ISSN 1870-4069

Reconstrucción de los parámetros de la calidad del agua en el río Fuerte, Sinaloa, implementando técnicas de aprendizaje máquina

José Luis Medina-Jiménez, Héctor Rodríguez-Rangel,
Leonel Ernesto Amábilis-Sosa, Kimberly Mendivil-García,
Juan Carlos Gonzalez-Nava, Daniel Antonio Nieblas-Fuentes

Tecnólogico Nacional de México Campus Culiacán,
División de Posgrado,
México

{jose.mj, hector.rr, leonel.as}@culiacan.tecnm.mx,
{kimberly.mendivil, 19171353, 19170633}@itculiacan.edu.mx

Resumen. La comisión nacional del agua (CONAGUA), a través de los años, mediante una red de monitoreo (RENAMECA), ha generado conjuntos de datos de parámetros que miden la calidad del agua, donde el objetivo es realizar tomas de muestras en distintos puntos estratégicos de los cuerpos de agua. Por diversos factores, como falla de captura, la dificultad de entrar a la zona de muestreo, falta de personal capacitado, tiempos de análisis de laboratorios, el registro de estos parámetros de calidad de agua no se realizan. De tal manera que, el conjunto de datos generado por la RENAMECA cuenta con una de las problemáticas más comunes dentro de la ciencia de datos, los valores perdidos dentro de cada uno de los parámetros en distintos niveles de porcentaje de perdida. Históricamente, el avance de la estadística y de la inteligencia artificial, han generado distintas técnicas capaces de recuperar estos valores faltantes, llamándolas métodos de imputación. En el artículo se describe la implementación de tres distintos métodos de imputación en un conjunto de datos de CONAGUA del río Fuerte en Sinaloa de 38 parámetros, donde la reconstrucción se realiza en siete parámetros. La validación de la reconstrucción se hizo mediante la simulación de distintos porcentajes de valores perdidos de 10, 20 y 30 %, obteniendo distintos resultados en valores de MSE y RMSE.

Palabras clave: Imputación, reconstrucción de datos, vecinos más cercanos K, regresión lineal.

Reconstruction of Water Quality Parameters in the Fuerte River, Sinaloa, Using Machine Learning Techniques

Abstract. The National Water Commission (CONAGUA), over the years, has generated datasets of parameters that measure water quality through a monitoring network (RENAMECA), where the objective is to take samples at strategic points in bodies of water. Due to various factors, such as capture failure,

difficulty accessing the sampling area, lack of trained personnel, and laboratory analysis times, the recording of these water quality parameters is not always possible. As a result, the dataset generated by RENAMECA has one of the most common problems in data science, missing values within each parameter at different levels of percentage loss. Historically, the advancement of statistics and artificial intelligence has generated different techniques capable of recovering these missing values, calling them imputation methods. This article describes the implementation of three different imputation methods on a CONAGUA dataset from the Fuerte River in Sinaloa with 38 parameters, where reconstruction is carried out on seven parameters. The validation of the reconstruction was done by simulating different percentages of missing values of 10, 20, and 30%, obtaining different results in MSE and RMSE values.

Keywords: Imputation, data reconstruction, K nearest neighbors, linear regression.

1. Introducción

En el campo de la ciencia y análisis de datos, la calidad de los conjuntos de datos es un tema central de gran importancia. Actualmente, diversos campos como las ciencias de la salud, finanzas, medio ambiente, etc., generan conjuntos de datos para mantener registro de los comportamientos de objetos de estudio en sus áreas y así realizar análisis estadísticos, inferencias y modelados [1, 2].

Los conjuntos de datos son colecciones de información almacenadas y organizadas en filas y columnas. Los datos pueden ser de cualquier tipo, como números, texto, imágenes, audio, etc. Los conjunto de datos son generados de diversas maneras, como la recopilación manual de datos mediante encuestas, entrevistas, observaciones, cuestionarios.

Por otro lado, se puede implementar la extracción de datos en las bases de datos mediante consultas,y por último existen los sensores, que son capaces de medir y recopilar información en tiempo real sobre el entorno, como temperatura, presión, humedad, entre otros. En el área de la inteligencia artificial la utilización de conjuntos de datos es esencial, ya que gracias a estos son capaces de generar modelos que puedan estimar o clasificar los comportamientos que se desea analizar.

Es así que, la cantidad y calidad de datos con las que se cuenta en un conjunto de datos son cruciales por el hecho de que estos determinan la precisión y la capacidad de generalizar modelos de aprendizaje [3]. Una de las ventajas de contar con conjuntos de datos diverso, es la ayuda para generar modelos eficaces donde se evitan problemáticas comunes como el sobreajuste y subajuste.

En México existe un organismo denominado CONAGUA el cual se encarga, de medir la calidad en los diversos cuerpos de agua con los que cuenta. En él se implementa una red nacional de monitoreo (RENAMECA), en el cual se colocaron puntos de muestreo estratégicamente para la medición de distintos parámetros de la calidad del agua [4].

Tabla 1. Parámetros de la calidad del agua y sus porcentajes de valores perdidos.

Parámetro	Porcentaje de datos perdidos
N_NO3	0.3205
N_TOT	0.3205
TOX_V_15_UT	0.3205
OD_mg/L	0.3205
TURBIEDAD	0.3205
AS_TOT	0.3205
CD_TOT	0.3205
CR_TOT	0.3205
COT	0.6410
PB_TOT	0.6410
TEMP_AMB	0.6410
TEMP_AGUA	0.6410
HG_TOT	0.9615
NI_TOT	0.9615
SDT	1.9231
CONDUC_CAMPO	1.9231
ABS_UV	2.5641
COT_SOL	3.2051
DUR_TOT	3.2051
E_COLI	3.5256
DBO_TOT	3.8462
DQO_TOT	4.1667
TOX_D_48_UT	4.1667
COLI_TOT	5.7692
DBO_SOL	6.4103
DQO_SOL	6.7308
CN_TOT	16.0256
CAUDAL	22.4359

El conjunto de datos generado por la RENAMECA de los parámetros de la calidad del agua en todo México, por diversos factores como error de captura de datos, traslado al punto de muestreo, tiempo de obtención de resultados, presenta una de las problemáticas más usuales dentro de la ciencia de datos, el cual es la falta de información o la perdida de información en diversos parámetros y en distintos niveles de porcentaje de faltantes [1].

De tal manera que la implementación de modelos estadísticos y de aprendizaje máquina en conjuntos de datos con valores faltantes no son robustos para el manejo de esta problemática, generando así modelos de estimación pocos eficientes [5].

En el Fuerte, Sinaloa, la minería es una de las actividades más intensas en desarrollo, la cual es causa principal de los cambios de concentración de metales pesados en cuerpos de agua.

Los parámetros del conjunto de datos de la calidad del agua en el río Fuerte obtenidos por la RENAMECA presentan datos perdidos, donde es indispensable la imputación de todos esos valores por los pocos registros para la implementación de estimación de los metales pesados.

Actualmente, existen técnicas para el manejo de datos perdidos, donde se hace uso de la eliminación directa de los registros, hasta la imputación de los valores, con la finalidad de obtener más datos de entrada para el modelo a generar de estimación [6].

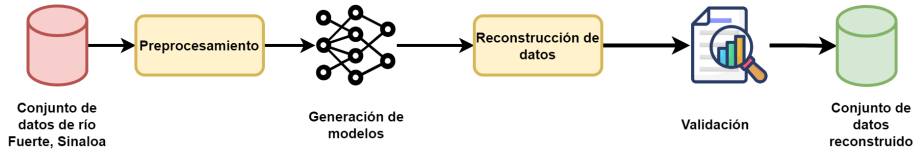

Fig. 1. Diagrama de imputación de conjunto de datos.

Entre las técnicas de imputación, existen las estadísticas como la media, la moda y la regresión lineal, las cuales fueron las primeras en implementarse. Por otra parte, existe la técnica capaz de imputar de manera múltiple los valores faltantes de diversos parámetros dentro del conjunto de datos denominado, imputación multivariada por ecuaciones encadenadas (MICE, por sus siglas en inglés).

Actualmente, la implementación de vecinos más cercanos K(KNN, por sus siglas en inglés), redes neuronales, árboles de decisiones y el aprendizaje profundo son los algoritmos más ampliamente utilizados [7].

Ante estas problemáticas, este artículo propone la implementación de tres modelos basados en vecinos más cercanos K y una red neuronal combinada con redes convolucionales para la reconstrucción individual de cada parámetro y en comparativa con la imputación múltiple mediante imputación multivariada por ecuaciones encadenadas (MICE).

Realizando la imputación de tres modelos aplicados en siete parámetros distintos de la calidad del agua, donde se simuló el borrado artificial de 10, 20 y 30 % para obtener mediante las métricas de MSE y RMSE el desempeño de estos modelos al comparar el conjunto de datos completo con el reconstruido. El siguiente artículo está organizado de la siguiente manera: Estado del arte, Materiales y Métodos, Resultados, Conclusiones y Trabajos Futuros.

2. Estado del arte

La imputación de los datos faltantes llevan casi 100 años de estudio, donde los principales métodos de reemplazo de valores perdidos fue la implementación de la media y moda, donde Milks (1932) se le atribuye la idea de la implementación de ellos.

Conforme fueron avanzando los años, surgieron diversas técnicas de imputación como imputación múltiple o métodos de regresión lineal y múltiple ([8]).

Actualmente, en las últimas décadas, el aprendizaje automático ha tomado gran relevancia en la imputación de datos faltantes, donde la implementación de redes neuronales y métodos de regresión basados en bosques aleatorios son ampliamente utilizados, permitiendo analizar conjuntos de datos más grandes y de mayor complejidad con mayor precisión y eficiencia.

Menéndez et al. [9] proponen la implementación de imputación controlada, es decir, imputan valores conocidos de un conjunto de datos de manera aleatoria para posteriormente compararlos con los valores reales conocidos. El conjunto de datos implementado son sobre contaminantes en una región de Polonia, implementando técnicas como imputación lineal, bosques aleatorios y vecinos más cercanos K.

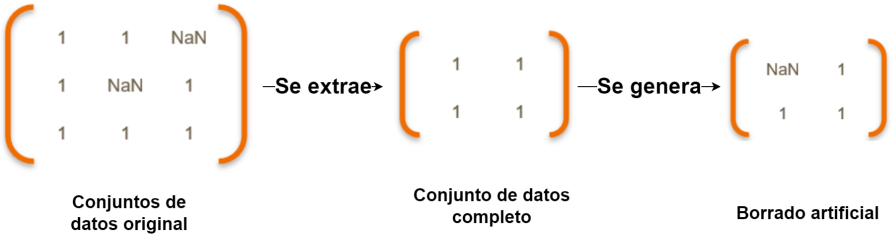

Fig. 2. Diagrama de borrado artificial.

Existen otros trabajos como el de Lin et al. [10], donde se hace la implementación de redes neuronales profundas, como el perceptrón multicapa y las redes de creencia profunda. En dicho trabajo, realizan la comparación con distintas técnicas como la media, KNN, árbol de clasificación y regresión (CART, por sus siglas en inglés) y máquina de soporte vectoriales (SVM, por sus siglas en inglés), llegando a la conclusión de que las redes de creencia profunda tienen mejor desempeño.

Camino et al. [11] proponen una solución de imputación de datos perdidos basada en modelos generativos profundos, donde se implementan autocodificadores de variación con distintas variantes implementados en conjuntos de datos con distintas características, donde la metodología para obtener los resultados fue la eliminación de valores al azar y reconstruirlos con varias técnicas.

Otras técnicas, aparte de las ya mencionadas para la imputación de datos faltantes, es el modelo bayesiano, el cual se propone en el trabajo de Zhai & Gutman [12], el cual se basa en los componentes de descomposición del valor singular de una matriz.

Por otra parte, en el desarrollo de Ratolojanahary et al. [13] realizan la combinación de técnicas de imputación multivariadas por ecuaciones encadenadas con métodos de aprendizaje automático para tratar la relación entre 200 parámetros de la calidad del agua. Los modelos con los que se combinaron fue, bosques aleatorios, árboles de regresión, vecinos más cercanos () y regresión vectorial de soporte.

La combinación de técnicas en distintos enfoques ha dado como resultado buenos algoritmos, en el caso de la imputación de datos faltantes, en el trabajo de Silva & Cabrera [14] se desarrolló un sistema de iferencia neuro-difusa co-activo llamado CANFIS-ART en el cual se automatiza el procedimiento de imputación de datos.

El modelo se construye a partir de las capacidades adaptativas de la red neuronal multicapa y el enfoque cualitativo de la lógica difusa utilizando el algoritmo fuzzy-ART. En el trabajo implementan la reconstrucción en 18 bases de datos, llegando al resultado que el enfoque propuesto por los autores superan totalmente los métodos de vanguardia y adicionalmente demuestran un mayor nivel de capacidad de generalización.

3. Metodología

Para la implementación de reconstrucción de datos faltantes, se hizo uso de un conjunto de datos proporcionado por CONAGUA a partir del río Fuerte de Sinaloa, el cual tienen 38 parámetros de la calidad del agua, en donde 28 de ellos cuentan con valores perdidos, los cuales se muestran en la Tabla 1.

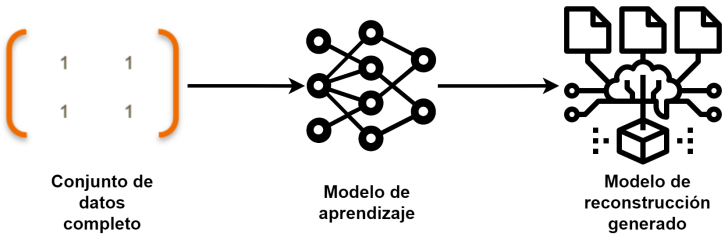

Fig. 3. Diagrama de generación de modelos.

3.1. Imputación de datos perdidos

La imputación de datos perdidos es una tarea que requiere distintos procesos para obtener una buena reconstrucción de los valores. La metodología planteada para el desarrollo de dicha tarea, es el generar un modelo por cada parámetro con datos faltantes y posteriormente realizar la imputación múltiple, con la finalidad de realizar una comparativa general, de cuál método resulta más eficiente.

El proceso de imputación de los valores de la calidad del agua está basado en el diagrama mostrado en la Figura 1:

El proceso de imputación será descrito por cada etapa para conocer a detalle que sucede en cada una de ellas.

Etapa de preprocesamiento

Para generar un modelo de aprendizaje es necesario ingresar los datos de tal manera que el método a implementar sea capaz de entender los datos ingresados.

Es decir, proveer los datos que sean útiles para que el modelo aprenda de ellos. Antes de ingresar los datos para generar los modelos, el conjunto de datos fue limpiada y normalizada por columna, para que tengan la mejor calidad posible de entrada,proceso de filtrado del conjunto de datos se muestra en la Figura 2.

La etapa de preprocesamiento está dividido en distintos procesos, donde se parte del conjunto de datos original del río Fuerte de Sinaloa y posteriormente generar un conjunto de datos completos y posteriormente realizar el borrado artificial con distintos porcentajes para su reconstrucción.

- **Conjunto de datos original:**
  El conjunto de datos original, son los registros del río Fuerte, en Sinaloa, de 38 parámetros en total, con alrededor de 357 registros de ellos. Cada parámetro tiene un porcentaje distinto de datos faltantes, es decir, valores que no fueron registrados en el conjunto de datos.
- **Generación de conjunto de datos completo:**
  La generación de conjunto de datos completo, consiste en que partiendo del conjunto de datos original, generar un conjunto de datos donde se realice la eliminación de aquellas columnas o también conocido como parámetros, con porcentajes altos de datos faltantes y posteriormente eliminar aquellas filas que tengan faltante de registros en alguno de los parámetros.

Tabla 2. Arquitectura de red convolucional con red neuronal artificial.

Capas	Kernel	Filtro	Dimensión de salida
Conv1	2x2	32	37x1
Flatten	-	-	37
Dense1	-	-	32
Dense2	-	-	1

Por otra parte, otras de las condiciones implementadas para la eliminación de columnas, parte de eliminar aquellos parámetros que tengan un comportamiento constante, debido a que no aportan nada al generar modelos donde se requiere reconstruir o estimar.

De tal manera que, se obtuvo un conjunto de datos sin valores faltantes, con el cual se redujo el de 357 registros a 273 y de 38 parámetros se redujo a 32 en total. A partir de este conjunto de datos completo, se implementó la normalización de datos por cada columna y tener estandarizado los formatos de los datos entrada para mayor eficiencia de procesamiento y mejores resultados.

– **Borrado artificial:**
Para realizar validaciones de la reconstrucción de datos faltantes, es necesario generar, a partir del conjunto de datos completo, un segundo conjunto de datos con la finalidad de eliminar o generar de manera aleatoria distintos porcentajes de pérdidas en los parámetros que se desea evaluar su reconstrucción.

De esta manera, se cuenta con dos conjuntos de datos, una con los datos completos y la segunda con el conjunto de datos con borrado artificial de manera aleatoria, para posteriormente realizar la reconstrucción en ella y compararla con el primer conjunto de datos y así obtener la métrica de evaluación.

Generación de modelos

Durante esta etapa se determinó el conjunto de parámetros a reconstruir, implementando aquellos que tuvieran un porcentaje de como cerca del 10 % y posteriormente implementar dos modelos de reconstrucción de manera individual por cada parámetro y realizar la reconstrucción múltiple de estos. La Figura 3 muestra el proceso que se siguió desde la entrada de los datos completos hasta la generación de un modelo de imputación o reconstrucción.

– **Modelo de aprendizaje:**
Se hizo uso de distintos modelos de aprendizaje para realizar la reconstrucción de datos, con la finalidad de tener una comparativa entre los diversos métodos que existen para generar la imputación de valores faltantes.

Los modelos implementados para la imputación fue el uso de vecinos más cercarnos K (KNN por sus siglas en inglés), debido a que es uno de los métodos más ampliamente utilizado para problemas de regresión, donde el valor de K solamente se exploró el 3.

Por otra parte, con la finalidad de explorar técnicas de aprendizaje profundo, se implementó las redes convolucionales combinada con red neuronal artificial,

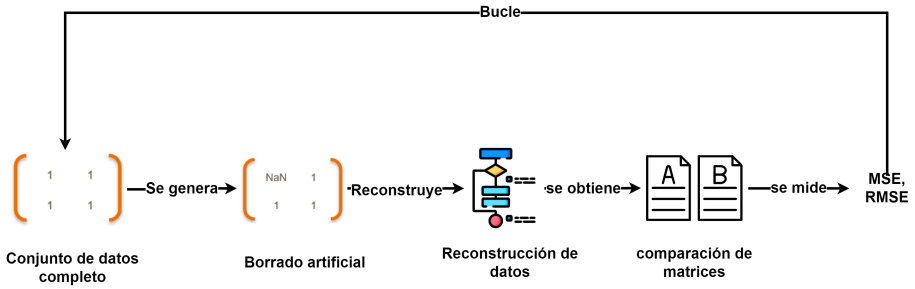

Fig. 4. Diagrama de bucle de obtención de resultados.

la configuración esta dada por la Tabla 2, se hizo uso de una de las técnicas de imputación múltiple denominada imputación múltiple por ecuación en cadena (MICE por sus siglas en inglés), esto con la finalidad de comparar la imputación de todos los parámetros a la vez y la reconstrucción individual por parámetro.

– **Modelo de reconstrucción generado:**
Los modelos de reconstrucción de cada método de aprendizaje generados por cada parámetro se implementaron en el conjunto de datos con borrado artificial, con la finalidad de realizar la comparación de ambas matrices, lo cual lleva a la etapa de obtención de resultados.

Validación:

La etapa de validación, se enfoca en realizar repetitivamente la generación de un conjunto de datos con borrado artificial en distintos porcentajes y como paso siguiente realizar la reconstrucción con los distintos modelos generados.

De tal manera que, para las pruebas, se implementó el borrado artificial en los parámetros en porcentajes de 10 %, 20 % y 30 % para posteriormente realizar la reconstrucción en el conjunto de datos con borrado artificial.

Una vez realizada la reconstrucción, se realiza una comparación entre la matriz de datos completa y la reconstruida para así calcular las métricas del error cuadrático medio (MSE por sus siglas en inglés) y raíz cuadrada del error cuadrático medio (RMSE por sus siglas en inglés).

Las métricas de MSE y RMSE son implementadas ampliamente para la evaluación de modelos de regresión, donde ambas describen el rendimiento de los modelos generados y a su vez realizar comparación diferentes modelos de distintas técnicas implementadas.

El MSE está definido por la ecuación 1, la cual describe la diferencia promedio al cuadrado que existe entre los valores estimados del parámetro reconstruido de la calidad del agua y el valor real, en donde, entre más cercano sea el valor a 0, indica una menor diferencia entre ambos valores:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \langle y_i + \hat{y} \rangle^2 . \tag{1}$$

Tabla 3. Resultados de imputación con valores faltantes de 10 %.

Parámetro	Modelo	MSE	RMSE
COLI_TOT	KNN	0.0536	0.2314
	MICE	0.0524	0.2288
	ANN + CNN	0.0608	0.2465
E_COLI	KNN	0.0109	0.1042
	MICE	0.0102	0.1008
	ANN + CNN	0.0072	0.0851
DBO_SOL	KNN	0.0144	0.1200
	MICE	0.0113	0.1061
	ANN + CNN	0.0127	0.1129
DBO_TOT	KNN	0.0207	0.1439
	MICE	0.0185	0.1359
	ANN + CNN	0.0231	0.1520
DQO_SOL	KNN	0.0013	0.0361
	MICE	0.0012	0.0340
	ANN + CNN	0.0023	0.0047
DQO_TOT	KNN	0.0081	0.0898
	MICE	0.0061	0.0783
	ANN + CNN	0.0060	0.0775
DUR_TOT	KNN	0.0099	0.0993
	MICE	0.0051	0.0717
	ANN + CNN	0.0040	0.0634

Por otra parte, el RMSE está definido por la ecuación 2, el cual es la obtención de la raíz cuadrada de MSE e indica el error promedio de estimación en relación con la variable que se estima. En este caso particular, al ser normalizados los datos entre 0 y 1, el RMSE nos indica en un porcentaje cuál es la tasa de error entre el resultado obtenido con el real:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\langle y_i + \hat{y}\rangle^2}. \tag{2}$$

Ahora bien, el proceso de reconstrucción se iteró 10 veces para generar un promedio entre los resultados del MSE y RMSE, con la finalidad de tener un valor más acertado al momento de realizar la reconstrucción de parámetros. La Figura 4 muestra el proceso que se sigue para la obtención de resultados.

4. Resultados

Los resultados obtenidos están enfocados en la aplicación de tres modelos distintos para la imputación de datos con distintos porcentajes para siete parámetros en total, donde estos parámetros se seleccionaron con la condición de tener un máximo de 10 % de faltantes de manera real.

Como primera parte, la Tabla 3 presenta los resultados de la reconstrucción de los siete parámetros con el borrado artificial de 10 % faltante, donde, las métricas de MSE y RMSE indican el desempeño de la imputación con las condiciones dadas.

Tabla 4. Resultados de imputación con valores faltantes de 20 %.

Parámetro	Modelo	MSE	RMSE
COLI_TOT	KNN	0.0603	0.2456
	MICE	0.0563	0.2372
	ANN + CNN	0.0704	0.2653
E_COLI	KNN	0.0080	0.0895
	MICE	0.0073	0.0853
	ANN + CNN	0.0037	0.0606
DBO_SOL	KNN	0.0144	0.1199
	MICE	0.0107	0.1032
	ANN + CNN	0.0088	0.0937
DBO_TOT	KNN	0.0205	0.1431
	MICE	0.0188	0.1371
	ANN + CNN	0.0198	0.1406
DQO_SOL	KNN	0.0023	0.0475
	MICE	0.0020	0.0449
	ANN + CNN	0.0019	0.0433
DQO_TOT	KNN	0.0053	0.0727
	MICE	0.0045	0.0674
	ANN + CNN	0.0060	0.0776
DUR_TOT	KNN	0.0044	0.0664
	MICE	0.0039	0.0627
	ANN + CNN	0.0047	0.0689

El análisis de los resultados de MSE en los distintos parámetros reconstruidos a un 10 % de faltantes, existe muy poca diferencia en el resultado por cada modelo generado, obteniendo MSE muy cercanos a 0, indicando que la diferencia entre los valores reales y los imputados son muy similares.

Por otra parte, al observar el RMSE indica el porcentaje de error con el que tiende a equivocarse entre estos parámetros, donde en la mayoría de los parámetros el error radica entre un ± 10 %.

La Tabla 4 presenta los resultados obtenidos de la imputación de valores en un 20 %. Aunque, se aumentó el número de valores faltantes, los valores de RMSE y MSE entre los modelos generados y cada uno de los parámetros se mantienen con una ligera diferencia, manteniendo seis de los parámetros con una diferencia cerca del 10 % a excepción de Coliformes Totales que presenta un RMSE relativamente alto entre un 25 %.

Por último, los resultados obtenidos para la imputación del 30 % se muestran en la Tabla 5, donde los modelos y parámetros mostraron un comportamiento similar de error en el RMSE, que en los anteriores porcentajes de datos faltantes.

Obteniendo así, las pruebas de tres distintos modelos como Vecinos más cercanos, combinación de una red neuronal artificial con redes convolucionales para la imputación individual de cada parámetro y posteriormente el MICE, que representa la imputación múltiple de las siete variables a la vez.

Tabla 5. Resultados de imputación con valores faltantes de 30 %.

Parámetro	Modelo	MSE	RMSE
COLI_TOT	KNN	0.0660	0.2569
	MICE	0.0558	0.2362
	ANN + CNN	0.0690	0.2626
E_COLI	KNN	0.0064	0.0799
	MICE	0.0058	0.0764
	ANN + CNN	0.0052	0.0724
DBO_SOL	KNN	0.0110	0.1047
	MICE	0.0086	0.0930
	ANN + CNN	0.0105	0.1025
DBO_TOT	KNN	0.0258	0.1606
	MICE	0.0231	0.1521
	ANN + CNN	0.0232	0.1522
DQO_SOL	KNN	0.0017	0.0416
	MICE	0.0020	0.0450
	ANN + CNN	0.0038	0.0615
DQO_TOT	KNN	0.0078	0.0884
	MICE	0.0055	0.0741
	ANN + CNN	0.0056	0.0746
DUR_TOT	KNN	0.0039	0.0628
	MICE	0.0034	0.0579
	ANN + CNN	0.0065	0.0806

Comparando los resultados de los tres distintos borrados artificiales para su reconstrucción, se puede observar que no hay una diferencia relevante entre los distintos porcentajes de faltantes, es decir, aunque se tenga 10, 20 o 30 % de datos perdidos, la imputación de dichas variables suele tener un margen de error muy similar entre los distintos modelos de cada parámetro.

De tal manera que, los distintos porcentajes de datos faltantes no influyen en que tan preciso va a ser la imputación, sino la correlación que existe entre los parámetros de entrada para tomar en cuenta la reconstrucción.

El contexto de la información aportada por la base de datos es la que determina que tan buena imputación va a realizar al momento de implementar cualquiera de las técnicas.

5. Conclusiones

El determinar qué método es el mejor para la imputación de valores perdidos en cualquier tipo de variable es difícil. Se necesita explorar todos los métodos existentes de imputación.

Ahora bien, entre los métodos explorados, no existe uno mejor o peor. Debido a los resultados obtenidos, se observa que los resultados no se relacionan con el porcentaje de valores perdidos, sino del hecho de las correlaciones que existen entre los valores a imputar con respecto a las variables de entrada.

Uno de los puntos más importantes dentro de las tareas de aprendizaje máquina, es contar con buenos conjuntos de datos que sean capaces de describir las características de una problemática dada, el hecho de contar con conjunto de datos son valores faltantes se vuelve un reto, ya que no pueden ser descartados simplemente, sino que es necesario encontrar aquellas técnicas que ayuden aportar más información de entrada a las técnicas de aprendizaje máquina y así obtener resultados favorables.

6. Trabajo futuro

Con los resultados obtenidos, se planea explorar otras técnicas del estado del arte, como modelos bayesianos e implementación de bosques aleatorios, para así determinar si existe una mejora en la imputación de datos perdidos mediante estas técnicas.

Posteriormente, es necesario terminar la reconstrucción de todos los parámetros con datos faltantes de la calidad del agua para así realizar una comparativa de predicción de metales pesados en el río fuerte de Sinaloa, con un conjunto de datos reconstruido y sin reconstruir, para determinar si existe una gran mejora con los valores imputados.

Una vez obtenido los resultados, la misma metodología de estimación de metales, se va a extrapolar a la predicción de otros parámetros con las entradas reconstruidas, como pH, turbidez, coliformes fecales.

Referencias

1. Jäger, S., Allhorn, A., Bießmann, F.: A benchmark for data imputation methods. Frontiers In Big Data, vol. 4 (2021) doi: 10.3389/fdata.2021.693674
2. Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma-Mittal, R., Munigala, V.: Overview and importance of data quality for machine learning tasks. In: Proceedings Of The 26th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, pp. 3561–3562 (2020) doi: 10.1145/3394486.3406477
3. Gudivada, V., Apon, A., Ding, J.: Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal On Advances In Software, vol. 10, no. 1-2 (2017)
4. Comisión Nacional del Agua: Calidad del agua en México (2023) https://www.gob.mx/conagua/articulos/calidad-del-agua
5. Jadhav, A., Pramod, D., Ramanathan, K.: Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, vol. 33, no. 10, pp. 913–933 (2019) doi: 10.1080/08839514.2019.1637138
6. Lin, W., Tsai, C.: Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, vol. 53, pp. 1487–1509 (2020) doi: 10.1007/s10462-019-09709-4
7. Enders, C. K.: Applied missing data analysis. Guilford Publications (2022)
8. McKnight, P., McKnight, K., Sidani, S., Figueredo, A. J.: Missing data: A gentle introduction. Guilford Press (2007)
9. Menéndez-García, L., Menéndez-Fernández, M., Sokoła-Szewioła, V., Prado, L., Ortiz-Marqués, A., Fernández-López, D., Sánchez, A. B.: A method of pruning and random replacing of known values for comparing missing data imputation models for incomplete air quality time series. Applied Sciences, vol. 12, no. 6465 (2022) doi: 10.3390/app12136465

10. Lin, W., Tsai, C., Zhong, J.: Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems, vol.239, no. 108079 (2022) doi: 10.1016/j.knosys.2021.108079
11. Camino, R., Hammerschmidt, C., State, R.: Improving missing data imputation with deep generative models. (2019)
12. Zhai, R., Gutman, R.: A Bayesian singular value decomposition procedure for missing data imputation. Journal Of Computational And Graphical Statistics, vol. 32, no. 2, pp. 470–482 (2022) doi: 10.1080/10618600.2022.2107534
13. Ratolojanahary, R., Ngouna, R., Medjaher, K., Junca-Bourié, J., Dauriac, F., Sebilo, M.: Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. Expert Systems With Applications, vol. 131, pp. 299–307 (2019) doi: 10.1016/j.eswa.2019.04.049
14. Silva-Ramírez, E., Cabrera-Sánchez, J. F.: Co-active neuro-fuzzy inference system model as single imputation approach for non-monotone pattern of missing data. Neural Computing And Applications, vol. 33, pp. 8981–9004 (2021) doi: 10.1007/s00521-020-05661-5