

Tipo de artículo: Artículo original

Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo

Prediction of the main diseases that affect health in Ecuador based on risk factors

Shabely Avellán Valdés^{1*} , <https://orcid.org/0000-0002-3078-4120>

Cristopher Agustín Holguín Intriago² , <https://orcid.org/0000-0003-1013-2237>

Marely del Rosario Cruz Felipe³ , <https://orcid.org/0000-0003-1937-1568>

¹ Departamento Tecnología de la Información y Comunicación, Facultad de Ciencias Informáticas, Universidad Técnica de Manabí. savellan1826@utm.edu.ec

² Departamento Tecnología de la Información y Comunicación, Facultad de Ciencias Informáticas, Universidad Técnica de Manabí. cholguin5277@utm.edu.ec

³ Departamento Tecnología de la Información y Comunicación, Facultad de Ciencias Informáticas, Universidad Técnica de Manabí. marely.cruz@utm.edu.ec

* Autor para correspondencia: savellan1826@utm.edu.ec

Resumen

Las principales enfermedades que afectan a los ecuatorianos son: las enfermedades cardiovasculares, la diabetes mellitus, las enfermedades cerebrovasculares y la hipertensión arterial, es por esto que el control del crecimiento de las mismas se ha vuelto un problema tanto de salud como social, que es necesario afrontar, ya que, además de las campañas sobre el cuidado para su prevención, no existe una herramienta tecnológica que permita identificar y detectar, de forma temprana y eficiente, estas enfermedades en personas independientemente de sus edades. En esta investigación se realizó un análisis de diferentes modelos de machine learning capaces de predecir el riesgo a padecer estas 4 enfermedades en el Ecuador. Se implementaron las metodologías Crisp-DM (Cross – Industry Standard Process for Data Mining) y el proceso KDD (Knowledge Discovery Databases). Los resultados obtenidos mostraron que el modelo Random Forest es el que posee mejor rendimiento para las enfermedades cerebrovasculares, con un valor F1_Score de 0,927; el modelo de Regresión Logística tiene un mayor desempeño en la diabetes mellitus, con un F1_Score de 0,781 y el modelo Bayes Naives en las enfermedades cardiovasculares, con un F1_Score de 0,836. El índice de Shapley fue usado para explicar los factores de riesgo que más influyeron en la aparición de las mismas.

Palabras clave: enfermedades; salud en Ecuador; machine learning; Crisp-DM; Random Forest

Abstract

The main diseases that affect Ecuadorians are: cardiovascular diseases, diabetes mellitus, cerebrovascular diseases and high blood pressure, which is why the control of their growth has become both a health and social problem, which is necessary to face, since, in addition to campaigns on care for its prevention, there is no technological tool that allows early and efficient identification and detection of these diseases in people regardless of their ages. In this research, an analysis of different machine learning models capable of predicting the risk of suffering from these 4 diseases in Ecuador was carried out. The Crisp-DM (Cross-Industry Standard Process for Data Mining) methodologies and the KDD (Knowledge Discovery Databases) process were implemented. The results obtained showed that the Random Forest model is the one with the best performance for cerebrovascular diseases, with an F1_Score value of 0.927; the Logistic Regression model has a better performance in diabetes mellitus, with an F1_Score of 0.781, and the Bayes Naives model in cardiovascular diseases, with an F1_Score of 0.836. The Shapley index was used to explain the risk factors that most influenced their appearance.



Esta obra está bajo una licencia *Creative Commons* de tipo **Atribución 4.0 Internacional**
(CC BY 4.0)

Keywords: *diseases; health in Ecuador; machine learning; Crisp-DM; Random Forest*

Recibido: 15/05/2022

Aceptado: 29/07/2022

En línea: 01/08/2022

Introducción

Las enfermedades crónicas no transmisibles (ENT) son la principal causa de muerte y discapacidad en todo el mundo. Estas incluyen el cáncer, las enfermedades cardiovasculares (ECV), la diabetes y las enfermedades pulmonares crónicas. Muchas de ellas pueden prevenirse reduciendo los factores de riesgo más comunes, como el consumo de tabaco, de alcohol, de alimentos poco saludables y la inactividad. Las ENT matan alrededor de 41 millones de personas cada año, lo que equivale al 71% de las muertes que se producen en el mundo. Cada año mueren 15 millones de personas de entre 30 y 69 años de edad y más del 85% de estas muertes "prematargas" ocurren en países de ingresos bajos y medianos.(Organización Mundial de la Salud OMS, 2021b).

Las ECV representan la primera causa de muerte en el mundo según lo evidencia la Organización Mundial de la Salud (OMS). Estas enfermedades se asocian al estilo de vida de las personas, estrés personal y laboral, falta de actividad física, así como con aspectos de nutrición. Este último es relevante ya que la alimentación está asociada a las ECV y a sus niveles de mortalidad (WHO Global Health Estimates, 2020). Estas constituyen la mayoría de las muertes por ENT, con 17,9 millones cada año (Organización Mundial de la Salud OMS, 2021b).

La hipertensión arterial (HTA) es el principal factor de riesgo para sufrir una ECV, padeciéndola una de cada tres adultos (Laluisa, 2014). En 2019, la prevalencia global estandarizada por edad de HTA en adultos de 30 a 79 años fue del 32% en mujeres y 34% en hombres (Zhou et al., 2021).

Por otro lado, la diabetes mellitus (DM) posee a nivel mundial un índice de muerte de 1,6 millones cada año (Organización Mundial de la Salud OMS, 2021b). Esta es considerada como una de las principales causas de ceguera, insuficiencia renal, infarto de miocardio, accidente cerebrovascular y amputación de miembros inferiores. Se estima que 62 millones de personas en las Américas viven DM tipo2. Este número se ha triplicado en la región desde 1980 y podrá alcanzar los 109 millones para el 2040 (Organización Mundial de la Salud OMS, 2021a).

Según el Global Burden of Disease publicado en el 2017, a nivel mundial, la enfermedad cerebrovascular (EC) fue la tercera causa principal de muerte y la segunda causa principal de discapacidad, afectando principalmente a adultos de mediana edad y ancianos. Se estima que hubo 5,5 millones de muertes por EC en el año anterior (Naghavi et al., 2017).

En el Ecuador las ECV ocupan el primer lugar de muerte, el 23,17% de las muertes por EC son la tercera causa de muerte en la población general, la segunda causa en mujeres y la cuarta en hombres (Solange Núñez-González et al., 2018). En cuanto a la DM, se evidencia un incremento significativo de la mortalidad, debido al aumento de sus



complicaciones, con un total de 4895 defunciones en el año 2017, 2289 hombres y 2606 mujeres; donde esta enfermedad representa el 90,3% del total, lo cual podría incrementar debido a los factores de riesgo (Calahorrano & Fernández, 2018).

En la última década la incidencia de HTA en el país ha ido en ascenso, llegando a cifras de 256 a 1084 casos por cada 100 mil habitantes; asimismo, en la ciudad de Cuenca se observó una prevalencia de HTA del 25,8 % (Navarro et al., 2020).

Estos datos generan preocupación a la comunidad médica ecuatoriana, por lo que resulta primordial abordar y entender este problema desde la perspectiva de la ciencia de datos para poder desarrollar modelos que permitan identificar la aparición temprana de las principales enfermedades que afectan la salud de los ecuatorianos, así como determinar cuáles variables afectan mayormente el desarrollo de la enfermedad.

Para la creación de modelos estadísticos se utilizan diferentes técnicas, entre las que se destacan los estudios transversales descriptivos, redes bayesianas, random forest, modelo lineal, redes bayesianas, el algoritmo Auto Prognosis. (Macías Sánchez & Cevallos Tumbaco, 2019)(Hasan et al., 2020)(Mujumdar & Vaidehi, 2019)(Sisodia & Sisodia, 2018)(Weng et al., 2017).

En la literatura consultada se puede inferir que no existen muchos estudios en el país que propongan el uso de variables que combinen resultados de diferentes exámenes, pruebas de rutina y cuestionarios directos que permitan encontrar la combinación de variables de mayor peso entre todas estas técnicas. En una investigación realizada por (Cárdenas-Ruperti et al., 2020) se destaca el uso de algoritmos de machine learning para la detección de factores de riesgo de la DM.

En esta investigación se traza como objetivo hacer una predicción de las principales enfermedades que afectan la salud en Ecuador, a partir de factores de riesgo, que sirva de insumo médico para el personal de salud, con el uso de modelos de machine learning, para así, dar soporte a la toma de decisiones médicas. Logrando también una detección rápida y/o temprana de este tipo de estas enfermedades que han significado a lo largo de la historia una inseguridad tan grande para el bienestar de la población.

Materiales y métodos

Se implementan las metodologías Crisp-DM (Cross – Industry Standard Process for Data Mining) debido a su orden y especificidad a la asignación de tareas y el manejo de fases y el proceso KDD (Knowledge Discovery Databases), ya que se adapta mejor al análisis de minería de datos lo que permitirá la caracterización de los pacientes. Al no ser de secuencia rígida estas metodologías permiten que se pueda volver atrás en las fases para la mejora de procesos y al ser



cíclica se pueden integrar los nuevos resultados de experimentación con la data. Esta comprende 4 fases: análisis del contexto, recopilación de datos, modelado y evaluación.

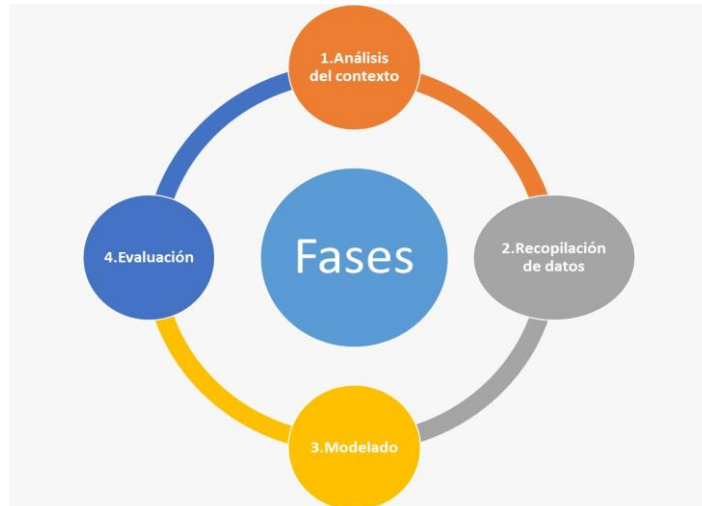


Figura 1. Esquema del ciclo de las metodologías Crisp-DM y proceso KDD

Estas fases se detallan a continuación.

Fase 1 Análisis del contexto

Se utilizó una base de datos (BD) proveniente del Instituto Nacional de Estadísticas y Censos (INEC). Con 5235 registros y 847 atributos, esta BD fue realizada en 15 provincias del país continental, de acuerdo a las Regiones Costa y Sierra, área urbana y rural, Quito y Guayaquil a personas mayores de 60 años (INEC (Instituto Nacional de Estadísticas y Censos), n.d.).

Fase 2 Recopilación de datos

De la data presente en la base de datos se realizó una preselección por un especialista en el área de la salud, para determinar cuáles eran considerados factores de riesgo, quedando así 340 columnas, las cuales fueron verificadas con una exploración de los datos para que no existieron datos incompletos o faltantes. La selección de las variables a usar fue llevada a cabo con el uso de la herramienta Collabratory de Google Researche, permitiéndonos así ejecutar código de Python; esta herramienta es muy útil para tareas de aprendizaje automático y análisis de datos.

Con Python se procedió con la limpieza de los datos; eliminando registros que contenían valores nulos, valores incompletos, valores que excedieron el rango normal, variables que no brindan información y en algunos casos fue necesario cambiar la etiqueta de las columnas para comprender qué valores estaban involucrados. Posteriormente, se



continuó con la selección de los atributos que aportaron la mayor cantidad de información relevante sobre el tipo de datos y los posibles valores.

Fase 3 Modelado

Para el modelado de los datos fue preciso utilizar la librería SKlearn, ya que la misma permite hacer análisis predictivo, así como en el preprocesamiento, la reducción de dimensionalidad (selección de parámetros), la clasificación, la regresión, la agrupación y la selección de modelos, ayudando a separar la data en conjuntos de 25% prueba y 75% entrenamiento. Se seleccionaron los siguientes algoritmos para el modelado de los datos: Bayes Nayves, Regresión logística y Random Forest, ya que son los más destacados de acuerdo a la literatura para la realización de este tipo de estudios.

Fase 4 Evaluación

En esta investigación se desarrolló modelos de evaluación de riesgos para las principales enfermedades que afectan la salud en el Ecuador, utilizando los 3 algoritmos de Machine Learning ya mencionados en la fase de Modelado. Posteriormente se ingresan un número creciente de variables, que son determinadas automáticamente por el propio algoritmo. Todos los algoritmos fueron entrenados y validados 10 veces durante cada proceso de iteración, que se repitió 100 veces, el rendimiento de los modelos fue calculado con las muestras de prueba, todos los parámetros de los modelos fueron determinados mediante validación cruzada más de 10 veces y una exploración en todos los datos de entrenamiento.

Resultados y discusión

En la determinación de los factores de riesgo de las 4 enfermedades que trata el estudio, ECV, EC, DM y HTA, se obtuvieron los siguientes resultados:

Mediante el uso de las herramientas de Sklearn se realizó la división de los datos en entrenamiento y prueba, usando una partición de 75% para entrenamiento y 25% para prueba. Preparándolos para realizar las predicciones con los algoritmos seleccionados. Los resultados mostrados en la tabla 1 se obtuvieron usando PyCaret statsmodels.



Tabla1. Configuración y división de la data

	Description	Value
2	Target Type	Binary
3	Original Data	5235, 17
4	Numeric Features	2
5	Categorical Features	14
6	Transformed Train Set	3926, 55
7	Transformed Test Set	1309, 55
8	Shuffle Train-Test	True
9	Stratify Train-Test	False

La tabla 1 muestra el resumen del dataset utilizado para realizar el procesamiento de datos final, donde se puede observar, el número de variables a utilizar, en este caso 17; la variable dependiente, que en este caso de estudio son 4 enfermedades, las cuales son: ECV, DM, EC, HTA. Dejando como variables independientes 16 datos, entre los que se tienen 2 numéricos, que son calculados usando los datos de cada paciente y 14 categóricos.

Para el proceso de limpieza de datos se realizó una revisión de la data general, la cual contaba con 847 columnas, que tenían presente preguntas generales sobre el diagnóstico de los pacientes encuestados. Las mismas comprenden preguntas que fueron eliminadas, por ejemplo: si el paciente presenta enfermedades del hígado, corazón, cerebrales, problemas de nutrición, entre otras.

Se procedió a eliminar los datos que no existían. Para ello se descartaron todas aquellas columnas que presentan una importante presencia de valores nulos. Luego se agregaron las variables pertinentes a los hábitos alimenticios, como pueden ser el consumo de carnes, frutas, verduras, alcohol, para determinar su importancia en el algoritmo. Estas últimas son consideradas factores de riesgo, ya que fueron indicadas por los especialistas médicos.

Estudios como el de (Zhang et al., 2020), hace un análisis de interacciones complejas entre variables masivas, que estudia la importancia de las variables, en el mismo se estima utilizar el enfoque de explicación aditiva de Shapley. Este enfoque unificado permite explicar el resultado de cualquier modelo de machine learning, ya que tiene en cuenta no sólo la viabilidad sino también el valor de las variables.

Se decidió aplicar el índice de Shapley para explicar los factores de riesgo que más influyen en cada una de las enfermedades, y se obtuvieron los resultados mostrados en los siguientes gráficos. Los colores mostrados en la figura



2 describen si la variable posee un valor alto o positivo, representado por el color rojo, o un valor bajo o negativo, representado por el color azul.

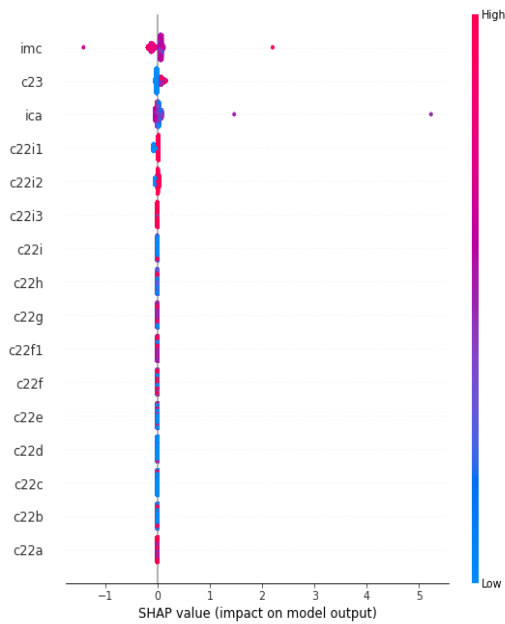


Figura 2 a. Variables de mayor importancia para las EC



Figura 2 b. Variables de mayor importancia para las DM

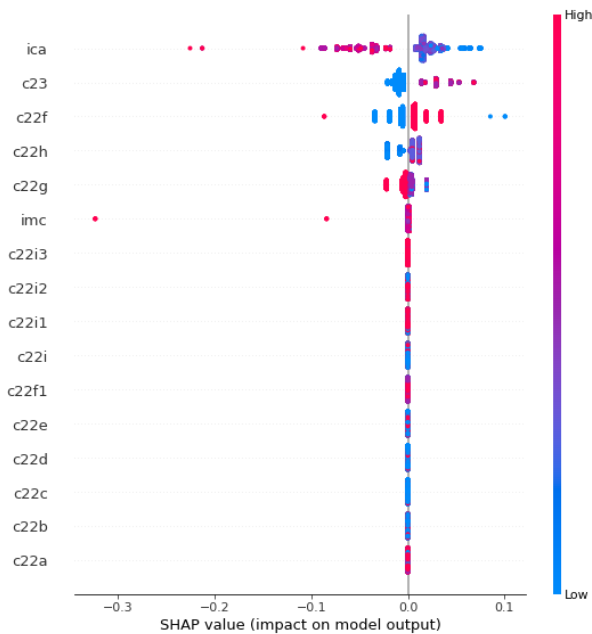


Figura 2 c. Variables de mayor importancia para las ECV

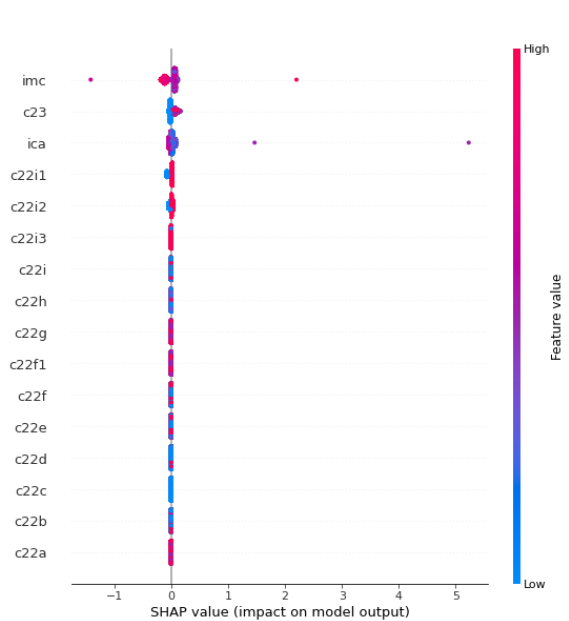


Figura 2 d. Variables de mayor importancia para la HTA



En la figura 2 a, b, c, d se muestran las variables que mayor impacto tienen en cada una de las enfermedades consideradas como variables dependientes en este estudio, estas son: c22a (Cuántas comidas completas come al día), c22b (Consume leche, queso o yogurt al menos 1 vez al día), c22c (Come huevos, fréjol, lenteja, al menos 1 vez por semana), c22d (Come carne, pescado, al menos 3 veces por semana), c22e (Consume frutas o verduras al menos 2 veces al día), c22f (Come menos que antes por problemas digestivos), c22f1 (En últimos 3 meses usted come:), c22g (Cuántos vasos de líquidos consume diariamente), c22h (En último año ha disminuido de peso sin dieta), c22i (Se considera bien nutrido), c22i1 (Cómo es su movilidad), c22i2 (En últimos 3 meses ha tenido alguna enfermedad), c22i3 (Actualmente presenta úlceras y/o escaras), c23 (En últimos 3 meses en promedio días por semana que toma alcohol). Por lo que se puede concluir a partir de estos datos que estas son las variables que mayor importancia tienen para las ECV, DM, EC y HTA.

Además de las variables seleccionadas para incluir en la data a emplear, a partir del análisis de los especialistas médicos, se realiza un control de los pacientes divididos entre aquellos que presentan cada una de las enfermedades del estudio, por género y seccionados por su índice de masa corporal (imc), para determinar si son necesarias a incluir otras variables. Este análisis se muestra en la figura 3.

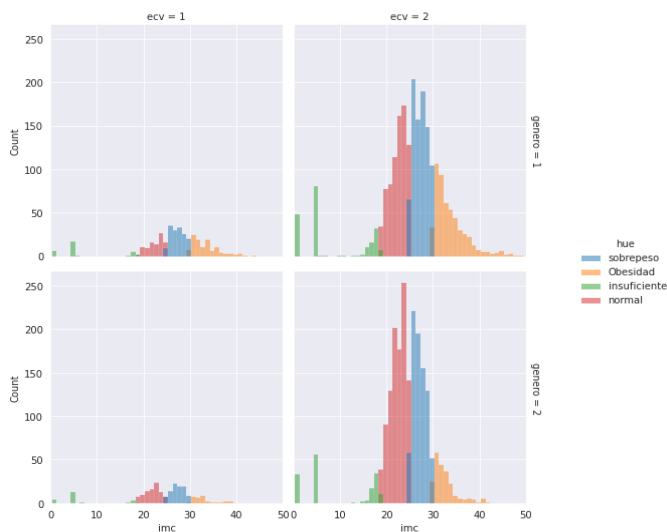


Figura 3 a. Representación del imc dividido entre los pacientes con ECV y por género

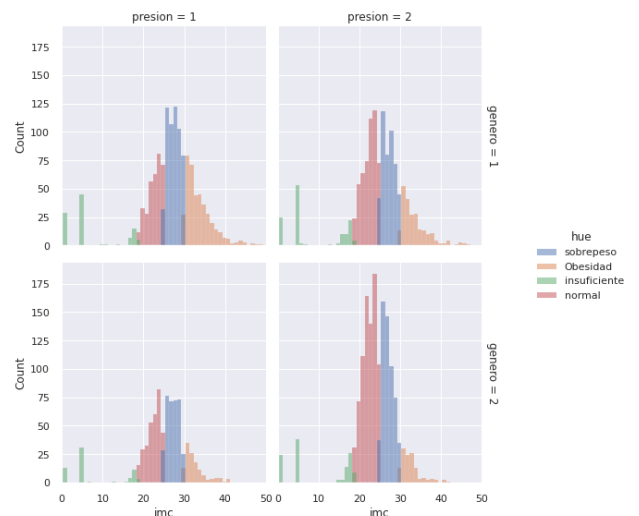


Figura 3 b. Representación del imc dividido entre los pacientes con HTA y por género



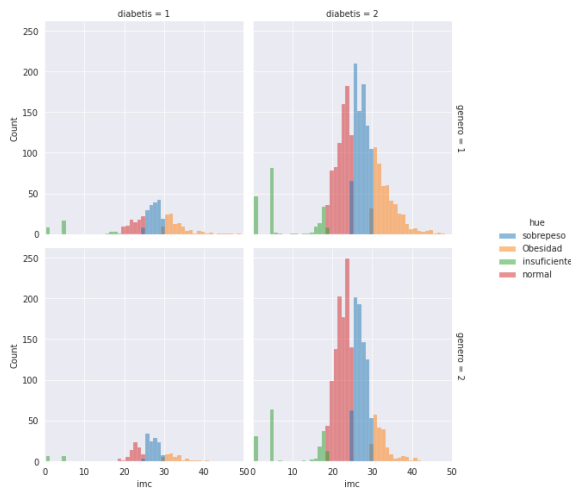


Figura 3 c. Representación del imc dividido entre los pacientes con DM y por género

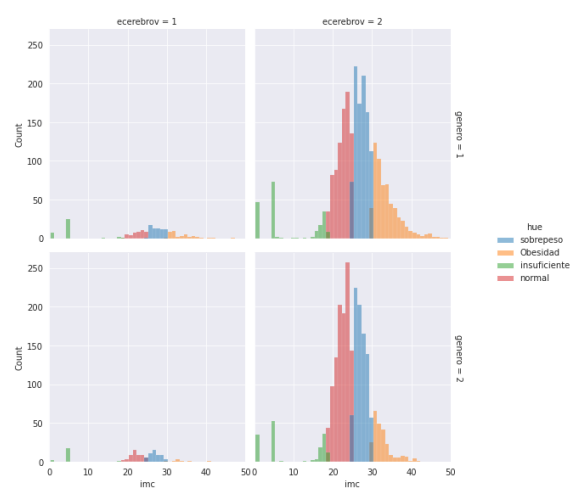


Figura 3 d. Representación del imc dividido entre los pacientes con EC y por género

Al revisar como el imc interacciona con las diferentes enfermedades, se puede denotar, que por sí sola no es una variable que pueda determinar totalmente estas enfermedades, esto tiene varias explicaciones. El imc es una métrica que siempre tiene que ser acompañada por otra de control que diferencie el peso magro con el peso de grasa. La métrica más exacta para esto suele ser el índice de grasa corporal, más es una variable que necesita de exámenes especiales para un resultado exacto, para buena suerte del estudio, contamos con las métricas necesarias para utilizar el índice de cintura altura (ica), la cual es útil en la diferenciación del peso magro y el peso en grasa.

Aun cuando esta métrica no es un punto de inflexión en la capacidad predictiva de los modelos, las figuras 3 a, b, c, d nos presentan varios datos de interés. Como principal factor tenemos la tendencia que muestran los grupos de pacientes con las enfermedades a mostrar un mayor porcentaje con sobrepeso y obesidad, que el grupo que no presenta enfermedades. Otro hallazgo que fascina al ver estas gráficas, es la similitud de cantidad de pacientes, con y sin HTA, la figura 3 c muestra cómo sigue la tendencia de mayor proporción de pacientes con sobrepeso especialmente entre los pacientes masculinos.

Se continuó con el proceso de separación de la data, que en este caso fue un 75/25 para los procesos de entrenamiento y prueba, ya que fue lo más conveniente debido a los datos.

Posteriormente se modeló con 3 algoritmos de clasificación, los cuales son: Bayes Naives, Random Forest y Regresión Logística, para luego proceder a la comparación de los mismos, de acuerdo a cada enfermedad.

Comparación de los modelos



Se utiliza el módulo sklearn.metrics, ya que este implementa muchas funciones y utilidades necesarias para medir el rendimiento de la clasificación, de los 3 modelos que se emplean. La función entrena todos los modelos y evalúa su rendimiento de acuerdo con 4 métricas de clasificación comunes. Las mismas se muestran en cada una de las siguientes tablas.

Tabla 2 Comparación de modelos para las ECV

Modelo	Accuracy	Precisión	Recall	F1_Score
Bayes Naives	0.865546218487395	0.816271799705746	0.865546218487395	0.8365443617579156
Random Forest	0.8854087089381207	0.791635568620806	0.8854087089381207	0.8359004423929585
Regresión Logística	0.8691499522445081	0.756974767203391	0.8691499522445081	0.8091932616195675

Según muestra la tabla 2, en la comparación de los algoritmos, todos los modelos para predecir el riesgo de ECV demostraron un rendimiento predictivo con una Accuracy que oscilaba entre 0,865 y 0,885, uno de los modelos dio resultados superiores con un valor predictivo positivo de Precision (0,816) y Accuracy (0,865), lo que permite afirmar que el modelo de Bayes Naives mantuvo un fuerte desempeño predictivo.

Tabla 3 Comparación de modelos para la DM

Modelo	Accuracy	Precisión	Recall	F1_Score
Bayes Naives	0.5301757066462949	0.751129759714196	0.5301757066462949	0.6192489765665218
Random Forest	0.8487394957983193	0.7560654355786202	0.8487394957983193	0.7821356666448954
Regresión Logística	0.8500477554918816	0.723271990810492	0.8500477554918816	0.7815521718881969

La tabla 3 muestra, en la comparación de los algoritmos, todos los modelos para predecir el riesgo de DM demostraron un rendimiento predictivo con una Accuracy que oscilaba entre 0,530 y 0,850, uno de los modelos dio resultados superiores con un valor predictivo positivo de Precision (0,723) y Accuracy (0,850), lo que permite afirmar que el modelo de Regresión Logística mantuvo un fuerte desempeño predictivo.

Tabla 4 Comparación de modelos para las EC

Modelo	Accuracy	Precisión	Recall	F1_Score
Bayes Naives	0.45989304812834225	0.9185342517092019	0.45989304812834225	0.6123641569236116
Random Forest	0.9472880061115355	0.9164423811482635	0.9472880061115355	0.9279950951625515
Regresión Logística	0.9398280802292264	0.8850197505026653	0.9398280802292264	0.911600845914119



De acuerdo a la tabla 4, en la comparación de los algoritmos, todos los modelos para predecir el riesgo de EC demostraron un rendimiento predictivo con una Accuracy que oscilaba entre 0,459 y 0,947, uno de los modelos dio resultados superiores con un valor predictivo positivo de Precision (0,916) y Accuracy (0,947), lo que permite afirmar que el modelo de Random Forest mantuvo un fuerte desempeño predictivo.

Tabla 5 Comparación de modelos para la HTA

Modelo	Accuracy	Precisión	Recall	F1_Score
Bayes Naives	0.5905271199388846	0.586673604848638	0.5905271199388846	0.5795862377505997
Random Forest	0.5905271199388846	0.5836178433381494	0.5905271199388846	0.5840130879712858
Regresión Logística	0.600764087870105	0.6014735200932516	0.600764087870105	0.5776458340328186

En la tabla 5, en la comparación de los algoritmos, todos los modelos para predecir el riesgo de HTA demostraron un rendimiento predictivo con una Accuracy que oscilaba entre 0,59 y 0,60, Todos los modelos dieron resultados con un bajo desempeño predictivo, lo que nos permite afirmar que la data no es suficiente para realizar una buena predicción de esta enfermedad.

Para el proceso de predicción se emplearon modelos de aprendizaje automático (machine learning), esta investigación logró gracias al desarrollo de varios modelos de evaluación, caracterizar el riesgo de padecer ECV, DM, EC y HTA. Para cada enfermedad, todos los modelos lograron un alto rendimiento predictivo, exceptuando para la HTA, ya que los datos presentes en la BD son insuficientes para hacer una buena predicción de la misma. En comparación con otros modelos, el modelo Random Forest tuvo un mejor desempeño, con un valor F1_Score de 0,927 para las EC; el modelo Bayes Naives fue el mejor para las ECV, con un F1_Score de 0,836 y el modelo de Regresión Logística logró mayor desempeño predictivo para la DM, con un F1_Score de 0,781. Este estudio demostró que las tecnologías de machine learning poseen una perspectiva única para identificar factores de riesgo significativos en estudios epidemiológicos a gran escala.

Conclusiones

Se evaluaron 3 modelos de machine learning que fueron ajustados a las variables consideradas factores de riesgos para las principales enfermedades que afectan la salud en el Ecuador. Esto permitió que se llegara a la conclusión de que el modelo Random Forest es el que posee mejor rendimiento para las EC, con un valor F1_Score de 0,927; el modelo de Regresión Logística tiene un mayor desempeño en la DM, con un F1_Score de 0,781 y el modelo Bayes Naives en las ECV, con un F1_Score de 0,836. Para la HTA, todos los modelos dieron resultados muy bajos, lo que nos permite afirmar que la data no es suficiente para realizar una buena predicción de esta enfermedad. Se demostró



que las variables o factores de riesgo que más influyen en cada una de las 4 enfermedades, en el país, según los datos analizados, son imc (índice de masa corporal), ica (índice cintura-altura), c22a (Cuántas comidas completas come al día), c22b (Consume leche, queso o yogurt al menos 1 vez al día), c22c (Come huevos, fréjol, lenteja, al menos 1 vez por semana), c22d (Come carne, pescado, al menos 3 veces por semana), c22e (Consume frutas o verduras al menos 2 veces al día), c22f (Come menos que antes por problemas digestivos), c22f1 (En últimos 3 meses usted come:), c22g (Cuántos vasos de líquidos consume diariamente), c22h En último año ha disminuido de peso sin dieta), c22i (Se considera bien nutrido), c22i1 (Cómo es su movilidad), c22i2 (En últimos 3 meses ha tenido alguna enfermedad), c22i3 (Actualmente presenta úlceras y/o escaras), c23 (En últimos 3 meses en promedio días por semana que toma alcohol). El índice de Shapley fue utilizado para validar estos resultados, y mostró que estos eran los factores más significativos.

Conflictos de intereses

En la confección de este artículo no hubo conflicto de interés entre los autores.

Contribución de los autores

1. Conceptualización: Marely del Rosario Cruz Felipe
2. Curación de datos: Christopher Agustín Holguín Intriago
3. Análisis formal: Christopher Agustín Holguín Intriago
4. Investigación: Shabely Avellán Valdés
5. Metodología: Shabely Avellán Valdés
6. Administración del proyecto: Shabely Avellán Valdés
7. Recursos: Marely del Rosario Cruz Felipe
8. Software: Christopher Agustín Holguín Intriago
9. Supervisión: Marely del Rosario Cruz Felipe
10. Validación: Christopher Agustín Holguín Intriago
11. Visualización: Shabely Avellán Valdés
12. Redacción – borrador original: Shabely Avellán Valdés
13. Redacción – revisión y edición: Marely del Rosario Cruz Felipe

Financiamiento



Esta obra está bajo una licencia *Creative Commons* de tipo **Atribución 4.0 Internacional**
(CC BY 4.0)

Esta investigación no requirió ningún tipo de financiamiento.

Referencias

- Zhou, B., Carrillo-Larco, R. M., Danaei, G., Riley, L. M., Paciorek, C. J., Stevens, G. A., Gregg, E. W., Bennett, J. E., Solomon, B., Singleton, R. K., Sophiea, M. K., Iurilli, M. L. C., Lhoste, V. P. F., Cowan, M. J., Savin, S., Woodward, M., Balanova, Y., Cifkova, R., Damasceno, A., ... Zuñiga Cisneros, J. (2021). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. *Lancet (London, England)*, 398(10304), 957–980. [https://doi.org/10.1016/S0140-6736\(21\)01330-1](https://doi.org/10.1016/S0140-6736(21)01330-1)
- Naghavi, M., Abajobir, A. A., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abera, S. F., Aboyans, V., Adetokunboh, O., Ärnlov, J., Afshin, A., Agrawal, A., Kiadaliri, A. A., Ahmadi, A., Ahmed, M. B., Aichour, A. N., Aichour, I., Aichour, M. T. E., Aiyar, S., Al-Eyadhy, A., ... Murray, C. J. L. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1151–1210. [https://doi.org/10.1016/S0140-6736\(17\)32152-9/ATTACHMENT/C3945D9D-7391-4E78-AEB4-8EBF47C0C63B/MMC2.PDF](https://doi.org/10.1016/S0140-6736(17)32152-9/ATTACHMENT/C3945D9D-7391-4E78-AEB4-8EBF47C0C63B/MMC2.PDF)
- Núñez-González, S., Aulestia-Ortiz, S., Borja-Villacrés, E., & Simancas-Racine, D. (2018). Mortalidad por enfermedades isquémicas del corazón en Ecuador, 2001-2016: estudio de tendencias. *Rev Med Chile*, 146, 850–856.
- Solange Núñez-González, Aglae Duplat, & Daniel Simancas. (2018). Mortalidad por enfermedades cerebrovasculares en Ecuador 2001- 2015: Estudio de tendencias, aplicación del modelo de regresión joinpoint. *Revista Ecuatoriana de Neurología / Vol. 27, No 1, 2018, 27(No 1)*. <http://revecuatneurol.com/wp-content/uploads/2018/09/Mortalidad-por-enfermedades-cerebrovasculares.pdf>
- Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8, 76516–76531. <https://doi.org/10.1109/ACCESS.2020.2989857>
- Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165, 292–299. <https://doi.org/10.1016/J.PROCS.2020.01.047>
- Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132, 1578–1585. <https://doi.org/10.1016/J.PROCS.2018.05.122>
- Weng, S. F., Reys, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, 12(4), e0174944.



<https://doi.org/10.1371/JOURNAL.PONE.0174944>

- Cárdenas-Ruperti, J. P., Durán-Pérez, F. S., & Padilla-Álvarez, L. A. (2020). A Machine Learning approach and probabilistic reasoning in the prog-nosis of risk factors for diabetes. *Ecuadorian Science Journal*, 4(1), 62–67. <https://doi.org/10.46480/esj.4.1.95>
- Calahorrano, A. Z., & Fernández, E. (2018). Diabetes mellitus tipo 2 en el Ecuador: revisión epidemiológica. *Mediciencias UTA*, 2(4), 3–9. <https://doi.org/10.31243/MDC.UTA.V2I4.1219.2018>
- Navarro, A. A. M., Navarro, J. C. M., & Tenorio, S. M. (2020). Frecuencia de hipertensión arterial en personas adultas del Barrio México, Puyo, Pastaza, Ecuador. *Revista Cubana de Reumatología.*, 22(2), 115. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1817-59962020000200008
- Zhang, L., Wang, Y., Niu, M., Wang, C., & Wang, Z. (2020). Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Scientific Reports* 2020 10:1, 10(1), 1–10. <https://doi.org/10.1038/s41598-020-61123-x>
- Laluisa, A. Prevalencia y factores de riesgo para hipertensión arterial en adultos mayores en áreas urbanas y rurales del cantón Ambato. Universidad técnica de ambato facultad de ciencias de la salud carrera de enfermería. *Repositorio UTA*, 76, 2014.
- Macías Sánchez, A. M., & Cevallos Tumbaco, C. M. Modelo basado en aprendizaje de máquina supervisado, para el análisis de datos cardiovasculares y su aplicación en el diagnóstico y pronóstico médico mediante el uso de una página web dinámica. Tesis para obtener el título de ingeniero en sistemas computacionales. Universidad de Guayaquil. Guayaquil – Ecuador. 2019
- WHO Global Health Estimates, "Las 10 principales causas de defunción", Who.int, 2021. [Online]. Disponible: <https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death>. [Accedido: 06- Jul- 2021].
- Organizacion Mundial de la Salud OMS. (2021b). Enfermedades no transmisibles. <https://www.who.int/es/news-room/fact-sheets/detail/noncommunicable-diseases>. [Accedido: 19- Jun- 2022].
- Organizacion Mundial de la Salud OMS. (2021a). Diabetes. <https://www.who.int/es/news-room/fact-sheets/detail/diabetes>. [Accedido: 19- Jun- 2022]
- INEC (Instituto Nacional de Estadísticas y Censos). (n.d.). Encuesta de Salud, Bienestar del Adulto Mayor |. Retrieved June 20, 2022, from <https://www.ecuadorencifras.gob.ec/encuesta-de-salud-bienestar-del-adulto-mayor/>

