Utilidad de los modelos de machine learning en el diagnóstico de enfermedad renal crónica

Sol María Carriazo Julio

Resumen— La enfermedad renal crónica (ERC) es un problema de salud pública que afecta a más de 850 millones de personas en el mundo, con importantes complicaciones y mayor mortalidad, comparado con la población general. Por lo tanto, es imprescindible su diagnóstico temprano para su tratamiento oportuno. Este estudio aplica técnicas de machine learning, con el fin de identificar patrons y construir un modelo predictivo de ERC, utilizando una base de datos de acceso público. Tras una rigurosa preselección y preprocesamiento de variables, se emplearon algoritmos de clasificación y análisis de clústeres en la plataforma BigML, evaluando su efectividad para agrupar y clasificar a pacientes basado en sus características clínicas y demográficas. Los resultados destacan la eficacia de los modelos supervisados, especialmente el modelo de deepnet, que mostró alta precisión en la clasificación de pacientes en riesgo. Pese a que los hallazgos destacan el potencial de ML para mejorar el diagnóstico de ERC, el estudio enfrenta limitaciones inherentes a los datos de acceso público, incluyendo la falta de la variable sexo, mitigada mediante predicción algorítmica. Este trabajo sugiere que ML puede despeñar un rol importante en la detección temprana de ERC, promoviendo estudios colaborativos que mejoren la representación y precisión de los datos.

Palabras clave — Enfermedad renal crónica, machine learning, aprendizaje supervisado, aprendizaje no supervisado.

Abstract— Chronic kidney disease (CKD) is a public health concern impacting near to 850 million people worldwide, resulting in a higher mortality rate compared to the general population. Given this elevated risk, an early diagnosis plays a critical role in effective treatment. This study explores the use of machine learning techniques applied to a publicly accessible dataset to identify patterns and develop predictive models for CKD diagnosis. Following a rigorous selection and preprocessing of variables, classification algorithms and cluster analysis were applied using the BigML platform, evaluating their effectiveness in grouping and classifying patients based on clinical and demographic characteristics. The results highlight the effectiveness of supervised models, particularly the deepnet model, which showed high accuracy in classifying patients at risk. While the findings underscore ML's potential to improve CKD diagnosis, the study faces inherent limitations of public-access data, including the lack of a gender variable, which was addressed through algorithmic prediction. This work suggests that ML could play a crucial role in early CKD detection, encouraging collaborative studies to enhance data representation and accuracy.

Index Terms—Chronic kidney disease, machine learning, supervised learning, unsupervised learning.

1 INTRODUCCIÓN

La enfermedad renal crónica (ERC) es un problema de salud pública que afecta a más de 850 millones de indivuos a nivel mundial, y está asociado con aumento de mortaliad y morbilidad (1). El diagnóstico temprano es crucial para enlentecer su progresión y reducir las complicaciones como la, la necesidad de terapia de reemplazo renal y enfermedades cardiovasculares asociadas. Sin embargo, los métodos tradicionales, basados principalmente en parámetros como la tasa de filtración glomerular y la creatinina sérica, no siempre son suficientes para una detección temprana y precisa.

La ERC se define, de acuerdo con las guías KDIGO 2012, como anormalidades en la estructura o función del riñón presentes durante más de 3 meses, con implicaciones para la salud. Sólo un criterio que identifique una alteración estructural o función renal anormal permite el diagnóstico de ERC. Los criterios incluyen una tasa de filtración glomerular estimada baja (TFGe <60 mL/min/1.73 m²) o evidencia de daño renal, como albuminuria patológica [relación albúmina en orina (UACR) ≥30 mg/g]; sedimento urinario, histología o imágenes anormales; anormalidades debido a trastornos tubulares o trasplante renal. En la práctica clínica, esto significa que para diagnosticar ERC cuando la TFGe es ≥60 mL/min/1.73 m² requiere un análisis de orina o imágenes renales (2).

Los valores de TFGe menores de 60 mL/min/1.73 m² indican ERC porque reflejan una pérdida promedio de al menos 50% del filtrado glomerular en un adulto jóven. Múltiples estudios han encontrado que filtrados glomerulares más bajos se encuentran asociados con un incremento en el riesgo relativo de eventos adversos en adultos incluyendo eventos cardiovasculares y mortalidad en todas las

- *E-mail de contacto: somacaju@hotmail.comm*
- *Trabajo tutorizado por: Remo Suppi*
- *Curso MO61505*

categorías de edad (3).

Pese a que es universalmente aceptado que la presencia de albuminuria anormal indica ERC, el uso apropiado de un único límite de TFGe para todas las edades, independientemente de la albuminuria es un tema de debate. Esto surge debido a que el filtrado glomerular disminuye con la edad avanzada, por lo que una definición basada en un valor fijo podría conllevar a un subdiagnóstico en individuos jóvenes, y un sobrediagnóstico en individuos mayores, en los que el filtrado glomerular disminuye con la edad (4).

Por lo tanto, se ha propuesto una definición de ERC adaptada a la edad, con puntos de corte de 75, 60 y 45 mL/min/1.73 m^2 para menores de 40, 40-64 y 65 años, respectivamente. Recientemente, Liu et al., (5) realizaron un estudio poblacional para identificar adultos con ERC incidente de acuerdo con el umbral tradicional de <60 ml/min/1.73m^2, y al corte adaptado para la edad, estimando la incidencia de ERC y el riesgo de fallo renal y muerte. Los autores encontraron que en utilizando el umbral tradicional, el 75% de los adultos mayores de 65 años o más con TFGe de 45 a 59 mL/min/1.73 m^2 con albuminuria leve o normal, tenían un riesgo de fallo renal y de muerte similar a aquellos controles sin ERC, sugiriendo que podría ser de utilidad utilizar valores adaptados para la edad.

En las últimas décadas, la inteligencia artificial y, más específicamente, los modelos de *machine learning* (ML) han mostrado un gran potencial para transformar el diagnóstico médico. El uso de algoritmos con la capacidad de aprender de grandes cantidades de datos ha permitido desarrollar modelos predictivos capaces de identificar patrones complejos que los métodos estadísticos tradicionales no pueden detectar. Estos avances han demostrado ser especialmente útiles en la identificación de biomarcadores y en la clasificación de enfermedades complejas como la ERC (6).

En el mundo contemporáneo caracterizado por el flujo y disponibilidad de grandes datos, la creación de modelos diagnósticos que puedan predecir el riesgo de ERC es vital, ya que pueden permitir el apoyo de profesionales de la salud para realizar un diagnóstico adecuado y predecir el deterioro de la función renal.

En este trabajo evaluaremos la utilidad de una base de datos de acceso libre y el uso de técnicas de ML para el diagnóstico del daño renal.

1.1 Estado del arte: ERC y Machine learning

Diversos estudios recientes han explorado el uso de ML para mejorar la precisión en el diagnóstico de ERC.

Cao et al. (7), utilizó diferentes algoritmos de ML para predecir el filtrado glomerular estimado basado en 24 variables en una base de datos de screening sanitario, encontrando buen rendimiento. El AUROC de los diferentes modelos fue de 0.85 y el algoritmo de gradient boosting exhibió la mejor precisión, con un AUROC de 0.914 dentro de los algoritmos validados. En otro estudio, Salekin y Stankovic (8), evaluaron clasificadores como KNN, random forest y redes neuronales en un conjunto de datos de 400 pacientes. Implementaron una selección de características con el método "wrapper" y seleccionaron cinco características para la construcción del modelo. La mayor precisión fue obtenida por el modelo Random Forest (98%) con un error de raíz cuadrada media (RMSE) de 0.11.

Almasoud y Ward (9) evaluaron la capacidad de distintos algoritmos de ML para predecir ERC usando una selección de características basada en correlación de Pearson, ANOVA y Cramer's V. Concluyeron que el algoritmo de Gradient Boosting (GB) tuvo la mayor precisión con un F-1 score de 99.1

Más allá del diagnóstico de ERC, diferentes autores han publicado el uso de modelos de ML para predecir enfermedades renales específicas, y el uso de biomarcadores para el diagnóstico de complicaciones relacionadas a la ERC, como trastornos del metabolismo mineral óseo. Dichos estudios ilustran cómo los modelos de ML podrían ser de utilidad para el enfoque diagnóstico de la ERC, mejorando no solo la precisión, sino también la capacidad de personalizar el tratamiento para diferentes pacientes. (10)

A pesar de los avances en el uso de ML para el diagnóstico de ERC, persisten algunos desafíos, como la necesidad de bases de datos de alta calidad y la dificultad para interpretar los resultados obtenidos de los modelos más complejos. Sin embargo, la integración de ML en la práctica clínica promete una mejora continua en la atención de los pacientes con daño renal.

2 OBJETIVOS

2.1 Objetivo general

El objetivo principal de este estudio es explorar la utilidad de métodos de ML para el diagnóstico del daño renal, utilizando bases de datos de acceso público.

2.2 Objetivos específicos

Objetivo 1: Evaluar la calidad y utilidad de un dataset de acceso público para entrenar modelos de ML.

Objetivo 2: Agrupar a los pacientes según sus características clínicas y demográficas para identificar patrones y subgrupos, mediante un análisis de *clusters*.

Objetivo 3: Desarrollar un modelo predictivo que permita identificar a los pacientes en riesgo de tener daño renal.

3 MATERIALES Y MÉTODOS

3.1 Conjunto de datos

El conjunto de datos fue descargado de una fuente pública, en la plataforma Kaggle. Los datos fueron recolectados en un hospital en Karaikudi, Tamilnadu, India, y está compuesto por 400 instancias y 25 atributos. Se trata de un dataset estructurado, diseñado para predecir el *outcome* ERC. Incluye variables relacionados con datos demográficos y otras correspondientes a analíticas de sangre y orina tales como: edad, presion arterial; en orina: densidad, albumina, glucosa, globulos rojos, células de pus, grupos de células de pus y bacterias. Parámetros sanguíneos: glucos al azar, urea, creatinina, sodio, potasio, hemoglobina, hematocrito, glóbulos blancos, glóbulos rojos. Adicionalmente incluía variables categóricas como presencia de hipertensión, diabetes mellitus, enfermedad coronaria, edema de MMII, anemia y la clasificación (presencia o no de ERC). En el **Anexo 1** se muestran las características de las variables.

3.2 Preprocesado de datos

Selección de variables

Para este trabajo, se realizó una seleccion de variables de forma estratificada, basado en varios criterios:

-*Missing values:* Se calculó el número y porcentaje de missing values por cada variable y se decidió eliminar las variables con más de 30%. El número de missing values se encuentra representado en el **Anexo 1.**

-*Anomaly detector y outliers:* Se calculó un *Anomaly detector* utilizando BigML. Este modelo calcula un "anomaly score" para cada instancia, e indica qué tan lejos está esa instancia de lo que se considera "normal en el conjunto de datos". Dentro de las anomalias detectadas, se encontraron edades muy bajas, y que algunas variables categóricas como hipertension tenían números, o otros caracteres. La variable apetito tenia una entrada de "no", la variable edema tenia una entrada "good", y la variable class tenía entradas de "no" y de "noCKD", entre otros. La variable creatinina presentaba valores muy altos, decidiendo eliminar a aquellos valores por encima de 20 mg/dl. Adicionalmente, la variable potasio tenía valores extremadamente altos de 47 y 39 mmol/L. (**Anexo 1 y 3**)

Otras anomalías detectadas, correspondían a la naturaleza de la enfermedad, y no eran considerados anómalos para este dataset. Los valores que eran considerados como anomalías según criterio clínico fueron tratados como missing values.

-*Colinealidad:* Se evaluó la colinealidad entre las variables para detectar si había alta correlación entre ellas, con el fin de eliminar variables con valores por encima de 0.8 o por debajo de -0.8. No se detectaron valores que cruzaran ese límite.

-*Criterio clínico:* Se decidió eliminar la variable presión arterial ya que no era claro si correspondía a valores de presión arterial sistólica, diastólica o media, por lo que se seleccionó la variable categórica hipertensión arterial, e igualmente la variable grupos de células de pus. La variable hemoglobina, se encontraba relacionada con la variable anemia, por lo que se decidió eliminarla y utilizar la variable categórica.

Finalmente, se eliminaron a aquellos individuos con edad menor de 18 años, por las diferencias en los métodos para calcular la función renal, con respecto a los adultos. Las variables seleccionadas para el análisis inicial fueron: edad, densidad urinaria, albúmina urinaria, glucosa urinaria, células de pus y bacterias urinarias. Así como glucosa, urea, creatinina, sodio, potasio, leucocitos, y variables categóricas de hipertensión, diabetes mellitus, enfermedad coronaria, edema de miembros inferiores, anemia y la clase (ERC).

Transformación de variables

Los *missing values* de las variables numéricas fueron imputados por su mediana, y los de las variables categóricas fueron imputados por su moda. Las variables categóricas del dataset fueron transformadas a formato binario.

Reducción de dimensionalidad

Para valorar la posibilidad de reducir dimensionalidad del dataset, se condujo un análisis de componentes principales (PCA) mediante BigML **(Anexo 4)**. Para este análisis, se excluyó la variable *outcome* (ERC). En el análisis se obtuvieron 31 componentes principales, observando que el PC1 explicaba el 19.51% de la varianza, mientras que el 90% de la varianza acumulada del dataset, se encontraba explicada por 23 componentes principales. Por lo tanto, se decidió descartar el uso de PCA para este estudio y mantener las variables originales.

Evaluación de la variable resultado

Este dataset presentaba etiquetas de clasificación de ERC (si, no). Sin embargo, llamaba la atención que existía un número de pacientes con valores de creatinina elevados, con diagnóstico negativo de ERC. Por lo tanto, se decidió realizar un análisis inicial para valorar la calidad de las etiquetas de los datos. Desafortunadamente, el dataset no incluía la variable sexo, la cual es de vital importancia en el análisis de datos clínicos, debido a las diferencias biológicas entre hombres y mujeres, con gran relevancia en el diagnóstico de ERC.

Como se mecionó previamente, el diagnóstico de ERC está basado principalmente en el cálculo de la TFGe (<60 ml/min/1.73m^2) en ausencia de de otros parámetros. Sin embargo, el cálculo de la TFGe no era posible en este dataset, ante la ausencia de la variable sexo.

Por lo tanto, se decidió entrenar un modelo para predecir la variable sexo **(Anexo 5)**, y añadirla como variable para construir el outcome TFGe baja para la edad.

Finalmente, al dataset se añadió la variable sexo (estimado), y se asignó el filtrado glomerular correspondiente para dicho sexo, creando posteriormente la variable outcome: TFGe baja para la edad.

3.3 Estrategia de análisis

En este estudio, se aplicaron herramientas de machine learning, utilizando como técnica de aprendizaje no supervisado el estudio de *clusters*, y exploración de diversos algoritmos de clasificación y regresión para analizar nuestra base de datos, en la plataforma BigML (https://bigml.com/).

Los algoritmos utilizados para este estudio fueron los siguientes: *No supervisado:* Estudio de *clusters*. *Supervisados*: regresión logística, decision tree, random forest, boosted trees, deepnet. (11)

- *Estudio de clusters:* Permite agrupar los datos del dataset en conjuntos que comparten características similares, lo que facilita la identificación de patrones con diferencias claras entre sí.

- *Regresión logística:* Ideal en este dataset, ya que estima la probabilidad de que un evento ocurra en función de una o más variables predictoras, siendo útil en problemas de clasificación binaria.

- *Decision tree:* Permite clasificar o predecir el resultado del conjunto de datos, basados en un conjunto de reglas. Se construye dividiendo repetidamente los datos en subconjuntos basados en características que maximizan la ganancia de información o minimizan la impureza.

- *Random Forest:* Permite la combinación de múltiples árboles de decisión para mejorar la precisión y robustez del modelo. Cada árbol se entrena con una muestra aleatoria del dataset, y el resultado final se obtiene promediando/votando los resultados de todos los árboles.

- *Boosted trees:* Es una variante de los árboles de decisiones donde los modelos se entrenan de forma secuencial, y cada modelo nuevo intenta corregir los errores cometidos por los anteriores.
  Este enfoque mejora la precisión del modelo final al enfocarse en las observaciones que fueron más difíciles de predecir.

- *Deepnet:* Es un modelo compuesto por múltiples capas de nodos que se entrenan para detectar patrones complejos en los datos. Son particularmente útiles en este tipo de tareas ya que los datos presentan relaciones no lineales o de alta dimensionalidad.

Modelo no supervisado

Para decidir el número óptimo de *clusters*, se utilizó el método G means, el cual es una extensión del algoritmo de clustering K-Means, diseñado específicamente para determinar automáticamente el número óptimo de *clusters* en el conjunto de datos. A diferencia de K-Means, donde el número de *clusters* debe especificarse previamente, G-Means utiliza un enfoque estadístico para ajustarlo de manera dinámica. Su principal ventaja es que automatiza su determinación, lo que es útil cuando no se conoce de antemano la estructura de los datos.

Es adecuado para conjuntos de datos grandes y complejos donde los *clusters* pueden no ser evidentes o varían en tamaño y forma. Para hacer el análisis de *clusters*, se excluyó la variable filtrado glomerular estimado para la edad, TFGe, clase. Posteriormente, se analizaron los centroides, y las diferencias entre cada *cluster*.

Modelos supervisados

El conjunto de datos se dividió en un 80% para entrenamiento (training), y 20% para evaluación (test).

Tras realizar la división de los datos, se evaluaron los diferentes algoritmos para identificar cuál presentaba mejores rendimientos. Las métricas de rendimiento utilizadas fueron: recall, precisión, F1 score, accuracy. Se utilizó la curva ROC (Receiver Operating Characteristic) para visualizar los verdaderos positivos y verdaderos negativos. El **Anexo 5** muestra las definiciones de cada una de las métricas a utilizar.

Predicción

Por último, se pidió a un nefrólogo de Guadalajara, México, que proporcionara dos casos clínicos de su consulta que sirvieran como ejemplo de predicciones del modelo seleccionado.

4 RESULTADOS

Evaluación de la calidad del dataset

Como características generales, el dataset tenía una edad media de 53.65 años, el 48.8% eran hombres, y 38 y 34% tenían antecedentes de hipertensión y diabetes, respectivamente. La media de creatinina era 2,61 mg/dl, con TFGe de 55.8 ml/min/1.73m^2. El 37% de los individuos tenían albuminuria positiva. La **tabla 1** muestra las principales características generales del dataset.

Al haber calculado la TFGe, con el dataset resultante fue posible estimar la fiabilidad de la clasificación inicial de ERC.

La **tabla 2** muestra la clasificación inicial de ERC comparada con la clasificación basada en TFGe calculada para cada sexo. Dentro del grupo inicialmente clasificados

como no ERC, se puede observar que 52 individuos tenían una TFGe <60 ml/min/1.73m². Por el contrario, 19 individuos tenían alguna TFGe >60 ml/min/1.73m², aunque, como se mencionó previamente, esto no sería excluyente con el criterio de ERC basada en la clasificación de KDIGO.

Debido al riesgo de misclasificación en el dataset original, ante la ausencia de más información de la recolección de los datos y condiciones clínicas de los pacientes, finalmente se decidió utilizar el outcome TFGe baja para la edad, para efectos de este trabajo.

Variable	
Edad (media, SD)	53,65+- 14,46
Sexo masculino (n;%)	187 (48,83)
Hipertensión (n;%)	144 (37,60)
Diabetes mellitus (n;%)	131 (34,20)
Enfermedad coronaria (n;%)	40 (10,44)
Pérdida de apetito (n;%)	90 (23,50)
Edema de miembros inferiores (n;%)	70 (18,28)
Anemia (n;%)	61 (15,93)
Glóbulos blancos (media, SD)	8280,68 (2507,21)
Glucosa al azar (media, SD)	146,96 +-49,95
Sodio (media, SD)	138,07 +-6,46
Potasio (media, SD)	4,40 +-0,67
Creatinina sérica (media, SD)	2,61 +-3,18
TFGe (media, SD)	55,85 +-40,45
TFG disminuida para la edad (n;%)	213 (55,61)
Albuminuria (cruces) (n;%)	
0	240 (62,66)
1	41 (10,70)
2	42 (10,97)
3	38 (9,92)
4	21 (5,48)
5	1 (0,26)

Tabla 1. *Características generales de la población.*

	TFGe <60			TFGe >60		
	Hombre	Mujer	Cualquiera	Hombre	Mujer	Cualquiera
No ERC (158)	11	52	52	147	106	106
Si ERC (225)	181	19	206	44	206	19

	TFGe normal para la edad	TFGe baja para la edad	Total
No ERC	163	3	166
Si ERC	7	210	217
Total	170	213	

Tabla 2. *Diagnóstico de enfermedad renal crónica según el criterio clásico y según el criterio ajustado para la edad.*

Un total de 213 individuos en el dataset presentaban una TFGe baja para la edad, de los cuales 173 eran mujeres y 39 eran hombres.

Clusters

Se realizó un análisis de G-means excluyendo las variables TFGe, urea, creatinina, e incluyendo la variable albúmina urinaria, excluyendo el resto de variables urinarias. Se obtuvio en total 4 clusters distribuidos de la siguiente manera. **(Figura 1).**

-*Cluster 0 (C0)*: 28 instancias (7.31%)
-*Cluster 1 (C1):* 22 instancias (5.74%)
-*Cluster 2 (C2):* 128 instancias (33.42%)
-*Cluster 3 (C3):* 205 instancias (53.52%)

La **figura 2** muestra los centroides de cada clúster, proporcionando información clave sobre las diferencias entre ellos.

Figura 1. *Distribución de los clusters.*

La mayoría de las instancias se encuentran agrupadas en el C3 y el C2, mientras que los C0 y C1 contienen un número significativamente menor de instancias, sugiriendo que las características en los C2 y C3 son más comunes en los datos.

Métricas del análisis de *Cluster*:

Total suma de cuadrados: 25.37
Suma de cuadrados intracluster: 16.97
***Suma de cuadrados entre clusters*:** 8.40
Ratio de la suma de cuadrados (ratio_ss): 0.33

El *ratio_ss* es relativamente bajo, lo que indica que una parte considerable de la variación total en los datos se explica por la variación dentro de los clusters en lugar de entre ellos. Esto sugiere que los *clusters* pueden estar solapados, lo que podría reflejar similitudes entre los grupos.

Figura 2. *Composición de variables de cada cluters.*

Centroides:

Cluster 0: Es totalmente de sexo femenino, con niveles de albuminuria positiva, hiponatremia, con mayor porcentaje de hipertensons.

Cluster 1: Mayor edad media, con albuminuria positiva, hiponatremia leve, pero con niveles de potasio elevado (5.9). En este *cluster* predomina la hipertensión, diabetes mellitus, y otros singos clínicos como edema de MMII y anemia. Los integrantes de este *cluster* también son predominantemente mujeres.

Cluster 2: Este grupo tiene la edad promedio más alta (64.3 años), pero no presenta albuminuria. Se evidencia normonatremia y normokalemia, con una prevalencia de hipertensión y diabetes mellitus, pero sin otros síntomas significativos. Este *cluster* también presenta un predominio de sexo femenino.

Cluster 3: Este es el grupo más joven (46.3 años), sin albuminuria, sin hipertensión, o diabetes, y todos los integrantes son de sexo masculino.

El *C0* agrupa sujetos con hipertensión, pero sin otras comorbilidades importantes, y con apetito normal. Este grupo tiene una edad ligeramente mayor y está integrado por mujeres. El C1 combina varios factores de riesgo, como hipertensión, diabetes, y anemia, con mayores niveles de potasio. El C2 muestra individuos mayores, con hipertensión y diabetes. El C3 incluye sujetos más jóvenes, sin comorbilidades relevantes, y todos son hombres. Este grupo parece ser el más saludable en comparación con los otros.

La distancia entre los *clusters* indica qué tanta separación hay entre ellos: El C0 está más cerca del C1 y el C2. El C3, por su parte, está más alejado de los otros *clusters*, reforzando que este grupo es diferente en términos de edad y características clínicas.

Batch centroid

Tras obtener los resultados de los *clusters*, se utilizó la opción *batch centroid*, proporcionada por BigML, la cual permite añadir los *clusters* como una nueva variable al dataset y permite analizar los datos de cada cluster.

Figura 3. *Filtrado glomerular según la edad, de acuerdo a cada cluster.*

Con el fin de comprender cómo se encontraban relacionados los *clusters* con las variables renales, se evaluó el filtrado glomerular para cada *cluster* **(Figura 3).**

El C3, que es el grupo más jóven, tiene el filtrado glomerular más alto con una media de 82.35 ml/min/1.73m². El C0 y C1, son mayores, con peor función renal, con TFGe de 22.57 y 8.98 ml/min/1.73m², respectivamente. El C2, presenta una TFGe de 28.74 ml/min/1.73m². Las diferencias encontradas son estadísticamente significativas (p=5.19e-58) **(Figura 4).**

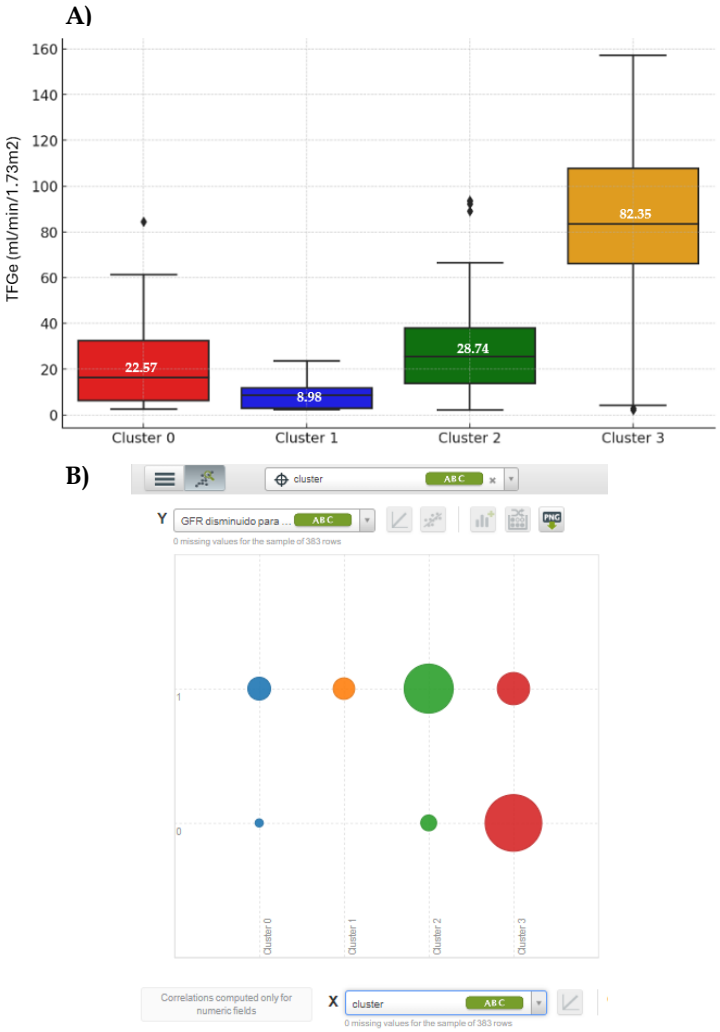


Figura 4. ***A)*** *Distribución de filtrado glomerular por cluster.* ***B)*** *Distribución del outcome filtrado glomerular bajo para la edad por cluster.*

Modelos de aprendizaje supervisado

Durante el análisis, se entrenaron un total de 5 modelos teniendo como outcome el filtrado glomerular bajo para la edad. La **tabla 3** muestra los resultados de los rendimientos de las evaluaciones de cada modelo.

En general, los modelos demostraron un buen rendimiento.

Los modelos de *Decision Tree*, *Random Forest*, y *Boosted Trees* presentan resultados bastante similares en con respecto a precisión, recall, F1-Score, aunque con un AUC ligeramente diferente.

Tanto la regresión logística como el deepnet destacan, mostrando los mejores rendimientos, con un recall alto (97.3%) y un AUC también notable (0.90), siendo, por lo tanto, las opciones más sólida. El **Anexo 6** muestra los resultados de los modelos supervisados.

	Accuracy	Recall	F1-Score	Precisión	AUC
Regresión logística	88.31	97.3	0.88	81.8	0.90
Decision tree	84.42	89.19	0.846	80.49	0.90
Random Forest	84.42	89.19	0.846	80.49	0.888
Boosted trees	84.42	89.19	0.846	80.49	0.893
Deepnet	87.01	97.3	0.88	80	0.91

Cross Validation

Con el fin de determinar si el modelo de deepnet presentaba *overfitting*, se realizó el proceso de *crossvalidation* en BigML Para ello se dividió el dataset de test en 5 partes y posteriormente se calculó la media de cada parámetro evaluado. El **Anexo 7** muestra los resultados del promedio de las evaluaciones del K-fold, y los resultados de la evaluación inicial.

De acuerdo a los resultados, los promedios del K-fold y del score F-1 son ligeramente más bajos que la evaluación inicial, sugiriendo que el modelo puede tener ligeras variaciones dependiendo del *fold*. La precisión, se mantiene de forma constante, indicando que el modelo sigue siendo confiable para detectar los verdaderos positivos. El *Recall*, por su parte, es ligeramente inferior en el K-fold, lo que puede inidcar que el modelo tiene un poco más de dificultad para captar todos los casos positivos en algunas particiones. El AUC, es alto en ambos casos, pero ligeramente inferior en el K-fold (0.896 vs 0.917).

La consistencia entre las métricas sugiere que no hay *overfitting*. Y las diferencias identificadas son relativamente pequeñas y esperables debido a la variabilidad natural entre las diferentes particiones de los datos.

Predicciones:

Tras entrenar los modelos, se decidió mostrar ejemplos de predicciones de dos pacientes de la vida real. Dichos pacientes fueron valorados en consulta externa de Nefrología en Guadalajara, México.

Ejemplo 1

Se trata de un paciente de sexo masculino, de 42 años sin antecedentes de HTA o diabetes o enfermedad coronaria. El paciente acude con edema de miembros inferiores. Su analítica mostró Hb: 15.3 g/dl, leucocitos de 5.300, con una albuminuria de de 3.21 g/día. Se puede observar en la **figura 5A** que el algoritmo clasifica al paciente con 60% de probabilidad de tener un filtrado glomerular adecuado para su edad. Esta clasificación es adecuada, ya que el filtrado de esta paciente es adecuado para su edad (TFGe real: 109 ml/min/1.73m^2).

Ejemplo 2

Mujer de 65 años con antecedentes de hipertensión arterial, sin diabetes o enfermedad coronaria. Al examen físico presenta edema de miembros inferiores, y sus laboratorios muestran Hb 10.1 g/dl, y leucocitos normales. En la **figura 5B** se observa que el modelo clasifica a la paciente en el grupo TFGe baja para la edad, con una probabilidad de 99.14%. (TFGe real: 24 ml/min/1.73m^2).

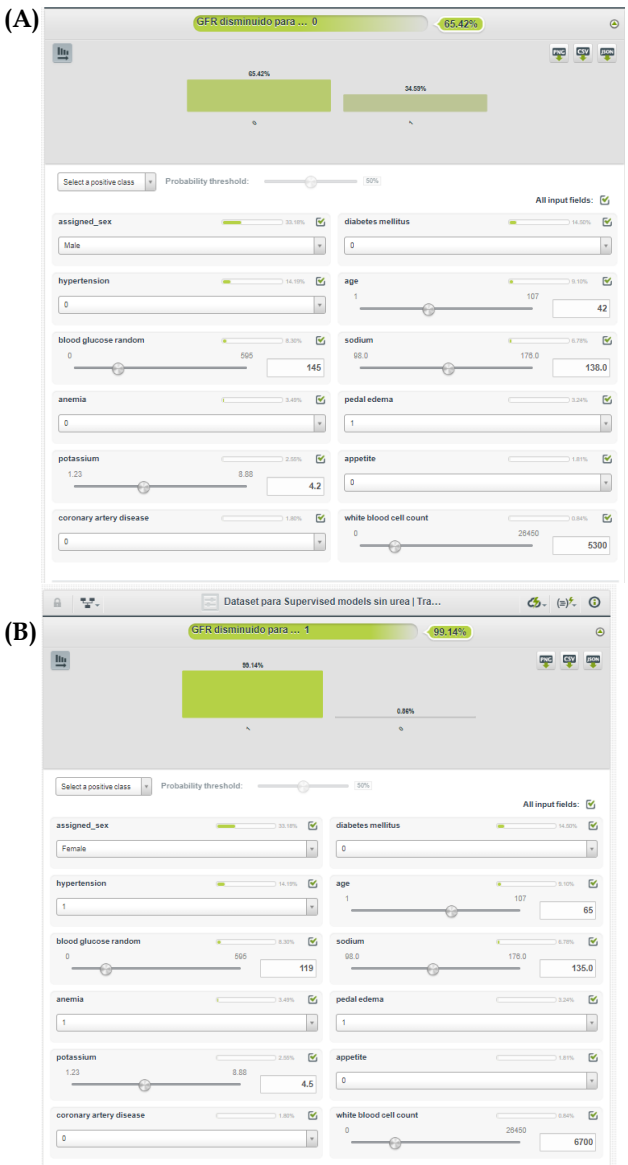


Figura 5. *Predicciones. A) Ejemplo 1. B) Ejemplo 2.*

5 DISCUSIÓN

En este estudio se ha evaluado la utilidad de una base de datos de uso público para entrenar modelos *de machine learning* que permitan clasificar a un grupo de individuos según la presencia de deterioro de la función renal.

En un primer paso se evaluó la calidad de la base de datos encontrando como desventajas la ausencia de la variable sexo y de la variable filtrado glomerular. Dicha limitación conllevó a modificar el protocolo y crear una base de datos semisintética, que incluyera valores de sexo y filtrado glomerular como estimador de la función renal. Para ello, en un primer paso se calculó el filtrado glomerular suponiendo ambos sexos, y se asignó el sexo de la etiqueta de ERC correspondiente. El grupo resultante, sirvió como base para la creación de un árbol de decisión que permitió predecir la variable sexo en los individuos restantes.

Diferentes estudios han reportado que las mujeres tienen menor riesgo de progresión a ERCT que los hombres, incluso tras ajustar por diferentes factores. Se sugiere que podría estar relacionado con factores biológicos, como los estrógenos representando un factor protector, el acceso a cuidados de la salud, entre otros, lo que hace que sea de vital importancia la inclusión de la variable sexo para un análisis más acertado en el campo de la ERC. Adicionalmente, es necesario para el cálculo del filtrado glomerular estimado (12).

Al predecir el sexo a partir de la variable resultado (ERC), surgió la limitación de condicionar los modelos de entrenamiento a presentar *overfitting*. Por lo tanto, en lugar de utilizar la variable filtrado glomerular, se tomó la decisión de utilizar el *outcome* filtrado glomerular bajo para la edad.

El concepto de filtrado glomerular ajustado para la edad está incrementando su aceptación en el campo de la nefrología, basado en la premisa de que la función renal disminuye fisiológicamente con la edad, no siendo adecuado utilizar el mismo punto de corte para todas las edades (13). Este concepto fue validado en un estudio poblacional realizado en Alberta, Canadá, en el que se reportó que aquellos individuos mayores de 65 años con una TFGe entre 45 y 60 ml/min/1.73m^2, no presentaban diferencias en mortalidad y efectos adversos comparados con aquellos sin ERC (5). Por lo tanto, se decidió utilizar esta variable como resultado.

Tras constituir la base de datos semisintética, se realizó un estudio de *clusters* que permitió agrupar los datos según sus características, permitiendo obtener grupos separados entre si con sentido clínico. El análisis arrojó 4 *clusters*, que presentaban diferencias en edad, y presencia de variables como hipertensión y diabetes. Al analizar los centroides en conjunto con la base de datos original, se pudo observar cómo ciertas características clínicas se asociaban con la presencia de un filltrado glomerular normal o con un filtrado glomerular bajo. Pese a que los resultados fueron llamativos, el análisis no agrupó adecuadamente a aquellos con

filtrados glomerulares intermedios, lo que sugiere que sería útil añadir nuevas variables que hacen parte del riesgo de aparición de ERC. Adicionalmente, no se puede descartar de que parte de la agrupación relacionada con el sexo, haya estado relacionado con la generación de la variable sexo estimado por el modelo entrenado.

Los modelos supervisados permitieron establecer el problema de clasificación de acuerdo con el *outcome* filtrado glomerular bajo para la edad. Los resultados de la evaluación mostraron un buen rendimiento, siendo la red neuronal, el que ofrece mejor rendimiento general con un balance sólido entre *precisión, recall y Score F1*, aunque la regresión logística también mostró un buen rendimiento. El análisis posterior de *crossvalidation* permitió descartar que los resultados se debieran a *overfitting* en el modelo.

Otros estudios han evaluado el uso de modelos de *machine learning* para predecir el deterioro de función renal en pacientes con ERC. (**Anexo 8)** (14). De igual forma, diferentes autores han analizado el dataset utilizado en este estudio para entrenar modelos de ML enfocados en la predicción de ERC. Tekale et al. (15) empleó SVM y *Decision trees* para predecir el diagnóstico de ERC, consiguiendo una precisión de 91.75% y 96.75%, respectivamente. Salekin y Stankovic(8), por su parte, aplicaron *K-Nearest Neighbours, Random Forest* y redes neuronales, alcanzando una precisión del 99.3% en F1-score con Random Forest, junto con una significativa reducción del error cuadrático medio (RMSE). Por otra parte, Singh (21) implementó una serie de algoritmos, entre ellos *Decision Tree, SVM, KNN, Naive Bayes* y regresión logística, observando una precisión del 100% con Decision Tree y Random Forest. Similar a este trabajo, todos los autores resaltaron la necesidad de un número mayor de individuos, añadiendo como limitación, el número de datos faltantes, y errores, que podían afectar la confiabilidad de los resultados.

A diferencia de los estudios descritos, que en su mayoría utilizan el dataset sin modificaciones significativas y se enfocan en predicciones binarias de presencia o ausencia de ERC, en este estudio se abordó la ausencia de variables claves como el sexo y la TFGe. Adicionalmente, este es el primer estudio que utiliza modelos de ML para predecir el outcome TFGe baja para la edad, proporcionando una perspectiva distinta a la de los estudios previos, y destacando la importancia de la selección adecuada de parámetros clínicos de acuerdo a la enfermedad evaluada.

Este trabajo sugiere que los modelos de *machine learning* podrían ser adecuados para clasificar pacientes en riesgo de presentar deterioro de la función renal y abre la puerta para proponer futuros estudios colaborativos, que permitan aumentar el número de pacientes y añadir múltiples factores de riesgo conocidos de ERC, y biomarcadores novedosos como estudios genéticos y estudios de imágen. Sería de interés comprender cómo la suma de estas variables podría interactuar con la aparición del daño renal que pueda tener implicaciones terapéuticas para ofrecer un enfoque más personalizado a los pacientes.

Este trabajo también presenta limitaciones significativas. El hecho de ser una base de datos de acceso público, con poca información sobre la recogida de los datos, puede conllevar a errores en la interpretación de los datos. La ausencia de la variable sexo, puede que no permita tener en cuenta las diferencias biológicas de la población, por lo que es necesario recogerla a la hora de la creación de un dataset. Sin embargo, la predicción de la variable sexo sobre el dataset, puede facilitar el *overfitting*, también nublando los resultados de los modelos, principalmente debido a la presencia de la variable nitrógeno ureico. La ausencia de datos del dataset original, y la necesidad de imputación de algunas variables, también representa una limitación.

Por último, aunque los rendimientos de los modelos hayan sido altos, en un siguiente paso se entrenaran múltiples modelos con la inclusión y exclusión de variables que reduzcan el potencial *overfitting*.

6 CONCLUSIÓN

El uso de herramientas de *machine learning*, representa una forma innovadora de diagnosticar a pacientes en riesgo de ERC, y de predecir la progresión de la enfermedad. Sin embargo, la elaboración de bases de datos que contengan variables relevantes para el diagnóstico es primordial a la hora de entrenar dichos modelos, ya que puede conllevar a errores de interpretación. En este estudio también se muestra la importancia de la correlación clínica a la hora de realizar estudios de *machine learning* en el campo de la salud y sienta las bases para la realización de estudios colaborativos que faciliten la inclusión de un mayor número de factores de riesgos y biomarcadores, así como un mayor número de sujetos en el estudio.

En estudios futuros, seleccionaremos diferentes variables para entrenar modelos que nos permitan reducir el overfitting, y analizaremos la utilidad de herramientas de machine learning para enfermedades renales específicas como lo es la poliquistosis hepatorrenal autosómica dominate.

7 AGRADECIMIENTOS

Agradezco inmensamente a los profesores del máster, en especial a Remo y Lola por su apoyo invaluable. También agradezco a mis compañeras Leti, Laura, Matías y Ariel, por todo su apoyo y colaboración durante todo el proceso.

8 REFERENCIAS

1. Francis A, Harhay MN, Ong ACM, Tummalapalli SL, Ortiz A, Fogo AB, et al. Chronic kidney disease and the global public health agenda: an international consensus. Nat Rev Nephrol. 2024;20(7):473-85.
2. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guidelines for the evaluation and managment of chronic kidney disease. . Kidney Int Suppl.2013;3:1–150.
3. Perez-Gomez MV, Bartsch LA, Castillo-Rodriguez E, Fernandez-Prado R, Fernandez-Fernandez B, Martin-Cleary C, et al. Clarifying the concept of chronic kidney disease for non-nephrologists. Clin Kidney J. 2019;12(2):258-61.
4. Glassock R, Delanaye P, El Nahas M. An Age-Calibrated Classification of Chronic Kidney Disease. JAMA. 2015;314(6):559-60.
5. Liu P, Quinn RR, Lam NN, Elliott MJ, Xu Y, James MT, et al. Accounting for Age in the Definition of Chronic Kidney Disease. JAMA Intern Med. 2021;181(10):1359-66.
6. Sanmarchi F, Fanconi C, Golinelli D, Gori D, Hernandez-Boussard T, Capodici A. Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. J Nephrol. 2023;36(4):1101-17.
7. Cao X, Lin Y, Yang B, Li Y, Zhou J. Comparison Between Statistical Model and Machine Learning Methods for Predicting the Risk of Renal Function Decline Using Routine Clinical Data in Health Screening. Risk Manag Healthc Policy. 2022;15:817-26.
8. Salekin A, Stankovic J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In: Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016, pp. 262–270, 2016.
9. Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Adv Computer. 2019;10(8):89–96.
10. Delrue C, De Bruyne S, Speeckaert MM. Application of Machine Learning in Chronic Kidney Disease: Current Status and Future Prospects. Biomedicines. 2024;12(3).
11. Chan L, Vaid A, Nadkarni GN. Applications of machine learning methods in kidney disease: hope or hype? Curr Opin Nephrol Hypertens. 2020;29(3):319-26.
12. Kattah AG, Garovic VD. Understanding sex differences in progression and prognosis of chronic kidney disease. Ann Transl Med. 2020;8(14):897.
13. Delanaye P, Glassock RJ, Pottel H, Rule AD. An Age-Calibrated Definition of Chronic Kidney Disease: Rationale and Benefits. Clin Biochem Rev. 2016;37(1):17-26.
14. Debal, D.A., Sitote, T.M. Chronic kidney disease prediction using machine learning techniques. J Big Data 9, 109 (2022).
15. Tekale S, Shingavi P, Wandhekar S, Chatorikar A. Prediction of chronic kidney disease using machine learning algorithm. Disease. 2018;7(10):92–6.
16. Su Y, Yuan D, Chen DG, Ng RH, Wang K, Choi J, et al. Multiple early factors anticipate post-acute COVID-19 sequelae. Cell. 2022;185(5):881-95.e20.
17. Yashfi SY. Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms. 2020.
18. Rady EA, Anwar AS. Informatics in Medicine Unlocked Prediction of kidney disease stages using data mining algorithms. Informatics Med. 2019;15(2018):100178.
19. Alsuhibany SA, et al. Ensemble of deep learning based clinical decision support system for chronic kidney disease diagnosis in medical internet of things environment. Comput Intell Neurosci. 2021;3:2021.
20. Poonia RC, et al. Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease. Healthcare. 2022;10:2.
21. Singh HV, et al. Chronic Kidney Disease Prediction Using Different Algorithms. Int J Sci Res Comput Sci Eng Inf Technol. 2020;6(5):6-13. doi:10.32628/CSEIT20652.

ANEXOS

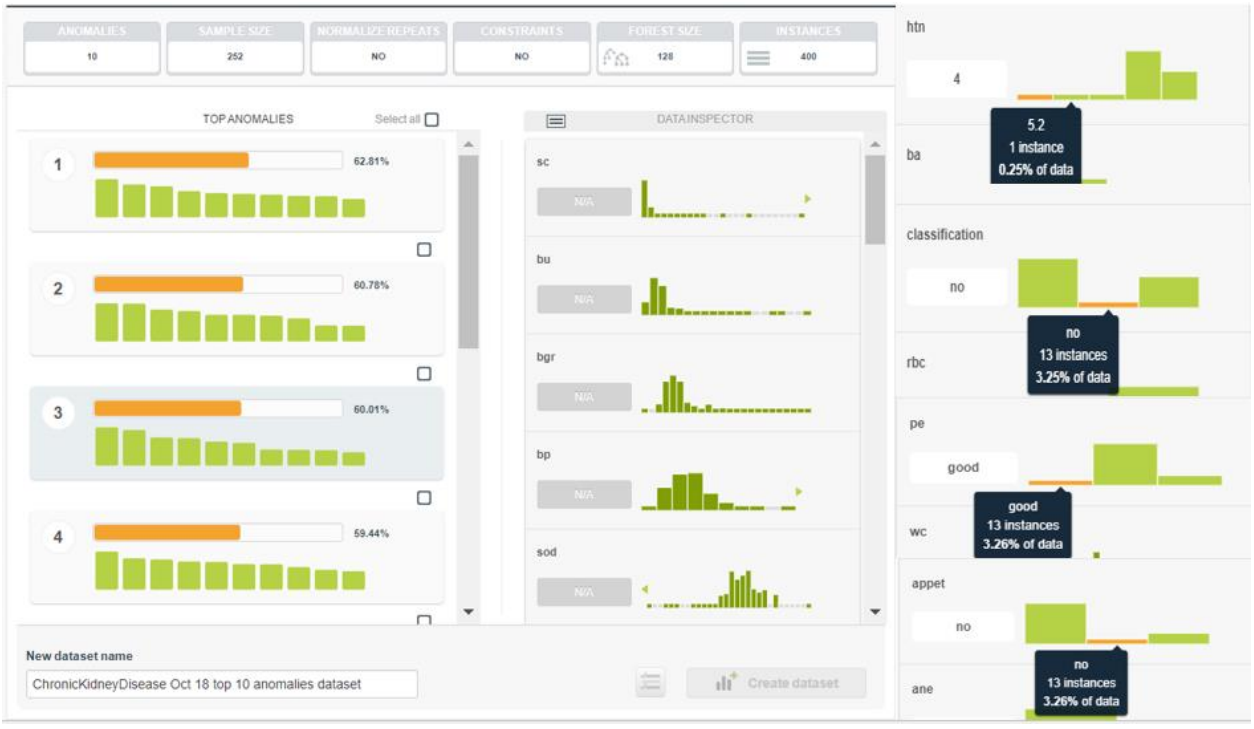
Anexo 1. Definición de las variables del dataset inicial, número de valores faltantes y ejemplo de valores atípicos.

Nombre de la variable	Tipo	Descripcion	Unidades	Valores faltantes	Porcentaje	Ej. Outliers/ valores atipicos
age	Numérica	Edad	year	9	2,25	2, 3, 5, 6, y 8 años
bp	Numerica	Presion arterial	mm/Hg	12	3	50, 100, 110 y 180.
sg	Categorico	Specific Gravity		47	11,75	1,02, 1,01
al	Categorico	Albumina		46	11,5	
su	Categorico	Azucar		49	12,25	
rbc	Categórcia Binaria	Globulos rojos		152	38	
pc	Categórcia Binaria	Celulas de pus		28	16,25	
pcc	Categórcia Binaria	pus cell clumps		4	1	
ba	Categórcia Binaria	Bacteria		44	1	
bgr	Numérica	Glucosa al azar	mgs/dl	44	11	
bu	Numérica	Urea	mgs/dl	19	4,75	
sc	Numérica	Creatinina	mgs/dl	17	4,25	
sod	Numérica	Sodio	mEq/L	87	21,75	
pot	Numérica	Potasio	mEq/L	88	22	39, 47
hemo	Numérica	Hemoglobina	gms	52	13	
pcv	Numérica	packed cell volume		72	18	
wbcc	Numérica	Globulos blancos	cells/cmm	108	26,75	?, 43
rbcc	Numérica	Globulos rojos	millions/cmm	131	32,5	
htn	Categórcia Binaria	Hipertension		5	1	"4", "8", "5.2" y "?"
dm	Categórcia Binaria	Diabetes mellitus		7	1,75	
cad	Categórcia Binaria	Enfermedad coronaria		4	1	
appet	Categórcia Binaria	Apetito		1	0,25	"no"
pe	Categórcia Binaria	Edema de MMII		1	0,25	Good
ane	Categórcia Binaria	Anemia		1	0,25	
class	Categórica Binaria	Enfermedad renal cronica		0	0	no

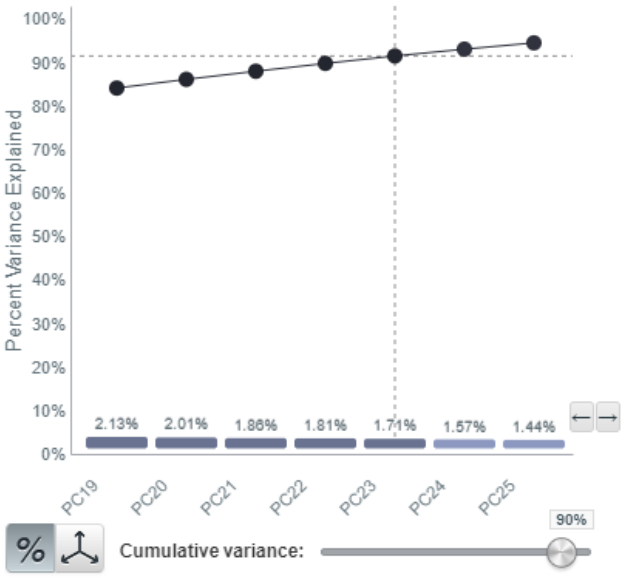
Anexo 2. Definición de métricas para evaluar rendimiento de los modelos entrenados.

	Definición	Cálculo
Recall (sensibilidad)	Mide la capacidad del modelo para identificar correctamente a las instancias positivas (verdaderos positivo). Proporción de verdaderos positivos sobre el total de instancias pertenecientes a la clase positiva.	Verdaderos positivos/Verdaderos positivos + Falsos negativos
Precisión	Mide la exactitud de las predicciones positivas del modelo, es decir, la proporción de instancias correctamente clasificadas como positivas sobre el total de predicciones positivas (verdaderos positivos + falsos positivos)	Verdaderos positivos/Verdaderos positivos+Falsos positivos
Score F1	Es la media armónica entre la precisión y el recall. Es útil cuando exite desbalance entre clases y se busca un equilibro entre precisión y recall.	2x (Precisión x Recall)/Precisión + Recall
Accuracy (Exactitud)	Proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de predicciones del modelo.	Verdaderos positivos + verdaderos negativos/Total de instancias

Anexo 3. Anomaly Detector



Anexo 4. *Principal component Analysis*



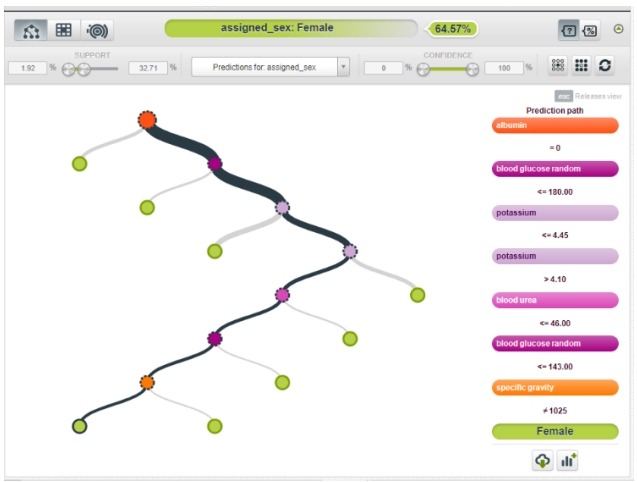
PREDICCIÓN DE LA VARIABLE SEXO

Para predecir la variable sexo se realizaron los siguientes pasos:

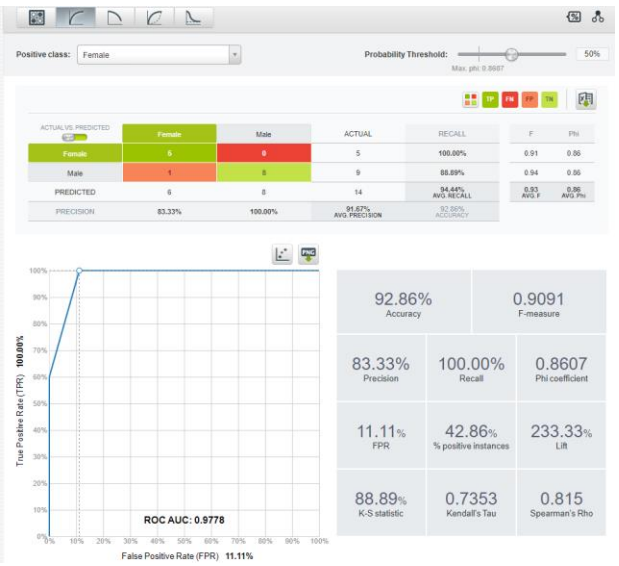
1. Cálculo de la TFGe asumiendo ambos sexos: Se utilizó la fórmula CKD-EPI para calcular la TFGe con base a la edad y a la creatinina sérica asumiendo sexo femenino y otro asumiendo sexo masculino.
2. Se crearon dos columnas para clasificar TFGe < (1) o > (0) de 60 ml/min/1.73m² para ambos sexos.
3. Se generó una columna donde se identificó si había diferencia en la clasificación de TFGe entre el sexo femenino y masculino. Si éste era diferente, se asignó un valor de 1, si eran iguales, se asignó el valor de 0.
4. Identificación de sexo según filtrado glomerular: Par los casos donde la diferencia de TFGe era 1, se revisó la columna clase (ERC), y si la clase es igual a 1, se asignó el sexo, cuya TFGe concordara con la asignación de ERC. De esta manera identificando qué sexo sería más probable en función de los resultados y la clasificación del paciente.
5. Se identificaron 66 individuos a quienes se les pudo asignar el sexo con esta estrategia. Se creó un dataset con estos individuos, y otros con aquellos en los que el sexo seguía como interrogante.
6. En BigML se subió el dataset con sexo asignado, sin incluir la variable clase y se entrenó un árbol de decisión que permitiera predecir el sexo. Realizamos un Split training:test de 80:20%, consiguiendo como rendimientos una Accuracy de 92,86% para ambos sexos, con recall de 88.89% y 100% para el sexo masculino, y femenino, respectivamente. **(Anexo 5).**
7. Predicción de sexo: Finalmente, se utilizó el modelo entrenado, para predecir la variable sexo del dataset con sexo desconocido.
8. Filtrado glomerular: En el dataset resultante, se asignó el nuevo parámetro de TFGe, según el sexo estimado, y se creó la variable TFGe baja para la edad.

Anexo 5. *Decision tree* para predicción de variable sexo.

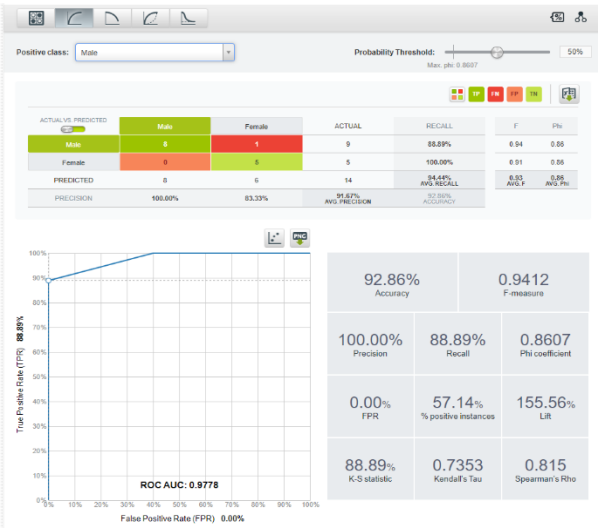
(A)



(B)

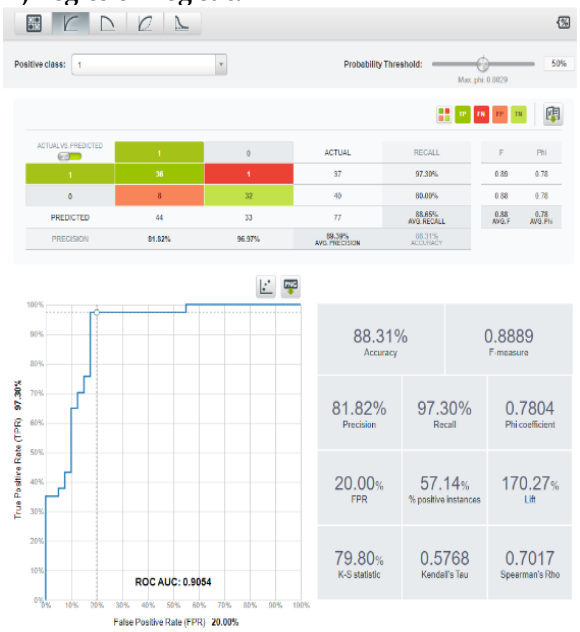


(C)

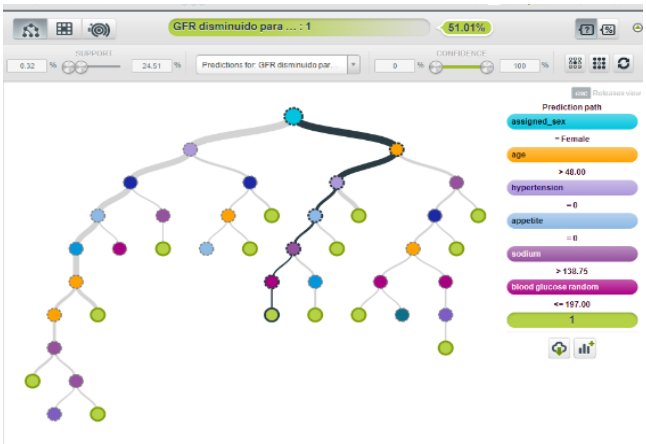


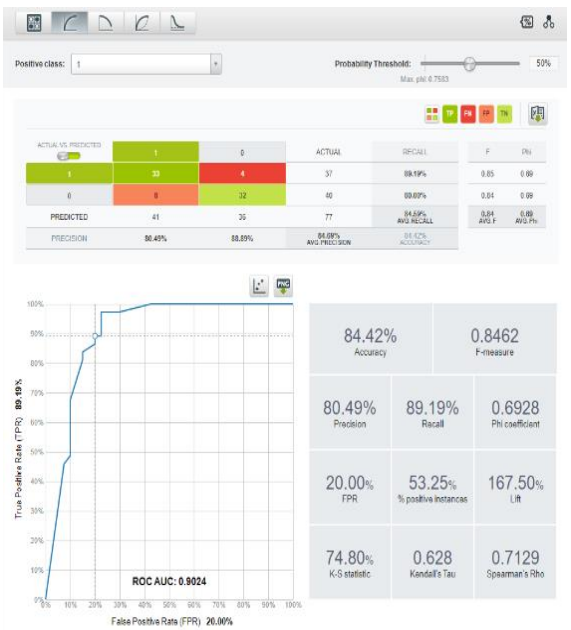
Anexo 6. *Modelos supervisados*

A) Regresión Logística

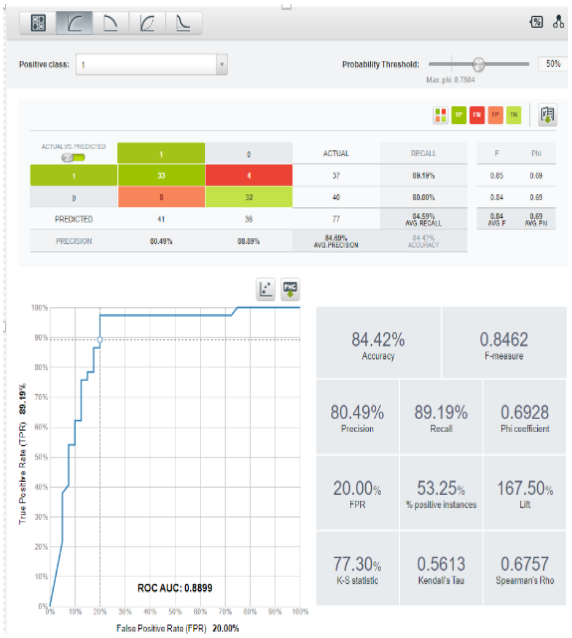


B) Árbol de decisión

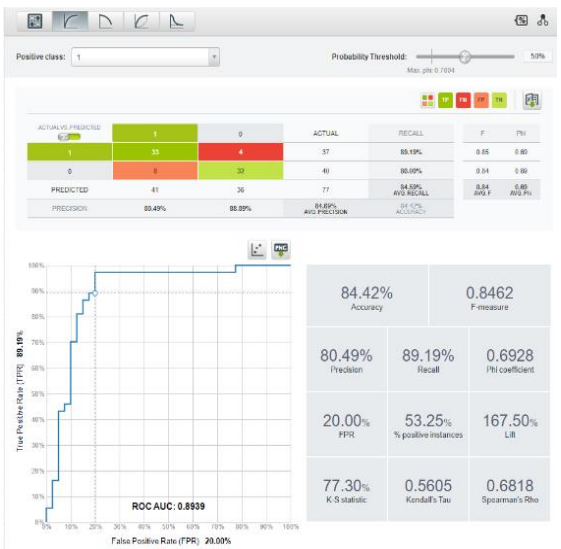




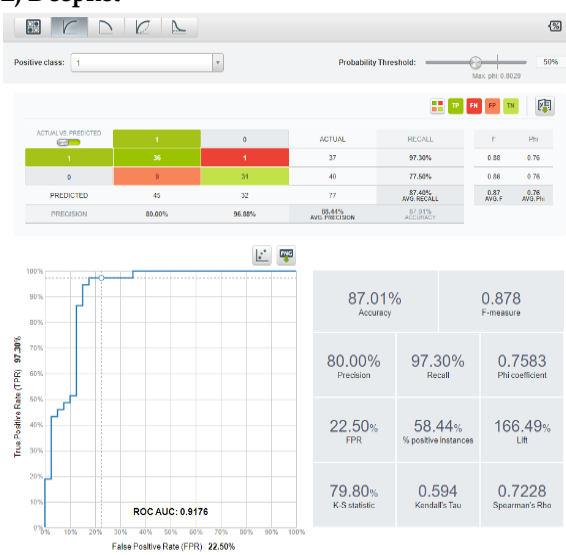
C) Random Forest



D) Boosted trees



E) Deepnet



Anexo 7. *5-fold crossvalidation*

Métrica	Evaluación Inicial (Deepnet)	K-fold Promedio
Accuracy	87.01%	84.00%
F-measure	80.00%	75.23%
Precision	97.30%	97.14%
Recall	0.7583	0.7148
Phi coefficient	22.50%	26.39%
FPR	58.44%	53.33%
% positive instances	166.49%	171.71%
Lift	79.80%	80.00%
K-S statistic	0.594	0.5738
Kendall's Tau	0.7228	0.6797
AUC	0.9176	0.8956

Anexo 8. Estudios evaluando el uso de *machine learning* para el diagnóstico de ERC. Adaptado de Debal, et al. (14)

No.	Técnica aplicada	Resultado reportado	Limitaciones	Autor
1	K-NN, RF, y NN, enfoque Wrapper y Enfoque Embebido	Score F1 de RF: 99.8	Tamaño de dataset pequeño con valores faltantes; no se incluyó predicción de nivel de severidad	Salekin y Stankovic (8)
2	DT y SVM	Precisión de DT: 91.75% y SVM: 96.75%	Necesita un aumento en el tamaño del dataset; no se incluyó predicción de severidad. Solo se compararon resultados de dos clasificadores	Tekale et al. (15)
3	NB, KNN, SVM, DT y ANN.	Precisión de NB: 94.6%	Dataset pequeño; no se realizó predicción de etapas; la precisión de la clasificación necesita mejora	Priyanka et al. (16)
4	RF y ANN	Precisión de RF: 97.12% y de ANN: 94.5%	Dataset pequeño y sin predicción de etapas	Yashfi (17)
5	PNN, MLP, SVM, RBF	Precisión de PNN: 96.7%, MLP: 60.7%, SVM: 87%, RBF: 51.5%	Dataset pequeño; se usaron algoritmos no adecuados para conjuntos de datos pequeños	Rady y Anwar (18)
6	Técnica EDL-CDSS, detección de outliers con ADASYN y ajuste de hiperparámetros con QOBOA	Sensibilidad de EDL-CDSS: 0.9680 y especificidad: 0.9702 comparado con ACO, FNC, KELM, CNN-GRU, DBN, DT, MLP y D-ACO	Tamaño de dataset de referencia pequeño; integración de IoT y datos de referencia no está clara; el estudio solo se centra en clasificación binaria (ERC o no-ERC)	Alsuhibany et al. (19)
7	KNN, ANN, SVM, NB y LR con Chi-Cuadrado y RFE	Precisión de KNN: 66.25%, ANN: 65%, SVM: 97.5%, NB: 95%, LR: 97.5%	Dataset pequeño y sin predicción de etapas; la precisión necesita mejora	Poonia et al. (20)

Abreviaciones: K-NN: K-Nearest Neighbors, RF: Random Forest, NN: Neural Network, DT: Decision Tree, SVM: Support Vector Machine, NB: Naive Bayes, PNN: Probabilistic Neural Network, MLP: Multi-layer Perceptron, RBF: Radial Basis Function, ADASYN: Adaptive Synthetic Sampling.