

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial
15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física
Maestría en Ciencia de Datos



Acceso al agua potable y relación con la pobreza en la zona metropolitana de Guadalajara

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
María Luisa Peralta Morfín

Director:
Mtro. Arturo Silva Gálvez

Tlaquepaque, Jalisco, 3 de junio de 2025

Acceso al agua potable y relación con la pobreza en la zona metropolitana de Guadalajara

María Luisa Peralta Morfín

Resumen

Este trabajo analiza la relación entre el nivel socioeconómico (NSE) y la disponibilidad de agua potable en la Zona Metropolitana de Guadalajara. Se parte del problema del acceso desigual al agua y su impacto en comunidades con menores recursos. El objetivo es evaluar si existe una correlación entre estos factores, utilizando modelos estadísticos y de aprendizaje automático.

Se emplearon datos del Censo de Población y Vivienda 2020 del INEGI y del índice de NSE calculado por la AMAI. A partir de estos, se aplicaron distintos modelos: regresión lineal, regresión logística, árboles de decisión (regresión y clasificación) y una red neuronal multicapa. Las variables consideradas incluyeron indicadores de acceso al agua como presencia de agua entubada, tinacos y cisternas.

Los modelos de regresión mostraron un bajo ajuste, con coeficientes R^2 cercanos a cero. Los modelos de clasificación aumentaron considerablemente en desempeño, siendo la red neuronal la que alcanzó la mayor precisión con un valor de 62.74 %. Los modelos sugieren una relación entre el NSE y el acceso al agua, pero sí revelan diferencias en el tipo de abastecimiento.

La variable con mayor peso fue el acceso a agua entubada (VIV_3), el acceso a agua entubada y saneamiento está más presente en zonas con NSE alto, mientras que las zonas con menor NSE muestran una mayor dependencia de tinacos y cisternas. Los resultados también indican la falta de indicadores clave que limitan la capacidad para establecer una relación clara entre el NSE y el acceso al agua. Por ello, se recomienda ampliar la recopilación de datos para incluir variables adicionales que permitan un análisis más profundo y robusto de esta problemática.

Tabla de Contenidos

	Página
1	Introducción 13
1.1.	Contexto 13
1.2.	Justificación 14
1.3.	Problema 15
1.4.	Objetivos 16
1.4.1.	Objetivo general 16
1.4.2.	Objetivos específicos 16
2	Metodología 17
2.1.	Descripción de los datos 17
2.2.	Análisis exploratorio 19
2.3.	Descripción de los modelos 24
2.3.1.	Regresión Lineal 24
2.3.2.	Regresión Logística 24
2.3.3.	Árboles de Decisión (Decision Trees) 25
2.3.4.	Redes Neuronales 25
2.4.	Descripción de las métricas 26
2.5.	Descripción de los experimentos o simulaciones 27
2.5.1.	Pre-procesamiento 27
2.5.2.	Regresión Lineal 28
2.5.3.	Árbol de decisión de regresión 28
2.5.4.	Regresión Logística 28
2.5.5.	Árbol de decisión de clasificación 28
2.5.6.	Red Neuronal 28
3	Resultados y discusión. 31
3.1.	Resultados 31
3.1.1.	Regresión lineal 31
3.1.2.	Árbol de decisión de regresión 31
3.1.3.	Regresión logística 32
3.1.4.	Árbol de decisión de clasificación 33
3.1.5.	Red neuronal 34
3.2.	Discusión 35
4	Conclusiones y trabajo futuro. 37
4.1.	Conclusiones 37
4.2.	Trabajo futuro 37

Índice de figuras

	Página
2.1. a) Ejemplo de datos e indicadores del Censo de población y vivienda 2020.	18
2.2. b) Ejemplo de datos e indicadores del Censo de población y vivienda 2020.	18
2.3. Ejemplo de datos e indicadores NSE calculado por AMAI en 2020.	19
2.4. Histogramas para variables numéricas del Censo de Población y Vivienda 2020.	20
2.5. Boxplot para variables numéricas del Censo de población y vivienda 2020.	21
2.6. Distribución del NSE calculado por AMAI en 2020.	22
2.7. Distribución del NSE calculado por AMAI en 2020.	22
2.8. Mapa del indicador NSE calculado por AMAI en 2020.	23
2.9. Red Neuronal de retropropagación	26
3.1. Curva de convergencia del modelo de regresión logística.	32
3.2. Análisis de pesos de las variables en la regresión logística.	32
3.3. Análisis de pesos de las variables en el árbol de decisión de clasificación.	33
3.4. Curva de convergencia del modelo de red neuronal.	34
3.5. Análisis de pesos de las variables en la red neuronal	35

Índice de tablas

	Página
2.1. Primera sección de indicadores seleccionados del Censo de población y vivienda 2020.	17
2.2. Segunda sección de indicadores seleccionados del Censo de población y vivienda 2020.	18
2.3. Indicadores seleccionados del NSE calculado por AMAI en 2020.	19
3.1. Comparación de modelos de regresión de acuerdo al coeficiente de determinación R^2	35
3.2. Comparación de modelos de clasificación de acuerdo a la precisión.	36

Dedicado a mi familia. A mis padres, quienes han sido una inspiración constante no solo en el ámbito personal, sino también profesional. Su constante superación y dedicación a su comunidad. Siempre buscando que su trabajo tenga un impacto social, ayudando a comunidades menos privilegiadas. También se lo dedico a mi hermana, que ha estado a mi lado siempre, apoyando mis decisiones y cuidándome en los días difíciles.

1 Introducción

1.1 Contexto

Durante los últimos años hemos experimentado las afectaciones del cambio climático y el daño medioambiental. Una de estas es la falta de agua, que observamos en los cortes de suministro en las ciudades y las sequías, al igual que el conteo del "día cero", el día en que ciertas regiones se quedarán sin agua.

En 2022, la Organización Mundial de la Salud y UNICEF reportaron que 2.2 mil millones de personas en el mundo no cuentan con acceso seguro a agua potable. [1] En México, de acuerdo con la Encuesta Nacional de Calidad e Impacto Gubernamental, el 52.3 % de la población en zonas urbanas considera que el suministro de agua potable es constante, y solo el 20.9 % a nivel nacional percibe que el agua que recibe es apta para beber sin causar daño a la salud. [2] Aunque la Norma Oficial Mexicana NOM-127-SSA1-2021 [3] establece los límites permisibles de calidad que debe cumplir el agua destinada al uso y consumo humano, la percepción ciudadana refleja desconfianza en su potabilidad.

Ha incrementado el interés de la relación entre la falta de acceso al agua y su importancia como recurso vital en la lucha contra la pobreza [4]. Diversos estudios han señalado que la carencia de agua potable limita las oportunidades de desarrollo, afecta la salud, impide la asistencia escolar y reduce el tiempo disponible para actividades productivas, lo que refuerza las condiciones de pobreza. [5, 6] Por ello, la falta de acceso a servicios básicos, como el agua potable, debe considerarse un indicador estructural de pobreza. En este contexto, surgió el término "pobreza hídrica", definido como "la situación en la que una nación o región no puede permitirse el costo del agua potable sostenible para todas las personas en todo momento". [7]

Un artículo que analiza las implicaciones del desabasto de agua mediante una encuesta realizada en el Área Metropolitana de Guadalajara (AMG) durante la crisis hídrica de 2021 muestra que más del 50% de los encuestados no contaron con acceso a agua potable. Esto afectó gravemente su calidad de vida y generó gastos

imprevistos. La encuesta también revela insatisfacción con las medidas gubernamentales y la falta de organización vecinal, destacando la necesidad de un conocimiento profundo para una gestión adecuada del agua. [8]

De acuerdo con la UNESCO, en Latinoamérica y el Caribe, las personas sin acceso adecuado al agua en las zonas urbanas suelen vivir en las periferias. [9] Estas zonas presentan grandes desafíos que limitan el acceso al agua segura, lo que refuerza las desigualdades sociales y económicas.

Este documento se estructura de la siguiente manera:

Introducción: Este capítulo proporciona un trasfondo del problema relacionado con el acceso al agua y su vínculo con la pobreza, una justificación para el estudio y una descripción general de los objetivos de la investigación en la Zona Metropolitana de Guadalajara.

Metodología: Este capítulo describe el enfoque metodológico de la investigación, incluyendo la recolección de datos, el análisis exploratorio de los mismos, y la explicación de los modelos y métricas utilizados para evaluar la relación entre el acceso al agua y la pobreza en la Zona Metropolitana de Guadalajara.

Resultados y Discusión: En este capítulo se presentan y discuten los hallazgos del análisis, proporcionando una reflexión crítica sobre los resultados en el contexto del acceso al agua y su impacto en la pobreza en la región.

Conclusiones: En este capítulo final se ofrecen conclusiones basadas en los hallazgos de la investigación, así como recomendaciones para políticas públicas y futuras investigaciones que aborden la relación entre el agua y la pobreza en la Zona Metropolitana de Guadalajara.

1.2 *Justificación*

De acuerdo con un análisis cuantitativo del agua, se determinó que el acceso limitado al agua incrementa las probabilidades de que una persona caiga en situación de pobreza. Esto sucede debido al tiempo que se debe invertir para conseguirla, reduciendo el tiempo disponible para otras actividades como el trabajo o la educación.[10] La falta de acceso a servicios básicos como el agua perpetúa un ciclo de pobreza, dificultando la mejora de las condiciones de vida.

Abordar estos problemas es complejo, ya que implica decisiones políticas sobre el uso del agua, la privatización de servicios y la regulación de la industria y los intereses económicos prevalecen sobre las necesidades de las comunidades más afectadas [11]. Las industrias responsables de la degradación ambiental no han enfrentado suficientes sanciones, mientras que el crecimiento poblacional sin una adecuada

planificación urbana exagera la falta de acceso a servicios [12].

Es fundamental reconocer la correlación entre el acceso al agua y otros factores como la salud y la higiene de la población. El agua es esencial para prevenir enfermedades y garantizar un saneamiento adecuado. El análisis de datos permite monitorear en tiempo real la calidad y disponibilidad de agua, así como realizar mapeos del uso local y flujos de agua subterránea [13]. Los modelos estadísticos, matemáticos y de aprendizaje automático serán útiles para identificar la relación entre las áreas con escasez de agua y la disparidad económica.

Actualmente el 70% del agua global se destina a la agricultura, lo que tiene un impacto significativo en la disponibilidad del recurso para otras actividades.[14] El uso de aguas subterráneas a mayores profundidades ha sido clave para reducir la escasez en algunas regiones, pero también presenta riesgos de sobreexplotación. Nuevas tecnologías, como la desalinización y las herramientas de información geográfica, han permitido ofrecer soluciones más innovadoras, promoviendo una gestión del agua más eficiente y equitativa.[14]

Se puede demostrar la relación entre la escasez de agua y la pobreza en la Zona Metropolitana de Guadalajara (ZMG) usando datos del INEGI, para identificar con mayor precisión las áreas más afectadas y generar un marco de referencia para la implementación de políticas públicas más equitativas. Además, los resultados de este análisis podrían servir como base para futuras investigaciones en otras regiones urbanas con características similares, ampliando el impacto de este trabajo tanto a nivel local como nacional.

1.3 *Problema*

El acceso al agua es un derecho humano vital que requiere atención urgente. La gestión del agua y las condiciones socioeconómicas están entrelazadas, lo que hace necesario abordarlas mediante políticas públicas integrales. En 2010, las Naciones Unidas reconocieron el acceso al agua como un derecho humano universal, subrayando su importancia fundamental para la vida [15].

A partir de este reconocimiento, diversos países han reformado sus marcos legales y creado programas enfocados en mejorar el acceso al agua. En México se incorporó este derecho en el artículo 4° constitucional ese mismo año, lo que impulsó la creación de políticas como el Programa Nacional Hídrico y estrategias de cobertura en comunidades marginadas. Sin embargo, los avances han sido desiguales, y persisten brechas importantes en la calidad, disponibilidad y accesibilidad del agua, especialmente en zonas rurales y periurbanas [16].

En la ZMG, la dependencia del agua proviene principalmente del Lago de Chapala y diversas presas, pozos y fuentes subterráneas. El ritmo actual de extracción pone en riesgo la sostenibilidad de estos recursos, y pronto serán insuficientes para abastecer a la población [17]. Uno de los mayores retos en una ciudad es crear soluciones a largo plazo que consideren la sostenibilidad de los recursos y sean sistémicas, asegurando el abastecimiento de agua.

El estudio busca aplicar distintos modelos de aprendizaje automático para evaluar la relación entre el acceso al agua y el nivel socioeconómico (NSE), e identificar las características de las viviendas con escasez de agua. Debido a que la estructura y calidad de los datos pueden dificultar el ajuste de ciertos modelos, se seleccionarán enfoques con características y capacidades diversas. Se pretende encontrar patrones e indicadores relevantes en las viviendas que enfrentan dificultades en el acceso al agua. Con el fin de generar información que contribuya a futuro para diseñar políticas públicas más efectivas y focalizadas para atender esta problemática.

1.4 *Objetivos*

1.4.1 *Objetivo general*

Analizar el acceso al agua potable de la Zona Metropolitana de Guadalajara a través de indicadores de INEGI e identificar su relación con el nivel socioeconómico utilizando modelos de aprendizaje automático.

1.4.2 *Objetivos específicos*

1. Recopilar datos sobre la disponibilidad de agua potable en la Zona Metropolitana de Guadalajara, utilizando fuentes como el INEGI.
2. Definir los indicadores que representarán el acceso al agua potable.
3. Seleccionar modelos de aprendizaje automático.
4. Definir las métricas con las que se evaluará el desempeño de cada modelo para poder comparar los resultados.
5. Aplicar modelos de aprendizaje automático para determinar la relación entre el acceso al agua y el nivel socioeconómico.
6. Analizar la relevancia de los indicadores de acceso al agua en relación con el nivel socioeconómico.

2 Metodología

2.1 Descripción de los datos

La base de datos presenta los resultados del Censo de Población y Vivienda realizado en México en 2020, y la información está disponible en el Sistema de Consulta de Integración Territorial (SCITEL). [18] A través de esta herramienta, los usuarios pueden descargar la información filtrando por entidad o seleccionando los indicadores deseados entre los 222 disponibles. El Instituto Nacional de Estadística y Geografía (INEGI) presenta los resultados por AGEB (Área Geoestadística Básica), que es una división geográfica establecida por el INEGI, y por manzana urbana.

La consulta se realizó para la entidad federativa de Jalisco, seleccionando los indicadores de identificación geográfica y vivienda. Se utilizó el filtro avanzado para obtener exclusivamente los datos de la Zona Metropolitana de Guadalajara (ZMG), la cual incluye los municipios de Guadalajara, Tonalá, Tlajomulco de Zúñiga, Tlaquepaque y Zapopan. Una vez generada la consulta, los datos se pueden descargar en formato XLSX o CSV. El total de registros es de 51,794 para los 5 municipios de la ZMG.

Los indicadores a utilizar se muestran en las siguientes tablas, 2.1 y 2.2.

#	Clave	Nombre	Tipo de campo
1	MUN	Clave del municipio o demarcación territorial	Número
2	NOM_MUN	Nombre del municipio o demarcación territorial	Texto
3	LOC	Clave de Localidad	Número
4	NOM_LOC	Nombre de la localidad	Texto
5	AGEB	Clave de AGEB	Número y Texto
6	CLAVE_M	Clave de la manzana	Número
7	T_VIV	Total de viviendas	Número
8	OCUP_VIV	Ocupantes en viviendas particulares habitadas	Número
9	PROM_OC	Promedio de ocupantes por cuarto en viviendas particulares habitadas	Número

Tabla 2.1: Primera sección de indicadores seleccionados del Censo de población y vivienda 2020.

#	Clave	Nombre	Tipo de campo
10	VIV_1	Viviendas particulares habitadas con piso de tierra	Número
11	VIV_2	Viviendas particulares habitadas que no disponen de energía eléctrica	Número
12	VIV_3	Viviendas particulares habitadas que disponen de agua entubada en el ámbito de la vivienda	Número
13	VIV_4	Viviendas particulares habitadas que disponen de agua entubada y se abastecen del servicio público de agua	Número
14	VIV_5	Viviendas particulares habitadas que disponen de tinaco	Número
15	VIV_6	Viviendas particulares habitadas que disponen de cisterna o aljibe	Número
16	VIV_7	Viviendas particulares habitadas que disponen de excusado o sanitario	Número
17	VIV_8	Viviendas particulares habitadas que disponen de drenaje	Número
18	VIV_9	Viviendas particulares que disponen de drenaje y sanitario con admisión de agua	Número

Tabla 2.2: Segunda sección de indicadores seleccionados del Censo de población y vivienda 2020.

Una muestra de este conjunto de datos se muestra en las siguientes figuras 2.1 y 2.2.

MUN	NOM_MUN	LOC	NOM_LOC	AGEB	CLAVE_M	T_VIV	OCUP_VIV
039	Guadalajara	0001	Guadalajara	005A	001	32	82
039	Guadalajara	0001	Guadalajara	037A	010	21	48
097	Tlajomulco de Zúñiga	0823	[Fraccionamiento	4003	001	105	202
097	Tlajomulco de Zúñiga	0827	nta Anita [Fraccic	4130	020	16	39
098	San Pedro Tlaquepaque	0001	Tlaquepaque	2644	001	84	317
098	San Pedro Tlaquepaque	0001	Tlaquepaque	2659	027	18	44
101	Tonalá	0001	Tonalá	1554	047	0	0
101	Tonalá	0001	Tonalá	0378	024	54	225
120	Zapopan	0001	Zapopan	187A	018	16	46
120	Zapopan	0001	Zapopan	1899	039	32	107

Figura 2.1: a) Ejemplo de datos e indicadores del Censo de población y vivienda 2020.

PROM_OCUP	VIV_1	VIV_2	VIV_3	VIV_4	VIV_5	VIV_6
0.57	*	0	30	30	30	24
0.62	0	0	19	19	19	14
1.28	0	0	61	59	5	*
0.6	0	0	13	13	13	12
1.08	3	0	77	77	57	44
0.76	0	0	15	15	14	*
0	0	0	0	0	0	0
1.15	*	0	52	52	44	14
0.94	0	0	13	13	*	*
0.88	0	0	27	27	27	20

Figura 2.2: b) Ejemplo de datos e indicadores del Censo de población y vivienda 2020.

La segunda fuente de datos corresponde al nivel socioeconómico (NSE) calculado a nivel AGEB y municipal en Jalisco para el año

2020 [19] Este índice fue desarrollado por la Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión AC (AMAI) a partir de los resultados de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2020. La clasificación del NSE segmenta a la población en siete categorías que reflejan el poder adquisitivo, desde el nivel A/B (el más alto) hasta el nivel E (el más bajo): A/B, C+, C, C-, D+, D y E.

Esta clasificación proporciona una visión detallada de la distribución socioeconómica en las distintas regiones del estado, lo que permite identificar patrones de desigualdad y relacionarlos con otros factores, como el acceso al agua potable. En el contexto del análisis, los datos del NSE sirven para explorar posibles asociaciones entre el nivel socioeconómico y la disponibilidad de servicios básicos en la Zona Metropolitana de Guadalajara.

Los indicadores a utilizar se muestran en la tabla 2.3.

#	Clave	Nombre	Tipo de campo
1	MUN	Clave del municipio	Número
2	NOM_MUN	Nombre del municipio	Texto
3	LOC	Clave de Localidad	Número
4	NOM_LOC	Nombre de la localidad	Texto
5	AGEB	Clave de AGEB	Número y Texto
6	NSE	Nivel predominante	Texto
7	T_VIV	Total de viviendas	Número

Tabla 2.3: Indicadores seleccionados del NSE calculado por AMAI en 2020.

Una muestra de este conjunto de datos se muestra en la figura 2.3.

MUN	NOM_MUN	LOC	NOM_LOC	AGEB	NSE	T_VIV
39	Guadalajara	1	Guadalajara	0740	A/B	809
39	Guadalajara	1	Guadalajara	5489	C	279
97	Tlajomulco de Zúñiga	1	Tlajomulco de Zúñiga	0195	D	457
97	Tlajomulco de Zúñiga	1	Tlajomulco de Zúñiga	1117	C+	52
98	San Pedro Tlaquepaque	1	Tlaquepaque	3356	E	3
98	San Pedro Tlaquepaque	14	Santa Anita	0027	D	846
101	Tonalá	1	Tonalá	0503	D+	773
101	Tonalá	1	Tonalá	0645	C-	525
120	Zapopan	1	Zapopan	7027	A/B	315
120	Zapopan	22	La Primavera	0142	C	566

Figura 2.3: Ejemplo de datos e indicadores NSE calculado por AMAI en 2020.

2.2 Análisis exploratorio

El análisis exploratorio de datos es un paso fundamental para comprender la estructura, distribución y características principales de los datos, permitiendo así justificar las decisiones de modelado. Este análisis se llevó a cabo en dos etapas principales: limpieza de datos y análisis descriptivo.

Se seleccionaron los datos correspondientes a los cinco municipios de la Zona Metropolitana de Guadalajara (ZMG) de los conjuntos de

datos estatales. Para estos municipios se van a analizar los valores de VIV_3 a VIV_9 que son los indicadores relacionados al agua. En esta etapa, se identificaron y transformaron los valores nulos, representados como *, N/D y N/A. Luego, se calculó el porcentaje de datos faltantes para evaluar si era conveniente incluir o excluir indicadores con una alta proporción de valores nulos. La mayoría de los indicadores tuvieron un porcentaje de datos nulos inferior al 5%, salvo 'Viviendas particulares habitadas que disponen de cisterna o aljibe', que presenta un 15% de valores nulos.

Se revisaron los tipos de columna y se convirtieron aquellos indicadores numéricos que estaban marcados como categóricos a valores numéricos, lo que permitió realizar los análisis estadísticos correctamente. Se generaron histogramas, cuyos resultados se muestran en la figura 2.4.

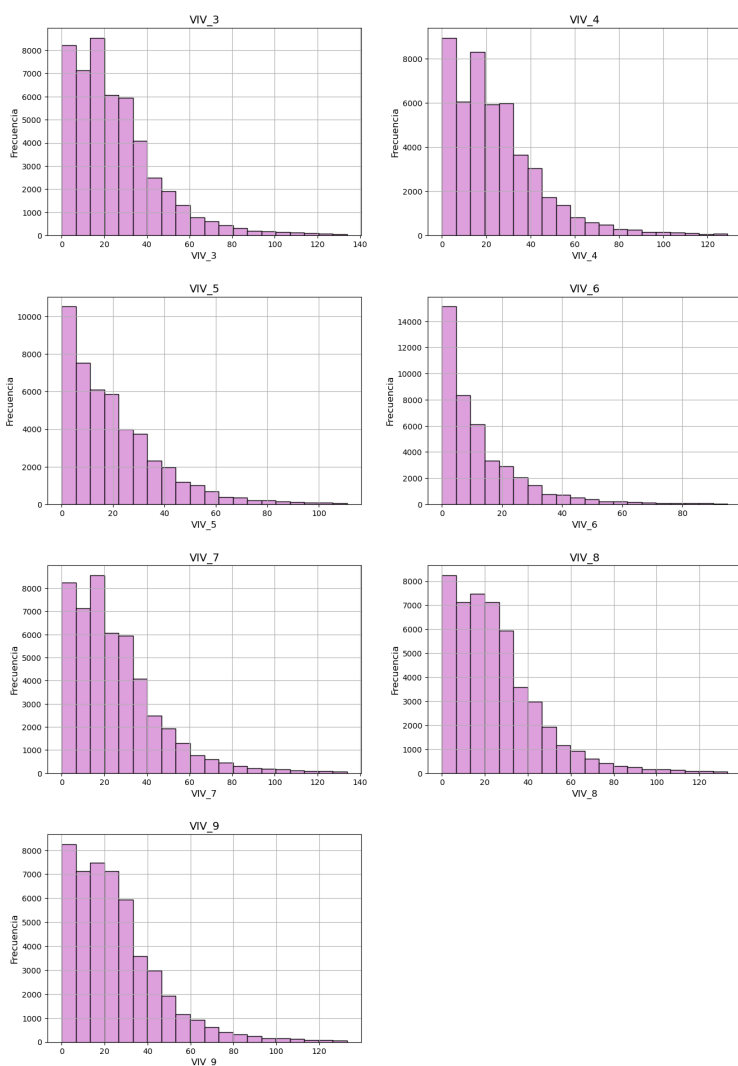


Figura 2.4: Histogramas para variables numéricas del Censo de Población y Vivienda 2020.

En ellos se observa que la mayoría de las variables numéricas presentan una distribución altamente sesgada a la izquierda, con frecuencias concentradas en valores bajos y una larga cola hacia la derecha. Todos los histogramas muestran esta asimetría positiva, y se continuará evaluando el manejo de los valores atípicos.

Para identificar valores atípicos se generaron boxplots para cada variable (Figura 2.5). Se puede ver en todas las variables que existen valores que se alejan considerablemente del rango intercuartílico. Esto refuerza la necesidad de considerar técnicas de normalización o transformación de los datos en etapas posteriores del análisis.

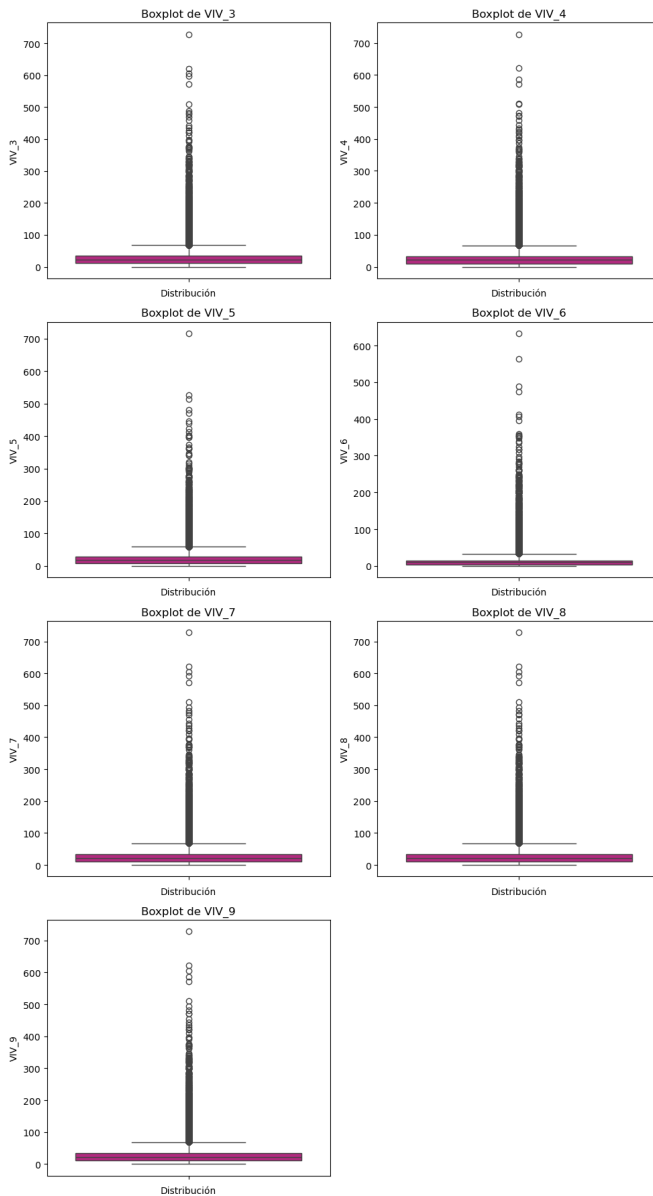


Figura 2.5: Boxplot para variables numéricas del Censo de población y vivienda 2020.

Para el indicador NSE, se generó un gráfico de distribución que se muestra en la figura 2.6. En este se observa que la categoría D es la más frecuente, mientras que D+ es la menos común.

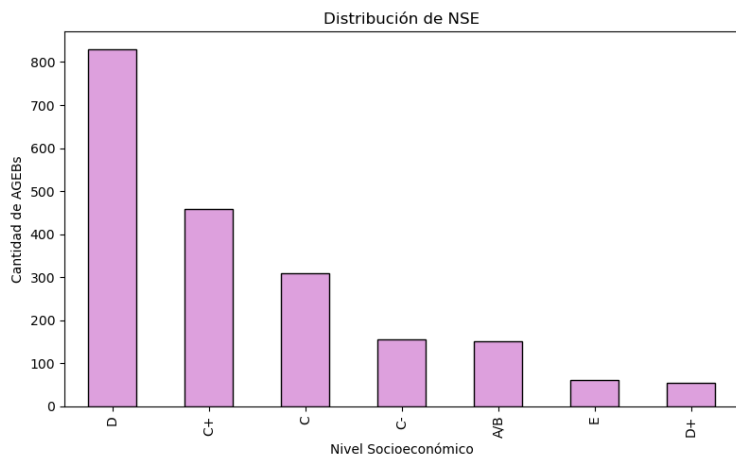


Figura 2.6: Distribución del NSE calculado por AMAI en 2020.

Para poder obtener una idea inicial de la relación lineal entre las variables se utilizó una matriz de correlación (Figura 2.7). La matriz de correlación permite identificar grupos de variables altamente correlacionadas ($r > 0.9$), lo que indica redundancia de información. Para evitar colinealidad, se seleccionan únicamente aquellas variables que representaban de forma más clara el acceso a servicios básicos.



Figura 2.7: Distribución del NSE calculado por AMAI en 2020.

Las variables VIV_4, VIV_7, VIV_8 y VIV_9 presentan una correlación perfecta entre sí (valor de 1.00) y alta correlación con VIV_3, lo que indica que capturan un comportamiento similar. Se conserva únicamente VIV_3 como representante de esta dimensión del acceso al agua. Asimismo, se seleccionan VIV_5 y VIV_6, que presentan una correlación más moderada entre ellas (0.82) y con el resto de las variables, además de mostrar la mayor asociación con la variable objetivo NSE.

Esta selección permite reducir la redundancia y conservar información representativa para el análisis posterior. Una vez seleccionadas estas variables, se corrigió el sesgo previamente visto en los histogramas con Box-Cox.

Finalmente, se generó un mapa del NSE en la ZMG para explorar su distribución geoespacial en ArcGIS Pro (Figura 2.8). Se observó una concentración de niveles socioeconómicos más altos en zonas centrales de municipios como Zapopan y Guadalajara, mientras que las periferias tienden a concentrar niveles bajos como D y D+. Este patrón coincide con estudios previos sobre segregación urbana y puede ser relevante para entender la distribución del acceso a servicios básicos como el agua.

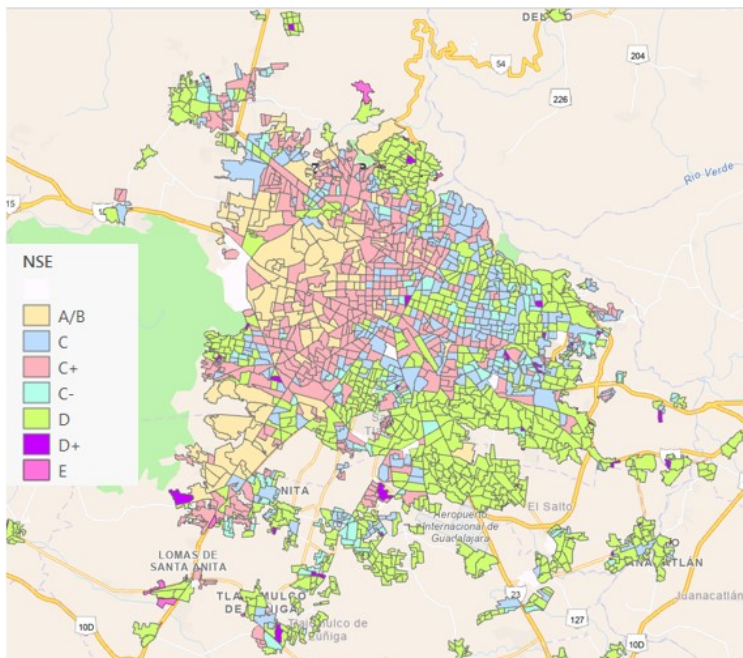


Figura 2.8: Mapa del indicador NSE calculado por AMAI en 2020.

2.3 Descripción de los modelos

En esta sección se describe el conjunto de modelos utilizados para analizar la correlación entre la pobreza y el acceso al agua en la ZMG. Se aplicarán diferentes metodologías de aprendizaje automático para explorar las relaciones entre estos dos factores. El objetivo de esta sección es justificar la selección de estos modelos en el contexto de la investigación, destacando su aplicabilidad y la capacidad para extraer patrones significativos del conjunto de datos.

2.3.1 Regresión Lineal

Este es un modelo estadístico que examina la relación lineal entre una variable dependiente y una o más variables independientes, lo que permite predecir el valor de la variable dependiente. La regresión lineal ajusta una línea recta a los datos, minimizando la diferencia entre las predicciones y los valores reales. [20] Es un modelo simple, pero se ha demostrado que es muy eficiente por su bajo costo computacional y rapidez de ejecución. Aunque no captura relaciones no lineales complejas, la regresión lineal puede servir como base para compararse con modelos más avanzados. Matemáticamente, podemos escribir la relación con la siguiente ecuación:

$$Y = \beta_0 + \beta_1 x_1 \quad (2.1)$$

Donde:

- Y es la variable dependiente
- β_0 es la intersección
- β_1 es el coeficiente de la variable independiente
- x_1 es la variable independiente

2.3.2 Regresión Logística

La regresión logística es un modelo estadístico utilizado para predecir una variable dependiente binaria a partir de una o más variables independientes. En lugar de predecir valores continuos, como en la regresión lineal, predice la probabilidad de que la variable dependiente pertenezca a una categoría específica. Por ejemplo, en un modelo de default (sí/no), se estima la probabilidad de que ocurra un default dado un balance. Esta probabilidad se encuentra entre 0 y 1, y se puede predecir un default si la probabilidad es mayor a un umbral, como 0.5. [20] Matemáticamente, la relación se expresa mediante la siguiente ecuación:

$$\log \left(\frac{p(X)}{1-p(X)} \right) = \beta_0 + \beta_1 X \quad (2.2)$$

Donde:

- $P(X)$ s la probabilidad de que el evento ocurra
- β_0 es la intersección
- β_1 es el coeficiente de la variable independiente
- X es la variable independiente

2.3.3 Árboles de Decisión (Decision Trees)

El árbol de decisión es un modelo que se puede utilizar tanto para la regresión como la clasificación. Su estructura es de jerarquía, donde se dividen los datos en subconjuntos más pequeños basados en características relevantes, estas son representadas en los nodos internos. Las predicciones finales se asignan en las hojas, ya sea categorías (clasificación) o valores numéricos (regresión). Para regresión, el modelo divide el espacio de características X en regiones R_1, R_2 , etc. [20] Cada región predice la media de los valores de la variable objetivo:

$$\hat{y}_m = \frac{1}{N_m} \sum_{i \in R_m} y_i \quad (2.3)$$

En clasificación, las divisiones maximizan las regiones mediante métricas como la entropía. El proceso se detiene al alcanzar ciertos criterios, como un número mínimo de observaciones por hoja. El criterio de división minimiza la suma de los errores cuadráticos (MSE): [20]

$$MSE = \sum_{m=1}^M \sum_{i \in R_m} (y_i - \hat{y}_m)^2 \quad (2.4)$$

2.3.4 Redes Neuronales

Las redes neuronales son un modelo de aprendizaje automático inspirado en el cerebro humano, compuesto por capas de neuronas artificiales interconectadas y con capacidad para capturar patrones complejos de grandes volúmenes de datos. Cada neurona realiza una operación matemática. La salida de cada capa se usa como entrada para la siguiente, permitiendo que el modelo aprenda de los datos de forma jerárquica. El modelo ajusta los pesos durante el entrenamiento mediante un proceso de retropropagación, utilizando un algoritmo como el descenso del gradiente para minimizar la función de costo. Matemáticamente, una red neuronal simple puede representarse como: [20]

$$y = f \left(\sum_{i=1}^n w_i x_i + b \right) \quad (2.5)$$

Donde:

- x_i son las entradas
- w_i son los pesos
- b es el sesgo
- f es la función de activación

En la figura 2.9 un ejemplo gráfico de una red neuronal que tiene dos entradas, dos capas ocultas con tres nodos cada una, y una salida. Esta red utiliza el algoritmo de retropropagación, lo que significa que el error se propaga hacia atrás desde la salida hacia las capas anteriores. Esto permite que la salida de ciertos nodos influya en las entradas de los nodos anteriores, ajustando los pesos y mejorando el aprendizaje del modelo.

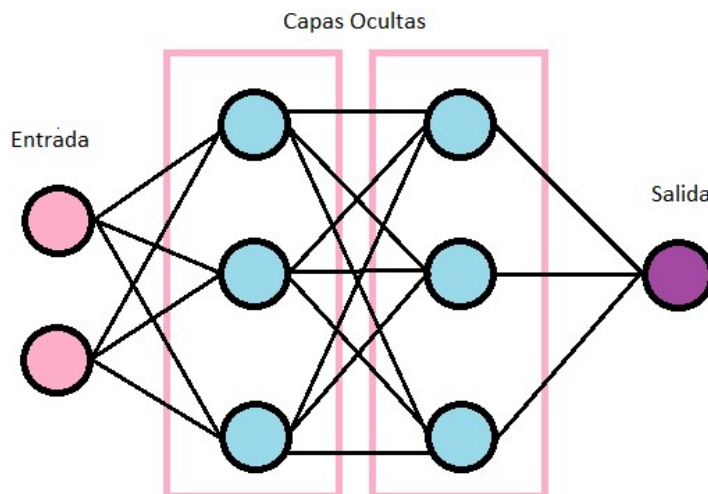


Figura 2.9: Red Neuronal de retropropagación

2.4 Descripción de las métricas

La selección de estas métricas se basa en su capacidad para demostrar el rendimiento de cada modelo. La comparación entre modelos es fundamental para identificar el más adecuado, y las métricas permiten una comparación objetiva de los resultados. De esta manera, se podrá seleccionar el modelo que mejor capture la relación entre la pobreza y el acceso al agua.

- **Precisión:** La precisión mide la proporción de predicciones correctas entre todas las predicciones realizadas, es decir, cuántas de las predicciones positivas fueron realmente correctas. Esta métrica ayuda a evaluar el desempeño del modelo de manera más detallada. Se eligió porque proporciona una evaluación clara de cuántas predicciones fueron realmente acertadas. Para realizar un análisis más preciso del desempeño de cada modelo de clasificación.
- El R^2 indica qué proporción de la variabilidad en los datos de acceso al agua y pobreza es explicada por el modelo. Un R^2 cercano a 1 indica un buen ajuste, mientras que valores cercanos a 0 sugieren que el modelo no logra explicar bien los datos. Se eligió esta métrica porque es ampliamente conocida y fácil de interpretar en modelos de regresión. Además, permite evaluar y comparar de manera objetiva los resultados de los modelos de regresión.

2.5 Descripción de los experimentos o simulaciones

Los experimentos y simulaciones buscan explorar y evaluar diferentes enfoques metodológicos para identificar relaciones significativas entre las variables y seleccionar el modelo que mejor capture esta relación. Cada modelo se detallará con el procedimiento planteado.

2.5.1 Pre-procesamiento

Se trabajó en Python y se utilizaron librerías como scikit-learn, numpy y matplotlib. Todos los modelos usaron como variables predictoras los indicadores de acceso al agua (VIV_3, VIV_5, VIV_6) y como variable objetivo el nivel socioeconómico (NSE). Antes del entrenamiento, los datos fueron normalizados utilizando la estandarización Z-score, restando la media y dividiendo entre la desviación estándar de cada variable. Esta estandarización se realiza para que todas las variables estén en la misma escala.

Posteriormente, los datos se dividieron en conjuntos de entrenamiento y prueba utilizando la función `train_test_split` de scikit-learn, con una proporción de 80/20. Es decir, el 80% de los datos se utilizó para entrenar los modelos y el 20% restante para evaluar su desempeño. Se fijó una semilla aleatoria (`random_state=42`) para asegurar la reproducibilidad de los resultados. Además, se mantuvo la distribución original de las clases en ambos subconjuntos para evitar sesgos en el aprendizaje.

2.5.2 *Regresión Lineal*

Este modelo se utilizó como línea base para explorar la relación entre el nivel socioeconómico (NSE) y los indicadores de acceso al agua. Se entrenó con `LinearRegression()` de `scikit-learn`, usando los datos de entrenamiento, y se evaluó en el conjunto de prueba mediante el coeficiente de determinación R^2 .

2.5.3 *Árbol de decisión de regresión*

El árbol de decisión de regresión es un modelo con mayor complejidad. Se construyó desde cero sin utilizar librerías externas. En cada nodo, el modelo selecciona la división que maximiza la reducción de varianza en la variable objetivo. El conjunto de entrenamiento se entrenó probando diferentes profundidades, desde 1 hasta 5 en incrementos de 1. El rendimiento del modelo fue evaluado en el conjunto de prueba con el coeficiente R^2 con la función `r2_score` de `scikit-learn`.

2.5.4 *Regresión Logística*

El modelo fue implementado desde cero utilizando `numpy`, con función de activación sigmoide y entrenamiento mediante gradiente descendente. Con una tasa de aprendizaje de 0.1, durante 1,000 iteraciones. Para la predicción, se utilizó un umbral de 0.5 donde los valores de salida iguales o mayores se clasificaron como 1, y los menores como 0. La métrica utilizada para evaluar el desempeño fue la precisión, calculada con `sklearn.metrics`. Además, se graficó la curva de convergencia de la función de pérdida y se analizaron los pesos finales de las variables para interpretar su relevancia en el modelo.

2.5.5 *Árbol de decisión de clasificación*

El modelo se construyó utilizando el criterio de ganancia de entropía como medida de pureza en las divisiones. Se programó desde cero una función recursiva para generar el árbol, con una profundidad máxima configurable. Se probaron distintos niveles de profundidad, desde 1 hasta 8 en incrementos de 1. La predicción se realizó recorriendo el árbol para cada observación del conjunto de prueba, y se evaluó con la métrica de precisión. Además, se graficaron los pesos finales de las variables.

2.5.6 *Red Neuronal*

La red neuronal multicapa fue codificada únicamente con `numpy`. Los datos fueron normalizados y divididos en entrenamiento y prueba

como en los demás modelos. La red neuronal tiene tres neuronas en la capa de entrada, una por cada variable predictora, cuatro neuronas en la capa oculta y una en la capa de salida. Utiliza la función de activación sigmoide. Se probaron diferentes tamaños de capa oculta, cuidando no sobreajustar el modelo, y este número resultó ser el más adecuado. También se comparó la sigmoide con otras funciones como ReLU.

El modelo fue entrenado utilizando retropropagación durante 10,000 iteraciones, utilizando el error cuadrático medio como función de pérdida y gradiente descendente como optimizador, con una tasa de aprendizaje de 0.0001. La evaluación se hizo con la precisión. Por último, se graficó la curva de convergencia de la función de pérdida y se analizaron los pesos finales de las variables para interpretar su relevancia en el modelo.

3 Resultados y discusión

3.1 Resultados

A continuación se presentan los resultados obtenidos por cada uno de los modelos, junto con la evaluación de su desempeño y una breve interpretación del comportamiento de las variables más relevantes.

3.1.1 Regresión lineal

La regresión lineal se utilizó como punto de partida. El valor obtenido de $R^2 = 0.0208$ indica que el modelo no logró explicar la variabilidad del NSE a partir de las variables de acceso al agua. Los coeficientes obtenidos para las variables VIV_3 (agua entubada), VIV_5 (tinaco) y VIV_6 (cisterna o aljibe) fueron bajos y sin un patrón destacable.

El bajo valor de R^2 indica que la relación entre el NSE y las variables de acceso al agua no puede explicarse de forma lineal, lo que confirma lo observado en el análisis exploratorio con la matriz de correlación (Figura 2.7). Esto sugiere que el modelo no tiene la capacidad suficiente para capturar la complejidad de los datos.

3.1.2 Árbol de decisión de regresión

El árbol de regresión obtuvo un desempeño aún menor al de la regresión lineal, con un $R^2 = 0.0104$ utilizando una profundidad máxima de 3. El segundo valor más alto de R^2 se obtuvo con una profundidad máxima de 2, con un valor de 0.0097. No se identificó una división predominante por alguna variable específica, lo que sugiere que el modelo no logró establecer reglas efectivas para predecir el NSE a partir de las variables VIV_3, VIV_5 y VIV_6.

Este modelo mantiene un R^2 bajo, lo que indica que, aunque la complejidad del modelo incrementa, no se logra encontrar una relación entre el NSE y las variables de acceso al agua de forma lineal. Aunque el árbol puede dividir los datos según ciertos valores de las variables, los resultados muestran que el acceso al agua no está determinado únicamente por el valor numérico del NSE.

3.1.3 Regresión logística

La Figura 3.1 muestra la curva de convergencia del modelo, donde la pérdida disminuye de forma estable sin señales de sobreajuste.

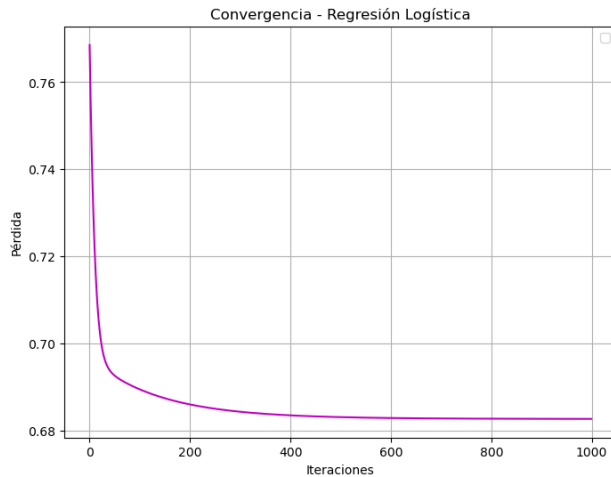


Figura 3.1: Curva de convergencia del modelo de regresión logística.

Este modelo obtuvo una precisión de 0.5403, lo que representa una mejora respecto a los modelos de regresión. Los pesos finales (Figura 3.2) en la regresión logística muestran que la variable VIV_3 tiene la mayor magnitud, seguida de VIV_6 y que la menor magnitud es de VIV_5.

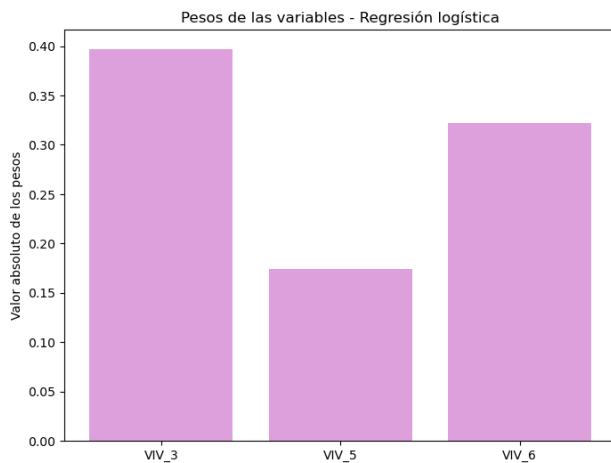


Figura 3.2: Análisis de pesos de las variables en la regresión logística.

La regresión logística tuvo un mejor desempeño que los modelos de regresión, lo que indica que tratar el NSE como una variable categórica es más adecuado. Al analizar los pesos finales, se observa que el acceso a agua entubada está más presente en AGEBs con NSE alto, mientras que las viviendas con alternativas como cisternas o tinacos se asocian más a niveles socioeconómicos bajos. Aunque el modelo no alcanza una precisión alta, sí muestra que hay diferencias en el tipo de abastecimiento de acuerdo con el NSE.

3.1.4 Árbol de decisión de clasificación

El árbol de clasificación obtuvo una precisión de 54.68% con profundidad máxima de 6. La diferencia con otras profundidades no fue tan significativa, siendo el segundo valor más alto de precisión que se obtuvo con una profundidad máxima de 7, con un valor de 54.67%.

La figura 3.3 muestra la frecuencia con la que cada variable fue utilizada en las divisiones del árbol, lo cual se interpretó como una medida de su relevancia. La variable VIV_3 (viviendas con agua entubada en el ámbito de la vivienda) fue la más usada, seguida de VIV_5 (viviendas con tinaco). La variable VIV_6 (viviendas con cisterna o aljibe) tuvo un peso menor.

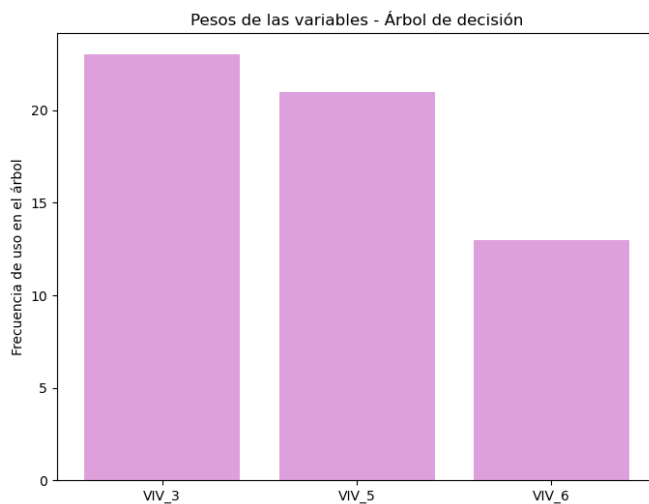


Figura 3.3: Análisis de pesos de las variables en el árbol de decisión de clasificación.

Para este modelo, la precisión se encuentra poco por encima de la regresión logística. Aunque es capaz de generar reglas a partir de las variables, su desempeño sigue siendo limitado. A partir de los pesos, se observa que el árbol depende principalmente de una variable (VIV_3), pero también incorpora con cierta frecuencia a VIV_5. La variable VIV_6 presenta un peso considerablemente menor, es decir, que tiene menos influencia.

3.1.5 Red neuronal

La Figura 3.4 corresponde al conjunto de entrenamiento y muestra una disminución constante de la pérdida, lo que indica que el modelo sí logró aprender durante el entrenamiento.

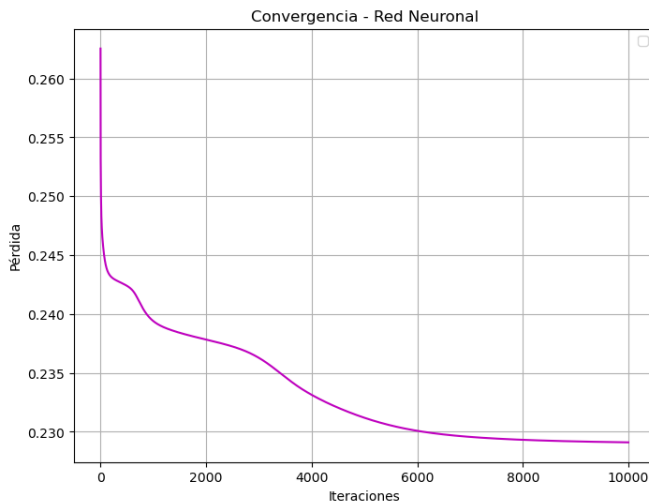


Figura 3.4: Curva de convergencia del modelo de red neuronal.

Este modelo obtuvo la mayor precisión entre los modelos probados, alcanzando un 62.74% en el conjunto de prueba con una capa oculta de tamaño 3. Se evaluaron diferentes tamaños para la capa oculta, desde 1 hasta 10, incrementando de uno en uno. No se continuó más allá porque se buscaba evitar el sobreajuste y la diferencia en precisión no resultaba significativa. Con una capa oculta de tamaño 10, la precisión fue de 62.37%.

Los pesos de la red neuronal como se ven en la figura 3.4 indican que VIV_3 tiene la mayor influencia y el acceso a agua entubada es un factor clave ligado a un NSE más alto al igual que en los modelos anteriores. Las variables VIV_5 (tinaco) y VIV_6 (cisterna o aljibe) tienen pesos de menor magnitud e impacto en la predicción del NSE.

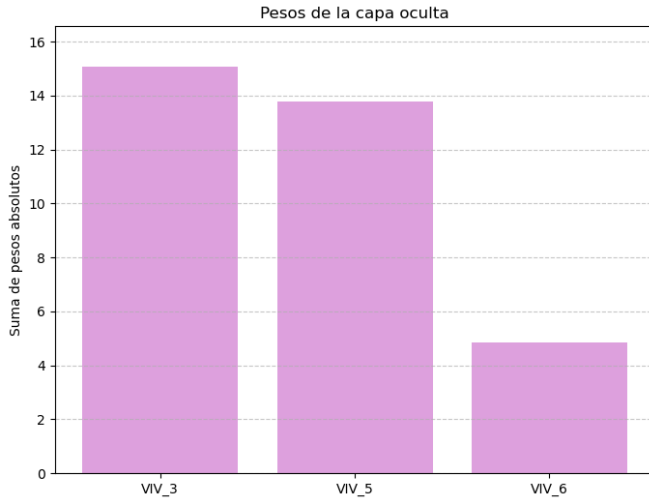


Figura 3.5: Análisis de pesos de las variables en la red neuronal

La red neuronal multicapa fue el modelo con mejor desempeño, aunque su precisión continúa siendo limitada. El modelo logró identificar diferencias en el tipo de abastecimiento según el NSE, pero su utilidad práctica para clasificar con precisión sigue siendo reducida. Esto sugiere que los indicadores de INEGI no son suficientes para relacionar el nivel socioeconómico a partir del acceso al agua.

3.2 Discusión

Aunque los diferentes modelos utilizados mostraron un desempeño limitado, permiten señalar diferencias en el tipo de abastecimiento según el NSE. Se evidenció que el acceso al agua no depende únicamente de una relación directa con el NSE, sino que está condicionado por dinámicas más complejas que los modelos lineales no logran capturar.

La tabla (3.1) muestra la comparación entre los modelos de regresión según su valor de R^2 .

Modelo	R^2
Regresión lineal	0.0208
Árbol de decisión (regresión)	0.0104

Tabla 3.1: Comparación de modelos de regresión de acuerdo al coeficiente de determinación R^2 .

El análisis sugiere que el uso de modelos de clasificación es más adecuado para este tipo de problemas, pero también que las variables disponibles no son suficientes para representar completamente las condiciones de escasez. Esto abre la posibilidad de incorporar nuevos indicadores en futuros estudios y refuerza la necesidad de abordar el acceso al agua como un fenómeno multifactorial.

La tabla (3.2) muestra la comparación entre los modelos de clasificación según su valor de precisión.

Modelo	Precisión
Regresión logística	0.5403
Árbol de decisión (clasificación)	0.5100
Red neuronal (sigmoide)	0.6237

Tabla 3.2: Comparación de modelos de clasificación de acuerdo a la precisión.

Aunque las variables VIV_4 (agua entubada de servicio público), VIV_7 (excusado), VIV_8 (drenaje) y VIV_9 (excusado y drenaje con agua) no fueron evaluadas en los modelos por su alta correlación con VIV_3 (agua entubada) y para evitar redundancia estadística, su contenido sigue siendo relevante. Esto refuerza la idea de que las viviendas con mayor acceso al agua entubada suelen contar también con condiciones de saneamiento adecuadas.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

A partir del análisis realizado, no se identificó una relación lineal entre el nivel socioeconómico (NSE) y los indicadores de acceso al agua. Los modelos de regresión (lineal y de árbol de decisión) presentaron un coeficiente de determinación (R^2) prácticamente cero, lo que indica que no hay una relación lineal entre las variables.

Los modelos de clasificación, particularmente la red neuronal, alcanzaron un mayor nivel de precisión, siendo la red neuronal la más destacada con un 62.74 %, lo que indica un mejor ajuste al identificar patrones no lineales en los datos.

Los pesos de las variables sugieren que las viviendas ubicadas en zonas de menor NSE tienden a depender más de soluciones alternativas de abastecimiento como tinacos o cisternas, lo que indica una condición de acceso menos favorable. La variable VIV_3, viviendas con agua entubada en el ámbito de la vivienda, tuvo mayor peso en los modelos de clasificación, lo que implica que su presencia está más asociada con niveles socioeconómicos relativamente más altos. Algunas variables relacionadas con el saneamiento no fueron incluidas en los modelos por su alta correlación con VIV_3, pero indican que el acceso al agua suele estar ligado a mejores condiciones básicas en la vivienda.

Los resultados sugieren una relación entre el nivel socioeconómico (NSE) y los indicadores de acceso al agua. Sin embargo, los indicadores captados de INEGI no son suficientes para definir esta relación. El estudio invita a la recolección de nuevos indicadores para encontrar el problema del acceso al agua potable por niveles socioeconómicos bajos.

4.2 Trabajo futuro

El indicador de suministro intermitente, o nuevos indicadores, podría ayudar a evaluar mejor la relación entre el nivel socioeconómico (NSE) y el acceso al agua. Esto es especialmente importante porque, en algunas zonas o temporadas, aunque exista conexión a la red pública, el abastecimiento no es constante ni garantizado. Futuros estudios

deberían incluir esta variable para un análisis más preciso. Además, dado que el último censo de INEGI se realizó en 2024, será posible comparar los resultados de los modelos con estos datos más recientes.

Por otro lado, al sugerir una relación entre la pobreza y el acceso al agua, surge la necesidad de discutir las acciones necesarias para atender a los grupos con menor disponibilidad, así como las implicaciones sociales y económicas que esto conlleva.

Bibliografía

- [1] U. Water, "Who/unicef joint monitoring program for water supply, sanitation and hygiene.." <https://www.unwater.org/publications/who/unicef-joint-monitoring-program-update-report-2023>, 2023.
- [2] INEGI, "Encuesta nacional de calidad e impacto gubernamental (encig) 2023.." <http://ctan.org/pkg/geometry>, March 2024.
- [3] S. de Salud, "Nom-127-ssa1-2021. agua para uso y consumo humano. límites permisibles de la calidad del agua," 2021. Diario Oficial de la Federación, México.
- [4] . M. P. Molle, F., *Water poverty indicators: conceptual problems and policy issues*. Water Policy, 5, 529-544 ed., May 2003.
- [5] U. N. D. Programme, *Human Development Report 2006: Beyond scarcity – Power, poverty and the global water crisis*. Palgrave Macmillan, 2006.
- [6] UNICEF and World Health Organization, "Progress on drinking water, sanitation and hygiene: 2017 update and sdg baselines," 2017. Geneva: WHO and UNICEF.
- [7] . C. J. Feitelson, E, *Water poverty: Towards a meaningful indicator*, vol. 4. Water Policy, December 2002.
- [8] C. G. E., "Desabasto de agua potable en algunas colonias en el área metropolitana de guadalajara; incumplimientos y consecuencias políticas.," *Opera*, vol. 31, Jan 2001.
- [9] W. U. W. W. A. Programme), *The United Nations World Water Development Report 2019: Leaving No One Behind*. UNESCO, 2019.
- [10] . L. P. J. A. Guevara Sanginés, A., "Agua, pobreza y uso del tiempo en México: Análisis cuantitativo como sustento del diseño de una política pública de doble dividendo.," *Nova Scientia*, vol. 7, 2015.
- [11] M. De Los Ángeles Mendieta, "Agua y pobreza, un vínculo lamentable," *Revista Especificar*, Nov 2017.

- [12] E. Figueroa, "La población en riesgo y la calidad del agua al sur de la zona metropolitana de guadalajara," *Agua y Territorio / Water and Landscape*, Jan 2021.
- [13] M. in Data Science, "How data can tackle the clean water crisis." <https://www.mastersindatascience.org/resources/how-data-science-helps-conquer-the-global-water-crisis/>, Mar 2020.
- [14] . L. M. R. Lopez-Gunn, E., "Re-thinking water scarcity: Can science and technology solve the global water crisis?," *Natural Resources Forum*, vol. 32, Aug 2008.
- [15] U. Nations, "Resolution adopted by the general assembly on 28 july 2010." <https://undocs.org/Home/Mobile?FinalSymbol=A%2FRes%2F64%2F292&Language=E&DeviceType=Desktop&LangRequested=False>, Jul 2010.
- [16] CONAGUA, "Programa nacional hídrico 2020–2024," 2020. Gobierno de México.
- [17] G. Pérez-Peña, O. Torres-González, "La insaciable sed de agua de la zona metropolitana de guadalajara," *En Renglones, revista del ITESO*, vol. 49, Aug 2001.
- [18] INEGI, "Principales resultados por ageb y manzana urbana 2020." <https://www.inegi.org.mx/app/scitel/Default?ev=10>, 2020.
- [19] AMAI, "Estimaciones nse 2020 y regla amai 2022." <https://www.amai.org/NSE/index.php?queVeo=NSE2020>, 2020.
- [20] T. H. R. T. Gareth James, Daniela Witten, *An Introduction to Statistical Learning with Applications in Python*. Springer, 1 ed., uly 2023.