ARTÍCULO ORIGINAL
ISSN 1870-7203

Utilidad clínica del *Machine Learning* con Python para la predicción de factores de riesgo cardiovascular

Clinical utility of *Machine Learning* with Python for predicting cardiovascular risk factors

Francisco Delgado Ayala,*,‡ Enrique Juan Díaz Greene,*,§ Federico Leopoldo Rodríguez Weber*,¶

Citar como: Delgado AF, Díaz GEJ, Rodríguez WFL. Utilidad clínica del *Machine Learning* con Python para la predicción de factores de riesgo cardiovascular. Acta Med GA. 2025; 23 (4): 323-328. https://dx.doi.org/10.35366/120510

Resumen

Introducción: demostramos la facilidad y el potencial del *Machine Learning* para mejorar la detección y prevención de factores de riesgo cardiovascular; además, enfatizamos el análisis de variables numéricas continuas, frecuentemente omitidas en investigaciones, subrayando la innovación del enfoque de nuestro estudio. **Objetivo:** evaluar con Python mediante métricas de rendimiento, la aplicabilidad de métodos predictivos (árbol de decisión, bosque aleatorio y K-Nearest Neighbors [KNN]). **Material y métodos:** realizamos una búsqueda en PubMed, Google Scholar, cubriendo el periodo de 1995 a 2023, los términos incluyeron "aprendizaje automático" y *"machine learning prediction model"*. La bibliografía adicional se identificó mediante búsquedas complementarias en internet y medios físicos. Se incluyeron datos somatométricos y de laboratorio de una población de trabajadores de un hospital en México. **Resultados y conclusión:** el bosque aleatorio mostró el mejor desempeño tanto en variables numéricas como categóricas. Las numéricas se evaluaron con raíz del error cuadrático medio (RMSE), error absoluto medio (MAE), error cuadrático medio (MSE) y coeficiente de determinación (R^2), mientras que las variables categóricas con exactitud, precisión, sensibilidad, puntaje F1 y ROC/AUC. Esto demuestra la robustez del bosque aleatorio en el manejo de diferentes tipos de datos, sugiriendo un gran potencial en contextos clínicos para la detección temprana de riesgos y estrategias de intervención.

Palabras clave: aprendizaje automático, factores de riesgo cardiovascular, predicción clínica, modelos predictivos, Python.

Abstract

Introduction: we demonstrate the ease and potential of machine learning in enhancing the detection and prevention of cardiovascular risk factors. Our study emphasizes the analysis of continuous numerical variables, often overlooked in research, highlighting the innovative approach of our investigation. **Objective:** to assess, using Python and performance metrics, the applicability of predictive methods (decision tree, random forest, and K-Nearest Neighbors [KNN]). **Material and methods:** a search was conducted on PubMed and Google Scholar, covering 1995 to 2023. Terms included "aprendizaje automático" and "machine learning prediction model". Additional literature was sourced through supplementary online searches and physical media. The study includes somatometric and laboratory data from a population of hospital workers in Mexico. **Results and conclusion:** the Random Forest model exhibited superior performance for numerical and categorical variables. Numerical variables were evaluated using root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE), and coefficient of determination (R^2). In contrast, categorical variables were assessed with accuracy, precision, sensitivity, F1-score, and ROC/AUC. This proves the Random Forest's robustness in handling diverse data types, suggesting its significant potential for early risk detection and intervention strategies in clinical settings.

Keywords: machine learning, cardiovascular risk factors, clinical prediction, predictive models, Python.

* Hospital Angeles Pedregal. Facultad Mexicana de Medicina, Universidad La Salle México. Ciudad de México, México.
‡ Residente de Medicina Interna.
§ Profesor titular del curso de Medicina Interna.
¶ Profesor adjunto del curso de Medicina Interna.
ORCID: 0000-0001-5680-4743

Correspondencia:
Francisco Delgado Ayala
Correo electrónico: fco.delaya@outlook.com

Recibido: 23-03-2024. *Aceptado:* 09-08-2024.

www.medigraphic.com/actamedica

Abreviaturas:
KNN = K-Nearest Neighbors
MAE = error absoluto medio
ML = *Machine Learning*
MSE = error cuadrático medio
R^2 = coeficiente de determinación
RMSE = raíz del error cuadrático medio
ROC/AUC = característica operativa del receptor *(Receiver Operating Characteristic)*/área bajo la curva *(Area Under the Curve)*

INTRODUCCIÓN

La historia del aprendizaje automático, o *Machine Learning* (ML), data de más de 50 años. Los algoritmos de ML han demostrado su eficacia empleando una combinación de métodos estadísticos, probabilísticos y de optimización para aprender de datos que reflejan nuestro pasado y descubrir patrones y relaciones en grandes conjuntos de datos, ya sean estructurados o no. Estas tecnologías han encontrado aplicaciones en diversas áreas, incluyendo la categorización automática de texto, determinación de vulnerabilidades en seguridad informática, la detección de correo spam, la prevención de fraudes con tarjetas de crédito, en *business intelligence* con el descubrimiento de patrones de compra de los clientes y, más importante, en el modelado de enfermedades.[1-9]

El ML puede dividirse en dos categorías generales:

1. Aprendizaje supervisado: involucra el desarrollo de modelos predictivos que aprenden de un conjunto de datos con etiquetas conocidas, lo que permite predecir resultados en ejemplos no etiquetados.
2. Aprendizaje no supervisado: se centra en crear modelos que identifican patrones o agrupaciones basándose en las características intrínsecas de los datos.[1,5,7]

El florecimiento del ML en la medicina clínica comenzó en la segunda mitad de la década de 1990 y la primera mitad de los 2000, destacándose en áreas como la genética, la oncología y la descripción de proteínas en bioquímica. Hoy en día, con nuevos algoritmos y mayor capacidad computacional, las posibilidades de impacto en la medicina diaria se han expandido enormemente.[1,10]

Árbol de decisión

Un árbol de decisión realiza pruebas y determina resultados para segmentar los datos en una estructura ramificada. Esta estructura está compuesta por varios tipos de nodos: el nodo "raíz", que inicia el árbol; los nodos "hoja", que representan las decisiones o resultados finales; y los nodos "internos", que corresponden a las pruebas realizadas para subdividir los datos. En estos últimos, se selecciona el nodo hijo que ofrezca la mejor calidad de información. El proceso de análisis y bifurcación se repite sucesivamente hasta alcanzar un nodo hoja, el cual presenta el resultado final.[1,2,8]

Bosque aleatorio

El bosque aleatorio es un conjunto de múltiples árboles de decisión que mejora la precisión de la predicción y reduce el riesgo de sobreajuste inherente a un solo árbol de decisión. En clasificación, cada árbol en el bosque emite un "voto" y la clase con la mayoría de votos es elegida como la salida del modelo. En regresión, se promedian los resultados de todos los árboles. Cada árbol se construye con una técnica llamada *bootstrap sampling* y consideran un subconjunto aleatorio de características para cada división.[1,2,8]

Los árboles individuales se construyen a partir de subconjuntos de datos seleccionados mediante muestreo con reemplazo *(bootstrap sampling)* y consideran un subconjunto aleatorio de características para cada división.[1,2,8]

Tabla 1: Comparativa de la capacidad predictiva para variables categóricas entre los algoritmos de árbol de decisión, bosque aleatorio y KNN.

Algoritmo	Exactitud	Precisión	Sensibilidad	F1	ROC/AUC
Árbol de decisión	0.61	0.41	0.61	0.48	0.77
Bosque aleatorio	0.61	• 0.57	0.61	• 0.55	0.77
KNN	• 0.62	0.39	• 0.62	0.48	• 0.78

AUC = área bajo la curva *(Area Under the Curve)*. KNN = K-Nearest Neighbors. ROC = característica operativa del receptor *(Receiver Operating Characteristic)*.
Los mejores resultados se muestran con un (•).

Tabla 2: Comparativa de la capacidad predictiva para variables numéricas por RMSE, MAE, R^2 y MSE, de los algoritmos de árbol de decisión, bosque aleatorio y KNN.

Variable/Algoritmo	RMSE	MAE	R^2	MSE
Tensión sistólica				
Árbol de decisión	1.18	0.89	-0.69	1.4
Bosque aleatorio	• 0.80	• 0.63	• 0.22	• 0.65
KNN	0.84	0.65	0.15	0.71
Tensión diastólica				
Árbol de decisión	0.87	0.68	0.3	0.76
Bosque aleatorio	• 0.64	• 0.46	• 0.62	• 0.41
KNN	0.83	0.62	0.37	0.68
Circunferencia de cintura				
Árbol de decisión	0.77	0.57	0.41	0.6
Bosque aleatorio	• 0.59	• 0.43	• 0.66	• 0.35
KNN	0.69	0.53	0.53	0.48
Perímetro de pantorrilla				
Árbol de decisión	1.37	0.77	-1.17	1.88
Bosque aleatorio	• 0.76	• 0.63	• 0.34	• 0.57
KNN	0.91	0.76	0.04	0.83

KNN = K-Nearest Neighbors. MAE = error absoluto medio. MSE = error cuadrático medio. RMSE = raíz del error cuadrático medio.
Los mejores resultados se muestran con un (•).

KNN (K vecinos cercanos)

Es un algoritmo de clasificación y regresión que predice la categoría o el valor de un dato, basándose en la mayoría de votos o en el promedio de los "K" datos más cercanos de un conjunto de entrenamiento. Para una tarea de clasificación, la etiqueta más frecuente entre los vecinos es la asignada al nuevo punto. Para regresión, se calcula un promedio de los valores de los "K" vecinos más cercanos.[1,2,8]

MATERIAL Y MÉTODOS

Este estudio unicéntrico, clínico-epidemiológico, observacional y retrospectivo[11] incluyó datos recopilados en 2017 de una población de trabajadores de un hospital en la Ciudad de México. La población de estudio consistió en personas mayores de 18 años que aceptaron la toma de muestras de laboratorio y somatometría, y que firmaron el consentimiento informado. Se excluyeron sujetos que presentaron más de cuatro variables con datos faltantes de las 12 variables principales: edad, sexo, turno laboral, peso, talla, tensión arterial sistólica, tensión arterial diastólica, glucosa sérica, colesterol total, colesterol HDL, circunferencia de cintura y perímetro de pantorrilla.

Además, se evaluaron variables compuestas como el índice de masa corporal, la razón circunferencia de cintura/talla, la resta colesterol total y HDL, y la razón colesterol total/HDL. Para la evaluación y selección del mejor algoritmo predictivo, ya sea para clasificación (variables categóricas) o regresión (variables numéricas) nos basamos en el rendimiento de sus métricas. Para clasificación se evaluó exactitud, precisión, sensibilidad, puntaje F1 y ROC/AUC; para regresión, raíz del error cuadrático medio (RMSE), error absoluto medio (MAE), el coeficiente de determinación (R^2) y el error cuadrático medio (MSE).

Datos y modelos de *Machine Learning*

Procesamiento previo de datos

Se realizaron cambios en el conjunto de datos de la población original para proteger la privacidad y se eliminaron participantes duplicados. También se descartaron variables no esenciales para la calidad del estudio.

De 872 participantes iniciales, se obtuvieron datos de 641 con 16 variables, sumando 7,064 registros. El porcentaje de datos faltantes fue diferente según la variable (turno laboral 0.78%, tensión arterial sistólica 0.62%, tensión arterial diastólica 0.62%, circunferencia de cintura 13.88% y perímetro de pantorrilla 82.06%).

Se utilizó el lenguaje Python y sus principales bibliotecas (Pandas, NumPy, matplotlib.pyplot, Seaborn, ydata_profiling, SciPy, Scikit-learn) para la manipulación y limpieza de datos, análisis exploratorio, análisis estadístico y desarrollo de modelos predictivos. Para asegurar la reproducibilidad y minimizar el sobreajuste, se emplearon tres divisiones aleatorias del conjunto de datos, distribuyendo los datos en grupos de entrenamiento y prueba en una proporción de 70 y 30%, respectivamente, mediante la función train_test_split de *sklearn*. El código de programación (repositorio) y análisis estadístico de este trabajo están disponibles para consulta en GitHub[12,13]: **https://github.com/paco-da/utilidad_clinica_ml**

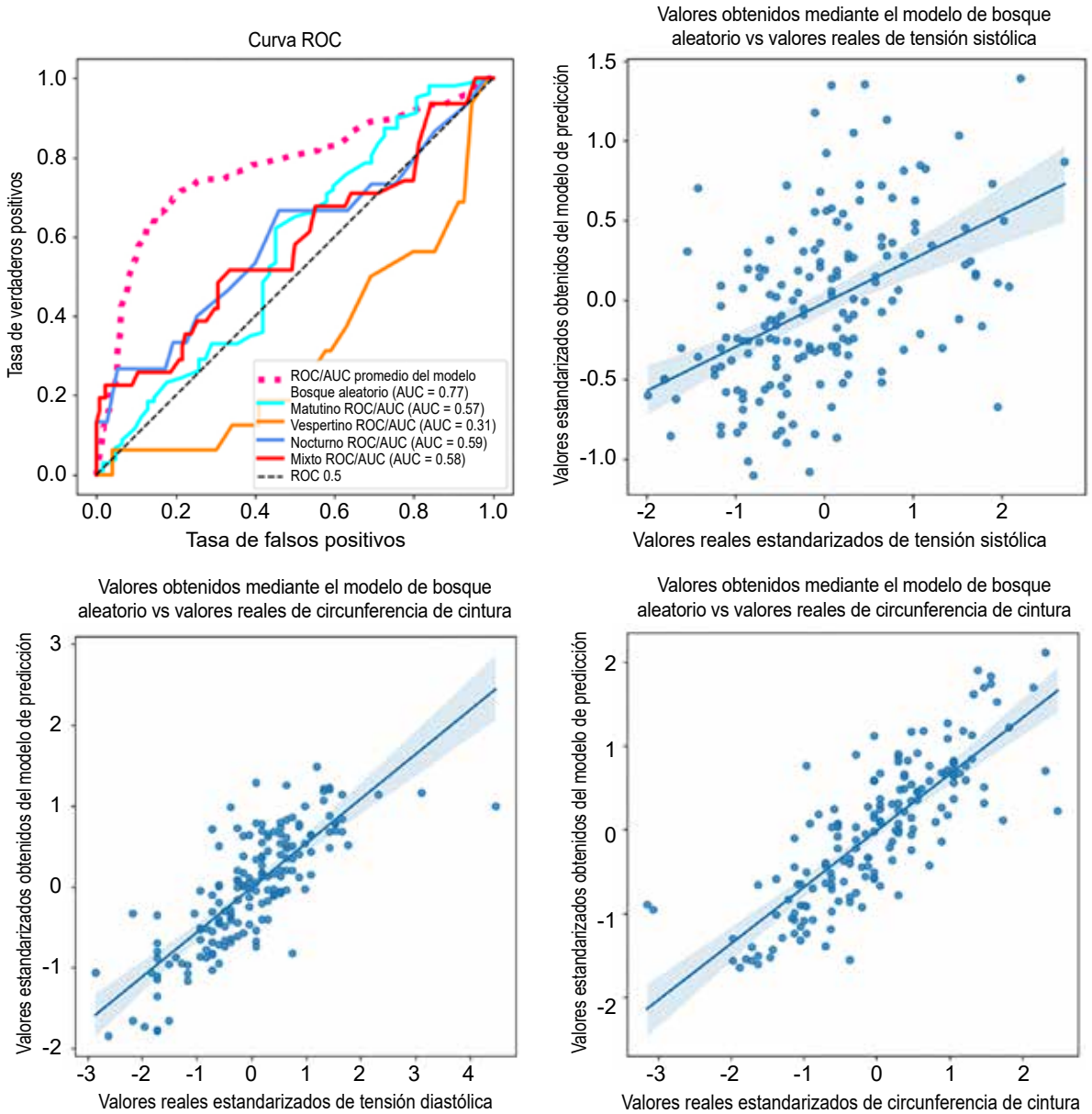

Figura 1: Resumen gráfico de las pruebas estadísticas aplicadas a los modelos de predicción.
AUC = área bajo la curva *(Area Under the Curve)*. ROC = característica operativa del receptor *(Receiver Operating Characteristic)*.
Gráfico elaborado con Matplotlib, Seaborn y Sklearn. En la imagen superior izquierda se observa la curva ROC/AUC del desempeño para las variables categóricas del modelo bosque aleatorio, el mejor desempeño se observa en el modelo bosque aleatorio general, con un AUC de 0.77, que indica un buen nivel de discriminación. En la imagen superior derecha y ambas imágenes en la parte inferior se observan gráficos de dispersión, donde cuanto más cerca se encuentren los puntos a la línea, mejor será la predicción del modelo, lo que indica que los algoritmos de bosque aleatorio están produciendo predicciones precisas.

Modelos de Machine Learning

El primer modelo predictivo se desarrolló para la única variable categórica, seguida del resto de variables numéricas. Para determinar el mejor algoritmo, se desarrollaron y analizaron árboles de decisión con profundidades de 3 a 25 nodos, bosques aleatorios con un rango de 100 a 125 árboles, y ajustando el número de vecinos de 3 a 35 para encontrar el óptimo en términos de capacidad predictiva en el algoritmo KNN.

Finalmente, la elección del modelo con la mejor capacidad de predicción se llevó a cabo mediante un enfoque jerárquico. En clasificación se estableció un orden descendente de importancia dándole mayor peso a la exactitud, seguido por la precisión, después sensibilidad, F1 y ROC/AUC. En regresión, se priorizó el RMSE, seguido por el MAE, R^2 y MSE *(Figura 1)*.

RESULTADOS

El bosque aleatorio exhibió el mejor desempeño global, debido a que superó en todas las predicciones de variables numéricas y fue ligeramente inferior que KNN para variables categóricas. En variables categóricas *(Tabla 1)* logró una exactitud y sensibilidad de 0.61; superó al árbol de decisión y KNN en valor predictivo positivo –o precisión– con 0.57, presentó un F1 de 0.55 y tuvo un ROC/AUC de 0.77.

El bosque aleatorio fue el mejor en todas las variables numéricas *(Tabla 2)*. En la predicción de tensión sistólica, el bosque aleatorio obtuvo un RMSE de 0.80, MAE 0.63, R^2 de 0.22 y un MSE de 0.65, lo que indica una menor variabilidad y refleja una menor dispersión de los errores. En tensión diastólica, de igual manera, el bosque aleatorio obtuvo el mejor RMSE, MAE, R^2 y MSE (0.64, 0.46, 0.62 y 0.41, respectivamente). Las predicciones en circunferencia de cintura (RMSE 0.59, MAE 0.46, R^2 0.66 y MSE 0.35) y perímetro de pantorrilla (RMSE 0.76, MAE 0.63, R^2 0.34 y MSE 0.57) también fueron las mejores.

DISCUSIÓN

Comparamos tres algoritmos de aprendizaje automático para la predicción de factores de riesgo cardiovascular, utilizando Python como herramienta analítica. El rendimiento superior del bosque aleatorio sugiere su aplicabilidad en la práctica clínica, especialmente por su precisión y robustez frente a las variables numéricas, que son fundamentales en la medición de todo tipo de variables en medicina. Debemos recalcar que un modelo altamente preciso no es sinónimo de perfección y que el sobreajuste del modelo lleva a estimación de datos erróneos. La estandarización de los valores y la exploración en la predicción de variables numéricas es lo que contrasta con la gran mayoría de los artículos científicos actuales, y dota de originalidad a nuestro estudio donde se compara la capacidad predictiva del *Machine Learning* en medicina.

CONCLUSIONES

El aprendizaje automático *(Machine Learning)* se emplea actualmente como herramienta en múltiples áreas, tales como finanzas, física, biología, videojuegos o en nuestra vida diaria, especialmente cuando utilizamos a ChatGPT como "multiusos"; y la medicina no debería ser una excepción. No obstante, existen diferentes obstáculos que deben ser superados para adoptar estas tecnologías con el objetivo de mejorar la salud de nuestros pacientes. La creación y publicación de este tipo de trabajos representa la superación de uno de esos obstáculos.

REFERENCIAS

1. Deo RC. Machine Learning in Medicine. Circulation. 2015; 132 (20): 1920-1930. doi: 10.1161/CIRCULATIONAHA.115.001593.
2. Géron A. Hands-on machine learning with scikit-learn, keras, and tensorflow. Third. In: Butterfield N, Taché N, Cronin M, Kelly B, Cofer K, Head R, et al., editors. Sebastopol: O'Reilly Media; 2023.
3. McKinney W. Python for Data Analysis. Third. In: Haberman J, Rufino A, Faucher C, Saruba S, Klefstad S, Futato D, et al., editors. Sebastopol: O'Reilly; 2022.
4. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. Stroke. 2019; 50 (5): 1263-1265. doi: 10.1161/STROKEAHA.118.024293.
5. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019; 19 (1): 281. doi: 10.1186/s12911-019-1004-8.
6. Theng D, Theng M. Machine learning algorithms for predictive analytics: a review and new perspectives. High Technology Letters. 2020; 26 (6): 537-345.
7. Park DJ, Park MW, Lee H, Kim YJ, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Sci Rep. 2021; 11 (1): 7567. doi: 10.1038/s41598-021-87171-5.
8. Wei C, Zhang L, Feng Y, Ma A, Kang Y. Machine learning model for predicting acute kidney injury progression in critically ill patients. BMC Med Inform Decis Mak. 2022; 22 (1): 17. doi: 10.1186/s12911-021-01740-2.
9. Rodríguez-Rivas JG, Rodríguez-Castillo S. Uso de Python para el análisis de datos aplicado en la investigación. Revista Incaing. 2022, 33-40.
10. Bhaskar H, Hoyle DC, Singh S. Machine learning in bioinformatics: a brief survey and recommendations for practitioners. Comput Biol Med. 2006; 36 (10): 1104-1125.
11. Martínez Montaño M del LC, Briones Rojas R, Cortés Riveroll JGR. Metodología de la investigación para el área de la salud. 2nd ed. de

León Fraga J, Guerrero Aguilar. Héctor F., Manjarrez de la Vega JJ, editores. México: McGraw-Hill; 2013.
12. Rule A, Birmingham A, Zuniga C, Altintas I, Huang SC, Knight R et al. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLoS Comput Biol. 2019; 15 (7): e1007007. doi: 10.1371/journal.pcbi.1007007.
13. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J et al. Ten simple rules for taking advantage of git and GitHub. PLoS Comput Biol. 2016; 12 (7): e1004947. doi: 10.1371/journal.pcbi.1004947.

Cumplimiento de las directrices éticas: los autores declaramos que no tenemos ningún tipo de conflicto de intereses. Todos los procedimientos siguieron los estándares éticos del Comité de Experimentación Humana y la Declaración de Helsinki de 1975, revisada en 2000.

Si desea consultar los datos complementarios de este artículo, favor de dirigirse a **editorial.actamedica@saludangeles.mx**