Revista Ingenio
UFPS Ocaña
"Journal of Engineering Sciences"

Original Article Doi: https://doi.org/10.22463/2011642X.4684

Análisis predictivo del riesgo de hipertensión en adultos mexicanos basado en indicadores nutricionales y calóricos

Predictive analysis of hypertension risk in Mexican adults based on nutritional and caloric indicators

MHPE-TE. Héctor Alejandro Acuña-Cid[1], PhD. Alejandro Mauricio-González[2], PhD. Roberto Solis- Robles[3], MSP. Anayancin Acuña-Ruiz[4], PhD. Luis Carlos Reveles-Gómez[5]

[1] *Maestría en Ciencias en Procesamiento de la Información, Universidad Autónoma de Zacatecas, Zacatecas, México, Orcid: https://orcid.org/0000-0002-8477-0963, Email: hacuna@uaz.edu.mx*
[2] *Maestría en Ciencias en Procesamiento de la Información, Universidad Autónoma de Zacatecas, Zacatecas, México, Orcid: https://orcid.org/0000-0002-2882-9633, Email: amgdark@uaz.edu.mx*
[3] *Maestría en Ciencias en Procesamiento de la Información, Universidad Autónoma de Zacatecas, Zacatecas, México, Orcid: https://orcid.org/0000-0001-6629-1048, Email: rsolis@uaz.edu.mx*
[4] *Licenciatura en Nutrición, Universidad Autónoma de Zacatecas, Zacatecas, México, Orcid: https://orcid.org/0000-0003-2630-1752, Email: anayancin.acuna@uaz.edu.mx*
[5] *Maestría en Ciencias en Procesamiento de la Información, Universidad Autónoma de Zacatecas, Zacatecas, México, Orcid: https://orcid.org/0009-0001-5406-1656, Email: luiscarlosreveles@uaz.edu.mx*

Cómo citar: H.A. Acuña-Cid, A. M. González, R. Solis-Robles y A. Acuña-Ruiz y L.C. Reveles-Gómez, "Análisis predictivo del riesgo de hipertensión en adultos mexicanos basado en indicadores nutricionales y calóricos", Rev. Ingenio, vol. 22, n º1, pp.24-32, 2025, doi: https://doi.org/10.22463/2011642X.4684

Fecha de recibido:1 de agosto de 2024
Fecha aprobación:13 de diciembre de 2024

RESUMEN

Palabras clave:

aprendizaje automático, nutrición, riesgos cardiovasculares, salud pública, predicción de enfermedades crónicas.

Este artículo desarrolla un análisis predictivo del riesgo de hipertensión en adultos mexicanos basado en indicadores nutricionales y calóricos. La hipertensión, una condición con serias implicaciones para la salud, requiere la identificación de factores de riesgo predictivos para su prevención y manejo efectivo. Se evaluaron varios modelos de aprendizaje automático, encontrando que el modelo Random Forest destaca por su alta precisión y robustez, mientras que el XGBoost sobresale por su eficiencia en conjuntos de datos grandes. En contraste, el modelo Naive Bayes mostró el rendimiento más bajo. Además, el estudio enfatiza la importancia de los macronutrientes y la ingesta calórica total en la predicción de la hipertensión, con proteínas, carbohidratos y lípidos como factores relevantes en el riesgo, especialmente en adultos jóvenes en México. Este hallazgo resalta la necesidad de integrar múltiples factores nutricionales en la evaluación del riesgo.

ABSTRACT

Keywords:

machine learning, nutrition, cardiovascular risks, public health, chronic disease prediction

This article develops a predictive analysis of hypertension risk in Mexican adults based on nutritional and caloric indicators. Hypertension, a condition with serious health implications, requires the identification of predictive risk factors for its prevention and effective management. Several machine learning models were evaluated, with the Random Forest model standing out for its high accuracy and robustness, while XGBoost excelled in efficiency with large datasets. In contrast, the Naive Bayes model showed the lowest performance. Additionally, the study emphasizes the importance of macronutrients and total caloric intake in predicting hypertension, with proteins, carbohydrates, and lipids being relevant risk factors, especially in young adults in Mexico. This finding highlights the need to integrate multiple nutritional factors in risk assessment.

1. Introducción

La hipertensión arterial se ha establecido como una de las enfermedades crónicas más comunes a nivel mundial y constituye un factor de riesgo clave para las enfermedades cardiovasculares. Este padecimiento plantea un reto significativo para los sistemas de salud globales, manifestándose en una alta carga de morbilidad y en los costos elevados relacionados con el tratamiento y manejo de sus complicaciones [1].

En México, la situación con respecto a la hipertensión es particularmente alarmante. En las últimas décadas, se ha observado un aumento considerable en la prevalencia de esta enfermedad entre los adultos. Este incremento destaca la urgencia de enfocar los esfuerzos en identificar y manejar los factores que contribuyen a esta tendencia preocupante [2]. Los estudios han demostrado que los patrones alimenticios, incluyendo el consumo de macronutrientes y calorías, tienen un impacto significativo en el control de la presión arterial. Esto resalta el rol crucial que la dieta puede desempeñar en el riesgo de desarrollar hipertensión, señalando una relación directa entre la nutrición y esta condición [3], [4].

2. Relación con teorías y trabajos existentes

La relación entre los hábitos alimenticios y el riesgo de hipertensión ha sido estudiada en la población adulta

Corresponding Author
Email: hacuna@uaz.edu.mx (Héctor Alejandro Acuña Cid)

mexicana, abarcando un rango de edad de 20 a 78 años, según datos de la Encuesta Nacional de Salud y Nutrición (ENSANUT) [5], [6]. La ENSANUT Continua de 2022, dirigida por Ismael Campos et al., revela que la prevalencia de hipertensión arterial en adultos mayores de 20 años es del 29.4%, con una incidencia un 27% más alta en hombres que en mujeres. Este hallazgo subraya la importancia de continuar actualizando y ampliando el conocimiento sobre esta condición, aprovechando la implementación de técnicas y herramientas avanzadas que facilitan estos análisis [5].

En este contexto, el desarrollo de modelos predictivos en el campo de la salud se ha demostrado como un ejemplo particularmente valioso. Según Hasdeu et al., estos modelos son esenciales para probar hipótesis y cuantificar riesgos, beneficios y costos en el manejo de enfermedades [7].

En 2005, Delen, Walker y Kadam realizaron una comparación entre tres modelos predictivos —redes neuronales artificiales, árboles de decisión y regresión logística— para predecir la supervivencia del cáncer de mama. Utilizaron medidas de desempeño como la exactitud, sensibilidad y especificidad [8].

En 2006, Bellazzi y Zupan abordaron la aplicación de la minería de datos en el ámbito médico, destacando modelos predictivos como SVM, clasificador naive-bayes y redes bayesianas. Propusieron una guía para implementar la minería de datos en medicina y resaltaron su contribución en este campo [9].

La importancia del uso de modelos en el contexto de la hipertensión ha sido destacada en diversas investigaciones. Por ejemplo, Araujo-Castro et al. lograron entrenar un modelo para predecir la resolución de la hipertensión post-adrenalectomía en pacientes con aldosteronismo primario (PA), ofreciendo una herramienta útil para el asesoramiento preoperatorio [10].

Además, el modelo IberScore desarrollado por Ruilope et al., que estima el riesgo cardiovascular en una población laboral española joven y saludable, ha sido aplicado exitosamente en estrategias de prevención cardiovascular. Carlos Fernández et al. combinaron los resultados de IberScore con la estrategia LIFE'S SIMPLE 7 de la American Heart Association, logrando una prevención más efectiva en la población laboral de España [11], [12].

Por otro lado, Luo et al. publicaron en 2023 que los síntomas de depresión aumentaron en un 19% el riesgo de incidencia de hipertensión, utilizando el modelo de Cox programado en R [13]. Además, en el mismo año, Argoty-Pantoja et al. presentaron la relación entre el índice de glucosa y la presión arterial, utilizando análisis de varianza (ANOVA) y modelos de regresión lineal, así como el modelo de Cox [14].

Por último, Kendale et al. en un artículo reciente emplearon diversos modelos predictivos, como regresión lineal, bosques aleatorios, SVM, naive bayes, k-vecinos más cercanos, análisis discriminante lineal, redes neuronales y máquina de aumento de gradiente, para predecir el riesgo de hipotensión postinducción en pacientes sometidos a anestesia general [15].

No obstante, a pesar de estos avances, la creación de modelos predictivos específicos para la hipertensión en México ha sido limitada. Chuen-Den Tseng et al. advierten que los resultados de estudios realizados en una población pueden no ser aplicables a otros grupos étnicos [16]. Por ello, es importante el desarrollo de modelos predictivos basados en datos de la población mexicana, esto permitiría implementar estrategias de prevención y control más precisas y adaptadas.

En particular, este proyecto busca abordar este vacío al desarrollar un algoritmo predictivo que, a través del análisis detallado del consumo de macronutrientes y calorías, identifique el riesgo de hipertensión la población adulta mexicana, integrando estos datos con el Sistema Mexicano de Alimentos Equivalentes (SMAE) [17] y utilizando técnicas avanzadas de aprendizaje automático, con el propósito de contribuir a la prevención temprana y personalizada de la hipertensión, buscando así una mejora en la salud y bienestar de la población mexicana.

3. Metodología

Para llevar a cabo este trabajo, se implementó la metodología Cross-Industry Standard Process for Data Mining (CRISP-DM). Según IBM, utilizar CRISP-DM como metodología permite describir las fases habituales de un proyecto de minería de datos, incluyendo las tareas necesarias en cada fase y la explicación de las relaciones entre estas tareas. Por otro lado, al aplicarlo como modelo de proceso, se obtiene una visión resumida del ciclo de vida completo de la minería de datos, tal como se ilustra en la Figura 1 [18].

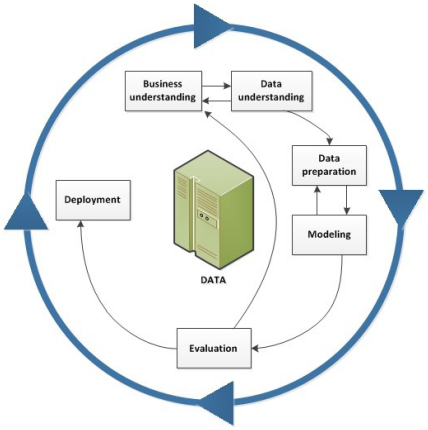

Figura 1. Diagrama CRISP-DM [18]

Esta implementación se dividió en seis fases principales y cada una con diversas actividades específicas. La primera fase, Comprensión del negocio, comenzó con una revisión de la literatura existente sobre hipertensión, nutrición y los riesgos asociados. Esta etapa permitió definir los alcances del proyecto, clarificando el problema de negocio y especificando los objetivos del proyecto.

En la segunda fase, Comprensión de los datos, se utilizó la información ya existente de la ENSANUT CONTINUA 2022 y del SMAE 5ta edición. Inicialmente, se llevó a cabo un análisis detallado para determinar qué instrumentos de la encuesta eran esenciales, enfocándose específicamente en aquellos que contenían las variables relevantes para el estudio de la hipertensión. Es importante destacar que la población de estudio seleccionada para este análisis oscila entre los 20 y 59 años, permitiendo así enfocarse en un grupo demográfico clave para la evaluación de la prevalencia y los factores de riesgo asociados a esta condición.

En la Tabla 1, se presentan los nombres de los instrumentos de la ENSANUT CONTINUA 2022 utilizados en este análisis, junto con las variables específicas extraídas de cada uno de ellos. Aunque variables como la calidad del sueño y la escala de estrés son reconocidas en la literatura como fundamentales en la detección y manejo de la hipertensión [19], [20], fueron omitidas en este estudio porque no estaban disponibles con la misma consistencia y completitud en la ENSANUT CONTINUA 2022. La selección final se centró en las variables con mayor robustez y validez dentro del contexto del estudio de la hipertensión, garantizando así la solidez de los datos y su ajuste a la realidad de la población mexicana analizada.

Tabla 1. Conjunto de datos de la ENSANUT CONTINUA 2022.

Instrumento	Variable
Cuestionario de salud de adultos (20 años o más)	Folio ¿Cuál es el sexo del encuestado? ¿Cuántos años cumplidos tiene actualmente? ¿Algún médico le ha dicho que tiene la presión alta?
Cuestionario de antropometría y tensión arterial.	Peso Talla Circunferencia de cintura Tensión arterial: sístole Tensión arterial: diástole
Actividad física - Adolescente y adultos	¿Cuántas horas en promedio duerme en un día? Durante los últimos 7 días, ¿Cuánto tiempo en total estuvo sentado(a) en uno de esos días de la semana?
Cuestionario de recordatorio de 24 horas	¿Me puede decir todo lo que comió y bebió el día de ayer, desde que se levantó hasta antes de dormir? Por favor incluya cualquier alimento o bebida. ¿Cuántas veces ha consumido este alimento en el día? ¿Cuántas veces ha consumido este alimento en la semana?"

Fuente: [5].

El Cuestionario de Recordatorio de 24 Horas jugó un papel crucial en el estudio al proporcionar un detallado registro de los alimentos y bebidas consumidos el día anterior a la encuesta. Sin embargo, este cuestionario no especificaba las cantidades de proteínas, lípidos, carbohidratos, ni el total calórico de los consumos registrados. Para abordar esta limitación y garantizar una evaluación nutricional precisa, se utilizó el Sistema Mexicano de Alimentos Equivalentes (SMAE) 5ta edición, estableciendo así una relación exacta entre cada alimento y bebida reportados y su correspondiente contenido de macronutrientes. De esta manera, el cuestionario cumplió su cometido al ser complementado con el SMAE para obtener una evaluación completa y precisa. Este proceso se realizó de manera manual, identificando un total de 158 alimentos y bebidas diferentes en el CR24.

Asimismo, se utilizaron otras dos columnas del cuestionario para determinar la frecuencia de consumo, las cuales preguntaban ¿Cuántas veces ha consumido este alimento en el día? y/o ¿Cuántas veces ha consumido este alimento en la semana?, en situaciones donde las columnas de frecuencia de consumo estaban vacías, pero había un registro de alimento o bebida, se asumió que dicho elemento se consumió una sola vez. Este método facilitó el cálculo de las calorías, proteínas, lípidos y carbohidratos de cada producto, utilizando los valores especificados por el SMAE.

Una vez calculados los macronutrientes para cada registro, se sumaron todos los valores correspondientes a cada folio para obtener un total que sería utilizado en el análisis posterior. Al final de este proceso meticuloso, se consolidaron 922 elementos para considerar en las fases subsiguientes del estudio.

La tercera fase fue la Preparación de los datos, donde se llevó a cabo la limpieza, integración y construcción de variables importantes como el rango de edad, calorías, proteínas, lípidos, carbohidratos, entre otras. Esta preparación fue crucial para facilitar el análisis y modelado posterior.

Se utilizó el instrumento Cuestionario de salud de adultos (20 años o más), que contenía datos sociodemográficos y de salud, se procedió a la categorización del sexo en 0 y 1. Además, se establecieron categorías de edad en intervalos de 20 a 29, 30 a 39, 40 a 49 y 50 a 59 años. La hipertensión se codificó como 1 para indicar presencia y 0 para su ausencia.

Desde el Cuestionario de antropometría y tensión arterial, se crearon nuevas variables a partir de los datos originales de peso, talla, circunferencia de cintura y las mediciones de tensión arterial sistólica y diastólica, calculando el promedio de estas mediciones para cada individuo.

Con respecto al instrumento Actividad física - Adolescente y adultos, se categorizaron las respuestas de las preguntas sobre el tiempo promedio de sueño diario y el tiempo total sentado en un día típico de la semana pasada. Las categorías utilizadas fueron Menos de 5 horas, 5-8 horas, y Más de 8 horas.

Finalmente, todos estos datos se consolidaron en un único conjunto de datos que incluyó 15 variables y 658 filas, tras la eliminación de registros vacíos o nulos. Este conjunto de datos resultante, enfocado en la hipertensión como variable principal.

En la cuarta fase, Modelado, se realizó un análisis estadístico descriptivo de las variables para proporcionar una comprensión sólida y fundamentada de los datos con los que se estaba trabajando. Este paso es crucial en cualquier proceso de análisis de datos, ya que ofrece una vista preliminar de las tendencias, patrones y anomalías dentro del conjunto de datos, permitiendo tomar decisiones sobre los pasos a seguir en el modelado y en la interpretación de los resultados [21]. En la Tabla 2 se muestra una parte del análisis descriptivo de las columnas calorías, lípidos, proteínas e hidratos.

Tabla 2. Análisis descriptivo de las variables

	Calorías	Proteínas	Lípidos	Hidratos
Count	658.00	658.00	658.00	658.00
Mean	9237.92	350.38	321.30	1292.9
Std	5858.64	232.34	234.31	812.98
Min	830.95	21.19	18.59	128.44
25%	5369.47	192.64	171.06	728.77
50%	7650.58	286.71	263.17	1082.3
75%	11873.3	428.87	399.99	1635.9
Max	44075.6	1540.71	1826.01	6321.84

Los resultados del análisis estadístico descriptivo revelaron una notable variabilidad en la ingesta de macronutrientes, lo que sugiere una diversidad significativa en las preferencias dietéticas y necesidades nutricionales de los individuos estudiados. Esta variabilidad en los datos, junto con el tamaño adecuado de la muestra, proporciona una base sólida para considerar la viabilidad de llevar a cabo un análisis predictivo. Además, se identificó la necesidad de normalizar los datos para asegurar la comparabilidad y precisión de los modelos predictivos. En este proceso, se crearon nuevas variables basadas en estas normalizaciones, lo que permitió una mejor interpretación y uso de los datos en los análisis subsecuentes.

En la quinta fase, Evaluación, se llevaron a cabo dos rondas distintas de evaluación del modelo. Inicialmente, se aplicaron diversas métricas utilizando todas las variables disponibles para revisar los resultados de rendimiento, lo que facilitó la identificación de los modelos más precisos y eficaces. Dado que el análisis involucraba dos posibles resultados (0 y 1), se optó por emplear modelos adecuados para clasificación binaria, ya que estos permiten diferenciar entre dos categorías o clases específicas. En este caso, los modelos de clasificación binaria resultaron ideales para identificar la presencia o ausencia de hipertensión en los individuos analizados, proporcionando una predicción clara sobre el riesgo de esta condición [22]. Los métodos seleccionados fueron la regresión logística (RL), árboles de decisión (DT), k-vecinos más cercanos (KNN), Random Forest (RF), Bayes Naives y XGBoost (XGB).

Durante la evaluación, se consideraron métricas como la exactitud, que proporciona una visión general del desempeño del modelo, aunque puede ser engañosa en conjuntos de datos desbalanceados [23]. Se midió también la precisión, que evalúa la capacidad del modelo para identificar correctamente los casos de hipertensión (1) y no hipertensión (0) [24]. También se evaluó la sensibilidad, crucial para determinar la capacidad del modelo de detectar todos los casos positivos [24]. El F1-score, que combina precisión y sensibilidad, ofrece una visión más completa del rendimiento del modelo [25]. Asimismo, se utilizó el área bajo la curva (AUC), que proporciona una medida agregada del rendimiento del modelo a lo largo de todos los posibles umbrales de clasificación, añadiendo una capa adicional de evaluación de la efectividad del modelo [25].

En el contexto de este estudio, el F1-score se considera la mejor métrica de desempeño. Esto se debe a que el F1-score proporciona un equilibrio entre la precisión y la sensibilidad, lo que es crucial en estudios de salud pública donde es vital minimizar tanto los falsos positivos como los falsos negativos. Minimizar los falsos negativos es especialmente importante, ya que no detectar un caso de hipertensión podría tener graves consecuencias para la salud del individuo. Por otro lado, un alto número de falsos positivos podría llevar a un uso ineficiente de los recursos y una carga innecesaria sobre el sistema de salud.

Posteriormente, se llevó a cabo una segunda evaluación, esta vez utilizando únicamente las variables relacionadas con los macronutrientes y calorías, para determinar cómo estos factores específicos afectaban la predicción de la hipertensión.

En ambas etapas se aplicaron métricas estándar de rendimiento, lo que permitió evaluar de manera efectiva si los modelos alcanzaban los objetivos propuestos o si era necesario realizar ajustes para optimizar su desempeño. Esta evaluación detallada ayudó a identificar las fortalezas y limitaciones de cada enfoque, garantizando que las intervenciones futuras estuvieran mejor informadas y fueran más dirigidas.

Finalmente, la última fase fue el Despliegue. Una vez aceptado el modelo, se preparó la implementación final. Los resultados y hallazgos fueron documentados.

4. Resultados

Uno de los resultados que se obtuvieron fue un conjunto de datos limpio y de alta calidad listo para el análisis. Esta etapa fue una de las más exhaustivas debido a la necesidad de transformar variables significativas. Una tarea crucial fue el cálculo de las calorías y macronutrientes basados en la información proporcionada por el Cuestionario de recordatorio de 24 horas (CR24).

En la Figura 2, se presentan los resultados obtenidos con cada modelo. Es importante mencionar que para el entrenamiento de los datos se empleó la técnica de balanceo SMOTE, debido a que originalmente los datos no estaban equilibrados [26]. Esta técnica de balanceo sintético ayudó a mejorar la precisión de los modelos y a evitar sesgos en las predicciones, asegurando una representación más justa de todas las clases en el conjunto de datos.

Tabla 3. Resultados de los modelos evaluados.

Modelo	Exactitud %	AUC %		Precisión %	Sensibilidad %	F1-score %
RL	83	82	0	79	92	85
			1	89	72	80
DT	78	78	0	81	77	79
			1	75	80	78
KNN	79	79	0	88	70	78
			1	73	89	80
RF	90	90	0	90	92	91
			1	91	88	90
BN	66	67	0	80	48	60
			1	60	87	71
XGB	88	89	0	86	93	90
			1	92	84	87

El modelo de Random Forest (RF) demostró ser el más eficaz, logrando optimización significativa, esto mediante ajustes en el número de árboles y la profundidad máxima. Se probaron configuraciones de 100, 200 y 300 árboles con profundidades de 3, 5, 7 y 9, identificando que 200 árboles a una profundidad de 9 era el equilibrio ideal para maximizar la precisión y la generalización del modelo. Esto resultó en una exactitud y un AUC del 90%. La precisión fue del 90% para la clase 0 y del 91% para la clase 1, con sensibilidades del 92% y 88% respectivamente, y F1-scores del 91% para la clase 0 y del 90% para la clase 1. Este modelo es robusto y menos susceptible al sobreajuste comparado con los árboles de decisión individuales, adecuado para manejar grandes volúmenes de datos y múltiples características, y proporciona indicadores de la importancia de cada característica.

El modelo de XGBoost (XGB) fue meticulosamente ajustado en términos de profundidad máxima de los árboles, tasa de aprendizaje y número de árboles, explorando profundidades de 3, 5 y 7, tasas de 0.01, 0.1 y 0.2, y configuraciones de 100, 200 y 300 árboles para maximizar el rendimiento mientras se controlaba el sobreajuste. El mejor rendimiento se alcanzó con una tasa de aprendizaje de 0.2, una profundidad máxima de 7 y 100 árboles, lo que permitió que el modelo aprendiera rápidamente y se adaptara a las complejidades de los datos sin sobreajustarse. Este ajuste condujo a una exactitud del 88% y un AUC del 89%, con una precisión del 86% para la clase 0 y del 92% para la clase 1, sensibilidad del 93% para la clase 0 y del 84% para la clase 1, y F1-scores del 90% y 87% respectivamente. Este modelo es eficiente y rápido, ideal para grandes conjuntos de datos, e incluye términos de regularización que ayudan a prevenir el sobreajuste, además de proporcionar medidas de importancia de las características y manejar bien datos heterogéneos y no lineales.

El modelo de Regresión Logística (RL) fue cuidadosamente ajustado experimentando con diferentes valores del parámetro de fuerza de regularización, incluyendo

0.01, 0.1, 1, 10 y 100, para hallar el mejor equilibrio entre sesgo y varianza. El valor óptimo se determinó en 0.1, que proporciona suficiente penalización para mantener el modelo generalizable sin ser demasiado restrictivo. Este ajuste resultó en una exactitud del 83% y un AUC del 82%, con una precisión del 79% para la clase 0 y del 89% para la clase 1, sensibilidades del 92% para la clase 0 y del 72% para la clase 1, y F1-scores del 85% para la clase 0 y del 80% para la clase 1. Este modelo es altamente interpretable, ofreciendo coeficientes que indican la dirección y magnitud del impacto de cada característica en la probabilidad de hipertensión, eficiente en términos computacionales y adecuado para manejar grandes conjuntos de datos, facilitando el uso de regularización para evitar el sobreajuste.

El modelo de K-vecinos más Cercanos (KNN) fue ajustado meticulosamente, alterando el número de vecinos y el método de ponderación para evaluar su impacto en la clasificación final. Tras probar configuraciones de 3, 5, 7 y 9 vecinos con métodos de ponderación uniforme y por distancia, se determinó que la configuración óptima consistía en utilizar 7 vecinos con ponderación por distancia. Este ajuste permitió que el modelo priorizara a los vecinos más cercanos, lo que mejoró notablemente la precisión de la clasificación. Como resultado de estos ajustes, el modelo KNN alcanzó una exactitud del 79% y un AUC del 79%, con una precisión del 88% para la clase 0 y del 73% para la clase 1, sensibilidades del 70% para la clase 0 y del 89% para la clase 1, y F1-scores del 78% para la clase 0 y del 80% para la clase 1. Este modelo es conocido por su simplicidad de implementación y comprensión, siendo eficiente en conjuntos de datos pequeños y medianos, aunque su desempeño varía significativamente según el número de vecinos y la métrica de distancia empleada.

El modelo de Árbol de Decisión (DT) fue cuidadosamente ajustado para equilibrar su complejidad y evitar el sobreajuste. Se experimentó con diferentes profundidades máximas de árbol y números mínimos de muestras requeridas para dividir un nodo, probando profundidades de 3, 5, 7 y 9, y mínimos de muestras de 2, 5 y 10. La configuración que mejor manejó la varianza y previno el sobreajuste involucró una profundidad máxima de 7 y un mínimo de 10 muestras por nodo. Este ajuste estratégico permitió que el árbol capturara patrones esenciales sin aprender ruido innecesario. Como resultado, el modelo DT alcanzó una exactitud del 78% y un AUC del 78%, con una precisión del 81% para la clase 0 y del 75% para la clase 1, sensibilidad del 77% para la clase 0 y del 80% para la clase 1, y F1-scores del 79% para la clase 0 y del 78% para la clase 1. Este modelo es valorado por su facilidad de interpretación visual a través de la representación gráfica del árbol y su capacidad para capturar relaciones no lineales entre las características y la hipertensión, aunque requiere atención cuidadosa para evitar el sobreajuste.

Finalmente, el modelo de Naive Bayes (BN) tuvo el peor desempeño, con una exactitud del 66% y un AUC del 67%. La precisión para la clase 0 fue del 80% y para la clase 1 del 60%, con una sensibilidad del 48% para la clase 0 y del 87% para la clase 1. Los F1-scores fueron del 60% para la clase 0 y del 71% para la clase 1. Aunque este modelo es muy fácil de implementar y rápido para entrenar, debido a que asume independencia condicional entre las características y determina las probabilidades directamente de los datos, mostró dificultades significativas en la identificación correcta de la clase 0. Esta simplicidad también implica que no se requieren ajustes de hiperparámetros, lo que hace al Naive Bayes menos flexible pero extremadamente eficiente para el procesamiento rápido en grandes volúmenes de datos, aunque esta característica puede limitar su efectividad en escenarios donde las dependencias entre características son cruciales para hacer predicciones precisas.

La figura 2 ofrece una representación visual detallada del rendimiento de los modelos implementados en la predicción de hipertensión al utilizar las cinco métricas clave: Accuracy, AUC, Precision, Recall y F1-score. Cada uno de estos ejes representa una métrica específica, y cada línea del gráfico corresponde a un modelo particular, formando una figura que revela las fortalezas y debilidades de cada modelo en comparación con los demás.

Los modelos que muestran valores elevados en todas las métricas tienen líneas más amplias y puntos más cercanos al borde exterior del gráfico, lo que indica un rendimiento superior. Este fue el caso de RF y XGB, cuyos puntos en cada métrica están consistentemente cerca del límite externo. Esta proximidad al borde refleja un desempeño alto en todas las áreas evaluadas, tanto en la capacidad de discriminación entre clases (AUC), como en la exactitud (Accuracy), la precisión y la capacidad de recuperación de predicciones correctas (Recall), así como en el equilibrio entre precisión y recall (F1-score). La forma más amplia y el posicionamiento exterior de estos modelos en el gráfico destacan su robustez y fiabilidad para la predicción de hipertensión, como ya se habia mencionado.

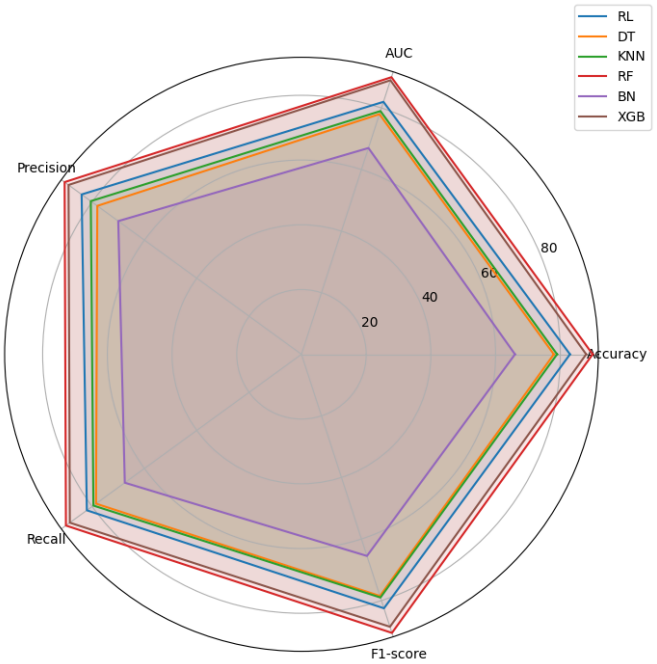

Figura 2. Comparación de Modelos en Predicción de Hipertensión (Gráfico de Radar)

En cambio, los modelos con puntos más próximos al centro, como BN, exhiben limitaciones en su rendimiento general. La cercanía de los puntos de BN al centro indica valores más bajos en varias métricas clave, particularmente en AUC y F1-score, lo que sugiere que este modelo tiene una capacidad limitada para clasificar correctamente las instancias de hipertensión. Además, un valor bajo en estas métricas implica que el modelo es menos consistente en mantener un buen balance entre precisión y sensibilidad, lo cual es fundamental en aplicaciones de salud donde la precisión en la predicción y la detección adecuada son esenciales para reducir riesgos.

Una vez examinados y probados los modelos propuestos, se determinó que el modelo Random Forest ofrecía el mejor desempeño en la predicción del riesgo de hipertensión. A partir de este modelo, se realizó un análisis de la importancia de las características. En la Figura 3, se presenta la importancia relativa de las variables en relación con la hipertensión.

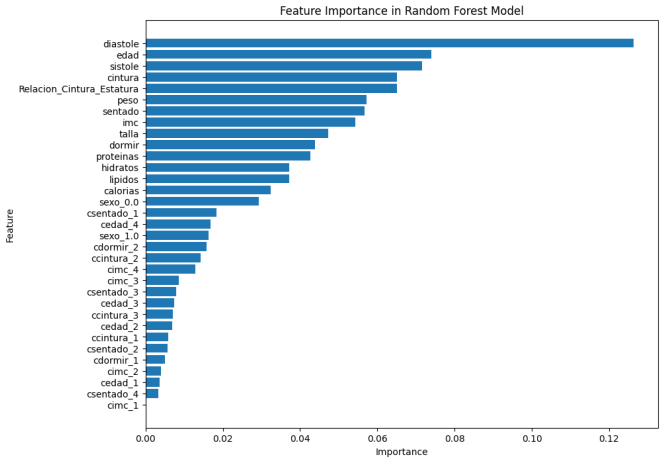

Figura 3. Importancia de características en el modelo Random Forest.

En el análisis realizado, se identificó que otros parámetros están más estrechamente relacionados con la predicción de la hipertensión, pero los estudiados en este proyecto también tienen una importancia significativa.

En cuanto a la importancia de los macronutrientes, se observó que las proteínas tienen una importancia de 4.26%, los hidratos de carbono 3.71% y los lípidos 3.71%. Estos resultados sugieren que, aunque los macronutrientes no son los factores más influyentes, su consumo tiene un impacto relevante en el riesgo de desarrollar hipertensión en adultos jóvenes en México.

El análisis de la importancia de las calorías mostró un valor de 3.23%. Esto indica que el consumo calórico total ejerce una influencia moderada en el riesgo de hipertensión en la población joven adulta mexicana. Aunque no es la característica más importante, su influencia no puede ser ignorada en el contexto del riesgo de hipertensión. En conjunto, estos hallazgos subrayan la necesidad de considerar múltiples factores nutricionales al evaluar el riesgo de hipertensión, aunque algunos parámetros pueden ser más determinantes que otros.

5. Discusión

Los resultados obtenidos en este estudio revelan importantes hallazgos sobre la predicción del riesgo de hipertensión en adultos mexicanos mediante el uso de modelos de aprendizaje automático y datos nutricionales. El Random Forest (RF) se destacó como el modelo más eficaz, no solo por su alta exactitud (90%) y capacidad predictiva, sino también por su habilidad para manejar grandes volúmenes de datos y múltiples características de forma robusta. Esto demuestra que los modelos basados en bosques aleatorios son menos propensos al sobreajuste en comparación con los árboles de decisión individuales, lo que es un desafío común en análisis complejos.

El análisis de importancia de características con RF mostró que, aunque los factores tradicionales como la edad y la presión arterial siguen siendo los más influyentes en la predicción de la hipertensión, los macronutrientes y la ingesta calórica también tienen un papel relevante. Este resultado es consistente con estudios previos que han indicado una relación entre la dieta y el riesgo de hipertensión, lo que refuerza la necesidad de incluir un enfoque nutricional integral en la evaluación de riesgo. En particular, las proteínas, carbohidratos y lípidos demostraron tener una influencia moderada, lo que sugiere que ciertos hábitos alimenticios pueden aumentar el riesgo de hipertensión en adultos jóvenes.

El modelo XGBoost (XGB) mostró un rendimiento cercano al de RF, con una exactitud del 88%. Sin embargo, su principal fortaleza radica en su capacidad para manejar datos heterogéneos y no lineales de manera eficiente, gracias a sus mecanismos de regularización que evitan el sobreajuste. Esto lo convierte en una opción especialmente adecuada para escenarios con conjuntos de datos complejos y variados, por lo que podría ser útil en investigaciones futuras que busquen optimizar modelos predictivos en el ámbito de la salud pública.

En contraste, el Naive Bayes (BN) presentó el rendimiento más bajo, con una exactitud del 66%, lo que indica que su simplicidad y las suposiciones de independencia entre las variables limitan su capacidad para captar relaciones más complejas en los datos. Esto confirma que su uso puede estar limitado a problemas más simples, y debe considerarse cuidadosamente al seleccionar modelos para estudios que involucren múltiples factores de riesgo.

Estos resultados abren nuevas posibilidades para investigar en mayor profundidad la interacción entre los factores nutricionales y la hipertensión, especialmente en poblaciones jóvenes. Además, el enfoque predictivo utilizado en este estudio puede servir de base para diseñar intervenciones de salud pública más personalizadas, que incluyan no solo los factores clínicos tradicionales, sino también la dieta y el estilo de vida. De esta manera, los modelos de aprendizaje automático como RF y XGB pueden ser herramientas útiles para mejorar la precisión en la prevención de enfermedades crónicas como la hipertensión.

6. Conclusión

Este estudio demuestra que los modelos de aprendizaje automático, como Random Forest y XGBoost, son herramientas eficaces para predecir el riesgo de hipertensión en adultos mexicanos. Estos modelos permiten integrar tanto factores clínicos como nutricionales en las evaluaciones de riesgo, lo que ofrece una nueva perspectiva para abordar esta enfermedad. Además, los hallazgos destacan la importancia de desarrollar estrategias preventivas más personalizadas que incluyan variables dietéticas. El enfoque metodológico presentado puede servir de base para futuras investigaciones y sugiere que la combinación de factores clínicos y nutricionales es clave para mejorar la prevención y el manejo de la hipertensión.

7. Agradecimientos

Este estudio no hubiera sido posible sin el apoyo constante y las oportunidades brindadas por la Universidad Autónoma de Zacatecas (UAZ) y el Instituto Politécnico Nacional (IPN). Estoy profundamente agradecido con la UAZ por su excelente ambiente académico y recursos que fueron esenciales para mi formación. Asimismo, extiendo mi gratitud al IPN por permitirme compatibilizar mis estudios con mi vida laboral, facilitando así mi desarrollo profesional y personal durante este periodo.

8. Referencias

[1] O. G. Montero Cadena, G. J. Guzmán Kure, R. C. Acosta Bravo, and M. B. Peñafiel Peñafiel, "Principales factores de riesgo de la hipertensión arterial," RECIMUNDO, vol. 7, no. 2, pp. 89–97, Jun. 2023, doi: 10.26820/recimundo/7.(2).jun.2023.89-97.

[2] J. M. Baglietto-Hernández, A. Mateos-Bear, J. P. Nava-Sánchez, P. Rodríguez-García, and F. Rodríguez-Weber, "Nivel de conocimiento en hipertensión arterial en pacientes con esta enfermedad de la Ciudad de México," Medicina Interna de México, vol. 36, no. 1, pp. 1–14, 2020, doi: https://doi.org/10.24245/mim.v36i1.2844.

[3] J. Plaza-Torres, A. Martínez-Sánchez, and R. Navarro-Suay, "Hábitos alimenticios, estilos de vida y riesgos para la salud. Estudio en una población militar," Sanidad Militar, 2022, doi: https://dx.doi.org/10.4321/s1887-85712022000200004.

[4] N. Díaz Wever, H. Herrera Mogollón, Z. Fajardo, and A. Galbán Chinchilla, "Consumo de macronutrientes y micronutrientes en adolescentes," TAYACAJA, vol. 4, no. 1, pp. 180–192, Jun. 2021, doi: 10.46908/tayacaja.v4i1.163.

[5] I. Campos-Nonato et al., "Prevalencia, tratamiento y control de la hipertensión arterial en adultos mexicanos: resultados de la Ensanut 2022," Salud Publica Mex, vol. 65, pp. 169– 180, 2023.

[6] L. Alcocer et al., "A reflection on the results of the ENSANUT 2022 on high blood pressure in Mexican adults," Cardiovascular and Metabolic Science, vol. 34, no. 3, pp. 85–93, 2023.

[7] S. Hasdeu, L. LAMFRE, P. CARO, and Federico. HORNE, "Revisión narrativa: modelos predictivos sobre la evolución de la pandemia por COVID-19," Rev Argent Salud Publica, vol. 12, p. 3, [Online]. Available: http://www.scielo.org.ar/scielo.php?script=sci_abstract&pid=S1853-810X20200003

00003&lng=es&nrm=iso&tlng=es

[8] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif Intell Med, vol. 34, no. 2, pp. 113–127, Jun. 2005, doi: 10.1016/j.artmed.2004.07.002.

[9] R. BELLAZZI and B. ZUPAN, "Predictive data mining in clinical medicine: Current issues and guidelines," Int J Med Inform, vol. 77, no. 2, pp. 81–97, Feb. 2008, doi: 10.1016/j.ijmedinf.2006.11.006.

[10] M. Araujo-Castro et al., "Predictive model of hypertension resolution after adrenalectomy in primary aldosteronism: the SPAIN-ALDO score," J Hypertens, vol. 40, no. 12, pp. 2486–2493, Dec. 2022, doi: 10.1097/HJH.0000000000003284.

[11] L. M. Ruilope et al., "PREDICTIVE PERFORMANCE AND CLINICAL UTILITY OF A NEW PREDICTIVE MODEL OF CARDIOVASCULAR RISK FOR YOUNG AND MIDDLE-AGED WORKING POPULATION," J Hypertens, vol. 36, no. Supplement 1, p. e221, Jun. 2018, doi: 10.1097/01.hjh.0000539624.20457.e0.

[12] C. Fernández-Labandera Ramos et al., "Estrategia clínica para reducir la enfermedad cardiovascular: integración de Iberscore y Life's simple 7 de la American Heart Association," Rev Esp Cardiol, vol. 71, no. Suplemento 1, p. 1306, 2018, [Online]. Available: http://www.revespcardiol.org/es-congresos-sec-2018-el-congreso-76-sesion-biomarcadores-escalas-riesgo-4438-estrategia-clinica-reducir-enfermedad-cardiovascular-52373

[13] Q. Luo, K. Bao, W. Gao, Y. Xiang, M. Li, and Y. Zhang, "Joint effects of depressive status and body mass index on the risk of incident hypertension in aging population: evidence from a nationwide population-based cohort study," BMC Psychiatry, vol. 23, no. 1, p. 608, Aug. 2023, doi: 10.1186/s12888-023-05105-z.

[14] A. D. Argoty-Pantoja, R. Velázquez-Cruz, J. Meneses-León, J. Salmerón, and B. Rivera-Paredez, "Triglyceride-glucose index is associated with hypertension incidence up to 13 years of follow-up in mexican adults," Lipids Health Dis, vol. 22, no. 1, p. 162, Sep. 2023, doi: 10.1186/s12944-023-01925-w.

[15] S. Kendale, P. Kulkarni, A. D. Rosenberg, and J. Wang, "Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension," Anesthesiology, vol. 129, no. 4, pp. 675–688, Oct. 2018, doi: 10.1097/ALN.0000000000002374.

[16] C.-D. Tseng, A. M.-F. Yen, S. Y.-H. Chiu, L.-S. Chen, H.-H. Chen, and S.-H. Chang, "A Predictive Model for Risk of Prehypertension and Hypertension and Expected Benefit After Population-Based Life-Style Modification (KCIS No. 24)," Am J Hypertens, vol. 25, no. 2, pp. 171–179, Feb. 2012, doi: 10.1038/ajh.2011.122.

[17] Martínez de Castro. Georgina Toussaint, A. B. Pérez Lizaur, H. R. García Hernández, and A. F. García Martínez, "Contenido nutrimental de los alimentos," in Nutrición y gastroenterología pediátrica, 2013.

[18] I.B.M., "Conceptos básicos de ayuda de CRISP-DM," in Obtenido de Guía de CRIPS-DM de IBM SPSS Modeler, 2021. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[19] D. J. Gottlieb, S. Redline, F. J. Nieto, C. M. Baldwin, A. B. Newman, H. E. Resnick, ... and N. M. Punjabi, "Association of usual sleep duration with hypertension: the Sleep Heart Health Study," Sleep, vol. 33, no. 8, pp. 1011-1018, 2010.

[20] T. M. Spruill, "Chronic psychosocial stress and hypertension," Current Hypertension Reports, vol. 12, pp. 10-16, 2010.

[21] D. Berrar, "Cross-Validation," in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[22] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation," Caspian Journal of Internal Medicine, vol. 4, no. 2, pp. 627-635, 2013.

[23] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," Electronics (Basel), vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/electronics8080832.

[24] M. Kuhn and K. Johnson, Applied predictive modeling. New York: Springer, 2013.

[25] J. A. Martínez Pérez and P. S. Pérez Martin, "La curva ROC," Medicina de Familia. SEMERGEN, vol. 49, no. 1, p. 101821, Jan. 2023, doi: 10.1016/j.semerg.2022.101821.

[26] V. Morales-Oñate, L. Moreta, and B. Morales-Oñate, "SMOTEMD: UN ALGORITMO DE BALANCEO DE DATOS MIXTOS PARA BIG DATA EN R," Perfiles, vol. 1, no. 24, 2020.