DOI: https://doi.org/10.5281/zenodo.16944357

Ciencia de datos aplicada al análisis de asociación entre la necesidad de servicios de salud en población dependiente y el índice de marginación urbana por AGEB en 2020

Data science applied to the analysis of the association between the need for health services in dependent population and the index of urban marginalization by AGEB in 2020

Angel Fernando Argüello Ortiz[1] *y Ángel J. Sánchez-García*[2]

Resumen: En este estudio se realiza un análisis de los microdatos del Censo de Población y Vivienda 2020, en el que se procesaron 387,205,920 datos estatales a nivel de AGEB, que, bajo un proceso de curación, culminaron en una base de 49 mil 586 registros con 19 variables, con las que se generaron los indicadores del Índice de Necesidad de Servicios de Salud (INSS), el Índice de Marginación Urbana 2020 (IMU) y el Promedio de Ocupantes por Vivienda. Utilizando técnicas de análisis espacial, ciencia de datos y aprendizaje automático, se implementaron cuatro algoritmos: *multiple linear regression, support vector regression, random forest* y *gradient boosting*, con validación cruzada K-Fold de los resultados. Se identifica una relación estadísticamente significativa entre los indicadores, pero débil, sugiriendo la consideración de otros factores y la aplicación de técnicas de ciencia de datos, destacando el uso de algoritmos ensamblados para mejorar niveles de precisión predictiva en explicabilidad y mayor exigencia computacional.

Palabras clave: área geoestadística básica (AGEB), machine learning, random forest, marginación.

Abstract: In this study, an analysis of microdata from the 2020 Population and Housing Census is conducted, processing 387,205,920 data points at the state level by AGEB (Basic Geostatistical Area), which through a curation process culminated in a database of 49,586 records with 19 variables. These were used to generate indicators for the Health Services Need Index (INSS), the 2020 Urban Marginalization Index (IMU), and the Average Occupants per Dwelling. Using spatial analysis techniques, data science, and machine learning, four algorithms were implemented: Multiple Linear Regression, Support Vector Regression, Random Forest, and Gradient Boosting, validating modelos by K-Fold cross-validation. A statistically significant but weak relationship between the indicators is identified, suggesting the consideration of other factors and the application

1 Docente de tiempo completo de la Facultad de Estadística e Informática, Universidad Veracruzana: aarguello@uv.mx
2 Docente de tiempo completo de la Facultad de Estadística e Informática, Universidad Veracruzana: angesanchez@uv.mx

of data science techniques, highlighting the importance of using ensemble algorithms to improve levels of predictive precision with explainability and greater computational demands.

Keywords: basic geostatistics area (AGEB), machine learning, random forest, marginalization.

Introducción

En la actualidad, el procesamiento, manejo y análisis de grandes volúmenes de información es una garantía para obtener mayor precisión en los resultados de diversos estudios, así como también para identificar la presencia o ausencia patrones de comportamiento; si bien es cierto que requiere de una gran inversión en tiempo de procesamiento, también es cierto que el avance tecnológico nos ha colocado, actualmente, en un punto sin retorno en que el elemento más importante está en las grandes cantidades datos, junto con la amplia capacidad de integración y análisis y la experiencia de programación o cómputo intensivo.

Así, al introducirnos a la era digital, cada uno de los registros de nuestra fuente de información se transforman en granos de arena que abonan a la estrategia de recopilación, acceso, almacenamiento, procesamiento y modelación de los datos, guiándonos en el proceso de transformación digital de la información y al nuevo enfoque del manejo y aplicación de técnicas estadísticas, definido como aprendizaje automático o *machine learning*. Este proceso de transformación también ha modificado la forma de aplicar la estadística, combinándose con la programación o cómputo intensivo y con infraestructura física, marcando un paradigma en la gestión de la información.

En el presente trabajo se conjugaron estos elementos, iniciando con la integración de los microdatos del Censo de Población y Vivienda 2020 a nivel de Área Geoestadística Básica urbana que son producidos por el Instituto Nacional de Estadística y Geografía (INEGI) (INEGI, Instituto Nacional de Estadística y Geografía, 2023) y que concentraron en el apartado de AGEB y manzana urbana un millón 683 mil 504 registros, dando un total de 387 millones 205 mil 920 datos, que fueron procesados para la integración final de una base de datos a nivel de AGEB de 49 mil 586 registros con 19 variables.

Con base en estos datos, se estimó el índice de necesidad de servicios de salud para analizarlo mediante un proceso de modelación, a partir de una regresión lineal múltiple, empleando como variables independientes el Índice de Marginación Urbana 2020 (IMU) estimado por el Consejo Nacional de Población (CONAPO) (CONAPO, 2021) y el Promedio de Ocupantes por Vivienda registrado por el INEGI. Se integraron las bases de datos correspondientes mediante el lenguaje de programación R y se realizaron los modelos de Machine Learning con las diferentes bibliotecas que el lenguaje de

programación Python provee. Cuatro modelos fueron probados y validados para este trabajo: regresión lineal múltiple, support vector regression, random forest y gradient boosting.

Problema

El presente capítulo se considera como una estrategia de análisis conjunto entre las metodologías y herramientas para el desarrollo de proyectos de ciencia de datos enfocado en el manejo de grandes volúmenes de información sociodemográfica a nivel nacional, implementación de metodologías estadísticas y desarrollo de técnicas de cómputo intensivo, con lo cual se destacan dos elementos importantes: el primero de ellos es la necesidad de implementar técnicas de ciencia de datos para el manejo de grandes volúmenes de información, como en este caso, de los 387 millones 205 mil 920 datos del Censo de Población y Vivienda 2020 a nivel de AGEB y manzana urbana producidos por el INEGI en un millón 683 mil 504 de registros y 19 variables; y el segundo es la aplicación de metodologías estadísticas para el procesamiento y análisis de dichos datos, que conjuntamente, dotan al estado de herramientas y criterios para la planeación y ejecución de políticas públicas en materia de salud y desarrollo.

Existe una gran diversidad de necesidades sociales insatisfechas en la población que generan un desequilibrio en el binomio gobierno-sociedad, plasmado en los niveles de atención de la gobernabilidad democrática, tales como: salud, educación, empleo, obra pública y seguridad.

Es así como se considera importante analizar las condiciones de necesidad de disponer de servicios de salud entre la población en edad dependiente, relacionado con los niveles de marginación urbana a nivel de Área Geoestadística Básica (AGEB), como una limitante significativa en materia de rezago. Para ello se implementan elementos y metodologías de ciencia de datos para el procesamiento de 387 millones 205 mil 920 datos para la conformación de las bases de información que permitirán destacar la necesidad de implementar procesos de ciencia de datos como *machine learning* y abonar en los criterios de planeación de las políticas públicas a niveles nacional, estatal y municipal.

Fundamentación

El acceso a los servicios de salud por parte de la población es un derecho humano consagrado en el artículo 4° de la Constitución Política de los Estados Unidos Mexicanos, en el que se establece que “Toda persona tiene derecho a la protección de la salud” y que es una atribución y responsabilidad estricta del Estado de proporcionar dichos servicios y definir las bases bajo las cuales se dará cumplimiento (Cámara de Diputados del H. Congreso de la Unión, 2021); mediante la Ley General de Salud en su artículo 2° se definen las ocho finalidades del derecho a la protección de la salud

de la población, destacando entre ellas no por más importantes sino relevantes para el estudio: el bienestar físico y mental de la persona y el disfrute de servicios de salud y de asistencia social que satisfagan eficaz y oportunamente las necesidades de la población al tratarse de personas que carezcan de seguridad social, la prestación gratuita de servicios de salud, medicamentos y demás insumos asociados, entre otros 6 (DOF, 2024); dicho acceso está condicionado a la afiliación a una institución pública que brinde dicho servicio público; de otro modo, se tendría que satisfacer esta necesidad mediante servicios privados en los cuales no se requiere ningún tipo de afiliación; y aún a pesar de que el Estado se esfuerza por garantizar a la sociedad un servicio de salud gratuito, expedito y de calidad, no se han alcanzado los niveles deseados, principalmente por los diferentes indicadores socioeconómicos relacionado con los niveles de eficiencia del sistema de salud; por lo que el impacto de la implementación de metodología estadística se basa en manejo, integración y análisis de la información disponible con altos niveles de desagregación por AGEB y manzana urbana, asumiendo que existe una relación entre la necesidad de acceso a los servicios de salud y el nivel de marginación urbana (CONAPO, 2021) junto con el promedio de ocupantes por vivienda, respondiendo parcialmente a la necesidad de estudios con altos niveles de desagregación y a las implicaciones de atender necesidades de derechos humanos y políticas públicas como se indica en el artículo 8 de la Constitución Política del Estado Libre y Soberano de Veracruz de Ignacio de la Llave, que implica la salud como un derecho fundamental para vivir y crecer en un ambiente saludable, ecológicamente equilibrado y sustentable (LEGISVER, 2025).

En otras palabras, mediante la implementación de metodologías y herramientas actualizadas se ubica el presente trabajo como un proceso de *machine learning*, que clasifica el algoritmo empleado como un modelo de ajuste/entrenamiento, ya que se espera que mediante el algoritmo el modelo aprenda; dicho proceso es parte integral de la ciencia de datos (figura 1), misma que es considerada un constructo derivado de la interacción entre la estadística, la computación y el conocimiento del medio, orientado a la extracción de información en grandes volúmenes de datos, lo que le da un enfoque multi, inter y transdisciplinario al relacionar principios y prácticas de diferentes campos tales como las matemáticas, la estadística, demografía, actuaría, inteligencia artificial, ingeniería de datos e ingeniería computacional.

Figura 1. Diagrama de ciencia de datos. Recuperado de https://diegokoz.github.io/intro_ds_bookdown/

Hipótesis

De investigación

La necesidad de servicios de salud en población dependiente está relacionada

con el índice de marginación urbana a nivel de AGEB y el promedio de ocupantes por vivienda.

De machine learning
Para el problema de necesidad de servicios de salud con los microdatos del INEGI, los algoritmos ensamblados de aprendizaje-máquina mejoran estadísticamente la precisión de modelos individuales.

Metodología
Con base en Méndez y otros (2011), este proyecto se encuentra delimitado como una revisión de casos, de tipo observacional, retrospectivo, transversal y descriptivo en el que las bases de datos por AGEB y manzana urbana de los microdatos se encuentran desagregadas por entidad federativa, de tal forma que se fusionaron las 32 bases para integrar una sola a nivel nacional con desagregación a nivel de manzana, quedando una base de datos de un millón 683 mil 504 registros con 230 variables; es decir, un procesamiento de 387 millones 205 mil 920 datos. A partir de la base de datos resultante se seleccionaron 5 variables para la estimación de dos indicadores que dieron origen al índice de necesidad de servicios de salud. De manera conjunta, se integró una variable nueva denominada clave geográfica, integrada por las claves de la entidad federativa, municipio, localidad, área geoestadística básica y manzana, excluyendo los registros con los totales correspondientes a entidad, municipio y localidad.

Bajo la presencia de registros faltantes por entidad se identificó que el estado con menor porcentaje fue Ciudad de México (33.5%) y con mayor porcentaje, Chihuahua (71.8%) (ver anexo 1); esto, aunado al comportamiento de las variables seleccionadas, influyó para decidir integrar una base de datos con 64 mil 313 registros y 5 variables de interés a nivel de AGEB para las estimaciones, calculando el número de registros incompletos en alguna o algunas de las variables, dando como resultado que la variable con menor cantidad de datos faltantes fue POB15_64 y la que registró mayor cantidad fue POB65_MAS; finalmente se concluyó con una base de datos con 49 mil 586 registros y tres variables –las que integrarían el modelo–, los cuales están registrados y coinciden totalmente en la base de datos del INEGI para las AGEB urbanas nacionales; sin embargo, la del índice de marginación urbana del CONAPO no coincide en su totalidad con la del INEGI, es por ello que se emplearon los microdatos.

Como parte del proceso de curación de las base de datos resultó que dos bases de datos –Hidalgo (14) y Quintana Roo (22)– tenían una estructura diferente en su variables regionalizadas; lo que evitaba un proceso inmediato de integración de las bases de datos; posteriormente se identificó que derivado del comportamiento de las variables, no podía utilizarse la base de datos a nivel de manzana, documentando los casos de datos publicados, disponibles y que de manera complementaria indicaban registros completos para la estimación de los indicadores e índice, quedando la base

de datos final a nivel de AGEB conformada con 49 mil 586 registros, comparables con los datos del CONAPO y del INEGI.

Es importante destacar que con el apoyo de análisis espacial se identificó la presencia de 16 mil 915 AGEB urbanas que no registró el Consejo Nacional de Población y que en algunos casos sí se logró asociar en el análisis del Índice de Necesidad de Servicios de Salud mediante los microdatos del INEGI; esto, bajo la consideración de que de acuerdo con el marco geoestadístico nacional, el territorio está divido en 81 mil 451 AGEB, de las cuales 17 mil 469 son rurales y 63 mil 982 son urbanas; sin embargo, en términos de distribución del territorio, el ámbito rural abarca un área de un millón 931 mil km2 y el ámbito urbano tan solo 24 mil 274.5 km2, lo que da una idea de la mala regionalización y amanzanamiento de la población de México (figuras 2 y 3).

Figura 2. Distribución territorial por ámbito urbano-rural

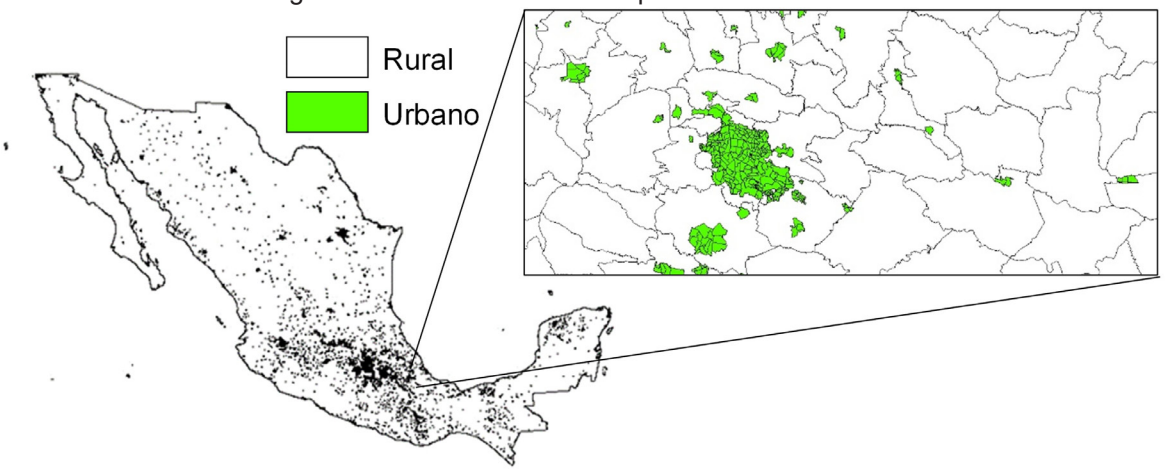
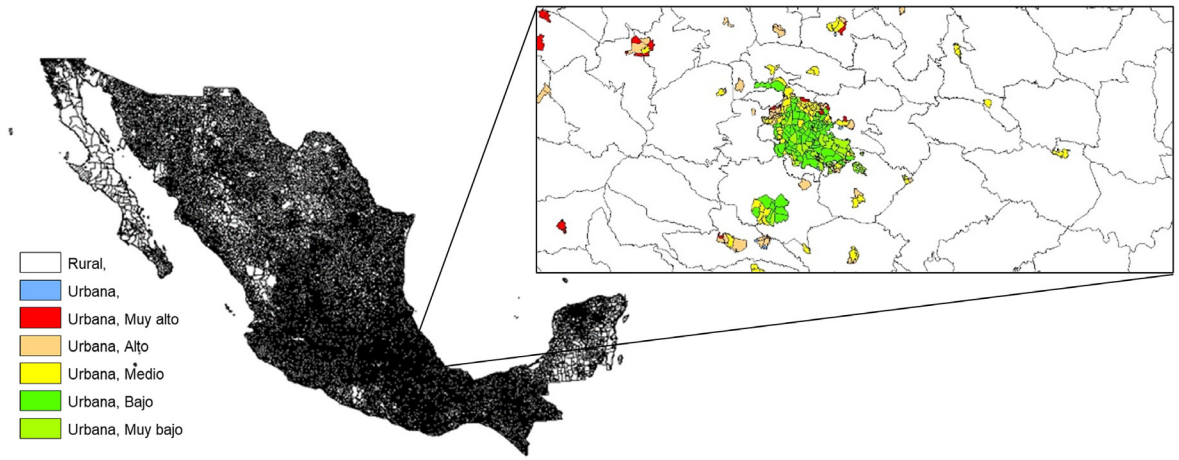


Figura 3. Regionalización urbano-rural por grado de marginación urbana



Las variables empleadas correspondieron a la población total (POB_TOT), la proporción de población sin afiliación a servicios de salud (PPSINDER) y el promedio de ocupantes por vivienda (PROM_OCUP) y la razón de dependencia que explica la carga económica presente en la población productiva (INEGI, 1997), derivadas del censo 2020. Para el caso de la razón de dependencia se integra con tres variables de la siguiente manera:

$$RD = \frac{P_{0-14} + P_{65+}}{P_{15-64}}$$

Donde:

P_{0-14} *es la población de* 0 *a* 14 *años*

P_{15-64} *es la población de* 15 *a* 64 *años*

P_{65+} *es la población de* 65 *años y más*

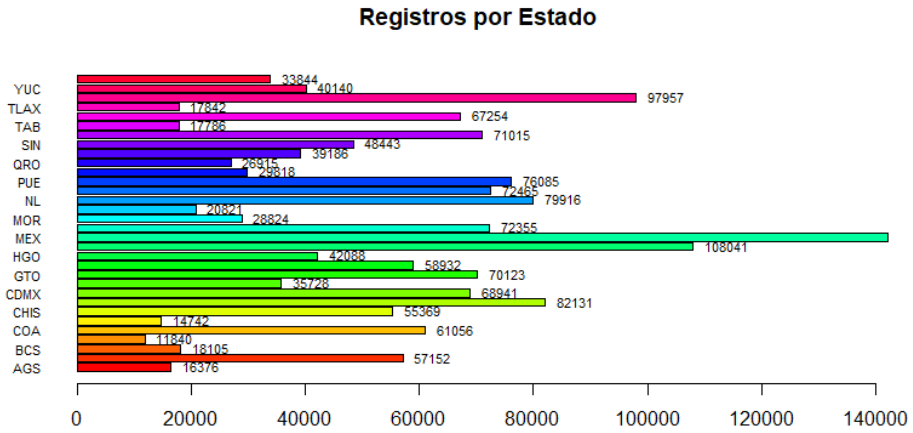
Con la integración y estimaciones correspondientes, se estimó el índice de necesidad a los servicios de salud (INSS) mediante:

$$PPSINDER = PSINDE\,R/P\,OB\_TOT$$

$$INSS = RD * PPSINDER$$

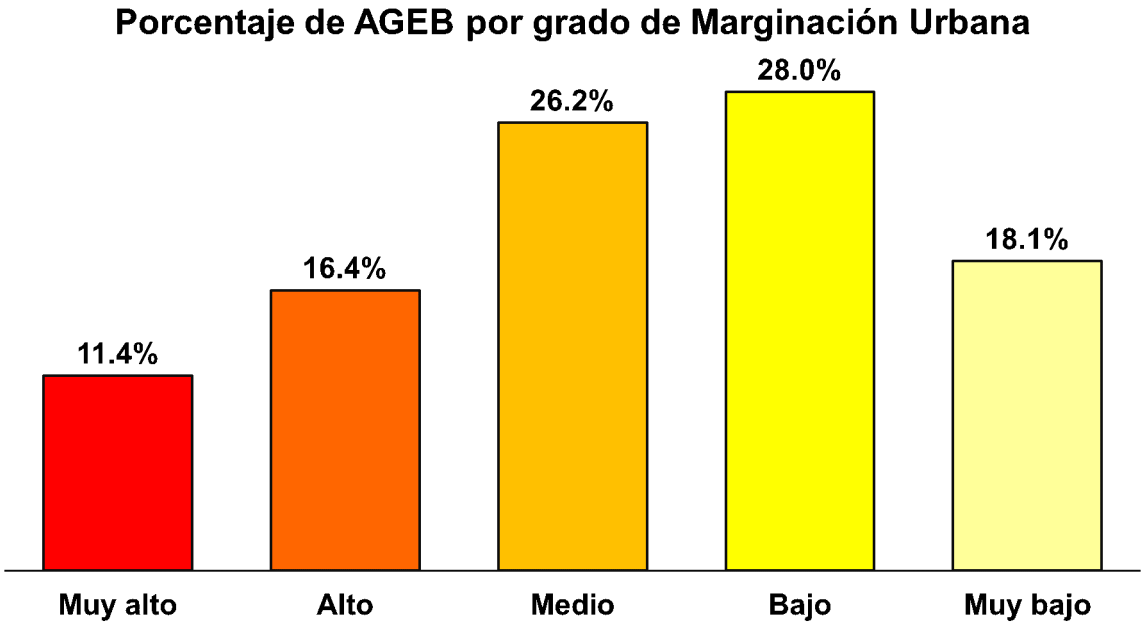
Derivado de estas estimaciones, se aplicaron técnicas de análisis exploratorio para la identificación de registros disponibles, excluidos, integrados por AGEB, conteos de datos faltantes por variable de estudio que provocaban registros incompletos, análisis de dispersión de los datos, correlación y estimación de densidades por variable iniciando con el conteo de registros por estado donde la mayor cantidad de registros a nivel de manzana fue en el estado de México (142,214) y con menor cantidad en Campeche (11,840) como se muestra en la figura 4.

Figura 4. Registros a nivel de manzana por estado



En materia de datos faltantes, los estados con los porcentaje más altos y muy cercanos fueron Chihuahua (71.8%) Baja California Sur (71.5%) y Zacatecas (71.1%), mientras que el estado con menor porcentaje fue Ciudad de México (33.5%); lo que llevó a que la integración de los registros seleccionados para cada estado estuvo entre el 2.5% y el 3.4%; es decir, a pesar de haberse registrado altos porcentajes de valores faltantes y de registros incompletos, la proporción fue representativa, más aún cuando se identificaron los casos por variable. Con relación a la clasificación que realiza el CONAPO del índice de marginación urbana como muy alto, alto, medio, bajo y muy bajo, las AGEB seleccionadas para el análisis resultaron proporcionalmente significativas como se muestra en la figura 5.

Figura 5. Porcentaje de AGEB seleccionados por grado de marginación urbana 2020



La representación de la dispersión de variables se realizó de manera estadística y gráfica. En el primer caso, como se muestra en la tabla 1, se identificaron varianzas muy altas en las variables del índice de necesidad a los servicios de salud (INSS) y el índice de marginación urbana 2020 (IMU_2020), no así en la variable del promedio de ocupantes por vivienda (PROM_OCUP).

Tabla 1. Varianza y desviación estándar de las variables

Variable	Varianza	Desviación estándar
Índice de necesidad a los servicios de salud (INSS)	49.20212	7.014422
Índice de marginación urbana 2020 (IMU_2020)	20.567254	4.535113
Promedio de ocupantes por vivienda (PROM_OCUP)	0.2376682	0.4875122

De manera gráfica, se puede corroborar lo que se identificó de manera estadística; lo cual se ilustra mediante tres representaciones gráficas (figuras 6, 7 y 8), que también se expresa en una de ellas el coeficiente de correlación, en la que se puede apreciar que se registran asociaciones de regular a mala; sin embargo, eso no significa un resultado negativo sino más bien un enfoque a descartar dentro de los posibles criterios de modelación que se refuerzan mediante *machine learning*.

Figura 6. Gráfico de cajas y alambres de las variables de estudio

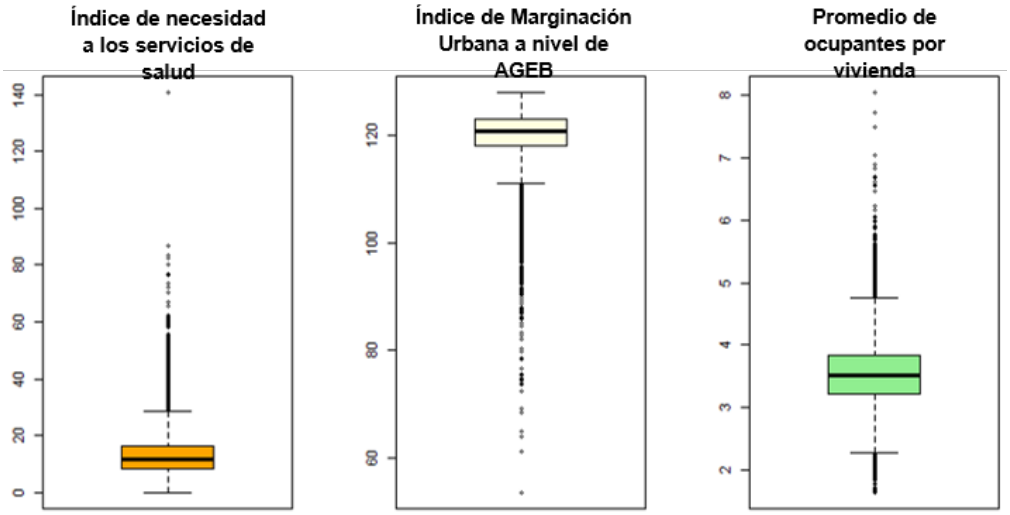


Figura 7. Diagrama de pares con dispersión y densidad de las variables de estudio

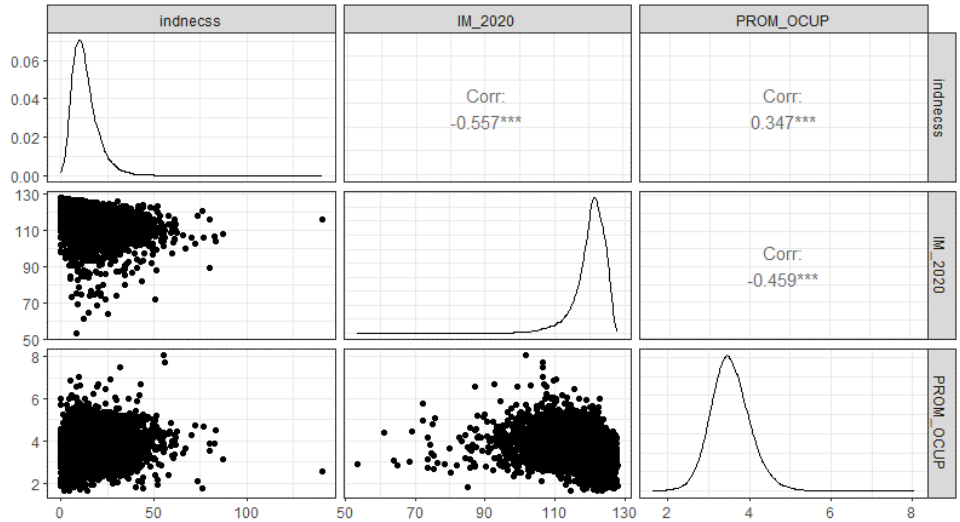
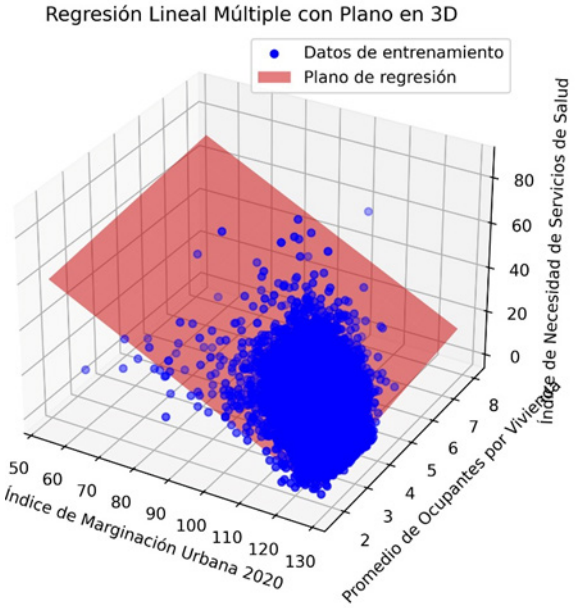


Figura 8. Gráfico 3d de dispersión de las variables bajo estudio



Bajo el enfoque de *machine learning* cada problema tiene una naturaleza específica al momento de seleccionar, plantear, diseñar o programar un algoritmo para su solución y que no existe un algoritmo aplicable a todos los problemas similares o iguales y que a su vez sea el mejor, surge la necesidad de elegir el método más adecuado dadas las características y restricciones del problema en cuestión. Esto se explica mejor mediante los teoremas No Free Lunch (NFL) (Wolpert, 1997), que establecen que no existe ningún algoritmo de optimización o aprendizaje automático que funcione mejor en todos los problemas. En otras palabras, este teorema resalta la importancia de elegir y adaptar la optimización y algoritmos de *machine learning* para problemas específicos en lugar de buscar uno que resuelva todos los problemas; estos algoritmos de *machine learning* se dividen en tres categorías, siendo las dos primeras las más comunes.

- Aprendizaje supervisado.
- Aprendizaje no supervisado.
- Aprendizaje por refuerzo.

El aprendizaje supervisado se compone de dos tareas principales: la clasificación y la regresión, misma que difieren principalmente en el tipo de salida, ya que están diseñados para predecir un valor numérico continuo; mientras que la clasificación se utiliza para predecir la pertenencia a una categoría o clase (el resultado de la predicción es una etiqueta de clase que pertenece a un conjunto discreto y finito de categorías).

Desde el enfoque de la ciencia de datos, para este estudio se emplearon como métodos individuales a la regresión lineal múltiple dada la variable respuesta INSS, asumiendo

la existencia de una relación lineal; posteriormente, se seleccionaron support vector regression (SVR) que es una modificación a support vector machine (SVM), así como árboles de decisión (usualmente utilizados para métodos ensamblados).

La aplicación de un análisis de regresión lineal múltiple para modelar el índice de necesidad de servicios de salud con base en las variables independientes de índice de marginación urbana 2020 y el promedio de ocupantes por vivienda, arrojó un coeficiente de determinación de 0.321, con una correlación múltiple de 0.57, la cual no podría considerarse como buena, por lo que mediante técnicas de *machine learning* se buscará un mejor ajuste. El modelo de regresión quedó estructurado como:

Índice de Necesidad de Servicios de Salud = 100.772491 + -0.779927*Índice de Marginación Urbana 2020+ 1.661177* Promedio de ocupantes por vivienda

La regresión de vectores de soporte o *support vector regression* (SVR) es una técnica de aprendizaje supervisado utilizada para la regresión, la cual se basa en el concepto de *support vector machines* (SVM) y es extendida para el problema de regresión (Cortes, 1995). Al igual que en SVM, los puntos de datos que caen dentro del margen o en el lado incorrecto del margen son considerados vectores de soporte. Estos son los puntos más importantes para la definición de la función de regresión y contribuyen significativamente al modelo; sin embargo, a diferencia de SVM, en SVR la idea principal es encontrar una función que se ajuste a los datos de entrenamiento y, al mismo tiempo, minimice la brecha entre los puntos de datos y la función.

El enfoque en los vectores de soporte y el uso de un margen de tolerancia permiten que el modelo sea robusto y capaz de manejar datos ruidosos o no lineales. SVR brinda la flexibilidad de definir el nivel de error aceptable en el modelo y encontrar una línea apropiada (o hiperplano en dimensiones superiores) para ajustarse a los datos.

Es importante destacar que SVR tiene parámetros sensibles, como el parámetro de regularización y la elección del kernel. La selección adecuada de estos parámetros puede requerir ajuste y validación cruzada o algún mecanismo de optimización. Los kernels proporcionados en “sklearn” son *linear* (lineal), *poly* (polinomial), *rbf* (Radial Function Basis), *sigmoid* (Sigmoide), *precomputed* (Matri Gram). Además, es posible controlar el parámetro ϵ.

Se pueden emplear diferentes métodos de regresión para hacer predicciones basadas en las variables explicativas. Los métodos ensamblados se han utilizado en diferentes dominios (Nti, 2020), y sus rendimientos han superado al de los regresores individuales (Hoc, 2023a) (Hoc, 2023b). Los métodos de regresión ensamblados de aprendizaje automático son técnicas que combinan múltiples modelos de regresión único para construir un modelo más potente y preciso (Lu, 2019). Estos métodos se basan en el principio de que en la combinación de múltiples modelos débiles pueden superar

las limitaciones individuales de cada uno y proporcionar predicciones más precisas y sólidas.

En lugar de depender de un único modelo de regresión, los métodos ensamblados aprovechan el conocimiento de diversidad de modelos y el colectivo para mejorar el rendimiento general del modelo. Los métodos ensamblados, como *bagging*, *boosting* y *stacking* son estrategias que combinan múltiples modelos de aprendizaje para mejorar el rendimiento predictivo en comparación con un solo modelo.

El enfoque de *bagging* se basa en la idea de adiestrar múltiples modelos independientes en conjuntos de datos de entrenamiento generados mediante muestreo con reemplazo (*bootstrap*). Cada modelo se entrena en un conjunto de datos diferente, y las predicciones se promedian (para problemas de regresión) o se votan (para problemas de clasificación) para obtener la predicción final. La esencia del *bagging* es entrenar modelos base en conjuntos de datos diferentes, lo que ayuda a reducir la varianza y mejorar la generalización del modelo ensamblado. *Random forest* es un ejemplo popular de un método basado en *bagging* que utiliza árboles de decisión como base. El algoritmo *random forest* (Breiman, 2001) combina métodos de aprendizaje ensamblados para crear múltiples árboles de decisión extraídos aleatoriamente a partir de los datos, promediando los resultados para generar un nuevo resultado que a menudo conduce a predicciones sólidas.

En este proyecto se utilizó el módulo *sklearn* de Python para entrenar el modelo de regresión de *random forest*, específicamente la función RandomForestRegressor (Geurts, 2006). La documentación RandomForestRegressor presenta muchos parámetros diferentes que se pueden ajustar para el modelo. Algunos de los parámetros importantes son el número de estimadores, la profundidad de cada estimador, número de muestras, entre otros. Por otro lado, a diferencia de *bagging, boosting* asigna pesos a las observaciones durante el proceso de entrenamiento; es decir, las observaciones mal predichas por los modelos anteriores tienen un peso mayor en los modelos subsiguientes, permitiendo que el modelo se enfoque más en los casos difíciles. Bajo este enfoque, los modelos se entrenan secuencialmente, y cada modelo intenta corregir los errores de los modelos anteriores, por lo que la predicción final se obtiene combinando las predicciones ponderadas de cada modelo. Ejemplos de algoritmos: *adaboost, gradient boosting* (por ejemplo, XGBoost, LightGBM).

Gradient boosting es una técnica de ensamblado que construye un modelo predictivo fuerte a partir de varios modelos predictivos débiles, generalmente árboles de decisión poco profundos. Se construye de manera secuencial, donde cada árbol corrige los errores de los árboles anteriores. Es un algoritmo de tipo *boosting*, lo que significa que se enfoca en corregir los errores cometidos por modelos anteriores para mejorar la precisión general.

El algoritmo *gradient boosting* fue propuesto inicialmente por Jerome H. Friedman en 2001. Friedman introdujo el concepto de *gradient boosting* como una técnica de optimización basada en funciones de pérdida. Propuso un enfoque de optimización secuencial donde se ajustan modelos débiles (generalmente árboles de decisión) para corregir los errores de los modelos anteriores. La idea central era construir un modelo aditivo que minimice una función de pérdida. Este algoritmo proporciona un rendimiento predictivo excepcional y es uno de los algoritmos más potentes en términos de precisión. Debido a su naturaleza, es robusto frente a valores atípicos y ruido en los datos debido a la combinación de múltiples árboles.

Para la verificación de modelos, se aplica la validación cruzada (*cross-validation*) que se utiliza en *machine learning* para evaluar el rendimiento de un modelo y mitigar problemas de sobreajuste, siguiendo los siguientes pasos:

- Reservar un conjunto de datos de prueba (test).
- Entrenar el modelo utilizando la parte restante del conjunto de datos.
- Utilizar la muestra de reserva del conjunto de prueba (validación). Esto ayuda a medir la eficacia del rendimiento de un modelo. Si el modelo arroja un resultado positivo en los datos de validación, se continúa con el modelo actual.

Hay varias formas de llevar a cabo la validación cruzada, entre ellas el *k-fold cross-validation* que, en cuanto a su eficiencia computacional, es más rápido al ejecutarse k veces en lugar de *n* veces (n es el número total de datos) y apoya la selección del modelo más adecuado, y considerando la cantidad de datos de este estudio es la mejor técnica con k = 10, con la finalidad de dividir el conjunto en 5 subconjuntos y en cada una de las cinco iteraciones se utilizó el 90% para entrenamiento y 10% para prueba.

Resultados

Para la ejecución de los cuatro algoritmos seleccionados: *multiple linear regression, support vector regression, random forest* y *gradient boosting*, se utilizaron los siguientes parámetros (tabla 2).

Tabla 2. Parámetros de los algoritmos de prueba

Multiple Linear Regression	Support Vector Regression	Random Forest	Gradient Boosting
-	• Kernel = Lineal • Grado de la función de kernel polinomial = 3 • Epsilon = 0.1	• Num de árboles = 50 • Función de calidad de separación = Error cuadrático • Bootstrap = True • Máxima profundidad de árboles = 2	• Num de árboles = 50 • Función de pérdida = Error cuadrático • Función de calidad de separación = Friedman MSE • Tasa de aprendizaje = 0.1

Los resultados obtenidos de la ejecución de la validación cruzada de k-Fold, tomando como métrica de evaluación el R^2 y k=10 (es decir, en cada iteración se utiliza el 90% para entrenamiento y el 10% de datos para prueba). Se aprecia que en general los mejores resultados se obtienen mediante *gradient bossting*; sin embargo, no ocurre en todos los *folds*. Se resaltan en negritas los mejores resultados. En promedio *gradient bossting* tienen mejor R^2 y *random forest* tiene menos variabilidad (tabla 3).

Tabla 3. Resultados de R^2 a través de k-fold como método de validación de modelos

Fold	Multiple Linear Regression	Support Vector Regression	Random Forest	Gradient Boosting
1	0.02486369	0.19228474	0.14054651	0.19629745
2	0.3878044	0.43220827	0.40412577	0.44011975
3	0.2369374	0.27533143	0.27053005	0.26651675
4	0.03900454	0.11073966	0.19296311	0.20512673
5	0.27227532	0.2958289	0.37611637	0.43250114
6	0.28558423	0.32610802	0.33738045	0.36903639
7	0.26794884	0.30046699	0.31130087	0.34364901
8	0.23377252	0.29064073	0.25234514	0.27463718
9	0.19773035	0.32357287	0.30284795	0.27152555
10	0.05153418	0.09356886	0.17818625	0.16728504
Promedio	0.19974554	0.26407504	0.27663424	0.2966694
Desviación Std.	0.11557670	0.0982649	0.08220083	0.09148445

Si bien, tanto *random forest* como gradient *boosting superan* el R^2 de los algoritmos individuales, es necesario validar que esa mejora sea estadísticamente significativa. Pasar identificar qué prueba utilizar, se realizaron pruebas de normalidad (tabla 4), donde se aprecia que todas las R^2 se distribuyen de forma normal.

Tabla 4. Pruebas de normalidad de variables

Modelo	X^2	Shapiro	Skewness	Kutosis
Multiple Linear Regression	Estadístico: 0.741 p-value: 0.6900	Estadístico: 0.893 p-value: 0.1839	Estadístico: -0.541 p-value: 0.5883	Estadístico: -0.670 p-value: 0.5027
Support Vector Regression	Estadístico: 0.485 p-value: 0.7845	Estadístico: 0.910 p-value: 0.2869	Estadístico: -0.680 p-value: 0.4959	Estadístico: 0.147 p-value: 0.8826
Random Forest	Estadístico: 0.690 p-value: 0.7372	Estadístico: 0.968 p-value: 0.8771	Estadístico: -0.236 p-value: 0.8131	Estadístico: -0.744 p-value: 0.4567
Gradient Boosting	Estadístico: 1.148 p-value: 0.5631	Estadístico: 0.929 p-value: 0.4461	Estadístico: 0.452 p-value: 0.650	Estadístico: -0.977 p-value: 0.3314

Finalmente, al verificarse que todas las predicciones tienen una distribución normal, se aplicó una prueba t-student para muestras emparejadas, con el fin de corroborar si existe una diferencia significativa al 95%. *Support vector regression* y ambos algoritmos ensamblados muestran una diferencia significativa con respecto a la regresión lineal (tabla 5).

Tabla 5. Prueba t-student para diferencias				
Modelo	**Multiple Linear Regression**	**Support Vector Regression**	**Random Forest**	**Gradient Boosting**
Multiple Linear Regression		Estadístico: 4.389 p-value: 0.0017	Estadístico: 4.901 p-value: 0.0074	Estadístico: 5.754 p-value: 0.0002
Support Vector Regression			Estadístico: 0.762 p-value: 0.465	Estadístico: 1.808 p-value: 0.1039
Random Forest				Estadístico: 2.212 p-value: 0.0541
Gradient Boosting				

Discusión

En este proyecto se experimentó con algoritmos ensamblados de regresión para mejorar la precisión del modelo para calcular el índice de necesidad de servicios de salud.

Si bien los algoritmos ensamblados arrojan mejores resultados, es importante mencionar que la explicabilidad y el poder de cómputo son una desventaja respecto a los algoritmos individuales. La elección del algoritmo dependerá de los objetivos de la investigación, pues si lo que se busca es un buen nivel de precisión, *machine learning*, y en especial los métodos ensamblados, demuestran ser una mejor opción.

Además, existe la posibilidad de explorar más algoritmos, pues si bien demuestran ser mejores que los individuales, siempre dependerá del conjunto de datos y sus características para una mejor o peor precisión.

Conclusiones

Con base en los resultados iniciales obtenidos, se puede concluir lo siguiente:

1. Los resultados obtenidos de los coeficientes de determinación de 0.321 y de correlación múltiple de 0.57, mediante el análisis de regresión lineal múltiple no arrojaron un nivel aceptable para la explicación del índice de necesidad a los servicios de salud, lo que dio origen a la implementación de técnicas de *machine learning* en la búsqueda de un mejor ajuste.
2. Se acepta la hipótesis de *machine learning* que establece que, para el problema de necesidad de servicios de salud con los microdatos del INEGI, los algoritmos ensamblados de machine learning mejoran estadísticamente la precisión de modelos individuales.
3. Para este contexto, el algoritmo ensamblado *gradient boosting* arrojó mejores resultados en cuanto al R^2.

4. Los resultados de los algoritmos ensamblados en validación cruzada tienen menos variabilidad.
5. Respecto a la hipótesis de investigación se identificó que, con base en el R^2, sí existe relación entre la necesidad de servicios de salud está relacionada con el índice de marginación urbana y el promedio de ocupantes por vivienda.

Referencias

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93-115.

Anselin, L. S. (2024, Enero 10). GeoDa. Retrieved from GeoDa: An introduction to spatial data analysis [Software]: https://geodacenter.github.io

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chasco, C. (2003). Métodos gráficos del análisis exploratorio de datos espaciales. Madrid, España, : Universidad Autónoma de Madrid.

CONAPO. (2021). Consejo Nacional de Población. Retrieved from Índices de Marginación 2020: https://www.gob.mx/conapo/documentos/indices-de-marginacion-2020-284372

Cortes, C. &. (1995). Support vector networks. Machine Learning, 20, 273-297.

Crespo, J. A. (2018). Elecciones y transición democrática en México (2000-2018). México: Editorial Porrúa.

De Corso, G. B. (2017). Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas. Cuadernos Latinoamericanos de Administración 13(25), 92-104.

DOF. (7 de Junio de 2024). LEY GENERAL DE SALUD. Obtenido de: https://www.diputados.gob.mx/LeyesBiblio/pdf/LGS.pdf

Electoral, I. N. (2024). Instituto Nacional Electoral. Retrieved from Sistema de Consulta de la Estadística de las Elecciones: https://siceen21.ine.mx/home

esri. (2022, Abril 9). ArcGis Pro. Retrieved from Environmental Systems Research Institute (ESRI). (2024). ArcGIS Pro [Software]: https://pro.arcgis.com

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-123.

Geurts, P. E. (2006). Extremely randomized trees. Machine Learning, 63(1), 3-42.

Hoc, H. e. (2023). Comparing Stacking Ensemble and Deep Learning for Software Project Effort Estimation. IEEE Access.

Hoc, H. e. (2023). Heterogeneous Ensemble Model to Optimize Software Effort Estimation Accuracy. IEEE Access.

INE. (2024). Elecciones Federales 2024. Retrieved from Base de Datos: https://computos2024.ine.mx/presidencia/base-de-datos

INE. (2024, Marzo 3). Instituto Nacional Electoral. Retrieved from Partidos que perdieron el registro: https://ine.mx/actores-politicos/partidos-politicos-nacionales/partidos-perdieron-registro/

INEGI. (1997). Manual de Medidas Sociodemográficas. Ciudad de México: INEGI.

INEGI. (2023, Noviembre 27). Instituto Nacional de Estadística y Geografía. Retrieved from Subsistema de Información Demográfica y Social: https://www.inegi.org.mx/programas/ccpv/2020/#microdatos

LEGISVER. (13 de Enero de 2025). H. CONGRESO DEL ESTADO DE VERACRUZ. Obtenido de: https://www.legisver.gob.mx/leyes/LeyesPDF/CONSTITUCI%C3%93N13012025.pdf

Lu, X. e. (2019). Ensemble learning regression for estimating unconfined compressive strength of cemented paste backfill. IEEE access, vol. 7, pp. 72125–72133.

Moran, P. (1948). The Interpretation of Statistical Maps. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 10, No. 2(, 243-251.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

Nti, K. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. Journal of Big Data, vol. 7, no. 1, 1–40.

Ordóñez, J. (2024, Agosto 3). Tribunal Electoral del Poder Judicial de la Federación. Retrieved from Representatividad de partidos. Vigilancia del padrón electoral y listas nominales [Documento PDF]: https://www.te.gob.mx/editorial_service/media/pdf/Representatividad_de_partidos.

Pedregosa, F. V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Woldenberg, J. (2012). Historia mínima de la transición democrática en México. México: El Colegio de México.

Wolpert, D. H. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, vol. 1, no. 1, 68–82.