CLASIFICACIÓN DEL ESTADIO DE PACIENTES DIAGNÓSTICADOS CON ENFERMEDAD RENAL CRÓNICA EN UN RÉGIMEN DE SALUD EXCEPTUADO EN COLOMBIA MEDIANTE ALGORITMOS DE APRENDIZAJE SUPERVISADO

Anlly Casas Rendon
Dulzura Rodríguez Caviedes

Director: Ricardo Borda

1. RESUMEN

La Enfermedad Renal Crónica (ERC) se caracteriza por la pérdida gradual de la función renal, siendo una condición médica sin cura conocida hasta la fecha. El enfoque para el manejo de esta enfermedad se dirige a frenar su progresión y mitigar los síntomas asociados. Actualmente se están desarrollando diversas soluciones basadas en modelos de aprendizaje supervisado con el objetivo de diagnosticar y detectar anomalías y variables que faciliten el control del deterioro y progresión de la enfermedad. En este artículo, se presentan varios algoritmos de aprendizaje supervisado tradicionales y modernos, incluidos la regresión logística multinomial, el Ramdon Forest, el XGBoost, CatBoost, Corn-Ordinal-NeuralNet, las redes neuronales Feedforward y las máquinas soporte vectorial (SVM). Utilizando un conjunto final de 14 variables y 3.189 observaciones, se realiza una clasificación predictiva con el fin de identificar el nivel de progresión de la enfermedad. El modelo con mejor rendimiento resulta ser XGBoost, el cual se utiliza en la fase de implementación con observaciones no utilizadas en las fases de entrenamiento y prueba para la generación de métricas. El propósito de este modelo es realizar clasificaciones de forma predictiva acerca de los estadios de la enfermedad en que se encontraran los pacientes diagnosticados con la condición con el fin de servir como herramienta para proyectar los recursos para el tratamiento médico y brindar servicios de salud dentro de los parámetros establecidos en los modelos de atención.

ABSTRACT

Chronic Kidney Disease (CKD) is characterized by the gradual loss of kidney function, being a medical condition with no known cure to date. The approach to the management of this disease is aimed at slowing its progression and mitigating the associated symptoms. Various solutions are currently being developed based on supervised learning models with the aim of diagnosing and detecting anomalies and variables that facilitate the control of the deterioration and progression of the disease. In this article, several traditional and

modern supervised learning algorithms are presented, including multinomial logistic regression, Ramdon Forest, XGBoost, CatBoost, Corn-Ordinal-NeuralNet, Feedforward neural networks, and support vector machines (SVM). Using a final set of 14 variables and 3.189 observations, a predictive classification is performed in order to identify the level of disease progression. The model with the best performance turns out to be XGBoost, which is used in the implementation phase with observations not used in the training and testing phases for the generation of metrics. The purpose of this model is to make predictive classifications about the stages of the disease in which patients diagnosed with the condition will be found in order to serve as a tool to project the resources for medical treatment and provide health services within the parameters established in the care models

2. Introducción

La Enfermedad Renal Crónica (ERC) se define como la disminución progresiva de la función renal, una condición médica que actualmente no tiene cura, brindándose el manejo a los pacientes con la única finalidad retrasar la progresión de la misma y reducir los síntomas. Esta condición se clasifica en cinco estadios, siendo la etapa 1 la más leve, en esta fase, a pesar del fallo renal, los riñones mantienen la capacidad de depurar toxinas en la sangre. A medida que la deficiencia progresivamente aumenta la tasa de filtración glomerular (TFGe) alcanza valores entre 60 ml/min a 89 ml/min, que es un daño renal leve, se le denomina estadio 2. Conforme a que la capacidad de filtración (TFGe) va disminuyendo entre valores de 30 ml/min a 59 ml/min se denomina etapa 3, por lo que el riñón no desecha adecuadamente toxinas y exceso de agua en sangre. En la etapa 4, la TFGe llega a valores de 15 ml/min a 29 ml/min, reflejando una afectación renal significativa que se manifiesta en la eliminación muy deficiente de desechos. Finalmente, en el estadio 5, la TFGe es menor a 15 ml/min, siendo la etapa más severa donde los riñones carecen de funcionalidad, siendo un fallo renal total (1,2).

De acuerdo con estudios del 2022, esta condición afecta a más del 10% de la población mundial. En consecuencia, se ha convertido a lo largo de los años como una de las principales causas de mortalidad (2). En Colombia para el 2022, se reportaron 790.117 diagnosticados con la enfermedad, en donde el 33.90% de estos se clasificaron en los estadios 1 y 2, adicionalmente el 4.54% se encontraban en estadio 5 que también se denominan población con ERC terminal, es de resaltar que la mortalidad fue un 19.85% menor que el 2021, siendo estas en su mayoría individuos en estadio 3, así como reflejo una reducción de casos en 11.14 % respecto al año anterior(3).

Como se enunció anteriormente, los pacientes con enfermedad renal crónica requieren una atención constante para el manejo paliativo de su condición, que en la actualidad es considerada como una patología de alto costo. Esta atención es monitorea por el organismo técnico no gubernamental denominado Cuenta de Alto Costo (CAC), encargado de gestionar el manejo estandarizado en el país de condiciones tales como VIH/SIDA, Cáncer, Hemofilia, Artritis Reumatoide, Hepatitis C y ERC (4). En el caso específico del ERC, cuenta con un modelo de atención en salud que ha sido considerado como uno de los mejores en Latinoamérica, tanto por su nivel de cobertura casi universal como por la información detallada que se tiene sobre la población afectada (5).

Es crucial señalar que, para proporcionar servicios de salud dentro de parámetros establecidos en los modelos de atención, resulta indispensable asegurar la asignación de recursos para el tratamiento médico. El esquema varía en función del estadio en que se encuentre el paciente con ERC, y en este mismo sentido el costo del procedimiento (6). En este contexto toma relevancia para la Entidad encargada de brindar el aseguramiento en salud acceder a las herramientas que permitan estimar el estadio en que se encontrara la población diagnosticada con la enfermedad con el propósito de brindar información para la proyección de costos asociados. Por lo tanto, desde la perspectiva de la estadística aplicada y el análisis de datos, se propone explorar diferentes modelos de aprendizaje automático para la predicción del estadio de pacientes con ERC adscritos a un Subsistema de Salud de Régimen Especial en Colombia.

3. Conceptos básicos

Actualmente, existen diversos documentos que abordan la aplicación de algoritmos de aprendizaje automático para diagnosticar, detectar y pronosticar la enfermedad renal crónica, mediante técnicas clasificatorias. Estos estudios, tanto a nivel local como internacional, sirven como punto de partida para la selección de los modelos a explorar, en concordancia con los datos disponibles (7–19). A continuación, se presenta una descripción breve de cada uno de ellos:

Iniciando con los métodos más simples, se encuentra la *Regresión Logística Múltiple*, que consiste en la expansión del modelo de regresión logística simple diseñado para problemas con una variable respuesta binaria, por lo que típicamente evalúan presencia o ausencia de un evento, siendo ampliamente aplicado en el ámbito de la salud para determinar si una condición está presente o no. En

cambio, la Regresión Logística Multinomial se utiliza cuando la variable dependiente tiene más de dos categorías, en consecuencia, la ecuación para este tipo de modelos incluye múltiples funciones logísticas, una para cada categoría que está evaluando (15,20).

Seguido están las *Máquinas de soporte vectorial*, por sus siglas en inglés (SVMs), tienen como objetivo identificar el hiperplano que divida en clases los datos de entrenamiento, por lo que se tiene una frontera de decisión que maximiza la distancia entre la data que se denominan vectores de soporte. Es un algoritmo ampliamente utilizado en las investigaciones en salud y bioinformática en problemas de clasificación y regresión (8,21) .

En los modelos de ensamble inicialmente se halla el *Random Forest*, para comprenderlo es fundamental definir el concepto de árbol de decisión, que es un método de aprendizaje que se basa en observaciones y construye de forma lógica categorías sucesivas representadas gráficamente en forma de árbol. Inicia con un nodo principal que realiza la primera clasificación, y a partir de este, se generan nodos adicionales que contienen preguntas centradas en la variable objetivo. Este ciclo continúa hasta llegar a un nodo final, también conocido como hoja, que muestra la decisión final del modelo (22). Siendo el *Random Forest* la amalgama aleatoria de varios de estos árboles de decisión, cada uno de ellos se entrenan utilizando conjuntos de datos diferentes, empleando un vector de entrada que indica a cada árbol las decisiones tomadas. Al final el modelo toma la decisión final seleccionado la respuesta más frecuente entre los árboles (8,15,22).

Posteriormente el modelo *XG Boost*, considerado de ensamble al ser un boosting de distintos arboles de decisión, es decir, mejora paulatinamente la precisión del modelos menos complejos al considerar el error, factor de regularización para evitar sobre ajustes y combina esta información en términos de la función objetivo, por lo que cada nuevo árbol de decisión creado tiende a corregir el error del anteriormente creado, logrando un entrenamiento secuencial de estos, su utilidad abarca problemas de regresión y clasificación (7).

El último modelo de ensamble *CatBoos*, es un conjunto de árboles de decisión, donde su característica principal radica en la simetría de las divisiones en cada nodo del mismo nivel. El modelo se destaca especialmente en el manejo de los datos categóricos al usar la codificación One-Hot para organizar las covariables en función de la variable respuesta; al igual que el XGBoost, el CatBoost se

caracteriza por un entrenamiento que se basa en la corrección de errores del árbol de decisión anterior (23).

En cuanto a redes neurales la denominada *Corn*, es un modelo que usa conjuntos de entrenamiento condicional obteniendo predicciones de rango incondicionales utilizando la aplicación de la regla de la cadena para distribuciones de predicciones condicional. logrando consistencia en los rangos mediante un esquema de entrenamiento novedoso, siendo recientemente desarrollado (24).

Finalmente, la red neuronal *feedforward*: Se distingue por el flujo unidireccional de la información, que se propaga hacia adelante entre capas durante el entrenamiento o la inferencia sin fase de retroalimentación. Su estructura se compone de una capa de entrada, una o varias capas ocultas y una capa de salida, sin la presencia de bucles en la arquitectura. Los cálculos se realizan en los nodos, utilizando la información de las entradas y funciones de activación, en el cual se ajustan los pesos de las conexiones para corregir el error de la predicción. Lo que la hace aplicable a problemas de regresión y clasificación (25).

En cuanto a las métricas utilizadas para evaluar el rendimiento de la clasificación proporcionada por cada modelo, el análisis se basará en los parámetros de precisión, Recall y F1- Score. La precisión representa la proporción de datos verdaderamente positivos dentro del grupo que fue inicialmente clasificado como positivo, sin enfocarse en falsos negativos. Para abordar los falsos negativos, se aplica la métrica conocida como Recall, la cual calcula la proporción de datos clasificados como positivos dentro del grupo que se sabe que es positivo (26).

El F1 – Score combina los cálculos de las métricas explicadas anteriormente en un solo valor, lo que permite evaluar el desempeño del modelo en términos de falsos positivos y falsos negativos con igual nivel de importancia, dado que (ß=1). Adicionalmente cabe destacar la métrica "Accuracy" también conocida como la exactitud del modelo, esta métrica abarca aspectos de otras métricas mencionadas anteriormente como precisión y recall, ya que representa la proporción de predicciones correctas en relación con el total de predicciones realizadas por el modelo. A continuación, se muestras las fórmulas para calcular cada una de las métricas mencionadas (26,27)

$$\boldsymbol{Precision} = \frac{TP}{TP + \boldsymbol{FP}}$$

$$\boldsymbol{Recall} = \frac{TP}{TP + \boldsymbol{FN}}$$

$$\boldsymbol{F1-Score} = \frac{Precision * Recall}{Precision + Recall}$$

$$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + FN + TN + FN}$$

Fuente: Referencia tomada de (26)

En cuanto a los métodos de optimización, el de búsqueda por grilla consiste en probar de manera exhaustiva todos los posibles valores dentro de un conjunto de datos discretos definidos por el usuario. Se realiza el entrenamiento de un modelo con cada combinación de valores, y se evalúan los modelos generados. La combinación de hiperparámetros que logra el mayor rendimiento en el conjunto de validación se selecciona como la configuración óptima. Este método asegura encontrar el óptimo global dentro del conjunto de hiperparámetros predeterminado. Sin embargo, su principal inconveniente es el alto costo computacional requerido. Por esta razón, en la literatura se han propuesto diversas formas de optimizar este método (28).

La búsqueda aleatoria es una evolución del método de búsqueda en grilla. Esta técnica consiste en muestrear valores seleccionados de manera aleatoria dentro del dominio definido para los hiperparámetros. En lugar de establecer un conjunto específico de puntos para cada hiperparámetros, el usuario define un rango donde buscar valores para estos parámetros. El éxito de este método es su eficacia al trabajar en espacios de alta dimensionalidad. Para un número suficientemente alto de dimensiones, es poco probable que todos los hiperparámetros tengan el mismo impacto en la métrica a optimizar (28).

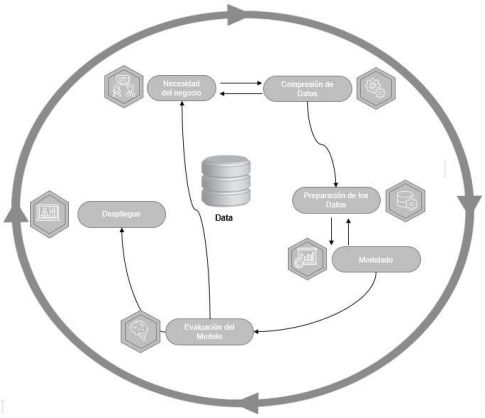
La optimización bayesiana es un método eficaz para abordar problemas complejos desde el punto de vista computacional, esta optimización se basa en modelos secuenciales de configuración adaptativa. Este enfoque se deriva del teorema de Bayes y realiza predicciones utilizando una función que es considerablemente menos costosa de evaluar. De manera automática, el método aborda el problema como una tarea de regresión, intentando estimar la distribución del rendimiento del algoritmo en función de sus hiperparámetros. Inicia con una distribución inicial que se ajusta de manera iterativa

basándose en los resultados obtenidos con cada combinación probada (28).

4. Metodología

La metodología CRISP-DM, cuyas siglas en inglés corresponden a "Cross Industry Standard Process for Data Mining", es ampliamente utilizada en proyectos de analítica de datos. Diversas investigaciones en el campo de la salud la han tomado como referencia, razón por la cual ha sido seleccionada para aplicarse en el presente proyecto (18,22,29,30) .

Un requisito fundamental que surge en los proyectos en analítica de datos es el proceso iterativo de la ejecución, evaluación y retroalimentación para obtener el mejor modelo. La metodología CRISP-DM satisface estas necesidades a lo largo de sus fases, las cuales incluyen: Comprensión del negocio, de los datos y preparación de estos, modelado, evaluación y despliegue, las cuales se abordan en detalle en el presente documento (31).



Fuente: Elaboración propia, referencia tomada de (31)

La aplicación de los modelos propuestos y el procesamiento de los datos se llevaron a cabo en la herramienta colaborativa en línea "Colaboratory", comúnmente conocida como "Colab", en conjunto con el lenguaje de programación Python, y sus respectivas librerías. Entre las más utilizadas se encuentran Pandas (para trabajar con datos tabulares), Seaborn y Matplotlib (para visualización y gráficos), junto con otras específicas según la función, las cuales se detallan a lo largo del documento (32).

5. Comprensión y preparación de la data

5.1. Descripción

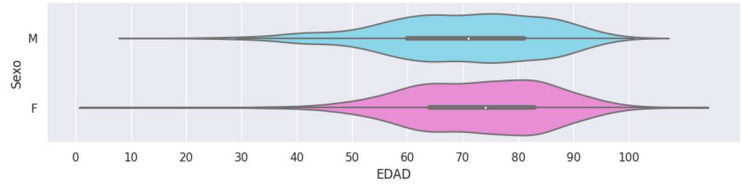
La Cuenta de Alto Costo (CAC) se encarga al interior del Sistema General de Seguridad Social en Salud del país de orientar acciones que garanticen la sostenibilidad del sistema frente a las enfermedades de VIH/SIDA, Cáncer, Hemofilia, Artritis Reumatoide, Hepatitis C y ERC, las cuales son consideradas de Alto Costo (4). Con este propósito se estableció la obligatoriedad a través de la Resolución 2463 del 2014 de enviar anualmente información sobre este tipo de afecciones (3,33).

En este contexto, desde el año 2015, la CAC diseñó un formato específico para el informe de individuos con enfermedad renal crónica (ERC) – hipertensión (HTA) – diabetes (DM). Este formulario consta de 113 preguntas dirigidas a pacientes diagnosticados con las condiciones mencionadas, los cuales están afiliados a un Subsistema de Salud (33). Por lo tanto, la base de datos a utilizar corresponde a la información contenida en el reporte de pacientes ubicados en todo el territorio colombiano, abarcando el periodo comprendido del 1 de Julio del 2022 al 30 de junio del 2023. De esta manera, se puede afirmar que la base de datos con la cual se desarrolla la presente investigación es del tipo estructurada y contiene todos los individuos objeto de interés.

5.2. Exploración

En la exploración de la data se observó que la distribución de la variable edad es asimétrica hacia la izquierda, donde la mayor concentración de pacientes se encuentra alrededor de los 75 años, con aproximadamente 800 pacientes en ese rango. Respecto a los valores extremos, se encontraron 28 individuos que están por encima de los 99 años, ubicándose en el extremo derecho del histograma, considerándose atípicos al superar considerablemente la esperanza de vida al nacer, que para Colombia es de 79,3 años y a nivel mundial de 73,3 años (34). En el extremo izquierdo, se observan 19 pacientes con edades inferiores a los 30 años, catalogándose como outlier dado que es inusual que pacientes jóvenes sean diagnosticados con esta enfermedad (Ver Anexo A – Gráfico No. 1). En cuanto a la densidad de hombres y mujeres de acuerdo al gráfico No.1, muestra que es equivalente, reflejando una distribución equitativa entre ambos sexos, de igual forma para la edad, el rango intercuartílico es similar, con una mediana de 75 años para los individuos masculinos y 70 años para los femeninos.

Gráfico No.1 – Diagrama de Violín Edad y Sexo de individuos



Fuente: Elaboración propia.

En cuanto a la variable de interés Tasa de filtración glomerular "TFGe" el grafico de dispersión muestra el efecto esperado en relación con la edad, (Ver Anexo A – Grafico No.3), ya que a medida que aumenta la edad de los individuos disminuye los valores registrados para la TFGe, lo que se relaciona directamente con la enfermedad renal crónica. Con respecto a los estadios, la mayor concentración de individuos están en el estadio 2, seguido del estadio 3 y 1, no obstante, para el estadio 4 se registran pocos casos (Ver Anexo A – Gráfico No. 4), siendo esta la condición más crítica de la enfermedad, dado que el estadio 5 se considera la fase terminal (1). Los individuos de la base de datos presentan una mediana de TFGe en los estadios del 1 al 5 en valores de 100 ml/min, 75 ml/min, 50 ml/min, 20 ml/min y 10 ml/min respectivamente, observándose una proximidad notable entre las etapas 4 y 5, es decir, de critica a terminal (Ver Anexo A – Gráfico No. 5).

5.3. Análisis de calidad

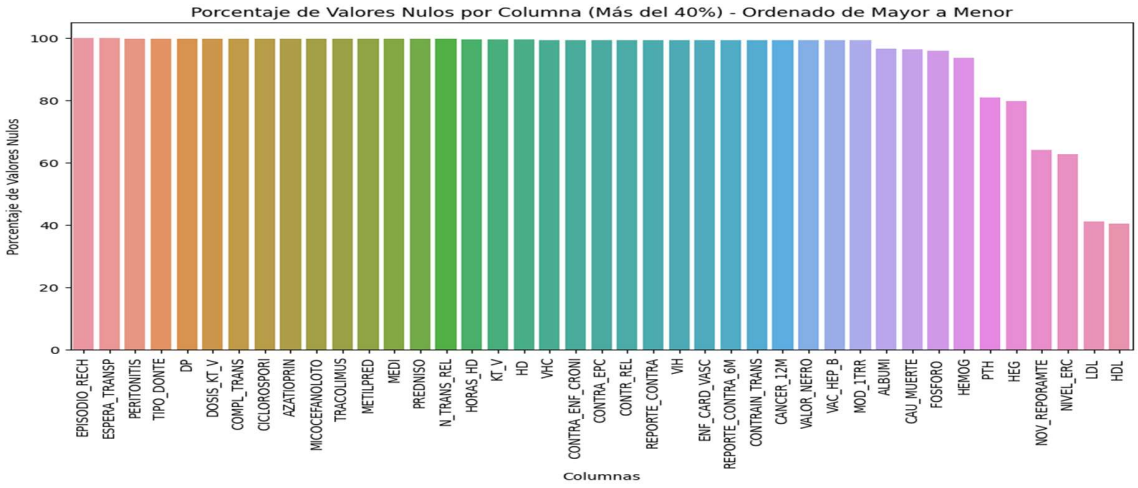
La base de datos de ERC es estructurada, estando compuesta por 4.967 registros y 114 variables, una cantidad considerada adecuada para el proceso de entrenamiento. Con el objetivo de simplificar la interpretación a nivel epidemiológico, se llevó a cabo una transformación de las variables "talla" y "peso" en una nueva variable denominada IMC (Índice de Masa Corporal), estrategia que busca reducir la dimensionalidad del conjunto de datos y facilitar la interpretación entre las mismas.

Como parte del proceso de limpieza y preparación de los datos, inicialmente se excluyen las variables denominadas talla, peso y aquellas que contienen información no relevante en términos de la variable respuesta, tales como fechas, código único de identificación, costos, códigos de EPS e IPS, al no ser información relevante desde el punto de vista clínico para el tema de interés, logrando así una reducción a 65 columnas, posteriormente se identifican que 41 de estas columnas presentan datos faltantes en más del 40% de las observaciones como se muestra en el gráfico No.2. El método "isnull" de la librería Pandas en Python calculó la proporción de valores faltantes en cada columna (35). Este porcentaje se ha establecido como referencia para eliminación de covariables,

considerando que la variable objetivo también presenta valores nulos en esta proporción, convirtiéndose en el segundo criterio de selección de datos a eliminar.

Finalmente, el repositorio se reduce a 24 variables manteniendo las 4.967 observaciones iniciales.

Gráfico No.2 – Porcentaje de valores nulos por columna mayor 40%



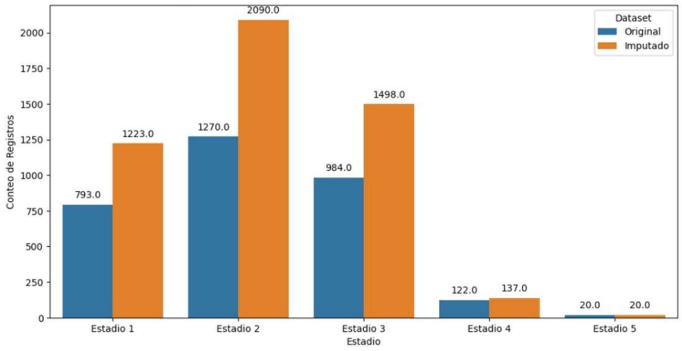
Fuente: Elaboración propia

Para abordar los datos faltantes se aplica técnicas de imputación, empleando la librería scikit-learn, con el método "IterativeImputer", cuyo diseño se basa en la estructura del paquete MICE (Multivariate Imputation by Chained Equations) de R studio (36), la función modela cada característica con el propósito de predecir los valores faltantes en las 27 columnas previamente filtradas. Este proceso se lleva a cabo de forma iterativa empleando modelos de regresión entrenados en función de otras observaciones en el conjunto de datos, en esta fase se especificó un número de 10 iteraciones mediante el parámetro "max_iter" (37).

Los autores Useche L y Mesa D destacan la utilización de tablas de contingencia y la preservación de la distribución marginal de las

variables como medidas para evaluar la técnica de imputación aplicada (38). Por lo que inicialmente se compararon los estadios 1,2,3,4 y 5 de la base de datos original con la base de datos imputada, como se muestra en la Gráfica No. 3, para la clase 1 el número de individuos aumentó de 793 a 1.223 despúes de la imputación, lo que representa un incremento del 54%. Se observó una situación similar para la clase 2 con un incremento del 64%, para la clase 3 con un incremento del 52% y para la clase 4 con un aumento del 12%, por otro lado la clase 5 se mantuvo sin variaciones.

Gráfico No.3 – Comparativo observaciones de Estadios imputados y sin imputación



Fuente: Autoras

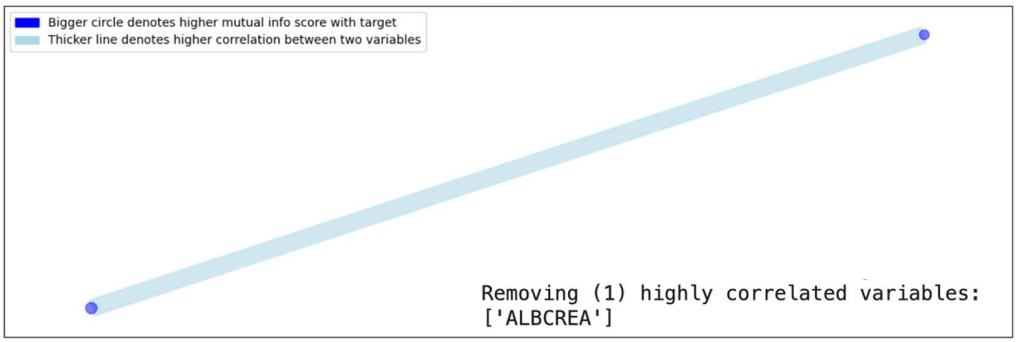
Partiendo de las observaciones en el análisis comparativo de la imputación por clases, que presenta el Gráfico 3, se procede a realizar una comparación de las métricas de los modelos de aprendizaje supervisado entrenados con la base de datos imputada y sin imputar, como muestra la Tabla No.2. Se observa que, al utilizar el marco de datos original, es posible clasificar individuos en el estadio 5 en algunos modelos, a diferencia de cuando se entrena con datos imputados. En consecuencia, de lo descrito, se ha tomado la decisión para efectos de la investigación, de trabajar con la base de datos sin imputar.

Por lo anterior, se efectúa un análisis de correlación entre las variables explicativas de la base de datos sin tratamiento de imputación, en un principio se exploró aplicando la función "corr" de Pandas (ver Anexo B – Grafico No.1), que por defecto calcula la

correlación de Pearson, sin embargo solo se identifican dos correlaciones negativas fuertes y una positiva al arrojar valores del coeficiente en el rango del (± 0.6 al ± 0.7), pero no identificaron coeficientes con valores superiores al 90% que puedan ser consideradas muy fuertes o directas (39,40). Por lo tanto, se decide realizar un enfoque más amplio que permita evaluar todos los tipos de variables previamente imputados aplicando el método de selección de variables "featurewiz".

La aplicación del paquete "featurewiz" facilita la selección de las covariables más relevantes en relación con la variable objetivo, que en el contexto del trabajo de investigación es la tasa de filtración glomerular. Durante la fase de modelado, esta variable se transformará para representar los cinco estadios de la Enfermedad Renal Crónica (1,2,3,4 y 5). En una primera instancia la función aplica el algoritmo SULOV que identifica pares de variables altamente correlacionadas y elimina aquella que tiene el puntaje de información más bajo en relación con la variable objetivo (41), como se muestra en la gráfica No. 3, el método selecciona el grupo de variables (albumina sérica y Albumina/Creatinuria), optando por remover la albumina sérica y seleccionar la relación albumina/creatinuria debido a que obtiene el puntaje más alto en términos de la tasa de filtración glomerular, que es la variable objetivo.

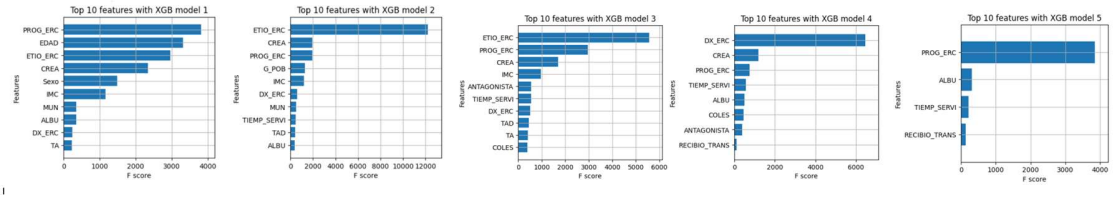
Gráfico No.3- Método SULOV



Fuente: Elaboración propia

Después de aplicar el método SULOV, aplica el método de XGBoost recursivo, el cual se encarga de seleccionar las variables restantes, dividiéndolas en conjuntos de entrenamiento y validación, con grupos de 10 características o menos. El proceso se realiza de forma iterativa hasta 5 veces, seleccionado en cada iteración aquellas variables que muestran un mejor rendimiento en función de la variable objetivo (41).

Gráfico No.4– Método XG Boost recursivo



Fuente: Elaboración propia

La función arrojo como resultado una lista de 14 características que son: Edad, , Creatinina, IMC, Sexo, Grupo Poblacional, Colesterol, Etiología ERC, Diagnóstico ERC, Tensión Arterial, tiempo de servicio, sí el usuario recibe Antagonista de los receptores de angiotensina II (ARA II), albumina/creatinuria, sí la persona se encuentra en un programa de atención de ERC (renoprotección, nefroprotección, protección renal, prediálisis) y si ha recibido o no trasplante renal, obteniendo un marco de datos final de 14 variables y 3.189 observaciones para la fase de modelado de datos. La reducción de observaciones iniciales de 4.968 a 3.189, es producto de la eliminación de los datos faltantes en la variable respuesta, los cuales suman 1.770.

6. Modelado de Datos

Con el propósito de clasificar pacientes en las cuatro etapas de la enfermedad renal crónica que han sido previamente diagnosticados y que pertenecen a un régimen de salud exceptuado del Sistema General de Seguridad Social en Colombia, se aplicaron siete modelos de aprendizaje automático supervisado los cuales se abordaron en el numeral 3 del presente documento, todos los modelos tienen en común como variable respuesta denominada "Estadio", es decir, la progresión de la enfermedad en los pacientes que se clasifica en estadios 1,2,3, 4 y 5.

Baştanlar et al, recomienda trabajar para los modelos de machine learning con un set de entrenamiento del 70%, y el 30% restante como set de prueba (42). Cabe destacar que el marco de datos que se va a emplear para el entrenamiento de los modelos corresponde a la base sin imputación, y que contiene 14 variables denominadas: 1) Edad, 2) Creatinina en sangre (mg/dl), 3) Índice de masa corporal (IMC), 4) Sexo, 5) Grupo Poblacional, 6) Colesterol Total, 7) Etiología ERC, 8) Diagnóstico ERC, 9) Tensión Arterial, 10) Tiempo de prestación de servicios, 11) Sí el Usuario recibe Antagonista de los receptores de angiotensina II (ARA II), 12) Albumina/creatinuria, 13) Sí, la persona encuentra en un programa de

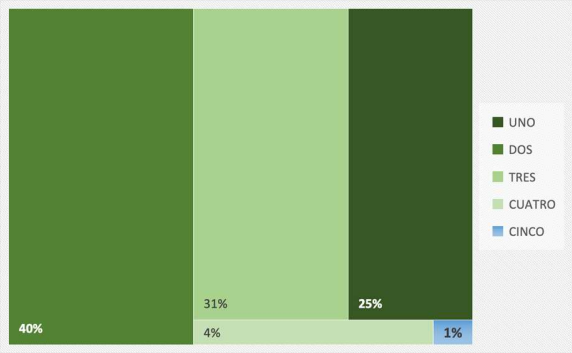
atención de ERC (renoprotección, nefroprotección, protección renal, prediálisis), 14) si ha recibido o no trasplante renal, adicionalmente se retiran las filas que contienen NaN en la variable respuesta, por lo que se obtiene un marco de datos final de 14 variables y 3.189 observaciones.

En el marco de datos final se encuentra variables discretas tales como: Edad, Tiempo de prestación de servicios, siendo las continuas: Creatinina en sangre (mg/dl), Índice de masa corporal (IMC), Diagnóstico ERC, Tensión Arterial, Sí el Usuario recibe Antagonista de los receptores de angiotensina II (ARA II), Albumina/creatinuria y por ultimo nominales sexo, Grupo Poblacional, Etiología ERC, Sí, la persona encuentra en un programa de atención de ERC (renoprotección, nefroprotección, protección renal, prediálisis), si ha recibido o no trasplante renal,

La generación de los modelos se lleva a cabo empleando los algoritmos de clasificación previamente seleccionados (Regresión Logística Multinomial, Máquinas de Soporte Vectorial, Modelos de ensamble, Redes neuronales), ejecutándolos en el entorno web Colab, programados en lenguaje Python y con la división de datos del 70/30. Como resultado se genera el informe de clasificación en las métricas denominadas en inglés, Precisión, Recall, F1-Score, además de la matriz de confusión para evaluar la clasificación de individuos.

En cuanto a los estadios, existe un desbalance toda vez que presenta una distribución desigual entre sus clases, como se observa en el gráfico No. 5, por ejemplo, la clase 2 corresponde al 40% de las observaciones de la variable estadio, considerándose como mayoritaria, mientras que la clase 5 es el 1%, siendo la minoritaria, por lo que se espera que los modelos entrenados tiendan a clasificar mejor las clases mayoritarias, como son el estadio 1,2 y 3 (43).

Gráfico No.5– Distribución de clases en el data set



Fuente: Elaboración propia

7. Evaluación

Al comparar el rendimiento de los modelos en términos de métricas, utilizando la base de datos imputada y la base de datos sin imputar, se observa que para el estadio 5 se logró la clasificación de individuos en los modelos Random Forest y Red Neuronal Feed Forward con la base de datos sin imputar a diferencia de la imputada, así como un incremento en los valores de precisión y recall del estadio 4 y 5, como se aprecia en la tabla No.2, siendo estos dos estadios los que mayor desbalance presentan, por lo que sumado a las pruebas de significancia estadística descritas en el numeral 5.3, los algoritmos definitivos fueron entrenados con la base de datos sin imputar, siendo estos los modelos a evaluar.

Para evaluar los modelos, otra consideración fundamental es el propósito de la investigación o el objetivo del negocio que consiste en clasificar a los pacientes diagnosticados con enfermedad renal crónica en los estadios 1,2,3,4 o 5 mediante el uso de algoritmos de aprendizaje supervisado. En este contexto el interés radica en captar la mayoría de los casos positivos posibles disminuyendo los falsos negativos, lo que refleja la métrica recall, así como considerar la proporción de predicciones correctas en relación con el total de los individuos clasificados, representado por el accuracy. No obstante, se abordarán las métricas del F1-Score y Precisión (44).

El modelo de Regresión Logística Multinomial (MRL) presentó un "accuracy" general del 67%, mostró equilibrio entre las métricas de recall y precisión para los estadios 1,2 y 3 , ya que obtuvo valores cercanos entre ellos, no obstante, para las clases 4 y 5, el modelo no clasificó a ningún individuo como positivo, ya que el recall fue 0, y a su vez no identificó correctamente casos como positivo (ver tabla No. 2).

En relación con las máquinas de soporte vectorial, se observa un desequilibrio entre las métricas precisión y recall para las clases 1,2, 3 y 4, siendo más evidente en el caso del estadio 4, toda vez que la precisión alcanzó un valor de 1.00 indicando que clasifica correctamente como positivos todos los individuos que realmente lo son, sin embargo el recall fue de 0.03, lo que implica que clasificó a 83 pacientes como negativos (83) cuando realmente no lo eran y solo 3 como verdaderos positivos (Ver Anexo C). En lo que refiere al estadio 5, muestra similar comportamiento en sus métricas al observado en el modelo MRL (Ver tabla No. 2).

En los modelos de ensamble, el Random Forest tiene un accuracy general del 76%, para los estadios 1,2,3 muestra valores cercanos entre sí para las métricas de precisión y recall, los cuales son más

altos que los mostrados en el MRL, pese a esto, para el estadio 4 registra una sensibilidad (recall) del 33%, indicando que clasifica correctamente una tercera parte de los casos positivos reales, lo que resulta en un numero alto de falsos negativos. Aunque el Random Forest logra clasificar individuos en el estadio 5 con un 100% en sus métricas, lo que equivale a no cometer errores en la clasificación , se infiere que el modelo puede tener un sobre ajuste en este aspecto debido a la poca cantidad de pacientes que se tienen en la clase 5 (Ver Tabla. No.2).

El segundo modelo de ensamble trabajado es el XGBoost, el cual logró una tasa correctamente de clasificación del 78% sobre todo el conjunto de datos, posicionándose en primer lugar en esta métrica entre los modelos entrenados. Es importante destacar que el conjunto de datos presenta desbalanceo en todas las clases, por lo que se realiza una evaluación individual del recall que arrojó valores para los estadios 1,2,3,4, y 5 de 80%, 77%, 81%, 53% y 100% respectivamente.

Estos resultados muestran mejoras en 1 punto porcentual en la precisión de las dos primeras clases, siendo el estadio 3 el que muestra el mejor valor de sensibilidad en el modelo, adicionalmente se observa una mejora significativa para el estadio 4 en la reducción de falsos negativos, comparado con los modelos anteriores registrando un recall del 53%. Sin embargo, al igual que el Random Forest, el XGBoost muestra un posible sobre ajuste al ser 100% todas las métricas del estadio 5.

El último modelo de ensamble es el CatBoost, al igual que el RLM y MSV no clasificó a ningún individuo como positivo, obteniendo un recall y precisión de 0, no logrando identificar de forma acertada los individuos positivos. Para los estadios 3 y 4, su mejor métrica es el recall mientras que para los estadios 1 y 2 es para la precisión. En general el F1-Score para las clases 1,2,3 y 4 es de 69%, 67%, 68% y 48% respectivamente.

Incrementado la complejidad de los modelos, se encuentran las redes neuronales, el primer modelo denominado “Corn”, cuyo comportamiento para la clase 5, es similar a los modelos RLM, MSV y Catboost, obteniendo el valor de 0 en todas sus métricas. Sin embargo, destaca su capacidad de capturar positivos y reducir falsos negativos en la clase 3, logrando un recall del 83%, además muestra un rendimiento importante en la precisión de clasificar verdaderos positivos alcanzando una precisión del 81%. En cuanto al accuracy general tiene un valor del 76% situándose en este aspecto al mismo nivel que el Random Forest.

Finalmente, la red neuronal Feed Foodward, clasifica todos los cinco estadios, el quinto estadio tiene una precisión, recall y F1-Score de 1, indicando posiblemente sobre ajuste, la clase 2 tiene el recall 0.82, y la clase 4 la precisión de 0.84 más elevada de 0.84, obteniendo en general una proporción de predicciones correctas del 77% (accuracy), estos resultados siguen la tendencia establecida por el XGBoost.

En la siguiente tabla se encuentra el resumen de las métricas de cada uno de los modelos aplicados en la investigación, discriminado por estadio de la enfermedad renal crónica, así como el comparativo con base de datos imputada y sin imputar.

Tabla No. *2 Métricas de los modelos aplicados*

TIPO	MODELO	E	PRECISIÓN		RECALL		F1-SCORE		ACCURACY		VAR	BIAS
			IMP	SIN	IMP	SIN	IMP	SIN	IMP	SIN	SIN	SIN
Modelos de clasificación	Regresión Logística Multinomial	1	0.80	0.66	0.62	0.65	0.70	0.65	0.66	0.67	0,44	-0,046
		2	0.64	0.59	0.69	0.61	0.67	0.60				
		3	0.64	0.60	0.71	0.67	0.68	0.63				
		4	0.06	0.00	0.03	0.00	0.04	0.00				
		5	0.00	0.00	0.00	0.00	0.00	0.00				
	Máquinas de soporte vectorial	1	0.56	0.76	0.32	0.44	0.41	0.55	0.52	0.59	0,49	0,082
		2	0.49	0.53	0.76	0.72	0.60	0.61				
		3	0.58	0.61	0.38	0.63	0.46	0.62				
		4	0.00	1.00	0.00	0.03	0.00	0.07				
		5	0.00	0.00	0.00	0.00	0.00	0.00				
Modelos de ensamble	Random Forest	1	0.86	0.79	0.80	0.80	0.83	0.79	0.81	0.76	0,27	-0,033
		2	0.78	0.75	0.85	0.76	0.81	0.75				
		3	0.81	0.75	0.80	0.78	0.81	0.76				
		4	0.76	0.80	0.35	0.33	0.48	0.47				
		5	0.00	1.00	0.00	1.00	0.00	1.00				
	Modelo de ensamble/ XG Boost	1	0.87	0.81	0.83	0.80	0.85	0.80	0.82	0.78	0,24	-0,002
		2	0.81	0.78	0.84	0.77	0.83	0.77				
		3	0.82	0.76	0.83	0.81	0.83	0.78				
		4	0.68	0.79	0.51	0.53	0.58	0.63				
		5	0.1	1.00	0.29	1.00	0.44	1.00				
	CATBOOST	1	0.75	0.70	0.70	0.69	0.73	0.69	0.67	0.67	0,24	-0,003
		2	0.70	0.69	0.78	0.66	0.74	0.67				

TIPO	MODELO	E	PRECISIÓN		RECALL		F1-SCORE		ACCURACY		VAR	BIAS
			IMP	SIN	IMP	SIN	IMP	SIN	IMP	SIN	SIN	SIN
		3	0.74	0.66	0.70	0.70	0.72	0.68				
		4	0.67	0.45	0.33	0.52	0.44	0.48				
		5	0.00	0.00	0.00	0.00	0.00	0.00				
Red Neuronal	Red Neuronal Corn	1	0.84	0.73	0.78	0.78	0.81	0.75				
		2	0.80	0.75	0.80	0.73	0.80	0.74				
		3	0.80	0.81	0.86	0.83	0.83	0.82	0.80	0.76	0,2	-0,067
		4	0.32	0.72	0.22	0.54	0.26	0.62				
		5	0.25	0.00	0.25	0.00	0.25	0.00				
	Red Neuronal Feed Foodward	1	0.91	0.83	0.76	0.67	0.83	0,74				
		2	0.82	0.73	0.76	0.82	0.79	0.77				
		3	0.71	0.79	0.95	0.81	0.81	0.80	0.79	0.77	0,49	0,082
		4	0.00	0.84	0.00	0.58	0.00	0.69				
		5	0.00	1.00	0.00	1.00	0.00	1.00				

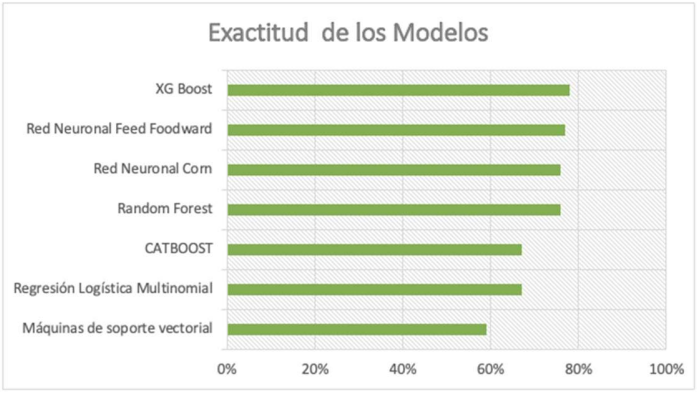
Fuente: Elaboración Autoras.

8. Resultados

Para abordar el problema de clasificación de individuos en los cinco estadios de la enfermedad renal crónica, involucró la implementación de dos modelos de clasificación, tres de ensamble y dos de redes neuronales. Para evaluar su rendimiento se utilizaron métricas como precisión, recall, F1- Score y accuracy, con el objetivo de identificar el mejor modelo en función de la base de datos estructurada que comprende el reporte de información sobre enfermedad renal crónica, hipertensión, diabetes asociada a la cuenta de alto costo y el objetivo del trabajo de investigación.

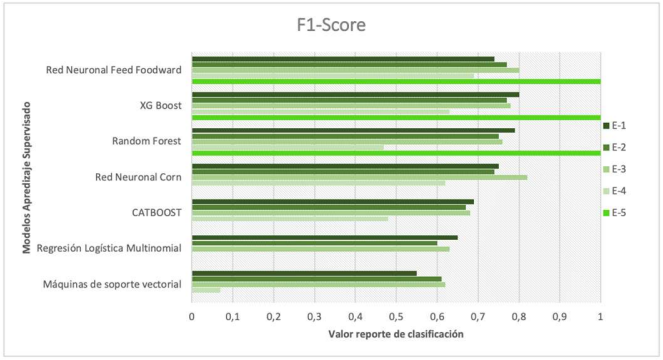
El gráfico 7 representa la exactitud de cada modelo (Accuracy), una métrica que engloba el desempeño de los cinco estadios en un solo valor en términos de predicciones positivas correctas, los modelos de ensamble y las redes neuronales presentan un desempeño superior al 0.7, en su orden el XGBoost liderando en un 78%, seguido de las Redes Neuronales Feed Forward con 77%, Corn con 76% y el Random Forest con 76%. En cuanto a la métrica recall fundamental que los algoritmos logren clasificar individuos en todas las clases en términos de recall, precisión y F1-Score, los modelos que cumplen con este requisito son XGBoost, Random Forest y Red Neuronal Feed Forward, como se visualiza en los gráficos del 8 al 10.

Gráfico No.7- Comparativo modelos Exactitud (Accuracy)



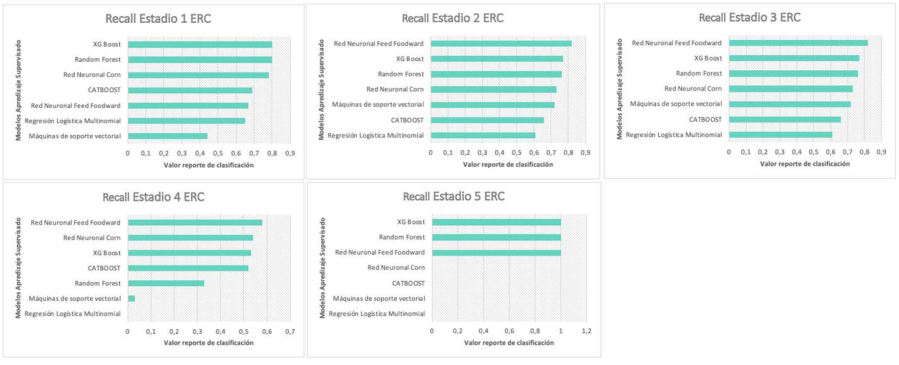
Fuente: Elaboración propia

Gráfico No.10- Comparativo modelos F1-Score



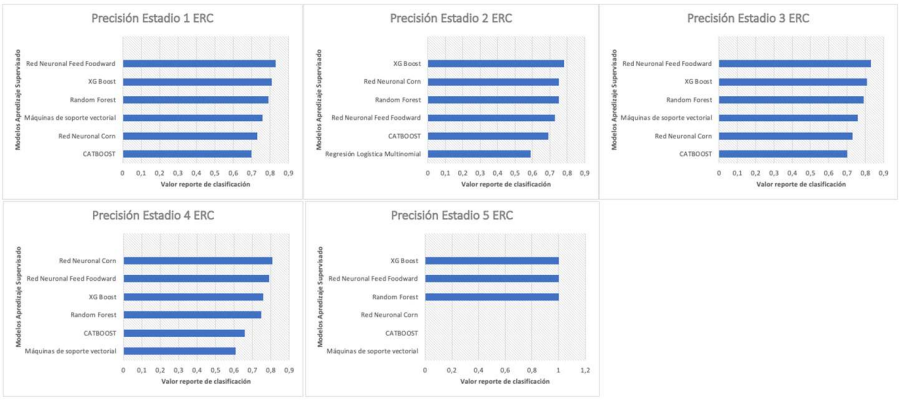
Fuente: Elaboración propia

Gráfico No.8- Comparativo modelos Recall



Fuente: Elaboración propia

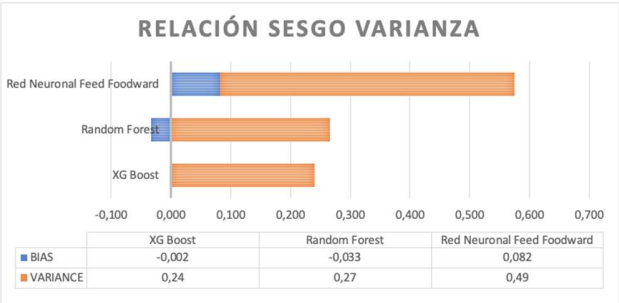
Gráfico No.9- Comparativo modelos Precisión



Fuente: Elaboración propia

Por último, al comparar los tres modelos que logran clasificar los cinco estadios en términos de sesgo y varianza, los autores Baştanlar et al, señala que un modelo con un BIAS (sesgo) elevado corresponde a un modelo subajustado, asimismo, indican que en caso de tener una varianza elevada, es sinónimo de sobre ajuste, por lo tanto, idealmente se busca que estas dos métricas tiendan a cero (42). En los modelos entrenados, se observa que XGBoost tiende obtener los valores más bajos, en cuanto al BIAS, a pesar de obtener un valor negativo de -0,002 es el que más se aproxima a cero, por lo que su clasificación es más precisa, en cuanto al varianza, al ser el valor de 0,24 siendo el más bajo entre los modelos, por lo que la clasificación presenta menor variación entre estadios.

Gráfico No.11- Comparativo BIAS - Variance



Fuente: Elaboración propia

Por lo tanto, tomando en cuenta el modelo de ensamble XGBoost, entre todos los modelos entrenados tiene el accurancy más elevado (78%), permite clasificar individuos en los cinco estadios de interés, ubicándose siempre en los tres primeros modelos con resultados más altos en términos de Recall, esto indica que captura adecuadamente

los casos positivos, y que presenta resultados de sesgo y varianza que tienden más a cero. Por ende, es el algoritmo seleccionado para efectuar la fase de optimización y despliegue, como indica la metodología CRISP-DM.

8.1. Optimización

Se llevó a cabo una optimización de parámetros con el objetivo de potenciar el rendimiento del modelo, aumentar la precisión y prevenir el sobreajuste, que ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y pierde su capacidad de generalización a nuevos datos. Se emplearon tres métodos para esta optimización: búsqueda por grilla, búsqueda aleatoria y optimización bayesiana Tras evaluar los resultados, se optó por seleccionar el segundo método, la búsqueda aleatoria, utilizando la función "RandomizedSearch" de la biblioteca scikit-learn.

La búsqueda aleatoria demostró una mejor identificación de la mejora de los hiperparámetros y, por ende, de las métricas del modelo. Entre los hiperparámetros optimizados se encuentran learning_rate con un valor de 0.083, max_depth configurado en 5, n_estimators ajustado a 172, y subsample establecido en 0.597.

Al implementar el modelo XGBoost con estos nuevos parámetros optimizados, se observa un avance significativo en las métricas del modelo. La precisión supera el 80% en todas las clases, el recall para las clases 1, 2, 3 y 5 es superior al 80%, mientras que para la clase 4 alcanza el 67%. El F1 Score en todas las clases es superior al 70%. En general, se logró una precisión del 85%, representando un aumento del 7% en comparación con el rendimiento anterior.

Este enfoque de optimización, junto con los resultados de las métricas, muestra una clasificación más precisa para cada una de las clases, logrando un equilibrio entre precisión y generalización. Esto contribuye a la creación de un modelo más sólido y adaptable a conjuntos de datos nuevos

8.2. Despliegue

Para el despliegue, se utilizó Google Colab para la ejecución y carga de datos con el fin de probar las predicciones con el modelo seleccionado (XGBoost) con información ajena a la base de datos con la cual se entrenó y se realizó el test del modelo, logrando así una validación externa. Para utilizar el modelo final estructurado, es necesario seguir los siguientes pasos para su ejecución.

1. Se debe abrir el notebook llamado "Deployment".
2. En la parte izquierda, en la carpeta llamada "files", dentro de la subcarpeta "content", se deben cargar la base de datos para ejecutar el modelo seleccionado en formato .xlsx®, la cual se denomina "Imputada.xlsx".
3. Posteriormente cargar en formato .xlsx® la información de cada una de las 11 variables de los nuevos casos a los que se quiere hacer una clasificación predictiva, para efectos del ejemplo el archivo se encuentra identificado como "variables.xlsx", como se muestra en la imagen No. 1.

Imagen No. 1 – Base de datos con las observaciones de los nuevos individuos a clasificar.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
EDA	CR	IN	ETIO_E	Se	G_P(	COL		DX_E	ALB/CR	TIEMP_SER	PROG_E	ANTAGONIS	RECIBIO_TRA
96	5,48	18	98	2	61	173	130	2	9	1	2	2	5
59,509589	0,58	25,6311675	98	2	61	147	117	0	7	12	2	2	5
71,1972603	0,97	26,0261749	8	2	31	221	118	1	1,7	0	2	1	5
101,556164	0,8	27,471384	8	1	31	183	140	1	7,2	12	2	1	5
83,3589041	3,75	26,7299275	5	1	31	112	140	1	223	12	2	1	5
71,5424658	1	26,0789715	8	1	31	150	123	1	8,2	12	2	1	5
85,5945205	1,81	22,2063307	8	2	31	199	150	1	153,2	12	2	1	5
55,6219178	0,6	35,2507611	98	2	5	232	133	0	3	12	2	1	5
95,169863	0,75	22,7694384	8	1	31	155	140	1	41,4	8	2	1	5

Fuente: Elaboración Propia

4. La primera parte se debe ejecutar y contiene el código del modelo seleccionado y sus resultados obtenidos.
5. Para la predicción, se utilizó la biblioteca Pandas y Requests, que nos permite cargar el archivo XLSX donde se encuentran los nuevos casos. Mediante un bucle for, se lee línea por línea para realizar las predicciones de los estadios, como se muestra en la imagen No.2

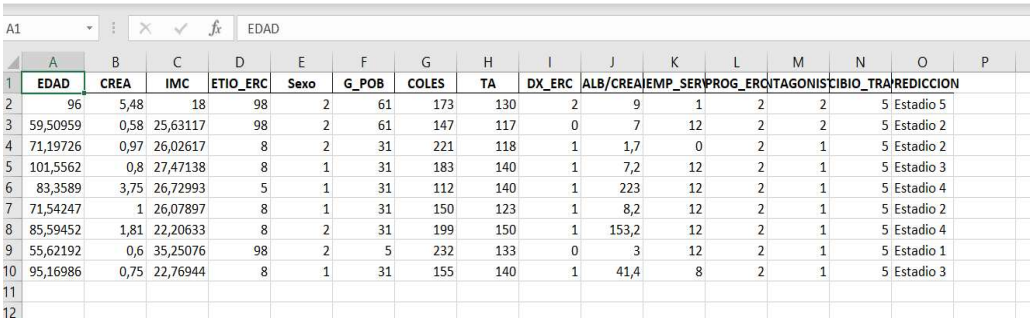
Imagen No. 2 – Código en lenguaje Python para el despliegue

```python
#Prediccion del modelo
import pandas as pd
import requests

# Definir la URL del servidor de predicción
url = "http://localhost:5000/predict"

# Ruta al archivo Excel
excel_file_path = "/content/variables.xlsx"

# Leer el archivo Excel en un DataFrame de pandas
df = pd.read_excel(excel_file_path)

# Mostrar el DataFrame leído del archivo Excel
print("Datos del archivo Excel:")
print(df)

# Convertir el DataFrame a una lista de listas
multiple_data = df.values.tolist()

# Lista para almacenar las predicciones
all_predictions = []

# Hacer predicciones para cada instancia de datos
for data_instance in multiple_data:
    predictions = model.predict([data_instance])
    all_predictions.append(predictions[0])

# Agregar las predicciones al DataFrame
df['PREDICCION'] = all_predictions


# Mapear códigos numéricos a nombres de estadios
mapeo_estadios = {0: 'Estadio 1', 1: 'Estadio 2', 2: 'Estadio 3', 3: 'Estadio 4',4: 'Estadio 5'}

# Aplicar el mapeo a la columna 'PREDICCION'
df['PREDICCION'] = df['PREDICCION'].map(mapeo_estadios)

# Guardar el DataFrame modificado en un nuevo archivo Excel
output_excel_path = "/content/variables_con_prediccion.xlsx"
df.to_excel(output_excel_path, index=False)

# Mostrar el DataFrame con la nueva columna de predicciones
print("Datos con predicciones:")
print(df)
```

Fuente: Elaboración Propia

6. Finalmente, se guarda un archivo llamado "variables_con_predicciones.xlsx", que contiene las líneas cargadas con sus respectivos estadios predichos, como se muestra en la imagen No. 3.

Imagen No. 3 – Base de datos con el resultado de la clasificación



EDAD	CREA	IMC	ETIO_ERC	Sexo	G_POB	COLES	TA	DX_ERC	ALB/CREA	EMP_SER	PROG_ERC	ANTAGONIS	CIBIO_TRA	REDICCION
96	5,48	18	98	2	61	173	130	2	9	1	2	2	5	Estadio 5
59,50959	0,58	25,63117	98	2	61	147	117	0	7	12	2	2	5	Estadio 2
71,19726	0,97	26,02617	8	2	31	221	118	1	1,7	0	2	1	5	Estadio 2
101,5562	0,8	27,47138	8	1	31	183	140	1	7,2	12	2	1	5	Estadio 3
83,3589	3,75	26,72993	5	1	31	112	140	1	223	12	2	1	5	Estadio 4
71,54247	1	26,07897	8	1	31	150	123	1	8,2	12	2	1	5	Estadio 2
85,59452	1,81	22,20633	8	2	31	199	150	1	153,2	12	2	1	5	Estadio 4
55,62192	0,6	35,25076	98	2	5	232	133	0	3	12	2	1	5	Estadio 1
95,16986	0,75	22,76944	8	1	31	155	140	1	41,4	8	2	1	5	Estadio 3

Fuente: Elaboración Propia

Este proceso de despliegue facilita la aplicación práctica del modelo a nuevos conjuntos de datos, permitiendo realizar la clasificación predictiva de forma eficientes y guardar los resultados para su posterior análisis.

9. Discusión:

En literatura que aborda estudios de comparación de modelos de aprendizaje supervisado para la detección de enfermedad renal, específicamente en términos de la exactitud (accuracy) de los algoritmos empleados, se destaca que el Random Forest arroja resultados cercanos al 100%. Aunque esto podría sugerir un buen rendimiento, señala una tendencia hacia el sobre ajuste (8). Sin embargo, en el contexto de este ejercicio de clasificación en cuatro grupos, el modelo logra resultados de 80% en esta métrica, evidenciando una robusta capacidad para actividades de clasificación que involucran más de dos condiciones.

Addullah et al, señala que la red neuronal Feedforward presenta un rendimiento sobresaliente en ejercicios de clasificación binaria de pacientes (tiene/ no tiene) la enfermedad renal crónica (25). Al evaluar el mismo modelo en una tarea de clasificación múltiple para

la misma condición médica, presenta valores elevados en sus métricas en comparación con los otros tipos de modelos de clasificación y ensamble. También demuestra que responde adecuadamente al entrenamiento con bases de datos desbalanceadas, toda vez logró clasificar individuos de la clase minoritaria, lo que confirma su idoneidad para abordar problemas de clasificación en el ámbito de la salud.

En relación con el XGBoost, los autores Ogunleye & Wang describen a este modelo como el mejor en su categoría para realizar clasificación predictiva de individuos que padezcan o no la ERC (7). Este rendimiento se confirma y se amplía aún más en el presente estudio, especialmente en problemas de clasificación multiclase, toda vez que demuestra estabilidad en sus métricas, y destaca como el mejor modelo entre los siete evaluados, especialmente en la clasificación de grupos con escasos individuos, siendo un modelo estable en términos de sesgo y varianza.

Siguiendo con el modelo XGBoost, en futuros trabajos se puede abordar su capacidad adaptación a otras poblaciones, toda vez que es importante considerar que la estructura de la base de datos empleada en el presente trabajo de investigación está estandarizada por parte del organismo Cuenta de Alto Costo, lo que podría hacer que este trabajo de investigación sea altamente replicable en otras Instituciones Prestadoras de Servicios de Salud, sin embargo, es relevante tener en cuenta que esto puede variar debido a la calidad y cantidad de la información registrada los marcos de datos de cada Entidad, lo que abre nuevas posibilidades para aplicación de los métodos de aprendizaje supervisado en el área de la salud, no solamente para diagnóstico de enfermedades, sino también para evaluar progresión de las mismas con un enfoque para la proyección de recursos como parte del aseguramiento de la atención en salud.

10. Consideraciones

Éticas: Las autoras declaran que los datos utilizados en el presente trabajo no contienen información que hagan identificable el sujeto, ni muestra referencias de pacientes, adicionalmente no se realizaron experimentos con humanos o animales.

Financiación: Las autoras declaran que el presente trabajo de investigación fue realizado con recursos propios, por lo cual carece de algún tipo de financiación.

Conflicto de interés: Las autoras declaran que no tener conflicto de interés.

REFERENCIAS

1. AKF´s Medical Advisory Committee. Etapas o estadios de la enfermedad renal [Internet]. 2023 [citado el 30 de noviembre de 2023]. Disponible en: https://www.kidneyfund.org/es/todo-sobre-los-rinones/etapas-o-estadios-de-la-enfermedad-renal#:~:text=La%20enfermedad%20renal%20cr%C3%B3nica%20(ERC)%20se%20divide%20en%20cinco%20etapas,y%20peor%20funcionan%20los%20ri%C3%B1ones.
2. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Vol. 12, Kidney International Supplements. Elsevier B.V.; 2022. p. 7–11.
3. Cuenta de Alto Costo (CAC). Situación de la Enfermedad Renal Crónica, la hipertesión arterial y la diabetes mellitus en Colombia 2022. Bogotá D.C.; 2023.
4. Cuenta de Alto Costo (CAC). Quienes somos - CAC [Internet]. 2023 [citado el 9 de noviembre de 2023]. Disponible en: https://cuentadealtocosto.org/quienes-somos/
5. Universidad de los Andes. Uniandes.edu.co Noticias / Ciencia, tecnología y salud / Salud y medicina /. 2018 [citado el 10 de octubre de 2023]. The Economist destaca modelo de atención renal en Colombia como uno de los mejores de Latinoamérica. Disponible en: https://uniandes.edu.co/es/noticias/gobierno-y-politica/modelo-de-atencion-renal-en-colombia-uno-de-los-mejores-de-latinoamerica-segun-the-economist
6. Ministerio de Hacienda y Crédito Público. Decreto 2644 de 2022. 2644 Colombia; dic 30, 2022 p. 1–73.
7. Ogunleye A, Wang QG. XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM Trans Comput Biol Bioinform. el 1 de noviembre de 2020;17(6):2131–40.
8. Roy MS, Ghosh R, Goswami D, Karthik R. Comparative Analysis of Machine Learning Methods to Detect Chronic Kidney Disease. En: Journal of Physics: Conference Series. IOP Publishing Ltd; 2021.
9. Whsd G, Kdm P, Kadcp K. Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD). En: IEEE Compute Society, editor. 17Th International Conference on Bioinformatics and Bioengineering. IEEE; 2017. p. 1–6.
10. Thongprayoon C, Kaewput W, Choudhury A, Hansrivijit P, Mao MA, Cheungpasitporn W. Is it time for machine learning algorithms to predict the risk of kidney failure in patients with chronic kidney disease? Vol. 10, Journal of Clinical Medicine. MDPI; 2021. p. 1–3.
11. Ghosh P, Javed Mehedi Shamrat FM, Shultana S, Afrin S, Anjum AA, Khan AA. Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm. En: Proceedings - 2020 15th International Joint Symposium on Artificial Intelligence and

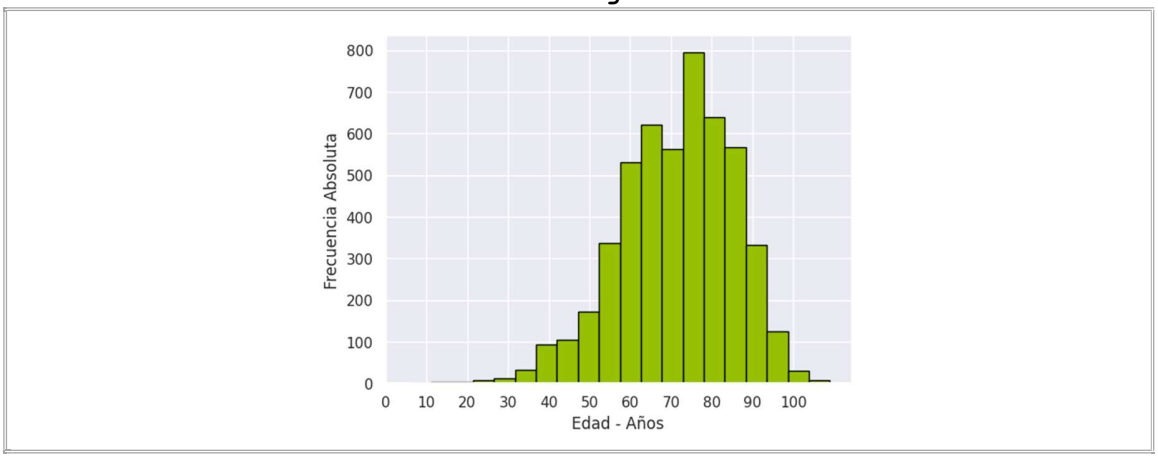
Natural Language Processing, iSAI-NLP 2020. Institute of Electrical and Electronics Engineers Inc.; 2020.
12. Charris L, Henriquez C, Hernandez S, Jimeno L, Guillen O, Moreno S. Comparative analysis of algorithms of decision trees in the processing of biological data. Revista I+D en TIC [Internet]. 2017;9:26–34. Disponible en: http://revistas.unisimon.edu.co/index.php/identic
13. Chittora P, Chaurasia S, Chakrabarti P, Kumawat G, Chakrabarti T, Leonowicz Z, et al. Prediction of Chronic Kidney Disease - A Machine Learning Perspective. Vol. 9, IEEE Access. Institute of Electrical and Electronics Engineers Inc.; 2021. p. 17312–34.
14. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA. el 20 de abril de 2011;305(15):1553–9.
15. Iftikhar H, Khan M, Khan Z, Khan F, Alshanbari HM, Ahmad Z. A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease. Sustainability (Switzerland). el 1 de febrero de 2023;15(3).
16. Ventrella P, Delgrossi G, Ferrario G, Righetti M, Masseroli M. Supervised machine learning for the assessment of Chronic Kidney Disease advancement. Comput Methods Programs Biomed. el 1 de septiembre de 2021;209.
17. Xie G, Chen T, Li Y, Chen T, Li X, Liu Z. Artificial Intelligence in Nephrology: How Can Artificial Intelligence Augment Nephrologists' Intelligence? Kidney Diseases. 2020;6(1):1–6.
18. Morales V, Ricardo G. Clasificador con redes neuronales para el pronóstico de la enfermedad renal crónica en la población colombiana. [Bogotá D.C.]: Universidad Internacional de La Rioja (UNIR); 2019.
19. Sánchez Gómez C. Desarrollo de soluciones software mediante Aprendizaje Automático en el ámbito de la Salud: situación tecnológica y perspectivas. [Cartagena]: Universidad Politécnica de Cartagena; 2019.
20. Kearns B, Gallagher H, De Lusignan S. Predicting the prevalence of chronic kidney disease in the English population: A cross-sectional study. BMC Nephrol. 2013;14(1).
21. Payal M, Ajagbe SA, Ananth Kumar T. Support Vector Machines (SVMS) Based Advanced Health Care System Using Machine Learning Techniques. International Journal of Innovate Research in Computer and Communication Engineering [Internet]. mayo de 2022;10(5):1–9. Disponible en: https://www.researchgate.net/publication/360947162
22. Charris L, Henriquez C, Hernandez S, Jimeno L, Guillen O, Moreno S. Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos. Revista I+D en TIC [Internet]. 2014;9(1):26–34. Disponible en: http://revistas.unisimon.edu.co/index.php/identic
23. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. J Pathol Inform. el 1 de enero de 2023;14.

24. Shi X, Cao W, Raschka S. Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities. el 16 de noviembre de 2021; Disponible en: http://arxiv.org/abs/2111.08851
25. Addullah AI, Md Nur A, Fatema TJ. Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide & Deep Learning. Institute of Electrical and Electronics Engineers. 2018;27–9.
26. Al-Juboori SAM, Hazzaa F, Jabbar ZS, Salih S, Gheni HM. Man-in-the-middle and denial of service attacks detection using machine learning algorithms. Bulletin of Electrical Engineering and Informatics. el 1 de febrero de 2023;12(1):418–26.
27. Shiuh Tong Lim et al. Prediction of Thyroid Disease using Machine Learning Approaches and Featurewiz Selection. Journal of Telecommunication Electronic and Computer Engineering. 2023;15(3):9–15.
28. Ruiz Sarrias O. Curvas de aprendizaje en la optimización Bayesiana de Hiperparámetros. [Madrid]: Universidad Nacional de Educación a Distancia; 2021.
29. Amador Martínez WD, Flórez Valencia L. Modelos de aprendizaje automático para la predicción de la progresión de la enfermedad renal crónica. [Bogotá D.C.]: Pontificia Universidad Javeriana; 20221.
30. Oróstica Tapia KY. Implementación de modelos de clasificación en cáncer basados en datos mutacionales y clínicos. [Santiago de Chile]: Universidad de Chile; 2021.
31. International Business Machines (IBM). Guía de CRISP-DM de IBM SPSS Modeler. 2021 [citado el 14 de octubre de 2023]. Conceptos básicos de ayuda de CRISP-DM. Disponible en: https://www.ibm.com/docs/es/spss-modeler/18.4.0?topic=dm-crisp-help-overview
32. Google. Colaboratory "Colab" [Internet]. 2023 [citado el 1 de diciembre de 2023]. Disponible en: https://colab.research.google.com/?hl=es-419#scrollTo=5fCEDCU_qrC0
33. Fondo Colombiano de Enfermedades de Alto Costo. Instructivo para el reporte de información según Resolución 2463/14. Bogotá D.C- : Organización Cuenta de Alto Costo ; 2016. p. 1–12.
34. World Health Organization. data.who.int. 2019 [citado el 30 de noviembre de 2023]. Esperanza de vida al nacer (años). Disponible en: https://data.who.int/es/indicators/i/90E2E48
35. NumFOCUS Inc. 6.2.1. 2023 [citado el 1 de diciembre de 2023]. pandas.isnull — pandas 2.1.3 documentation. Disponible en: https://pandas.pydata.org/docs/reference/api/pandas.isnull.html
36. Scikit Learn Developers. 6.4. Imputation of missing values — scikit-learn 1.3.2 documentation [Internet]. 2023 [citado el 1 de diciembre de 2023]. Disponible en: https://scikit-learn.org/stable/modules/impute.html
37. Reyes Osorio T. Evaluación de metodología de imputación de datos en motores Diesel para el desarrollo de sistemas de diagnóstico inteligente de fallas. [Santiago de Chile]: Universidad de Chile; 2023.

38. Useche L, Mesa D. Una introducción a la imputación de valores perdidos. Terra. 2006;XXII(31):127–52.
39. NumFOCUS Inc. 6.2.1. 2023 [citado el 3 de diciembre de 2023]. pandas.DataFrame.corr — pandas 2.1.4 documentation. Disponible en: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html
40. Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesth Analg. el 1 de mayo de 2018;126(5):1763–8.
41. AutoViML. Github.com. 2020 [citado el 2 de diciembre de 2023]. featurewiz/README.md. Disponible en: https://github.com/AutoViML/featurewiz/blob/main/README.md
42. Baştanlar Y, Özuysal M. Introduction to machine learning. En: Methods in Molecular Biology. Humana Press Inc.; 2014. p. 105–28.
43. Hoyos Osorio JK. Metodología de clasificación de datos desbalanceados basado en métodos de sub muestreo. [Pereira]: Universidad Tecnológica de Pereira; 2019.
44. Rubini Lj, Assistant Professor K. Generating comparative analysis of early stage prediction of Chronic Kidney Disease. International Journal of Modern Engineering Research [Internet]. 2015;5(7):49–55. Disponible en: www.ijmer.com

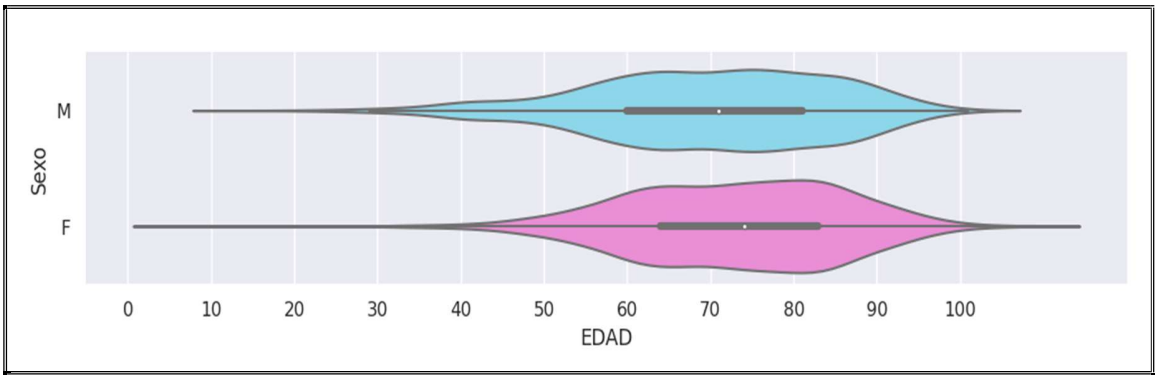
ANEXO A
GRÁFICOS EXPLORACIÓN DE DATOS

No. 1 – Histograma Edad



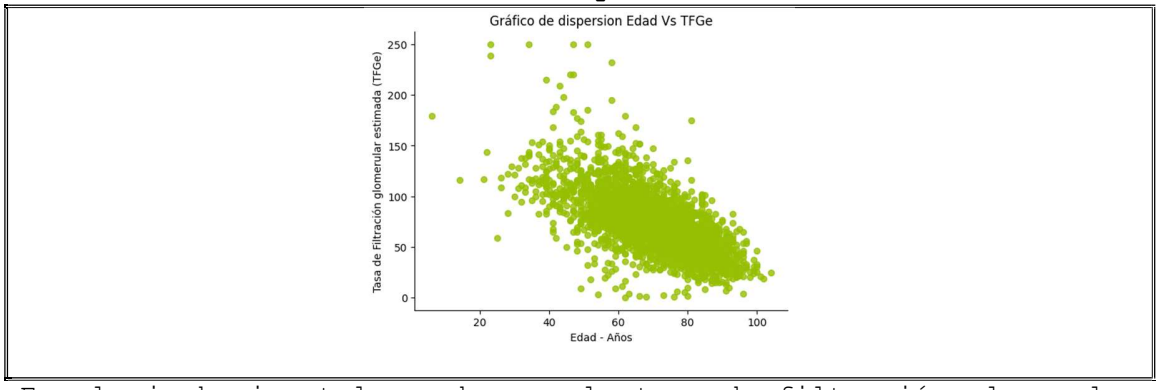
El histograma muestra una distribución asimétrica hacia la izquierda de las edades de los pacientes con afecciones de ERC, HTA y DM. La mayoría de los datos se concentra en el rango de 65 a 85 años, con un pico máximo cercano a los 75 años, que corresponde aproximadamente a 800. Adicionalmente, se observa en sus valores extremos individuos que han superado los 100 años y otros menores de 20 años.

No.2 – Gráfico Violín Edad Vs Sexo



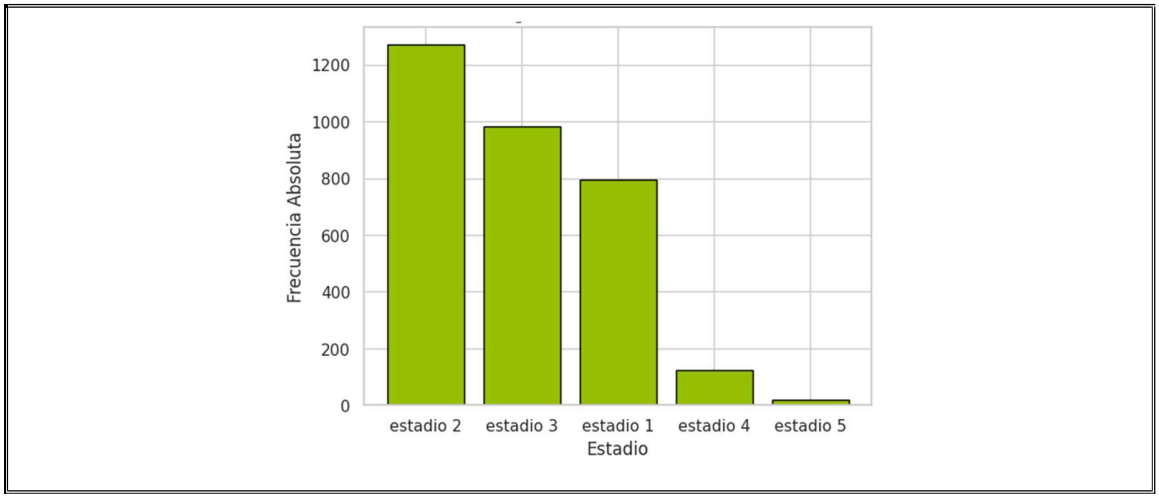
Alineado con lo anterior, el grafico de violín muestra que la mayor concentración de observaciones se sitúa entre las edades de 60 a 85 años. Para el grupo masculino, la moda es sutil y se encuentra a los 85 años, mientras que la mediana de 75 años. En el caso de las mujeres, la moda es de 75 años y la mediana de 70 años. Aunque existe una diferencia en la moda de las edades de ambos sexos, el rango intercuartílico con la dispersión de los datos presenta un comportamiento similar tanto para el sexo femenino y masculino.

No. 3 – Gráfico de dispersión Edad Vs TFGe



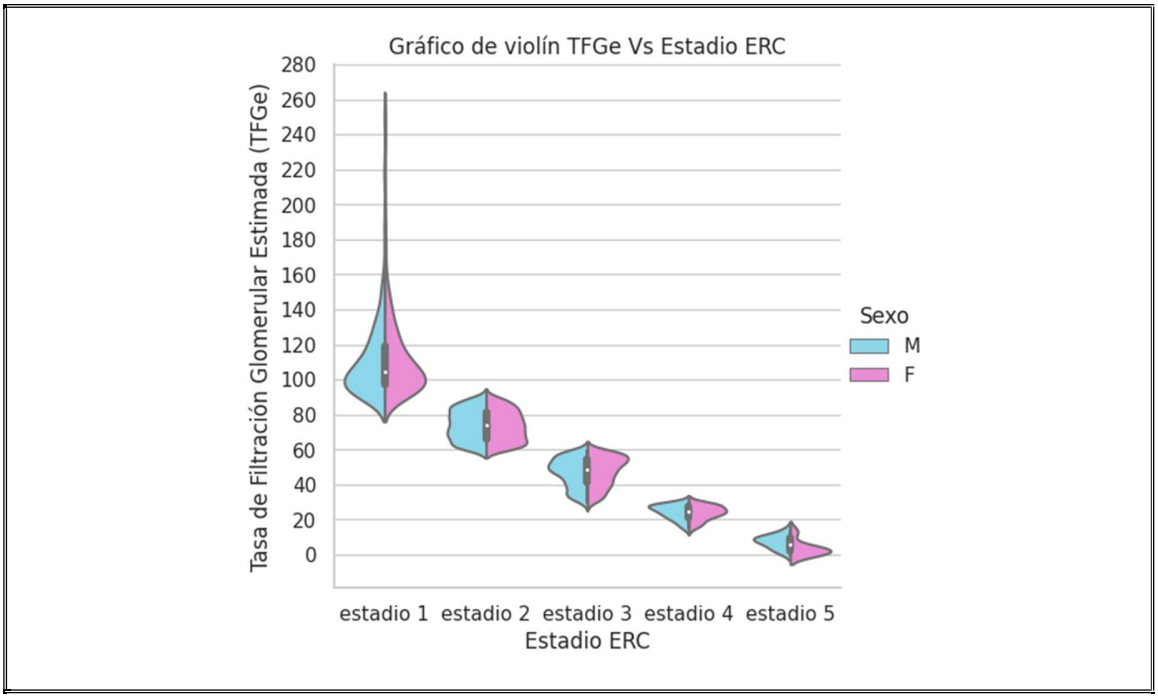
En el eje horizontal se observa la tasa de filtración glomerular estimada TFGe que tiene permitido tomar valores de 1 a 250 (33) y sirve como medida de la función renal, estando asociada directamente con los estadios 1,2,3,4 y 5 de la Enfermedad Renal Crónica (1). El grafico muestra la disminución del al TFGe en relación con el aumento de la edad, siendo un efecto esperado desde el punto de vista biológico. La mayoría de los individuos se concentran en el rango de edades de 50 a 90 años, con valores de TFGe distribuidos de 10 a 100, demostrando falla renal.

No. 4 – Diagrama de barras Estadio ERC



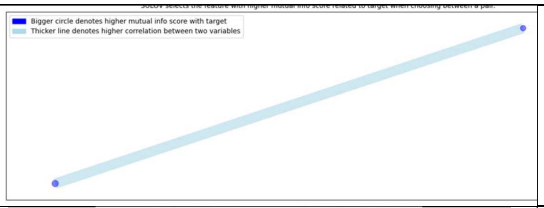
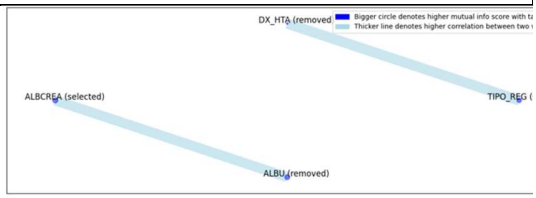
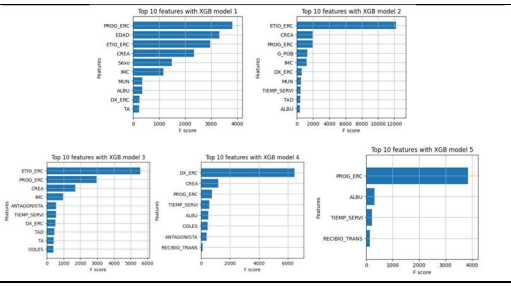
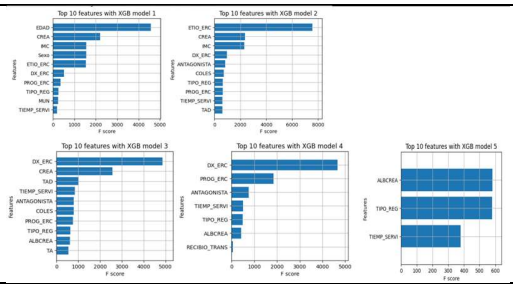
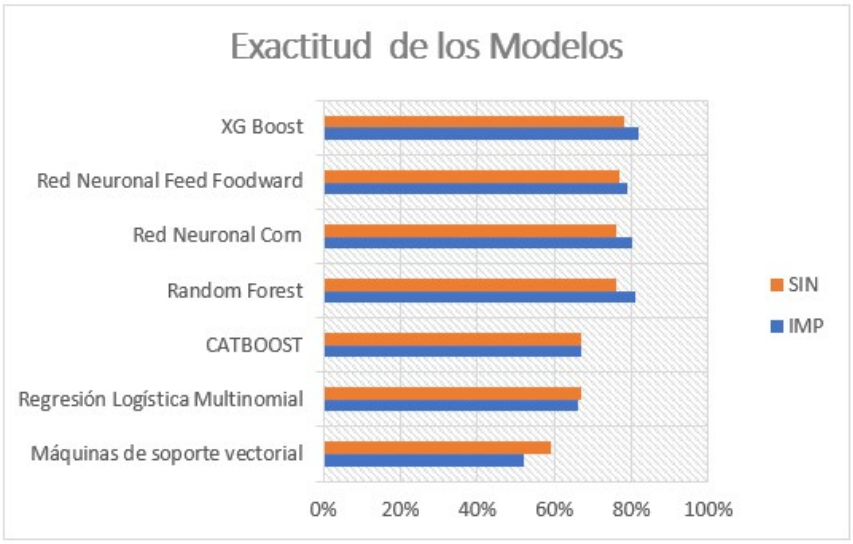
El diagrama muestra que, en la fase exploratoria la mayor concentración de individuos están en el estadio 2, seguido del estadio 3 y estadio 1. Sin embargo, se muestra una disminución significativa para el estadio 4 que representa el nivel más critico de la condición. Adicionalmente se registran muy pocos pacientes en el estadio5, siendo aproximadamente unos 20, lo que constituye una cantidad insuficiente para el entrenamiento. Asimismo, este estadio no se considera en el estudio ya que es la fase terminal de la enfermedad y la alternativa de tratamiento es el transplante renal, el cual depende de factores externos a la capacidad de aseguramiento por parte del administrador en recursos en salud.

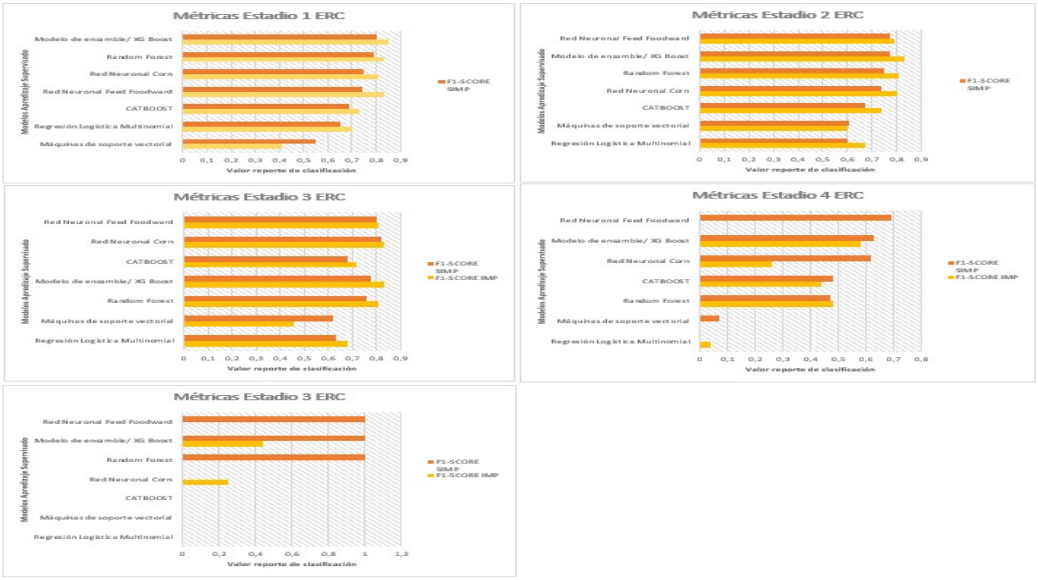
No. 5 – Gráfico de Violín Estadio ERC Vs TFGe



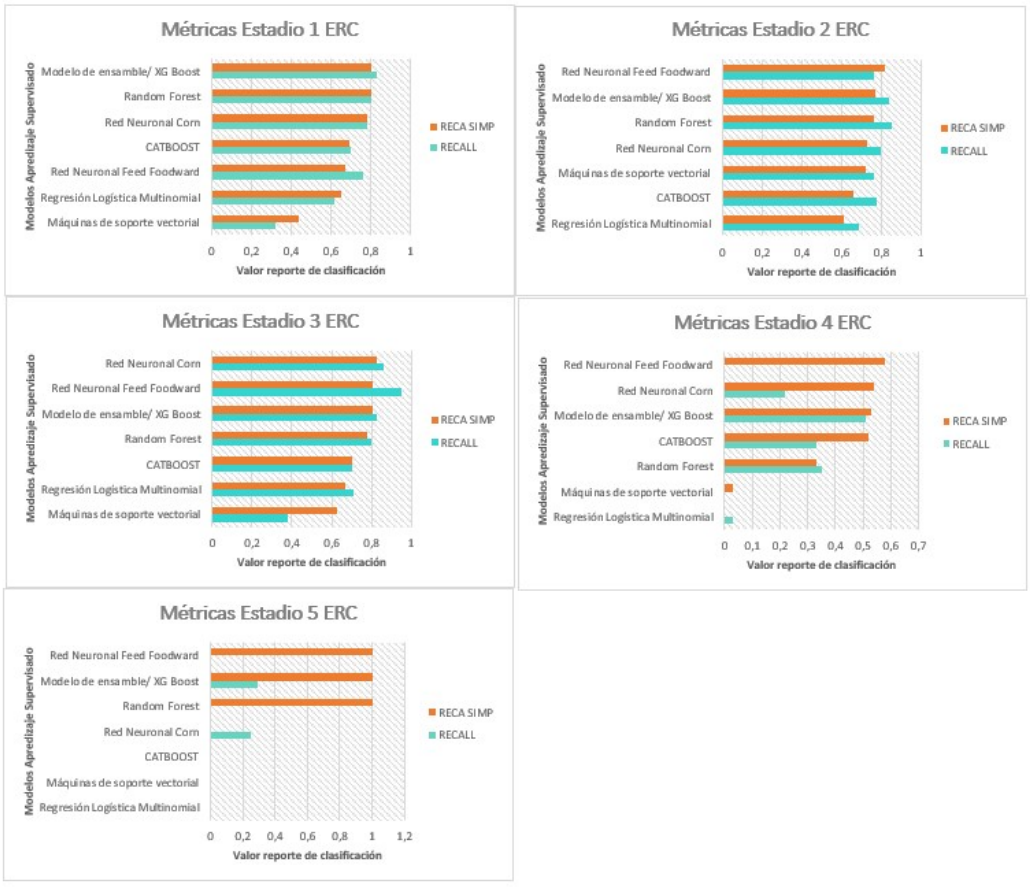
En cuanto al estadio 1 se observa la mayor concentración de datos tanto para hombres y para mujeres alrededor del valor de 100 ml/min de la tasa de filtración glomerular, lo que indica que se encuentra en el límite extremo inferior para considerarse con falla renal leve. No obstante, se evidencia registros con valores de 250 que pueden presentar falla renal por otras causas no atribuible a la TFGe. El estadio 2 tiene la mediana de 75 ml/min, mientras que el estadio 3 presenta una mediana de 50 ml/min. Sin embargo, los individuos del estadio 4 muestran una media cercana a los 20 ml/min siendo próxima al estadio 5, fase terminal.

ANEXO B
COMPARACIÓN RESULTADOS DATA ORIGINAL Y SIN IMPUTAR

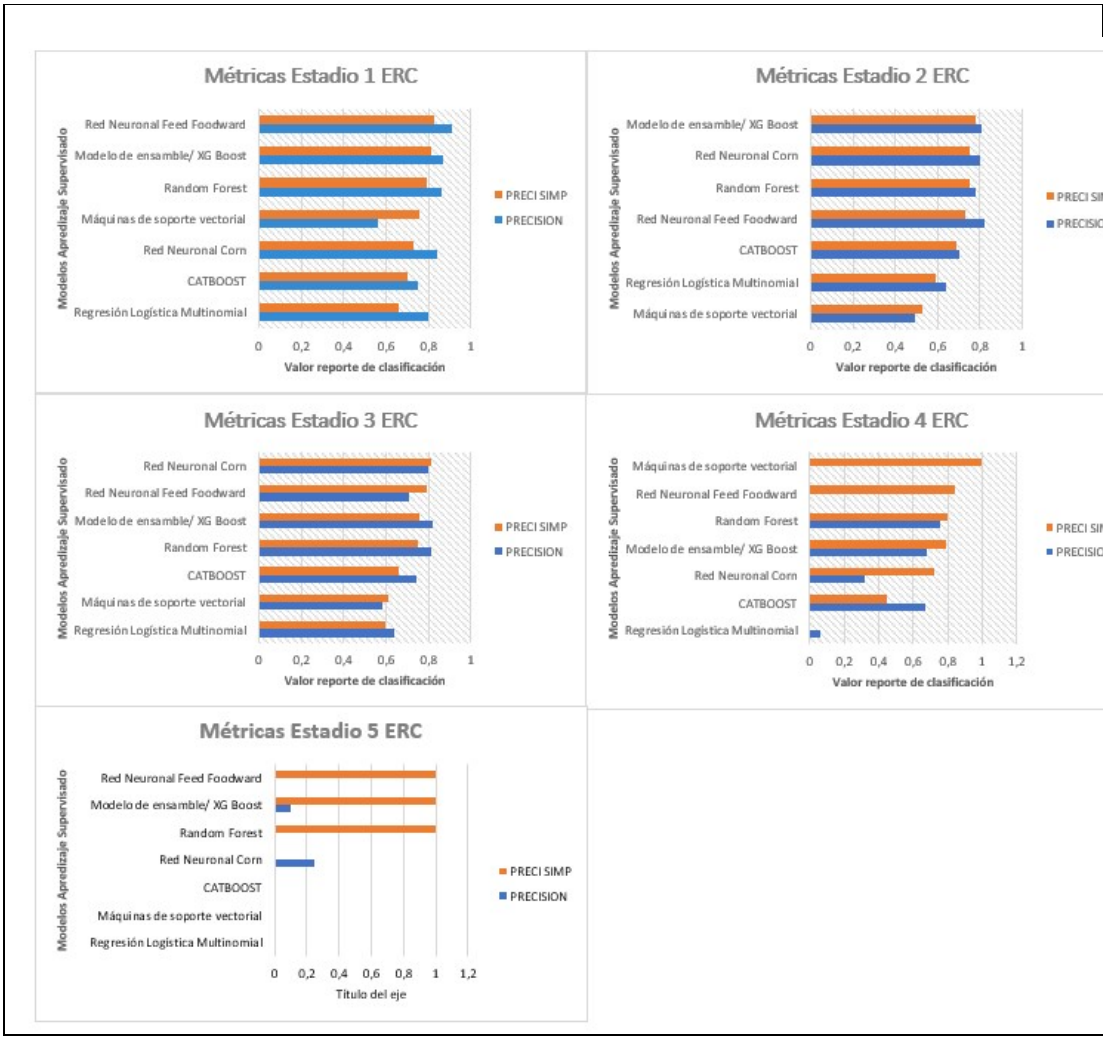
SIN IMPUTAR	IMPUTADA
Selección de variables	
	
	
(['EDAD', 'CREA ', 'IMC' , 'ETIO_ERC', 'Sexo', 'G_POB', 'COLES', 'TA','DX_ERC, 'ALB/CREA','TIEMP_SERVI','PROG_ERC', 'ANTAGONISTA', 'RECIBIO_TRANS']	(['EDAD','CREA ', 'IMC', 'ETIO_ERC', 'Sexo','G_POB', 'COLES', 'TAD','DX_ERC', 'ALB/CREA', 'TIEMP_SERVI', 'ANTAGONISTA', 'PROG_ERC', 'RECIBIO_TRANS']
COMPARATIVO MÉTRICAS	
ACCURACY	
	
F1 SCORE	



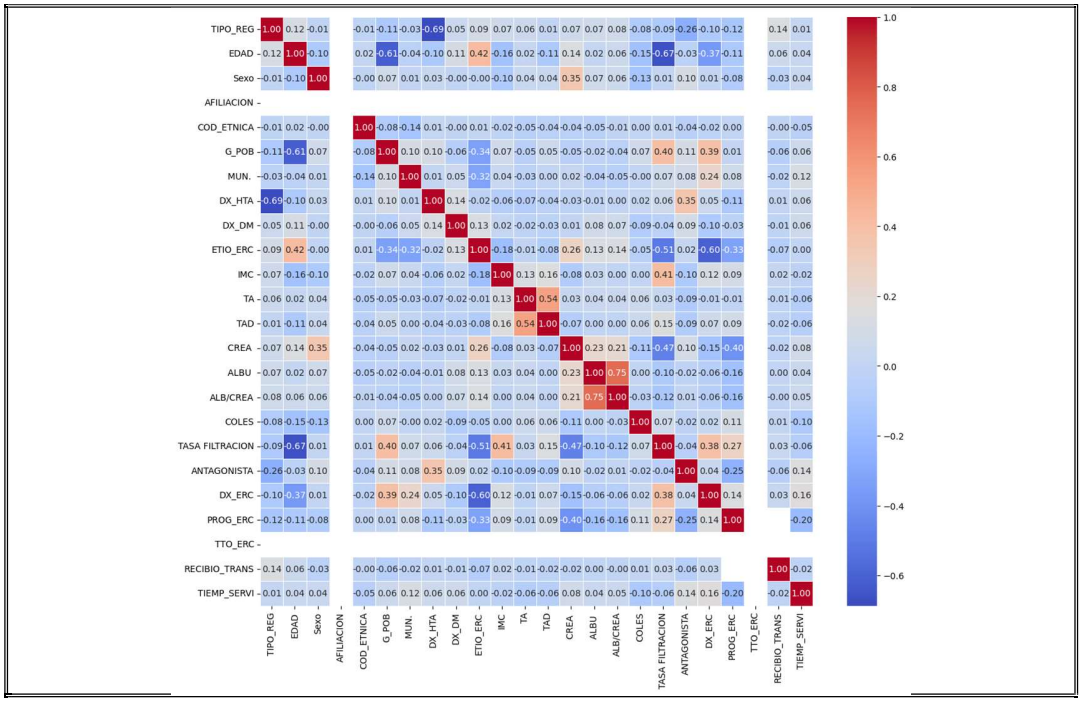
RECALL



PRECISIÓN



Mapa de calor de correlación

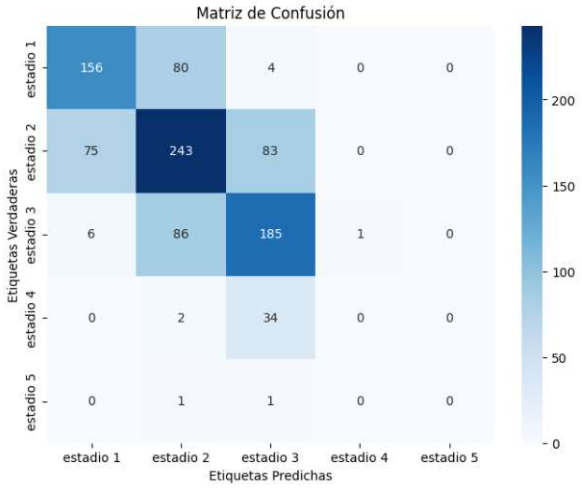


De la biblioteca Pandas se emplea la función "corr" para calcular la matriz de correlación de las variables de la base de datos, que previamente fue imputada. La función por defecto aplica el coeficiente de correlación de Pearson. En el mapa de calor se observa una correlación negativa del 67% para la variable edad y tasa de filtración, siendo coincidente con lo observado en la fase exploratoria, así como una correlación de -0.69 entre el diagnóstico de hipertensión arterial y el tipo de registro. En cuanto al coeficiente de microalbúmia y creatinina muestra una correlación positiva fuerte del 75%. Sin embargo, para las variables de tratamiento médico no dialítico para ERC estadio 5, y afiliación, no se obtuvo respuesta adecuada en la aplicación del método de correlación, a pesar de ser variables numéricas para efecto del análisis.

ANEXO C
REPORTES DE CLASIFICACIÓN Y MATRIZ DE CONFUSIÓN DE MODELOS

No. 1 – Regresión Logística Múltiple

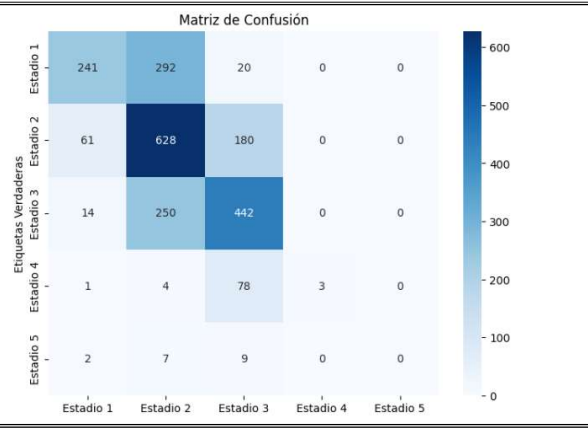
	precision	recall	f1-score	support
estadio 1	0.66	0.65	0.65	240
estadio 2	0.59	0.61	0.60	401
estadio 3	0.60	0.67	0.63	278
estadio 4	0.00	0.00	0.00	36
estadio 5	0.00	0.00	0.00	2
accuracy			0.61	957
macro avg	0.37	0.38	0.38	957
weighted avg	0.59	0.61	0.60	957



La clasificación más efectiva de individuos como verdaderos positivos se observa en el estadio 2, con 243 casos identificados correctamente, seguido por el estadio 3, con 185 casos, y el estadio 1, con 156 casos. Esta tendencia se refleja de manera similar en la métrica de 'recall', que indica la reducción de la tasa de falsos negativos. En términos de mitigación de falsos positivos, el mejor rendimiento se logra en el estadio 3, con un valor del 65%. Los estadios 4 y 5 exhiben un comportamiento muy similar, alcanzando su máximo número de identificaciones, más de 900 casos como verdaderos negativos .

No.2 – Maquinas de Soporte Vectorial

	precision	recall	f1-score	support
Estadio 1	0.76	0.44	0.55	553
Estadio 2	0.53	0.72	0.61	869
Estadio 3	0.61	0.63	0.62	706
Estadio 4	1.00	0.03	0.07	86
Estadio 5	0.00	0.00	0.00	18
accuracy			0.59	2232
macro avg	0.58	0.36	0.37	2232
weighted avg	0.62	0.59	0.57	2232

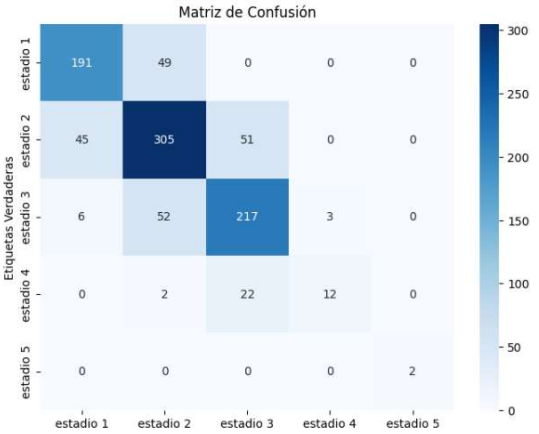


La MSV clasificó en el estadio 2 a 1.601 individuos de un total de 2.232 casos reales, obteniendo un recall del 72%, lo que presenta la tasa más baja de falsos negativos. Le sigue la clase 3 con un 63% y la clase 1 con un 44%. La métrica más alta de precisión es en el estadio 1 con un 76%. Sin embargo, al evaluar el F1-Score el estadio 3 muestra el equilibrio entre precisión y recall.

No. 3 - Random Forest

Reporte de Clasificación:

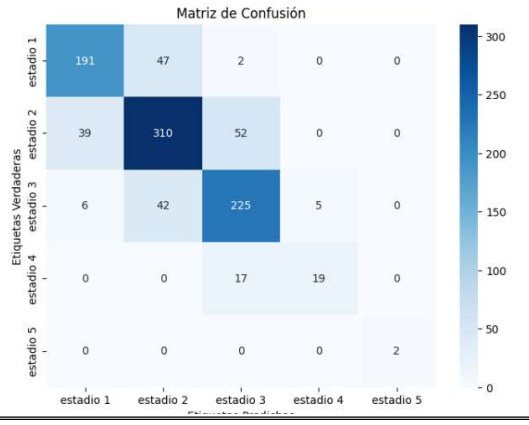
	precision	recall	f1-score	support
estadio 1	0.79	0.80	0.79	240
estadio 2	0.75	0.76	0.75	401
estadio 3	0.75	0.78	0.76	278
estadio 4	0.80	0.33	0.47	36
estadio 5	1.00	1.00	1.00	2
accuracy			0.76	957
macro avg	0.82	0.73	0.76	957
weighted avg	0.76	0.76	0.76	957



El Bosque Aleatorio clasificó a un total de 191,305,217 individuos en los estadios 1, 2, 3, 4 y 5, logrando identificar correctamente 664,453,604,918 individuos como verdaderos positivos y 953 como verdaderos negativos. Estos resultados se traducen en valores elevados de F1 Score para los estadios 1, 2 y 3, que se mantienen estables. Además, para el estadio 5, el modelo no clasificó ningún falso negativo ni falso positivo, alcanzando un recall y precisión del 100%. para el estadio 4 y 5 obtiene un comportamiento favorable con respecto a los otros modelos.

No. 4 – XG Boost

	precision	recall	f1-score	support
estadio 1	0.81	0.80	0.80	240
estadio 2	0.78	0.77	0.77	401
estadio 3	0.76	0.81	0.78	278
estadio 4	0.79	0.53	0.63	36
estadio 5	1.00	1.00	1.00	2
accuracy			0.78	957
macro avg	0.83	0.78	0.80	957
weighted avg	0.78	0.78	0.78	957

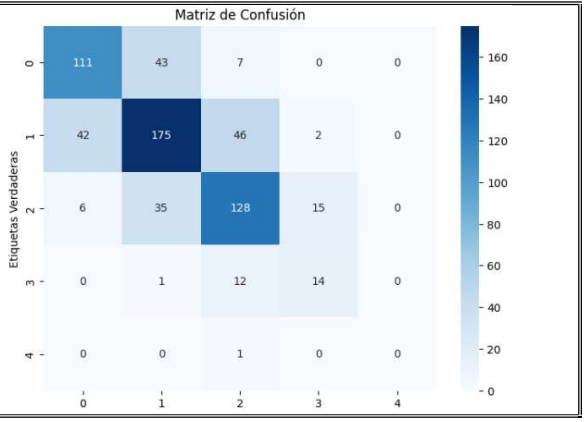


El desempeño general del modelo XG Boost muestra un accuracy del 78%, logrando identificar correctamente el mayor número de verdaderos positivos en los estadios 4 (916) y 5 (955). Para los estadios 1, 2 y 3, se obtienen valores de recall mayores al 77%, lo que significa que el modelo identifica la mayoría de los casos positivos y tiene muy pocos casos de falsos negativos. En el caso del estadio 5, el modelo no clasifica falsos negativos ni falsos positivos. En general, el comportamiento de Recall y Precisión es muy similar para todas las clases, lo que resulta en un F1 Score estable.

No. 5 - CatBoos

Reporte de Clasificación:

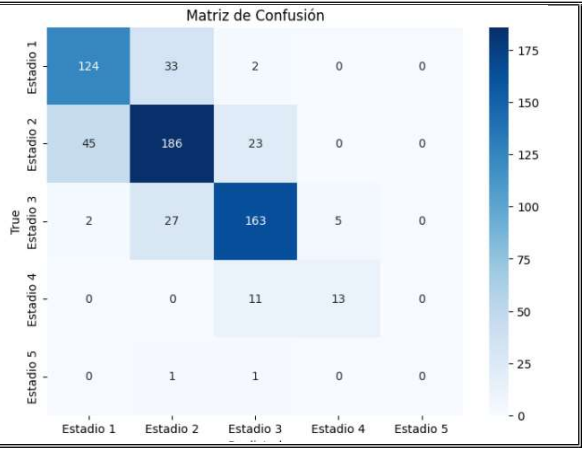
	precision	recall	f1-score	support
Estadio 1	0.70	0.69	0.69	161
Estadio 2	0.69	0.66	0.67	265
Estadio 3	0.66	0.70	0.68	184
Estadio 4	0.45	0.52	0.48	27
Estadio 5	0.00	0.00	0.00	1
accuracy			0.67	638
macro avg	0.50	0.51	0.51	638
weighted avg	0.67	0.67	0.67	638



El modelo CatBoost clasificó a 14 individuos como estadio 4 de un total de 31 casos reales, obteniendo una precisión del 45%, siendo la más baja después del estadio 5, el cual, tanto en el recall como en la precisión, obtiene un porcentaje del 0%. Para el resto de los estadios, las métricas de precisión, recall y F1 Score se encuentran por encima del 65%. En conjunto, el modelo alcanza un accuracy general del 83%

No. 6 – Red Neuronal Corn

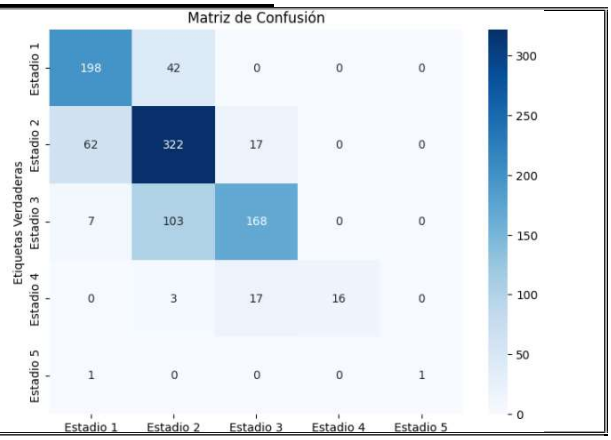
	precision	recall	f1-score	support
Estadio 1	0.73	0.78	0.75	159
Estadio 2	0.75	0.73	0.74	254
Estadio 3	0.81	0.83	0.82	197
Estadio 4	0.72	0.54	0.62	24
Estadio 5	0.00	0.00	0.00	4
micro avg	0.76	0.76	0.76	638
macro avg	0.60	0.58	0.59	638
weighted avg	0.76	0.76	0.76	638



La red neuronal convolucional clasificó a 186 individuos como estadio 2 de un total de 247 casos reales, obteniendo una precisión del 75% y un recall del 73%, lo que se traduce en un F1-Score del 74%. Para los estadios 1 y 2, las métricas de precisión, recall y F1-Score se encuentran por encima del 73%. Para el estadio 4, el resultado de recall disminuye a un 54%, donde no se obtienen métricas para el estadio 5, ya que el recall y la precisión son del 0%, clasificando solo verdaderos negativos (634) y 2 falsos negativos, teniendo un comportamiento desfavorable en esta clase. En conjunto, el modelo alcanza un accuracy general del 76%

No. 7 - Red *neuronal feedforward*:

	precision	recall	f1-score	support
Estadio 1	0.74	0.82	0.78	240
Estadio 2	0.69	0.80	0.74	401
Estadio 3	0.83	0.60	0.70	278
Estadio 4	1.00	0.44	0.62	36
Estadio 5	1.00	0.50	0.67	2
accuracy			0.74	957
macro avg	0.85	0.64	0.70	957
weighted avg	0.75	0.74	0.73	957



La red neuronal FeedForward realiza una clasificación efectiva para todos los estadios, clasificando 198, 322, 168, 16 y 1 casos como verdaderos positivos, con una precisión mayor al 69%. En el estadio 4, obtiene un recall del 44%, siendo el más bajo en este modelo con respecto a las demás clases. El F1 Score es estable, siendo mayor al 62% para todas las clases