5. Estrategia de predicción para mantener la calidad de vida en la región Tula, Hidalgo

Silvia Soledad Moreno Gutiérrez*
Yoali Trejo Ambrosio**

Resumen

La calidad de vida (CV) consiste en la percepción del individuo respecto al cumplimiento de sus metas y expectativas con base en su sistema de valores, acceso a bienes y servicios e importancia que otorga a la salud y diversos aspectos sociales, materiales y ambientales. Según su interacción con estos factores el individuo adquiere autonomía e independencia como cualidades esenciales para incrementar su bienestar y sus condiciones de vida. No obstante, la CV como fenómeno complejo y multidimensional en zonas industriales ha sido afectada por diversas causas, entre ellas, la contaminación ambiental que amenaza la vida. La zona metropolitana (ZM) de Tula, Hidalgo, por ejemplo, hoy reconocida por su desarrollo económico y productivo se ubica entre las más contaminadas del planeta. Con base en lo anterior y en el potencial preventivo de las tecnologías inteligentes, se expone la aplicación de los modelos neuronales artificiales como estrategias tecnológicas de fortalecimiento de la toma de decisiones desde un enfoque predictivo basado en la opinión de 2 000 familias de la zona metropolitana, con datos recabados en el año 2022 mediante el instrumento WHOQOL-BREF enfocado

* Doctora en Planeación Estratégica y Dirección de Tecnologías por la Universidad Popular Autónoma del Estado de Puebla (UPAEP), México. Obtuvo la Maestría en Ciencias Computacionales, Especialidad en Computación Inteligente y Licenciatura en Computación en la Universidad Autónoma del Estado de Hidalgo (UAEH), México. https://orcid.org/0000-0002-8957-3707

** Ingeniero de software egresada de la Escuela Superior de Tlahuelilpan de la Universidad Autónoma del Estado de Hidalgo (UAEH), México.

en salud física, psicológica, relaciones sociales y medio ambiente. El desempeño de los modelos inteligentes demostró capacidad para representar el fenómeno y predecir el puntaje de CV por categoría con resultados favorables: MSE .011 y .022, RMSE .01062 y .075 y R2 .9982 y .999. Por lo tanto, se consideran herramientas estratégicas de alto potencial para identificar factores clave con implicaciones favorables en la tarea de conservar la CV.

Palabras clave: *redes neuronales artificiales, calidad de vida, zona industrial Tula.*

Introducción

La calida de vida (CV) es una percepción del individuo basada en aspectos propios y de su interacción con las personas y el ambiente. A su vez, esta interacción de forma permanente, tanto en términos favorables y/o desfavorables, representa una labor que contribuye al fortalecimiento de dos habilidades fundamentales: la autonomía y la independencia del individuo, habilidades que resultan esenciales para incrementar el bienestar y mantener o elevar las condiciones de vida (WHO, 1999; Ordoñez-Aquino *et al.,* 2023).

La CV constituye un fenómeno que considera fundamentalmente la razón de ser de las personas (García, 2020), no obstante, esta condición podría verse afectada por múltiples razones. Una de ellas es el aspecto ambiental al constituir uno de los problemas de impacto global con mayor relevancia debido al efecto adverso que ha provocado sobre los sectores de la sociedad a raíz de la emisión excesiva de gases de efecto invernadero (GEI). Esto da lugar a diversas problemáticas para la población que ha visto afectada su CV (Rico *et al.,* 2022). De hecho, aquellas poblaciones residentes de zonas contaminadas se consideran vulnerables debido al efecto adverso ocasionado a su salud física, psicológica, social y ambiental. La zona metropolitana (ZM) de Tula, Hidalgo, por ejemplo, ha alcanzado un alto desarrollo económico y productivo, no obstante, se encuentra entre las más contaminadas del planeta, con un alto índice de enfermedades crónicas que continúa en aumento (Moreno-Gutiérrez *et al.,* 2024). Por lo anterior, la premisa actual se ha enfocado en construir estrategias que contribuyan a mantener la CV

en general ante el fenómeno global del cambio climático (CC) derivado de la presencia de GEI, cuyo impacto se observa de manera permanente, sobre todo en zonas industriales.

La literatura expone diversas estrategias para afrontar el problema ambiental y sus consecuencias en la salud. No obstante, la presencia de la inteligencia artificial (IA) ha sido frecuente por su potencial para apoyar la prevención de situaciones diversas mediante la aplicación de técnicas de aprendizaje automático (AA) que han cobrado relevancia por su alta precisión y potencial en las tareas de adelantarse al futuro y construir mecanismos destinados a evitar o reducir los riesgos, así como la incertidumbre (Ishaq *et al.,* 2021). Con base en lo anterior el presente documento expone hallazgos en cuanto a la aplicación de diferentes modelos de AA, tanto de clasificación como de regresión, entre ellos modelos neuronales orientados a fortalecer la toma de decisiones con base en los datos recabados de la encuesta de opinión aplicada a habitantes de la ZM, considerando cuatro dimensiones que son: salud física, salud psicológica, relaciones sociales y medio ambiente. El instrumento aplicado se conoce como WHOQOL-BREF, validado por la Organización Mundial de la Salud (OMS) y orientado a cuantificar la CV de las personas a través de su opinión en las dimensiones antes mencionadas (Hidalgo-Rasmussen *et al.,* 2021).

Respecto al desempeño de los modelos con mejores resultados, es decir, las redes neuronales artificiales (RNA), se demostró su capacidad para representar el fenómeno y predecir el puntaje de CV por dimensión alcanzando resultados que fueron favorables: MSE .011, RMSE .01062 y R2 .9982 para el primero, mientras que el segundo alcanzó .022, y .075 y .999, respectivamente. Por lo tanto, se consideran herramientas estratégicas potenciales debido a que son capaces de identificar factores clave de la percepción de la población, lo cual contribuye al mejoramiento de la toma de decisiones enfocada en la atención de las personas y el incremento de la CV.

Marco teórico y revisión de la literatura

Para el análisis de datos el AA, mejor conocido como machine learning (ML) constituye una disciplina valiosa que permite a través de algoritmos

identificar patrones y realizar predicciones. Su capacidad inherente para descubrir patrones ocultos en los datos históricos ha impulsado la comprensión de fenómenos complejos (Hastie *et al.,* 2021). Este proceso puede ser guiado por diversas metodologías, entre ellas el CRISP (Cross-Industry Standard Process for Data Mining) para el desarrollo de modelos predictivos mediante 6 fases con un enfoque iterativo (Espinosa-Zúñiga, 2020):

1. *Entendimiento del negocio.* Se identifican la problemática, los objetivos y requerimientos del proyecto.
2. *Entendimiento de los Datos.* Se analizan exhaustivamente los datos históricos disponibles para asegurar su comprensión y relevancia.
3. *Preparación de datos.* se enfoca en identificar y transformar los datos relevantes para el modelo.
4. *Modelado.* Consiste en la aplicación de algoritmos de AA para construir modelos predictivos a partir de los datos preparados y con base en los objetivos establecidos.
5. *Validación.* Juega un papel fundamental en determinar el rendimiento y capacidad del modelo para generalizar y predecir con precisión a partir de nuevos datos.
6. *Implementación.* Se traducen los resultados del análisis de datos en acciones prácticas. Involucra el desarrollo de aplicaciones, la automatización de procesos y la integración de sistemas basados en los modelos creados (Timarán-Pereira *et al.,* 2019; Batti *et al.,* 2019).

Algunos de los algoritmos de AA se exponen en el presente apartado con el propósito de ofrecer claridad al lector:

- Bosque aleatorio o *Random Forest* (RF) es un método que integra árboles de decisión independientes para generar predicciones, utiliza técnica de bagging para entrenar múltiples árboles mediante muestras bootstrap, cada una se formada por subconjuntos de datos del conjunto de entrenamiento elegidos aleatoriamente y manteniendo el tamaño del original. La combinación de diferentes predicciones mejora la generalización del modelo, de igual forma, se reduce el riesgo de sobreajuste e incrementa la precisión (Ishaq *et al.,* 2021).

- Árbol de decisión extra o *Extra Tree Decision* (ETD), es una variante de los árboles de decisión y comparte similitudes con el RF. Se orienta a la elección de características y umbrales en forma aleatoria por cada nodo, como resultado, incrementa su variabilidad y diversidad.
- Máquina de soporte vectorial (MVS) es un algoritmo que está basado en modelos matemáticos, adecuada para regresión y clasificación, construye hiperplanos de alta dimensión en busca de la mejor separación lineal. Emplea funciones de kernel para mapear espacios superiores y abordar relaciones complejas (Rojas-Rubio *et al.,* 2022).
- Red neuronal artificial, constituye un algoritmo inteligente que permite modelar fenómenos de comportamiento no lineal y consta de múltiples capas y nodos llamados neuronas, los cuales se encuentran interconectados para procesar datos mediante funciones de activación (Huang *et al.,* 2020).
- Perceptron Multicapa (MLP, por sus siglas en inglés), es un modelo de RNA, el cual podría resolver problemas de clasificación o regresión (Rojas-Rubio *et al.,* 2022).
- La función de activación ReLU (Rectified Linear Unit) es de aplicación frecuente en modelos recientes. Posee la capacidad para mitigar problemas de desvanecimiento de gradientes con eficiencia computacional (Grande *et al.,* 2023)

Respecto a la revisión de la literatura y la construcción de alternativas innovadoras para abordar el fenómeno de la CV, durante la revisión de la literatura no se encontraron trabajos que apliquen AA para valorarla en la ZM, sin embargo, diversos estudios la han explorado y se exponen en este apartado.

Un estudio aplicó un análisis de componentes principales para sintetizar variables económicas, sociales y físicas en tres dimensiones clave que explican la calidad de vida: 1) estabilidad económica, 2) composición de los hogares y 3) conciencia ambiental, a través de las cuales se construyó un índice de bienestar, el análisis identificó patrones heterogéneos en Pachuca, Hidalgo, resaltando la necesidad de construir políticas públicas focalizadas para el desarrollo urbano (Acuña *et al.,* 2021).

En salud, para las personas con diabetes mellitus tipo 2, se analizó el impacto de la enfermedad en la CV, el acceso a la atención médica, los efectos emocionales y sociales (Guzman-Acostupa y Zarate-Anastares, 2023). De igual forma, Ortiz *et al.,* (2022) coinciden en el impacto de la enfermedad en aspectos cognitivos de los adultos mayores en la CV. Flores-Ramírez *et al.,* (2020) por su parte, emplearon el cuestionario WHOQOL-BREF y PROQOL-IV para evaluar la CV de profesionales de enfermería en áreas críticas. Los hallazgos exponen la relevancia del aspecto psicológico y contextual. Otro estudio analizó la percepción respecto al control de peso por género en estudiantes universitarios adolescentes en México y su influencia en la CV.

Por su parte, Moreno-Gutiérrez *et al.,* (2024), desarrollaron modelos de predicción de AA para apoyar el diagnostico médico predictivo de enfermedades crónicas en habitantes de la zona. Obtuvieron resultados de alta precisión luego de aplicar modelos de clasificación y regresión que alcanzaron un rendimiento superior al 90%.

La CV, desde el enfoque predictivo con AA, ha sido escasamente estudiada en la ZM, no obstante, estas técnicas de IA son herramientas idóneas para analizar el futuro a través de los datos y de esta manera contribuir a la prevención de problemática relacionada con la CV.

Metodología

El presente trabajo consistió en construir diversos modelos de AA para la predicción de puntaje de CV, considerando los datos referentes a las percepciones de los residentes de la ZM de Tula de Allende, Hidalgo. Considerando la relevancia de los resultados, únicamente se exponen los modelos que mostraron un mejor desempeño, siendo estos los de RNA. Siguiendo la metodología para ciencia de datos CRISP-DM, se desarrollaron las fases que se exponen en el presente apartado para cada modelo.

Fase 1

A nivel global la puntuación máxima de CV alcanzó un máximo de 87.9 puntos. En el caso de México, para el año 2020 este índice logró una puntuación de 45.8 y ubicó al país en el lugar 70 global, lo cual es motivo de preocupación. No obstante, la Organización para la Cooperación y el Desarrollo Económico (OCDE) lo ubicó en el lugar 36 (A.K. García, 2020).

En cuanto al estado de Hidalgo, lugar donde se encuentran diferentes zonas industriales entre ellas la ZM de Tula, ha sido reconocidp tanto por su importancia económica como por su excesiva emisión de GEI. En las décadas recientes ha tenido logros y desafíos en el ámbito de la CV (Márquez, 2020). El Índice de Desarrollo Humano Municipal se clasifica como alto con 0.753 (PNUD, 2023), no obstante, la esperanza de vida en 2020 fue inferior al promedio nacional con 74.1 años (Salda, 2020). Los indicadores de desarrollo humano y salud superan la media, sin embargo, la seguridad es un desafío en términos sociales (SESNSP, 2021).

Fase 2

En cuanto a la comprensión de datos, la fase inicia con la identificación estratégica de municipios que integran la ZM de Tula. Estos son: Tula de Allende, Atitalaquia, Tlahuelilpan, Tepetitlan, Tlaxcoapan y Atotonilco, cada uno reportando altos índices de contaminación ambiental. La recopilación de los datos inició con la aplicación del cuestionario WHOQOL-BREF y consistió en una actividad que reunió un total de 2000 participantes. Una vez reunido el banco de datos, se llevó a cabo un análisis exhaustivo para comprender su comportamiento y explorar cada una de las 26 variables que integran el instrumento aplicado, las cuales se organizan en cuatro dimensiones: salud física, salud psicológica, relaciones sociales y medio ambiente (ver tabla 1).

Tabla 1. *Distribución de los ítems con cálculo de resultados por dominio*

	Dominio	Variables: Cambios [(1 = 5) (2 = 4) (3 = 3) (4 = 2) (5 = 1)]		Cálculo de resultados (Score 4-20)
INTERNOS	Salud física	03	ImpedimentosDolorFísico	$\frac{(3+4+10+15+16+17+18)\times 4}{7}$
		04	TratamientoMédico	
		10	Energía	
		15	Desplazamiento	
		16	Sueño	
		17	ActividadesHabil	
		18	CapacidadTrabajo	
	Salud psicológica	05	VidaDisfrute	$\frac{(5+6+7+11+19+26)\times 4}{6}$
		06	VidaSentido	
		07	Concentración	
		11	AceptaApariencia	
		19	SelfEsteem	
		26	SentimientosNegativos	
EXTERNOS	Relaciones sociales	20	RelacionesPersonales	$\frac{(20+21+22)\times 4}{3}$
		21	VidaSexual	
		22	ApoyoSatisf	
	Medio ambiente	08	VidaSeguridad	$\frac{(8+9+12+13+14+23+24+25)\times 4}{8}$
		09	SaludAmbiente	
		12	Dinero	
		13	Información	
		14	ActividadesOcio	
		23	CondicionesVivienda	
		24	ServiciosSanit	
		25	ServiciosTransp	

Fuente: WHOQOL (2012).

Aplicando el ambiente de desarrollo de Jupyter Notebook, se efectuó la conversión del archivo de datos a formato .csv con codificación UTF-8, con el propósito de asegurar la integridad y comprensión adecuadas. El análisis de correlación expresó la relación entre las variables con las 4 variables objetivo en este caso: salud física, salud psicológica, salud social y medio ambiente, las cuales fueron construidas mediante los puntajes recabados de las variables que componen cada dimensión.

Según demostraron los datos, en la población predominan las mujeres con una ventaja de 441 personas. Se identificó que el grupo con la mejor CV, en términos de género, son los hombres, representando un 13.83%, en comparación con las mujeres que alcanzaron un 12.70%. En relación con

las afecciones presentes, se encontró que los problemas respiratorios afectaron al 1.87% de la población, seguidos de la diabetes mellitus con un 1.77% y, por último, la hipertensión arterial con un 1.32%. Estas cifras suman al 19.05% de las personas que declararon tener algún tipo de condición médica adversa.

Es interesante notar que la presencia de una enfermedad influye en la percepción de la CV, ya que el 0.093% de las personas con afecciones reportaron tener una buena opinión. La autoestima demostró una fuerte dependencia con las relaciones personales, la capacidad laboral y la habilidad para llevar a cabo las actividades cotidianas.

Fase 3

Esta fase se concentra en transformar los datos con el propósito de mejorar su calidad y, con ello, el rendimiento del modelo. Los datos fueron categorizados como se muestra en la siguiente tabla:

Tabla 2. *Factores sociodemográficos categorías*

Variable	Descripción	Categoría
Sexo	Sexo	0 Hombre, 1 Mujer
Nacimiento	Año de nacimiento	_ _ _ _
Nivel de estudios	Grado de estudios	0 Ninguno, 1 Primarios, 2 Medios, 3 Universitarios
Estado civil	Estado civil	1 Soltero/a, 2 Separado/a, 3 Casado/a, 4 Divorciado/a, 5 En pareja, 6 Unión libre, 7 Viudo/a
Enfermo	¿Está enfermo?	0 No, 1 Si
Enfermedad, problema	Si está enfermo, ¿qué padece?	0 Ninguno, 1 Enfermedad/Problema, 2 Diabetes, 3 Diabetes y rodilla, 4 Diabetes e Hipertensión, 5 Hipertensión, 6 Parkinson; Hipertensión Arterial; Problemas Cardiacos, 7 Hipertensión y Artrosis Reumatoide, 8 Hipertensión y problemas gástricos, 9 Problemas gástricos, 10 DM e Hipertensión, 11 Hipotiroidismo e Hipertensión, 12 Hipertiroidismo, 13 DM, 14 DM; HTA;EPOC, 15 Alcoholismo, 16 Problemas de vías respiratorias, 17 Sinusitis, 18 Alergias, 19 Rinitis, 20 Asma, 21 Problemas cardiacos, 22 Hepatitis, 23 Insuficiencia renal, 24 Hiperuricemia, 25 Artritis, 26 Fractura, 27 EPOC; Columna, 28 Nervio Asiático; Lumbalgia, 29 Osteoporosis,

Enfermedad problema	Si está enfermo, ¿qué padece?	30 Osteoporosis y Vejiga caída, 31 Retención a insulina, 32 Conjuntivitis, 33 Cáncer Páncreas, 34 Anemia, 35 Sobrepeso/Obesidad, 36 Presión alta, 37 Migraña, 38 Estrés, 39 Depresión, 40 Insomnio, 41 Estilo de vida, 42 Covid; HTA, 43 Nob, 44 Mo
Calidad de Vida Calif	Evaluación subjetiva del encuestado	0 Sin respuesta, 1 Muy mala, 2 Regular, 3 Normal, 4 Bastante buena, 5 Muy buena
Salud Satisf	Satisfacción de salud presente	0 Sin respuesta, 1 Muy insatisfecho/a, 2 Un poco insatisfecho/a, 3 Lo normal, 4 Bastante satisfecho/a, 5 Muy satisfecho/a

Fuente: elaboración propia con datos de (WHOQOL, 2012).

Durante esta fase las variables se dividieron en: numéricas, ordinales y nominales, se ajustaron valores nulos en la variable nacimiento con el valor de la moda, se mantuvo el nivel educativo más alto de cada persona. 'sexo' y 'enfermo' se transformaron en binarias, en la variable 'enfermedadProblema' las enfermedades se agruparon para reducir categorías de 78 a 44.

El banco de datos inicial no incluía los puntajes de CV según WHOQOL-BREF, por ello, se calculó la variable objetivo como numérica y como categórica con el fin de realizar un análisis comparativo del rendimiento de los modelos.

Las 26 variables se agruparon en cuatro dominios (ver tabla 2). Para cada uno se sumaron los puntajes de los registros de las variables que los componen, tal como se mencionó previamente.

Tabla 3. *Transformación de puntaje por dominio en una escala de 0 a 100*

S		Dominio 2		Dominio 3		Dominio 4	
Puntaje	**Transformación**	**Puntaje**	**Transformación**	**Puntaje**	**Transformación**	**Puntaje**	**Transformación**
7	0	6	0	3	0	8	0
8	6	7	6	4	6	9	6
9	6	8	6	5	19	10	6
10	13	9	13	6	25	11	13
11	13	10	19	7	31	12	13
12	19	11	19	8	44	13	19
13	19	12	25	9	50	14	19
14	25	13	31	10	56	15	25
15	31	14	31	11	69	16	25
16	31	15	38	12	75	17	31
17	38	16	44	13	81	18	31
18	38	17	44	14	94	19	38
19	44	18	50	15	100	20	38

20	44	19	56	21	44
21	50	20	56	22	44
22	56	21	63	23	50
23	56	22	69	24	50
24	63	23	69	25	56
25	63	24	75	26	56
26	69	25	81	27	63
27	69	26	81	28	63
28	75	27	88	29	69
29	81	28	94	30	69
30	81	29	94	31	75
31	88	30	100	32	75
32	88			33	81
33	94			34	81
34	94			35	88
35	100			36	88
				37	94
				38	94
				39	100
				40	100

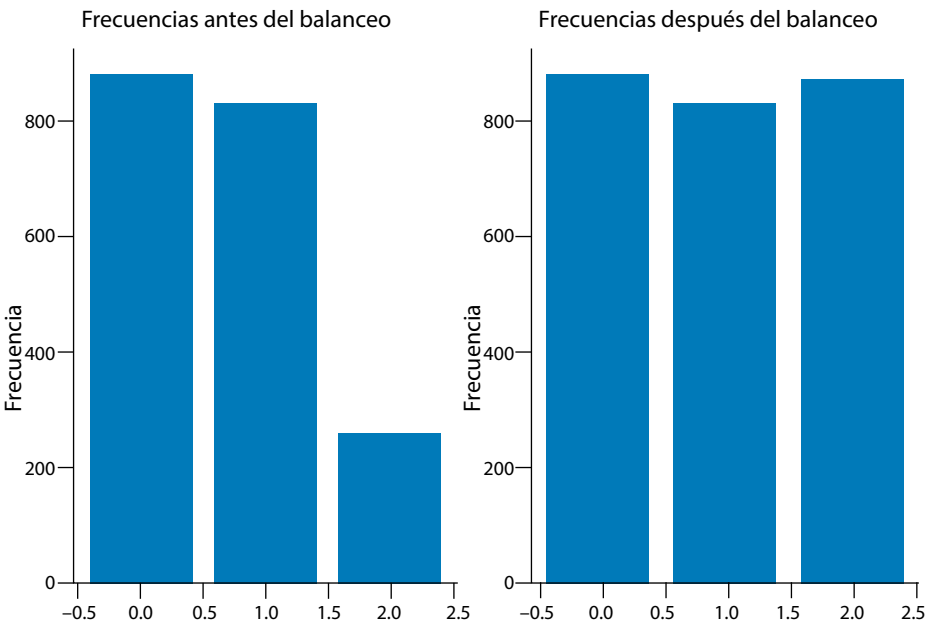
Fuente: elaboración propia con datos de (WHOQOL, 2012).

En el primer enfoque se aplicó un modelo de clasificación y se calculó una puntuación de CV, transformando las puntuaciones de cada dimensión a una escala de 0 a 100 según la tabla 3. eEstos resultados se dividieron en tres categorías: "ninguna calidad de vida" (0) para puntuaciones inferiores a 50, "calidad de vida media" (1) para puntuaciones entre 51 y 80, y "buena calidad de vida" (2) para puntuaciones superiores a 80, estos valores se agruparon en la variable 'CV'. Con el segundo enfoque se crearon cuatro variables numéricas: 'salud_física', 'salud_psicológica', 'relaciones_sociales' y 'medio_ambiente', estas variables capturaron las puntuaciones de cada uno de los dominios evaluados en el cuestionario WHOQOL-BREF. Cada una representa la puntuación obtenida, lo que permitió un análisis detallado y descriptivo de cada dimensión.

En cuanto a la normalización de los datos, para el enfoque 1 se aplicó la StandardScaler, adecuada en caso de distribuciones no gaussianas o valores atípicos, obteniendo datos con media de 0 y desviación estándar de 1. En el enfoque 2 se aplicó z-score, el cual estandariza los datos utilizando la media y la desviación estándar, adecuada para modelos de regresión.

Se realizó el balanceo de datos con el propósito de evitar sesgos en el rendimiento del modelo. Se utilizó una técnica de sobre muestreo para el caso de la clasificación (ver figura 1).

Figura 1. *Balanceo de datos categóricos*



Fuente: elaboración propia.

Fase 4

Para la construcción del modelo se aplicaron diferentes algoritmos tanto de clasificación como de regresión.

Los algoritmos de clasificación se exponen en este apartado. El algoritmo RF se construyó a través de un total de 700 árboles con profundidad máxima de 10 y se recurrió a la validación cruzada con 10 divisiones. Con esta información se creó una nueva instancia del clasificador RF, utilizando los valores de hiperparámetros óptimos, ya mencionados. Este enfoque permitió aprovechar al máximo los datos de entrenamiento al dividirlos en 10 partes, de las cuales una se reservó como conjunto de validación mientras

que las 9 restantes se emplearon para entrenamiento. La métrica de rendimiento se calculó mediante el promedio de los resultados de las 10 épocas de validación cruzada.

La optimización se realizó mediante el uso de la técnica GridSearchCV, que permitió explorar exhaustivamente un conjunto diverso de combinaciones de hiperparámetros para identificar los de mejor desempeño. Previamente se eligieron tres posibles valores para el número de árboles: 300, 500 y 700, con la intención de explorar un rango diverso.

– Algoritmo ETD, para este caso se realizó la optimización del rendimiento del modelo aplicando la técnica del caso anterior, es decir, RF, la cual explora diferentes combinaciones de parámetros para identificar los óptimos para el modelo.

Se exploraron varios aspectos clave: 1) número de estimadores 50, 100 y 150, estas propuestas representan la cantidad de árboles que se construirán en el bosque de clasificación, 2) profundidad máxima de los árboles, donde se consideraron tres alternativas, ninguna restricción, 5 niveles y 10 niveles, esta variación permite entender hasta qué punto los árboles deben explorar la estructura de los datos, 3) número mínimo de muestras para dividir un nodo, donde se probaron 3 opciones 2, 5 y 10, la variación en este número permite adaptar la complejidad del modelo según la disponibilidad de datos. Después de la búsqueda de hiperparámetros, los valores seleccionados fueron: 150 estimadores, profundidad máxima de 5 niveles y un mínimo de 5 muestras para dividir un nodo.

– MSV, en este contexto, se desarrolló una instancia del clasificador MVS con un kernel de base radial con el objetivo de efectuar una separación no lineal de los datos. Esta elección de kernel es crucial, ya que permite al modelo tratar con problemas donde las clases no se pueden separar de manera lineal en el espacio original de características.

Durante el entrenamiento, el algoritmo ajusta los parámetros del modelo para encontrar la mejor separación entre las clases en el espacio trans-

formado. Esta separación óptima se logra al maximizar el margen entre las clases y minimizar la clasificación incorrecta (Borja-Robalino *et al.,* 2020).

– MLP. Se procedió a la aplicación del algoritmo, el modelo se instanció con los hiperparámetros que se mencionan: tres capas ocultas con 150, 100 y 50 neuronas respectivamente, 500 épocas, y la función de activación ReLU. Esta función es comúnmente considerada en las capas ocultas de RNA debido a su capacidad para manejar problemas de activación no lineales (Natalia *et al.,* 2023).

Se entrenó aplicando el optimizador Adam, el cual es popular en el aprendizaje profundo debido a su eficiencia en la adaptación de tasas de aprendizaje.

Con el propósito de evitar el sobreajuste del modelo, se aplicó la técnica de *early stopping,* la cual permite habilitar la detención temprana del proceso de entrenamiento. Es decir, el entrenamiento se detendrá si no se observa una mejora en la función de pérdida ante el conjunto de validación (Bentoumi *et al.,* 2022).

Por otra parte, los algoritmos de regresión aplicados consistieron en RNA, los cuales se exponen a continuación:

RNA, *versión 1.* Tal como en casos anteriores, previo a la construcción de modelo se realizó un proceso de escalado y normalización mediante Z-score. La creación del modelo se lleva a cabo utilizando la biblioteca Keras, que ofrece una interfaz intuitiva para construir redes neuronales. La red consta de cuatro capas ocultas, cada una con la misma cantidad de neuronas que la dimensión de los datos de entrenamiento, aplicando función de activación ReLU en todas las capas ocultas para introducir no linealidades en la red. Se compila utilizando el optimizador Adam, además, de la regularización L2 en las capas ocultas para evitar el sobreajuste. La función de pérdida fue error cuadrático medio (MSE, por sus siglas en inglés).

Como se menciona en los casos previos, la técnica de GridSearchCV también fue aplicada para la identificación de los mejores hiperparametros, posteriormente, el modelo fue entrenado.

Versión 2. La normalización de los datos se realizó mediante MinMaxScaler y se estableció un modelo secuencial que consta de varias capas ocultas con función de activación ReLU. La capa de salida tiene 4 neuronas igual al número de variables que se desea predecir. Se aplicó el optimizador adam, el cual ajustó automáticamente la tasa de aprendizaje durante el entrenamiento, lo que aceleró la convergencia.

Para la función de pérdida se estableció MSE, pues se consideró apropiado para problemas de regresión, ya que penaliza las diferencias entre las predicciones y los valores reales al cuadrado. De igual forma se aplicó *early stopping* para combatir el sobreajuste del modelo.

Fase 5

Para los modelos de clasificación la validación del rendimiento se efectuó mediante la construcción de matrices de confusión. Estas matrices permitieron un análisis detallado de métricas como la exactitud, la precisión, la sensibilidad y la especificidad (Düntsch *et al.,* 2019). Los modelos demostraron ser apropiados al evidenciar un sólido desempeño, no obstante, algunos modelos superaron a otros, como se muestra en el apartado de los resultados.

Resultados

Respecto a los modelos de clasificación, los resultados se muestran en la tabla 4.

Tabla 4. *Resultados de los modelos categóricos*

	Validación		
Modelo	**Precisión**	**Sensibilidad**	**F1-score**
MLP	90%	90%	90%
RF	91%	91%	91%
SVM	90%	90%	90%
ETD	92%	92%	92%

Fuente: elaboración propia.

El mejor resultado se expone en la tabla donde se observa que el ETD alcanzó un mejor desempeño. No obstante, los modelos en general se consideran adecuados para la predicción de CV en las categorías antes mencionadas: 0 para ninguna calidad de vida, 1 para la calidad de vida media y 2 para una buena calidad de vida.

Respecto a los modelos de regresión, los resultados se muestran a continuación.

En la RNA, versión 1 los resultados se consideran favorables y se muestran en la tabla 5:

Tabla 5. *Valores promedio obtenidos por validación cruzada*

Métrica	Resultado
Valor promedio de la pérdida	1.38806653793
Valor promedio de la pérdida de validación	1.38806653793

Fuente: elaboración propia.

En cuanto a las métricas de validación del rendimiento, se obtuvieron resultados que demuestran un buen desempeño y se exhiben en la tabla 6.

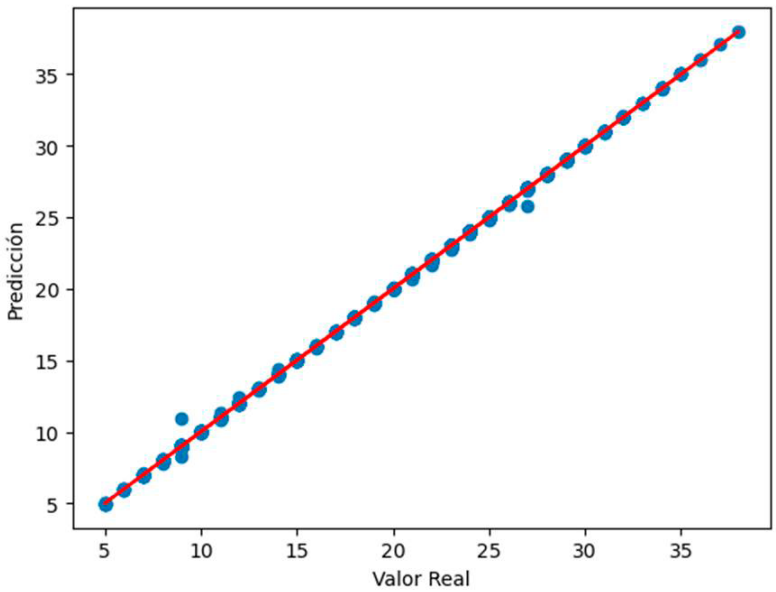
Tabla 6. *Métricas de evaluación del rendimiento, RNA versión 1*

Métrica	Resultado
MSE	0.27932153
RMSE	0.52850878
R2	0.96014878

Fuente: Elaboración propia.

Los resultados se exponen de forma gráfica en la figura 2.

Figura 2. *Gráfica real/predicción*



Fuente: Elaboración propia.

Para la RNA, versión 2, se muestran los resultados de pérdida y precisión (ver figura 3).

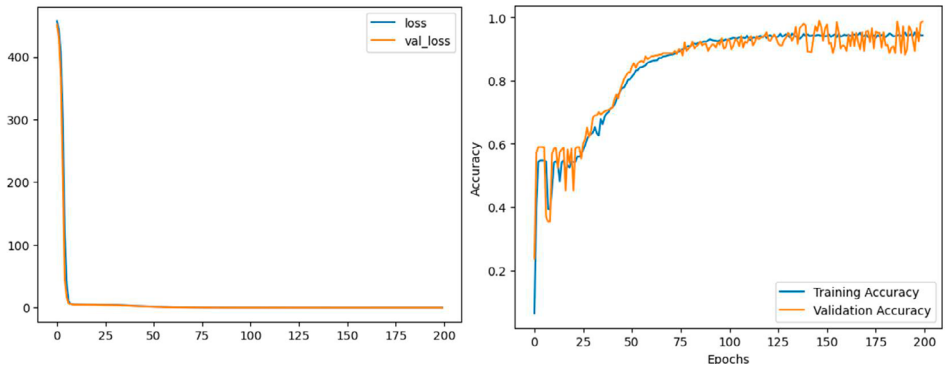
Tabla 7. *Valores de métricas de pérdida y precisión*

Métrica	Resultado
Pérdida	0.008788
Pérdida validación	0.011293
Precisión	0.943350
Precisión validación	0.987715

Fuente: Elaboración propia.

De forma gráfica y para mayor comprensión, estos resultados se exponen en la figura 3.

Figura 3. *Gráfica de pérdida/precisión*



Fuente: Elaboración propia.

Las métricas de evaluación incluyeron el MSE, la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Un valor más bajo de MSE indica un mejor rendimiento del modelo, el RMSE se interpreta en el contexto del problema y un valor de R^2 más cercano a 1 indica un mejor ajuste del modelo a los datos de prueba (Chicco *et al.*, 2021; Hodson, 2022).

Tabla 8. *Resultados de los modelos de regresión*

	Validación		
Modelo	**MSE**	**RMSE**	**R2**
RNA V_1	0.2793	0.5285	0.9601
RNA V_2	0.0112	0.1062	0.9982

Conclusiones

Los modelos aplicado de AA se consideraron adecuados para la predicción de CV tanto en aspectos de clasificación como de regresión. La diferencia en cuanto a desempeño en el caso de los modelos de clasificación evidencia una diferencia mínima entre cada algoritmo, siendo el mejor el ADT.

Respecto a los modelos neuronales para regresión, es decir, a través de los cuales se obtiene el puntaje por cada una de las 4 dimensiones de CV, los modelos también resultaron aptos, con un rendimiento claramente superior a los de clasificación y entre sí, con una diferencia pequeña. Por lo anterior, la

CV se observa como un fenómeno adecuado para su representación mediante técnicas de AA y, más aún, con modelos neuronales inteligentes.

El modelo de MLP para la clasificación alcanzó un desempeño favorable como lo demostró el resultado de sus métricas que alcanzaron 90% de precisión. No obstante, los modelos neuronales de regresión lo superaron con un R2 de 0.96 y 0.99, respectivamente de la versión 1 y versión 2.

La CV constituye un fenómeno de alta relevancia para la sociedad, sin embargo, el desarrollo económico, el crecimiento demográfico desmedido y la excesiva emisión de GEI han afectado la buena percepción de CV de habitantes en la ZM Tula, tal como demuestran los datos revisados. Los casos de enfermedades crónicas se han incrementado, la calidad del aire ha disminuido, la contaminación ambiental se incrementa día a día, y el número de estudios es escaso. Por lo tanto, se considera importante ahondar en las propuestas y estrategias orientadas a mantener la CV en la zona, aprovechando las tecnologías emergentes que ofrecen alto potencial y precisión.

Por otra parte, la propuesta se considera adecuada como apoyo a la toma de decisiones del personal médico y las áreas gubernamentales, quienes son responsables de afrontar las problemáticas de salud tanto física como psicológica, y ambiental, en pro del incremento de la CV de los habitantes que enfrentan una situación de vulnerabilidad por las características de la zona donde habitan.

Referencias

Bentoumi, M., Daoud, M., Benaouali, M., & Taleb Ahmed, A. (2022). Improvement of emotion recognition from facial images using deep learning and early stopping cross validation. Multimedia Tools and applications, 81(21), 29887-29917.

Bhatti, A., Javed, A. R., & Riaz, M. N. (2019). *A Cross Industry Standard Process for Data Mining Implementation and its application in Health Doma*in. International Journal of Computer Applications, 180(43), 36-42.

Borja-Robalino, R., Monleon-Getino, A., & Rodellar, J. (2020). *Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning*. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E30), 184-196.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation*. PeerJ Computer Science, *7*, e623.

Düntsch, I., & Gediga, G. (2019). *Confusion matrices and rough set data analysis*. In Journal of Physics: Conference Series (Vol. 1229, No. 1, p. 012055). IOP Publishing.

Espinosa-Zúñiga, J. J. (2020). *Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública*. Ingeniería, investigación y tecnología, 21(1).

Flores-Ramírez M., Escalante-Hernández C., Ortiz-López G., Chico-Barba G. *Calidad de vida en profesionales de enfermería que laboran en áreas críticas en una institución de tercer nivel de atención*. Rev Enferm Neurol. 2020;19(2):pp. 53-65.

García, A.K. (2020). *México, en el top 5 de mayor calidad de vida en América Latina, según los datos de Numbeo.* El economista. Recuperado 28 de marzo de 2023, de https://www.eleconomista.com.mx/economia/Mexico-en-el-top-5-de-mayor-calidad-de-vida-en-America-Latina-segun-los-datos-de-Numbeo-20210718-0010.html

García, K. (2020). *México, en el top 5 de mayor calidad de vida en América Latina, según los datos de Numbeo*. El economista. Recuperado 28 de marzo de 2023, de https://www.eleconomista.com.mx/economia/Mexico-en-el-top-5-de-mayor-calidad-de-vida-en-America-Latina-segun-los-datos-de-Numbeo-20210718-0010.html.

Grande, R. E., & Bonilla, M. N. I.(2023). *Red neuronal convolucional (CNN) de reconocimiento de plantas de maíz para un sistema de visión artificial*. Página Editorial, 356.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hidalgo-Rasmussen, C. A., Morales, G., Ortiz, M. S., Rojas, M. J., Balboa-Castillo, T., Lanuza, F., & Muñoz, S. (2021). *Propiedades psicométricas de la versión chilena del WhoqolBref para la calidad de vida*. Psicología Conductual, *29*(2), 383-398.

Hodson, T. O. (2022). *Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not*. Geoscientific Model Development, *15*(14), 5481-5487.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2020). *CondenseNet: An efficient DenseNet using learned group convolutions*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), 1710-1723.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). *Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques*. IEEE access, *9*, 39707-39716.

Moreno-Gutiérrez, S. S., Tlanepantla-Pantoja, D., López-Pérez, S., & Siliceo-Cantero, H. H. (2024). Aprendizaje automático para el diagnóstico médico predictivo: aplicación en zonas industriales. DYNA-Ingeniería e Industria, 99(2).

Natalia, I. B. M., & Grande, E. (2023). Red neuronal convolucional para la clasificación de piezas mecánicas usando un sistema de visión artifi-cial. Revista Ingeniantes, 10(2), 2.

Ordoñez-Aquino, C., & Gonzales, G. F. (2023). Calidad del aire en Perú no se ajusta a los valores recomendados por la Organización Mundial de la Salud (OMS). Revista Médica Herediana, 34(4), 236-238.

Ortiz, A. P., Diaz, M., Díaz, J. M. M., Gálvez, A. L. B., Rangel, A. L. M. G. C., & Arévalo, R. V. (2022). *Online Cognitive-Behavioral Intervention on Adherence and Quality of Life in Elderly Adults with Diabetes: Two cases study*. International journal of psychology and psychological therapy. *22*(3), 331-344

Programa de las Naciones Unidas para el Desarrollo. PNUD. (2023). *Informe de Desarrollo Humano Municipal 2010-2020: Una Década de Transformaciones Locales Para El Desarrollo De México*. UNDP.org. PNUD México. Recuperado 29 de marzo de 2023, de https://www.undp.org/es/mexico/publicaciones/informe-de-desarrollo-humano-municipal-2010-2020-una-decada-de-transformaciones-locales-en-mexico-0

Rico, R. M., Carrasco-Gallegos, B. V., & Némiga, X. A. (2022). *Importancia de las áreas verdes en zonas urbanas con alta contaminación*. El caso de Atitalaquia, Atotonilco de Tula y Apaxco, México. CONTEXTO. Revista de la Facultad de Arquitectura de la Universidad Autónoma de Nuevo León, 16(24), 40-53.

Rojas-Rubio, L., & Meneses Villegas, C. (2022). *Una comparación empírica de algoritmos de aprendizaje automático versus aprendizaje profundo para la detección de noticias falsas en redes sociales*. Ingeniare. Revista chilena de ingeniería, *30*(2), 403-415.

Salda, R. (2020). *Hidalgo*: *Economía, empleo, equidad, calidad de vida, educación, salud y seguridad pública*. Data México. Recuperado de https://datamexico.org/es/profile/geo/hidalgo-hg

Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). *Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11*. *Revista de investigación, desarrollo e innovación*, *9*(2), 363-378

WHO (World Health Organization). (1999). *World Atlas of Aging*, *World Health Organization Center for Health development*. WHO Press, Kobe, pp. 1-138

WHOQOL Group. (2012). *The World Health Organization Quality of Life (WHOQOL) (WHO/HIS/HSI Rev.2012.03, Vol. 106*). World Health Organization. https://www.who.int/publications/i/item/WHO-HIS-HSI-Rev.2012.