computacional de clasificación de aprendizaje de máquina supervisado, para el análisis de datos cardiovasculares y pronóstico médico. *ECUADORIAN SCIENCE JOURNAL*, *4*(2), 71–79.

Bustan, M., & Poerwanto, B. (2021). Logistic Regression Model of Relationship between Breast Cancer Pathology Diagnosis with Metastasis. *Journal of Physics: Conference Series*, *1752*.

Campos, L., Sánchez, D., & Abuchar, A. (2019). Machine Learning y el control de hipertensión arterial. *Revista Hashtag*, 47–58.

Castillo, B., Tarazona, J., Tarazona, C., Hurtado, C., & Cornelio, F. (2023). Automatización del análisis exploratorio de datos y procesamiento geoquímico univariado empleando Python. *Revista Del Instituto de Investigación de La Facultad de Minas, Metalurgia y Ciencias Geográficas.*, *26*(51).

Daniel Moreno-Zambrano, C., Moreno-Zambrano, D., Santamaría, D., Ludeña, C., Barco, A., Vásquez, D., & Santibáñez-Vásquez, R. (2016). Enfermedad

Cerebrovascular en el Ecuador: Análisis de los Últimos 25 Años de Mortalidad, Realidad Actual y Recomendaciones. In *Revista Ecuatoriana de Neurología 17 Rev. Ecuat. Neurol* (Vol. 25, Issue 3).

Enciclopedia Médica A.D.A.M. (2024, February 28). *Manejo de su glucemia*. Https://Medlineplus.Gov/Spanish/Ency/Patientinstructions/000086.Htm.

Eske, J. (2021). *Cómo medir los niveles normales de glucosa en la sangre.* Https://Www.Medicalnewstoday.Com/Articles/Es/Prueba-de-Glucosa-En-Sangre.

Fede, S. (2021). *Kaggle*. Stroke Prediction Dataset.

Feigin, V. L., Stark, B. A., Johnson, C. O., Roth, G. A., Bisignano, C., Abady, G. G., Abbasifard, M., Abbasi-Kangevari, M., Abd-Allah, F., Abedi, V., Abualhasan, A., Abu-Rmeileh, N. M., Abushouk, A. I., Adebayo, O. M., Agarwal, G., Agasthi, P., Ahinkorah, B. O., Ahmad, S., Ahmadi, S., … Murray, C. J. L. (2021). Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Neurology*, *20*(10), 795–820. https://doi.org/10.1016/S1474-4422(21)00252-0

Hernández, J., & Orlandis, N. (2021). Índice de masa corporal elevado y la predicción de disglucemias. *Revista Cubana de Endocrinología*, *31*(3), 1–12.

Jerez, M., & Madero-Cabib, I. (2021). Trayectorias de estrés familiar y laboral y su asociación con accidentes cerebrovasculares. *Saúde Pública*, *55*.

López, A., Sánchez, E., & Loeza, C. (2023). Identificación efectiva de patologías ginecológicas aplicando Random Forest. Caso: Hospital de la mujer, Tabasco, México. *Komputer Sapiens: Revista de Divulgación de La Sociedad Mexicana de Inteligencia Artificial*, *2*, 1–11.

Martínez, J., & Pérez, P. (2024). Regresión logística. *Medicina de Familia. SEMERGEN*, *50*(1).

Olascoaga, L., & Ascue, S. (2020). *Desarrollo de un algoritmo con redes neuronales para la predicción de ACV en pacientes diabéticos.* Universidad Autónoma del Perú.

Organización Mundial de la Salud / OMS. (2023). *Diabetes*. Https://Www.Who.Int/Es/News-Room/Fact-Sheets/Detail/Diabetes.

Organización Mundial de la Salud / OMS. (2024). *Obesidad y sobrepeso.* Https://Www.Who.Int/Es/News-Room/Fact-Sheets/Detail/Obesity-and-Overweight.

Piloto, A., Suarez, B., Belaunde, A., & Castro, M. (2020). La enfermedad cerebrovascular y sus factores de riesgo. *Revista Cubana de Medicina Militar*, *49*(3). http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0138-65572020000300009&lng=es&tlng=es.

Ramos, D., Toapanta, M., & Alcívar, C. (2024). Predicción de accidente cerebrovascular utilizando regresión logística. *Revista Científica Interdisciplinaria Investigación y Saberes*, *14*(1), 158–177.

Rodríguez, W. (2024). *Análisis de los Factores de Riesgo del Bajo Peso al Nacer mediante Modelos de Regresión Logística Binomial y Polinomial, Centro de Salud Toribia Castro Chirinos. Lambayeque-2016* [Doctorado en Ciencias de la Salud]. Universidad Nacional Pedro Ruiz Gallo.

Sarmiento, J. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Revista UIS Ingenierías*, *19*(4), 1–18.

Yangón, G., & Reyes, A. (2023). *Aplicación de estrategias de transformación digital e innovación en el sector empresarial del Ecuador* [Escuela de Negocios]. Universidad de las Américas.

Yupari, I., Bardales, L., Rodriguez, J., Barros, S., & Rodríguez, Á. (2021). Factores de riesgo de mortalidad por COVID-19 en pacientes hospitalizados: Un modelo de regresión logística. *Revista de La Facultad de Medicina Humana*, *21*(1), 19–27.

Yurieski Pérez, Alián Pérez, & Alberto Caballero. (2023). Caracterización de pacientes con enfermedad cerebrovascular isquémica atendidos en el

Hospital General Docente Guillermo Domínguez de Las Tunas. *Revista Finlay*, *13*(1).