ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2023.1.1.720

Logistic regression: an example for infarct prediction

Henry Silva-Marchan, Mg[1], Gerardo Ortiz-Castro, Mg[1],
Oscar Jhan Marcos Peña-Cáceres, Mg[2], Manuel Alejandro More-More, Dr[3]
[1]Universidad Nacional de Tumbes, Perú, hsilvam@untumbes.edu.pe; gortizc@untumbes.edu.pe
[2]Universidad Cesar Vallejo, Piura, Perú, ojpenac@ucvvirtual.edu.pe
[3]Universidad Nacional de Piura, Piura, Perú, mmorem@unp.edu.pe

Abstract*– *Technological advances have allowed the development and availability of specialized tools for the use of historical data. In many cases, these tools are used for decision-making in multidisciplinary institutions that require support for the development of their activities, particularly in the health sector. The purpose of this study is to use a machine learning algorithm to predict heart attacks using demographic and family health data. The methodology focused on the extraction of open data from the Demographic and Family Health Survey (ENDES) applied in 2021 in Peru, characterization and execution of machine learning techniques using Orange Data Mining software. In this first approach, the results show that the logistic regression model has an accuracy of 0.99% on the prediction of heart attacks under the use of ENDES Peru 2021 data. For future studies, it is suggested to incorporate unstructured data such as text documents, sensor data and images to strengthen the reliability of the model.

Keyword- Machine learning, logistic regression, Infarction, ENDES.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Regresión logística: un ejemplo para la predicción de infartos

Henry Silva-Marchan, Mg[1], Gerardo Ortiz-Castro, Mg[1],
Oscar Jhan Marcos Peña-Cáceres, Mg[2], Manuel Alejandro More-More, Dr[3]
[1]Universidad Nacional de Tumbes, Perú, hsilvam@untumbes.edu.pe; gortizc@untumbes.edu.pe
[2]Universidad Cesar Vallejo, Piura, Perú, ojpenac@ucvvirtual.edu.pe
[3]Universidad Nacional de Piura, Piura, Perú, mmorem@unp.edu.pe

Resumen*– Los avances tecnológicos han permitido desarrollar y disponer de herramientas especializadas para el uso de datos históricos. En muchas de las veces empleados para la adopción de decisiones en instituciones de carácter multidisciplinario y que requieren un soporte para el desarrollo de sus actividades en lo particular el sector salud. El propósito de este estudio es hacer uso de un algoritmo de aprendizaje automático para predecir infartos mediante el empleo de datos demográficos y de salud familiar. La metodología se centró en la extracción de datos abiertos de la Encuesta Demográfica y de Salud Familiar (ENDES) aplicada en 2021 en Perú, caracterización y ejecución de técnicas de aprendizaje automático mediante el software Orange Data Mining. En este primer acercamiento los resultados reflejan que el modelo de regresión logística tiene una precisión del 0,99%, sobre la predicción de infartos bajo en el empleo de datos ENDES Perú 2021. Para estudios futuros, se sugiere incorporar datos no estructurados como, documentos de texto, datos de sensores e imágenes que fortalezcan la confiabilidad del modelo.

Palabra clave- Aprendizaje automático, regresión logística, Infartos, ENDES.

I. Introducción

En Machine Learning (ML), la regresión logística representa un tipo de aprendizaje supervisado, específicamente de clasificación. Se utiliza en aquellos casos que cuentan con dos clases (0/1), (si/no), (A, B), (true/false), pueden tener una representación gráfica:

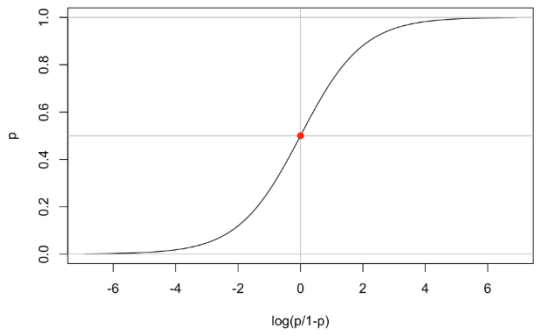

Fig. 1. Regresión lineal para predecir una variable numérica.

La regresión logística se dio origen alrededor de 1960 con la investigación de Cornfield, Gordon y Smith, luego en 1967 Walter y Duncan la aplicaron del modo como ahora se usa, es decir, para identificar la probabilidad de que ocurra un proceso o resultado en base a una serie de variables explicativas. Desde los inicios de los 80 su uso se incrementa debido a los avances acontecidos en el campo de las tecnologías de la información [1].

Los procedimientos de regresión de variable de resultados o dependiente, cualitativa alcanzan varias técnicas o formas de modelamiento que intentan explicar y pronosticar un atributo cualitativo basado en datos de otras variables conocidas, sean de tipo cuantitativo o cualitativo que se comporten como variables explicativas [2].

La regresión logística se usa también, para determinar si una o más variables interpretan una variable de resultados con características cualitativas.

Este hecho es muy común en medicina porque siempre tratamos de responder cuestiones que se formulan por la presencia o ausencia de ciertas propiedades no cuantificables, pero que representan la presencia o ausencia de un efecto de interés, como si un paciente en hospitalización fallece antes del alta, si se produce una readmisión y si el paciente desarrolla nefropatía diabética.

En este contexto es necesario indicar que los avances tecnológicos han dado cabida al uso de técnicas de aprendizaje automático, con la finalidad de proyectar, estimar y en muchos casos predecir valores futuros, en tal sentido, el presente trabajo tiene como propósito analizar los resultados del uso de la regresión logística en un conjunto de datos, para predecir qué tan propensa se encuentra una persona en Perú a un infarto, basándose en la encuesta demográfica y de salud familiar que el Instituto Nacional de Estadística e Informática (INEI) aplicó durante el año 2021.

La aparición del SARS-COV-2, que causó la COVID-19, en Perú, se reporta de manera oficial el 6 de marzo de 2020 y el 25 de marzo 2020 se promulga el decreto supremo 94, mediante el cual se dispusieron normas de aislamiento social que llevaron a un nuevo esquema de convivir socialmente. Estas medidas no solo desnudaron las limitaciones y carencias del sistema sanitario actual, si no también, evidenciaron el rol de la industria y la sociedad para enfrentar la COVID-19 [3].

Estas medidas de aislamiento influyeron en la población al no poder realizar sus actividades de manera normal, incidiendo en la reducción de realizar actividades físicas, al no poder desplazarse a sus centros laborales, lugares de esparcimiento, realizar deporte al aire libre o ejercitarse como lo venían realizando antes de las medidas de aislamiento, lo que como consecuencia perjudica la salud de la población.

La investigación tiene el objetivo determinar si la regresión logística permite pronosticar si una persona se encuentra propenso de tener un infarto.

II. Infarto en el Perú

La Organización Mundial de la Salud (OMS) reporta que una de las principales causa de mortalidad en el orbe es la enfermedad aterosclerótica y aproximadamente 30% de los fallecidos registrados por causa de cardiopatía isquémica con un efecto mucho mayor que el de trastornos infecciosos y el cáncer, estimándose que al 2030 la mortalidad aumentará a 36% [4].

La enfermedad cardiovascular es una dolencia no transmisibles con mayor prevalencia alrededor del orbe como origen de enfermedad y muerte; esta enfermedad amenaza para las próximas décadas en una pandemia con consecuencias devastadoras [5].

En todos los países se reporta las enfermedades cardiovasculares como una de las causas de mortalidad. Como los accidentes cerebrovasculares (ACV) que causan principalmente discapacidad y mortalidad [6].

Una de las causas más frecuentes es el infarto de miocardio con elevación del segmento ST (IMCEST). En Norteamérica, el IMCEST representa del 25 a 40% de ocurrencia de infarto de miocardio (3), con una mortalidad intrahospitalaria del 5 al 6% y del 7 al 18% al año del evento [4].

La pandemia ha permitido identificar que las personas mayores de 60 años son muy proclives a la ingesta de alcohol, lo que afecta significativamente su salud, más aún si ingiere fármacos junto con el alcohol debido a los males de salud que puede tener de forma crónica. Se suma a lo antes indicado que estos adultos suelen incrementar el riesgo de ingesta de tabaco, benzodiacepinas y opioids [7].

Ahmed et al. [8] reporta que los jóvenes tienden a compilar información de las redes sociales que está estrechamente relacionada con su consumo de sustancias adictivas, lo que los hace más propensos al estrés. Aparte, se ha conocido a otro segmento poblacional vulnerable al abuso de la ingesta de alcohol cuyo rango de edad se sitúa entre 21 a 40 años [9].

III. Regresión logística

Es uno de los algoritmos de mayor uso para clasificar eventos en la industria. Parecido al perceptrón y a Adeline, el modelo de regresión logística es también, en este caso, un modelo de clasificación binaria que puede ampliarse a la clasificación multiclase [10].

En regresión logística, la variable de resultados puede ser dicotómica, esto es, de dos categorías, polinómica nominal, enmarcada en tres o más categorías con un orden no natural, o polinómica ordinal, que se basa en tres o más categorías con un orden natural [11].

Desde una colección de datos de entrada, la salida, será discreta y no continua por el empleo del algoritmo la Regresión Logística [12]. La Regresión Logística es un algoritmo supervisado y se usa para la clasificación. En este caso se clasificó en valores de salida, donde el valor de 1 indica que la persona se encuentra propenso a tener un infarto y 0, estado de no encontrarse propenso. Constituye un método estadístico que propicia formular el modelamiento de una variable de resultado de naturaleza cualitativa, y define la posibilidad de asignar a una clase como una función de distribución logística[13].

Las aplicaciones de la regresión logística en las ciencias médicas y de la salud facilitan el análisis de los resultados en base a aspectos explicativos y predictivos, lográndose establecer la fortaleza de la asociación con el uso de odds ratio de los factores de riesgo con la variable resultado analizada de una forma independiente y así apreciar el valor predictivo de cada factor o bien del modelo en su totalidad [2].

Podemos definir a Y como una variable dependiente binaria que toma 2 valores probables etiquetados como 0 y 1.

Sean $X_1, \dots, X_k$ una serie de variables independientes observadas que van a explicar y/o predecir el valor de Y.

El objetivo es calcular:

$$P[Y = 1|X_1, \dots, X_k]$$
$$P[Y = 0|X_1, \dots, X_k] = 1 - P[Y = 1|X_1, \dots, X_k] \quad (1)$$

Se construye un modelo de la forma:

$$P[Y = 1|X_1, \dots, X_k] = p[X_1, \dots, X_k; \beta] \quad (2)$$

Donde $p(X_1, \dots, X_k; \beta)$ es una relación que recibe el nombre de función de enlace (función de probabilidad) cuyo valor depende de un vector de parámetros $\beta = (\beta_1, \dots, \beta_k)$.

El modelo regresivo logístico para una sola variable independiente se denomina regresión logística binaria simple, y se representa como:

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 \quad (3)$$

Donde P, constituye la probabilidad en que una persona exhiba o desarrolle la característica que interesa y X es la única variable explicativa.

En el caso que el modelo puede ser ampliado, añadiendo más variables independientes (continuas o categóricas). Lo que provocará entender mejor porque varía la respuesta entre un individuo y otro [14].

Se denomina regresión logística binaria múltiple, consideremos entonces la variable dicotómica, la cual puede tomar valores 0 o 1 dependiendo del estudio, y un conjunto de variables independientes $x_1,x_2,\ldots,x_k$ con probabilidad $P(Y=1|x)$.

El modelo logístico múltiple es:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (4)$$

Después de haber hecho hincapié en los casos de regresión logística binaria, y considerando el contexto de estudio se utilizará la regresión logística binaria múltiple.

IV. Extracción de los datos

Se utilizaron los datos que representa la encuesta ENDES de Perú, de año 2021. La indicada encuesta está conformada por 13 módulos, de los cuales fue necesario hacer uso de los datos que corresponden al módulo con código 1640 que corresponde a la encuesta de salud, debido a que brindan las características de los miembros del hogar, presión arterial, si le diagnosticaron hipertensión arterial, diabetes, sus hábitos alimenticios, consume de alcohol, cigarrillos. Además de ello fue necesario recurrir a la herramienta SPSS para obtener los datos y conocer la composición de cada una de las variables que se describen en la Tabla 1.

El origen de datos dispone de 34,115, índice de registros, donde el 80% se empleó para actividades de entrenamiento y el 20% para el conjunto de prueba.

TABLA I
Variables extraídas de la encuesta demográfica y de salud familiar

Variable	Módulo	Descripción
ID1	1640	Año
HHID		Identificación Cuestionario del Hogar
QS23		Años cumplidos
QS100		Algún profesional le ha medido la Presión Arterial
QS102		Le diagnosticaron Hipertensión Arterial o Presión Alta
QS104		Compraron medicamentos para controlar su Presión Alta
QS106		Tomo medicamentos tal cual indico le indico el médico
QS107		Le midieron el azúcar o glucosa en la sangre
QS109		Le diagnosticaron diabetes o azúcar alta
QS111		Ha comprado medicamentos para controlar la diabetes o azúcar alta
QS113		Tomó los medicamentos tal cual le indicó el médico
QS202		Fuma diariamente
QS206		Ha consumido alguna vez bebidas alcohólicas
QS208		En los últimos 12 meses ha consumido alguna bebida alcohólica
QS209		En los últimos 12 meses, tomó bebidas alcohólicas 12 o más días
QS210		En los últimos 30 días ha consumido bebidas alcohólicas
QS200		En los últimos 12 meses ha fumado cigarrillos
QS201		En los últimos 30 días ha fumado cigarrillos
TARGET		Persona propensa a infarto

Al tener una base de datos de gran capacidad, se utilizó para un primer análisis SPSS, para identificar la bondad del modelo, la correspondencia de las variables independientes con el target.

Obteniendo un primer resumen de procesamiento de casos que se describen en la Tabla 2. Estableciendo que el 100% de los registros fueron procesados, es decir se consideraron los 34,115 registros, los que figuran como casos seleccionados incluidos en el análisis, para los casos perdidos no se marcó ninguno por eso el valor 0 y en los casos no seleccionados nos reporta un valor de 0, teniendo hasta el momento ninguna observación en este primer procesamiento de nuestra data.

TABLA 2
RESUMEN DE PROCESAMIENTO DE CASOS

Casos sin ponderar		N	Porcentaje
Casos seleccionados	Incluido en el análisis	34115	100,0
	Casos perdidos	0	,0
	Total	34115	100,0
Casos no seleccionados		**0**	**,0**
Total		**34115**	**100,0**

En cuanto a la bondad del modelo, en el bloque 1 obtenido se cuenta con información que precisa que el modelo ayuda a explicar el evento como se describe en la Tabla 3.

TABLA 3
PRUEBAS ÓMNIBUS DE COEFICIENTES DE MODELO

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	15389,660	15	,000
	Bloque	15389,660	15	,000
	Modelo	15389,660	15	,000

De acuerdo a los resultados de las pruebas ómnibus de coeficientes del modelo la significación (Sig.) es inferior de 0,05 denota que el modelo apoya a interpretar el evento, indicando que, las variables independientes explican la variable dependiente.

Respecto a la bondad del modelo, especificado en la Tabla 4 y 5, el modelo es significativo, las variables independientes explican en 0,363 y 0,913 de la variable dependiente por lo que cataloga de modo adecuado a la variable dependiente en un 98,70%, este tanto por ciento establece el número de casos que el modelo puede pronosticar correctamente.

TABLA 4
RESUMEN DEL MODELO

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1905,518[a]	,363	,913

TABLA 5
TABLA DE CLASIFICACIÓN

Observado			Pronosticado		
			Persona propensa a infarto		Porcentaje correcto
			No	SI	
Paso 1	**Persona propensa a infarto**	No	31449	280	99,1
		SI	169	2217	92,9
	Porcentaje global				98,7

Los resultados mostrados en las tablas 4 y 5, precisan que es un modelo aceptable. Cuando el modelo clasifica correctamente más de 50% de los casos, se acepta el modelo, si no es así, simplemente se selecciona una nueva variable independiente o más.

En la relación de las variables explicativas con la variable resultado, se identifica si estas variables explican la variable dependiente, la dirección de la relación, la fortaleza de la relación, cuan más lejos de 1 sea más alta dicha relación.

TABLA 6
VARIABLES EN LA ECUACIÓN

Paso 1	B	Error estándar	Wald	gl	Sig.	Exp(B)
QS100	-3.803	0.220	299.814	1	0.000	0.022
QS102	10.698	0.343	970.380	1	0.000	44249.402
QS104	-9.729	0.840	134.179	1	0.000	0.000
QS106	-0.178	0.875	0.041	1	0.839	0.837
QS107	-1.154	0.158	52.981	1	0.000	0.316
QS109	11.213	0.392	817.380	1	0.000	74122.122
QS111	-9.575	1.057	82.077	1	0.000	0.000
QS113	-1.773	1.083	2.678	1	0.102	0.170
QS202	0.347	0.458	0.576	1	0.448	1.415
QS206	-0.170	0.368	0.213	1	0.644	0.844
QS208	-2.443	0.222	121.526	1	0.000	0.087
QS209	11.406	0.326	1223.384	1	0.000	89898.967
QS210	6.115	0.257	565.914	1	0.000	452.657
QS200	-0.445	0.200	4.942	1	0.026	0.641
QS201	0.379	0.241	2.483	1	0.115	1.461
Constante	-8.782	0.431	415.959	1	0.000	0.000

La relación de las variables explicativas con la variable precisa que las variables QS100, QS102, QS104, QS107, QS109, QS111, QS208, QS209, QS210, QS200 explican la predicción del infarto en este context estudiado, al tener una significación menor a 0,05. Mientras más se incrementan estas variables más probabilidad de que ocurra el infarto.

De todas las variables independientes aquellas que tuvieron una mayor fortaleza para explicar el evento de que ocurra el infarto son que se le haya diagnosticado hipertensión, diagnosticado diabetes o azúcar alta, que, en los últimos 12 meses, ingirió alcohol 12 o más días. Sus valores Exp(b), exponencial de b, más se aleja de 1.

V. MACHINE LEARNING: REGRESIÓN LOGÍSTICA

La expresión machine learning fue enunciado en 1959 por A.L. Samuel [15], y su finalidad es evolucionar algoritmos que apoyen a tomar decisiones y aprender de sus resultados, y con capacidad de aprender a hacer algo sin haberse diseñado expresamente para tal propósito.

Para el procesamiento de los datos en machine learning, se utilizó Orange, como aplicación para el proceso de nuestra data y realizaremos el procesamiento de los datos en un 80% datos de entrenamiento y un 20% de prueba.

Orange es un completo marco basado en componentes para el aprendizaje automático y la minería de datos. Está dirigido tanto a usuarios experimentados como a investigadores en aprendizaje automático que deseen escribir scripts en Python para crear prototipos de nuevos algoritmos reutilizando la mayor parte posible del código, y para aquellos que acaban de entrar en el aprendizaje automático, pueden emplear el entorno de programación visual potente y fácil de utilizar [16].

Construir un modelo de aprendizaje automático no se limita solo a usar un algoritmo de aprendizaje en una base de datos, sino que es una secuencia de pasos a seguir, como se muestran en la Fig 1.

Existe una diversidad de algoritmos de clasificación y regresión [17], desde modelos lineales comunes como la regresión logística o el análisis discriminante lineal, a otras más recientes como algoritmos de tipo agregación o bagging, como random forest y los de boosting o empaquetado. Igualmente, es posible usar técnicas con submuestras y sobremuestreo, como usualmente se da en eventos de salud, hay categorías con mayores casos que otras [18].

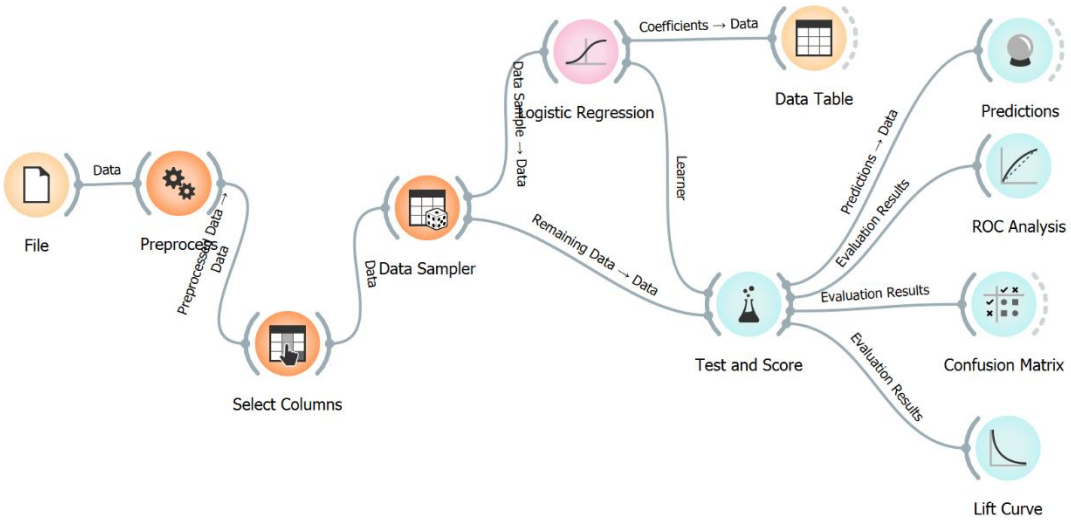

Fig 2. Procesamiento del modelo

La capacidad en un algoritmo, de obtener un buen ajuste depende de las cualidades de los datos (cantidad de variables, linealidad, distribución normal, valores perdidos, variables continuas o categóricas, etc.) [18].

Predicted

Actual	No	Si	Σ
No	6305	29	6334
Si	30	459	489
Σ	6335	488	6823

Fig 3. Matriz de confusión

VI. Discusión de los resultados

El modelo apoya a interpretar el suceso, indicando que, las variables independientes explican la variable dependiente y es significativo, las variables independientes explican en 0,363 y 0,913 de la variable dependiente por lo que cataloga de modo adecuado a la variable dependiente en un 98,70%.

La matriz de confusión es uno de los medios más empleados y útiles para evaluar la calidad de los modelos de clasificación que se basan en el aprendizaje automático. En particular, cuando una clase se confunde con otra, se puede mostrar claramente, lo que nos permite manejar diferentes tipos de errores por separado [19].

La precisión del modelo es esencialmente el cociente de la cantidad total de predicciones correctas entre la cantidad total de predicciones, se refiere a qué tan cerca está el resultado de la medición del valor real, en términos estadísticos [20].

Los resultados obtenidos en la matriz de confusión (Fig. 3) indican que la cantidad de Verdaderos Positivos (VP) que el modelo predice como positivos, es 6305; la cantidad de resultados negativos, Falsos Positivos (FP) que el modelo predice como positivos, es igual a 29; el número de resultados positivos que el modelo predice como negativos, Falsos Negativos (FN) es igual a 30; el número de ejemplos negativos, Verdadero Negativos (VN) que el modelo predice como negativos, es 459.

En los resultados, el FP indica que el individuo no está propenso a padecer de un infarto, pero el algoritmo ha diagnosticado que sí está propenso. En los resultados de FN el individuo sí está propenso a padecer un infarto, pero el algoritmo predice que no, este error del algoritmo se traduce en una falta de exploración anticipada de la enfermedad.

Con los resultados de la matriz de confusión podemos calcular una medida de precisión, midiendo la calidad del modelo de machine learning en tareas de clasificación, se calcula VP/(VP+FP). El valor obtenido es 0,94 lo que indica que se encuentra a un 94% de obtener el valor verdadero.

La matriz de confusión nos permite establecer la exactitud (accuracy) para medir la tasa de casos que el modelo ha clasificado correctamente, se calcula (VP+VN)/(VP+VN+FP+FN), el valor obtenido es 0,99, lo que indica que el modelo acierta el 99% de veces.

VII. Conclusiones

El software Orange Data Mining, es una de las alternativas vigentes y modernas para desarrollar estudios que pronostiquen, describan y precisen patrones de conducta sobre el estado de salud del ser humano.

A pesar de las desventajas que puede tener el aprendizaje automático, los aportes, son significativos. En este primer acercamiento el modelo de regresión logística tuvo un desempeño de precisión sobre el 99% en la predicción de infartos bajo el contexto de los datos ENDES Perú 2021.

Se precisa que, el modelo no incluye totalmente información médica de pacientes y otros determinantes sociales que pueden conducir a enfermedades cardíacas, como el estado de tabaquismo, nivel de actividad física, peso entre otras. Además de ello, la confiabilidad del estudio puede fortalecerse, integrando imágenes que se encuentren relacionadas a placas o ecográficas de los pacientes de acuerdo a su historial médico.

Por otro lado, los sensores que hoy se integran en dispositivos como Smart Watch, son de gran utilidad, debido que es posible conocer el ritmo de desplazamiento, monitoreo de sueño y en algunos de los casos el oxígeno en sangre.

Debido a la gran cantidad de datos relacionados con las enfermedades del corazón, se recomienda a priori realizar un estudio mixto con técnicas de aprendizaje automático y big data, escenario que podría traer algunos beneficios en el campo de la salud.

Referencias

[1] E. Domínguez Alonso and D. A. Padilla, “Regresion Logística. Un ejemplo de su uso en Endocrinología,” *Rev. Cuba. Endocrinol.*, vol. 12, no. 1, 2001.

[2] J. C. Fiuza Pérez, Mª Dolores, Rodríguez Pérez, “La regresión logística: una herramienta versátil,” *Nefrología*, vol. 20, 2000, [Online]. Available: https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-articulo-X0211699500035664.

[3] I. B. Barreto, R. M. S. Sánchez, and H. A. S. Marchan, “Consecuencias económicas y sociales de la inamovilidad humana bajo Covid – 19 caso de estudio Perú,” *Lect. Econ.*, no. 94, 2021, doi: 10.17533/UDEA.LE.N94A344397.

[4] M. Chacón-Diaz *et al.*, “Tratamiento del infarto agudo de miocardio en el Perú y su relación con eventos adversos intrahospitalarios: Resultados del Segundo Registro Peruano de Infarto de Miocardio con elevación del segmento ST (PERSTEMI-II),” *Arch. Peru. Cardiol. y Cirugía Cardiovasc.*, vol. 2, no. 2, 2021, doi: 10.47487/apcyccv.v2i2.132.

[5] J. L. Barrios Morocho and J. Valle Bayona, “Riesgo de infarto de miocardio en pacientes críticos mayores de 65 años,” *An. la Fac. Med.*, vol. 78, no. 2, 2017, doi: 10.15381/anales.v78i2.13187.

[6] A. Bernabé-Ortiz and R. M. Carrillo-Larco, “Tasa de incidencia del accidente cerebrovascular en el Perú,” *Rev. Peru. Med. Exp. Salud Publica*, vol. 38, no. 3, 2021, doi: 10.17843/rpmesp.2021.383.7804.

[7] D. D. Satre, M. E. Hirschtritt, M. J. Silverberg, and S. A. Sterling, “Addressing Problems With Alcohol and Other Substances Among Older Adults During the COVID-19 Pandemic,” *American Journal of Geriatric Psychiatry*, vol. 28, no. 7. 2020, doi: 10.1016/j.jagp.2020.04.012.

[8] M. Z. Ahmed, O. Ahmed, Z. Aibao, S. Hanbin, L. Siyu, and A. Ahmad, “Epidemic of COVID-19 in China and associated Psychological Problems,” *Asian J. Psychiatr.*, vol. 51, 2020, doi: 10.1016/j.ajp.2020.102092.

[9] N. A. Armendariz, “‘COVID 19 y su Impacto en el Consumo de Drogas: Revisión Sistemática,’” *Rev. Investig. Científica en Psicol. Órgano Of. Comun.*

Científica Monográfico, vol. 17, 2020.

[10] S. Raschka and V. Mirjalili, "Un recorrido por los clasificadores de aprendizaje automático con scikit-learn," in *Python Machine Learning*, Segunda., España, 2019, pp. 73–127.

[11] A. F. Godoy Viera, "Técnicas de aprendizaje de máquina utilizadas para la minería de texto," *Investig. Bibl.*, vol. 31, no. 71, pp. 103–126, Jan. 2017, doi: 10.22201/IIBI.0187358XP.2017.71.57812.

[12] J. I. Bagnato, "Regresión Logística," in *APRENDE MACHINE LEARNING en Español - Teoría + Práctica Python*, España, 2020, p. 43.

[13] B. Sánchez Martínez, V. Vega Falcón, and N. Gómez Martínez, "Predicción de la diabetes mellitus tipo 2 en pacientes adultos mediante regresión logística binaria.," *Dilemas Contemp. Educ. Política y Valores*, May 2021, doi: 10.46377/dilemas.v8i3.2675.

[14] Juan José Vindell, "Regresión Logística en Salud Pública," 2021. https://rstudio-pubs-static.s3.amazonaws.com/794646_8e890af7d90d495c83e55ed0344527ba.html (accessed Jan. 25, 2023).

[15] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 44, no. 1–2, 2000, doi: 10.1147/rd.441.0206.

[16] J. Demšar, B. Zupan, G. Leban, and T. Curk, "Orange: From experimental machine learning to interactive data mining," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3202, 2004, doi: 10.1007/978-3-540-30116-5_58.

[17] C. Robert, "Machine Learning, a Probabilistic Perspective ," *CHANCE*, vol. 27, no. 2, 2014, doi: 10.1080/09332480.2014.914768.

[18] P. I. Dorado-Díaz, J. Sampedro-Gómez, V. Vicente-Palacios, and P. L. Sánchez, "Aplicaciones de la inteligencia artificial en cardiología: el futuro ya está aquí," *Rev. Española Cardiol.*, vol. 72, no. 12, 2019, doi: 10.1016/j.recesp.2019.05.016.

[19] H. A. Silva Marchan, O. J. M. Peña Cáceres, D. M. Ricalde Moran, T. Samaniego-Cobo, and C. M. Perez-Espinoza, "A Machine Learning Study About the Vulnerability Level of Poverty in Perú," in *Technologies and Innovation*, 2022, pp. 3–14.

[20] J. I. Bagnato, "Métricas y Confusion Matrix," in *APRENDE MACHINE LEARNING en Español - Teoría + Práctica Python*, 2020, pp. 79–82.