FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

DESARROLLO DE UN ALGORITMO CON REDES NEURONALES PARA LA PREDICCIÓN DE ACV EN PACIENTES DIABÉTICOS

PARA OBTENER EL TÍTULO DE

INGENIERO DE SISTEMAS

AUTORES

LUIS ALONSO OLASCOAGA ROMAN

ORCID: 0000-0002-6815-5733

SAÚL SEBASTIAN ASCUE SILVA

ORCID: 0000-0002-4718-6602

ASESOR

ING. VICTOR MANUEL GUEVARA PONCE

ORCID: 0000-0003-1787-7549

LÍNEA DE INVESTIGACIÓN

DESARROLLO DE SOFTWARE

LIMA, PERÚ, SETIEMBRE DE 2020

DEDICATORIA

A nuestras familias que con el esfuerzo y dedicación han hecho que nuestros sueños se hagan realidad, apoyándonos y brindándonos un camino de superación; a su apoyo incondicional en todo momento.

AGRADECIMIENTOS

Agradecemos al Ing. Olascoaga Zavaleta Anderson, por brindarnos el equipo de software y hardware pertinente para la presente investigación, a nuestros amigos de la facultad de ingeniería y a nuestros docentes por su gran apoyo.

ÍNDICE

LISTA DE TABLAS

INDICE DE FIGURAS

DESARROLLO DE UN ALGORITMO CON REDES NEURONALES PARA LA PREDICCIÓN DE ACV EN PACIENTES DIABÉTICOS

LUIS ALONSO OLASCOAGA ROMAN
SAÚL SEBASTIAN ASCUE SILVA

UNIVERSIDAD AUTÓNOMA DEL PERÚ

RESUMEN

El objetivo de esta investigación fue desarrollar un algoritmo basado en redes neuronales para la predicción de ACV focalizado en pacientes diabéticos donde se tuvo como indicadores de predicción a la precisión, accuracy, sensibilidad y especificidad. El diseño de investigación fue transversal con un nivel de investigación correlacional. La data usada fue de 17372 registros Recuperados del repositorio Kaggle los cuales estaban conformados por 9 variables. El instrumento usado para la confiabilidad fue el Alfa de Cronbach el cual determinó una homogeneidad del 94%, dicho instrumento fue empleado para determinar el mínimo porcentaje de predicción aceptable. La investigación fue desarrollada con la metodología SEMMA (sample, explore, modify, model, assess), la cual permitió la creación del modelo de red neuronal, también se empleado la librería H2O para generar el modelo de Deep Learning. Los resultados que arrojó el modelo de red neuronal superaron el mínimo aceptable (88% para todos los indicadores), en donde la precisión obtuvo un 91%, accuracy 94%, sensibilidad 93% y la especificidad 94%. Finalmente, se obtuvieron mejores resultados de los previstos, dado que el modelo superó el mínimo aceptable.

***Palabras clave*:** redes neuronales, diabetes, accidentes cerebrovasculares.

DEVELOPMENT OF AN ALGORITHM WITH NEURAL NETWORKS FOR THE PREDICTION OF ACV IN DIABETIC PATIENTS

LUIS ALONSO OLASCOAGA ROMAN
SAÚL SEBASTIAN ASCUE SILVA

UNIVERSIDAD AUTÓNOMA DEL PERÚ

ABSTRACT

The objective of this research was to develop an algorithm based on neural networks for the prediction of focused stroke in diabetic patients where the precision, accuracy, sensitivity and specificity were taken as predictive indicators. The research design was cross-sectional with a level of correlational research. The data used was 17372 records retrieved from the Kaggle repository which were made up of 9 variables. The instrument used for reliability was Cronbach's Alpha, which determined a homogeneity of 94%. This instrument was used to determine the minimum percentage of acceptable prediction. The research was developed with the SEMMA methodology (sample, explore, modify, model, assess), which allowed the creation of the neural network model, the H2O library was also used to generate the Deep Learning model. The results obtained by the neural network model exceeded the minimum acceptable (88% for all indicators), where the precision obtained 91%, accuracy 94%, sensitivity 93% and specificity 94%. Finally, better results than expected were obtained, since the model exceeded the minimum acceptable.

***Keywords*:** neural network, diabetes, stroke.

INTRODUCCIÓN

El accidente cerebrovascular (ACV) es la tercera discapacidad y el segundo factor de mortalidad en todo el mundo, debido a que afecta a más de 15 millones de habitantes; y la pequeña población no mayor a los 5 millones que logra sobrevivir a un ACV padece de secuelas irreversibles que lo acompañaran a lo largo de vida.

La investigación abordó el tema de accidentes cerebrovasculares, el cual puede presentarse de 2 formas; ACV isquémico o hemorrágico, ambos pueden causar secuelas físicas o mentales hasta incluso ocasionar la muerte; sin embargo, hablar del ACV puede resultar muy extenso dado que presenta una gran variedad de múltiples factores, la investigación se centró en un grupo específico de pacientes diabéticos los cuales son más vulnerable o tienen mayor probabilidad de sufrir uno o dos ACV a lo largo de su vida si no tienen los cuidados y tratamientos necesarios.

Es importante mencionar que el ACV tiene 7 factores claves los cuales incrementan la probabilidad de sufrirlo; cualquier persona tiene la probabilidad de padecer un ACV, por ello esta investigación se centró en tales factores; los cuales son, el género, la edad, historial familiar, hipertensión, tabaquismo, diabetes, presión arterial.

El presente proyecto de investigación aplicó el uso de las redes neuronales, las cuales son perfectas para alinear grandes grupos de información y encontrar patrones ocultos mediante procesos de activación y el backpropagation (retro propagación). En esta situación se creó un modelo basado en redes neuronales mediante el Deep Learning con el objetivo de predecir con el mínimo error la probabilidad que una persona pueda padecer un ACV según su historial médico (Anamnesis).

Los resultados del modelo de redes neuronales arrojaron un accuracy del 94%; un 91% de precisión en cuanto a la predicción de ACV en los pacientes diabéticos, con una sensibilidad del 93% y una especificidad del 94%.

Esta investigación se realizó con una división en diferentes capítulos, los cuales se explica a continuación:

Capítulo I: Planteamiento metodológico; en este capítulo se abordó la descripción del problema, la realidad problemática, definición del problema y enunciado del problema; así como también se definió el tipo y nivel de investigación, la justificación, objetivos y la hipótesis.

Capítulo II: Marco referencial; llegados a este capítulo detallamos los antecedentes que tuvieron las relaciones con la investigación, así como las bases teóricas sobre las cuales fue basado esta investigación.

Capítulo III: Desarrollo de la solución; para este capítulo se hizo uso de la metodología SEMMA, la cual permitió el desarrollo de este modelo de red neuronal, así como también se hizo uso de la librería H2O para generar dicho modelo**;** en el transcurso de este capítulo se apreció como se va haciendo paso a paso el desarrollo**.**

Capítulo IV: Análisis de resultados y contrastación de hipótesis; en este capítulo se definió la población, así como también la muestra, el tipo de muestreo y la confiabilidad del instrumento usado. Los resultados fueron analizados mediante estadística descriptiva y para culminar el capítulo se hizo la contratación de la hipótesis.

Capítulo V: Discusiones, conclusiones y recomendaciones; llegados a este capítulo se detalló las discusiones basadas en los antecedentes; se detallaron las conclusiones y recomendaciones de la investigación.

CAPÍTULO I

PLANTEAMIENTO METODOLÓGICO

1.1 El problema

1.1.1 Realidad problemática

Realidad mundial

Actualmente un accidente cerebrovascular (ACV) es la tercera discapacidad y la segunda causa mortal en todo el mundo; afecta cada año a más de 15 millones de personas, de estos más de 5 millones de personas fallecen y otros pocos sobreviven con secuelas irreversibles. En los países sub desarrollados de medio o bajos ingresos económicos, durante los últimos años, la incidencia del ACV se ha incrementado considerablemente.

Realidad nacional

En nuestro país se registra a personas mayores de 65 años a más, con casos de ACV en donde el 6,8% son de zonas urbanas y 2,7% en zonas rurales, en donde este grupo representan el 29,2 % y 14,7% respectivamente, causas de muerte en nuestro país. Uno de los problemas resaltantes a la hora de tratar con el manejo de información de ACV por primera vez es la ausencia de un oportuno y adecuado diagnóstico, dando a conocer que entre 14% y 26% de personas con ACV mayores a 65 años de las zonas urbanas y rurales respectivamente, no son diagnosticados.

1.1.2 Definición del problema

Hoy en día no existe a ciencia cierta algo que permitan saber que tan relacionados están los ACV con respecto a los pacientes diabéticos y que complicaciones relacionadas a la diabetes pueden determinar qué tan propenso es un paciente diabético en sufrir ACV.

La mayoría de las personas que sufren ACV no poseen una cultura preventiva con respecto a su salud, la falta de chequeos anuales para descartar cualquier tipo

de enfermedad ocasiona que no puedan detectar a tiempo las complicaciones iniciales de ACV y su posterior tratamiento.

Por lo que en la actualidad se hace uso de herramientas tecnológicas de inteligencia artificial predictiva en el área de la medicina humana haciendo que sea una de las mejores opciones para poder ayudar al diagnóstico preventivo o en la detección de la etapa inicial del ACV.

1.1.3 Enunciado del problema

¿En qué medida el desarrollo de un algoritmo basado en redes neuronales contribuirá en la predicción de ACV en pacientes diabéticos?

1.2 Tipo y nivel de investigación

1.2.1 Tipo de investigación

Aplicada

Busca cómo es posible utilizar o aplicar los conocimientos aprendidos mientras que a la vez se aprenden nuevos conocimientos, luego sistematizar e implementar la practica obtenida en la investigación. El uso de los resultados y de los conocimientos adquiridos de la investigación da como producto una nueva forma más sólida, sistemática y organizada de mirar la realidad del caso.

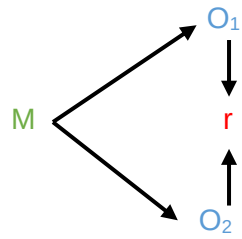
1.2.2 Nivel de investigación

Correlacional

Tiene como objetivo calcular la magnitud relacional entre las variables de una población de datos. Se busca visualizar la relación entre varios fenómenos, o comprobar que no existe relación entre ellos. La prioridad es entender y saber cómo es que las variables relacionadas se comportan entre sí; A continuación, presentamos el diseño correlacional en la tabla 1:

Tabla 1

Diagrama *de diseño correlacional*

Relación	Variables
	**M =** Muestra **O_1 =** Variable 1 **O_2 =** Variable 2 **r =** relación entre las dos variables

Fuente: Slideshare (s.f.)

1.3 Justificación de la investigación

Tecnológica

Se justifica tecnológicamente debido al desarrollo de un algoritmo basado en redes neuronales, trae consigo la exploración y el uso de nuevas herramientas que pueden mejorar la calidad de la investigación dando resultados muchos precisos.

La minería de datos es el procesamiento de una gran cantidad de datos con el objetivo de predecir las tendencias y patrones de los datos mediante el análisis matemático. Por lo general los patrones que son muy complejos no se pueden detectar mediante la exploración normal de los datos.

Metodológica

El desarrollo de un algoritmo basado en redes neuronales trae consigo nuevos métodos y formas de investigación que contribuyen con la obtención de información relevante y mucho más exacta que pueda ser usado en posteriores investigaciones.

Social

Justificamos de manera social el desarrollo de algoritmo basado en redes neuronales, con la finalidad de generar una herramienta que permita a las organizaciones o entidades trabajar con nuestro modelo y así ayudar a la sociedad a prevenir y diagnosticar con antelación la diabetes mejorando el estilo de vida de las personas y su calidad; Una gran mayoría de países no se constatan del contratiempo

social y económico de la diabetes. Esta carencia de comprensión es el mayor obstáculo para las estrategias de prevención efectivas que pueden aportar a reducir de manera significativa del incremento de la diabetes tipo 2.

1.4 Objetivos de la investigación

1.4.1 Objetivo general

Determinar en qué medida el desarrollar un algoritmo basado en redes neuronales contribuye en la predicción de ACV en pacientes diabéticos.

1.4.2 Objetivos específicos

- Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de precisión.
- Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de accuracy.
- Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales determinara el margen de sensibilidad.
- Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales determinara el margen de especificidad.

1.5 Hipótesis

1.5.1 Hipótesis general

Si se desarrolla un algoritmo basado en redes neuronales, entonces se podrá realizar la predicción de ACV de los pacientes diabéticos.

1.5.2 Hipótesis específica

- El desarrollo de un algoritmo basado en redes neuronales permite la identificación del índice de precisión.
- El desarrollo de un algoritmo basado en redes neuronales permite la identificación del índice de accuracy.

- El desarrollo de un algoritmo basado en redes neuronales permite la identificación del margen de sensibilidad.
- El desarrollo de un algoritmo basado en redes neuronales permite la identificación del margen de especificidad.

1.6 Variables e indicadores

1.6.1 Variable independiente

- El uso de redes neuronales

Tabla 2

Conceptualización de la variable independiente.

Indicador: Presencia - Ausencia
Cuando se hace uso del algoritmo basado en redes neuronales en la predicción de ACV en un grupo de estudio y cuando no existe presencia del mismo en otro grupo

Tabla 3

Indicador variable independiente.

Indicador	**Índice**
Presencia – Ausencia	Si - No

1.6.2 Variable dependiente

- Predicción de ACV en pacientes diabéticos

A continuación, presentamos en la tabla 4 los indicadores de la variable dependiente; así mismo en la tabla 5 presentamos los Indicadores de la variable dependiente.

Tabla 4

Indicador de la variable dependiente – todas las variables

Indicador	Descripción
Stroke	Derrame cerebral ocasionado por una trombosis (obstrucción de una arteria) o una hemorrágico (ruptura de una arteria) ocasionados en el cerebro.
Hipertensión	Es el aumento de la fuerza de presión que ejerce la sangre.
Glucosa	Es la medicación de la cantidad o nivel de glucosa que se encuentra en la sangre.
Presión Arterial	Es la presión de la sangre al empujar las paredes de las arterias y venas.
Sexo	Condición de distinción entre las personas.
Edad	Tiempo transcurrido desde el nacimiento de una persona.
Enfermedades cardiacas	Afecciones en los vasos sanguíneos, arterias coronarias, ritmo y defectos cardiacos.
Historial de Fumador	Registro sobre la progresión o disminución de consumo de tabaco
BMI (índice de masa corporal)	Es la asociación de la masa y talla de una persona.
Precisión	Determina la fracción de registros que realmente resulta ser positivo en el grupo que el clasificador ha declarado como una clase positiva.
Accuracy	Es la sumatoria de los registros con resultados correctos y erróneos en el modelo.
Sensibilidad	Corresponde a la fracción de registros positivos predichos correctamente en el modelo.
Especificidad	Corresponde a la fracción de registros negativos predichos correctamente en el modelo.

Tabla 5

Indicador de la variable dependiente

Indicador	Índice	Unidad de Medida	Unidad de Observación
hipertensión	[0 – 1]	[0 – 1]	Reporte
glucosa	[80 - 200]	Miligramos/ Decilitros (mg/dl)	Reporte
Presión Arterial	[60 – 200]	mmHg (mililitros de mercurio)	Reporte
Sexo	[Hombre-Mujer]	1 - 2	Reporte
Edad	[20 - 80]	años	Reporte
BMI	[20.00 - 88.72]	Kg/cm^2	Reporte
Enfermedades cardiacas	[0 – 1]	[1 – 0]	Reporte
Historial de Fumador	[never, not current, current former ever]	[1 – 5]	Reporte
Stroke	[0 – 1]	Presencia / Ausencia	Reporte
Precisión	> 88%	Porcentaje	Reporte
Accuracy	> 88%	Porcentaje	Reporte
Sensibilidad	> 88%	Porcentaje	Reporte
Especificidad	> 88%	Porcentaje	Reporte

1.7 Limitaciones

Temporal: El presente trabajo de investigación se realiza durante el periodo comprendido entre agosto del 2019 hasta julio del 2020.

Conceptual: El presente trabajo de investigación tiene como delimitación conceptual los modelos algorítmicos de redes neuronales, así como a la predicción de ACV en pacientes diabéticos.

1.8 Diseño de la investigación

Transversal

La investigación o estudio transversal tiene como definición un tipo de investigación observacional el cual tiene habilidad de analizar datos de variables recopiladas de una población, el cual dichos datos se seleccionan en función en variables particulares de interés; ya sea una muestra o subconjunto predeterminado en un periodo de tiempo. A continuación, presentamos en la tabla 6 los diseños de investigación.

Tabla 6

Diseño de la investigación

P	G	O
Población de pacientes diabéticos	clasificación de subgrupos de control experimental (train – test)	tratamiento experimental de los datos

Donde:

P (Población): cantidad total de pacientes con diferentes historiales médicos.

G (sub grupo o clasificación): Es el grupo de estudio o muestra.

O: Son los valores de los indicadores de la variable dependiente que serán observados.

Descripción:

La investigación inicia con una población (P), la cual será subclasificada en sub grupos (G) el cual está conformado por los grupos de entrenamiento (train) y de prueba (test); el cual está conformado por una base de datos de diferentes pacientes

diabéticos (con ACV y sin ACV), cuyos indicadores serán sometidos a una observación (O).

1.9 Técnicas e instrumentos para la recolección de información

Tabla 7

Técnicas e instrumentos de la investigación

Técnica	Aplicación	Instrumentos	Método	Indicador
Revisión de documentos	Historial del paciente (Anamnesis)	Reportes	1. Seleccionar los documentos a revisar. 2. Registrar datos relevantes. 3. Analizar la información.	• Stroke • Hipertensión • Glucosa • Enfermedades cardiacas • Presión arterial • BMI • Historial de fumador • Edad • Sexo
Observación	Resultados de la investigación	Reportes	1. Exploración de reportes.	• Precisión • Sensibilidad • Especificidad • Accuracy

CAPÍTULO II

MARCO REFERENCIAL

2.1 Antecedentes

Yang (2019) en su tesis titulada como *Study on the segmentation method of stroke in chronic stage*, tiene como objetivo lograr el diagnóstico preventivo de accidentes cerebrovasculares y la ubicación de las lecciones ocasionadas por estas, mediante la interpretación automática de las imágenes obtenidas por tomografías computarizadas (resonancia magnética T1 y T2) y procesadas por una red neuronal profunda para determinar si un paciente posee un accidente cerebrovascular leve o crónico.

Se implementará una arquitectura de red neuronal profunda (red de fusión profunda multiescala) de estructura piramidal encapsulada hueca para permitir para la extracción de diferentes patrones de escala de imágenes tomografías. Se hará uso de un método de red neuronal de codificador – decodificador profundo entre extremos, lo que garantizará una resolución del mapa de características.

Los resultados arrojaron que esta arquitectura de red neuronal profunda obtuvo un 58.10% de precisión, 4.10% más que la arquitectura U-Net que es la utilizada generalmente para la segmentación semiautomática de las imágenes tomografías.

Como conclusión la arquitectura de red neuronal profunda ofrece una mejor precisión a la hora de ubicar las lecciones por accidente cerebrovascular, ayudando al diagnóstico rápido y al tratamiento de ACV.

Zhanfeng (2016) en su tesis con el título de *Stroke evaluation and reflection with artificial neural network in haptic virtual environment* tiene como objetivo general poder evaluar los métodos de clasificación y de evaluación de los pacientes con accidente cerebrovasculares para determinar si se encuentran aún con signos de secuelas de ACV o si ya se encuentran sanos, también tiene objetivo diseñar y crear

un modelo de red neuronal capaz de evaluar y establecer los métodos de clasificación de los pacientes con ACV de manera automática y precisa.

Se construirá un modelado de red neuronal BP combinado con un entorno virtual, cual mediante el testeo de 2 acciones físicas específicas determinara la condición actual del paciente, la arquitectura de la red neuronal tendrá una capa oculta de 7 neuronas basada en la teoría de Brunnstrom. La prueba del modelo de red neuronal fue a una muestra de 37 pacientes para el muestreo de entrenamiento de la red neuronal en estado de rehabilitación por secuelas de ACV.Se utilizó la curvatura ROC para determinar el rendimiento del modelo de clasificación del estado actual del paciente.

El área bajo la curva ROC (AUC) arrojo un 0.876 en la prueba de la clasificación de los pacientes aumentante de manera significativa la precisión de predicción de 92.1% a 94.12%.

Como conclusión general el uso de la red neuronal BP para la clasificación del estado de los pacientes con problemas de ACV logra determinar con mucha precisión y significancia.

Chunxiao (2019) en su investigación denominada “Research on cerebrovascular disease prediction system based on the long short term memory neural network” tiene como objetivo el diseño y creación de un modelo predictivo para incidencias tempranas de enfermedades cerebrovasculares ayudando a la mejora en el tratamiento médico y el diagnostico preventivo de esta enfermedad. Para este estudio se basó en la teoría para determinar los indicadores de los cuales se obtuvo de conjuntos generales de dominio en la escala de puntuación de riesgo de Essen y otras escalas para proponer un nuevo modelo de diagnóstico para los accidentes cerebrovasculares en base a el diseño de una red neuronal de memoria de corto y

largo plazo (LSTM) y el algoritmo de Adam probando dicho modelo en pacientes en la muestra de entrenamiento y otra cantidad pacientes en la muestra prueba en el modelo predictivo.

Tras la comparación del modelo actual y el modelo basado en redes neuronales se obtuvo como resultado una disminución considerable en la tasa de recurrencia y en la tasa de mortalidad en pacientes con altos índices de sufrir un ACV.

Como conclusión general, el uso de modelos predictivos basados de redes neuronales muestra un mejora significativa a la hora de ayudar con el diagnostico preventivo y su tratamiento.

Masruriyah et al. (2019) en su investigación denominada “Predictive analytics for stroke disease” tiene como objetivo desarrollar un análisis predictivo para la enfermedad cerebrovascular, en donde se emplea una analítica para convertir los datos sin filtrar el conocimiento estratégico mediante las redes neuronales; en donde la población fue de 18,425 pacientes utilizando datos de BioMed Central; se empleó las pruebas mediante la validación cruzada de K-Fold dando como resultado una predicción del 95,15%.

El aporte de esta investigación es el desarrollo de un modelo de diagnóstico o pronóstico predictivo para determinar con anticipación la ocurrencia de esta enfermedad.

Hung et al. (2018) en su investigación “Improving young stroke prediction by learning with active data augmenter in a large-scale electronic medical claims database” tienen como objetivo mejorar el desempeño de predicción del accidente cerebrovascular de 2 grupos (de 25 a 45 años y 45 a 85 años), esta data es obtenida de la base de datos electrónica de reclamaciones médicas (EMC), cuya población de estudio será de 552,898 sujetos; La investigación aborda el modelo DNN

el cual se basa en el aprendizaje automático correlacionando características complejas, este modelo consta de un input igual a 200 aplicando la función de activación TANH, adicionalmente se compararon los enfoques TGL, MGL, SDA y ADA l; así como también diferentes grupos de datos de entrenamiento (10%, 20%, 40%, 80%) donde el enfoque ADA obtuvo un 80.2% con un grupo de datos del 20%, siendo el mejor resultado y adicionalmente la mejora de desempeño es del 9.3% y una mejora en la curva AUC del 8.2%.

Como aporte la propuesta de este trabajo fue acotar un método para prevenir el desequilibrio de los datos, es decir se propone la manera de aumentar el rendimiento al combinar datos según su relación, de tal manera que la investigación se centra en ACV tempranos (jóvenes).

Kyriacou et al. (2015) el estudio denominado “Prediction of the time period of stroke based on ultrasound image analysis of initially asymptomatic carotid plaques” tiene por objetivo desarrollar modelos de predicción temprana (≤ 3 años) y a largo plazo (> 3 años) mediante vectores estadísticos (modelo predictivo SVM), los datos fueron extraído de manera estadistica para realizar la clasificación de los 2 grupos (ACV a largo y corto plazo). Para el desarrollo de la investigación se dividio la data en 3 grupos (FS1: características clínicas y de placas, FS2: características de textura, FS3: características de textura - morfología). llegando el grupo FS2 a una clasificación correcta de 77,7%; sensibilidad y especificidad de 76,8% y 79,11% respectivamente. Adicionalmente la ponderación de los 3 grupos da un resultado de clasificación del 78,7%; sensibilidad y especificidad de 88,6% y 72, 6% respectivamente.

El principal aporte de esta investigación es la incursión al reconocimiento de clasificación y predicción de distintos grupos dentro de intervalos de tiempo para determinar alertas tempranas.

2.2 Base teórico científicas

2.2.1 Definición de diabetes

La diabetes es una enfermedad silenciosa el cual consiste en que una persona posee un nivel elevado de azúcar en su sangre. Esta enfermedad no se cura, pero si se puede tratar para evitar complicaciones a corto o largo plazo y llevar un estilo de vida normal. En nuestro país más de 3 millones de personas padecen esta enfermedad y que más de la mitad de los que la padecen no se percatan y desconocen que poseen dicha enfermedad.

2.2.2 Definición del ACV (accidente cerebro – vascular)

A. Clasificación del ACV

Desde un punto de vista fisiopatológico, los accidentes cerebrovasculares pueden ser clasificados en dos tipos: el isquémico y el hemorrágico.

a. ACV isquémico

Arauz y Ruíz (2012) sostienen que:

En el ataque isquémico transitorio (AIT) no existe daño neuronal permanente. La propuesta actual para definir al AIT establece un tiempo de duración de los síntomas no mayor a 60 min, recuperación espontánea, ad-integrum y estudios de imagen (de preferencia resonancia magnética), sin evidencia de lesión. Estudios recientes muestran que los pacientes con AIT tienen mayor riesgo de desarrollar un infarto cerebral (IC) en las 2 semanas posteriores, por lo que se han diseñado escalas de estratificación de riesgo. La escala ABCD27 se basa en 5 parámetros (por sus siglas en inglés), a los que se asigna un puntaje de entre 0 y 2, de acuerdo a si está o no presente: A, edad (> 60 años = 1 punto); B, presión arterial (= 1); C, características clínicas (hemiparesia = 2, alteración

del habla sin hemiparesia = 1, otros = 0); D, duración del AIT (> 60 min = 2; 10-59 min = 1; < 10 min = 0); D, diabetes (2 puntos si está presente). (p. 12).

b. ACV hemorrágico

Graeme (2016) nos brinda algunos alcances importantes sobre:

Se clasifica de acuerdo con el sitio anatómico o la presunta etiología. La mayoría de los sitios comunes de hemorragia intracerebral son supratentoriales (85-95%), incluyendo la zona profunda (50-75%) y lobar (25-40%). Las causas más comunes son la hipertensión (30-60%), la angiopatía amiloide cerebral (10-30%), la anticoagulación (1-20%), y las lesiones vasculares estructurales (3-8%). Aproximadamente en el 5-20% de los casos, la causa no puede determinarse. (p. 1).

B. Factores de riesgo

Peñafiel (2018) Nos hace una introduccion sobre el ACV:

El ictus o accidente cerebrovascular es una patología producida por el bloqueo abrupto de un vaso sanguíneo cerebral, generalmente una arteria y en casos más raros una vena. La falta de oxígeno y nutrientes provoca la muerte neuronal. La isquemia cerebral se debe a causas diversas como trombosis, embolismo o hipoperfusión sistémica. La trombosis implica la oclusión arterial causada por enfermedades como ateroesclerosis, disección o displasia fibromuscular. En el embolismo la obstrucción se produce por un émbolo liberado desde un lugar lejano, que al llegar a los pequeños vasos cerebrales impide el flujo sanguíneo. La hipoperfusión sistémica puede afectar tanto a cerebro como a otros órganos, que reciben una cantidad inadecuada de oxígeno y nutrientes. Una causa primaria, poco común, de accidente

> cerebrovascular son las patologías hematológicas, como la hipercoagulabilidad, policitemia vera, anemia drepanocítica y síndrome antifosfolipídico. (p. 1).

Dentro de los principales factores de riesgo tenemos los siguientes factores.

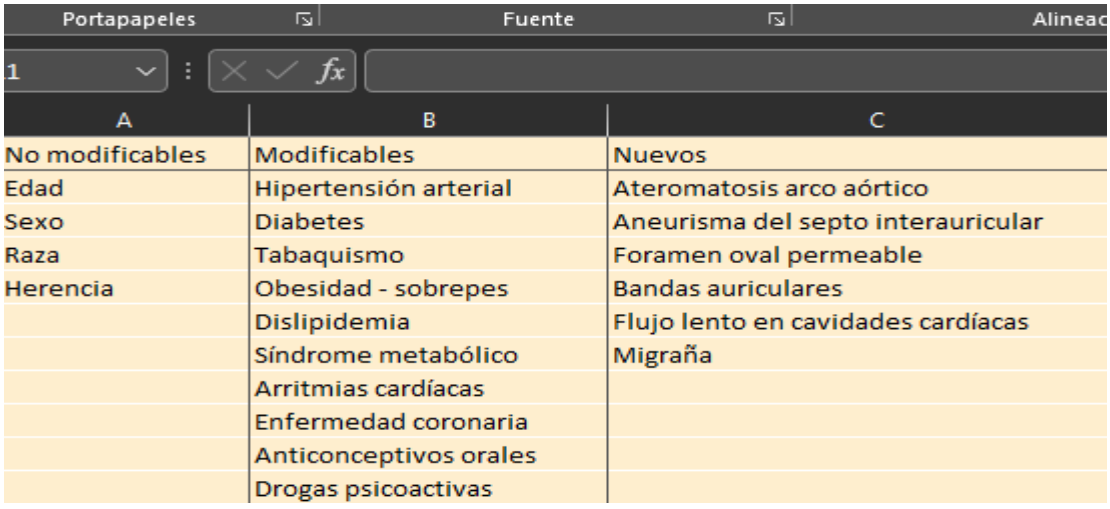
- **Edad:** 45 años a más, la probabilidad de padecer un ACV incrementa potencialmente.
- **Historia familiar**: es uno de los factores más cruciales, debido a que si algún familiar ha sufrido de ACV las probabilidades que las siguientes generaciones lo padezcan son muy altas.
- **Sexo:** Las mujeres tienen mayor riesgo de sufrir ictus debido a situaciones como el embarazo, historia de preeclampsia/eclampsia y diabetes gestacional, uso de anticonceptivos orales y terapia hormonal postmenopáusica.
- **Drogas:** es un asociado que estimula a un ACV.
- **Sueño:** Sobre exigir al cuerpo dándole pocas horas de reposo es un factor muy común.
- **Colesterol:** es un tipo de grasa que se encuentra en las células de nuestro cuerpo**.**
- **Diabetes**: Es el factor más resaltante por el que una persona sufre ACV.
- **Hipertensión:** es un factor que agravia la situación de una persona, siendo que, sumado otros factores mencionados previamente, puede resultar contraproducente en la recuperación o tratamiento preventivo.

El estudio realizado por los autores Celis, Hernandez y King Chio (2019) comprende los siguientes aspectos:

El centro nacional para la prevención de la enfermedad crónica y la promoción de la salud (CDC) en Estados Unidos, analizó los datos del sistema de vigilancia en el comportamiento de factores de riesgo (BRFSS) del 2003, con el fin de evaluar la prevalencia de los múltiples factores de riesgo para ACV e identificar las diferencias de riesgo entre los subgrupos de población (2). Este estudio encontró que 37 por ciento de la población tiene dos o más factores de riesgo para presentar un ACV y que existen diferencias considerables entre los grupos socioeconómico y la población racial y étnica. El estudio evalúo 256.155 participantes en 50 estados, mayores de 18 años, considerando factores de riesgo como hipertensión, hiperlipidemia, diabetes, tabaquismo, obesidad y sedentarismo. (p. 33).

Figura 1

Factores de riesgo de un ACV



A	B	C
No modificables	Modificables	Nuevos
Edad	Hipertensión arterial	Ateromatosis arco aórtico
Sexo	Diabetes	Aneurisma del septo interauricular
Raza	Tabaquismo	Foramen oval permeable
Herencia	Obesidad - sobrepes	Bandas auriculares
	Dislipidemia	Flujo lento en cavidades cardíacas
	Síndrome metabólico	Migraña
	Arritmias cardíacas	
	Enfermedad coronaria	
	Anticonceptivos orales	
	Drogas psicoactivas	

Fuente: Celis , J. I., Hernandez, D. L., & King , L. M. (2019)

2.2.3 Funciones de activación

Guo et al. (2019) definen sobre las funciones de activacion:

La función de activación es una parte importante de la red neuronal [3]. La función de activación se refiere a cómo retener y mapear las características de

las "neuronas activadas" a través de funciones no lineales, que es la clave para resolver problemas no lineales en redes neuronales. Cuando la función de activación es lineal, las combinaciones lineales de ecuaciones lineales solo tienen la capacidad de expresión lineal, incluso si la red tiene muchas capas, solo la red lineal con una capa oculta [15]. Esta red de representación lineal de entrada es solo el equivalente de un perceptrón multicapa. Y eso no será una aproximación no lineal a ninguna función. Como el rendimiento está lejos de ser requerido, intente usar una combinación no lineal. El uso de la función de activación aumenta la no linealidad del modelo de red neuronal y resuelve el problema de la expresión insuficiente de la operación lineal, lo que hace que la red neuronal profunda tenga un significado real [7]. La investigación muestra que las diferentes funciones de activación tienen una gran influencia en el rendimiento de la red. (p. 3582).

A. Función escalonada

Determina si un valor de entrada (input) es mayor al umbral, en donde si es mayor se le asigna el valor 1 y si es mejor el valor 0.

$$f(x) = \begin{cases} 0 \text{ for } x < 0 \\ 1 \text{ for } x \geq 0 \end{cases}$$

B. Función sigmoide

Genera una distorsión a los valores, en donde los más grandes tienden a 1 y los más pequeños tienden a 0; generando una deformación más pronunciada que favorece el aprendizaje de la red neuronal.

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$$

C. Función RELU (unidad rectificado lineal)

Es la función más común, se comporta de 2 maneras; de forma lineal cuando los valores son positivos y tiende a ser constante cuando el valor es negativo.

$$f(x) = \begin{cases} 0 \text{ for } x < 0 \\ x \text{ for } x \geq 0 \end{cases}$$

D. Función TANH (tangente hiperbólica)

Similar a la función sigmoide con la diferencia que sus valores varían de -1 hasta 1.

$$f(x) = thanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

2.2.4 Gráficos de análisis descriptivo

A. Gráfico de cajas y bigotes

Palladino (2011) Expone lo siguiente con respecto a los gráficos de cajas y bigotes:

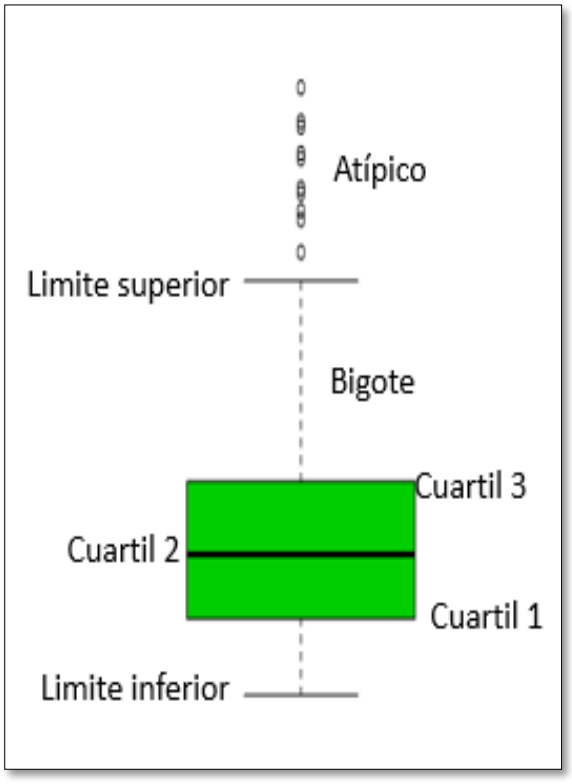
> El gráfico de caja ("box-plot" en inglés) es una forma de presentación estadística destinada, fundamentalmente, a resaltar aspectos de la distribución de las observaciones en una o más series de datos cuantitativos. Reemplaza, en consecuencia, al histograma y a la curva de distribución de frecuencias sobre los que tiene ventajas en cuanto a la información que brinda y a la apreciación global que surge de la lectura. Fue ideado por John Tukey, de la Universidad de Princeton (U.S.A.) en 1977 y los detalles que siguen corresponden a la descripción dada por este autor. Cabe destacar que en diferentes textos (y presentaciones del gráfico) se utilizan de manera diferente

a las señaladas por su creador algunos elementos de la presentación; lo que, en lo posible, se aclara en este documento. (p. 1).

El presente gráfico de cajas y bigotes es una forma visual de ver la distribución de los datos, la dispersión y la simetría; la caja muestra la acumulación de los datos en donde se puede hallar desde el 1er hasta el 3er cuartil de los mismos, mientras que los bigotes muestran una menor cantidad de datos (límite superior y límite inferior); además, este tipo de gráficos nos muestra donde se ubican los datos atípicos (si hubiera); otra ventaja de este gráfico es la de concatenar diferentes grupos para hacer una comparación general.

Figura 2

Modelo de cajas y bigotes

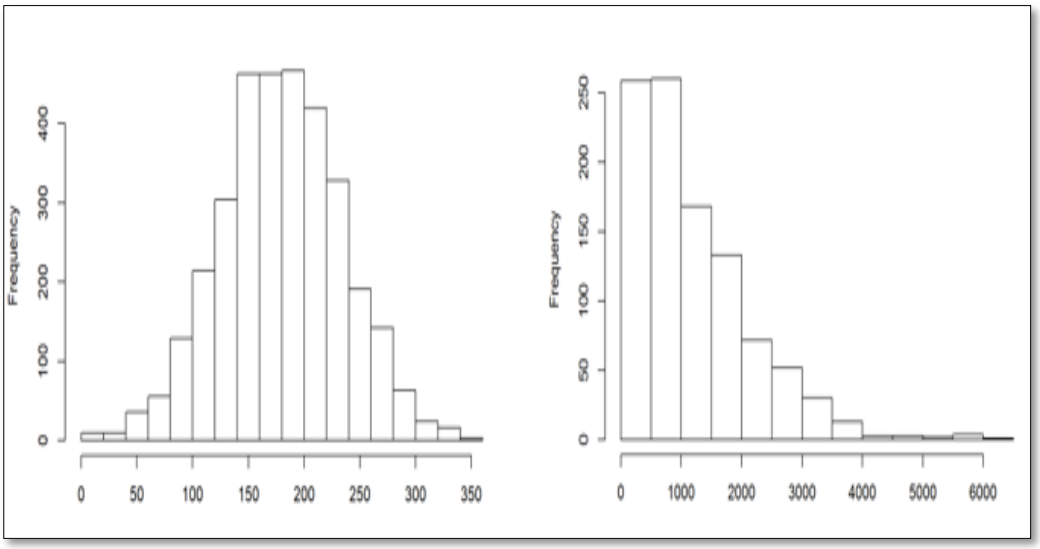


B. Histogramas

Gutiérrez y Cintas (2013) nos dan la siguiente definición sobre los histogramas: Por comodidad, generalmente se toman los intervalos de clase del mismo ancho y se omite el concepto de densidad empírica, pues en caso de intervalos de igual ancho, la forma del histograma es idéntica, si se toma como ordenada la densidad o si se asume como la frecuencia relativa. El software de estadística, refuerza esta costumbre, pues por defecto hace gráficos de histograma con intervalos del mismo ancho. Introduciendo el tema de la representación gráfica de los datos, usando intervalos de anchura desigual, se produce una ganancia conceptual importante, pues obliga a la representación del histograma como rectángulos que tienen como base el intervalo de clase y su área proporcional (o igual) a la frecuencia relativa. (p. 230).

Figura 3

Modelo de histogramas



2.2.5 Fórmulas para la identificación de la predicción

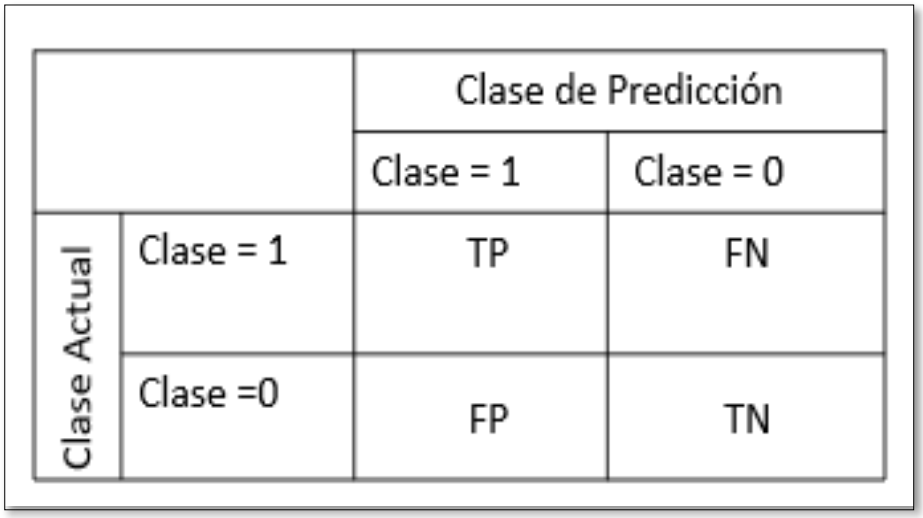
Según Tan, Steinbach y Kumar (2013) Afirman que:

> Dado que la medida de precisión trata a cada clase como igualmente importante, puede no ser adecuada para analizar conjuntos de datos desequilibrados, donde la clase rara se considera más interesante que la clase mayoritaria. Para la clasificación binaria, la clase rara a menudo se denota como la clase positiva, mientras que la clase mayoritaria se denota como la clase negativa (p.296).

La evaluación del desempeño de un modelo se basa en la sumatoria de registros con resultados positivos o negativos, dicha sumatoria es colocada en una tabla, la cual se le conoce como matriz de confusión (figura 4); dicha matriz representa la clasificación de un problema binario.

Figura 4

Modelo de matriz de confusión



De la Figura 4 Se tiene los siguientes significados

- TP; es la suma de datos predichos correctamente para la clase 1.
- FP; es la suma de datos predichos incorrectamente para la clase 1.
- TN; es la suma de datos predichos correctamente para la clase 0.
- FN; es la suma de datos predichos incorrectamente para la clase 0.

A. Precisión

Se basa en el recuento de los datos predichos correctamente en una determinada clase (0 o 1) divididos sobre el total de datos pertenecientes a dicha clase; de esta forma hallamos la tasa probabilística que posee al determinar exitosamente una predicción.

$$\text{Precisión Clase } 1 = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Precisión Clase } 0 = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

B. Accuracy

Es el indicador que muestra de forma general el rendimiento de un modelo para ser contrastado con diferentes modelos; dado que expresa el promedio de los datos predichos correctamente para ambas clases.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

C. Sensibilidad

La sensibilidad de un modelo se define como la tasa probabilística de datos predichos correctamente en el modelo basados en la clase 1.

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

D. Especificidad

Por el contrario, a la sensibilidad, la especificidad ve la tasa probabilista de datos predichos correctamente para la clase 0.

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

2.2.6 Metodología

A. SEMMA (sample, explore, modify, model, assess);

SAS Enterprise Miner (2006) describe:

Se refiere al proceso central de realizar minería de datos. Comenzando con una muestra estadísticamente representativa de sus datos, SEMMA facilita la aplicación de técnicas exploratorias estadísticas y de visualización, seleccionar y transformar las variables predictivas más significativas, modelar las variables para predecir resultados y confirmar la precisión de un modelo. (p.1).

A continuación, se describirá los 5 modelos de desarrollo de SEMMA:

- **Sample (**Muestreo**);** Es la primera etapa de la metodología SEMMA la cual consiste en la extracción de los datos que generan valor al análisis, con ello obtendremos una reducción de tiempos al momento de realizar las cargas.
- **Explore (**Exploración**);** en esta etapa se aplican diferentes técnicas de análisis en los datos con la finalidad de detectar anomalías o deficiencias que perjudiquen el minado de datos; para ello se debe proceder con el tratamiento de los mismos haciendo uso de algoritmos (o técnicas de estadística) que nos permitan descubrir estas anomalías.
- **Modify (**Modificación**);** En la etapa de modificación se realizará la transformación de los datos de acuerdo a los tipos de variables (números, textos, factores) basados en el análisis de la exploración, de ser oportuno, se realizarán agrupaciones, introducción de nuevos campos (variables) o inclusive la reducción (o eliminación) de variables (o datos) que generan ruido o no guardan pertinencia en el minado.
- **Model (**Modelado**);** En la presente etapa usaremos las diferentes herramientas (software), técnicas (métodos estadísticos), modelos de minería (algoritmos) que nos permitan realizar la búsqueda de patrones (asociaciones o combinaciones), para ello debemos tener en cuenta que

cada tipo de modelo, técnica o herramienta tiene un contexto en la cual llegan a proporcionar mejores resultados (por ejemplo, las redes neuronales nos facilitan el ajuste de relaciones no lineales complejas) para llegar a obtener predicciones confiables.

- **Assess (**Evaluación**);** Luego de haber construido el modelo se realiza la evaluación con la finalidad de determinar la validez, y la confiabilidad; para ello se debe usar el grupo de datos “Test” (creado en la primera fase). De obtener demasiados resultados deficientes, se debe proceder a modificar el modelo realizado o en su defecto, a desarrollar otro modelo.

2.3 Estado del arte

A. Redes neuronales

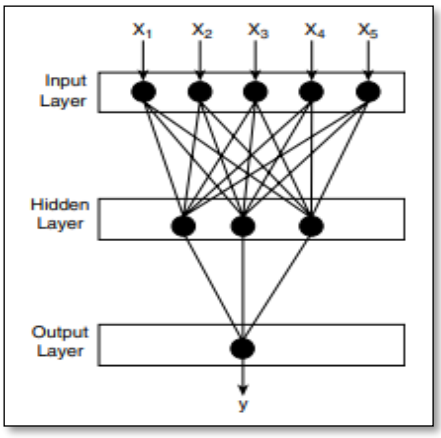
Las redes neuronales son modelos matemáticos que asemejan el comportamiento biológico de las neuronas y la sinapsis de las mismas dentro del cerebro humano. Estas redes neuronales están compuestas por la capa de entrada (input), capa salida (output), capas ocultas y las funciones de activación. (Morera, 2018).

Las capas ocultas según Matich (2001) sostiene que:

Son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de redes neuronales. (p.16).

Figura 5

Modelo de red neuronal



- **Etapa de aprendizaje**

  Según Montaño (2002) afirma:

  El objetivo que se persigue es hacer mínima la discrepancia o error entre la salida obtenida por la red y la salida deseada por el usuario ante la presentación de un conjunto de patrones denominados grupo de entrenamiento. Por este motivo, se dice que el aprendizaje en las redes backpropagation es de tipo supervisado, debido al usuario (o supervisor) determina la salida deseada ante la presentación de un determinado patrón de entrada. (p. 246).

- **Deep Learning**

Según LeCun y Hinton (2015) nos presenta una apreciación sobre la importancia y cualidades del Deep learning (aprendizaje profundo):

El aprendizaje profundo permite modelos computacionales que se componen de múltiples capas de procesamiento para aprender representaciones de datos con múltiples niveles de abstracción. Estos métodos han mejorado drásticamente el estado del arte en reconocimiento del habla, reconocimiento de objetos visuales, detección de objetos y muchos otros dominios, como el

descubrimiento de fármacos y la genómica. El aprendizaje profundo descubre una estructura compleja en grandes conjuntos de datos mediante el uso del algoritmo de retropropagación para indicar cómo una máquina debe cambiar sus parámetros internos que se utilizan para calcular la representación en cada capa a partir de la representación en la capa anterior. Las redes convolucionales profundas han producido avances en el procesamiento de imágenes, video, voz y audio, mientras que las redes recurrentes han iluminado datos secuenciales como el texto y la voz. (p. 436).

CAPÍTULO III

DESARROLLO DE LA SOLUCIÓN

3.1 Factibilidad del proyecto

3.1.1 Factibilidad técnica

Dadas las circunstancias de la investigación es debido tomar en cuenta las características necesarias de los recursos a utilizar con la finalidad de obtener un óptimo rendimiento de los procesos; a continuación, se describen los aspectos técnicos en la tabla 8.

Tabla 8

Descripción de los elementos

RECURSOS	DESCRIPCIÓN	UNIDADES
Hardware	Laptop	• I7-9750H • Disco: 256GB SSD • RAM: 8GB • Procesador: 2,6 GHz • Tarjeta gráfica: GTX 1660 Ti (4GB)
Software	Sistema operativo Microsoft office	• Windows 8.1 Professional 64bits • 2019 plus
Otros	Conexión a internet	• 2Mbps

3.1.2 Factibilidad económica

La investigación presenta una inversión centrada en costo/beneficio, en la cual se tiene las siguientes apreciaciones:

- Tratamiento de información confiable, dado la privacidad (baja exposición de datos).
- Facilidad de manejo de grandes volúmenes de información, debido la capacidad de procesamiento que se posee, lo que permite optimizar tiempo, y aumenta el nivel de pruebas.

A continuación, se presenta la relación de costos de desarrollo en la tabla 9.

Tabla 9

Costos de desarrollo de la solución

RECURSOS GENERALES	DESCRIPCIÓN	UNIDADES	COSTO (S/.)	TOTAL (S/.)
		Recursos humanos		
Ascue Silva, Saúl Sebastian	Persona	1 Und.	S/.2,000.00	S/.2,000.00
Olascoaga Roman, Luis Alonso	Persona	1 Und.	S/.2,000.00	S/.2,000.00
		Recursos hardware		
Laptop	-	2 Und.	S/.3,500.00	S/.7,000.00
		Recursos software		
Windows 8.1	-	2 Und.	S/.149.00	S/.298.00
Microsoft office	-	2 Und.	S/.99.00	S/.198.00
R Studio	V3	2 Und.	-	-
R	V3.6.1	2 Und	-	-
		Otros gastos		
Materiales de oficina	-	-	S/.50.00	S/.50.00
Hojas	-	1/2 millar	S/.18.00	S/.18.00
Total Presupuesto				**S/.11,564.00**

3.2 Metodología SEMMA

Para el desarrollo de este algoritmo se hace uso del entorno integrado RStudio que hace uso del lenguaje R; También se hace uso de las siguientes librerías presentadas en la tabla 10.

Tabla 10

Concepto de las librerías

Librería	Descripción
ggplot2	Crea visualizaciones de datos haciendo uso de la gramática de gráficos.
corrplot	Realiza Visualizaciones gráficas a partir de una matriz de correlación o matriz general.
e1071	Compendio de funciones orientada a la estadística y probabilidad.
gmodels	Diversidad de funciones orientada al ajuste de modelos.
caret	Herramientas para la clasificación y entrenamiento; trazados y modelos de regresión.
DMwR	Esta librería presenta una gran variedad de funciones; en la presente se usará una de sus funciones basada en el método SMOTE, que permite resolver problemas de desequilibrio de clases.
ROCR	Permite la visualización de rendimiento, como las curvas de sensibilidad, especificidad, gráficos de elevación y precisión.
dplyr	Facilita un rápido y consistente trabajo con marco de datos, objetos.

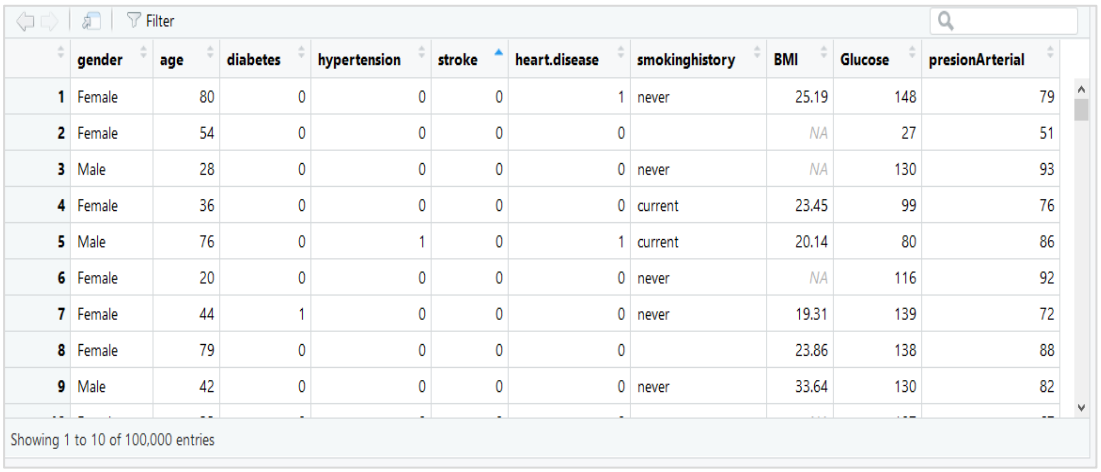
A continuación, iniciamos con el desarrollo de la metodología.

A. **Sample (**Muestreo**)**

A continuación, se muestra el conjunto de datos origen en la figura 6.

Figura 6

Tabla de variables



Filter

	gender	age	diabetes	hypertension	stroke	heart.disease	smokinghistory	BMI	Glucose	presionArterial
1	Female	80	0	0	0	1	never	25.19	148	79
2	Female	54	0	0	0	0		NA	27	51
3	Male	28	0	0	0	0	never	NA	130	93
4	Female	36	0	0	0	0	current	23.45	99	76
5	Male	76	0	1	0	1	current	20.14	80	86
6	Female	20	0	0	0	0	never	NA	116	92
7	Female	44	1	0	0	0	never	19.31	139	72
8	Female	79	0	0	0	0		23.86	138	88
9	Male	42	0	0	0	0	never	33.64	130	82

Showing 1 to 10 of 100,000 entries

De los cuales será descartado los pacientes que no tengan diabetes para el desarrollo del modelo de red neuronal. De igual manera para la división de datos se prepara un pequeño script en donde se tomará a modelo de entrenamiento (train) el 70% de los datos, y la diferencia serán los datos prueba (test); cabe destacar que de la data inicial es de 100 mil registros, lo cual no implicara que los datos de entrenamiento y prueba serán iguales al total inicial, debido a que sufrirán un tratamiento.

B. **Explore (**Exploración**)**

En la presente fase comenzaremos haciendo una exploración de los datos que tenemos, esto será gracias al comando summary y str, el cual nos retornará un resumen de las variables (dependiendo el tipo puede mostrar más o menos detalles), este será la vista general dentro de la indagación a los datos.

En la figura 7 podemos apreciar los tipos de datos de cada variable, así como también un pequeño recuento de los datos que se encuentran en ellos, a su vez se puede determinar que tanto smokinghistory (dentro de su resumen de datos se ve (“”) y BMI (se aprecian NA) se cuentan con datos nulos y blancos.

Figura 7

Resumen de los datos simplificado

```
> str(datos)
'data.frame':   8500 obs. of  9 variables:
 $ gender         : Factor w/ 3 levels "Female","Male",..: 1 2 2 2 1 2 1 2 1 1 ...
 $ age            : num  44 67 50 73 53 50 67 57 36 60 ...
 $ hypertension   : int  0 0 1 0 0 0 0 0 0 0 ...
 $ stroke         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ heart.disease  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ smokinghistory : Factor w/ 6 levels "","current","ever",..: 5 6 2 4 4 4 5 1 2 5 ...
 $ BMI            : num  19.3 NA NA 25.9 NA ...
 $ Glucose        : int  200 86 93 180 159 172 145 119 80 98 ...
 $ presionArterial: int  118 84 165 186 75 111 159 176 85 114 ...
```

A continuación, en la figura 8 vemos a más detalle los datos (con la excepción que no se cuenta con el tipo de datos), en donde se aprecian el mínimo, máximo, primer cuartil, segundo cuartil, media y mediana de los datos; en el caso del gender, smokinghistory y stroke por ser factores solo se aprecia un recuento de los datos por cada tipo. En la figura anterior observamos que en el BMI contenía valores nulos y el smokinghistory contenía valores blancos; sin embargo, solo vemos que se contabiliza los valores nulos para el BMI; además se puede apreciar que el valor mínimo del BMI es de 10.98 por lo que se considera como dato atípico. Otro factor que vemos se ubica en el stroke, como se puede apreciar la cantidad de datos para pacientes que ha sufrido de ACV es solo de 406 (4% de la data).

Figura 8

Resumen de los datos

```
> summary(datos)
    gender          age         hypertension    stroke   heart.disease        smokinghistory
 Female:4461   Min.   : 3.00   Min.   :0.0000   0:8094   Min.   :0.0000              :1454
 Male  :4039   1st Qu.:52.00   1st Qu.:0.0000   1: 406   1st Qu.:0.0000   current    : 948
 Other :   0   Median :62.00   Median :0.0000            Median :0.0000   ever       : 472
               Mean   :60.95   Mean   :0.2456            Mean   :0.1491   former     :1590
               3rd Qu.:72.00   3rd Qu.:0.0000            3rd Qu.:0.0000   never      :3346
               Max.   :80.00   Max.   :1.0000            Max.   :1.0000   not current: 690

      BMI           Glucose      presionArterial
 Min.   :10.98   Min.   : 80.0   Min.   : 60.0
 1st Qu.:27.50   1st Qu.:109.0   1st Qu.: 94.0
 Median :31.85   Median :139.0   Median :128.0
 Mean   :33.01   Mean   :139.4   Mean   :130.1
 3rd Qu.:37.36   3rd Qu.:170.0   3rd Qu.:165.0
 Max.   :88.72   Max.   :200.0   Max.   :200.0
 NA's   :1527
```

En la Figura 9 se puede apreciar 2 sentencias; una determina los valores nulos, en donde solo el BMI no da un resultado de 1527 valores; y la 2da sentencia nos proporciona los datos blancos en donde solo smokinghistory tiene un total de 1454 datos. Para las demás variables no se observa ningún valor atípico.

Figura 9

Resumen de los datos nulos y blancos

```
> apply(datos, 2,function(x) sum(is.na(x)))
        gender            age   hypertension         stroke  heart.disease smokinghistory
             0              0              0              0              0              0
           BMI        Glucose presionArterial
          1527              0              0
> apply(datos, 2,function(x) sum((x)==""))
        gender            age   hypertension         stroke  heart.disease smokinghistory
             0              0              0              0              0           1454
           BMI        Glucose presionArterial
            NA              0              0
```

Como conclusión de esta fase tenemos:

- Ejecutar una limpieza de datos nulos y blancos.
- Balancear los datos, dado que hay un gran desequilibrio en la variable stroke.
- Ajustar los valores del BMI.

C. **Modify (**Modificación**)**

A continuación, se modificarán los valores de las variables y los tipos de variables, esto se debe a que las redes neuronales trabajan con números, y no con cadenas o caracteres; También, se realizará un balance de los datos, véase en la figura 10.

Se iniciará haciendo el cambio de los valores en las siguientes variables o columnas:

gender; para esta variable, sus datos serán:

Tabla 11

Leyenda del sexo

Leyenda:	Gender
1 =	Male
2 =	Female
3 =	Other

Smokinghistory; en el caso de esta variable, sus datos de origen serán cambiados de la siguiente manera:

Tabla 12

Leyenda de la hipertensión

Leyenda:	hypertension
1 =	Never
2 =	Former
3 =	Not Current
4 =	Current
5 =	Ever

Figura 10

Modificación de contenido de variables

```
> #genero
> ACV$gender[ACV$gender == "Male"] = 1
> ACV$gender[ACV$gender == "Female"] = 2
> ACV$gender[ACV$gender =="Other"] = 0
>
> #Historial Fumador
> ACV$smokinghistory[ACV$smokinghistory == "current"] = 1
> ACV$smokinghistory[ACV$smokinghistory == "not current"] = 2
> ACV$smokinghistory[ACV$smokinghistory == "former"] = 3
> ACV$smokinghistory[ACV$smokinghistory == "ever"] = 4
> ACV$smokinghistory[ACV$smokinghistory == "never"] = 5
>
> distinct(ACV,smokinghistory, gender)
   smokinghistory gender
1               5      2
2               2      1
3               1      1
4               3      1
5               3      2
6                      1
7               1      2
8               5      1
9               2      2
10              4      1
11                     2
12              4      2
```

Adicionalmente se cambiará el tipo de variable de las siguientes variables, dado que ya se han hecho las modificaciones para el caso de gender y smokinghistory; en el caso de age su tipo de variable inicial es numeric como se puede apreciar en la Figura 11.

- **gender** = factor -> entero
- **smokinghistory** = factor -> entero
- **age** = numérico -> entero

Figura 11

Cambio de tipo de variable

```
> ####CAMBIO DE TIPOS DE VARIABLES
> ACV$gender =  as.integer(ACV$gender)
> ACV$smokinghistory =  as.integer(ACV$smokinghistory)
> ACV$age =  as.integer(ACV$age)
> str(ACV)
'data.frame':   8500 obs. of  9 variables:
 $ gender         : int  2 1 1 1 2 1 2 1 2 2 ...
 $ age            : int  44 67 50 73 53 50 67 57 36 60 ...
 $ hypertension   : int  0 0 1 0 0 0 0 0 0 0 ...
 $ stroke         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ heart.disease  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ smokinghistory : int  5 2 1 3 3 3 5 NA 1 5 ...
 $ BMI            : num  19.3 NA NA 25.9 NA ...
 $ Glucose        : int  200 86 93 180 159 172 145 119 80 98 ...
 $ presionArterial: int  118 84 165 186 75 111 159 176 85 114 ...
```

Luego, se deberá eliminar los valores atípicos de la data (NA's y blancos). En la figura 12 se realizan los respectivos filtros, comenzando con la eliminación de datos nulos en la variable BMI y luego se procede a delimitar el valor mínimo dentro del mismo (mayor a 20.00); para el caso del gender solo se toma los valores 1 y 2, descartando cualquier dato atípico (posible inclusión de dato other representado por 0); para el caso de los datos dentro de la variable smokinghistory se eliminan los valores blancos y por último, se definen valores mayores a 19 dentro de la variable age.

Figura 12

Limpieza de los datos

```
> #####LIMPIEZA :
> ACV_EDIT = ACV[!is.na(ACV$BMI),]
> ACV_EDIT = ACV_EDIT[ACV_EDIT$BMI >20.00,]
> ACV_EDIT = ACV_EDIT[ACV_EDIT$gender >0,]
> ACV_EDIT = ACV_EDIT[!is.na(ACV_EDIT$smokinghistory),]
> ACV_EDIT = ACV_EDIT[ACV_EDIT$age >19.00,]
>
> apply(ACV_EDIT, 2,function(x) sum(is.na(x)))
        gender            age   hypertension         stroke  heart.disease smokinghistory
             0              0              0              0              0              0
           BMI        Glucose presionArterial
             0              0              0
> apply(ACV_EDIT, 2,function(x) sum((x)==""))
        gender            age   hypertension         stroke  heart.disease smokinghistory
             0              0              0              0              0              0
           BMI        Glucose presionArterial
             0              0              0
> min(ACV_EDIT$BMI)
[1] 20.04
> min(ACV_EDIT$age)
[1] 20
```

El siguiente paso será balancear la data (figura 13), para ello usaremos la librería Smote que mediante un algoritmo incrementa los datos según los parámetros que se le den, para ello asignaremos la variables mediante una formula (la cual describe el factor "stroke", y las demás variables), se incluye la variable que almacena la data, "perc.over" va orientado al sobre muestreo que se incluirá la data (definido como la decisión de cuantos datos se generaran de la clase minoritaria), "k" es la cantidad de vecinos cercanos por el cual el algoritmo procesara la información para generar nueva data y por ultimo "perc.under" es la relación entre la clase mayoritaria

a generar conjuntamente con “perc.over”; antes de ejecutar este punto se tuvo un total de 5913 datos, luego de la ejecución se cuenta con 23076 datos de los cuales los resultados de pacientes con ACV llegan a conformar el 37.8% de la data. A su vez el balanceo de la data hace un cambio a los tipos de variables, convirtiendo a todos en numéricos a excepción del “stroke” que se conserva como factor.

Figura 13

Balanceo de datos

```
> #Balanceo de datos
> b = names(ACV_EDIT)
> Form.bal = as.formula(paste("stroke ~", paste(b[!b %in% c("stroke")], collapse ="+")))
> case.add = round(max(dim(ACV_EDIT))*0.50)
> ACV.bal = SMOTE(Form.bal,
+                 ACV_EDIT,
+                 perc.over = case.add,
+                 k = 35,
+                 perc.under = 170)
>
> summary(ACV.bal)
     gender           age         hypertension    stroke    heart.disease    smokinghistory
 Min.   :1.000   Min.   :20.00   Min.   :0.0000   0:14346   Min.   :0.0000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:56.06   1st Qu.:0.0000   1: 8730   1st Qu.:0.0000   1st Qu.:3.000
 Median :2.000   Median :66.00   Median :0.0000             Median :0.0000   Median :4.000
 Mean   :1.538   Mean   :64.11   Mean   :0.2922             Mean   :0.2002   Mean   :3.665
 3rd Qu.:2.000   3rd Qu.:74.39   3rd Qu.:1.0000             3rd Qu.:0.0000   3rd Qu.:5.000
 Max.   :2.000   Max.   :80.00   Max.   :1.0000             Max.   :1.0000   Max.   :5.000
      BMI            Glucose      presionArterial
 Min.   :20.04   Min.   : 80.0   Min.   : 60
 1st Qu.:27.77   1st Qu.:111.0   1st Qu.: 98
 Median :31.48   Median :138.0   Median :130
 Mean   :32.68   Mean   :138.8   Mean   :131
 3rd Qu.:36.47   3rd Qu.:166.0   3rd Qu.:163
 Max.   :88.72   Max.   :200.0   Max.   :200
> str(ACV.bal)
'data.frame':   23076 obs. of  9 variables:
 $ gender         : num  1 2 2 1 2 1 2 2 2 1 ...
 $ age            : num  65 46 53 74 80 73 48 29 39 67 ...
 $ hypertension   : num  0 0 1 0 0 0 0 1 0 0 ...
 $ stroke         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ heart.disease  : num  0 0 0 1 1 1 0 0 0 0 ...
 $ smokinghistory : num  5 2 5 2 4 5 3 5 5 5 ...
 $ BMI            : num  39.6 48.1 27.8 28.7 25.3 ...
 $ Glucose        : num  88 158 129 87 156 102 97 82 86 195 ...
 $ presionArterial: num  168 189 85 167 108 104 166 87 84 141 ...
>
```

D. **Model (**Modelado**)**

En esta fase iniciaremos viendo el comportamiento de la data respecto al stroke por cada indicador, para ello usaremos gráficos de cajas, gráfico de barras e histogramas.

Figura 14

Gráficos de Cajas 1

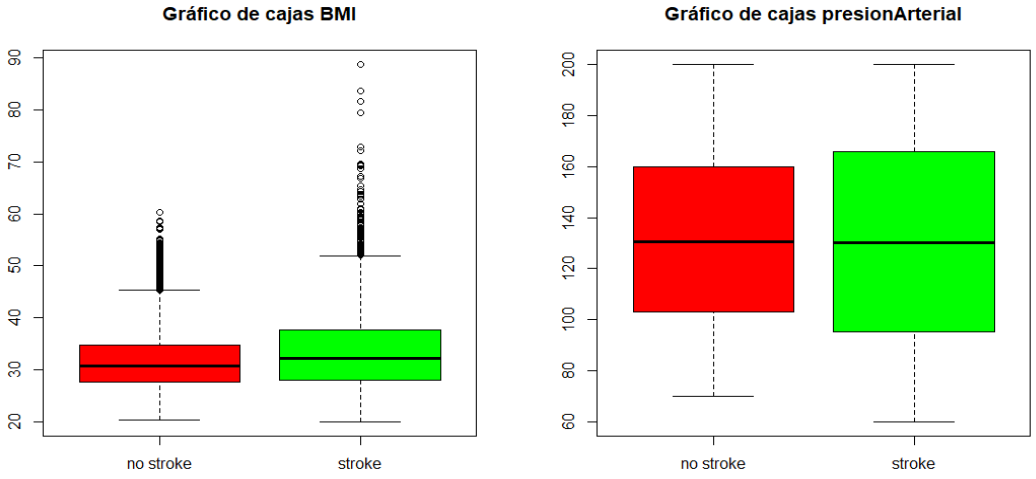
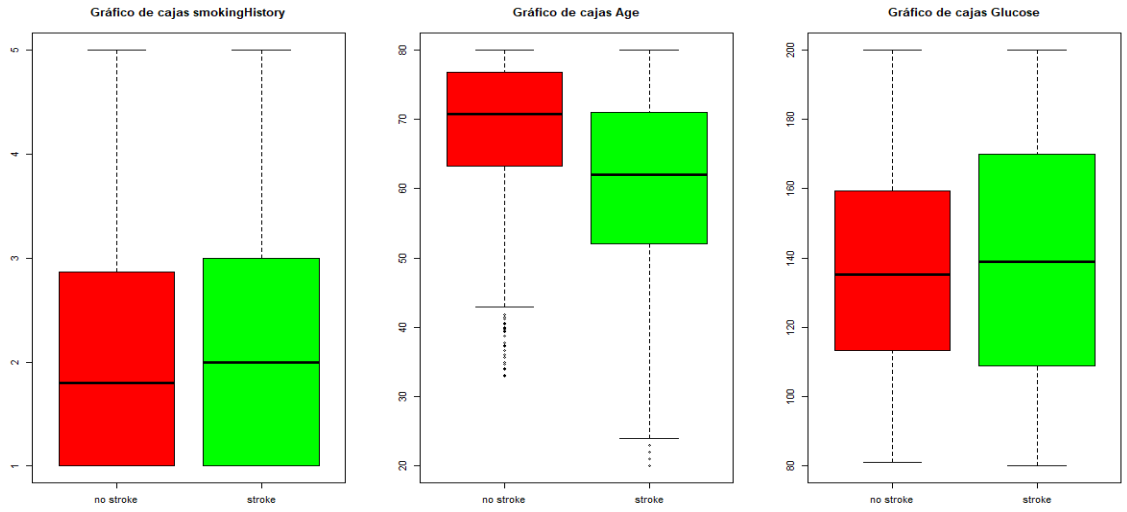


Figura 15

Gráficos de cajas 2



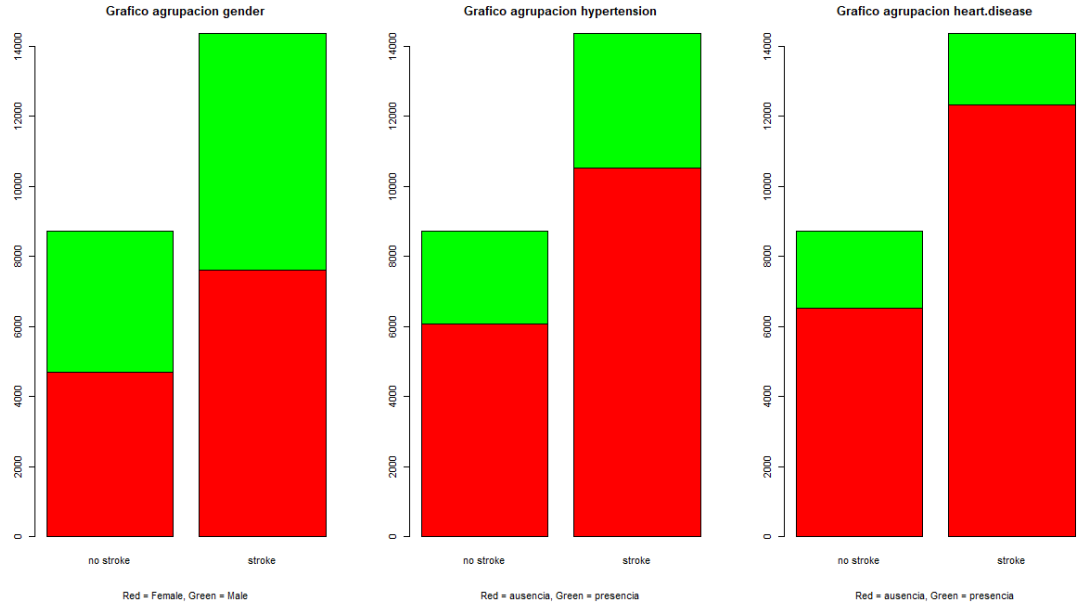
En la figura 14 y figura 15 se aprecian los diferentes indicadores en relación al stroke (nos interesa saber sobre las cajas verdes que representan un stroke), en donde se tiene la tendencia del stroke respecto a los mismos; sin embargo, también se aprecian en algunos de los gráficos una continuidad de circunferencias (atípicos)

en los diferentes límites de las cajas, estos nos muestran los datos atípicos que se presentan en el indicador para la correlación con el stroke.

- Gráfico de cajas BMI (Figura 14), la tendencia de un stroke para este indicador tiene un rango de 29 a 35, siendo la tendencia central el valor 30; sin embargo, se cuenta con una cantidad de datos atípicos en el rango de 50 a 90 regresivamente.
- Gráfico de cajas presiónArterial (Figura 14), la densidad de los datos se encuentra entre los valores 95 y 165, para este caso no se observan valores atípicos.
- Gráfico de cajas smokingHistory (Figura 15), la tendencia de los datos yace a la derecha abordando los valores 1; 2 y 3 del indicador.
- Gráfico de cajas Age (Figura 15),al igual que en el gráfico de cajas smokingHistory se tiene mayor densidad a la izquierda entre los datos 56 hasta 76; con unos ligeros datos atípicos a la derecha.
- Gráfico de cajas Glucosa (Figura 15), la densidad de los datos se encuentra entre los valores 118 hasta 172, sin la presencia de datos atípicos.

Figura 16

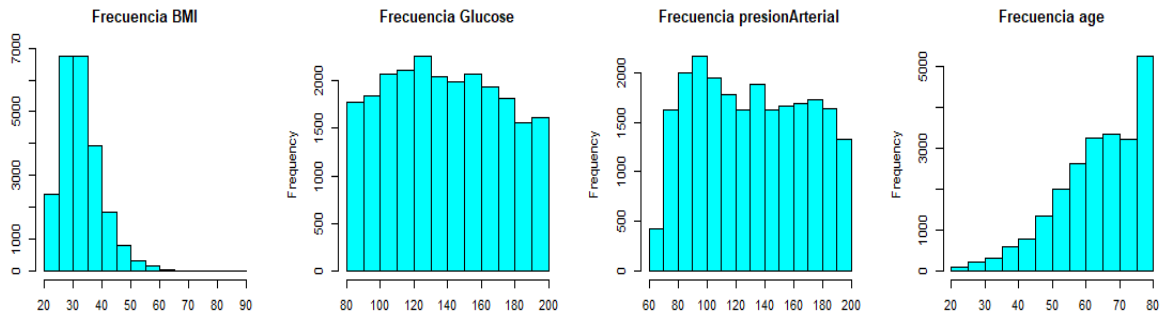
Cantidad de datos de stroke por categoría



A diferencia de los anteriores indicadores en la Figura 16 se aprecian gráficos de barras que indican la cantidad de datos referentes al stroke por cada indicador (cada uno posee su leyenda); esto se debe a que estos indicadores tienen una función más orientada a ser categorías (los datos representan 1 y 0) por ende, es oportuno saber la cantidad asociada a cada uno de ellos.

Figura 17

Ejemplo de histogramas



Adicionalmente se muestran histogramas en la figura 17, que guardan la misma relación con los gráficos de cajas (figura 14 y 15) con la finalidad de ver el

comportamiento de los datos. Luego se aplicará una normalización a los datos, para centralizar la data (el código se muestra en la figura 18).

Figura 18

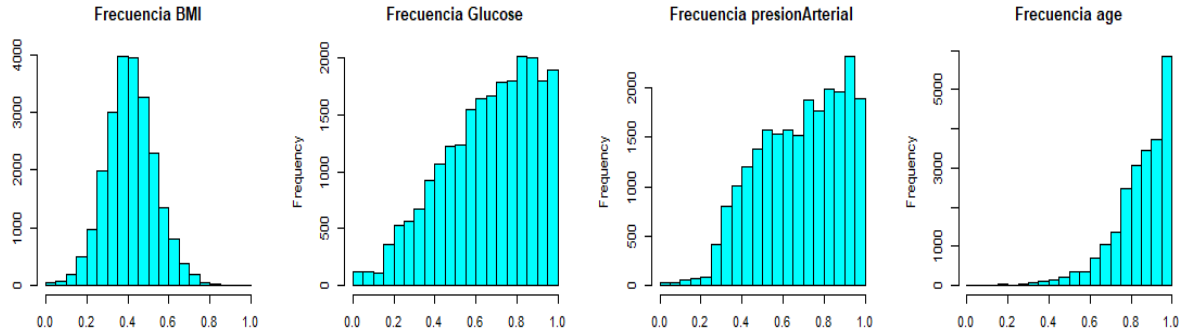
Normalización de los datos

```
> ###Normalización
> normaliza <- function(x) {return (sqrt((x-min(x))/(max(x)-min(x))))}
> ACV_norm <- as.data.frame(lapply(ACV.rel, normaliza))
> str(ACV_norm)
'data.frame':   23076 obs. of  9 variables:
 $ gender         : num  1 1 0 1 0 1 0 0 0 0 ...
 $ age            : num  0.516 1 0.904 0.983 0.94 ...
 $ hypertension   : num  1 0 0 1 0 0 1 1 1 1 ...
 $ stroke         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ heart.disease  : num  0 0 0 0 0 0 0 0 1 0 ...
 $ smokinghistory : num  1 0.5 1 0.866 0.5 ...
 $ BMI            : num  0.645 0.306 0.463 0.433 0.258 ...
 $ Glucose        : num  0.456 0.474 0.5 0.904 0.619 ...
 $ presionArterial: num  0.12 0.493 0.561 0.89 0.712 ...
```

Esta normalización se puede apreciar de mejor manera en los gráficos de cajas o histogramas; a continuación, se muestra un ejemplo de cómo queda la normalización de la data (figura 19).

Figura 19

Histogramas de la normalización de los datos



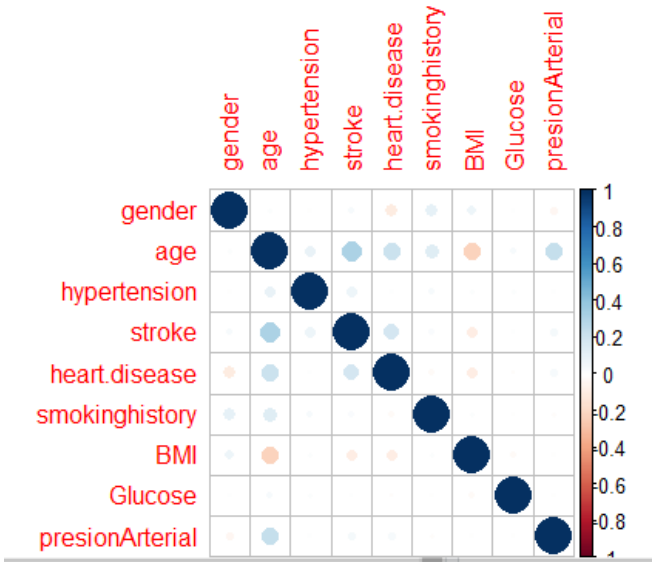
A continuación, se genera un gráfico de correlación para determinar la pertinencia de los indicadores con respecto al stroke.

En la figura 20 se observa las diferentes relaciones de los indicadores, para el caso del stroke, se tiene una fuerte relación con el indicador age, hypertension, heart.disease; una baja relación con gender, smokinghistory y la presionArterial; sin embargo, en el caso del BMI y glucose se ve una relación opuesta (tiende a ser

negativo). También debemos considerar las relaciones que se generan con los mismos indicadores, debido a que estos forman patrones.

Figura 20

Gráfico de correlación



Para concluir esta preparación de los datos (modelado) y comenzar con la creación de la red neuronal, se debe crear los grupos de entrenamiento (train) y prueba (test); En la Figura 21 se aprecia que para el grupo de entrenamiento (train) se tiene 16153 datos y para la prueba (test) un total de 6923 datos.

Figura 21

División de datos train/test

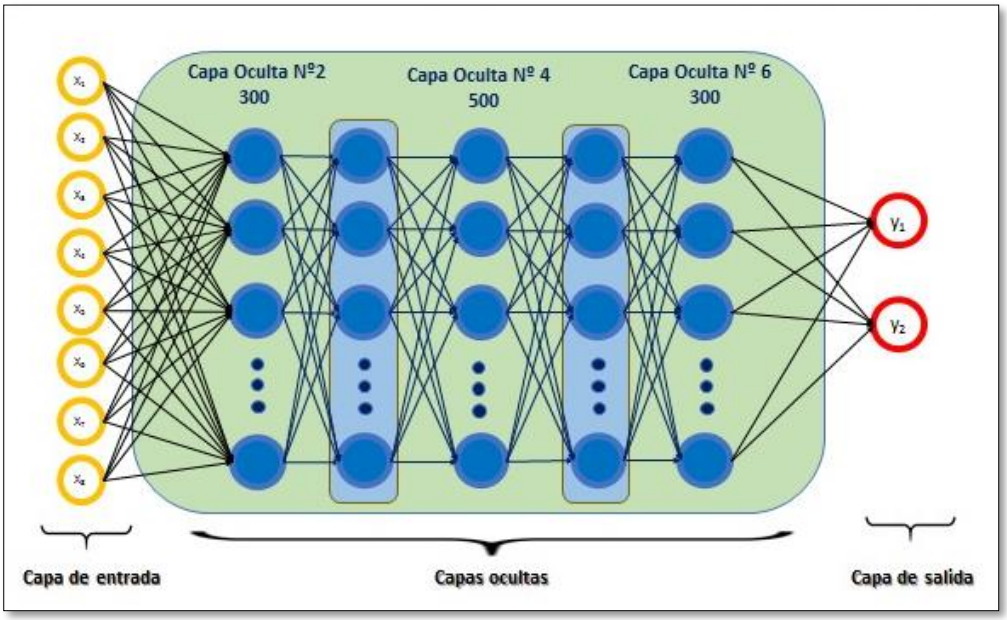
```
> #division de datos
> index = sample(1: nrow(ACV_norm), round(0.70 * nrow(ACV_norm)))
> train = ACV_norm[index,]
> test  = ACV_norm[-index,]
>
> dim(train)
[1] 16153     9
> dim(test)
[1] 6923    9
```

En la última fase de modelado de la red neuronal se aplica el Deep learning, por lo cual se usa el api de “H2O”; este api nos permite usar sus respectivas librerías

para el Deep learning (gran variedad de modelos). Este modelo usara la función de activación de RELU (unidad rectificada lineal) el cual favorece las clasificaciones, también se toma en cuenta el uso de 6 capas las cuales contienen 300; 300; 500; 500; 300; 300 neuronas respectivamente; por último, se estima un uso de 100 épocas (iteraciones) dentro de este algoritmo.

Figura 22

Modelo de red neuronal



E. **Assess (**Evaluación**)**

La figura 23 nos muestra el proceso de corrección de la predicción que surge en la red neuronal, desde el inicio se ve que hasta la época (interacción) 20 hay un gran ajuste; sin embargo, pasando las épocas 60 el ajuste del modelo es mínimo.

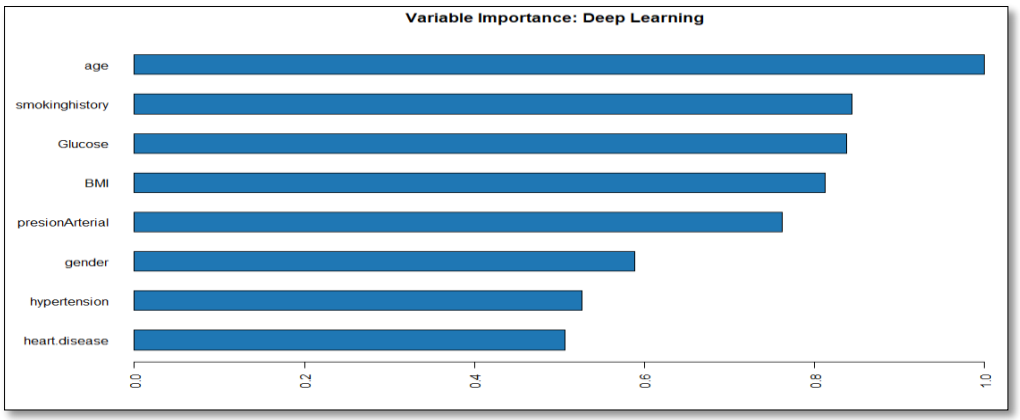
Figura 23

Historia de la puntuación de entrenamiento

También podemos apreciar en la Figura 24 la influencia o pertinencia dentro de la red neuronal de cada indicador o variable respecto a la importancia de predecir un Stroke.

Figura 24

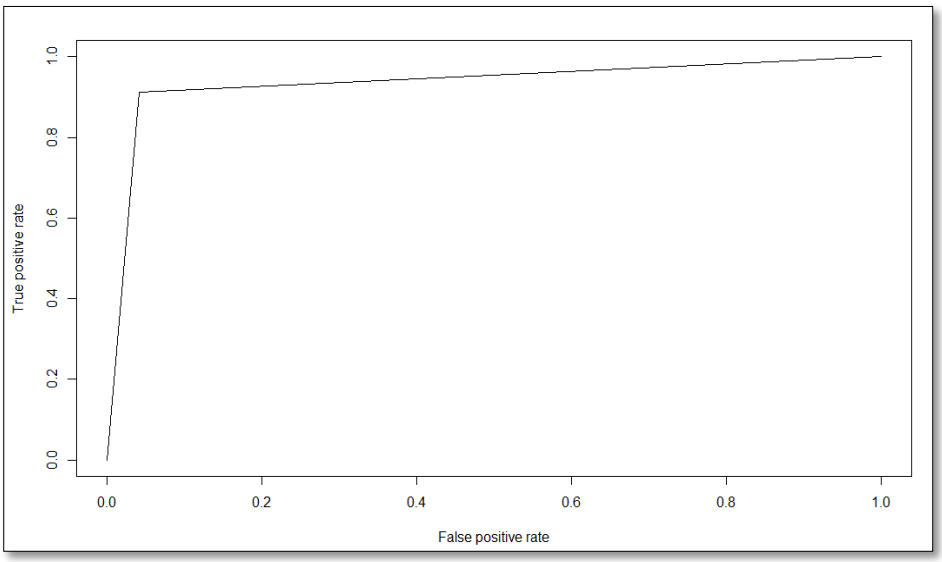
Importancia de variables según la predicción



La figura 25 representa la curva ROC, el gráfico muestra la sensibilidad en función a los falsos positivos, en la cual la curva debe tender a 1; para ello la línea debe formar una curva que supere el 0.6 (prueba útil) de lo contrario sería un pésimo modelo (prueba inútil). En la presente gráfica nos muestra que el modelo supera el mínimo (0.6), siendo el valor actual 93.51.

Figura 25

Curva ROC



CAPÍTULO IV

ANÁLISIS DE RESULTADOS Y CONTRASTACIÓN DE LA HIPÓTESIS

4.1 Población y muestra

4.1.1 Población

Se tomará en cuenta los datos del repositorio kaggle, los cuales comprenden un registro de ocho mil quinientos datos.

N = 17,372

4.1.2 Muestra

Para la presente investigación se tomará un nivel de confianza del 99% y un margen del 1%; entonces, se aplica la siguiente fórmula:

$$n = \frac{\mathrm{z} \wedge 2(\mathrm{p} * \mathrm{q})}{\mathrm{e} \wedge 2 + (\frac{\mathrm{z} \wedge 2(\mathrm{p} * \mathrm{q})}{\mathrm{N}})}$$

N = Tamaño de muestra
z = Nivel de confianza deseado
p = Proporción de la población con la característica deseada (éxito)
q = Proporción de la población sin la característica deseada (fracaso)
e = Nivel de error dispuesto a cometer
N =Tamaño de la población

Con un margen del 1%, un nivel de confianza del 99% y una población de 17,372 (ocho mil quinientos); nos da un tamaño de:

n = 8,500

4.1.3 Tipo de muestreo

Muestreo deliberado, crítico o por juicio; Es aquel tipo de muestreo que hace la selección de la población o propósito de estudio en base al conocimiento que tiene del mismo para lograr el estudio del conjunto de datos y su posterior uso.

4.2 Validez y confiabilidad de instrumento

4.2.1 Validez

La validación del instrumento se ha desarrollado gracias de tres expertos en sus respectivas áreas profesionales quienes han revisado la pertinencia, claridad y relevancia recomendando su aplicabilidad; apreciamos los datos de los involucrados en la tabla 13.

Tabla 13

Validez del instrumento por juicio de expertos

	Experto 1	**Experto 2**	**Experto 3**
Nombre	Jan Hermoza Dueñas	José Luis Herrera Salazar	Pedro Martin Lezama Gonzales
Especialidad	Medicina General	Dr. en Sistemas	Ing. En sistemas

4.2.2 Confiabilidad del instrumento

La confiabilidad del instrumento ha sido determinada a través del método Alfa de Cronbach que se aplicó en la encuesta, la cual consistió en determinar el mínimo margen de precisión para ser aceptable en la dimensión de las variables de predicción; donde se tiene presente los indicadores de precisión, accuracy, sensibilidad y especificidad.

A continuación, mediante el instrumento Alfa de Cronbach se determinó una homogeneidad de los datos que es igual al 94%, lo que resalta la confiabilidad de los datos.

Figura 26

Varianza de los indicadores

	KPI1	KPI2	KPI3	KPI4	Total
	5	5	5	5	20
	4	4	4	4	16
	4	4	4	4	16
	4	4	4	4	16
	5	4	5	5	19
	5	5	4	5	19
	4	4	4	4	16
	4	4	4	4	16
	4	4	4	4	16
	4	4	4	4	16
	5	5	5	5	20
VARI	0.254545455	0.21818182	0.21818182	0.25454545	3.21818182

Figura 27

Número de elementos del mínimo aceptable

Minimo aceptable		%
1.- 75%	0	0.0%
2.- 80%	0	0.0%
3.- 85%	0	0.0%
4.- 88%	30	68.2%
5.- 90%	14	31.8%
total	44	100%

Figura 28

Validez del margen mínimo de predicción para los indicadores

K	4
SUMVi	0.94545455
Vt	3.21818182
alfa	0.94161959

4.3 Análisis e interpretación de resultados

4.3.1 Resultados

Para la interpretación de resultados se utiliza la matriz de confusión, para determinar la cantidad de datos predichos correcta e incorrectamente.

Tabla 14

Matriz de confusión

	Stroke	No Stroke
Stroke	2445	173
No Stroke	239	4066

A continuación, se muestran los cálculos y resultados que nos genera la matriz de confusión.

Figura 29

Resultados de la matriz de confusión

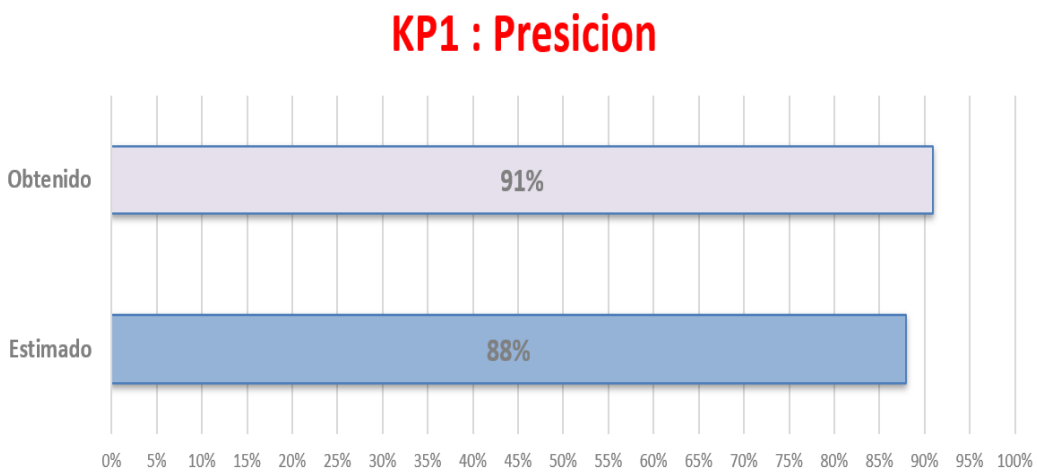
```
> PRECISION = TP/(TP+FP)
> ACCURACY = (TP+TN)/(TP+FN+TN+FP)
> SENSIBILIDAD =  TP/(TP+FN)
> ESPECIFICIDAD = TN/(TN+FP)
>
> data.frame(TP,FP,FN,TN)
    TP  FP  FN   TN
1 2445 239 173 4066
> data.frame(PRECISION, ACCURACY, SENSIBILIDAD, ESPECIFICIDAD)
  PRECISION  ACCURACY SENSIBILIDAD ESPECIFICIDAD
1 0.9109538 0.9404882     0.933919     0.9444832
```

A. Indicador 1: Precisión: KPI 1

Se estableció previamente que el nivel aceptable estimado para el presente indicador de precisión (KPI1) es de 88%, al hacer uso del modelo de redes neuronales se obtuvo como como resultado un porcentaje del 91% de precisión. Lo que indica que si hay una diferencia a la hora de implementar el modelo de redes neuronales ya que predice de una mejor manera la probabilidad de ACV en pacientes diabéticos.

Figura 30

Gráfico de la precisión estimado y recuperado

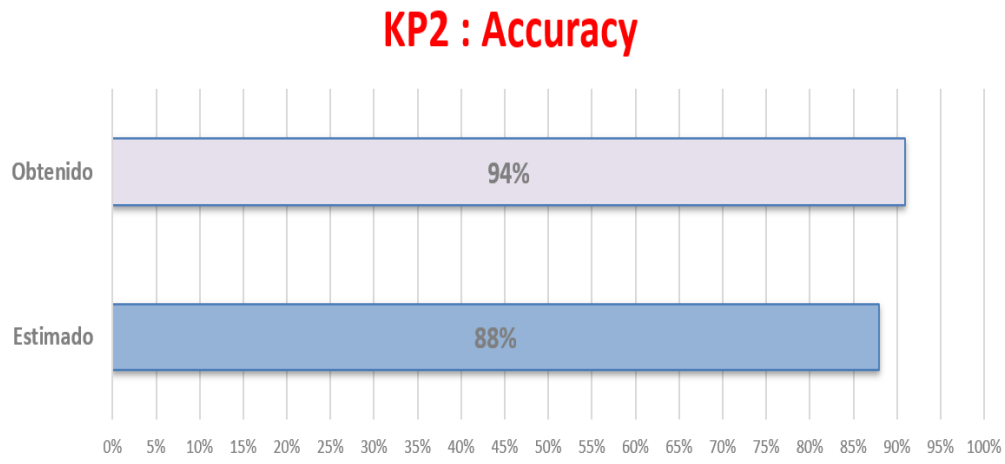


B. Indicador 2: Accuracy: KPI 2

Se estableció previamente que el nivel aceptable estimado para el presente indicador de accuracy (KPI2) es de 88%, al hacer uso del modelo de redes neuronales se obtuvo como como resultado un porcentaje del 94% de accuracy. Lo que indica que si el nivel estimado ha sido superado por los resultados recuperados a la hora de implementar el modelo de redes neuronales ya que predice de una mejor manera la probabilidad de ACV en pacientes diabéticos.

Figura 31

Gráfico del accuracy estimado y recuperado

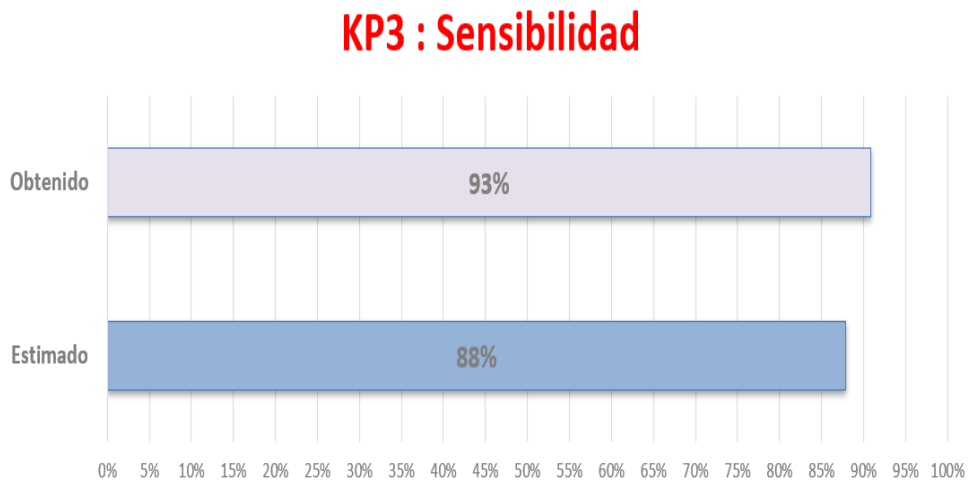


C. Indicador 3: Sensibilidad: KPI 3

Se estableció previamente que el nivel aceptable estimado para el presente indicador de sensibilidad (KPI3) es de 88%, al hacer uso del modelo de redes neuronales se obtuvo como como resultado un porcentaje del 93% de sensibilidad. Lo que indica que si el nivel estimado ha sido superado por los resultados recuperados a la hora de implementar el modelo de redes neuronales ya que predice de una mejor manera la probabilidad de ACV en pacientes diabéticos.

Figura 32

Gráfico de la sensibilidad estimado y recuperado

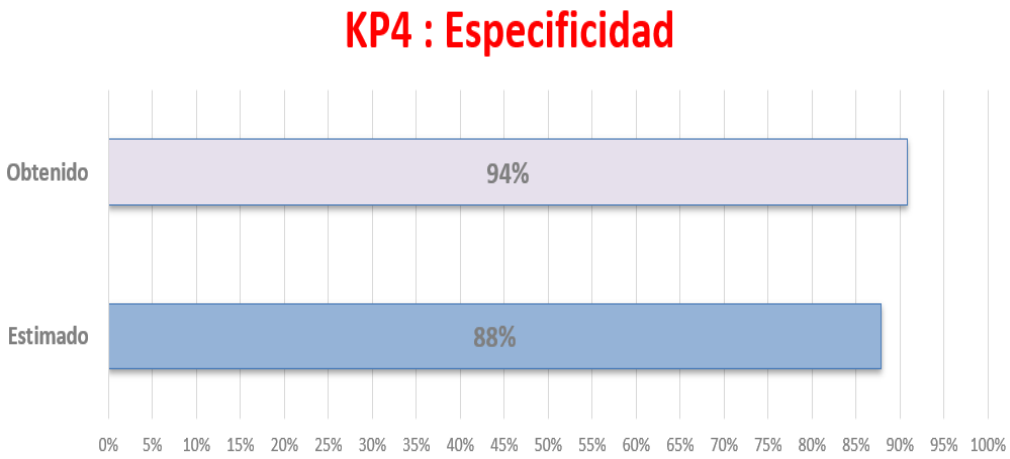


D. Indicador 4: Especificidad: KPI 4

Se estableció previamente que el nivel aceptable estimado para el presente indicador de especificidad (KPI4) es de 88%, al hacer uso del modelo de redes neuronales se obtuvo como como resultado un porcentaje del 94% de especificidad. Lo que indica que si el nivel estimado ha sido superado por los resultados Recuperados a la hora de implementar el modelo de redes neuronales ya que predice de una mejor manera la probabilidad de ACV en pacientes diabéticos.

Figura 33

Gráfico de la especificidad estimado y recuperado



4.4 Nivel de confianza y grado de significancia

La presente investigación cuenta con las siguientes consideraciones:

- Nivel de confianza: 99%
- Nivel de significancia: 1%

4.5 Contrastación de la hipótesis

Para la investigación se presentan 4 indicadores mostrados en la tabla 15:

Tabla 15

Indicadores para la contrastación de la hipótesis

Indicador	Mínimo aceptable	Rendimiento
I1: Precisión	88%	91.09%
I2: Accuracy	88%	94.04%
I3: Sensibilidad	88%	93.39%
I4: Especificidad	88%	94.44%

- **Contrastación para la precisión**

  H_0: El desarrollo de un algoritmo basado en redes neuronales no permite la identificación del índice de precisión.

  H_1: El desarrollo de un algoritmo basado en redes neuronales permite la identificación del índice de precisión.

Donde:

- RN = Rendimiento de la precisión de la red neuronal

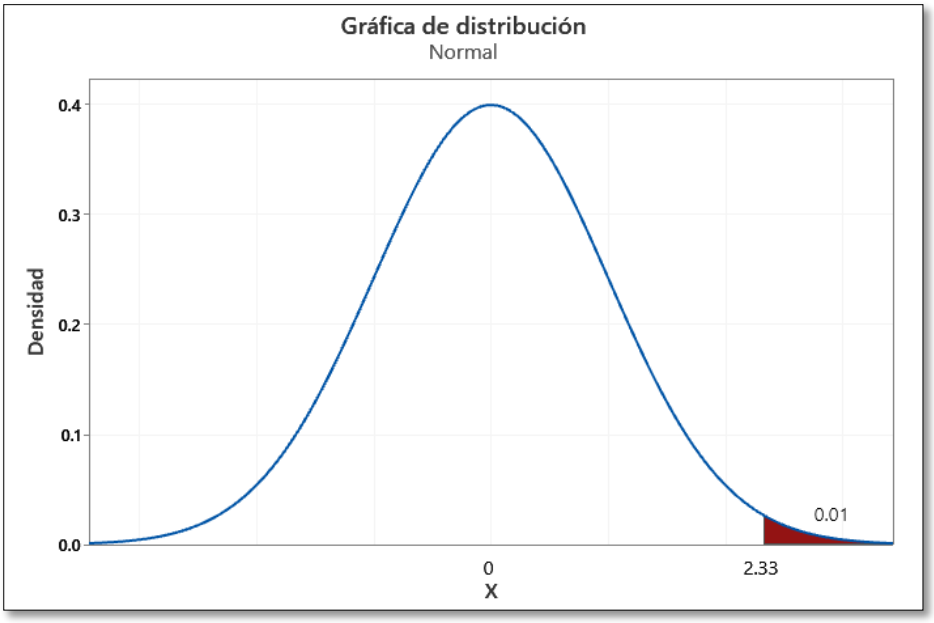
A continuación, dada la naturaleza del indicador (porcentaje) se aplicará el método estadístico de diferencias de proporciones.

$$ZC = \frac{0.9109 - 0.88}{\sqrt{\frac{0.9109(1-0.9109)}{8500} + \frac{0.88(1-0.88)}{8500}}} = 6.5921$$

Región crítica

Figura 34

Gráfica de distribución KPI 1



Decisión:

Dado que el resultado Recuperado 6.5921 > 2.33, se rechaza la hipótesis nula (H0) y se acepta la hipótesis alterna (H1).

Conclusión:

Se demuestra que usando un nivel de significancia del 1% y confianza del 99%, la precisión de redes neuronales para la predicción de ACV en pacientes diabéticos supera el mínimo estimado (88%).

- **Contrastación para el accuracy**

H_0: El desarrollo de un algoritmo basado en redes neuronales no permite la identificación del índice de accuracy.

H_1: El desarrollo de un algoritmo basado en redes neuronales permite la identificación del índice de accuracy.

H_0: RA ≤ 85%

H_1: RA > 85%

Donde:

- RA = Rendimiento del accuracy en la red neuronal

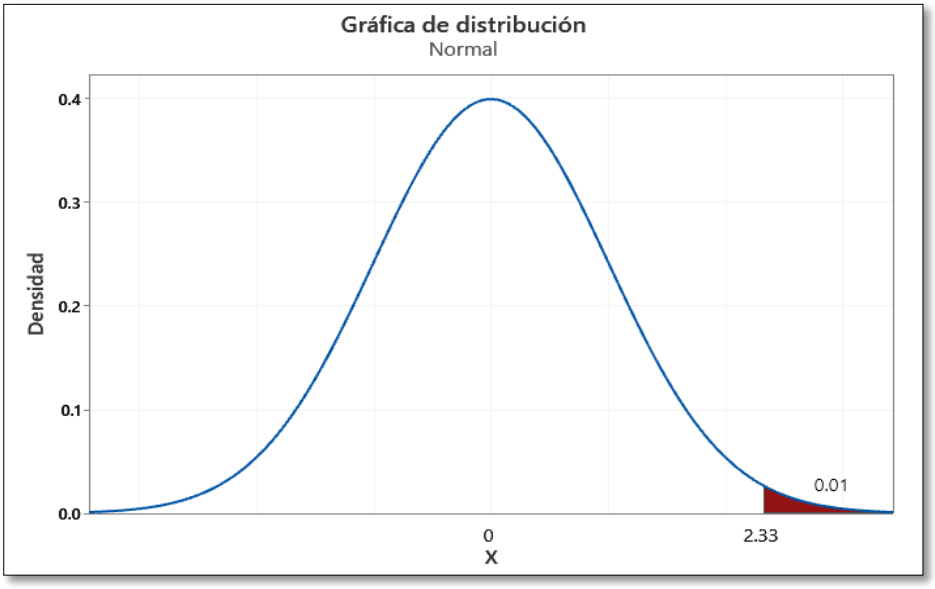
A continuación, dada la naturaleza del indicador (porcentaje) se aplicará el método estadístico de diferencias de proporciones.

$$ZC = \frac{0.9404 - 0.88}{\sqrt{\frac{0.9404(1-0.9404)}{8500} + \frac{0.88(1-0.88)}{8500}}} = 13.8503$$

Región crítica

Figura 35

Gráfica de distribución KPI 2



Decisión

Dado que el resultado Recuperado 13.8503 > 2.33, se rechaza la hipótesis nula (H0) y se acepta la hipótesis alterna (H1).

Conclusión:

Se demuestra que usando un nivel de significancia del 1% y confianza del 99%, la accuracy de redes neuronales para la predicción de ACV en pacientes diabéticos supera el mínimo estimado (88%).

- **Contrastación para la sensibilidad**

  H_0: El desarrollo de un algoritmo basado en redes neuronales no permite la identificación del margen de sensibilidad.

  H_1: El desarrollo de un algoritmo basado en redes neuronales permite la identificación del margen de sensibilidad.

  Donde:

  - RA = Rendimiento de la sensibilidad en la red neuronal

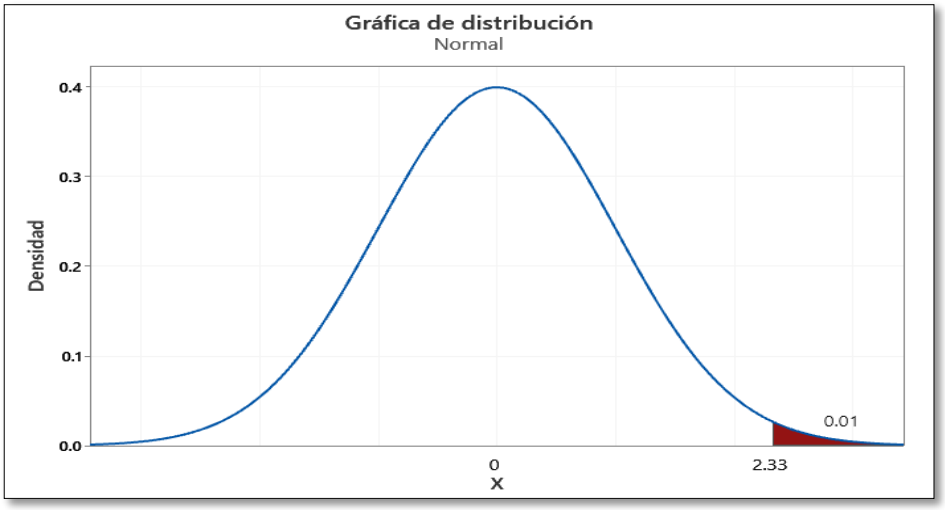
A continuación, dada la naturaleza del indicador (porcentaje) se aplicará el método estadístico de diferencias de proporciones.

$$\mathrm{ZC} = \frac{0.9339 - 0.88}{\sqrt{\frac{0.9339(1-0.9339)}{8500} + \frac{0.88(1-0.88)}{8500}}} = 12.1481$$

Región crítica

Figura 36

Gráfica de distribución KPI 3



Decisión

Dado que el resultado Recuperado 12.1481 > 2.33, se rechaza la hipótesis nula (H0) y se acepta la hipótesis alterna (H1).

Conclusión:

Se demuestra que usando un nivel de significancia del 1% y confianza del 99%, la sensibilidad de redes neuronales para la predicción de ACV en pacientes diabéticos supera el mínimo estimado (88%).

- **Contrastación para la especificidad**

  H_0: El desarrollo de un algoritmo basado en redes neuronales no permite la identificación del margen de especificidad.

  H_1: El desarrollo de un algoritmo basado en redes neuronales permite la identificación del margen de especificidad.

Donde:

- RE = Rendimiento de la especificidad en la red neuronal

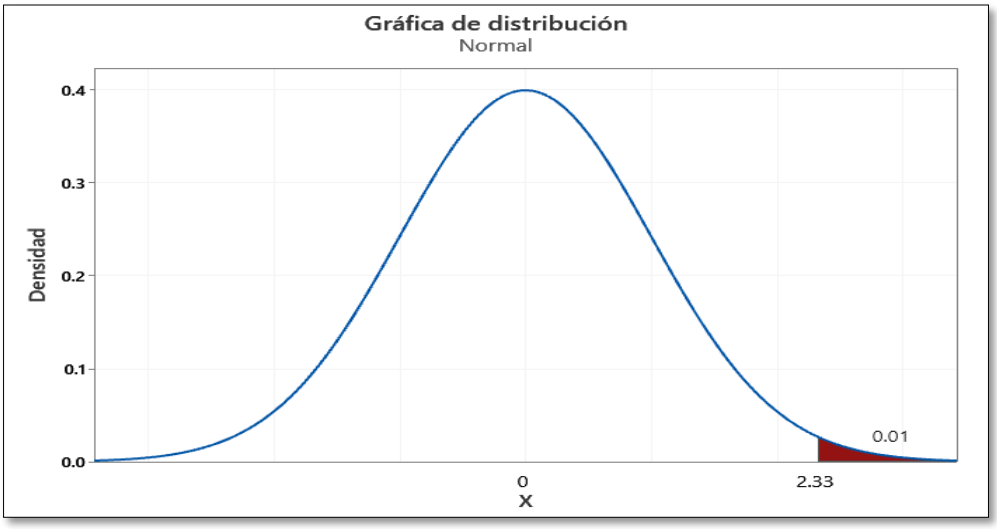
A continuación, dada la naturaleza del indicador (porcentaje) se aplicará el método estadístico de diferencias de proporciones.

$$ZC = \frac{0.9444 - 0.88}{\sqrt{\frac{0.9444(1-0.9444)}{8500} + \frac{0.88(1-0.88)}{8500}}} = 14.9319$$

Región crítica

Figura 37

Gráfica de distribución KPI 4



Decisión:

Dado que el resultado Recuperado 14.9319 > 2.33, se rechaza la hipótesis nula (H_0) y se acepta la hipótesis alterna (H_1).

Conclusión:

Se demuestra que usando un nivel de significancia del 1% y confianza del 99%, la especificidad de redes neuronales para la predicción de ACV en pacientes diabéticos supera el mínimo estimado (88%).

CAPÍTULO V

DISCUSIONES, CONCLUSIONES Y RECOMENDACIONES

5.1 Discusiones

Según con el objetivo general se planteó determinar en qué medida el desarrollar un algoritmo basado en redes neuronales contribuye en la predicción de ACV en pacientes diabéticos, los resultados recuperados en la figura 21, datos que al ser comparados con lo encontrado por Masruriyah et al. (2019) en su tesis titulada *Predictive analytics for stroke disease*, quien concluyó que el uso de un modelo predictivo combinado con la validación cruzada de K-Fold de una muestra de 18,425 pacientes arrojo una predicción de 95,15% para determinar de manera anticipada la ocurrencia de los accidentes cerebros vasculares en dichos pacientes. Además, Cheng & Chiu (2017) en su investigación “An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database”, menciona que el modelo basado en redes neuronales logro un buen desempeño en la predicción de casos de ACV en pacientes de alto riesgo y que podría funcionar como punto de referencia de comunicación a la hora de remitir pacientes.

Según el objetivo específico de esta investigación, se planteó determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de precisión de la predicción , los resultados recuperados en la Figura 21, datos que al ser comparados con lo presentado por Zhanfeng(2016) en su tesis titulada *Evaluación y pensamiento del accidente cerebrovascular basado en una red neuronal artificial en un entorno virtual táctil*, quien concluyó que el usar un modelo de red neuronal combinado con un entorno virtual táctil capaz de evaluar y establecer el estado de los pacientes logrando tener una precisión de predicción del 94%. Además, Nawal & Sreela (2015) en su investigación “Predictive model for transferring stroke in-patients to Intensive Care Unit” menciona que en su investigación que al comparar muchos predictivos como arboles de decisión, máquina de vectores, regresión

logística; el modelo de red neuronal arroja un 94% de precisión para determinar si un paciente hospitalizado por accidente cerebrovasculares necesita ser trasladado al área de UCI.

Según el objetivo específico de esta investigación, se planteó determinar en qué medida el desarrollo de los modelos predictivos para identificar el índice de sensibilidad, los resultados recuperados en el la figura 21, datos que al ser comparados con los presentado por Kyriacou et al. (2015) en su tesis titulada *Prediction of the time period of stroke based on ultrasound image analysis of initially asymptomatic carotid plaques* quien concluyó que la realización de modelos predictivos SVM para grupos tempranos de ACV de largo y corto plazo dio como resultado general un porcentaje de sensibilidad y especificidad de 88 ± 6% y 72 ± 6% respectivamente. Con estos resultados se corrobora que el modelo de redes neuronales para la predicción de ACV (accidentes cerebro-vasculares) arroja mejores resultados de la especificidad. Además, Kansadub et al. (2015) en su investigación titulada como “Stroke risk prediction model based on demographic data”, menciona que una de las mejores opciones para la predicción de ACV es el modelo de redes neuronales indicando de que este, posee un alto índice de especificidad, sensibilidad y precisión.

5.2 Conclusiones

- Se determinó en qué medida el uso de un modelo basado en redes neuronales determinara el margen de sensibilidad para la predicción de ACV en pacientes diabéticos. Esto se ve reflejado en los resultados recuperados con el incremento de la sensibilidad en un 93%, superando el rango mínimo estimado de 88% para esta variable (véase en la figura 24).
- Se determinó en qué medida el desarrollo de un modelo basado en redes neuronales determinara el margen de especificidad para la predicción de ACV en pacientes diabéticos. Esto se ve reflejado en los resultados recuperados con el incremento de la especificidad en un 94%, superando el rango mínimo estimado de 88% para esta variable (véase en la figura 27).
- Como conclusión general en esta tesis se determinó en qué medida el desarrollo un algoritmo basado en redes neuronales contribuye en la predicción de ACV en pacientes diabéticos. Esto confirma lo que dice N. Masruriyah et al. (2019).

5.3 Recomendaciones

- Se recomienda hacer un buen uso de criterio cuando se recopile la data para la red neuronal ya que para una mejor predicción es necesario información relevante, precisa y entendible; como la obtención de recetas médicas, lugar de residencia, etc.
- Se recomienda diseñar un modelo de red neuronal capaz de interpretar los historiales médicos (anamnesis), clasificarlos e introducirlos en un modelo de red neuronal más robusto; da tal modo que se automatice todo el proceso.
- Como recomendación final es necesario tener en claro que el modelo de redes neuronales por sí solo no resolverá el problema, por lo que se recomienda complementar la información obtenida por la red neuronal con acciones mucho más específicas para ayudar a resolución del problema.

REFERENCIAS

Arauz, A. y Ruíz, A. (2012). Enfermedad vascular cerebral. *Revista de la Facultad de Medicina de la UNAM* , *55*(3) , 11-21. https://cutt.ly/uEVXPBe

Behar , R. y Grima , P. (2013). El histograma como un instrumento para la comprensión de las funciones de densidad de probabilidad. *UPC Barcelona*, 229-235. https://cutt.ly/aEVXZaF

Celis, J., Hernandez, D. y King, L. (2019). Guía neurológica 8 enfermedad cerebrovascular.*Acta Neurologica Colombiana*. https://cutt.ly/iEVX9p1

Chan, M. (2016). *INFORME MUNDIAL SOBRE LA DIABETES.* Organizacion Mundial de la Salud. https://cutt.ly/PEVCzb8

Cheng, C. & Chiu, H. (2017). *An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database* [Sesión de conferencia]. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Jeju, Korea (South). doi: 10.1109/EMBC.2017.8037381

Chunxiao, Y. (2019). *Research on Cerebrovascular Disease Prediction System Based on the Long Short Term Memory Neural Network* [Tesis de maestria,Beijing Jiaotong University]. China Academic Journal. https://cutt.ly/WEVCQMn

Graeme, J. (2016). *Diagnóstico y tratamiento actualizado - Accidente cerebrovascular.* IntraMed. https://cutt.ly/FEVCI96

Guo, Y., Sun, L., Zhang, Z., & He, H. (2019). *Algorithm Research on Improving Activation Function of Convolutional Neural Networks* [Sesión de conferencia]. 2019 Chinese Control And Decision Conference (CCDC). Nanchang, China. doi: 10.1109/CCDC.2019.8833156

Gutiérrez, B., Cintas, G. (2013). El histograma como un instrumento para la comprensión de las funciones de densidad de probabilidad. *Discovery UPC*, 229-235. https://cutt.ly/hEBdTlO

Hung, C., Lin, C., & Lee, C. (2018). Improving Young Stroke Prediction by Learning with Active Data Augmenter in a Large-Scale Electronic Medical Claims Database [Sesión de conferencia]. 2018 40th Annual International Conference

of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, HI, USA. doi: 10.1109/EMBC.2018.8513479

Kansadub, T., Thammaboosadee, S., Kiattisin, S., & Jalayondeja, C. (2015). Stroke risk prediction model based on demographic data [Sesión de conferencia]. 2015 8th Biomedical Engineering International Conference (BMEiCON). Pattaya, Thailand. doi: 10.1109/BMEiCON.2015.7399556

Kyriacou, E., Vogazianos, P., Christodoulou, C., Loizou, C., Panayides, A.S., Petroudi, S., Pattichis, M., Pantziaris, M., Nicolaides, A. & Pattichis, C.S. (2015). *Prediction of the time period of stroke based on ultrasound image analysis of initially asymptomatic carotid plaques* [Sesión de conferencia]. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC*).* Milan, Italy. doi: 10.1109/EMBC.2015.7318367

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 436-444. https://cutt.ly/eEVC2NU

Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones* [Tesis Doctoral, Universidad Tecnológica Nacional]. Biblioteca Electrónica de la UTN https://cutt.ly/UEVC7U6

Montaño Moreno, J. J. (2002). *Redes Neuronales Artificiales aplicadas al Análisis de Datos* [Tesis Doctoral , Universidad de las Islas Baleares]*.* Tesis Doctorals en Xarxa https://cutt.ly/nEVVskB

Morera-Munt, A. (2018). *Introducción a los modelos de redes neuronales artificiales-El Perceptrón simple y multicapa* [Tesis de pregrado,Universidad Saragoza]*.* Saguan Repositorio Institucional de Documentos https://cutt.ly/lEVVf7v

Nawal , N., & Sreela, S. (2015). Predictive model for transferring stroke in-patients to Intensive Care Unit [Sesión de conferencia]. 2015 International Conference on Computing and Network Communications (CoCoNet). Trivandrum, India. doi: 10.1109/CoCoNet.2015.7411288

Nur-Masruriyah, A., Djatna, T., Dewi, M., Handayani, H., & Wahiddin, D. (2019). *Predictive Analytics For Stroke Disease* [Sesión de conferencia]. 2019 Fourth

International Conference on Informatics and Computing (ICIC*)*. Semarang, Indonesia. doi: 10.1109/ICIC47613.2019.8985716

Palladino, A. (2011). GRÁFICO DE CAJA*. Universidad Nacional del Nordeste,* 1-4. https://cutt.ly/vEVVJ55

Peñafiel, M. (2018). *9 factores de riesgo (modificables y no) de accidente cerebrovascular*. ELSEVIER. https://cutt.ly/kEVV7LA

SAS Enterprise Miner. (2006). *Enterprise MinerTM SEMMA*. SAS. https://cutt.ly/UEVBK7P

Slideshare (s.f.). *Diseños de investigación.* https://www.slideshare.net/AlbertAP/diseos-de-investigacin-69328635

Tan, P., Steinbach, M. & Kumar, V. (2013). *Introduction to Data Mining Tan Steinbach Kumar First Edition.* Pearson.

Vijaya-Kumar, D., & Rama-Krishniah, J. (2016). *An automated framework for stroke and hemorrhage detection using decision tree classifier* [Sesión de conferencia]. 2016 International Conference on Communication and Electronics Systems (ICCES*).* Coimbatore, India. doi: 10.1109/CESYS.2016.7889861

Yang, H. (2019). *Study on the Segmentation Method of Stroke in Chronic Stage* [Tesis de maestria, Universidad de la Academia de Ciencias de China]. Red de conocimiento de China (CNKI) https://cutt.ly/wEVMW38

Zhanfeng, W. (2016). *Stroke Evaluation and Reflection with Artificial Neural Network in Haptic Virtual Environment.* https://cutt.ly/zEBpbPc

ANEXOS

Anexo 1: Matriz de consistencia

<table>
<tr><th>PROBLEMA</th><th>OBJETIVOS</th><th>HIPÓTESIS</th><th>VARIABLES</th><th></th></tr>
<tr>
<td>Problema general
¿En qué medida el desarrollo de un algoritmo basado en redes neuronales contribuirá en la predicción de ACV en pacientes diabéticos?</td>
<td>Objetivo general
Determinar en qué medida el desarrollar un algoritmo basado en redes neuronales contribuye en la predicción de ACV en pacientes diabéticos.</td>
<td>Hipótesis general
Si se desarrolla un algoritmo basado en redes neuronales, entonces se podrá realizar la predicción de ACV de los pacientes diabéticos.</td>
<td rowspan="2">Independiente
X: Redes neuronales

Dependiente
Y: Predicción de ACV en pacientes diabéticos</td>
<td rowspan="2">Tipo de investigación
Aplicada

Nivel de investigación:
Correlacional

Diseño de investigación
Transversal

Población
17372 pacientes

Muestra:
8500 pacientes</td>
</tr>
<tr>
<td>Problemas específicos
¿En qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de precisión?
¿En qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de accuracy?
¿En qué medida el desarrollo de un algoritmo basado en redes neuronales determinara el margen de sensibilidad?
¿En qué medida el desarrollo de un algoritmo basado en redes neuronales determinara el margen de especificidad?</td>
<td>Objetivos específicos
Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de precisión.
Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales identificara el índice de accuracy.
Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales determinara el margen de sensibilidad.
Determinar en qué medida el desarrollo de un algoritmo basado en redes neuronales determinara el margen de especificidad.</td>
<td>Hipótesis específicas
El desarrollo de un algoritmo basado en redes neuronales permite la identificación del índice de precisión.
El desarrollo de un algoritmo basado en redes neuronales permite la identificación del índice de accuracy.
El desarrollo de un algoritmo basado en redes neuronales permite la identificación del margen de sensibilidad.
El desarrollo de un algoritmo basado en redes neuronales permite la identificación del margen de especificidad.</td>
</tr>
</table>

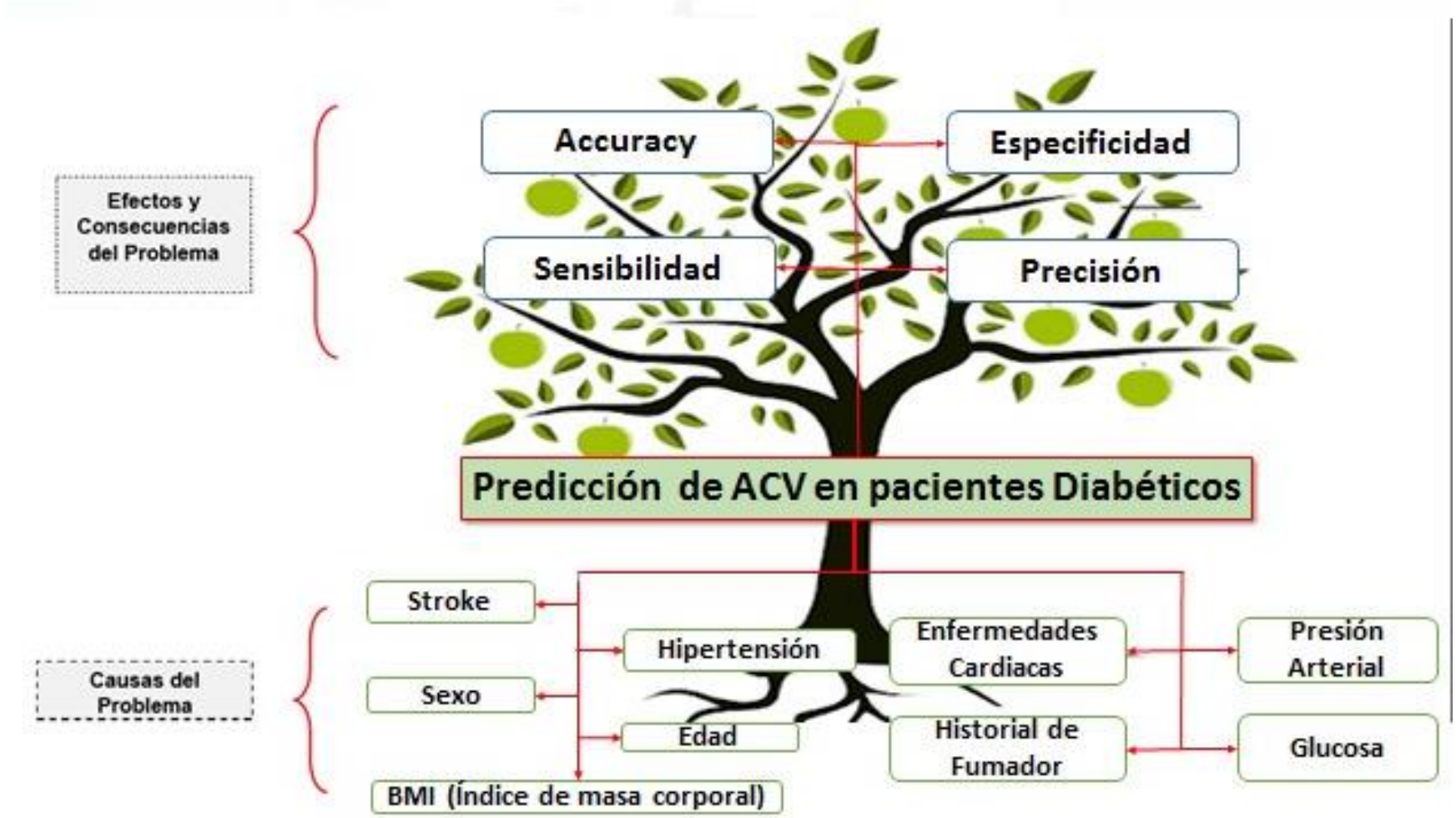
Anexo 2: Matriz de operacionalización de variables

Variables	Dimensión	Indicadores	Índice	Unidad de medida	Técnica	Unidad de investigación	Instrumento
Redes neuronales	Presencia- Ausencia		SI-NO			Pacientes diabéticos	Software
Predicción de ACV en pacientes diabéticos	Variables de Predicción	Precisión	> 88%	Porcentaje	Revisión de reporte del software	Probabilidad de ACV	Reporte
		Accuracy					
		Sensibilidad					
		Especificidad					
	Indicadores	Stroke	[0 - 1]	[0 - 1]	Revisión de documentos	Stroke	Reporte
		Hipertensión	[0 - 1]	[0 - 1]		Hipertensión	
		Glucosa	[80 - 200]	Decilitros (mg/dl)		Glucosa	
		Enfermedades cardiacas	[0 - 1]	[0 - 1]		Enfermedades cardiacas	
		Presión arterial	[60 - 200]	mmHg (mililitros de mercurio)		Presión arterial	
		BMI	[20.0 – 88.7]	Kg/cm^2		BMI	
		Historial de fumador	[nunca, no actualmente, ocasional recurrente siempre]	[1 – 5]		Historial de fumador	
		Edad	[20 - 80]	años		Edad	
		Sexo	[Hombre-Mujer]	[1 – 2]		Sexo	

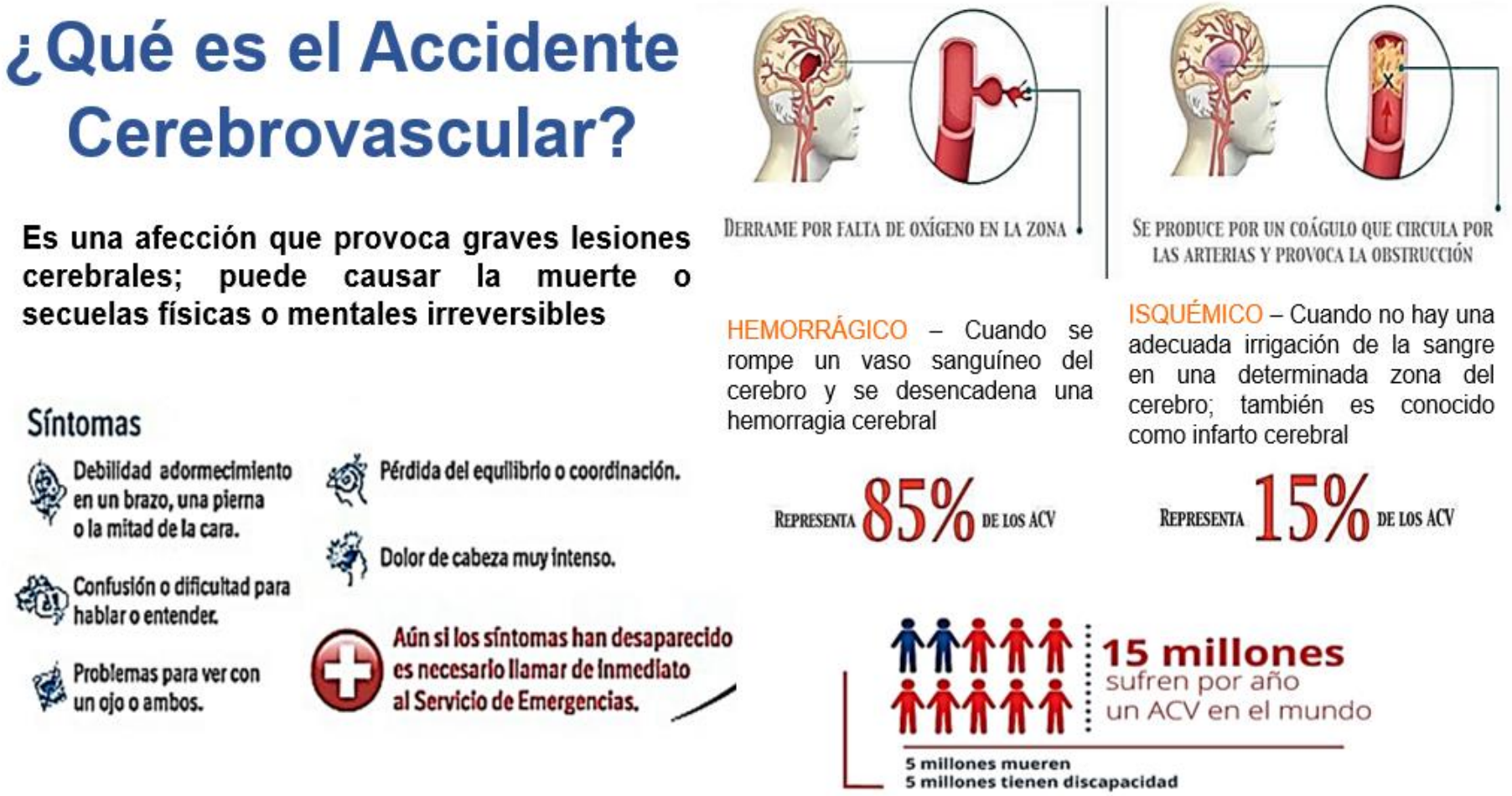
Anexo 3: Matriz de revisión de literatura

Título	Autor	Año	Tipo	Descripción
Study on the segmentation method of stroke in chronic stage	Yang, Hao	2019	Tesis	Universidad de la Academia China de Ciencias (Instituto Shenzhen de Tecnología Avanzada, Academia China de Ciencias)
Stroke evaluation and reflection with artificial neural network in haptic virtual environment	Zhanfeng, Wu	2016	Tesis	Universidad de Tecnología, Fabricación y Automatización de Maquinaria de China Oriental
Research on cerebrovascular disease prediction system based on the long short term memory neural network	Chunxiao, Yao	2019	Tesis	Universidad Beijing Jiaotong
Predictive analytics for stroke disease	Nur Masruriyah, Anis Fitri; Djatna, Taufik; Dewi Hardhienata, Medria Kusuma; Handayani, Hanny Hikmayanti; Wahiddin, Deden	2019	Artículo	2019 Fourth International Conference on Informatics and Computing (ICIC)
Improving young stroke prediction by learning with active data augmenter in a large-scale electronic medical claims database	Hung, Chen-Ying; Lin, Ching-Heng; Lee, Chi-Chun	2018	Artículo	2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)
Prediction of the time period of stroke based on ultrasound image analysis of initially asymptomatic carotid plaques	Kyriacou, Efthyvoulos; Vogazianos, P.; Christodoulou, Christodoulos; Loizou, Christos P.; Panayides, Andreas S.; Petroudi, S.; Pattichis, Marios S.; Pantziaris, Marios; Nicolaides, Andrew; Pattichis, Constantinos S.	2015	Artículo	2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

Anexo 4: Matriz del problema



Anexo 5: Imagen pictográfica del problema



Anexo 6: Cuestionario

CUESTIONARIO

Nombre y Apellidos: ……

Especialidad: ……. DNI: …….

El presente cuestionario tiene como finalidad identificar el **mínimo porcentaje** de precisión, accuracy, sensibilidad y especificidad con respecto a la predicción de un accidente cerebro-vascular en pacientes diabéticos.

A continuación, se define el concepto de los respectivos indicadores.

Precisión	determina la fracción de registros que realmente resulta ser positivo en el grupo que el clasificador ha declarado como una clase positiva.
Accuracy	Es la sumatoria de los registros con resultados correctos dentro de la predicción
Sensibilidad	Corresponde a la fracción de registros positivos predichos correctamente en el modelo.
Especificidad	Corresponde a la fracción de registros negativos predichos correctamente en el modelo.

Marco con un aspa **(x)** la opción que considere más optima.

		PORCENTAJE MÍNIMO ACEPTABLE				
		75%	80%	85%	88%	90%
1	¿Qué porcentaje de eficiencia es considerado aceptable para la precisión de ACV en pacientes diabéticos?					
2	¿Qué porcentaje de eficiencia es considerado aceptable para el accuracy de ACV en pacientes diabéticos?					
3	¿Qué porcentaje de eficiencia es considerado aceptable para la sensibilidad de ACV en pacientes diabéticos?					
4	¿Qué porcentaje de eficiencia es considerado aceptable para la especificidad de ACV en pacientes diabéticos?					

Anexo 7: Validación de juicio de expertos

CERTIFICADO DE VALIDEZ DE CONTENIDO DE INSTRUMENTOS A TRAVÉS DE JUICIO DE EXPERTO

Título de la investigación	DESARROLLO DE UN ALGORITMO CON REDES NEURONALES PARA LA PREDICCION DE ACV EN PACIENTES DIABÉTICOS
Nombre(s) del(os) instrumento(s)	Cuestionario
Autor(es) del instrumento	ASCUE SILVA SAÚL SEBASTIAN, OLASCOAGA ROMAN LUIS ALONSO

Nº	DIMENSIONES / Indicadores	Pertinencia[1]		Relevancia[2]		Claridad[3]		Sugerencias
	DIMENSIÓN 1: Variables de Predicción	Si	No	Si	No	Si	No	
1	Precisión	/		/		/		
2	Accuracy	/		/		/		
3	Sensibilidad	/		/		/		
4	Especificidad	/		/		/		
	DIMENSIÓN 2: Indicadores	Si	No	Si	No	Si	No	
5	Stroke	/		/		/		
6	Hipertensión	/		/		/		
7	Glucosa	/		/		/		
8	Presión Arterial	/		/		/		
9	Sexo	/		/		/		
10	Edad	/		/		/		
11	Enfermedades Cardiacas	/		/		/		
12	Historial de Fumador	/		/		/		
13	BMI (Indice de masa corporal)	/		/		/		

Observaciones (precisar si hay suficiencia): .....

Opinión de aplicabilidad: **Aplicable [/]** **Aplicable después de corregir [ ]** **No aplicable [ ]**

Apellidos y nombres del juez validador. Dr/ Mg: Jan Hermosa Dueñas **DNI:** 41922075

Especialidad del validador: Médico Cirujano

29 de julio del 2020

[1]**Pertinencia:** El ítem corresponda al concepto teórico formulado.
[2]**Relevancia:** El ítem es apropiado para representar al componente o dimensión específica del constructo
[3]**Claridad:** Se entiende sin dificultad alguna el enunciado del ítem, es conciso, exacto y directo

Nota: Suficiencia, se dice suficiencia cuando los ítems planteados son suficientes para medir la dimensión

Firma del Experto Informante.

CERTIFICADO DE VALIDEZ DE CONTENIDO DE INSTRUMENTOS A TRAVÉS DE JUICIO DE EXPERTO

Título de la investigación	DESARROLLO DE UN ALGORITMO CON REDES NEURONALES PARA LA PREDICCION DE ACV EN PACIENTES DIABÉTICOS
Nombre(s) del(os) instrumento(s)	Cuestionario
Autor(es) del instrumento	ASCUE SILVA SAÚL SEBASTIAN, OLASCOAGA ROMAN LUIS ALONSO

Nº	DIMENSIONES / Indicadores	Pertinencia[1]		Relevancia[2]		Claridad[3]		Sugerencias
DIMENSIÓN 1: Variables de Predicción		Si	No	Si	No	Si	No	
1	Precisión	X		X		X		
2	Accuracy	X		X		X		
3	Sensibilidad	X		X		X		
4	Especificidad	X		X		X		
DIMENSIÓN 2: Indicadores		Si	No	Si	No	Si	No	
5	Stroke	X		X		X		
6	Hipertensión	X		X		X		
7	Glucosa	X		X		X		
8	Presión Arterial	X		X		X		
9	Sexo	X		X		X		
10	Edad	X		X		X		
11	Enfermedades Cardiacas	X		X		X		
12	Historial de Fumador	X		X		X		
13	BMI (indice de masa corporal)	X		X		X		

Observaciones (precisar si hay suficiencia): .....

Opinión de aplicabilidad: **Aplicable [X]** **Aplicable después de corregir [ ]** **No aplicable [ ]**

Apellidos y nombres del juez validador. Dr/ Mg: **Pedro Martin Lezama Gonzales** **DNI:** 09656793

Especialidad del validador: Ingeniero de sistemas

26 de julio del 2020

Firma del Experto Informante.

[1]**Pertinencia:** El ítem corresponde al concepto teórico formulado.
[2]**Relevancia:** El ítem es apropiado para representar al componente o dimensión específica del constructo
[3]**Claridad:** Se entiende sin dificultad alguna el enunciado del ítem, es conciso, exacto y directo

Nota: Suficiencia, se dice suficiencia cuando los ítems planteados son suficientes para medir la dimensión

CERTIFICADO DE VALIDEZ DE CONTENIDO DE INSTRUMENTOS A TRAVÉS DE JUICIO DE EXPERTO

Título de la investigación	**DESARROLLO DE UN ALGORITMO CON REDES NEURONALES PARA LA PREDICCION DE ACV EN PACIENTES DIABÉTICOS**
Nombre(s) del(os) instrumento(s)	**Cuestionario**
Autor(es) del instrumento	**ASCUE SILVA SAÚL SEBASTIAN, OLASCOAGA ROMAN LUIS ALONSO**

Nº	DIMENSIONES / Indicadores	Pertinencia[1]		Relevancia[2]		Claridad[3]		Sugerencias
DIMENSIÓN 1: Variables de Predicción		Si	No	Si	No	Si	No	
1	Precisión	X		X		X		
2	Accuracy	X		X		X		
3	Sensibilidad	X		X		X		
4	Especificidad	X		X		X		
DIMENSIÓN 2: Indicadores		Si	No	Si	No	Si	No	
5	Stroke	X		X		X		
6	Hipertensión	X		X		X		
7	Glucosa	X		X		X		
8	Presión Arterial	X		X		X		
9	Sexo	X		X		X		
10	Edad	X		X		X		
11	Enfermedades Cardiacas	X		X		X		
12	Historial de Fumador	X		X		X		
13	BMI (indice de masa corporal)	X		X		X		

Observaciones (precisar si hay suficiencia): .....

Opinión de aplicabilidad: **Aplicable [X]** **Aplicable después de corregir [ ]** **No aplicable [ ]**

Apellidos y nombres del juez validador. Dr/ Mg: José Luis Herrera Salazar **DNI:** 41922075

Especialidad del validador: Dr. en Sistemas

27 de junio del 2020

Firma del Experto Informante.

[1]**Pertinencia:** El ítem corresponde al concepto teórico formulado.
[2]**Relevancia:** El ítem es apropiado para representar al componente o dimensión específica del constructo
[3]**Claridad:** Se entiende sin dificultad alguna el enunciado del ítem, es conciso, exacto y directo

Nota: Suficiencia, se dice suficiencia cuando los ítems planteados son suficientes para medir la dimensión

Anexo 8: Resolución de cuestionario

CUESTIONARIO

Nombre y Apellidos: .....

Especialidad: ..... DNI: .....

El presente cuestionario tiene como finalidad identificar el **mínimo porcentaje** de precisión, accuracy, sensibilidad y especificidad con respecto a la predicción de un accidente cerebro-vascular en pacientes diabéticos.

A continuación, se define el concepto de los respectivos indicadores.

Precisión	determina la fracción de registros que realmente resulta ser positivo en el grupo que el clasificador ha declarado como una clase positiva.
Accuracy	Es la sumatoria de los registros con resultados correctos dentro de la predicción
Sensibilidad	Corresponde a la fracción de registros positivos predichos correctamente en el modelo.
Especificidad	Corresponde a la fracción de registros negativos predichos correctamente en el modelo.

Marco con un aspa **(x)** la opción que considere más optima.

		PORCENTAJE MÍNIMO ACEPTABLE				
		75%	80%	85%	88%	90%
1	¿Qué porcentaje de eficiencia es considerado aceptable para la precisión de ACV en pacientes diabéticos?				X	
2	¿Qué porcentaje de eficiencia es considerado aceptable para el accuracy de ACV en pacientes diabéticos?				X	
3	¿Qué porcentaje de eficiencia es considerado aceptable para la sensibilidad de ACV en pacientes diabéticos?				X	
4	¿Qué porcentaje de eficiencia es considerado aceptable para la especificidad de ACV en pacientes diabéticos?				X	

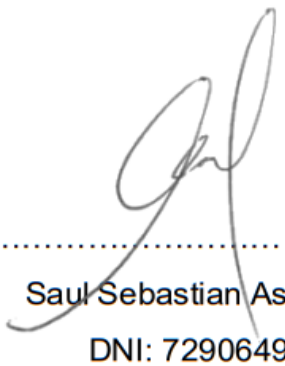
Anexo 9: Declaración Jurada

DECLARACIÓN JURADA

Yo Saúl Sebastian Ascue Silva identificado con DNI N° 72906491, código de estudiante N° 2151891920 y Luis Alonso Olascoaga Roman identificado con DNI N° 76906300, código de estudiante N° 2152891462; Egresados de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Autónoma del Perú.

Declaramos bajo juramento la responsabilidad de los datos usados en el desarrollo de la tesis "DESARROLLO DE UN ALGORITMO CON REDES NEURONALES PARA LA PREDICCIÓN DE ACV EN PACIENTES DIABÉTICOS" con la cual buscamos obtener el Titulo de Ingenieros de Sistemas.

Lima, 24 de noviembre del 2021



Saul Sebastian Ascue Silva	Luis Alonso Olascoaga Roman
DNI: 72906491	DNI: 76906300