PREDICCIÓN DE MORTALIDAD EN MÉXICO: UN ANÁLISIS COMPARATIVO DE ALGORITMOS DE *MACHINE LEARNING*

MORTALITY PREDICTION IN MEXICO: A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

Aldo Ruben Granillo Iturbe[1]
agranilloi001@alumno.uaemex.mx,
ORCID: 0009-0000-7048-3712

Doricela Gutiérrez Cruz[2]
dgutierrezcr@uaemex.mx
ORCID: 0000-0003-2843-3273

Resumen

La elaboración de proyecciones de mortalidad, resulta de una gran utilidad en el desarrollo de las naciones, ya que es gracias a ellas que se pueden tomar acciones para garantizar una mejor calidad y esperanza de vida. Con el fin de contribuir a este esfuerzo en el presente trabajo, se propone un análisis comparativo de algoritmos de *machine learning* para evaluar su rendimiento en tareas de predicción, evaluando un histórico de datos de defunciones en México desde 1950 hasta 2022, donde se pudo observar que la eficiencia de los algoritmos analizados puede verse afectada debido al orden de los datos analizados, asimismo, se proponen dos opciones para mejorar el desempeño: una basada en un mejor ajuste en los algoritmos y otra donde se pretende emplear algoritmos más potentes para llevar a cabo una predicción precisa y certera.

***Palabras clave*:** Machine learning, algoritmos, mortalidad.

[1] Ingeniería en Sistemas Inteligentes, Centro Universitario UAEM Nezahualcóyotl, Universidad Autónoma del Estado de México.
[2] Ingeniería en Sistemas Inteligentes, Centro Universitario UAEM Nezahualcóyotl, Universidad Autónoma del Estado de México.

Abstract

The preparation of mortality projections is very useful in the development of nations since it is thanks to them that actions can be taken to guarantee a better quality of life and life expectancy. In order to contribute to this effort, this paper proposes a comparative analysis of machine learning algorithms to evaluate their performance in prediction tasks, evaluating a history of death data in Mexico from 1950 to 2022, where it could be observed that the efficiency of the algorithms analyzed may be affected due to the order of the data analyzed. Likewise, two options are proposed to improve performance, one based on a better fit in the algorithms and another where it is intended to use more powerful algorithms to carry out an accurate and accurate prediction.

***Keywords*:** Machine learning, algorithms, mortality.

Fecha de envío: 06/06/2024
Fecha de aprobación: 10/11/2024
Fecha de publicación: 01/01/2025

Introducción

La aplicación de la Estadística en el estudio de la mortalidad, tiene una estrecha relación con la calidad de vida de las personas que habitan una región, permitiendo aplicar políticas que busquen mejorar órganos como el sistema público de salud, el acceso a sistemas de salud, el desarrollo económico del país hasta la calidad y el acceso a alimentos de la canasta básica entre otros. Es innegable la importancia de esta estadística para la implementación de diversas políticas, así como para realizar proyecciones demográficas, las cuales son de gran importancia para el desarrollo del país, por ejemplo, no tendría sentido desarrollar la actividad industrial en una entidad en la que la mayoría de sus habitantes no cuentan con una edad dentro del margen de productividad.

La mortalidad es un término usado en múltiples disciplinas, en Estadística se refiere a las muertes totales en un año en un cierto grupo poblacional esto sin importar las causas que la generan, esta métrica tiene como función la generación de información que permita identificar

patrones, tendencias y características (INEGI, 2023). Las proyecciones demográficas en México habitualmente se han realizado con métodos matemáticos como la ley de Gompertz-Makeham (Ogaz, 1991) en la cual se hacen predicciones basadas en probabilidades, algunas de estas presentan fallas en los extremos haciendo referencia en la niñez y ancianidad (Vignoli, 2012), donde, con el paso del tiempo, se esperaría que el grueso de defunciones esté en edades más avanzadas, pero esto no se ve reflejado en los datos donde el grueso de datos está en edades más tempranas, esta dinámica representa una gran dificultad en realizar las proyecciones debido a que no todas las entidades cuentan con las herramientas para hacer dichos cálculos, esto se ve mayormente reflejado en las regiones con mayor pobreza en las que no se cuenta con suficiente información y la que se encuentra resulta deficiente.

Por lo anterior su relevancia impacta a contribuir en el desarrollo de mejores técnicas para la predicción de mortalidad de la población, en el presente trabajo se generan conocimiento a fin de contribuir a través de técnicas de algoritmos de *machine learning* con el objetivo de hacer predicciones respecto a la mortalidad y de ese modo validar la eficiencia entre los algoritmos y métricas aplicadas en el presente documento.

Antecedentes

Galván y Meza (2014) aplicaron técnicas de minería de datos con los algoritmos de Soporte Vectorial Máquinas (SVM), K-Nearest Neighbors (KNN), Redes Neuronales Artificiales (ANN), Algoritmos Genéticos (AG) para identificar patrones y tendencias en la mortalidad por diferentes causas el estudio, donde se encontró que las principales causas de muerte en México son las enfermedades del corazón, la diabetes y los tumores, siendo esta con mayor incidencia en hombres.

Ortiz (2020) analizó las características sociodemográficas asociadas a tasas de mortalidad infantil utilizando datos reportados por el departamento administrativo nacional de estadística (DANE). Donde aplicaron un modelo de máquinas de soporte vectorial (SVM) mediante el cual se recolectaron datos sobre nacimientos y defunciones de menores donde las variables sociodemográficas más relevantes como fueron sexo, edad de la madre, nivel educativo, lugar de residencia donde, tras un análisis de diferentes modelos, se pudo concluir que las SVM fueron las que obtuvieron el mejor desempeño en la métrica de predicción de sucesos de nacido no vivo o vivo.

A la vez García *et al*.,(2011) realizaron un trabajo de Proyección estocástica de la mortalidad mexicana por medio del método de Lee-Carter, un modelo estocástico que permite estimar la evolución futura de las tasas de mortalidad por edad y sexo, donde obtuvieron proyecciones de la mortalidad. En los que se esperó una disminución de la mortalidad en todos los grupos de edad con ello prevén valores de confianza del 95%, donde destaca que el grupo de las mujeres para 2050 será más longevo que el actual. Un modelo de *machine learning* para predecir el riesgo de muerte por COVID-19, se basa en un conjunto de datos de pacientes de diferentes países, donde se obtuvo que el modelo fue capaz de identificar los factores de riesgo más importantes como la edad, sexo, destacando la importancia de considerar las causas de muerte y la estructura por edad de la población (Gil y Quintero, 2022).

Materiales y Métodos

Para el desarrollo de este artículo se utilizará la metodología Knowledge Discovery in Data bases (KDD) en el cual se lleva a cabo la extracción de manera iterativa de conocimiento partiendo de grandes volúmenes de datos donde se aplican técnicas de minería de datos con el fin de obtener nueva información para su análisis, y, de esta manera, generar nuevos conocimientos. Esta metodología desarrolla las siguientes fases:

Selección de Datos

En este estudio se llevará a cabo un análisis para comparar el rendimiento de los algoritmos K-Nearest Neighbors (KNN), RandomForest y regresión lineal para tareas de regresión. En este caso, en la predicción de la mortalidad en México se procesarán los datos proporcionados por el banco de datos abiertos de México (Proyecciones de la Población de México y de las Entidades Federativas, 2020-2070, 2023).

Preprocesamiento de los datos

En esta etapa nos centramos llevar a cabo una limpieza de los registros del *dataset,* el banco de datos contiene 737 661 registros de los cuales, solo se considerarán para este estudio 389 181

datos, siendo estos correspondientes a las defunciones totales en edades desde 0 hasta 109 años que es el rango de edad considerado por el INEGI, estos datos van desde el año 1950 al 2022 dentro de la república mexicana, ya que el 2023 será el año sometido al estudio. En estos registros encontramos las siguientes columnas: “renglón”, “año”,” entidad”, ”cv_geo”, “sexo”, “edad” y “defunciones”. Inicialmente los datos fueron preprocesados para eliminar las columnas que no son de interés para el estudio dejando únicamente “año”, “edad” y “defunciones” como se ilustra en la Tabla 1. La mayoría de los registros son de tipo numérico, por lo cual no fue necesario realizar un procesamiento adicional para poder hacer uso de estos. Para la limpieza del banco de datos se eliminaron datos nulos en caso de existir haciendo uso de funciones de pandas, además se emplearon *data frames* de pandas para dividir el banco de datos y almacenar los datos de interés en estos *data frames*.

Minería de datos

El análisis de los datos se lleva a cabo mediante la implementación de tres algoritmos de *machine learning*, los cuales fueron aplicados mediante la librería Sklearn, los algoritmos son los siguientes:

Regresión lineal, se basa en relaciones lineales entre variables dependientes que van de un objetivo a una o varias variables de destino, para este estudio se emplea una regresión lineal múltiple por el método de mínimos cuadrados ordinarios, este método es útil debido a que calcula los coeficientes de las rectas generadas para obtener una recta en base a la aproximación promedio de todos los puntos.

Random forest, es un algoritmo de regresión basado en árboles, trabaja con múltiples árboles de decisión para generar predicciones las cuales ajusta para obtener una sola predicción más precisa, en este estudio se generaron 100 árboles de decisión para este regresor considerando que ese número se considera optimo evitando así posibles sobreajustes además se consideró las limitaciones computacionales las cuales pueden verse reducidas al ampliar el número de árboles de decisión.

K-Nearest Neighbors* (*Knn), es empleado en tareas de clasificación como de regresión en esta última las predicciones son llevadas a cabo mediante el promedio de las K distancias más cercanas al registro, para este estudio se optó por usar la configuración por default del algoritmo quedando así para el valor de "K" como los 5 vecinos más cercanos debido a que ofrece un buen equilibro con respecto al sobreajuste y sub ajuste

Análisis de los datos

Para la evaluación del rendimiento de estos algoritmos, se emplearon métricas de precisión para el cálculo del error, las cuales son error medio cuadrático, error absoluto medio y el coeficiente de determinación.

Las predicciones obtenidas serán comparadas entre ellas y serán contrastadas contra los datos de defunciones totales durante el año 2023, estas fueron escogidas ya que permiten conocer de manera cuantificable la diferencia entre los valores reales y los valores predichos, estas, a su vez, son útiles no solo para evaluar los resultados, sino que también nos permite ajustar los hiperparámetros de los algoritmos para obtener resultados más óptimos en las predicciones.

El criterio para la aplicación de estos algoritmos consta en la creación de estructuras que contengan los datos del *dataset,* en este caso se usaron los *data frame* que son una estructura propia de pandas en *python*, en el *data frame* se carga el *dataset*, el cual será dividido en grupos que representan las variables independientes y dependientes, esto es necesario para entrenar a los modelos de regresión, ya que necesitan conocer cuál es la variable objetiva, en este caso "Defunciones" representa la variable dependiente, posteriormente se crearán los conjuntos de entrenamiento y prueba con una división de 70% para entrenamiento y 30% para prueba, dando así, una división simple para nuestros datos. Estos conjuntos de datos son pasados a los algoritmos para ser entrenados. Posteriormente se aplicaron las métricas de evaluación de error [Figura 1].

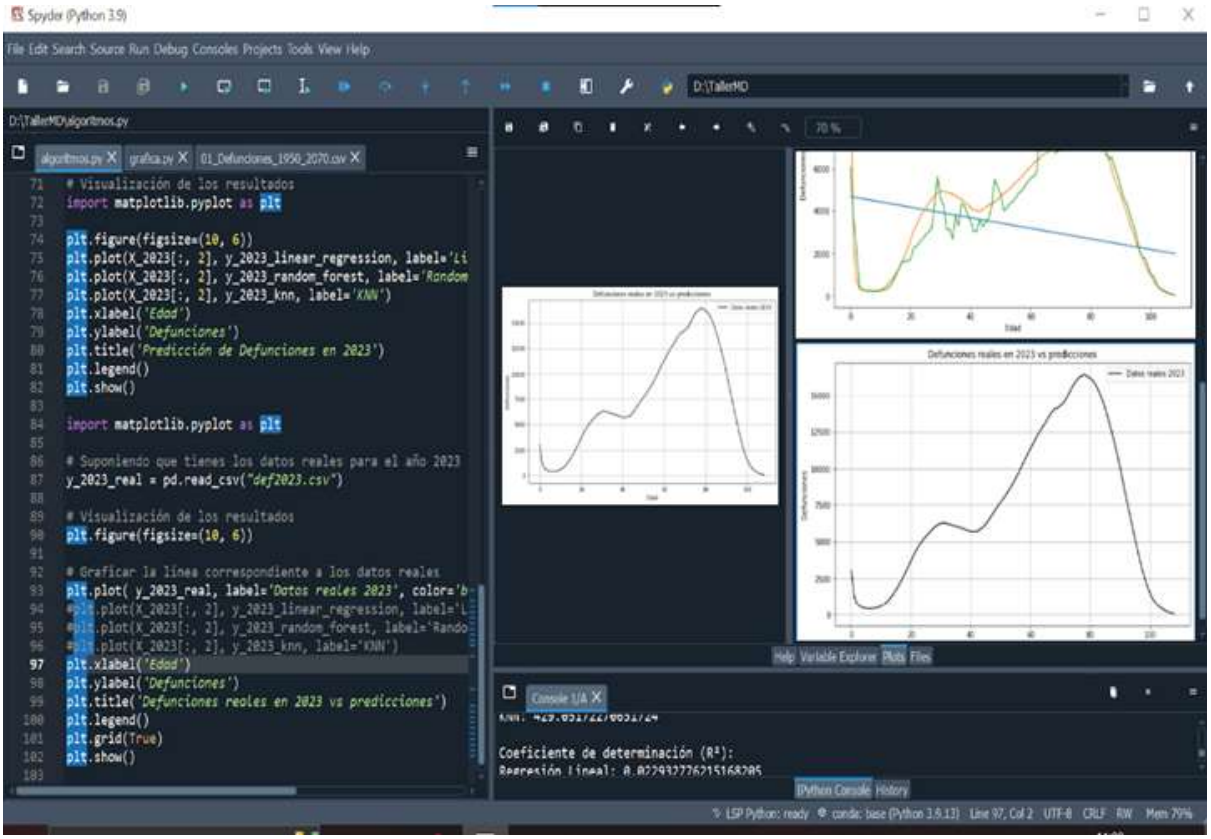

Fig. 1. visualización del programa, (elaboración propia 2024).

Estas métricas tienen diversas formas de evaluar los resultados de las predicciones de los algoritmos. El error medio cuadrático mide la diferencia promedio al cuadrado entre los valores reales y los predichos por el modelo, por otra parte, el error absoluto medio solo contempla la diferencia promedio de los valores reales entre los valores predichos a su vez el coeficiente de determinación proporción de variabilidad de la variable dependiente en base a la independiente, estas métricas fueron calculadas mediante su implementación en código. Posteriormente se llevó a cabo la predicción de defunciones totales para el año 2023 llevando a cabo una comparación entre los datos reales de defunciones totales de dicho año contra las predicciones hechas por cada uno de los modelos [Figura 2].

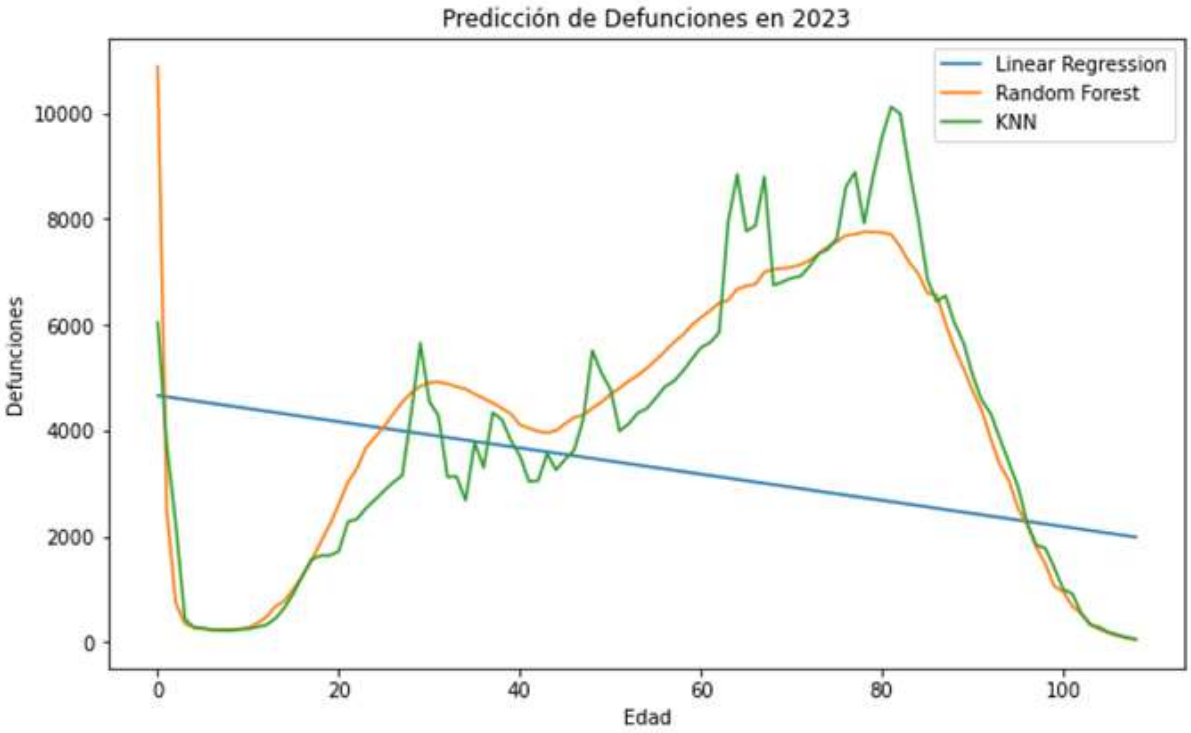

Fig. 2. Predicciones de los algoritmos, (elaboración propia 2024).

Resultados

En este estudio se llevó a cabo una comparativa para evaluar el rendimiento de tres algoritmos, generando así una predicción de mortalidad en México para el año 2023. En este estudio se contemplaron únicamente las muertes totales. Dichos algoritmos, fueron sometidos a una evaluación de su rendimiento mediante métricas de error (Tabla 2), estas son un indicativo de la calidad de predicciones que estos algoritmos realizan.

Tabla. 1. Tabla de errores de los algoritmos, (elaboración propia 2024).

Algoritmo	Error medio cuadrático	Error absoluto medio	Coeficiente de determinación
Regresión lineal	32,684,262.36	2036.83	0.02
Random Forest	61,943.92	53.97	0.99
Knn	5,150,698.03	429.05	0.84

Generalmente suelen considerarse rangos de 0 a 1 para considerar la exactitud de la predicción de un algoritmo, aunque eso no implica directamente que si se obtienen valores fuera de este rango no sean de utilidad, esto depende del trabajo que se esté realizando, es posible observar que para el error medio cuadrático y el error absoluto medio, se obtuvieron valores muy elevados, indicando que los valores predichos por los algoritmos presentan desviaciones considerables respecto a sus valores reales. A diferencia del coeficiente de determinación, se sugiere un mejor ajuste para los algoritmos, exceptuando a la regresión lineal que, hasta ahora, ha sido el que ha obtenido valores deficientes, mientras que, a su vez, el algoritmo *random forest* es el que mejor desempeño ha obtenido. Durante el año pasado 2023 en México fallecieron un total de 805 313 personas. Esto acorde a los registros del *dataset* en el cual se almacenaron las muertes totales del banco de datos original. Dichas defunciones fueron graficadas haciendo uso de *matplotlib*, esto se ve reflejado en la Figura 3.

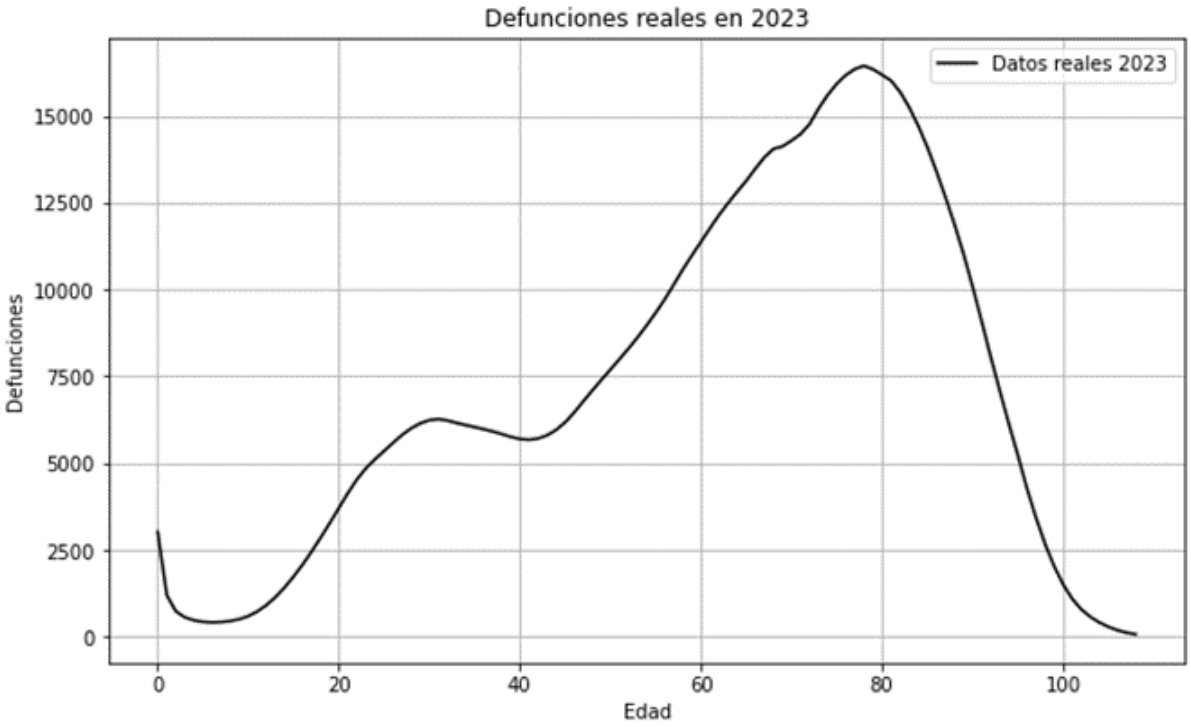

Fig. 3. Gráfica de defunciones totales en México en el año 2023, (elaboración propia 2024).

A continuación, se muestra en la Figura 4 las predicciones hechas por los algoritmos ya mencionados anteriormente.

Resulta interesante la visualización de las predicciones, nótese que la más deficiente, fue el algoritmo de regresión lineal, ya que no es capaz de representar los picos que aprecian respecto a los datos originales, esto puede deberse a la lógica propia del algoritmo, la cual se centra en encontrar una ecuación lineal que describa correctamente la correlación entre las variables.

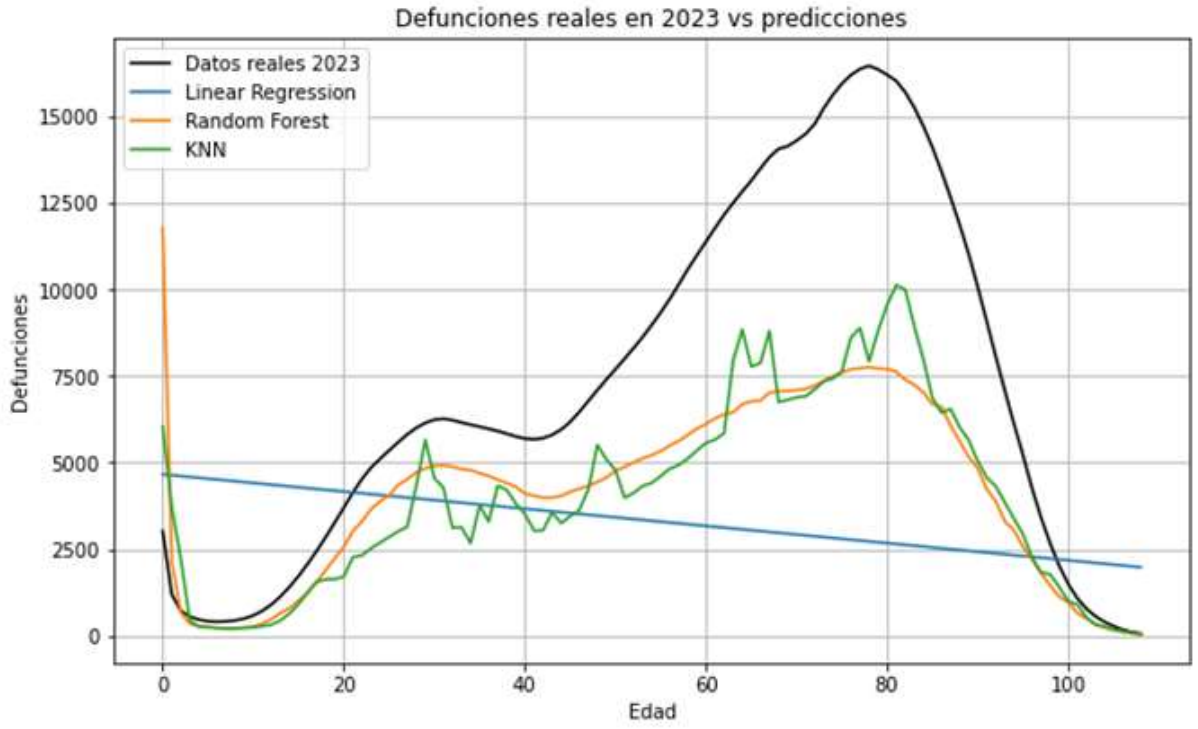

Fig. 4. Predicciones de los algoritmos respecto a los datos reales, (elaboración propia 2024).

Asimismo, los algoritmos *knn* y *random forest,* en específico este último, fueron los que obtuvieron mejores resultados a comparación de la regresión lineal, no obstante, aún los resultados distan bastante de la información real. El algoritmo de regresión lineal, predice un total de 333 414, siendo una diferencia de 471 899 muertes una diferencia del 58% con los datos reales. Por otra parte, los algoritmos de *random forest* y *knn* respectivamente predijeron 404 156 y 429 568 decesos con una diferencia respecto a los datos reales de 50% y 53% respectivamente, lo cual no es muy óptimo en donde se busca tener un índice de certeza considerable como podría ser un 90% de los datos. Si bien las predicciones de mortalidad son una expresión que surge al tratar de estimar el comportamiento de los decesos respecto a factores como la edad, género, entre otros, resulta imprescindible la implementación de técnicas que satisfagan de la manera más óptima dicha necesidad. Asimismo, las técnicas de *machine learning* representan aún una gran área de aprovechamiento, pero resulta imperante tomar en cuenta el diseño de los datos, el cómo son estructurados, y el algoritmo a implementar para que así las posibilidades de presentar una mala implementación de los algoritmos predictivos sea menor Dromundo (2018).

Cierre

En la presente investigación se compararon tres algoritmos de *machine learning* para realizar predicciones sobre la mortalidad en México para el pasado año 2023, los cuales se apoyaron del banco de datos abiertos con el *dataset* de defunciones en México.

Se obtuvieron valores de error mediante el uso de métricas para determinar el error de dichas predicciones, esto en base a los datos reales. Los resultados obtenidos son una evidencia de que, a pesar de que los algoritmos de *machine learning* son una herramienta muy potente y útil, en múltiples casos aún no pueden ser equiparables a las técnicas que comúnmente son empleadas o algoritmos específicamente creados para realizar estos cálculos como la tasa bruta de mortalidad Ogaz (1991) haciendo referencia a este estudio en particular.

Si bien se contaba con una amplia gama de datos, siendo estos recopilados desde 1950 hasta el 2022, resulta difícil la tarea predictiva, ya que las defunciones no son exclusivamente de un orden natural, debido a que no todas las defunciones registradas son provocadas por la misma causa; algunas son orden natural como la vejez, enfermedades, otras son por agentes externos como los fenómenos naturales. Es por esto que la tarea predictiva aumenta su complejidad, un ejemplo de esto es el SARS-COV 2 que fue un fenómeno externo que causó múltiples decesos (García y Pérez, 2021; Gil y Quintero, 2022). Existen un sinfín de elementos los cuales hay que considerar a evaluar, por ello, un trabajo a futuro podría considerarse la aplicación de algoritmos más potentes como las redes neuronales para realizar este tipo de predicciones. Esto debido a su capacidad de aprender patrones para simular con un comportamiento más natural.

Asimismo, evaluar en periodos más cortos de tiempo de un año, haciendo ajustes en los hiperparámetros de los algoritmos empleados, ya que, como se mencionó, múltiples factores pueden afectar el cálculo de la predicción a causa de múltiples factores que pueden intervenir en plazos tan largos de tiempo. A su vez, podría considerarse hacer un análisis más particular sobre algunas afecciones o periodos de tiempo específicos que sean propensos a generar interés para analizar donde se proponga llevar a cabo una segmentación más certera de los datos, donde la sección de los registros se ajuste a los periodos que serán sometidos a estudio, y, en base a ello, construir un mejor modelo para la tarea predictiva.

Referencias

Antequera, M. G., Sendra, A. L., & Lama, J. R. (2020). Algoritmos de machine learning y su aplicación al mantenimiento industrial en el sector agroalimentario. *Dialnet*, 165-169. https://dialnet.unirioja.es/servlet/articulo?codigo=7655416

Bermúdez, C & Bermúdez, S. (2019). Sistema de predicción de estadísticas de nacimientos y defunciones en Colombia soportadas por TIC. Universidad Cooperativa de Colombia, Facultad de Ingenierías, Maestría en Gestión de Tecnologías de la Información, Bucaramanga. https://hdl.handle.net/20.500.12494/10746

Dávila, C. A. (2012). Ajuste matemático de la mortalidad general en México 2000, 2005 y 2010. *Papeles de población*,*18* (74), 01-34. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-74252012000400006&lng=es&tlng=es.

Dromundo, A. X. U. (2024). *Machine learning y su importancia en la actualidad*. IPADE. https://www.ipade.mx/newsmedia/tendencias/machine-learning-y-su-importancia-en-la-actualidad/

Galván. P. & Meza, A. (2014). *Estudio de minería de datos para la información de mortalidad en México*. http://132.248.52.100:8080/xmlui/handle/132.248.52.100/2789

García, E y Pérez, H. (2021). La mortalidad por COVID-19 en México. https://dsp.facmed.unam.mx

García, V. M. & Ordorica, M. (2012). Proyección estocástica de la mortalidad mexicana por medio del método de Lee-Carter. *Estudios demográficos y urbanos*, *27*(2), 409-448. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0186-72102012000200409&lng=es&tlng=es.

Gil, V. & Quintero, C. (2022). Predicción del riesgo de muerte por COVID-19 con Machine Learning. https://www.researchgate.net/profile/Victor-Gil-Vera/publication/362781339_Prediccion_del_riesgo_de_muerte_por_COVID-19_con_Machine_Learning/links/62feb497ceb9764f7206d76c/Prediccion-del-riesgo-de-muerte-por-COVID-19-con-Machine-Learning.pdf

Instituto Nacional de Estadística y Geografía (INEGI). (2023). Mortalidad. https://www.inegi.org.mx/programas/mortalidad/?fbclid=IwAR3INMoyRI7YIPhdqHO5TR98oG6qmyncoDS_dlWvTv9Fwe1gGn1gBFtR5x4

Mina, A. (2006). Ley de mortalidad mexicana. Funciones de supervivencia. *Estudios Demográficos y Urbanos*, *21*(2), 431–456. https://doi.org/10.24201/edu.v21i2.1255

Ogaz, H. (1991). La función de Gompertz-Makeham en la descripción y proyección de fenómenos demográficos. E*studios Demográficos y Urbanos*, *6*(3), 485–520. https://doi.org/10.24201/edu.v6i3.820

Ortiz, V. (2020). *Aplicación de técnicas de aprendizaje automático para la segmentación y clasificación de características sociodemográficas asociadas a tasas de mortalidad infantil utilizando datos reportados por el DANE Colombia entre los años 2008 al 2017*. http://hdl.handle.net/20.500.12010/19171.

Vignoli, J. R. (2011). Reproducción adolescente y desigualdades. *Revista Latinoamericana de Población*, *5*(8), 87-113. https://doi.org/10.31406/relap2011.v5.i1.n8.6