Predicción de accidentes cerebrovasculares mediante K-Nearest Neighbor

Title of the article in english

Beatriz, Balan Fiuza[1], 0009-0004-7509-2146, beatrizbalan@upeu.edu.pe

Nombre(s) y Apellido(s) [2], ORCID iD, correo institucional

[1] Universidad Peruana Unión, Lima, Perú

[2] Institución de afiliación, ciudad, país (si los autores provienen de una misma institución, colocar solo un sub índice)

Autor de correspondencia: correo institucional

Resumen: Debe contener entre 150 a 200 palabras (100 para comunicaciones cortas y notas científicas), incluyendo: justificación (breve introducción), objetivo, materiales y métodos, resultados y conclusiones (principales). Debe estar escrito en un solo párrafo. El resumen debe estar escrito en tiempo pasado.

Palabras clave: Se deben colocar de tres a cinco palabras clave, en orden alfabético separados por una coma, que no estén incluidas en el título y que identifiquen el contenido del artículo.

Abstract: Traducción fiel del resumen.

Keywords: Traducción fiel de las palabras clave.

1. Introducción

El accidente cerebrovascular (ACV) es una patología caracterizada por ser una lesión encefálica aguda causada por la obstrucción o ruptura de vasos sanguíneos cerebrales Los ACV's se clasifican en 2 grandes grupos: a) ACV isquémico y b) ACV hemorrágico (1). Aunque haya habido avances cuanto, a la compresión y

tratamiento de los accidentes cerebrovasculares, esta patología sigue siendo una causa de defunciones en el mundo y en el Perú. ((2)

Accidente Cerebrovascular:

Aproximadamente, más de 13,7 millones accidentes cerebrovasculares son reportados por año, menciona la organización mundial de ACV. De este grupo de pacientes 60% aproximadamente son de <70 años y 25% > 25 años y más de 2.7 millones de personas llegan al fallecimiento. (3,4). En Perú, hay una vigilancia escasa para este problema, pero se estima que de cada 100 000 personas 117 por año desarrollarán un ACV. (5)

De una forma más específica, el ACV se define como el déficit neurológico focal agudo en cualquier parte del Sistema nerviosos central (SNC), debido a causas vasculares (2,6). Las características clínicas del ictus un déficit neurólogico focal de inicio súbito, cursando con disartria o Glasgow disminuido, migraña, convulsiones y etc., pero siempre cursando con factores de riesgo presentes en la historia clínica del paciente, como: diabetes mellitus (DM), hipertensión arterial (HTA), dislipidemia. (7)

La escala de NIHSS es la responsable por la evaluación de riesgo y gravedad del ACV, constituida de 11 ítems que permiten una evaluación de las funciones corticales, pares craneales, función motora, sensibilidad, coordinación y lenguaje. Los puntos específicos son: 1) valoración del nivel de conciencia (incluye también respuesta a preguntas y a órdenes motoras), 2) mirada conjugada, 3) campos visuales, 4)paresia facial, 5)paresia de extremidades superiores, 6) paresia de extremidades inferiores, 7) ataxia de las extremidades, 8) sensibilidad, 9) lenguaje, 10) disartria y 11) extinción-negligencia-inatención. (8–10)

Como notado anteriormente el riesgo de ACV se incrementan cuando hay avance de la edad, para ambos sexos. (11) Pero, el riesgo de que se produzca esta condición médica también se asocia a las patologías crónicas (DM, HTA y dislipidemia), presentes en la mayoría de los pacientes, pueden ser modificables o no. (12)

- *Modificables*: Hipertensión, diabetes, consumo de alcohol, cigarro y drogas, sedentarismo, hiperlipidemia, dieta.
- *No modificables*: Edad, sexo, etnia, genética

Principalmente, en decurrencia del desbalance de los factores de riesgo, se incrementa la susceptibilidad de ACV. (13) Se presenta como un problema de salud extenso, complejo y difícil de precisar debido a su inestabilidad. (14) Debido a esta gran importancia este trabajo tiene como objetivo predecir, utilizando el modelo de Machine Learning y K – Nearest Neighbor, los eventos de A. cerebrovasculares.

Machine Learning

Es una inteligencia artificial que permite el desarrollo y aprendizaje de las máquinas de una forma directa que debe ser adquirida de un conjunto de datos (conocido como data set), mediante un entrenamiento. El aprendizaje en este contexto está asociado con el reconocimiento de patrones complejos y la toma de decisiones inteligentes basadas en datos. Lo que lo vuelve complicado es que en el conjunto de todas las decisiones posibles dadas todas las entradas posibles son demasiadas complejas para describirlo. Para poder solucionar este problema, Machine Learning desarrolla algoritmos que descubren conocimientos a partir de datos concretos, basándose en sólidos principios estadísticos y computacionales. (15)

Estos algoritmos tienen la capacidad de predecir nuevos resultados siguiendo lo aprendido mediante los datos que fueron usados para su entrenamiento.(16)
Broussard (2018) clasifica a los tipos de aprendizaje en tres tipos:

- ***Supervisado:*** Este aprendizaje nos enseña a que cada entrada en el sistema debe tener una salida. Tiene como objetivo aprender la relación entre la entrada y la salida. El aprendizaje supervisado se divide en: Clasificación: las salidas del sistema son finitas y discretas y son interpretadas como la clase a la que pertenece. Por ejemplo: “0” o “1”; “Falso” o “Verdadero”; “Sí” o “No”. o Regresión: las salidas son continuas.
- ***No Supervisado:*** Nos enseña que para cada entrada no hay ninguna salida, por lo que se le permite a este encontrar algún tipo de estructura en la entrada.
- ***Reinforcement***: El programa interactúa con un ambiente dinámico en el cual debe realizar un objetivo específico. En este caso, se le brinda al algoritmo premios y castigos, a medida que va transitando el problema. Los premios y castigos son utilizados con el objetivo de que el algoritmo aprenda las consecuencias de sus decisiones.

El Machine Learning tiene el potencial de mejorar considerablemente aspectos del análisis empírico, a diferencia de los demás métodos estadísticos tradicionales.
Incluye avances en la predicción, mediante el procesamiento rápido de grandes cantidades de datos, detección de las relaciones no lineales y de orden superior entre exposiciones y factores de confusión; y en la mejora de la precisión de la predicción.

La categorización tradicional de los enfoques de ML, se pueden clasificar en 3 ramas principales:
1) El aprendizaje no supervisado es un término general para los algoritmos que aprender patrones a partir de datos no etiquetados, es decir, variables que no están etiquetadas por un ser humano. Por ejemplo, el aprendizaje no supervisado agrupará las instancias de datos en función de la similitud.
2) El aprendizaje supervisado comprende algoritmos que aprenden una función, que asigna una entrada a una salida, mediante el uso de datos etiquetados, es decir, a los

valores de las categorías de la variable de resultado se les asignan etiquetas o etiquetas significativas.
3) El aprendizaje por refuerzo tiene que ver con un agente inteligente tomando decisiones para un entorno y mejorando sobre la base de la noción de recompensa acumulativa, es decir, el agente variará y optimizará la entrada sobre la base de la retroalimentación del entorno.

K-Nearest Neighbor

Dentro de los algoritmos disponibles en machine learning, encontramos a KNN, que es un algoritmo supervisado donde se toma un conjunto de datos de prueba y se calcula la distancia que hay entre el dato y todo el conjunto de entrenamiento. Seguidamente, se buscan los k datos del entrenamiento que estén más cercanos al dato de prueba. Se realiza una votación en sustento a la clase de K datos de entrenamiento seleccionados, que se encuentran clasificados. La clase que fue más votada es la clase que se le asigna a dato de prueba que se estaba clasificando.

KNN es un método de aproximación simple no paramétrica basado en la regla del vecino más cercano, que consiste en estimar el valor de un dato desconocido a partir de las características del dato más próximo, según una medida de similitud o distancia. El método del vecino más cercano se puede extender utilizando no uno, sino un conjunto de datos más cercanos para predecir el valor de los nuevos datos, en lo que se conoce como los k-vecinos más cercanos

2. Materiales y Métodos:

El conjunto datos (data set) fue encontrado en la plataforma Kaggle, la fuente de datos es de confidencial, se denomina Conjunto de datos de predicción de accidentes cerebrovasculares. Este data set presenta 12 atributos y (?) instancias/benignos y malignos, los datos vacíos fueron excluídos, resultando en x registros, de los cuales x son positivos y x negativos.

Próximamente, en la tabla 1,se describe la información de los atributos presentes en el data set, tipo de dato y sus respectivos valores, al final de esto se creó el archivo “stroke_dataset_limpio.csv”

Atributo	Tipo de Atributo	Rango de Valores
Sexo	Nominal	Masculino, Femenino, Otro
Edad	Numérico Continuo	0 a 100+ años
HTA	Binario (Nominal)	0 = No hipertensión 1 = Sí hipertensión
Enf_Cardiaca	Binario (Nominal)	0 = No enfermedad cardiaca

		1 = Sí enfermedad cardiaca
Est_Civil	Binario (Nominal)	0 = No Casado 1 = Sí casado
Trabajo	Nominal	Private Self-employed Govt_job Children Never_worked
Residencia	Binario (Nominal)	0 = Rural 1 = Urbano
Glucosa	Numérico Continuo	55 a 270 mg/dL aproximadamente
IMC	Numérico Continuo	10 a 97 (Índice de Masa Corporal)
Tabaco	Nominal	formerly smoked never smoked smokes, Unknown
Stroke	Binario (Nominal)	0 = No sufrió ACV 1 = Sí sufrió ACV

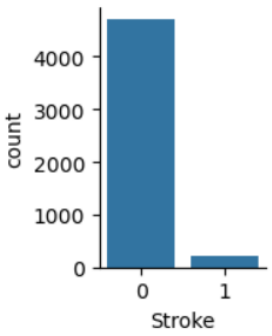
Tabla 1 - Información de los atributos del Data Set

Entrenamiento del Algoritmo con KNN

Se realizó el entrenamiento del algoritmo KNN y su implementación en Python (anaconda) importando las librerías Scikit Learn, de la siguiente manera:

Lo primero que se hizo fue subir el conjunto de datos a un data frame utilizando el comando pd.read_csv (' stroke_dataset_limpio.csv, sep=','), en la Fig.1, se muestra la distribución de datos agrupados por el resultado, donde No= 0 y Yes=

Figura 1. Distribución de datos por el resultado



Segundo, se importo las librerías from sklearn.model_selection import train_test_split; seleccionar aleatoriamente el 70% de los datos para entrenamiento y el 30% restante para las pruebas; tomar como valor inicial n_neighbors=1 y un random_state=42, entrenar el algoritmo llamado: knn_mm utilizando el método fit(x_train, y_train), hacer predicciones con los datos de prueba utilizando knn_mm.predict() y generar la matriz de confusión inicial como se aprecia en la Tabla 2 con el comando confusion_matrix() y el reporte de clasificación inicial como se observa en la Tabla 3 utilizando el comando classification_report.

En la matriz de confusión de la Tabla 3, se aprecia los valores: Verdadero Positivo (VP), Verdadero Negativo (VN), Falso Positivo (FP) y Falso Negativo (FN).

Tabla.2 Matriz de confusión inicial con n_neighbors=1

		Negativo	Positivo
Valores Reales	Negativo	VN= 900	FN=29
	Positivo	FN= 50	VP=3
		Predicción	

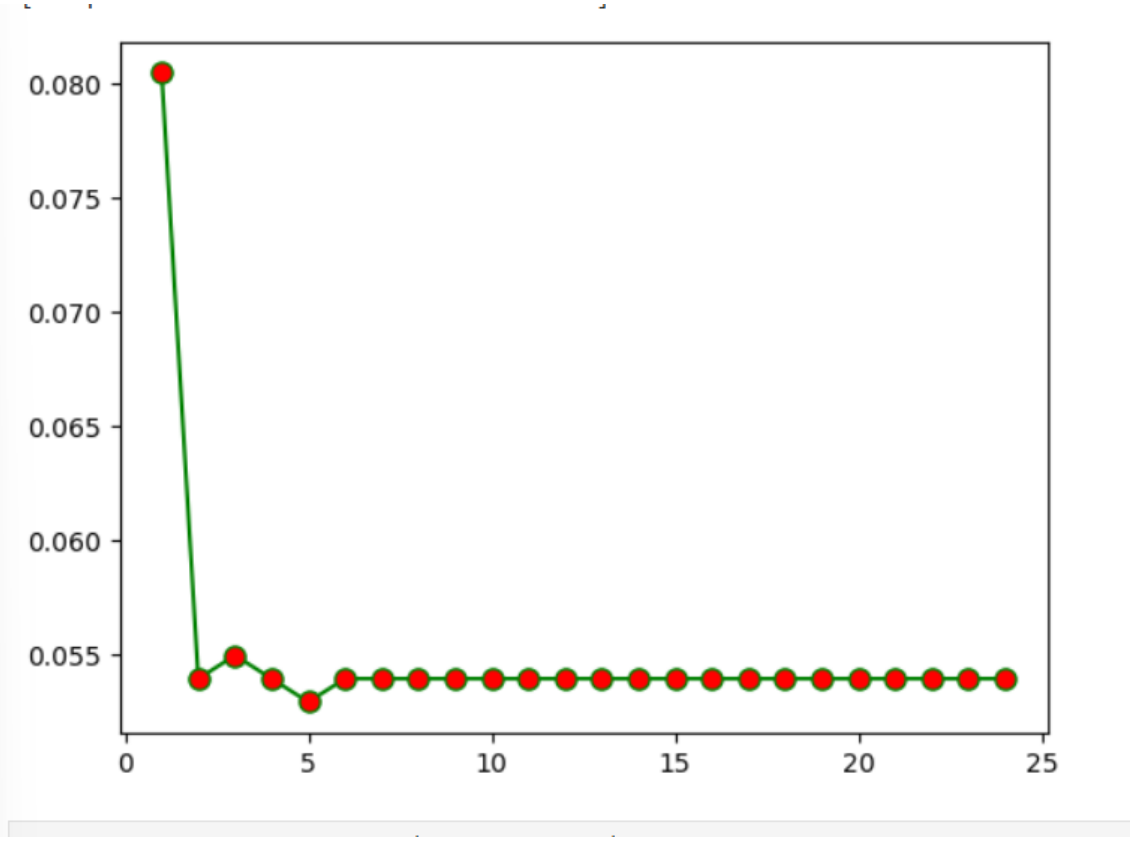
En el reporte de clasificación inicial con n_neighbors=1 de la Tabla 4 se observa una Precisión (Precision)=0.09%, sensibilidad (Recall)=0.06% y un Puntaje-F1(F1-score) = 0.07% para la clase de negativo, asimismo para la clase de positivo una Precisión (Precision)=95%, sensibilidad (Recall)=97% y un Puntaje-F1(F1-score) = 96% y finalmente se observa que el nivel de exactitud (Accuracy) del modelo es del 92%.

Tabla .3 Reporte de clasificación inicial con n_neighbors=1

	Presicion	Recall	F1-score	Support
Negativo	0.09	0.06	0.07	53
Positivo	0.95	0.97	0.96	929
Accuary			0.92	982
Macro AVG	0.52	0.51	0.51	982
Weighted	0.90	0.92	0.91	982

Selección de modelo

Para la selección de un modelo adecuado es necesario encontrar el mejor valor para n_neighbors, entonces, se logró diseñar un modelo repetitivo donde se muestra un gráfico con las diferentes tasas de error como se observa en la siguiente figura, en este caso, se logra ver que el mejor valor para n_neighbors=10



Entonces vuelve a generar un modelo knn_mm utilizando el método fit(x_train,y_train) y con el n_neighbors=10, y luego se debe generar la matriz de confusion de modelo optimo utilizand el comando: confusion_matrix() según se observa en la tabla 4.

		Negativo	Positivo
Valores Reales	Negativo	VN= 929	FN=0
	Positivo	FN= 53	VP=0
		Predicción	

Usando el comando classification_report de libraría ScikitLearn, se obtiene la Tabla 5. Reporte de clasificación óptimo; en esta tabla se logra observar una Precisión(Precision)=0%, sensibilidad(Recall)=0% y un Puntaje-F1(F1-score)= 0% para la clase de Negativo, asimismo se observa una Precisión(Precision)=95%, sensibilidad(Recall)=100% y un Puntaje-F1(F1-score)= 97% para la clase Positivo,

asimismo, se puede observar el nivel de exactitud(Accuracy) del no modelo knn_mm es del 95%.

	Presicion	Recall	F1-score	Support
Negativo	0.00	0.00	0.00	53
Positivo	0.95	1.00	0.97	929
Accuary			0.95	982
Macro AVG	0.47	0.50	0.49	982
Weighted	0.89	0.95	0.92	982

Bibliografía:

1. Dhiego Alves de Lacerda, Pedro Fechine Honorato, osé George Ferreira de Albuquerque. september 2024. 2024 [cited 2025 Apr 26]. Stroke: Symptom identification, FAST protocol and initial management. Available from: https://consensus.app/papers/stroke-symptom-identification-fast-protocol-and-initial-lacerda-honorato/cf3f15ee9e6953f6995b985ee377d4da/
2. Campbell BC V., De Silva DA, Macleod MR, Coutts SB, Schwamm LH, Davis SM, et al. Ischaemic stroke. Nat Rev Dis Primers. 2019 Oct 10;5(1):70.
3. Johnson CO, Nguyen M, Roth GA, Nichols E, Alam T, Abate D, et al. Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019 May;18(5):439–58.
4. Thayabaranathan T, Kim J, Cadilhac DA, Thrift AG, Donnan GA, Howard G, et al. Global stroke statistics 2022. International Journal of Stroke. 2022 Oct 19;17(9):946–56.
5. Bernabé-Ortiz A, Carrillo-Larco RM. Tasa de incidencia del accidente cerebrovascular en el Perú. Rev Peru Med Exp Salud Publica. 2021 Oct 11;38(3):399–405.
6. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJ (Buddy), Culebras A, et al. An Updated Definition of Stroke for the 21st Century. Stroke. 2013 Jul;44(7):2064–89.
7. Owolabi MO, Thrift AG, Mahal A, Ishida M, Martins S, Johnson WD, et al. Primary stroke prevention worldwide: translating evidence into action. Lancet Public Health. 2022 Jan;7(1):e74–85.
8. Wang K, Shou Q, Ma SJ, Liebeskind D, Qiao XJ, Saver J, et al. Deep Learning Detection of Penumbral Tissue on Arterial Spin Labeling in Stroke. Stroke. 2020 Feb;51(2):489–97.
9. Yedavalli VS, Tong E, Martin D, Yeom KW, Forkert ND. Artificial intelligence in stroke imaging: Current and future perspectives. Clin Imaging. 2021 Jan;69:246–54.

10. Pindado Carrasco S, Esteban Cornejo M, Morel Fernández S. Impacto de la escala NIHSS en la Unidad de Ictus del Hospital Universitario Ramón y Cajal: una herramienta para mejorar la calidad asistencial. Revista Científica de la Sociedad Española de Enfermería Neurológica. 2024 Jan;59:32–7.
11. Strambo D, Michel P, Nguyen TN, Abdalkader M, Qureshi MM, Strbian D, et al. Endovascular Versus Medical Therapy in Posterior Cerebral Artery Stroke: Role of Baseline NIHSS Score and Occlusion Site. Stroke. 2024 Jul;55(7):1787–97.
12. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJ (Buddy), Culebras A, et al. An Updated Definition of Stroke for the 21st Century. Stroke. 2013 Jul;44(7):2064–89.
13. Kuriakose D, Xiao Z. Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. Int J Mol Sci. 2020 Oct 15;21(20):7609.
14. Saceleanu VM, Toader C, Ples H, Covache-Busuioc RA, Costin HP, Bratu BG, et al. Integrative Approaches in Acute Ischemic Stroke: From Symptom Recognition to Future Innovations. Biomedicines. 2023 Sep 23;11(10):2617.
15. Lary DJ, Alavi AH, Gandomi AH, Walker AL. Machine learning in geosciences and remote sensing. Geoscience Frontiers. 2016 Jan;7(1):3–10.
16. Ray S. A Quick Review of Machine Learning Algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE; 2019. p. 35–9.