ANÁLISIS PREDICTIVO DE DATOS ABIERTOS PARA DETERMINAR EL CRECIMIENTO DE LA POBLACIÓN Y CONSUMO DE AGUA POTABLE EN LA REPÚBLICA MEXICANA

PREDICTIVE ANALYSIS OF OPEN DATA TO DETERMINE POPULATION GROWTH AND POTABLE WATER CONSUMPTION IN THE MEXICAN REPUBLIC

José Luis Hernández González
Tecnológico Nacional de México / IT de Apizaco, México
luis.hg@apizaco.tecnm.mx

Rodolfo Eleazar Pérez Loaiza
Tecnológico Nacional de México / IT de Apizaco, México
rodolfo.pl@apizaco.tecnm.mx

Perfecto Malaquías Quintero Flores
Tecnológico Nacional de México / IT de Apizaco, México
perfecto.qf@apizaco.tecnm.mx

Recepción: 4/noviembre/2024 **Aceptación:** 16/marzo/2025

Resumen

Se presenta el análisis exploratorio y predictivo de la base de datos del Consejo Nacional de Población (CONAPO), con la finalidad de identificar variables de interés para cuantificar el consumo de agua potable en los diferentes Estados de la República Mexicana, de acuerdo con la agenda 2030, el agua forma parte de los Objetivos del Desarrollo Sostenible (ODS) y México ha adoptado tales objetivos. Por otro lado, se requiere conocer el crecimiento poblacional de los Estados y las Ciudades, con la finalidad de establecer políticas públicas que ayuden a disminuir tanto la problemática del agua así como de la vivienda, es decir, el presente trabajo identifica la validez del aprendizaje supervisado con regresión polinomial y comparar la predicción con la base de datos abiertos: “Proyecciones de Población de Municipios de México”, con la finalidad de determinar el crecimiento poblacional de los Estados y tener información que permita estimar la cantidad de agua requerida por la población; es importante hacer notar que, aunque algunas ciudades cuentan

con medidores del consumo de agua potable, no hay registros con estadísticas reales, ni tampoco históricos del consumo. Asimismo, se realiza un análisis exploratorio de la base de datos por medio del software estadístico R y se construyen diagramas de puntos del crecimiento poblacional, también, se realiza un análisis de regresión polinomial hasta el año 2010 y su análisis predictivo al año al 2070, se muestran los Estados con mayor población para cada 10 años a partir de 1970 al 2070 y se calcula el error absoluto respecto a la base de la CONAPO y el modelo de regresión cúbico para cada Estado.

Palabras Clave: Agua Potable, Análisis Exploratorio, Datos Abiertos, Regresión, Población.

Abstract

The exploratory and predictive analysis of the National Population Council (CONAPO) database is presented, with the aim of identifying variables of interest to quantify potable water consumption in the different states of the Mexican Republic. According to the 2030 Agenda, water is part of the Sustainable Development Goals (SDGs), and Mexico has adopted these objectives. Additionally, it is necessary to understand the population growth of the states and cities to establish public policies that help mitigate both water and housing issues. In other words, this work assesses the validity of supervised learning using polynomial regression and compares the predictions with the open data source: "Population Projections for Municipalities of Mexico," in order to determine the population growth of the states and obtain information that allows for estimating the amount of water required by the population. It is important to note that, although some cities have potable water consumption meters, there are no real statistics or historical records of consumption.

Additionally, an exploratory analysis of the database is conducted using the statistical software R, scatter plots of population growth are constructed, and a polynomial regression analysis is performed up to the year 2010, with a predictive analysis extending to 2070. The states with the highest populations for every 10 years from 1970 to 2070 are shown, and the absolute error is calculated in relation to the CONAPO database and the cubic regression model for each state.

Keywords**:** *Potable Water, Exploratory Analysis, Open Data, Regression, Population.*

1. Introducción

De acuerdo con el Informe de las Naciones Unidas sobre Recursos Hídricos 2024, se estima que el consumo de agua para agricultura es del 70%, el 20% para uso industrial y el 10% para uso doméstico; sin embargo, la dinámica poblacional, los hábitos, los cambios de dieta y principalmente la urbanización de las poblaciones, generan nuevas demandas de uso de agua [UNESCO, 2024], el presente trabajo se basa en el consumo de agua potable para uso doméstico.

De acuerdo con la Organización de las Naciones Unidas (ONU) y según la agenda 2030 que es un plan de acción mundial con la finalidad de asegurar el progreso social, económico sostenible y asegurar la paz, la cual ha sido adoptada por el Gobierno de México, se establecen 17 Objetivos del Desarrollo sostenible, y ya se han seleccionado los siguientes objetivos para atender en el presente artículo [PND, 2024], [ONU, 2024]: Objetivo 6.- Agua limpia y saneamiento y Objetivo 11.- Ciudades y comunidades sostenibles.

Sin embargo, en las problemáticas de las ciudades pueden intervenir otras variables como la infraestructura, la desigualdad y el clima. Aunado a tales problemáticas, las ciudades no cuentan con mecanismos suficientes para la medición del consumo de agua; se han cuantificado las pérdidas de agua en fugas en distintas ciudades, pero no se tienen estadísticas del consumo real de agua a través de medidores.

Uno de los aspectos importantes a considerar en las políticas públicas, es el conocer la dinámica de las poblaciones, la urbanización juega un papel muy importante, ya que se debe dotar de servicios públicos a la población, se requiere atender nuevos problemas de carácter ambiental y los actuales desafíos sociales como son la movilidad, la recolección de basura, el saneamiento de espacios, entre otros, de aquí que se considera importante conocer esa dinámica poblacional que permita estimar la cantidad de recursos deseables para el bienestar del ser humano.

Actualmente se cuentan con algunas bases de datos abiertos que se considera pueden proporcionar información que permita conocer algunos rasgos del consumo

de agua y el crecimiento de vivienda urbana en los Estados, así como algunos hábitos de la población: censos de “población y vivienda” y las “proyecciones de la población de los municipios de México de 1950 a 2070” [CONAPO, 2024].

Se realiza el tratamiento de las bases de datos de la CONAPO, así como un análisis exploratorio y un análisis predictivo mediante regresión polinomial, se realiza un análisis de regresión y correlación para obtener modelos analíticos del crecimiento de los Estados. Se han generado gráficas de puntos a partir del año 1970 hasta el 2070 e identificar el tamaño de población por Estado, así como el análisis de regresión cúbico considerando la información al año 2010, se valida la ecuación con datos del año 2010 al año 2020 y se realiza la proyección. El propósito es establecer una estrategia que permita conocer el comportamiento del consumo de agua potable a través de datos abiertos como son las bases de datos del Consejo Nacional de Población, a través del crecimiento de la población, se requieren de $100\ litros$ de agua por día por persona.

2. Métodos

En el presente trabajo se describe el análisis exploratorio y predictivo realizado sobre una base de datos abierta referente a tema del agua se considera la población a mitad del año, 1950-2070, cuya fuente es la Base de datos con la estimación de la población desagregada por edad y sexo a nivel nacional y por entidad federativa (1970-2070) de la CONAPO. Existen diferentes bases de datos respecto a las proyecciones por municipios y corresponden a las proyecciones de la población de México y a las entidades federativas del 2016 al 2050, Bases de Datos de la Conciliación Demográfica 1950 a 2019 y Proyecciones de la población de México 2020 a 2070, tal información, se ha generado con la finalidad de contar con una herramienta que permita conocer características demográficas de la población, se basan en las variaciones de las tasas de nacimiento, de mortandad, así como cambios de la población debido a la migración. De acuerdo con el documento metodológico de las proyecciones de población de los municipios de México del 2010 al 2030 [CONAPO, 2024], se describe la siguiente metodología:

- Suavizamiento de las series de tiempo por medio del algoritmo de Gray.

- Prorrateo de información faltante o no especificada.
- Ajuste por prorrateo de la población por edad.

Se ha actualizado el estudio y se ha publicado las Bases de datos de la Conciliación Demográfica 1950 a 2019, Proyecciones de Población de los municipios de México 2010 a 2030 y Proyecciones de la población de México 2020 a 2070, se cuenta con cifras sobre la población al inicio y a mitad del año, para este trabajo, se consideró la información de mitad del año.

El objetivo del presente trabajo es estimar la población mediante métodos de aprendizaje supervisado de la Inteligencia Artificial a partir del año 2024. Mediante un modelo de regresión polinomial por lo que se selecciona como población, los datos históricos del año 1970 a 2010 y se utiliza como muestra para realizar el proceso de validación del modelo del año 2010 a 2020. Hussen describe métodos de regresión para el aprendizaje Automático [Hussen y Moshin, 2024].

Dado que la base de datos se encuentra estratificada por edades y sexo, se ha realizado una limpieza y adecuación del conjunto de datos, se han definido las variables Población Total, Población Hombre y Población Mujeres, así como la población total por cada Estado. Se han generado dos archivos de trabajo, tanto para realizar los análisis de regresión, como para la construcción de mapas interactivos para visualizar la información, los formatos utilizados son .cvs y el .shp es un formato shapefile para almacenar información geográfica con atributos de las entidades, formados por medio de puntos, líneas o polígonos.

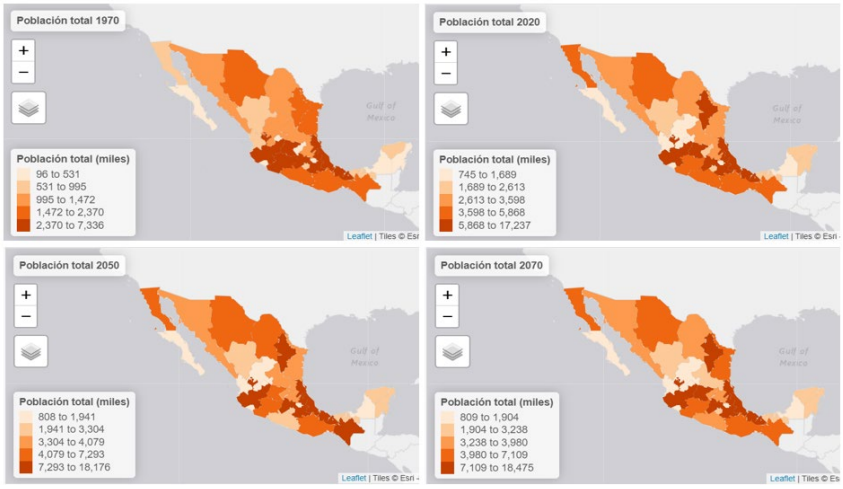
Se ha realizado un análisis exploratorio descriptivo de los Estados por año, asimismo, se han determinado valores mínimos y máximos de la población por año. También, se han generado algunas gráficas de cajas y gráficas de puntos (diagramas de dispersión). La Tabla 1 muestra el crecimiento de los Estados de la base de datos de la CONAPO y las proyecciones para el año 2070, se han seleccionado los 5 Estados con mayor población cada 10 años a partir de 1970 para observar el cambio en el crecimiento de los Estados, en realidad los cambios se presentan en diferentes años y no cada 10; sin embargo, se optó por la medida para simplificar la información. En la Tabla 1, se presenta la población en miles y se

ordenó el conjunto de datos del mayor al menor, de ahí se identificaron los Estados. La Figura 1 muestra en forma gráfica la información de la Tabla 1 a través de mapas de intensidad, tales mapas se han construido en R [R Core Team, 2024].

Tabla 1 Crecimiento cada 10 años de los 5 Estados con mayor población.

1970		**1980**		**1990**		**2000**	
7336.32	MX	8960.023	MX	10253.144	E. México	13311.647	E. México
4124.264	E. México	7619.093	E. México	8485.531	MX	8706.087	MX
4085.275	Veracruz	5482.746	Veracruz	6433.25	Veracruz	7045.73	Veracruz
3523.706	Jalisco	4498.784	Jalisco	5579.963	Jalisco	6467.806	Jalisco
2722.961	Puebla	3450.129	Puebla	4328.841	Puebla	5208.423	Puebla
2010		**2020**		**2024**		**2030**	
15529.76	E. México	17236.788	E. México	17616.018	E. México	18144.619	E. México
9044.151	MX	9332.593	MX	9204.018	MX	9259.907	Jalisco
7806.88	Veracruz	8507.662	Jalisco	8820.915	Jalisco	9032.705	MX
7517.154	Jalisco	8169.287	Veracruz	8127.727	Veracruz	8046.71	Veracruz
5935.838	Puebla	6705.217	Puebla	6989.402	Puebla	7443.37	Puebla
2040		**2050**		**2060**		**2070**	
18474.722	E. México	18175.667	E. México	17351.496	E. México	16133.593	E. México
9784.753	Jalisco	10051.671	Jalisco	10076.421	Jalisco	9868.59	Jalisco
8609.423	MX	8430.362	N. León	8869.543	N. León	9078.596	N. León
8018.826	Puebla	8365.647	Puebla	8513.178	Puebla	8461.898	Puebla
7769.371	N. León	8015.397	MX	7899.115	Chiapas	8074.413	Chiapas

Fuente: Elaboración propia.



Fuente: Elaboración propia.

Figura 1 Crecimiento de la población, años 1970, 2020, 2050, 2070.

Aunque se han seleccionado los mapas para cada 10 años, es deseable construir mapas para años intermedios. Los mapas son generados mediante la librería ggplot, tmap, entre otras y se pueden generan mapas estáticos o dinámicos, lo que posibilita la generación de informes o es su caso, el análisis interactivo, ya que se permite seleccionar alguno de los Estados y comprobar el dato graficado, si se

requiere, es posible agregar más información como la población del año seleccionado, o alguna de las variables de interés, además, es posible si se requiere trazar o dibujar polígonos o áreas de influencia en los mapas. Se pueden incluir casillas de verificación para ocultar o mostrar la información.

Los modelos de regresión lineal son considerados de aprendizaje supervisado, se dividen en dos fases principales: fase de entrenamiento y fase de prueba o de validación del modelo, la finalidad es realizar un análisis predictivo, lo que permitirá estimar la población a futuro, según la Comisión Nacional del Agua (CONAGUA) [CONAGUA, 2024], la agenda 2030, se considera “Cobertura universal” y es necesario garantizar el suministro de agua potable y alcantarillado a 20 años. Se requiere hacer un análisis predictivo de los periodos 2020 al 2040, dado que se cuenta con un análisis predictivo de la CONAPO, se validan los modelos de regresión obtenidos de 1970 a 2020.

Medina [Medina, 2024] presenta una propuesta que permite simplificar el modelo de crecimiento poblacional y, que tradicionalmente, se basan en el lineal, el exponencial y el logístico. En un análisis preliminar se observó que el análisis de regresión exponencial tiene un coeficiente más bajo de correlación que el lineal y el polinomial. Aunque el modelo lineal parece funcionar muy bien, se ha optado por usar regresión cuadrática y cúbica para elaborar los informes.

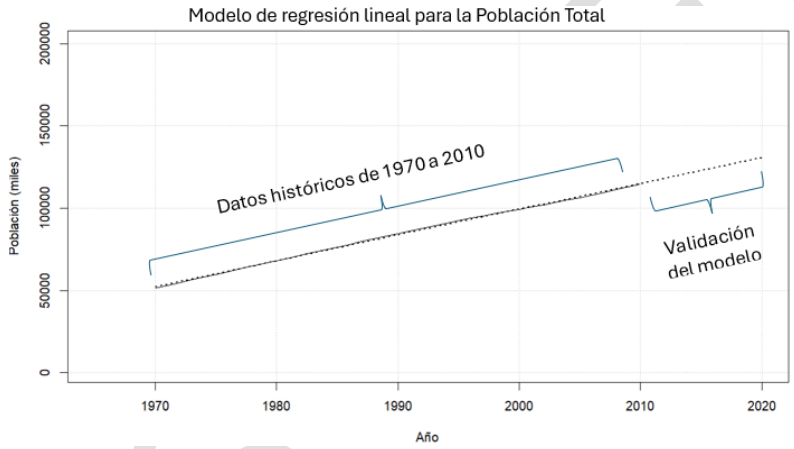
Fases del análisis de regresión y predicción

Para realizar el cálculo de proyecciones al año 2070, además de comparar las gráficas de dispersión y modelos ajustados con las proyecciones de la CONAPO, el trabajo que se explica en este documento, consistió principalmente en el desarrollo de un análisis de regresión basado en tres modelos (lineal, cuadrático y cúbico); así también, se explica el cálculo del error absoluto entre ambos conjuntos de datos, los planteados por la CONAPO y los generados por el modelo de regresión, con la finalidad de establecer un valor numérico que permita medir la sensibilidad de los ajustes encontrados:

- Se han definido tres períodos de tiempo: Se trabajó con la base de datos para los años desde 1970 al 2010 para realizar en análisis de regresión

polinomial, en tal análisis, se prueban tres ajustes: lineal cuadrático y cúbico, se grafican los diagramas de dispersión junto con el polinomio ajustado.

- Se seleccionaron datos del año 2010 a 2020 para validar el entrenamiento respecto a los datos originales (en algunos casos, el ajuste cuadrático es semejante al lineal, y al cúbico), dado el comportamiento que describe la CONAPO, se eligió el modelo cúbico. La Figura 2 muestra la gráfica del diagrama de dispersión y su ajuste de curva, además, se sobrepone la gráfica con el resultado de la validación del modelo.
- El análisis predictivo se realizó considerando los datos del año 2020 al 2070.



Fuente: Elaboración propia.

Figura 2 Proceso de validación del modelo.

El cálculo del error absoluto de las predicciones de la CONAPO y de las predicciones mediante regresión de la ecuación cúbica, se realiza para cada Estado, se considera como valor objetivo el de la CONAPO, dando lugar a la Ecuación 1 del cálculo de error absoluto.

$$Error\ absoluto\% = \sum_{i=1970}^{2070} \frac{|Población\ regresión_i - Población\ CONAPO_i|}{Población\ CONAPO_i} \qquad (1)$$

Se ha optado por considerar en el cálculo, “todos” los elementos de la serie, es decir del año 1970 a 2070, aunque lo ideal es realizar el cálculo a partir del año 2020; sin embargo, para los años donde se realizó el entrenamiento el error es cero. Terver y colaboradores [Terven, et al., 2024] en su artículo “Loss Functions and Metrics in

Deep Learning. A review" definen el error medio absoluto (MAE) usado comúnmente en funciones de pérdida para problemas de regresión, Ecuación 2. Donde MAE es el error medio absoluto, y valor observado (población), $\hat{y}$ pronóstico de la población y n total de años para la predicción

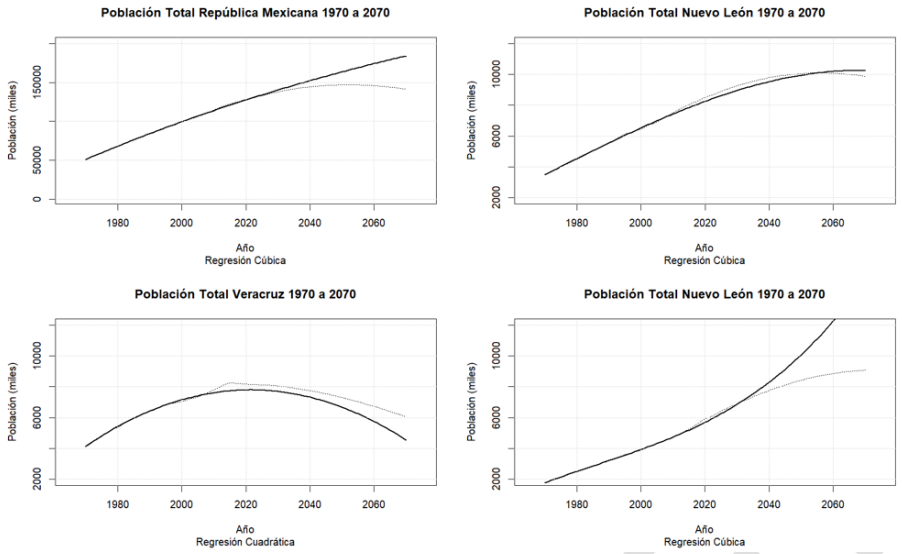
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (2)$$

Para este estudio se utilizó el concepto de error absoluto de los métodos numéricos y no dividir en el total de datos ya que los años no serán considerados como un promedio eliminando el cociente $\frac{1}{n}$, se incluyen las ecuaciones de regresión cúbica por Estado.

3. Resultados

La Figura 3 muestra algunos Estados seleccionados con modelos de regresión cúbica para la Población Total en el Estado de Jalisco, donde se aprecia que existe un buen comportamiento de acuerdo con las proyecciones de la CONAPO. En la Figura 3, también se muestra el caso del Estado de Veracruz, donde se ha optado por realizar un ajuste cuadrático, aunque se ha elaborado un fragmento del programa para graficar los modelos de regresión lineal, cuadrático y cúbicos, se optó por calcular el error absoluto considerando como real el dato de la proyección de la CONAPO y el valor de predicción del modelo cúbico que se muestra en la Tabla 2.

De acuerdo con la Ecuación 1, el error en el estado de Veracruz para regresión cuadrática es de 5.58% y el error para regresión cúbica es de 49.57%. La Tabla 3 muestra la población obtenida mediante regresión cúbica, aún no se identifica el modelo que muestra el menor error absoluto, ya que, en algunos Estados, la regresión cuadrática tiene mejor ajuste. Aun así, se considera que los comportamientos para los últimos 20 años (1950 a 1970), tienen una alta variación en el pronóstico, por lo que se valora la posibilidad de realizar un trabajo futuro incluyendo otros indicadores, así como más información para entrenar los modelos de regresión.



Fuente: Elaboración propia.

Figura 3 Modelos de regresión asociados a la gráfica de puntos.

Tabla 2 Cálculo de error absoluto respecto a la CONAPO, regresión cuadrática.

Estado	$Error\ absoluto\%$	Estado	$Error\ absoluto\%$	Estado	$Error\ absoluto\%$
Aguascalientes	17.46	Guerrero	10.10	Quintana Roo	42.04
Baja California	42.95	Hidalgo	7.55	San Luis Potosí	1.12
Baja California Sur	22.57	Jalisco	4.63	Sinaloa	12.86
Campeche	22.07	E. de México	13.09	Sonora	15.24
Chiapas	12.31	Michoacán	17.14	Tabasco	29.23
Chihuahua	15.7	Morelos	13.03	Tamaulipas	27.07
CDMX	11.44	Nayarit	10.01	Tlaxcala	20.82
Coahuila	2.84	Nuevo León	4.48	Veracruz	5.58
Colima	20.27	Oaxaca	4.83	Yucatán	4.14
Durango	10.75	Puebla	7.66	Zacatecas	15.58
Guanajuato	4.24	Querétaro	11.99	Población Total	6.65

Fuente: Elaboración propia

Tabla 3 Proyección de la población mediante regresión cúbica.

2040		**2050**		**2060**		**2070**	
21122	E. México	22383	E. México	23351	E. México	24026	E. México
9942	Jalisco	10685	Jalisco	11384	Jalisco	12038	Jalisco
8497	Chiapas	9340	Puebla	10185	Puebla	11812	Baja California

Fuente: Elaboración propia.

En la Tabla 2, se muestra el cálculo del error absoluto considerando como valor verdadero la proyección de la CONAPO, el ajuste calculado corresponde a un modelo de regresión cúbico, se muestra como en algunos Estados, se incrementa el valor del error como es el caso de Baja California o Quintana Roo, para esos

estados se ha repetido el análisis considerando un ajuste cuadrático. El incremento en el valor del error es conocido como overfitting y underfitting o en su defecto, subajustar o sobreajustar, tal efecto puede producir una predicción deficiente, una de las recomendaciones que se hace, es cambiar el grado del polinomio. La Tabla 4 presenta las ecuaciones de regresión obtenidas, donde $\hat{y}$ es el pronóstico de la población y x años (1970 a 2070)

Tabla 4 Ecuaciones de regresión obtenidas por estado.

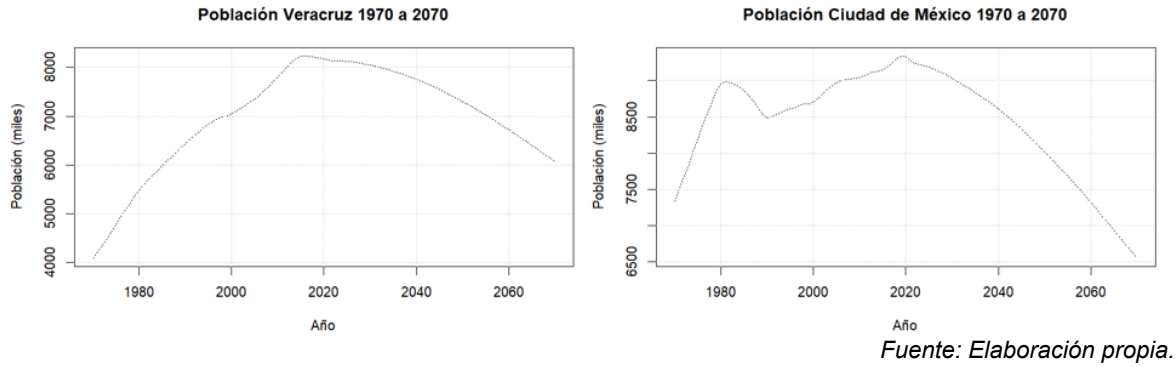
Aguascalientes	$\hat{y} = 294718.5 - 317.1288\,x + 0.08512796x^2$
Baja California	$\hat{y} = 3113102 - 3187.328\,x + 0.8160001x^2$
Baja California Sur	$\hat{y} = 600941.5 - 615.5796\,x + 0.1576678x^2$
Campeche	$\hat{y} = -108817 + 95.61291\,x - 0.02042557x^2$
Chiapas	$\hat{y} = 630248.5 - 714.4661\,x + 0.2006775x^2$
Chihuahua	$\hat{y} = 364820 - 411.4777\,x + 0.1152937x^2$
Ciudad de México	$\hat{y} = -4307674 + 4314.979\,x - 1.078347x^2$
Coahuila	$\hat{y} = -440496.2 + 404.7317\,x - 0.09164447x^2$
Colima	$\hat{y} = 55709.73 - 65.273\,x + 0.018845x^2$
Durango	$\hat{y} = -725518.9 + 714.7316\,x - 0.1756058x^2$
Guanajuato	$\hat{y} = -1410968 + 1342.41\,x - 0.3172652x^2$
Guerrero	$\hat{y} = -1589014 + 1555.513\,x - 0.3797233x^2$
Hidalgo	$\hat{y} = 125737.9 - 159.2178\,x + 0.04874832x^2$
Jalisco	$\hat{y} = -1076607 + 988.6465\,x - 0.2235414x^2$
México	$\hat{y} = -6355201 + 6110.602\,x - 1.463212x^2$
Michoacán	$\hat{y} = -2898758 + 2867.734\,x - 0.7081594x^2$
Morelos	$\hat{y} = -499240.5 + 473.7562\,x - 0.1116806x^2$
Nayarit	$\hat{y} = -301031.5 + 291.7026\,x - 0.07035369x^2$
Nuevo León	$\hat{y} = 278896.1 - 349.7186\,x + 0.1061226x^2$
Oaxaca	$\hat{y} = -761056.3 + 723.0658\,x - 0.1703954x^2$
Puebla	$\hat{y} = -103788.6 + 25.94431\,x + 0.01426365x^2$
Querétaro	$\hat{y} = 896908.6 - 934.1948\,x + 0.2432322x^2$
Quintana Roo	$\hat{y} = 1998722 - 2040.49\,x + 0.5207869x^2$
San Luis Potosí	$\hat{y} = -828913.8 + 803.3989\,x - 0.193879x^2$
Sinaloa	$\hat{y} = -2148219 + 2124.345\,x - 0.5244709x^2$
Sonora	$\hat{y} = 224568.7 - 261.5513\,x + 0.07520314x^2$
Tabasco	$\hat{y} = -206972.7 + 171.1525\,x - 0.03335599x^2$
Tamaulipas	$\hat{y} = 957374.6 - 1004.306\,x + 0.2635132x^2$
Tlaxcala	$\hat{y} = 305698.4 - 326.2295\,x + 0.08693784x^2$
Veracruz	$\hat{y} = -5635915 + 5583.633\,x - 1.381049x^2$
Yucatán	$\hat{y} = -157116.3 + 129.1803\,x - 0.02488367x^2$
Zacatecas	$\hat{y} = -724690.7 + 717.8418\,x - 0.1773962x^2$

Fuente: elaboración propia.

4. Discusión

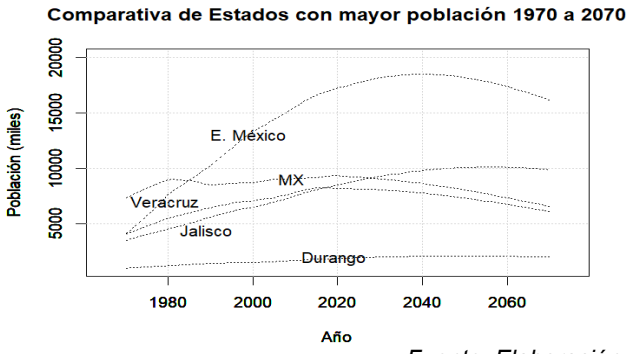
Las gráficas de puntos de 1970 a 2070 en los diferentes Estados de la base de datos de la CONAPO, mantienen las mismas características, es decir, aparentemente un crecimiento poblacional lineal hasta el año 2040 y

posteriormente, se muestra un descrecimiento de la población; sin embargo, hay casos específicos como el de la Ciudad de México, que presenta una disminución notable. Debido a que se ha considerado el total de la información, no se requiere el uso de pruebas estadísticas tanto para los coeficientes de polinomios, ni del coeficiente de regresión. La Figura 4 muestra que al graficar por separado el conjunto de datos y de acuerdo con la escala, se aprecia que algunas funciones tienen comportamiento de funciones cuadráticas, sin embargo, al graficarlas simultáneamente en la misma escala, se observa una tendencia lineal. Además, es importante notar que cuando el conjunto de puntos se ajusta a un modelo lineal, aunque se le aplique un modelo cuadrático o cúbico, los coeficientes se hacen prácticamente cero o son igual a cero, quedando nada más los coeficientes del modelo lineal. La Figura 5 muestra el cambia de las gráficas al modificar los límites de los ejes, dado que la población total es la suma de la población de todos los estados, el eje "y" es relativamente más grande que en el eje "y" para cada estado, se han realizado graficas por separado para todos los estados y se han sobrepuesto algunos estados, además se muestra las gráficas al 2020 que corresponden a los datos reales y se muestran las predicciones de la CONAPO al 2070. Se considera que los valores obtenidos durante la predicción son adecuados en los primeros $10\ a\tilde{n}os$ a partir del periodo de entrenamiento, a partir del año 2040, la CONAPO determinó que disminuye la población y en ese periodo, lo que puede variar es el concepto de *extrapolación*, el cual es un problema para realizar pronósticos, porque los valores de predicción no siempre coinciden con los valores reales.



Fuente: Elaboración propia.

Figura 4 Gráficas separadas, diferentes límites del eje "y".



Fuente: Elaboración propia.

Figura 5 Gráficas simultaneas con escala ajustada al mismo límite del eje y.

Se realizan algunas propuestas para mejorar los modelos de aprendizaje supervisado según Kowsher y otros [Kowsher, et al., 2020] proponen realizar predicción con modelos de métodos numéricos. Jiang y colaboradores [Jiang, et al., 2024] hacen una propuesta para interpolar y extrapolar valores en el aprendizaje supervisado considerando modelos híbridos para mejorar la predicción.

5. Conclusiones

Actualmente se ha dado un gran auge al análisis de datos mediante técnicas de inteligencia artificial, dentro de ellas se encuentra el análisis de datos por medio de aprendizaje supervisado y se hace uso del análisis de regresión, además de conocer el comportamiento del crecimiento poblacional, también se han generado modelos matemáticos lineales y polinomiales que permiten generar una predicción y en particular para pronosticar el crecimiento de las ciudades; en trabajos futuros, además de requerir el tamaño de la población se requiere determinar la cantidad de agua potable, así como la demanda de vivienda.

Se observa que, dados los datos históricos, el pronóstico a 10 y $20\ años$ es adecuado, por lo que es posible extender el análisis para predecir el comportamiento en municipios o ciudades; los scripts de R pueden ser reutilizados y únicamente se requiere cambiar el archivo de datos o incrementar la información del actual. Por otro lado, algunos estudios de infraestructura como dotación de agua y alcantarillado, proponen que la estimación sea de entre 30 a $50\ años$, se requiere incorporar más variables en los modelos de aprendizaje supervisado e incluir las

variaciones de la población que considera CONAPO, como son natalidad, mortandad y migración, aunado a esto, también sería factible incluir otras variables de interés como son las la productividad, fuentes de trabajo en los Estados o ciudades, variables económicas y otras para mejorar la proyección de los datos.

Se extenderá el análisis con la base de datos de Población y Vivienda del INEGI, para incorporar más información, ya se han seleccionado las bases de datos de los censos de los años 2000, 2010 y 2020, así como la base de datos del segundo conteo del 2005.

El archivo original, está conformado con la información de $276\ indicadores$ y más de $192{,}247\ localidades$ por $Estados$ y el total de la República Mexicana. Se han eliminado temporalmente las variables que se considera no aportan información al problema de consumo de agua, quedando como población por $Estado$ y con 88 indicadores de los 276 indicadores de la base original.

6. Bibliografía y Referencias

[1] CONAGUA. Agenda del agua 2030. Avances y logros 2012. https://www.conagua.gob.mx/CONAGUA07/Publicaciones/Publicaciones/SGP-10-12baja.pdf.

[2] CONAPO. Proyecciones de la población de México 2020 a 2070. https://www.gob.mx/conapo/documentos/bases-de-datos-de-la-conciliacion-demografica-1950-a-2019-y-proyecciones-de-la-poblacion-de-mexico-2020-a-2070?idiom=es.

[3] CONAPO. Proyecciones de los municipios de México 2010 a 2030. http://www.conapo.gob.mx/work/models/CONAPO/Resource/1529/1/images/Documento_Metodologico_Proyecciones_de_la_poblacion_de_Mexico_20102050.pdf.

[4] Hussen, D., Moshin, A. A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, vol. 01, No. 4, pp. 140-147. https://www.researchgate.net/publication/348111996_A_Review_on_Linear_Regression_Comprehensive_in_Machine_Learning.

[5] Jiang, B., Zhu, X., Tian, X., Yi, W., Wang, S. Integrating Interpolation and Extrapolation. A Hybrid Predictive Framework for Supervised Learning. Appl. Sci. 2024, 14, 6414: https://www.mdpi.com/2076-3417/14/15/6414.

[6] Kowsher, Md and Uddin, Md. Jashim and Moheuddin, Mir Md and Turaba, Mahbuba Yesmin, Two New Regression and Curve Fitting Techniques Using Numerical Methods (2020). Algorithms for Intelligent Systems, Springer 2020, Available at SSRN. https://ssrn.com/abstract=3590089 or http://dx.doi.org/10.2139/ssrn.3590089.

[7] Medina, I. Modelos simples de crecimiento poblacional desde la perspectiva de flujos-reservorios en la plataforma Stella como herramienta para visualizar elementos que regulan los sistemas dinámicos. Enseñanza y Comunicación de las Geociencias, v. 3, núm. 1, p. 33-40. http://encomunicacionct .geociencias.unam.mx/wp-content/uploads/2024/06/Medina-Gomez-v2.pdf.

[8] Organización de las Naciones Unidas. Página oficial, https://www.un.org/es.

[9] Plan Nacional de Desarrollo 2019-2024. Diario Oficial, Presidencia de la República, Gobierno de México. https://www.gob.mx/cms/uploads/ attachment/file/487316/PND_2019-2024.pdf.

[10] R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

[11] Terven, J., Cordova, D., Ramírez, A. Chavez, E. Loss Functions and Metrics in Deep Learning. A Review. Under Review in Computer Science Review. Computer Science. Machine Learning: https://arxiv.org/abs/2307.02694.

[12] UNESCO. Agua para la prosperidad y la paz. Informe mundial de las Naciones Unidas sobre el desarrollo de los Recurso Hídricos 2024. (2024). https://agua.org.mx/wp-content/uploads/2024/04/agua-para-la-paz-Resumen-ejecutivo_UNESCO.pdf.