DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI
ISSN: 2007-6363

Publicación semestral, Vol. 13, No. Especial (2025) 221–228

Aplicación de un método clasificador a datos de la enfermedad renal crónica y anemia
Application of a classification method to chronic kidney disease and anemia data

V. E. Ramos-Rivero [a]

[a] *Área Académica de Matemáticas y Física, Universidad Autónoma del Estado de Hidalgo, 42184, Mineral de la Reforma, Hidalgo, México.*

Resumen

El análisis de datos médicos es una herramienta fundamental en la toma de decisiones clínicas y en el diagnóstico de enfermedades. La presente investigación se enfoca en la implementación de una interfaz gráfica que permite la representación visual de datos de química sanguínea y la aplicación de un algoritmo de clasificación basado en K-Nearest Neighbors (KNN), con el objetivo de facilitar la interpretación de la información médica y mejorar la detección de anomalías en los parámetros sanguíneos. Para el desarrollo de esta interfaz se utilizó Python para la construcción de la interfaz y la representación gráfica de los datos en 2D y 3D, permitiendo importar archivos ".csv". Además, incorpora la funcionalidad de clasificación mediante KNN, mostrando la matriz de confusión y métricas de desempeño. Los resultados obtenidos muestran que la herramienta facilita la identificación de patrones en los datos y permite la exploración intuitiva de las relaciones entre variables.

Palabras Clave: Anemia, enfermedad renal crónica, K-vecinos más cercanos, matriz de confusión.

Abstract

Medical data analysis is a fundamental tool in clinical decision making and disease diagnosis. The present research focuses on the implementation of a graphical interface that allows the visual representation of blood chemistry data and the application of a classification algorithm based on K-Nearest Neighbors (KNN), with the aim of facilitating the interpretation of medical information and improving the detection of abnormalities in blood parameters. For the development of this interface, Python was used for the construction of the interface and the graphical representation of the data in 2D and 3D, allowing the import of ".csv" files. In addition, it incorporates the KNN classification functionality, showing the confusion matrix and performance metrics. The results obtained show that the tool facilitates the identification of patterns in the data and allows intuitive exploration of the relationships between variables.

Keywords: Anemia, chronic kidney disease, K-nearest neighbors, confusion matrix.

1. Introducción

Las enfermedades renales son afecciones comunes en la actualidad y, al igual que el ser humano, evolucionan con el tiempo. El desarrollo de nuevas patologías, así como la creciente resistencia de bacterias y virus a los medicamentos, ha impulsado la necesidad de investigaciones constantes para la detección, tratamiento y prevención de estas enfermedades. A nivel mundial, la anemia es una de las enfermedades más prevalentes entre mujeres y niños, mientras que la enfermedad renal crónica (ERC) afecta a millones de personas, convirtiéndose en una de las principales causas de mortalidad.

Según datos estadísticos de 2019, los países del continente americano con mayor tasa de mortalidad debido a enfermedades renales son Nicaragua, El Salvador, Bolivia, Guatemala, Surinam, Honduras y Ecuador (Organización Panamericana de la Salud, 2021). En México, la ERC figura como una causa de muerte a partir de los 25 años, registrándose en 2023 un total de 15,928 fallecimientos (INEGI, 2024), de los cuales el 44.3 % correspondió a mujeres y el 55.7 % a hombres.

Por su parte, la anemia afecta principalmente a niños pequeños, mujeres embarazadas y aquellas en edad reproductiva. En los países de ingresos bajos, la incidencia de esta enfermedad es mayor, sobre todo en comunidades rurales. En cifras, para 2019, aproximadamente 539 millones de mujeres no embara-

*Autor para correspondencia: ra421571@uaeh.edu.mx
Correo electrónico: ra421571@uaeh.edu.mx (Victor Efrain Ramos-Rivero);

Historial del manuscrito: recibido el 26/09/2024, última versión-revisada recibida el 25/02/2025, aceptado el 25/02/2025, publicado el 26/04/2025. **DOI:** https://doi.org/10.29057/icbi.v13iEspecial.13816

zadas padecían anemia en todo el mundo (World Health Organization: WHO and World Health Organization: WHO, 2023). En México, entre 2018 y 2019, una de cada cinco mujeres de entre 0 y 49 años presentó esta condición. La pandemia de COVID-19 contribuyó al incremento de la desnutrición y la mortalidad asociada a esta enfermedad. Aunque entre 2006 y 2012 se observó una disminución en la prevalencia de la anemia, en 2018 se registró un nuevo incremento, especialmente en mujeres embarazadas, lo que pudo derivar en complicaciones gestacionales o abortos espontáneos (Shamah Levy y Mejía Rodríguez, 2023).

Al igual que las enfermedades continúan desarrollándose y afectando a la población, la tecnología avanza con el propósito de mejorar el sector salud. Su aplicación no solo permite combatir y, en algunos casos, erradicar enfermedades, sino también optimizar el equipamiento médico y acelerar los procesos de diagnóstico. La inteligencia artificial (IA) ha demostrado ser una herramienta fundamental en este ámbito, contribuyendo al desarrollo de algoritmos de machine learning para el análisis médico, entre ellos el algoritmo de k-nearest neighbors (KNN), redes neuronales, entre otros. Asimismo, la IA permite el diseño de software especializado que mejora la precisión diagnóstica, agiliza la atención médica y facilita el acceso a información clave, lo que puede contribuir a salvar vidas (Moraguez, 2024).

Los algoritmos de clasificación en inteligencia artificial son herramientas que permiten asignar elementos a categorías específicas con base en una variable objetivo. Estos algoritmos analizan los datos de entrada e identifican patrones que facilitan su clasificación (Axolot, 2023). Uno de los métodos más utilizados es el algoritmo de clasificación de vecinos más cercanos el cual será empleado en esta investigación.

El algoritmo KNN es un modelo de aprendizaje supervisado, una subcategoría de machine learning e inteligencia artificial. A medida que se introducen nuevos datos, este modelo ajusta sus ponderaciones para mejorar su precisión (IBM, 2024c). Como en cualquier método estadístico, siempre existe un margen de error, por lo que es necesario aplicar técnicas de validación, como la validación cruzada, los clasificadores de bosques aleatorios o la matriz de confusión.

El propósito de esta investigación es implementar el algoritmo KNN para el análisis de datos relacionados con la anemia y la ERC, dos enfermedades de gran impacto en la sociedad. A través de este modelo, se busca proporcionar una herramienta eficiente para la clasificación de datos médicos, permitiendo obtener resultados rápidos y precisos incluso con grandes volúmenes de información. Finalmente, se pretende determinar el número óptimo de vecinos más cercanos (K) para cada caso y fijarlo como un parámetro definitivo dentro del código desarrollado en Python.

2. Antecedentes

2.1. Persona sana

Una persona sana se define como aquella que goza de bienestar físico, mental y social, más allá de la mera ausencia de enfermedad. Desde el punto de vista médico, una persona sana es aquella que puede desempeñar sus actividades diarias sin limitaciones físicas, con un funcionamiento óptimo de su organismo, resistencia a enfermedades y una adecuada capacidad de recuperación ante lesiones o afecciones (Clínica Universidad de Navarra, 2024).

En términos clínicos, la salud de una persona puede evaluarse mediante valores de referencia establecidos por laboratorios especializados en análisis clínicos. Estos valores permiten determinar el estado general del organismo y su correcto funcionamiento. Algunos de estos parámetros se presentan en la tabla 1.

Dado que no siempre es posible obtener valores de referencia de un solo laboratorio de manera accesible y pública, se optó por recopilar información de diversas fuentes. Como resultado, la tabla 1 incluye valores provenientes de distintas referencias, las cuales se especifican junto a cada parámetro correspondiente.

Persona sana	
Características	Valores de referencia
Presión arterial	120/80 mmHg (Padilla y Abadie, 2021)
Gravedad especifica de la sangre	1.05 - 1.06 (Padilla y Abadie, 2021)
Albumina	3.5 - 5.4 g/dL (Padilla y Abadie, 2021)
Azúcar (en ayunas)	64 - 107 mg/dL (Kratz *et al.*, 2004)
Eritrocitos (glóbulos rojos)	4.2 - 5.9 millones de células/μL (Padilla y Abadie, 2021)
Células de pus	0 - 5 células/HPF (Digital, 2024)
Aglomerados de células de pus	Ausentes
Urea en sangre	12 - 43 mg/dL (de 1 a 17 años); 20 - 68 mg/dL (de 18 a 70 años) (Bustamante, 2019)
Creatinina	0.74 - 1.35 mg/dL (hombres); 0.59 - 1.04 mg/dL (mujeres) (Kratz *et al.*, 2004)
Sodio	135 - 145 mEq/L (Kratz *et al.*, 2004)
Potasio	3.5 - 5.0 mEq/L (Kratz *et al.*, 2004)
Hemoglobina	14 - 17 g/dL (hombres); 12 - 16 g/dL (mujeres) (Padilla y Abadie, 2021)
Hematocritos	41 % - 51 % (hombres); 36 % - 47 % (mujeres) (Padilla y Abadie, 2021)
Recuento de glóbulos blancos	4500 - 11000 células/μL (Padilla y Abadie, 2021)

Tabla 1: Tabla cuyos valores mostrados son los usados como referencia para la presente investigación.

2.2. Anemia por falta de hierro

Este tipo de anemia ocurre cuando el cuerpo no cuenta con una cantidad suficiente de hierro para producir hemoglobina, la proteína de los glóbulos rojos encargada de transportar oxígeno en la sangre. Entre las principales causas se encuentran la pérdi-

da acelerada de glóbulos rojos sin una producción adecuada para reponerlos, o una deficiente absorción del hierro consumido. En el caso de las mujeres, la menstruación abundante o prolongada puede ser un factor determinante.

Los síntomas más comunes incluyen debilidad o fatiga frecuente, dolores de cabeza persistentes, palpitaciones intensas y mareos. A medida que la anemia progresa, estos síntomas pueden agravarse, provocando dificultad para respirar, palidez extrema y caída del cabello (Roberts, 2024).

Existen diversos estudios sobre el tratamiento de esta enfermedad. En particular, se ha realizado una encuesta a médicos en la que se documentan sus métodos para tratar la anemia ferropénica, proporcionando un panorama sobre las estrategias médicas más utilizadas (Manso, 2022).

Los glóbulos rojos tienen un ciclo de vida promedio de 90 a 120 días y representan aproximadamente el 99 % del contenido sanguíneo, mientras que el 1 % restante está compuesto por leucocitos y plaquetas (Lifeder, 2021). Una vez que estas células cumplen su función, el organismo las elimina y envía señales a la médula ósea para generar nuevas.

Los valores más característicos de esta enfermedad se presentan en la tabla 2. Cabe destacar que la anemia también puede derivar en enfermedades renales, lo que resalta la importancia de su diagnóstico y tratamiento oportunos.

Anemia	
Características	Valores de referencia
Eritrocitos (glóbulos rojos)	< 4.5 millones/μL (hombres); < 4 millones/μL (mujeres) (Braunstein, 2022)
Hemoglobina	< 14g/dL (hombres); < 12g/dL (mujeres) (Braunstein, 2022)
Hematocritos	< 42 % (hombres); < 37 % (mujeres) (Braunstein, 2022)

Tabla 2: Tabla que contiene valores de referencia para la anemia ferropénica.

2.3. Enfermedad renal crónica

Este daño crónico, conocido como enfermedad renal crónica (abreviado como *ERC*, o por sus siglas en inglés, *CKD*), es una condición progresiva de largo plazo (generalmente superior a tres meses), en la cual los riñones pierden gradualmente su capacidad para eliminar toxinas de la sangre. Asimismo, se considera ERC cuando existen alteraciones estructurales en los riñones o afectaciones en otras funciones renales, como la pérdida de proteínas en la orina. Además, los pacientes que han recibido un trasplante renal también son clasificados dentro de esta condición (Renal CAre Services, 2022).

Es importante distinguir la enfermedad renal crónica (ERC) de la insuficiencia renal crónica (IRC). La IRC se caracteriza por infecciones recurrentes o persistentes en los riñones, lo que puede derivar en un daño renal permanente. Una de sus principales causas es la ascensión de bacterias desde el tracto urinario hasta los riñones, provocando inflamación y afectaciones en la función renal (Wing y Schiffman, 2022).

Algunas de las principales causas de la enfermedad renal crónica (ERC) incluyen la diabetes, la presión arterial alta (hipertensión), enfermedades cardíacas, obesidad, edad superior a 60 años, antecedentes familiares o personales de la enfermedad y el consumo de tabaco (National Kidney Foundation, Inc., sf). Sin embargo, es posible que la ERC no presente síntomas evidentes durante un tiempo prolongado (Latif, 2023).

Como ocurre con cualquier enfermedad, la ERC presenta síntomas característicos. En sus primeras etapas, pueden incluir falta de apetito, dolores de cabeza y náuseas. A medida que la enfermedad progresa, pueden aparecer manifestaciones más graves, como dolor en los huesos, cambios anormales en la pigmentación de la piel, susceptibilidad a hematomas o incluso la presencia de sangre en las heces (Latif, 2023).

Enfermedad renal crónica	
Características	Valores de referencia
Diabetes	Sí (National Kidney Foundation, Inc., sf)
Hipertensión	Sí (National Kidney Foundation, Inc., sf)
Enfermedad cardíaca	Sí (National Kidney Foundation, Inc., sf)
Aglomerado de células de pus	Presentes

Tabla 3: Tabla que contiene los valores de referencia de la enfermedad renal crónica.

2.4. Método K-NN

El algoritmo KNN (k-nearest neighbors) es un clasificador de aprendizaje no supervisado que emplea la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual (IBM, 2024a). Este clasificador utiliza el concepto de ”voto mayoritario”, ampliamente reconocido en la literatura. Sin embargo, es importante destacar que el algoritmo no requiere necesariamente más del 50 % de los votos para asignar una categoría, ya que se puede establecer un umbral diferente, asignando el porcentaje máximo a una determinada clase.

2.4.1. Métricas de distancia

Para su correcto funcionamiento, el algoritmo KNN requiere la definición de una métrica de distancia que determine la proximidad entre los puntos de datos, en particular entre el punto de consulta y sus vecinos más cercanos. Existen diversas formas de calcular esta distancia, entre las cuales destacan:

- Distancia euclidiana: Representa la longitud de la línea recta entre dos puntos en un espacio.
- Distancia Manhattan: Calcula la suma de los valores absolutos de las diferencias entre dos puntos, considerando únicamente movimientos horizontales y verticales.
- Distancia Minkowski: Generaliza las distancias Euclidiana y Manhattan mediante un parámetro ajustable.
- Distancia de Hamming: Se utiliza para comparar vectores booleanos o cadenas de caracteres, identificando las posiciones en las que no coinciden (IBM, 2024a).

Matemáticamente, la distancia Minkowski se define como:

$$d_{Minkowski} = \left(\sum_{i=1}^{n} |x_i - y_i| \right)^{\frac{1}{p}} \quad (1)$$

donde x_i, y_i indican puntos en un espacio n-dimensional, valor p es un parámetro que determina el tipo de distancia a calcular.

- Para $p = 2$, la ecuación (1) se reduce a la distancia euclidiana.
- Para $p = 1$, se obtiene la distancia Manhattan.

De esta manera, al modificar el parámetro p, es posible adaptar el cálculo de la distancia a distintos contextos y tipos de datos.

2.4.2. *Definición del parámetro K*

Para seleccionar un valor adecuado de K, es fundamental considerar las características de los datos de entrada. En términos generales, se recomienda elegir un valor impar para K con el propósito de evitar empates en la clasificación. Además, es importante tener en cuenta que:

- Valores altos de K conducen a un sesgo alto y varianza baja, lo que hace que el modelo sea más general pero menos sensible a los datos individuales.
- Valores bajos de K resultan en una varianza alta y sesgo bajo, lo que puede hacer que el modelo se ajuste demasiado a los datos de entrenamiento (overfitting) (IBM, 2024a).

Una de las principales ventajas del clasificador KNN radica en su simplicidad de implementación, lo que facilita su uso en diversos problemas de clasificación. Además, su capacidad de adaptación permite incorporar nuevas muestras sin necesidad de reentrenar el modelo por completo. En comparación con otros clasificadores, KNN requiere únicamente definir una métrica de distancia y un único valor de K para su funcionamiento.

2.5. *Matriz de confusión*

La matriz de confusión es una herramienta utilizada para evaluar el desempeño de un modelo de clasificación en el contexto del aprendizaje automático. Su propósito es comparar los valores predichos por el modelo con los valores reales, proporcionando una visualización detallada de los aciertos y errores cometidos. Esta matriz desglosa el número de instancias correctamente clasificadas y las instancias clasificadas erróneamente dentro de cada categoría (IBM, 2024b). A partir de la matriz de confusión, es posible calcular métricas clave para evaluar el desempeño del modelo, tales como:

$$Precision = \frac{VP}{VP + FP} \quad (2)$$

y una coincidencia definida como

$$Coincidencia = \frac{VP}{VP + FN} \quad (3)$$

donde VP significa “verdaderos positivos”, VN significa “verdaderos negativos”, FP significa falsos positivos y FN significa falsos negativos. También la exactitud es definida como

$$Exactitud = \frac{Predicciones_{buenas}}{Predicciones_{totales}} \quad (4)$$

En el caso de la ecuación 4, las $Predicciones_{buenas}$ corresponde al número de aciertos del clasificador y $Predicciones_{totales}$ representa la cantidad total de predicciones realizadas.

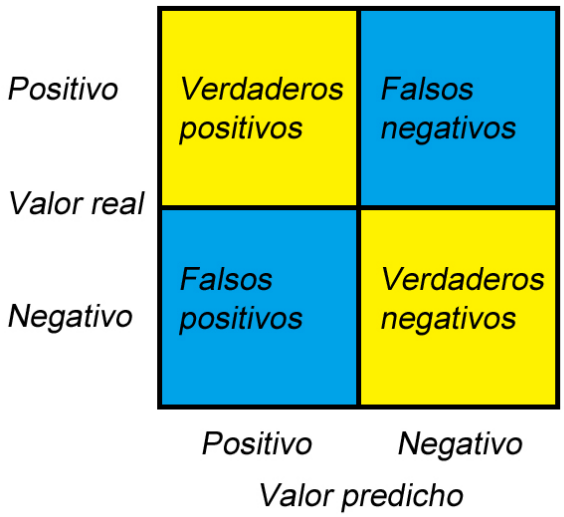

Figura 1: Representación gráfica de una matriz de confusión, en la que se incluyen las abreviaturas utilizadas en las ecuaciones (2) y (3).

3. Métodos

3.1. *Fuente de datos*

Los datos utilizados en esta investigación fueron obtenidos de OpenML (Open Machine Learning Foundation, 2013), una plataforma en línea que proporciona bases de datos, algoritmos y recursos de aprendizaje automático de manera abierta y gratuita. En particular, la base de datos seleccionada para este estudio (Manteo, 2021) contiene información relacionada con análisis de química sanguínea, incluyendo parámetros básicos y algunos valores específicos de interés.

Si bien la fuente original no especifica la procedencia geográfica de los datos recolectados, su organización estructurada facilita su manejo y análisis. Sin embargo, el autor de la base de datos no proporciona detalles sobre el significado o la justificación de cada columna. La elección de esta base de datos se basó en su relevancia para el objetivo del estudio, ya que, tras una búsqueda exhaustiva en diversas plataformas de datos médicos, se encontraron pocas opciones que incluyeran información específica sobre química sanguínea y anemia. Finalmente, OpenML fue la plataforma que mejor cumplió con los criterios establecidos para la investigación.

La base de datos consta de un total de 250 registros distribuidos aleatoriamente, con rangos de edad comprendidos entre 7 y 90 años. Inicialmente, los datos estaban tabulados en un archivo con formato ”.arff”, en el cual los valores de algunas

características se representaban como una combinación de palabras y números. Para facilitar su procesamiento, la información se transformó a un archivo ".csv", realizando las siguientes modificaciones:

- Los valores categóricos fueron convertidos a variables binarias, asignando 0 a aquellos atributos considerados negativos y 1 a los positivos.

  Algunos parámetros numéricos estaban expresados en múltiplos, por lo que fue necesario calcular su valor real antes de proceder con el análisis.

3.2. Interfaz

Se desarrolló una interfaz gráfica en Python con el objetivo de facilitar la visualización de datos médicos almacenados en archivos ".csv". Esta interfaz permite representar los datos en gráficas 2D y 3D, proporcionando botones intuitivos y listas desplegables que muestran los nombres de las columnas de la base de datos.

La herramienta está diseñada para ser fácil de usar, permitiendo a los usuarios seleccionar libremente las variables que desean graficar, sin restricciones a características específicas. Además, admite la posibilidad de graficar un mismo dato contra sí mismo, siempre y cuando se trate de valores numéricos, evitando así posibles confusiones durante la manipulación de los datos.

La interfaz consta de dos ventanas principales:

- Ventana de graficación en 3D
  - Representa los datos en un espacio tridimensional.
  - Permite la rotación de la gráfica en cualquier dirección para visualizar diferentes ángulos.
  - Los puntos se diferencian mediante colores al momento de la graficación.
  - Contiene un panel lateral con botones y listas desplegables.
  - Desde esta ventana, se puede acceder a la segunda ventana mediante el botón "Abrir ventana en 2D".
- Muestra gráficos bidimensionales según la selección del usuario.

La Figura 2 muestra la apariencia de la interfaz antes de importar un archivo ".csv". Por su parte, la Figura 3 ilustra la apariencia de la segunda ventana, correspondiente a la graficación en 2D. Para poder acceder al esta segunda ventana es necesario cargar un archivo antes, lo cual se observa en esta imagen, ya que los campos relacionados a las listas desplegables se muestran con la información "ID" y "Age".

Una vez importado un archivo .csv, las columnas del archivo de la base de datos se muestran como opciones en las listas desplegables, permitiendo su selección para graficación. La Figura 4 muestra la interfaz después de cargar un archivo de datos, dando la libertad de usar los botones disponibles de la forma que más le sea útil a un usuario, así como una elección distinta de columna en la lista desplegable.

Adicionalmente, la interfaz permite aplicar el algoritmo de clasificación K-Nearest Neighbors (KNN), generando una ventana emergente con información relevante, como el número de vecinos seleccionados, la matriz de confusión y detalles sobre las clasificaciones incorrectas. Esta funcionalidad puede resultar útil según el contexto y el propósito del análisis.

Cabe destacar que la interfaz aún se encuentra en desarrollo. Sin embargo, las figuras 3, 2 y 4 representan el estado actual del proyecto, y la implementación del clasificador KNN dentro de la herramienta ya es completamente funcional.

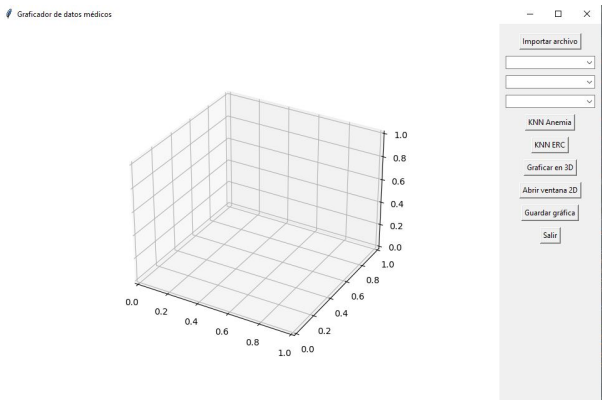

Figura 2: Representación gráfica de la interfaz antes de importar un archivo csv.

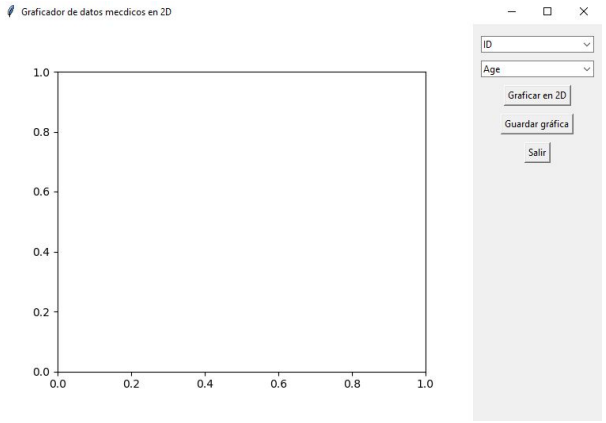

Figura 3: Representación gráfica de la interfaz de la ventana de graficación 2D.

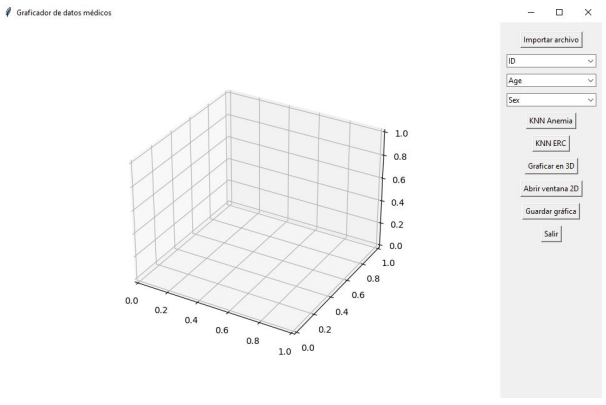

Figura 4: Representación gráfica de la interfaz después de importar un archivo csv.

3.3. Metodología

Dado que no se tuvo acceso a las bases de datos originales, los valores utilizados como "valores de referencia" en esta investigación corresponden a los presentados en la tabla 1. Para evaluar la efectividad del clasificador, se utilizó una matriz de confusión, la cual se muestra en la figura 1.

En las tablas 4, 5, 6, 7, 8 y 9, las columnas etiquetadas como "E", "P" y "C" representan "exactitud", "precisión" y "coincidencia", respectivamente. Las siglas "VP", "VN", "FP" y "FN" corresponden a "verdaderos positivos", "verdaderos negativos", "falsos positivos" y "falsos negativos", respectivamente.

Así mismo, "p" dentro de cada columna, indica el valor que toma el exponente en la ecuación 1, mientras que "K max" hace referencia al número de vecinos que obtuvo los mejores resultados en términos de "E", "P" y "C".

De los 250 registros disponibles en la base de datos utilizada, se seleccionaron 150 para evaluar el desempeño del clasificador. Posteriormente, se añadieron 50 registros adicionales con el fin de analizar la variación en los resultados, y finalmente se incorporaron los últimos 50 para verificar la eficiencia y exactitud del clasificador. A continuación se aplicaron las ecuaciones 2, 3 y 4, considerando distintas métricas con valores de $p = 1$, $p = 2$, $p = 3$ y $p = 4$. Así mismo, los datos fueron evaluados utilizando los valores de $K = 2, 3, 5, 7, 9, 11, 13$. Aunque se observa que las tablas 4, 5, 6, 7, 8 y 9 solo se presenta un valor de K y un valor de p, estos corresponden a la combinación que proporcionó la mayor exactitud.

3.3.1. Estudio para la anemia

A partir de los primeros 150 registros, se identificó que 35 corresponden a usuarios con anemia, mientras que los 115 restantes corresponden a usuarios sin anemia. El clasificador generó diferentes matrices de confusión en función de las variaciones en el valor de p y el número de vecinos considerados.

Es importante destacar que los valores presentados en la tabla 4 corresponden principalmente a verdaderos negativos, ya que los resultados muestran una mayor proximidad a estos en comparación con los verdaderos positivos dentro de la misma tabla. Esto sugiere que el clasificador tiene una mayor capacidad para identificar correctamente a las personas que no padecen la enfermedad.

p	K max	VP	VN	FP	FN	E	P	C
1	13	12	113	2	23	0.9826	0.8308	0.9826
2							0.8303	
3								
4							0.8308	

Tabla 4: Tabla de valores donde la exactitud (E) es máxima con 150 datos, usando distintos valores de p y K para la anemia. Se observa también que el valor del vecino más cercano es independiente del tamaño de p.

Considerando ahora un total de 200 registros, se observa que 47 corresponden a usuarios con anemia y 153 a usuarios sin anemia. Al aplicar el clasificador con distintos valores de p y K, se obtiene la tabla 5.

p	K max	VP	VN	FP	FN	E	P	C
1	13	19	149	4	28	0.9738	0.8418	0.9738
2		21	148	5	26	0.9673	0.8505	0.9673
3	11	22	148	5	25	0.9673	0.8554	0.9673
4	9	23	149	4	24	0.9738	0.8612	0.9738

Tabla 5: Tabla de valores donde la exactitud (E) es máxima con 200 datos, usando distintos valores de p y K para la anemia. Los valores de K se estabilizan a medida que se incrementa el valor de datos al igual que con la métrica.

Se reitera que los valores obtenidos para exactitud, precisión y coincidencia corresponden principalmente a los verdaderos negativos, ya que estos presentan una mayor proximidad a su valor real. En contraste, los verdaderos positivos no alcanzan siquiera la mitad de su valor real.

Considerando ahora un total de 250 registros, se identificó que 60 corresponden a usuarios con anemia y 190 a usuarios sin anemia. Aplicando el mismo procedimiento previamente descrito, se obtiene la tabla 6.

p	K max	VP	VN	FP	FN	E	P	C
1	9	30	182	8	30	0.9578	0.8584	0.9578
	11	28			32		0.8504	
2	9	32	181	9	28	0.9526	0.866	0.9526
3	7	32	181	9	28	0.9526	0.866	0.9526
	9							
4	7	32	181	9	28	0.9526	0.866	0.9526
	9							

Tabla 6: Tabla de valores donde la exactitud (E) es máxima con 250 datos, usando distintos valores de p y K para la anemia. Los valores de K se estabilizan en $K = 9$

Se destaca nuevamente que los resultados presentados en la tabla 6 se basan en los verdaderos negativos, los cuales, en este conjunto de datos, exhiben valores repetidos dentro de una misma métrica.

3.3.2. Estudio para la enfermedad renal crónica

Tomando los primeros 150 registros, de manera similar al caso de la anemia, se identificó que, según los cúmulos de células de pus, 27 corresponden a usuarios con enfermedad renal crónica (ERC), mientras que 123 no presentan la enfermedad. A partir de estos datos y utilizando los valores de p y K, se obtiene la tabla 7.

p	K max	VP	VN	FP	FN	E	P	C
1	9	0	122	1	27	0.9918	0.8187	0.9918
	11							
2	9							
	11							
3	9							
	11							
4	9							
	11							

Tabla 7: Tabla de valores donde la exactitud (E) es máxima con 150 datos, usando distintos valores de p y K para la ERC. Se observa que independientemente del valor de p, el valor del vecino más cercano tiene dos valores que pueden usarse según el contexto.

Tras evaluar los primeros 150 registros, se incorporaron 50 adicionales, obteniéndose un total de 36 usuarios diagnosticados con la enfermedad y 164 sin diagnóstico positivo. Como resultado, se generó la tabla 8.

p	K max	VP	VN	FP	FN	E	P	C
1								
2	13	3	162	2	33	0.9878	0.8307	0.9878
3								
4								

Tabla 8: Tabla de valores donde la exactitud (E) es máxima con 200 datos, usando distintos valores de p y K para la ERC. Se observa que el valor de K es independiente del tamaño de la métrica..

Finalmente, se analizaron los 250 registros de la base de datos utilizada, de los cuales 42 correspondían a usuarios con la enfermedad y 208 a usuarios sin diagnóstico positivo. Como resultado, se obtuvo la tabla 9.

p	K max	VP	VN	FP	FN	E	P	C
1								
2	13	0	206	2	42	0.9903	0.8306	0.9903
3								
4								

Tabla 9: Tabla de valores donde la exactitud (E) es máxima con 250 datos, usando distintos valores de p y K para la ERC. Se observa que el valor de K es independiente del tamaño de la métrica.

Es importante destacar que los valores de exactitud, precisión y coincidencia en las tablas 7, 8 y 9 corresponden a los verdaderos negativos, ya que, en cada caso, estos se aproximaron más a su valor real. En contraste, los verdaderos positivos presentaron una mayor desviación respecto a su valor real.

4. Resultados

Dentro de la metodología, la elección de los valores de K y p no fue completamente aleatoria. Los valores impares de K fueron seleccionados para garantizar que el número de vecinos más cercanos siempre fuera impar, evitando así posibles empates en la clasificación. Aunque se incluyó un caso particular con $K = 2$, su impacto en los resultados finales fue insignificante, por lo que no se consideró un candidato relevante para la selección del vecino más cercano.

Por otro lado, la elección de los valores de p se basó en la consideración de las primeras dos métricas de distancia más utilizadas: Manhattan ($p = 1$) y Euclidiana ($p = 2$). Adicionalmente, se incluyeron otros dos valores con el propósito de explorar su comportamiento en un espacio más amplio y evaluar su impacto en el desempeño del clasificador.

Los datos procesados para su clasificación mediante el algoritmo presentan valores que favorecen distintos escenarios. En el caso de la anemia, con un conjunto de 150 registros, los 13 vecinos más cercanos siempre alcanzan el valor máximo de exactitud, precisión y coincidencia, independientemente del valor de p utilizado en la métrica.

Cuando el número de registros aumenta a 200, el valor de los vecinos más cercanos varía en función del incremento de p. Sin embargo, los valores de exactitud, precisión y coincidencia presentan solo ligeras variaciones, lo que sugiere que la elección de K podría no ser determinante. No obstante, esta tendencia no se refleja en los verdaderos negativos, cuyos valores máximos se alcanzan específicamente cuando $K = 13$ o $K = 9$.

Dado que las tablas 4, 5 y 6 solo consideran los valores máximos dentro del conjunto de vecinos definidos en la metodología, se observa en la tabla 5 la presencia de distintos valores de exactitud con variaciones mínimas entre ellos. Además, conforme el valor de p aumenta, surge un nuevo valor de K menor que debe tomarse en cuenta.

Finalmente, al emplear el conjunto completo de datos para la clasificación de anemia, comienzan a aparecer múltiples valores repetidos, como se evidencia en la tabla 6. En esta, los valores de $p = 1, 2, 3, 4$ coinciden en que $K = 9$ es el vecino más cercano. Esto podría sugerir que, a medida que el número de datos aumenta, el valor óptimo de K tiende a estabilizarse en un valor más preciso. No obstante, esta hipótesis se desestima al analizar los verdaderos negativos, que alcanzan un valor de 182 cuando $p = 1$ y $K = 9$, logrando la mayor exactitud y coincidencia dentro de la tabla, pero con una menor precisión.

Con este tercer conjunto de datos, se puede concluir que el uso de 9 vecinos más cercanos es la opción más adecuada para la clasificación. Por lo tanto, dicho valor debería implementarse de manera definitiva en el código desarrollado.

En cuanto a la Enfermedad Renal Crónica (ERC), los primeros 150 datos presentan valores repetidos en todos los aspectos, como se muestra en la tabla 7. Independientemente del valor de p o de si $K = 9$ o $K = 11$, los valores de exactitud, precisión y coincidencia permanecen constantes, lo que indica que cualquiera de estos valores de K es válido.

Al aumentar el número de datos a 200, el clasificador determina que el número óptimo de vecinos más cercanos es $K = 13$, mostrando valores idénticos de exactitud, precisión y coincidencia para todos los valores de p. Finalmente, al emplear el conjunto completo de datos, se observa que el comportamiento se mantiene sin cambios, ya que $K = 13$ sigue siendo el valor óptimo para la clasificación.

Cabe destacar que la exactitud es un factor relevante al comparar distintos conjuntos de datos. Como se aprecia en la tabla 3, la clasificación obtenida se acerca a un resultado casi perfecto. Dado el comportamiento repetitivo del vecino más cercano en cada conjunto de datos, se concluye que el valor de $K = 13$ debe establecerse de manera fija en el código.

En conclusión, el objetivo de esta investigación se ha cumplido satisfactoriamente con el número de datos extraídos de la base de datos, logrando determinar con precisión los valores óptimos de K para la clasificación de anemia y ERC.

5. Discusión

Aunque los resultados presentados corresponden a un conjunto de datos relativamente pequeño, no se especifican las condiciones bajo las cuales fueron obtenidos, lo que deja un vacío en la interpretación de los valores generados por el clasificador. La falta de información sobre el origen de la base de datos, así como sobre la preparación o los estudios realizados a cada paciente, son aspectos fundamentales que no pueden ser ignorados al llevar a cabo este tipo de análisis, ya que su relevancia

práctica puede variar según el contexto. Como se mencionó en la introducción, los datos sobre la anemia y la Enfermedad Renal Crónica (ERC) adquieren significados distintos dependiendo del marco en el que se analicen.

Durante el desarrollo de la interfaz, el valor óptimo del vecino más cercano (K) se determinó a partir de un análisis detallado de cada conjunto de datos, seleccionando aquel que mejor se adaptara a la clasificación. No obstante, se prevé la incorporación de nuevos parámetros y funcionalidades con el objetivo de hacer la interfaz más comprensible y útil en aplicaciones reales dentro del ámbito hospitalario.

Así mismo, la presencia de valores repetidos dentro de un mismo valor de p en el clasificador podría indicar que los parámetros seleccionados no son los más adecuados. Además, el hecho de que en el apartado de ERC apenas se hayan obtenido verdaderos positivos sugiere la necesidad de incluir características adicionales en el modelo. En cualquiera de estos casos, es necesario realizar un análisis más exhaustivo del problema. Por el momento, los resultados obtenidos indican que el clasificador presenta un mejor desempeño en la identificación de individuos que no padecen la enfermedad.

En esta investigación, la principal limitante de la base de datos utilizada es la falta de información sobre los criterios empleados por el autor para la elección de los parámetros. Como consecuencia, aunque el clasificador desempeñe adecuadamente su función, la implementación de datos reales proporcionados por el sector salud podría generar resultados con un sesgo distinto al obtenido en este estudio.

Por ello, en futuros trabajos se espera contar con datos cuya selección tenga una justificación clara, lo que permitirá obtener valores más representativos y mejorar la precisión del clasificador. Además, se plantea la optimización del código para un mejor procesamiento y análisis de los datos, garantizando así una evaluación más robusta y confiable.

Agradecimientos

Agradezco a la Dra. Erika Elizabeth Rodríguez Torres y al cuerpo académico en sistemas dinámicos por el apoyo y guía durante la realización de este trabajo.

Referencias

Axolot (2023). Inteligencia artificial ¿qué es un algoritmo de clasificación? - axolot agencia.

Braunstein, E. M. (2022). Evaluación de la anemia.

Bustamante, J. M. (2019). Todo lo que tienes que saber sobre urea en la sangre normal.

Clínica Universidad de Navarra (2024). Qué es salud. diccionario médico. clínica u. navarra.

Digital, O. (2024). Pus cells in urine: Normal range, causes, symptoms, tests and treatment.

IBM (2024a). Knn.

IBM (2024b). Matriz de confusión.

IBM (2024c). Supervised learning.

INEGI (2024). Estadísticas de defunciones registradas (EDR).

Kratz, A., Ferraro, M., Sluss, P. M., y Lewandrowski, K. B. (2004). Normal reference laboratory values. *New England Journal of Medicine*, 351(15):1548–1563.

Latif, W. (2023). Enfermedad renal crónica: Medlineplus enciclopedia médica.

Lifeder (2021). Eritrocitos (glóbulos rojos).

Manso, G. L. H. (2022). Diagnóstico y tratamiento de la anemia ferropénica en la asistencia primaria de españa. *Medicina Clínica Práctica*, 5(4):100329.

Manteo, J. (2021). Base de datos: Chronic kidney disease.

Moraguez, E. R. (2024). Inteligencia artificial en medicina: Cómo la ia está salvando vidas.

National Kidney Foundation, Inc. (s.f.). Enfermedad renal crónica (erc).

Open Machine Learning Foundation (2013). Openml. `https://www.openml.org/`.

Organización Panamericana de la Salud (2021). Carga de enfermedes renales. `https://www.paho.org/es/enlace/carga-enfermedes-renales`.

Padilla, O. y Abadie, J. (2021). Valores normales de laboratorio.

Renal CAre Services (2022). Diferencias entre enfermedad renal crónica erc e insuficiencia renal crónica irc.

Roberts, J. (2024). Anemia ferropénica: Medlineplus enciclopedia médica.

Shamah Levy, T. y Mejía Rodríguez, F. (2023). La anemia entre las mujeres mexicanas en edad fértil.

Wing, E. J. y Schiffman, F. J. (2022). *Cecil. Principios de medicina interna*. Elsevier Health Sciences.

World Health Organization: WHO and World Health Organization: WHO (2023). Anemia.