Implementación De Modelos De Aprendizaje Automático Para La Prevención De Enfermedades Renales (ERC) O Sus Derivadas

Rafael Humberto Rodriguez Barrios, ✉ rafaelh.rodriguezb@utadeo.edu.co.com

Tesis de Maestría presentada para optar al título de Magíster en Ingeniería y Analítica de Datos

Asesor:
Fran Ernesto Romero Álvarez,

Universidad Jorge Tadeo Lozano
Facultad de ciencias naturales e ingeniería
Maestría en ingeniería y analítica de datos
Bogotá D.C., Colombia
2020

AGRADECIMIENTO

A todos mis seres queridos que me apoyaron desde niño y hasta hoy, pero especialmente, a mi madre Ismenia, mi esposa Niyiret, y mis hijas Heliana Isabel y Saray Sofia, porque una Maestría la hace el esfuerzo de toda una familia.

A mi asesor Fran Romero, por su confianza y deseo de trabajo en conjunto.

A la nefróloga Sandra Castelo, por su apoyo y colaboración en el desarrollo de esta investigación.

A la Universidad Jorge Tadeo Lozano, y finalmente,

A Dios, por darme la sabiduría y la capacidad para poder alcanzar este meta, gracias a la vida, porque ser Magíster, es un privilegio en este mundo, y especialmente en Colombia, por eso, el compromiso de retribuir todo lo aprendido

para el beneficio de nuestra sociedad.

DEDICADO

A todos los pacientes con ERC y sus familias

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

GLOSARIO

Aprendizaje Automático

El Aprendizaje Automático es un paso previo a la Inteligencia Artificial (IA) para que ésta se desarrolle en su totalidad. Concretamente, es un aprendizaje de máquinas que utilizan muchos datos con el objetivo de ser cada vez más inteligentes a la hora de realizar evaluaciones. El Aprendizaje Automático se alimenta de algoritmo e información, es algo así como llevar la inteligencia al dato. El Aprendizaje Automático está unido a la analítica, ésta es la que le dice a la máquina que cierto comportamiento es adecuado o no.

Conjunto de datos

Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones.

Confianza

Es la probabilidad de acierto que calcula el sistema para cada una de las predicciones.

Experimento

Es como se identifica un procedimiento de Aprendizaje Automático dentro de la plataforma *Azure de Microsoft.*

Algoritmos

Son una serie de pasos matemáticos u operacionales específicos para resolver un problema o realizar una tarea. En el contexto del Aprendizaje Automático, un algoritmo transforma o analiza datos para llevar a cabo las tareas de Análisis y Clasificación.

Modelo

La definición de un modelo es la representación matemática de las relaciones en un conjunto de datos. O lo que es lo mismo, es una forma simplificada y matemáticamente formalizada de aproximarse a la realidad y hacer predicciones a partir de esta aproximación.

Variables

Son elementos o dimensiones de un conjunto de datos; por lo tanto, elegir las informativas, discriminatorias e independientes es un paso crucial para lograr algoritmos efectivos.

Supervisión vs No Supervisión

El Aprendizaje Automático puede tener dos enfoques fundamentales. Por una parte, el Aprendizaje Supervisado**,** que es una forma de enseñar a un algoritmo cómo hacer su trabajo cuando tiene un conjunto de datos para los que sabe su respuesta.

Por otra parte, se denomina Aprendizaje no Supervisado cuando un algoritmo analiza el dato que no ha sido etiquetado con una respuesta para identificar patrones o correlaciones.

Red Neuronal

Una Red Neuronal es un modelo de computación cuya estructura de capas se asemeja a la estructura interconectada de las neuronas en el cerebro, con capas de nodos conectados. Una Red Neuronal puede aprender de los datos, de manera que se puede entrenar para que reconozca patrones, clasifique datos y pronostique eventos futuros.

Bosque de Decisiones

Es un modelo de Aprendizaje Automático utilizado para resolver problemas de regresión y clasificación. Consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto y pueden producir predicciones más precisas que cualquiera de las partes individuales.

Regresión Logística

Es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurrido como función de otros factores.

Jungla de Decisiones

Son una extensión reciente de los Bosques de Decisión. Una selva de decisión consiste en un conjunto de decisiones basadas en un grafo dirigido y acíclico.

Bagging

También conocido como empaquetado, es un procedimiento diseñado para mejorar la estabilidad y precisión de algoritmos de aprendizaje automático usados en clasificación. También sirve para reducir la varianza de aquellos algoritmos que tienen una alta varianza.

Diálisis

La diálisis es un proceso mediante el cual se extraen las toxinas y el exceso de agua de la sangre, que se utiliza como terapia renal sustitutiva tras la pérdida de la función renal en personas con falla renal.

Smote

Es una técnica estadística de sobre muestreo de minorías sintéticas para aumentar el número de casos de un conjunto de datos de forma equilibrada. El módulo funciona cuando genera nuevas instancias a partir de casos minoritarios existentes que se proporcionan como entrada.

RESUMEN

La Enfermedad Renal Crónica (ERC) es un problema de salud global con una alta tasa de morbilidad y mortalidad, que induce a padecer otras enfermedades. Dado que no hay síntomas visibles durante las primeras etapas de la ERC, los pacientes a menudo no notan la enfermedad. La detección temprana de la ERC permite a los pacientes recibir un tratamiento oportuno para mejorar la progresión de esta enfermedad. Los modelos de Aprendizaje Automático pueden ayudar eficazmente a los médicos a lograr este objetivo debido a su rendimiento de reconocimiento rápido y preciso. En este estudio, proponemos una metodología de Aprendizaje Automático para el diagnóstico de ERC.

El conjunto de datos que se utilizó fue otorgado por la Clínica Renal Colombiana, gracias a la colaboración de la nefróloga Sandra Castelo. Estos datos en su totalidad fueron anonimizados. Como guía de referencia se usó el modelo *CRISP-DM®* [1]. Los datos se trabajaron en su totalidad en la nube en la plataforma de *Azure* desde ahí se le realizaron los procesos para la exploración y el análisis, donde se encontró que los datos de las muestras estaban desbalanceados. Por lo que se utilizó la técnica *SMOTE* para balancear los datos.

Después de completar de manera efectiva el balanceo de datos, se utilizaron cuatros algoritmos de Aprendizaje Automático (Regresión Logística, Bosque de Decisión, Red Neuronal, y Jungla de Decisiones). Entre estos modelos de Aprendizaje Automático, el de Bosque de Decisión logró el mejor rendimiento con un 92%, que soporta una buena línea base para soluciones en producción en el enfoque empleado en esta investigación.

1. INTRODUCCIÓN

Según un informe de *Global Burden Disease* [2], la Enfermedad Renal Crónica (ERC) es de las causas de muerte que más ha aumentado en los últimos años. Uno de cada siete adultos sufre ERC, una de las patologías más desconocidas, sin embargo, de mayor impacto en la calidad de vida de los pacientes y que aumenta exponencialmente el riesgo de muerte.

La Enfermedad Renal Crónica (ERC) ha sido considerada en el Sistema General de Seguridad Social en Salud (SGSSS) [3] como una patología de alto costo, por generar un fuerte impacto económico sobre las finanzas del Sistema. Causando un dramático efecto sobre la calidad de vida del paciente y su familia, incluidas repercusiones laborales.

Para reducir la alta mortalidad de la ERC se debe profundizar en la investigación y dirigirla a los estadios iniciales de la enfermedad analizando su grupo de riesgo, con la ayuda de exámenes de laboratorio, buscando que los pacientes no lleguen a las etapas finales como son: diálisis, trasplante o muerte.

Se busca hallar mediante el Aprendizaje Automático, un aporte valioso para que de forma temprana se pueda realizar una clasificación de la enfermedad en sus etapas iniciales por medio de los resultados de los laboratorios clínicos aprovechando el gran potencial del Aprendizaje Automático, en el análisis y clasificación de los datos.

2. DESCRIPCION DEL PROBLEMA

En Colombia entre el 1 de julio del 2013 y el 30 de junio del 2014 fueron reportados 3.055.568 casos de personas con diagnóstico de enfermedad renal en cualquiera de sus estadios, 1.406.364 personas tienen ERC en etapa final, mostrando una tasa de mortalidad de 28.19% por cada cien mil personas [3].

Teniendo en cuenta las cifras descritas anteriormente y que por cada paciente en diálisis existen 18 personas con algún grado de probabilidad de sufrir una patología renal [3], estos hechos generan preocupación en el Sistema Nacional de Salud en Colombia, ya que estas cantidades pueden aumentar de gran manera en razón a que el mayor porcentaje del gasto público se dedica solo a financiar la atención médica mientras que, para prevención el presupuesto es mucho menor. Dicha situación se agrava por la ausencia del estado en la mayoría de las regiones. Por otro lado, los tiempos de espera en la asignación de la primera cita con un especialista, con lleva a que el tiempo necesario para un diagnóstico definitivo o tratamiento de la enfermedad pueda tomar de 3 a 6 meses [4].

El panorama descrito se traduce en enfermedades crónicas con procedimientos de alto costo y con un enorme sufrimiento para los pacientes. Partiendo del hecho que si las enfermedades renales son diagnosticadas en sus primeras etapas pueden recibir tratamientos a bajo costo y con una mejor calidad de vida para el paciente.

Por todo esto, se hace necesario la ayuda de herramientas tecnológicas que basadas en datos puedan soportar el proceso de toma de decisión en los diagnósticos iniciales de manera rápida, con alta precisión y a bajo costo. Con ellos, se reduce el tiempo requerido para el diagnóstico permitiendo al paciente recibir tratamiento de la enfermedad antes de que este avance a una etapa de no retorno.

3. OBJETIVOS

3.1. Objetivo general

Diseñar e implementar un modelo de Aprendizaje Automático que, a partir de los datos provenientes de laboratorios clínicos, permita predecir el posible diagnóstico de la ERC en sus etapas iniciales, contribuyendo a disminuir la tasa de mortalidad y costos para el sistema de salud.

3.2. Objetivos específicos

1. Realizar el proceso de recopilación y análisis exploratorio de los datos de laboratorios clínicos para identificar las variables a usar en la construcción del modelo de aprendizaje para la clasificación de ERC.

2. Llevar a cabo un proceso pre-procesamiento de los datos obtenidos, para mejorar la calidad de los mismos y su utilidad para la generación de modelos.

3. Seleccionar y aplicar las técnicas de Aprendizaje Automático que permitan predecir la ocurrencia de ERC.

4. Evaluar los diferentes modelos para establecer el más adecuado para predecir la ocurrencia de ERC.

4. PREGUNTA DE INVESTIGACIÓN

¿Cómo es posible implementar un modelo que, a partir de la aplicación de la técnica de Aprendizaje Automático, permita predecir la ocurrencia de ERC en sus etapas iniciales?

5. MARCO TEORICO

5.1 Enfermedad Renal Crónica

La Enfermedad Renal Crónica (ERC) es el estado clínico-patológico que puede conducir a una enfermedad renal a etapa terminal, cualquiera que sea la naturaleza del proceso fisiopatológico, desde enfermedades genéticas o inmunes específicas (como la Nefritis Lúpica) hasta lesiones más sistémicas (como la Nefropatía Diabética). La ERC se asocia con una disminución de la función renal la cual puede estar relacionada con la edad y se encuentra acelerada por la hipertensión, diabetes, obesidad y trastornos renales primarios [3].

La ERC es una condición que representa una elevada carga para el paciente, la familia, la sociedad y el sistema de salud. A nivel mundial, es la sexta causa de muerte de dinámico crecimiento que afecta a cerca del 10% de la población. Se estima que 850 millones de personas en el mundo padecen de enfermedad renal y esta es responsable de al menos 2,4 millones de muertes al año, mientras que la lesión renal aguda, importante impulsor de la ERC, afecta a más de 13 millones de personas en el mundo [2].

Para detectar la ERC existen varias pruebas, una de ellas es la Prueba de Creatinina. En esta se analiza la sangre en busca de un producto de desecho llamado creatinina, proviene del tejido muscular, cuando los riñones están dañados tienen dificultad para eliminar la creatinina de la sangre. Pero el análisis de creatinina es solo el primer paso a continuación, se usa el resultado de creatinina en una fórmula matemática para averiguar el índice de Filtración Glomerular (IFG) el número de IFG indica a su proveedor de atención médica la capacidad de funcionamiento de sus riñones [5] .

Para el periodo comprendido entre el 1º de julio de 2018 y el 30 junio de 2019, se reportó a la Cuenta de Alto Costo (CAC) de la presidencia de la república de Colombia, la información de 4.539.694 personas con ERC, Fig. 1. Se encontró que el 61% fueron mujeres, el 39% hombres y el promedio de edad para el total de la población reportada fue de 24-64 años.

Fig. 1. Personas reportadas al SGSS Enfermedad Renal Crónica según el régimen de afiliación, Colombia 2019 [6]

5.2 Aprendizaje Automático en la Nefrología

El Aprendizaje Automático, es un tipo de Inteligencia Artificial. La base del Aprendizaje Automático son los métodos algorítmicos, que permiten a la máquina resolver problemas sin una programación informática específica. La amplia aplicación del Aprendizaje Automático en el campo médico está ayudando a promover la innovación médica, reduciendo sus costos y mejorando la calidad en el diagnóstico médico.

El Aprendizaje Automático ayuda a las computadoras a poseer la misma capacidad de aprender, identificar y clasificar que los seres humanos [7]. Sin embargo, la investigación relacionada para resolver problemas clínicos a través del Aprendizaje Automático en la nefrología aún necesita mayor desarrollo.

El Aprendizaje Automático en términos generales se puede dividir en Aprendizaje Supervisado, Aprendizaje no Supervisado y Aprendizaje de Refuerzo [8]. El Aprendizaje Supervisado, es la forma más común de Aprendizaje Automático que se utiliza en las investigaciones del área médica [4]. Cada instancia de Aprendizaje Supervisado contiene un objeto de entrada (generalmente un vector) y un valor de salida deseado (también conocido como señal supervisada) [4]. Habitualmente entre los algoritmos que se aplican para el Aprendizaje Supervisado se

encuentran: Árboles de Decisión, Clasificación Ingenuo de Bayes, Regresión por Mínimos Cuadrados, Regresión Logística, Maquina de Soporte Vectorial (SVM), Métodos (Conjuntos de Clasificadores).

El Aprendizaje no Supervisado no contiene información de categoría, ni dispone de datos etiquetados para el entrenamiento. El Aprendizaje no Supervisado divide de manera óptima las muestras en diferentes categorías según las características de los datos de entrenamiento sin las etiquetas correspondientes [6]. Además, el Aprendizaje no Supervisado puede capturar patrones morfométricos intrínsecos en secciones de histología, que pueden desempeñar un papel clave en el diagnóstico patológico [8]. En el futuro, es probable que el Aprendizaje no Supervisado reduzca la brecha entre la Inteligencia Humana y la Inteligencia Artificial.

El Aprendizaje Profundo es un tipo específico de método del Aprendizaje Automático y sirve para procesar grandes cantidades de datos como entrada sin la necesidad de un paso claro de selección de características. Se puede entrenar para encontrar patrones complejos en *Big Data* con un alto grado de precisión [9]. El Aprendizaje Profundo en este estudio no se utiliza porque no se cuenta con una gran cantidad de datos, pero en una mejora de este podría ser muy útil.

Estudios recientes demuestran que las Redes Neuronales Profundas han logrado gran desempeño comparable a nivel de expertos en tareas de clasificación de imágenes biomédicas y naturales [10]. Esto sumado a la capacidad de generar suposiciones [11], la adaptabilidad al análisis heterogéneo de conjuntos de datos y los programas de Aprendizaje Profundo de código abierto de rápida difusión, está logrando que el Aprendizaje Profundo desempeñe un papel importante en la promoción del desarrollo médico [12].

Las Redes Neuronales Convolucionales (RNC) han ganado fuerza con el desarrollo del procesamiento de imágenes, en la clasificación de conjuntos de datos [13]. El proceso de construcción es similar al mecanismo de percepción visual de los organismos, gracias a esto el aprendizaje puede ser supervisado y sin supervisión. Estas redes de Aprendizaje Profundo son utilizadas en el diagnóstico y el tratamiento de diferentes enfermedades. También se pueden volver

a entrenar a través de conjuntos de datos específicos de la población [9, p. 11]. Las RNC han superado el rendimiento humano en el reconocimiento visual de ciertos objetivos [9].

Con el constante crecimiento de los datos digitales, relacionados a la prestación de los servicios en la atención médica y en el desarrollo de la Inteligencia Artificial, se ha logrado que el Aprendizaje Automático, basado en conjuntos de datos clínicos apropiados permitan llegar al diagnóstico de enfermedades [12]. Aplicando diferentes métodos en medicina y biología computacional, también ha presentado un excelente desempeño en el campo del análisis de imágenes médicas y la genómica. [10, p. 17].

A pesar del uso efectivo del Aprendizaje Automático en la medicina, la falta de evidencia y el poco alcance de la investigación en la enfermedad renal han llevado al hecho de que la nefrología aún no se beneficie ampliamente de esta, a diferencia en la enfermedad cardiovascular, el *Big Data* y el Aprendizaje Automático se encuentra totalmente madurado [8]. En el futuro la combinación del Aprendizaje Automático y *Big Data*, será un factor importante en la promoción de la medicina ayudando a mejorar la precisión en el estudio de la patología renal y la precisión del riesgo de enfermedad renal.

5.3 Regresión Logística

La Regresión Logística, gracias a que es un tipo de análisis de regresión, permite predecir el resultado de una variable categórica en función de las variables independientes o predictores. Este se basa en la regresión lineal y obtiene el peso de cada predictor y un sesgo. Si la suma de los efectos de todos los predictores excede un umbral, la categoría de la muestra se clasificará como uno de los valores de Estadio.

5.4 Redes Neuronales Artificiales

Las Redes Neuronales puede analizar relaciones no lineales en los conjuntos de datos debido a su estructura compleja, logrando aprender y formarse a sí mismos, en lugar de ser programados de forma explícita logrando escoger el estadio de mejor referencia.

5.5 Bosques de Decisión

El algoritmo de los Bosques de Decisión genera una gran cantidad de árboles de decisión mediante el muestreo aleatorio del entrenamiento y sus predictores. Cada árbol de decisión está entrenado para encontrar un límite que maximice la diferencia entre los estadios de la ERC. La decisión final está determinada por las predicciones de todos los árboles en el diagnóstico de la enfermedad. Máquinas de Soporte Vectorial (SVM) divide diferentes tipos de muestras estableciendo una superficie de decisión en un espacio multidimensional que comprende los predictores de las muestras.

5.6 Jungla de Decisiones

Jungla de Decisiones permite crear un modelo de clasificación multiclase, utilizando algoritmos haciendo que el modelo entrenado se utiliza para predecir un objetivo que tiene varios valores.

5.7 Azure ML Studio

Azure Machine Learning, es un entorno basado en la nube usado para entrenar, implementar, automatizar, administrar y realizar un seguimiento de los modelos de ML [14]. En el campo de la salud, su aporte a logrado maximizar la atención al paciente, minimizando los costos, con las características de Aprendizaje Automático más avanzadas logrando soluciones que pueden analizar imágenes, reconocer la voz, hacer predicciones con datos de laboratorios clínicos. En el desarrollo de este proyecto se utilizará Azure Machine Learning para la exportación de los datos, análisis de estos y creación de los algoritmos.

5.8 Power BI

Es una solución destinada a la inteligencia empresarial, que permite unir diferentes fuentes de datos [15] en este proyecto se importaron los datos a una base de *SQL Server* en el entorno *Azure Machine* la cual se conectó a *Power BI.* Donde se realiza el análisis de los datos para presentarlos de una

manera fácil y atractiva. Uno de los resultados de las gráficas se obtuvo ingresando lenguaje *Python* dentro de *Power BI*.

5.9 CRISP-DM

En la ejecución de proyectos de minería de datos se debe realizar procesos que planifiquen y que guíen su desarrollo. *CRISP-DM*® [1](Proceso estándar entre industrias para la minería de datos) permite el desarrollo de los proyectos, guiado por una serie de etapas relacionadas entre sí, logrando contemplar el proceso de análisis de datos como un proyecto profesional, y estableciendo así un contexto más rico que influye en la elaboración de los modelos.

El CRISP-DM Se divide en varias fases como se logra apreciar en la Fig. 2.

Fig. 2. Guía visual de la metodología *CRISP-DM Fuente: Elaboración propia.*

Fase I. Definición de necesidades del cliente (comprensión del negocio)

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase II. Estudio y comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos iniciales y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad,

descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos a partir de los datos en bruto iniciales.

Fase IV. Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema y se calibran sus parámetros a valores óptimos.

Fase V. Evaluación (obtención de resultados)

En esta fase del proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.

6. ESTADO DE ARTE

El interés en la aplicación de la técnica de Aprendizaje Automático en el área médica se ha incrementado desde hace más de tres décadas. Fue demostrado en la Conferencia Internacional de Minería de Datos realizada en el 2012, en California (EU) donde las aplicaciones médicas predominaron considerablemente.

Es así como en este mundo globalizado y con tendencia al uso de la tecnología en todos los campos, se evidencia que el Aprendizaje Automático es uno de esos pilares sobre los que avanza la aplicación del uso de herramientas digitales. Para obtener grandes beneficios en los problemas cotidianos de la humanidad, ayudando a pasar de ser seres reactivos a ser seres proactivos.

Directamente en el campo médico, son muchas las aplicaciones que se han realizado de técnicas de Aprendizaje Automático en predicciones clínicas basadas en datos. Investigaciones demuestran la importancia y fuerza que está tomando estas técnicas para el diagnóstico, tratamiento y pronóstico médico, aprovechando los grandes volúmenes de datos que actualmente se tienen.

Gracias a estos resultados se ha logrado aumentar por parte de la comunidad médica la confianza en este tipo de herramientas, logrando encontrar en la literatura médica estudios para detección de diferentes tipos de cáncer, desde las puntuaciones de riesgo para guiar la anticoagulación y el uso de medicamentos para el colesterol hasta la estratificación de riesgo de los pacientes en la unidad de cuidados intensivos. Hoy se cuenta con grandes comunidades dedicadas a alimentar repositorios para contribuir al aprendizaje y evaluación de técnicas de Aprendizaje Automático. Un ejemplo, lo constituye el caso específico de la Universidad de California, *Irvine* la cual Actualmente cuenta con 497 conjuntos de datos como un servicio para la comunidad de Aprendizaje Automático, que incluye resultados de biopsias de lesiones mamarias, predictores de cardiopatía, registros de supervivencia posquirúrgica y otras de diversas disciplinas para aprender.

A continuación, se relacionan los estudios más recientes en los que se han utilizado técnicas de Aprendizaje Automático, en prevención de la Enfermedad Renal Crónica.

Comparación y desarrollo de herramientas de aprendizaje automático en la predicción de la progresión de la ERC [16]**.**

Este estudio fue realizado entre agosto de 2015 a septiembre de 2018 en el Departamento de Nefrología en *Shanghai*, el objetivo de este estudio fue pronosticar rápidamente la gravedad de la ERC utilizando características demográficas y bioquímicas de la sangre. En él desarrollan y comparan varios modelos predictivos utilizando enfoques estadísticos, de aprendizaje automático y de redes neuronales. Para lograrlo utilizaron una muestra de 551 pacientes diagnosticados con ERC con proteinuria, de estos pacientes se tomaron 13 factores derivados de la sangre y 5 características demográficas que fueron usadas como variables, y con ellos se establecieron y compararon nueve modelos predictivos (Regresión Logística, Red Elástica, Regresión de Lazo, Regresión de Crestas, Maquina de Vectores de Soporte, Bosque de Decisiones Aleatorias, *XGBoost* (Aumento de Gradiente Extremo).

La conclusión de este estudio fue el desarrollo de una herramienta web *CKD* (Sistema de Predicción) para la práctica clínica que se puede utilizar ampliamente en la evaluación del progreso de la proteinuria en la nefrología y durante los exámenes de seguimiento.

Diagnóstico de la ERC basado en Máquinas de vectores de soporte por métodos de selección de características [17]**.**

En este estudio realizado en el Departamento de Ingeniería Informática de la Facultad de Tecnología de la universidad *Gazi* de Ankara Turquía, se utilizó el algoritmo de clasificación de la Máquina de Soporte Vectorial para diagnosticar la enfermedad renal crónica. Los datos fueron tomados del repositorio de aprendizaje automático UCI de pacientes ERC. Este conjunto de datos está formado por 24 características que excluyen el atributo de clase. Incluyeron un total de 400 instancias, en 250 con CKD y 150 instancias sin CKD. Se utilizaron 2 métodos de selección de características: El primero fue método de envoltura, el segundo fue el método de filtro para reducir la dimensión del conjunto de datos de la enfermedad renal crónica. En el método de envoltura, se utilizó un evaluador de subconjunto clasificador con un método de búsqueda por pasos y un evaluador de subconjunto de envoltura con el método de búsqueda El mejor primero. Los métodos

de envoltura y filtro basados en la búsqueda escalonada El mejor primero y Algoritmo Voraz se desarrollaron para evaluar los métodos de selección de características y la precisión de los algoritmos de clasificación. Con el método de filtro, se utilizó el evaluador de subconjunto de selección de características de correlación con un método de búsqueda gradual y un evaluador de subconjunto filtrado con el método de búsqueda El mejor primero.

Los resultados mostraron que el clasificador de la Máquina de Soporte Vectorial mediante el uso del evaluador de subconjuntos filtrados con el método de selección de características del método de búsqueda Primero el Mejor tiene una tasa de precisión más alta (98.5%) en el diagnóstico de enfermedad renal crónica en comparación con otros métodos seleccionados.

Estratificación de la ERC mediante registros de visitas al consultorio: manejo del desequilibrio de datos mediante Meta clasificación jerárquica [18]**.**

Para este estudio se recopilaron los datos de 13,111 pacientes de la ciudad *Delaware*, durante las visitas a atención primaria y prácticas especializadas. El conjunto de datos consta de 120,739 registros que comprenden información del paciente almacenado en el registro médico electrónico (EHR), el conjunto de datos comprende todos los registros de pacientes diagnosticado con ERC en las etapas 3,4 y 5. De los registros iniciales solo se usaron 93,218 que contenían toda la información completa. Como primera medida utilizaron varios métodos de clasificación comunes como referencia para la comparación (Ingenuo de Bayes, Regresión Logística, Árbol de Decisión y Bosques Aleatorios). Utilizando la estrategia de uno contra todos. para entrenarlos implementaron Scikit-learn que es la herramienta básica para desarrollar Ciencia de Datos en *Python*, orientado a objetos de alto nivel y ampliamente utilizado en aplicaciones de Inteligencia Artificial.

El método que propusieron fue Meta-clasificadores jerárquicos, desarrollando un enfoque jerárquico de meta-clasificación para asignar una etapa de ERC (en el rango 3-5) a un registro del paciente ante un desequilibrio de datos elevado.

Como conclusión se destaca el desarrollo enfocado en la clasificación por conjuntos basado en meta-clasificación jerárquica, para identificar etapas de ERC a partir de un conjunto de datos

altamente desbalanceado. El clasificador de conjunto basado en muestreo alcanzó una mayor sensibilidad con respecto a las etapas 4 y 5 en comparación con la obtenida por otros métodos, los resultados fueron exitosos logrando un porcentaje del 93% superando a los clasificadores de referencia. Los meta-clasificadores simples y los enfoques informados previamente para abordar el desequilibrio en la identificación de cada una de las dos etapas avanzadas de ERC (etapa 4 y etapa 5), manteniendo su alto nivel de rendimiento cuando el número de registros se trunca significativamente, lo que demuestra su estabilidad y generalización.

Algoritmo de Aprendizaje Automático para la detección temprana de la Enfermedad Renal en etapa terminal [19]

Este estudio analizó 10,000,000 reclamos de seguros médicos de 550,000 registros de pacientes utilizando una base de datos comercial de seguros médicos de pacientes de una de las mayores compañías de seguros de salud con sede en los Estados Unidos desde el 1 de enero de 2006 hasta el 31 de diciembre de 2018. Los criterios de inclusión fueron pacientes mayores de 18 años diagnosticados de ERC estadios 1-4. Como las principales etiologías subyacentes de la ERC son la diabetes y la hipertensión, se excluyeron los pacientes cuyas afecciones subyacentes fueran glomerulopatías agudas, anomalías congénitas o lesión renal traumática, porque el curso de la enfermedad en estas afecciones es diferente y puede interferir con la interpretación de los resultados. Se compilaron 240 candidatos a predictores, divididos en seis grupos de características: datos demográficos, afecciones crónicas, características de diagnóstico y procedimiento, características de medicamentos, costos médicos y recuento de episodios.

Se uso un método de incorporación de características basado en la implementación del algoritmo *Word2Vec* para capturar más información temporal para los tres componentes principales de los datos: diagnóstico, procedimientos y medicamentos. Para el análisis, se utilizó el algoritmo de árbol de decisión *Gradient Boosting*.

El modelo, fue basado en análisis de *big data*, mostrado valores predictivos muy altos con C-Estadística de 0,93, sensibilidad de 0,715 y especificidad de 0,958. El valor predictivo positivo (VPP) fue de 0,517 y el valor predictivo negativo (VPN) fue de 0,981.

Como conclusión se determina que este modelo logró mejores resultados en todas las métricas en la que fue probado, este modelo puede ser usado por organizaciones de mantenimiento de la salud y los hospitales, para que cuando un paciente se acerca al umbral de riesgo de ERC, se pueda enviar un mensaje de advertencia electrónicamente al médico para iniciar una derivación para una consulta de nefrología.

Una metodología de Aprendizaje Automático para diagnosticar la ERC [20]

En este estudio, se propuso una metodología de Aprendizaje Automático para el diagnóstico de ERC. El conjunto de datos de CKD se obtuvo del repositorio de Aprendizaje Automático de la Universidad de California en Irvine (UCI), que contaba con una gran cantidad de valores faltantes. Se utilizó la imputación KNN para completar los valores faltantes, que selecciona varias muestras completas con las medidas más similares para procesar los datos faltantes para cada muestra incompleta. Después de completar de manera efectiva el conjunto de datos incompleto, se utilizaron seis algoritmos de Aprendizaje Automático (Regresión Logística, Bosque Aleatorio, Máquina de Vectores de Soporte, Vecino más Cercano K, Clasificador de Bayes Ingenuo y Red Neuronal de Avance). Entre estos modelos de Aprendizaje Automático, el Bosque Aleatorio logró el mejor rendimiento con una precisión de diagnóstico del 99,75%. Al analizar los errores de juicio generados por los modelos establecidos, se propuso un modelo integrado que combina Regresión Logística y Bosque Aleatorio mediante el uso de perceptrón, que podría alcanzar una precisión promedio de 99. 83% después de diez veces de simulación. Se especula que esta metodología podría ser aplicable a datos clínicos más complicados para el diagnóstico de ERC.

Uso de modelos de Aprendizaje Automático para predecir el inicio de la terapia de reemplazo renal en pacientes con ERC [21]

Este estudio exploró las posibilidades de crear modelos de pronóstico para predecir la aparición de TSR (hemodiálisis, diálisis peritoneal o trasplante renal) a los 3, 6 y 12 meses desde el momento del primer diagnóstico del paciente con ERC, utilizando solo los datos de comorbilidades del Seguro Nacional de Salud de Taiwán. Con una cantidad limitada de datos incluidas las comorbilidades, pero sin considerar los valores de laboratorio. Utilizando datos de 8.492 pacientes,

se obtuvo el área bajo la curva característica operativa del receptor (AUC) de 0,773 para predecir el TSR dentro de los 12 meses desde el momento del diagnóstico de ERC. Retrospectivamente, el investigador determina los individuos que estuvieron expuestos al diagnóstico de ERC por primera vez en cada uno de los grupos de estudio. El subconjunto seleccionado de datos incluyó 23,948 pacientes. Se filtraron aquellos con el diagnóstico de ERC, que tenía que ocurrir antes del TSR. El tiempo de TSR se determinó con la primera ocurrencia de hemodiálisis, diálisis peritoneal o trasplante renal. Esto resultó en 19,954 pacientes.

Los datos fueron preprocesados con cuatro métodos para determinar si el preprocesamiento podría mejorar los resultados. Selección de características, filtrado y reducción de dimensionalidad, se aplicaron individualmente antes del ML, mientras que el cuarto enfoque, el equilibrio de datos, fue implementado por los algoritmos ML. Se evaluaron 10 algoritmos de aprendizaje automático que se implementan en los paquetes de *Python Scikit-learn y XGBboost*: Árbol de Decisión, Arboles de Decisión de Ensacado, Bosque Aleatorio, *XGBoost,* Máquinas de Vectores de Soporte, Gradiente Descendente Simple, Vecinos más Cercanos, Bayes Ingenuos de Gaus , Regresión Logística y Red Neuronal. Cada experimento consistió en un algoritmo ML y una combinación de enfoques de preprocesamiento de datos (filtrado, selección de características, etc.). Los mejores resultados se obtuvieron con Regresión Logística en combinación con características de tiempo y equilibrio de datos, y sin selección de características, filtrado o reducción de dimensionalidad.

Como conclusión el estudio demostró la ventaja en el uso de comorbilidades para la predicción de TSR, adicional el estudio permite considerar que los algoritmos de ML serian una posible herramienta de detección para predecir el marco de tiempo de progresión del paciente con ERC antes de que necesite TRS, también se demostró que no hay ninguna ventaja adicional en centrarse solo en pacientes con diabetes en términos de rendimiento de predicción. Aunque estos resultados no son adecuados para su adopción en la práctica clínica, el estudio proporciona una base sólida y una variedad de enfoques para estudios futuros de modelos de pronóstico en la atención médica.

7. METODOLOGÍA

En el desarrollo de este proyecto se emplea el modelo *CRISP-DM®* el cual es la guía de referencia más amplia utilizada en el desarrollo de proyectos de analítica y minería, a datos recolectados desde laboratorios clínicos. Para ello, se implementará cada una de las etapas propuestas.

7.1 Fase I. Comprensión del negocio

Esta fase se divide en 4 tareas que ayudarán a tener una mejor comprensión del negocio como muestra la Fig. 3.

Fig. 3. Tareas de la comprensión del negocio.

7.1.1 Determinación de los objetivos de negocio.

La enfermedad renal crónica (ERC) es el estado clínico-patológico que puede conducir esta enfermedad a etapa terminal, cualquiera que sea la naturaleza del proceso fisiopatológico, desde enfermedades genéticas o inmunes específicas, hasta lesiones más sistémicas.

La ERC, se asocia con una disminución de la función renal la cual puede estar relacionada con la edad y se encuentra acelerada en la hipertensión, diabetes, obesidad y trastornos renales primarios.

La ERC, es un problema de salud global con alta tasa de morbilidad y mortalidad, e induce otras enfermedades. Como no hay síntomas evidentes durante las primeras etapas de la ERC, los pacientes a menudo no notan la enfermedad, siendo esta la característica principal, logrando que eventualmente se genere una pérdida completa de la función renal.

La detección temprana de la ERC permite a los pacientes recibir un tratamiento oportuno para mejorar la progresión de esta enfermedad. Como se ha planteado en los objetivos del trabajo, se busca desarrollar un modelo de Aprendizaje Automático para la predicción en el diagnóstico de ERC, para aportar en la disminución de complicaciones mayores en la enfermedad como procesos de diálisis, trasplante renal o llegar a la muerte.

El principal criterio de éxito para el presente proyecto, con la ayuda del Aprendizaje Automático, es poder identificar en etapas iniciales la ERC conductas o patrones de comportamiento para mejorar la calidad de vida de los pacientes.

7.1.2 Evaluación de la situación.

La idea para el planteamiento del presente proyecto nace de la situación actual sobre el aumento de diagnóstico confirmatorio de ERC, que por su mal tratamiento o por desconocimiento del usuario de sus patologías, derivan de forma irreversible a las etapas finales de la ERC como lo es diálisis de por vida, afectando financieramente el sistema de salud, al ser un tratamiento muy costoso que genera la mayor cantidad de absorción de los recursos dispuestos para la salud en Colombia. Los cuales se podrían reducir al utilizar herramientas como el Aprendizaje Automático en la clasificación de ERC desde las etapas iniciales.

Si bien la aplicación del Aprendizaje Automático en el cuidado de la salud y otras áreas es favorable, todavía en el campo de la enfermedad renal no se ha explotado todo su potencial [22].

7.1.3 Determinación de los objetivos de la minería de datos.

El objetivo en términos técnicos de este proyecto como se referencia en el objetivo general es diseñar, implementar y desplegar un modelo de Aprendizaje Automático que a partir de los datos procedentes de laboratorios clínicos permita clasificar la posibilidad de un diagnóstico de la ERC. Mediante el análisis de los estudios de laboratorios que son de bajo costo para las entidades de salud, con estos datos disminuir la tasa de mortalidad y costos del sistema de salud.

Los antecedentes médicos junto con los exámenes de laboratorio dan indicio a identificar síntomas o signos que puedan ser utilizados como variables constitutivas del problema en pacientes de ERC, en una gran escala ya que se pueden manejar gran cantidad de datos sin inconvenientes. Con los datos iniciales se realiza una descripción y exploración de estos, verificando que se puedan utilizar o que tengan la información mínima para realizar la clasificación, por medio del análisis de estos datos y obtener los pacientes con una incidencia de ERC.

Con los datos obtenidos se moldea un conjunto para entrenamiento, se realizan varias pruebas las cuales definen o determinan la o las técnicas más relevantes para utilizarlas en el clasificador y que los resultados sean eficaces y eficientes.

Con el clasificador definido se entrenan y validan los modelos predictivos para establecer el modelo con mayor precisión para los datos, seleccionando el que ofrezca mejores resultados. Los modelos predictivos a menudo ejecutan cálculos durante las transacciones en curso, por ejemplo, para evaluar el riesgo o la oportunidad de un paciente en particular, de forma que aporte conocimiento a la hora de tomar una decisión en el tratamiento.

7.1.4 Producir el plan del proyecto.

El plan del proyecto se podrá encontrar en el Anexo cronograma – Plan de trabajo Proyecto. En él se describen todos los pasos necesarios, desde el planteamiento del problema, recolección de datos, hasta el análisis de este.

7.2 Fase II: Estudio y comprensión de los datos

En esta sección se describen los datos iniciales obtenidos, tales como número de registros y campos por registros, su identificación, el significado de cada campo y la descripción del formato inicial, como muestra la Fig. 4.

Fig. 4. Tareas del Estudio y comprensión de los datos

7.2.1 Recolección de datos de partida.

El conjunto de datos utilizado para este proyecto se obtuvo gracias a la Clínica Renal Colombiana, y a su Gerente y Representante legal la Nefróloga Sandra Castelo, quien permitió y Autorizó el tratamiento de estos datos. El Conjunto de datos contiene 373.770 muestras anonimizadas. En este conjunto de datos, cada muestra tiene 17 variables o características predictivas (11 variables numéricas y 6 variables categóricas (nominales)).

7.2.2 Descripción de los datos.

Aquí se encontrará la descripción de los datos recopilados para el presente estudio, en la Tabla 1 las variables con sus definiciones y en la Tabla 2 la descripción de las variables.

Tabla 1. Definiciones

VARIABLE	DEFINICIÓN
Municipio	Código DANE de los municipios del país.
Edad	Rango de años de vida de una persona.
Género	Identificación de Sexo de una persona.
HTA (Hipertensión)	Es una enfermedad crónica que se caracteriza por el incremento continuo de las cifras de la presión sanguínea, por encima de los límites sobre los cuales aumenta el riesgo cardiovascular.
Albúmina	Es una proteína producida por el hígado. El examen de albúmina en suero mide la cantidad de esta proteína en la parte líquida y transparente de la sangre y en la orina.

Azúcar (Glucosa)	Es el azúcar base que se encuentra en la sangre.
Glóbulos Rojos	Son los encargados de transportar el oxígeno a través de la sangre.
Bacterias	Microorganismos perjudiciales en la sangre.
Urea Sangre	Es uno de los productos de desecho que los riñones eliminan de la sangre.
Creatinina	Es un producto de desecho generado por los músculos que se elimina a través de la sangre o en la orina.
Sodio	Es un tipo de electrolito. Los electrolitos son minerales con carga eléctrica que ayudan a mantener los niveles de líquido, y el equilibrio de sustancias químicas del cuerpo llamadas ácidos y base.
Potasio	Es uno de los muchos electrolitos del organismo. Se encuentra dentro de las células. Los niveles normales de potasio son importantes para el funcionamiento del corazón y el sistema nervioso.
Hemoglobina	Es una molécula adherida a los glóbulos rojos, que ayuda a transportar el oxígeno y el dióxido de carbono por el cuerpo, y es malo cuando aparece en la orina.
DM (Diabetes Mellitus)	Es una enfermedad crónica en la cual el cuerpo no puede regular la cantidad de azúcar en la sangre.
Anemia	Es una afección por la cual el cuerpo no tiene suficientes glóbulos rojos sanos, síntoma más común de la enfermedad renal crónica. [23]
Peso	Valor de la masa corporal.
Estadio	Es el número que se utiliza para determinar la etapa de la enfermedad renal de una persona.

Tabla 2. Descripción de las variables

NO	VARIABLE	DATATYPE	DESCRIPCIÓN	UNIDAD O MEDIDA
1	Municipio	Numérica	Código De Ciudad	Código Postales
2	Edad	Numérica	Edad	Años
3	Género	Nominal	Tipificación Masculino Y Femenino	M, F
4	HTA	Nominal	Hipertensión	Si, No (0,1)
5	Albunia	Numérica	Albúmina	0,1,2,3,4,5
6	Azúcar	Numérica	Azúcar	0,1,2,3,4,5
7	Glóbulos rojos	Nominal	Recuento De Glóbulos Rojos	Normal, Anormal
8	Bacterias	Nominal	Bacterias	Notpresent, Present
9	Urea sangre	Numérica	Urea En Sangre	Mgs/Dl
10	Creatinina	Numérica	Creatinina En Sangre	Mgs/Dl
11	Sodio	Numérica	Sodio	Meq/L
12	Potasio	Numérica	Potasio	Meq/L
13	Hemoglobina	Numérica	Hemoglobina	Gms
14	DM	Nominal	Diabetes Mellitus	Si, No (0,1)
15	Anemia	Nominal	Anemia	Si, No
16	Peso	Numérica	Peso	Kilogramos
17	Estadio	Numérica	Estadio	1,2,3,4,5

7.2.3 Explorar los datos.

Para la exploración de los datos se creó una base de datos en la plataforma *Azure*, en la cual se importan los datos y se conectan a la herramienta de visualización *Power BI*, para realizar y visualizar de una manera más armónica todos los datos contenidos en el conjunto de datos. Inicialmente se realiza una estadística descriptiva de las variables que conforman los datos. En la Tabla 3, se identifican las características principales de las variables de los archivos empleados, dicho resultado se obtuvo de utilizar los comandos de *Python* dentro *Power BI*.

Tabla 3. Estadístico de las variables

.	MEAN	STD	MIN	25%	50%
MUNICIPIO	30328.268418	28219.829984	5001.0	8001.0	13001.0
EDAD	73.509500	12.013227	1.0	67.0	75.0
HTA	1.000000	0.000000	1.0	1.0	1.0
ALBUNIA	2.500494	1.501571	0.0	1.0	3.0
AZÚCAR	1.999239	1.224102	0.0	1.0	2.0
UREA_SANGRE	143.392474	86.160827	1.0	70.0	139.0
CREATININA	5.781614	4.888842	0.1	1.7	3.8
SODIO	134.952450	20.229694	100.0	117.0	135.0
POTASIO	4.748699	1.300653	2.5	3.6	4.8
HEMOGLOBINA	14.397278	1.310687	12.1	13.4	14.3
DM	0.000000	0.000000	0.0	0.0	0.0
PESO	77.753432	9.998848	60.0	70.5	77.2
ESTADIO	3.137247	0.437758	3.0	3.0	3.0

Antes de empezar a procesar el conjunto de datos, se realiza un conjunto de visualizaciones para que ayude a comprender mejor las características de la información con la que se trabaja y su correlación.

Primero se visualiza en formato de historial las cuatro características de entrada con nombres “Duración”, “Páginas”, “Acciones” y “Valor” se puede ver gráficamente qué valores comprenden los mínimos, máximos y en qué intervalos se concentra la mayor densidad de registros.

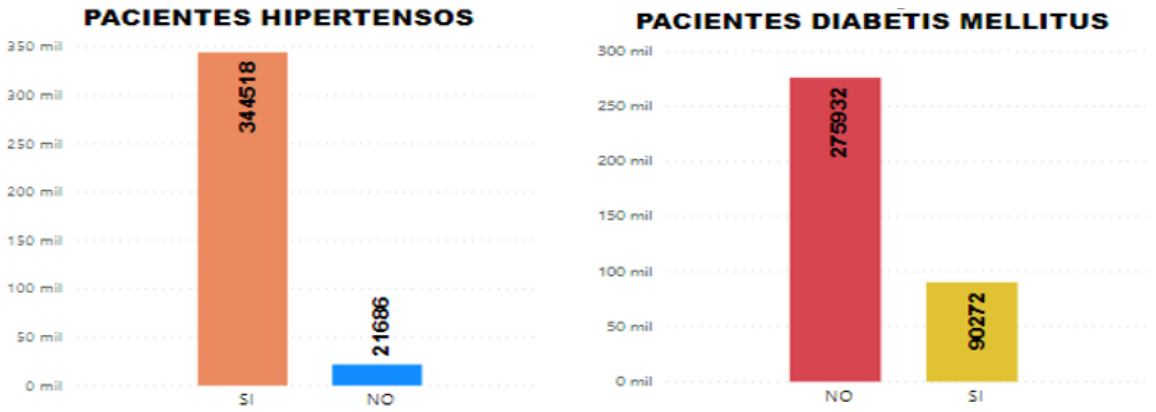

Fig. 5. Clasificación de las variables de pacientes

Como puede observarse en la grafica de Fig. 5, las enfermedades bases de la ERC son Hipertension y Diabetes. Los datos nos muestran que hay gran cantidad de pacientes con esta enfermedades pero que puntea la Hipertension, tambien conocida como enfermedad silenciosa.

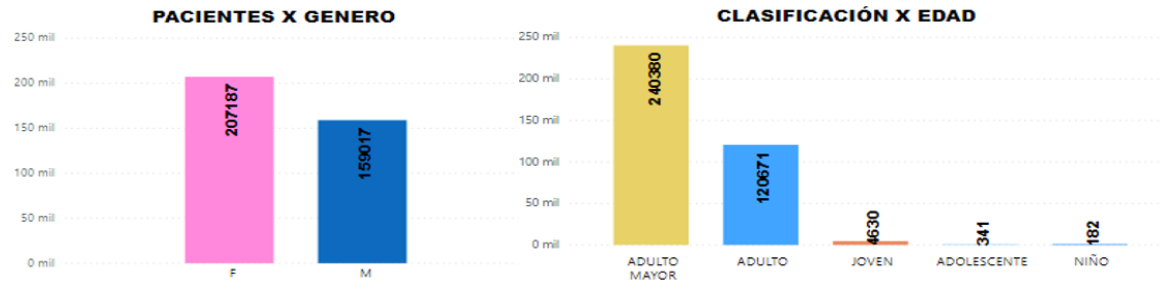

Fig. 6. Clasificación de las variables de pacientes según edad y género

Considerando la gráfica de la Fig. 6, la mayor prevalencia de la ERC es en mujeres, esto debido a su mayor esperanza de vida y llegada a la edad de riesgo de ERC (Adulto Mayor).

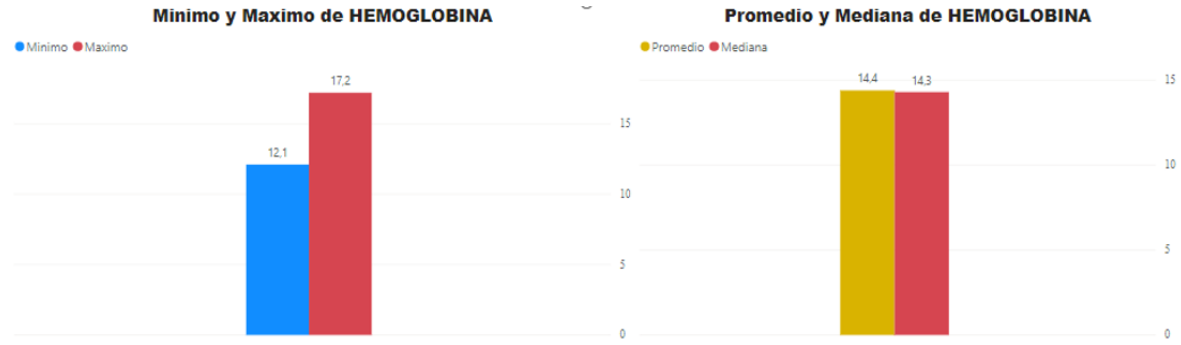

Fig. 7. Estadística clasificación Hemoglobina

En la gráfica de la Fig. 7 se presentan respetivamente los valores de la Hemoglobina, mínimo, máximo, promedio y mediana de los datos. Se estima que la cuarta parte de los pacientes crónicos se presenta con baja hemoglobina en los estadios iniciales de la enfermedad.

Fig. 8. Estadística clasificación Peso

Estudios epidemiológicos sitúan el exceso de peso como factor de riesgo para el desarrollo de la ERC junto otros factores. Como se observa en la gráfica de la Fig. 8 los datos obtenidos de las pacientes evidencian una alta tendencia al sobrepeso.

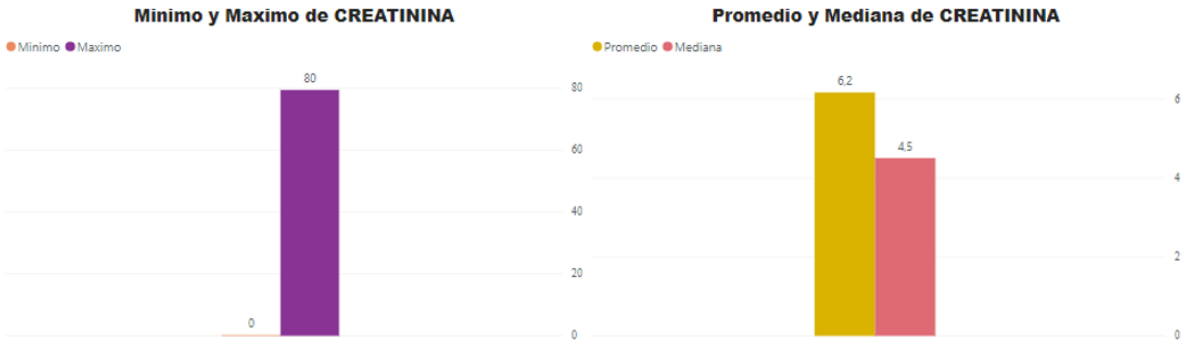

Fig. 9. Estadística, clasificación Promedio y Media Creatinina

A medida que la Creatinina sube su promedio en sangre, el porcentaje de la función renal baja. Por la anterior razón es una variable muy importante para el diagnóstico de la enfermedad renal. Fig. 9

Se realiza también un gráfico de dispersión que se obtuvo a partir del mismo conjunto de datos para cada par de variables. Las diagonales se tratan de manera diferente, dibujando una gráfica para mostrar la distribución univariada de los datos para la variable de cada columna.

Como se evidencia, las variables Edad y Peso presentan una alta correlación positiva, lo que no se presenta entre Hipertensión y Diabetes Mellitus que no se aprecia ninguna correlación entre estas variables. Para las variables de Peso e Hipertensión se presenta una baja correlación positiva, como pasa también entre las variables de Hipertensión con Edad y Diabetes Mellitus también con Edad. Es decir, la Hipertensión y Diabetes Mellitus son enfermedades que no dependen del factor edad, se puede padecer según riesgo como antecedentes familiares.

7.2.4 Verificar la calidad de los datos.

En esta sección se realizó la verificación de los datos para determinar la consistencia de los valores de los campos, la cantidad de distribución de los valores nulos y para encontrar valores fuera de rangos que pueden generar ruido para el proceso.

Este proceso de verificación se realizó en todo el conjunto de datos recibido. En los campos donde no se encontraban registros se cambió los campos vacíos por un valor *Null*.

7.3 Fase III: Análisis de los datos y selección de características

Fig. 10. Preparación de los datos

Al contar con la información de los datos se realiza el enfoque en la identificación de las variables que se utilizaran. Durante la revisión de los datos se encontraron un total de 58 variables, dentro de estas se identificaron 3 variables que se eliminaron para conservar la privacidad, seguridad y sensibilidad de los pacientes.

De las 55 variables restantes se eliminaron las que no tenían muestra, ya que si no aportan información en el entrenamiento no tendrían ninguna determinación. Después se realizó una clasificación entre las variables subjetivas y objetivas. Luego de realizar el análisis médico que se detalla en la Tabla 4 se eliminan 38 variables que no son relevantes para la predicción de la ERC.

Tabla 4. Variables descartadas

VARIABLES DESCARTADAS			
1	Periodo	**20**	Albuminuria Creatinuria
2	Última Fecha Atención	**21**	Albuminuria
3	UAP	**22**	BUN Orina
4	Enfermedad Renal Patológico	**23**	BUN Plasma
5	Tipo HC	**24**	Colesterol HDL
6	Fecha Historia	**25**	Colesterol LDL
7	Talla	**26**	Colesterol Total
8	Diámetro Cintura	**27**	Examen Orina
9	IMC	**28**	Glicemia
10	Hepatitis B	**29**	Hb Glicosilada
11	Hepatitis C	**30**	Hematocrito
12	Presión Arterial Sistólica	**31**	Hemoglobina
13	Presión Arterial Diastólica	**32**	Proteinuria

14	Presión Arterial Pulso	**33**	PTH
15	Presión Arterial Media	**34**	Triglicéridos
16	Frecuencia Cardiaca	**35**	Tasa Filtración
17	Frecuencia Respiratoria	**36**	Conducta
18	Obesidad	**37**	Fecha Próximo Control
19	Albúmina Sérica	**38**	Recomendaciones

Con la identificación de la escala que se utilizará para la clasificación de pacientes con ERC, se determinaron las variables a utilizar para evaluar un paciente según su grado de severidad. El listado inicial de 58 variables luego de realizar el análisis médico, el cual permite llegar a un conjunto de 17 variables, lo que requirió de la experticia del médico nefrólogo en el proceso de eliminación.

Dentro de todo este proceso se utilizarán criterios que permitan medir la severidad de la enfermedad renal, es importante y primordial tener claro que el juicio del experto (en este caso el médico) es crucial para tomar una decisión definitiva sobre el estado de un paciente, para ello el experto tendrá en cuenta los antecedentes base como pueden ser la Hipertensión y Diabetes.
Al tener el conjunto de datos listo, se realiza la importación sobre el experimento que se realizará en *Azure Machine Learning Studio* donde se hará la evaluación de los algoritmos. Seleccionando los atributos numéricos y no numéricos, lo que permite limitar las columnas disponibles para una operación posterior.

Al tener los atributos seleccionados, se indica qué columna tendrá los valores que se desean categorizar o predecir, en el caso de esta evaluación se utiliza la variable Estadio. Como se indica en la Tabla 5 es la que describe en qué estadio se encuentra la falla renal de un paciente.

Tabla 5. Clasificación del Estadio

ESTADIO	DESCRIPCIÓN
1	Daño renal con filtración normal o alta
2	Daño renal con disminución leve de la función renal
3	Daño renal con disminución moderada de la función renal
4	Daño renal con disminución severa de la función renal
5	Falla renal

Identificadas las características del conjunto de datos con el mayor poder predictivo, se realizan pruebas estadísticas con el Test de Correlación de *Pearson*, para determinar qué columnas son más predictivas.

Para esta tarea se utiliza el módulo de selección de características basado en filtros que trae *Azure Machine Learning*, como se observa en la Fig.11 el cual proporciona múltiples algoritmos de selección de características, que incluyen métodos de correlación identificando variables ordinales y no ordinales.

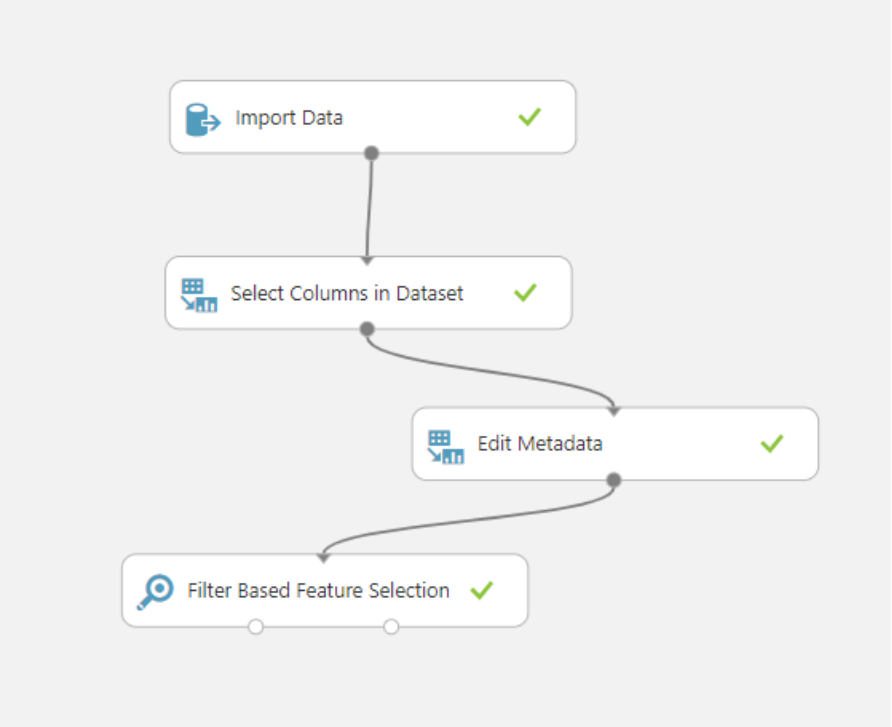

Fig. 11. Imagen Experimento de preparación de los datos.

7.4 Fase IV: Modelado

Fig. 12. Modelado.

Paso siguiente, se inicia el proceso de selección de la técnica de Aprendizaje Automático que se usará para desarrollar el Clasificador de pacientes objeto de esta investigación.

De acuerdo con la efectividad de los algoritmos más utilizados en estudios para el diagnóstico de enfermedades, se eligen los siguientes algoritmos.

- Regresión Logística: LR
- Redes Neuronales Artificiales: ANN
- Bosque de Decisiones: RF
- Jungla de Decisiones: DJ

Al contar con estos algoritmos, se divide el conjunto de datos de forma aleatoria en 2 partes: conjuntos de entrenamiento y conjuntos de pruebas. El experimento se realiza con el módulo *Split*, Fig.13 usando el 70% de los datos como datos de entrenamiento, y el 30% se reserva para evaluar la eficacia del modelo.

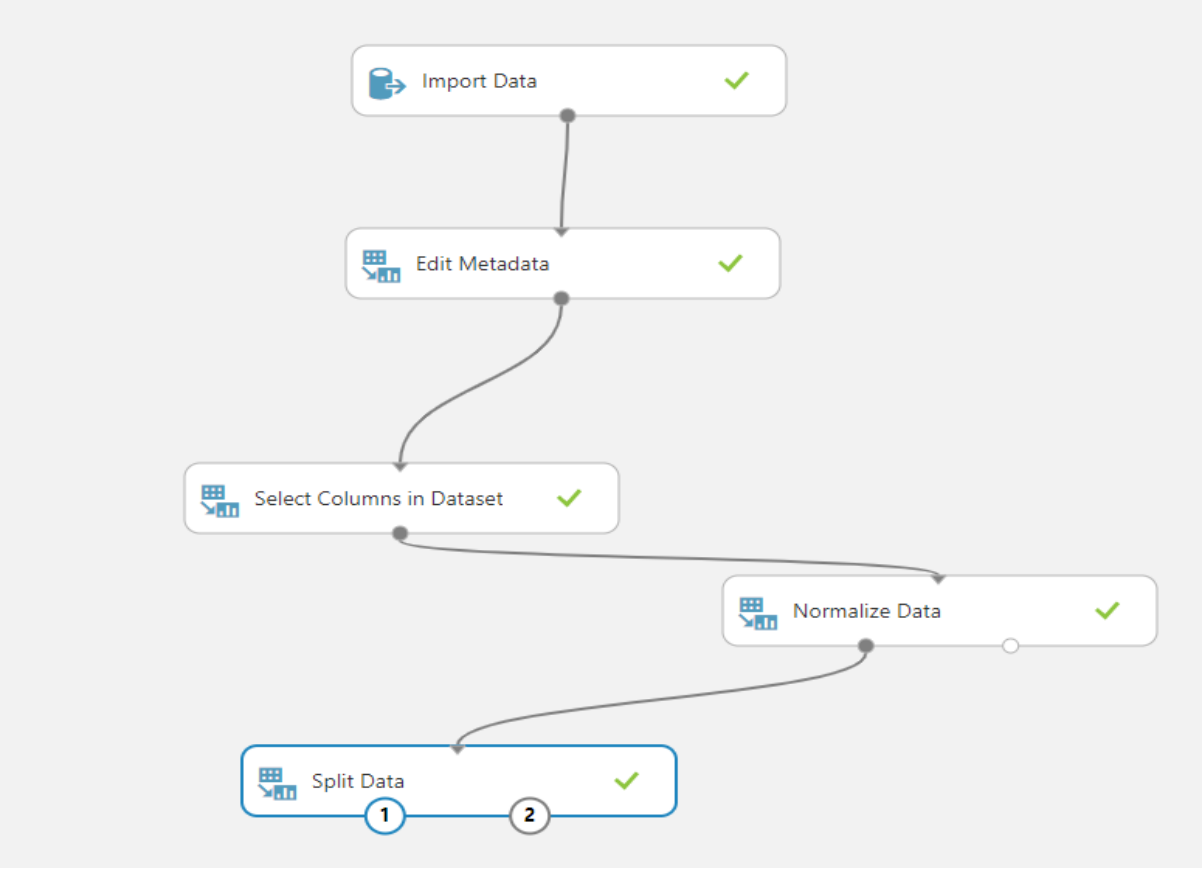

Fig. 13. Imagen Experimento de división de los datos.

Luego, se entrenan los modelos escogidos usando la herramienta *Train Model*, la cual es una herramienta de Azure que sirve para entrenar un modelo de manera supervisada con el conjunto de datos de entrenamiento como entrada. A estos modelos se agrega el *Score Model* que en Azure sirve para realizar punteo de predicciones de un modelo entrenado.

En cada combinación de modelo entrenado y los datos de prueba, se usa el modelo de clasificación *Evalúate Model* para calcular la matriz de confusión de resultados Fig. 14.

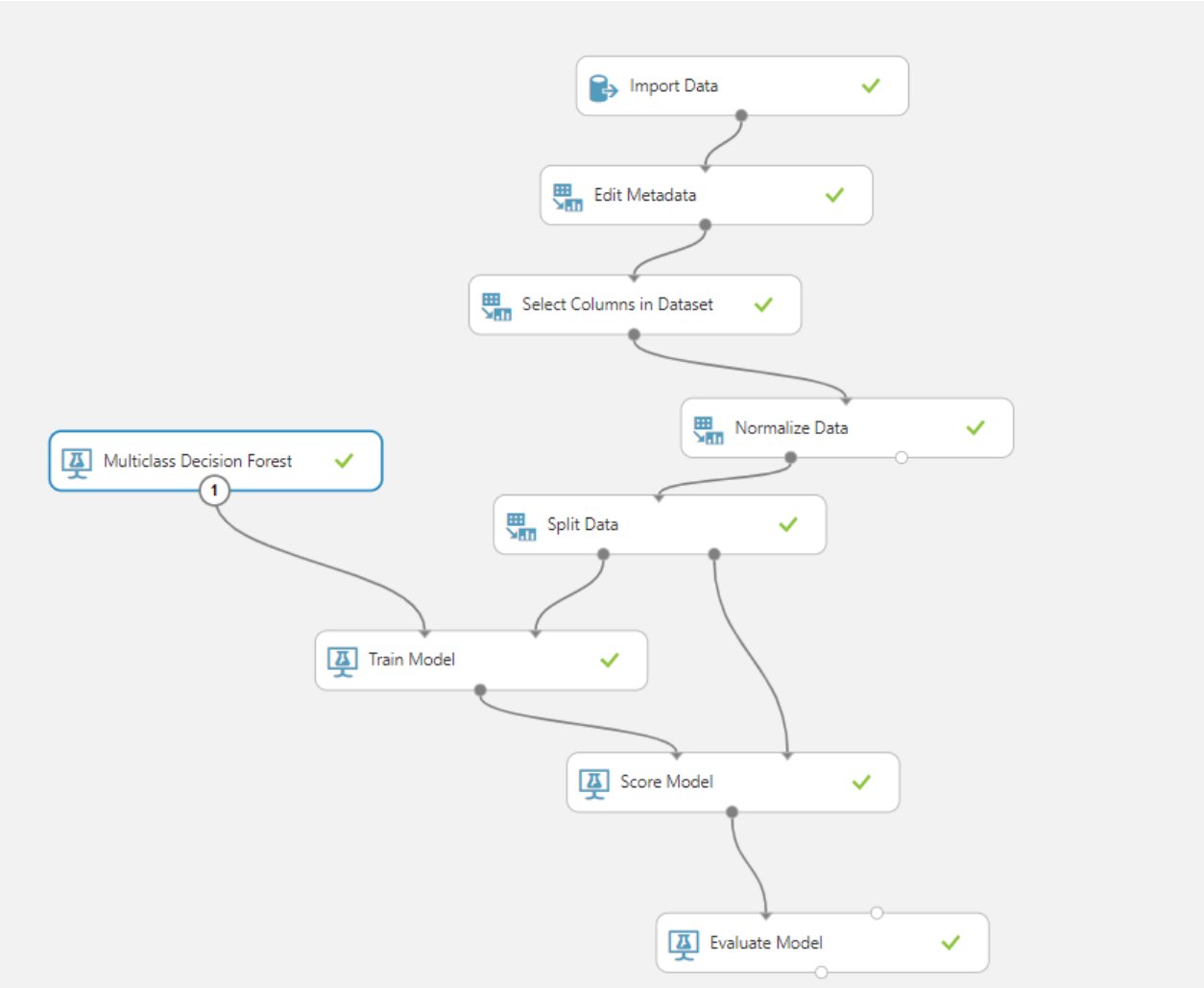

Fig. 14. Imagen Experimento de entrenamiento de los modelos.

Al realizar esta primera evaluación de los algoritmos escogidos, se verifica que los datos para los modelos se encuentran desbalanceados. Por ejemplo, en la precisión general que es la probabilidad de que un individuo sea clasificado correctamente mediante una prueba, es decir la suma de los

verdaderos positivos más los verdaderos negativos dividida por el número total de individuos evaluados, da valores aceptables como se observa en la Tabla 6, pero al verificar las matrices los valores siempre tienden a el estadio 3. Como se observa en las Tablas 7, 8, 9 y 10.

Tabla 6. Comparación de rendimiento entre los modelos

	Multiclase Red Neuronal	**Multiclase Bosques de Decision**	**Multiclase Regresión Logística**	**Multiclase Jungla de Decisiones**
Precisión General	0.784333	0.831205	0.772724	0.834632
Precisión Media	0.892167	0.88747	0.886362	0.889755
Precisión Micro-Promediada	0.784333	0.831205	0.772724	0.834632
Precisión Macro-Promediada	0.445861	0.691389	0.373977	0.816794
Sensibilidad Micro-Promediado	0.784333	0.831205	0.772724	0.834632
Sensibilidad Macro-Promediado	NaN	0.616839	NaN	0.536782

Tabla 7. Matriz de Redes Neuronales Artificiales

		3	**4**	**5**
Clase Actual	**3**	97.1%	0	2.9%
	4	97.2%	0.1%	2.8%
	5	82.9%	0.0%	17.0%

Tabla 8. Matriz de Bosques de Decisión

		3	4	5
Clase Actual	3	93.9%	3.3%	2.7%
	4	49.4%	41.7%	8.9%
	5	40.6%	10.0%	49.4%

Tabla 9. Matriz de Regresión Logística

		3	4	5
Clase Actual	3	94.0%	1.4%	4.6%
	4	79.9%	12.1%	8.0%
	5	59.5%	3.5%	37.0%

Tabla 10. Matriz de Jungla de Decisiones

		3	4	5
Clase Actual	3	98.7%	0.0%	1.3%
	4	77.8%	15.3%	6.9%
	5	50.7%	2.3%	47.1%

Para realizar la tarea del balanceo se agrega el módulo *SMOTE*, que es una técnica de estadística que permite aumentar el número de casos de forma equilibrada, tomando muestras del espacio de características para cada clase de destino y sus vecinos más cercanos. Generando nuevos ejemplos que combinan características de destino con características de sus vecinos Fig.15.

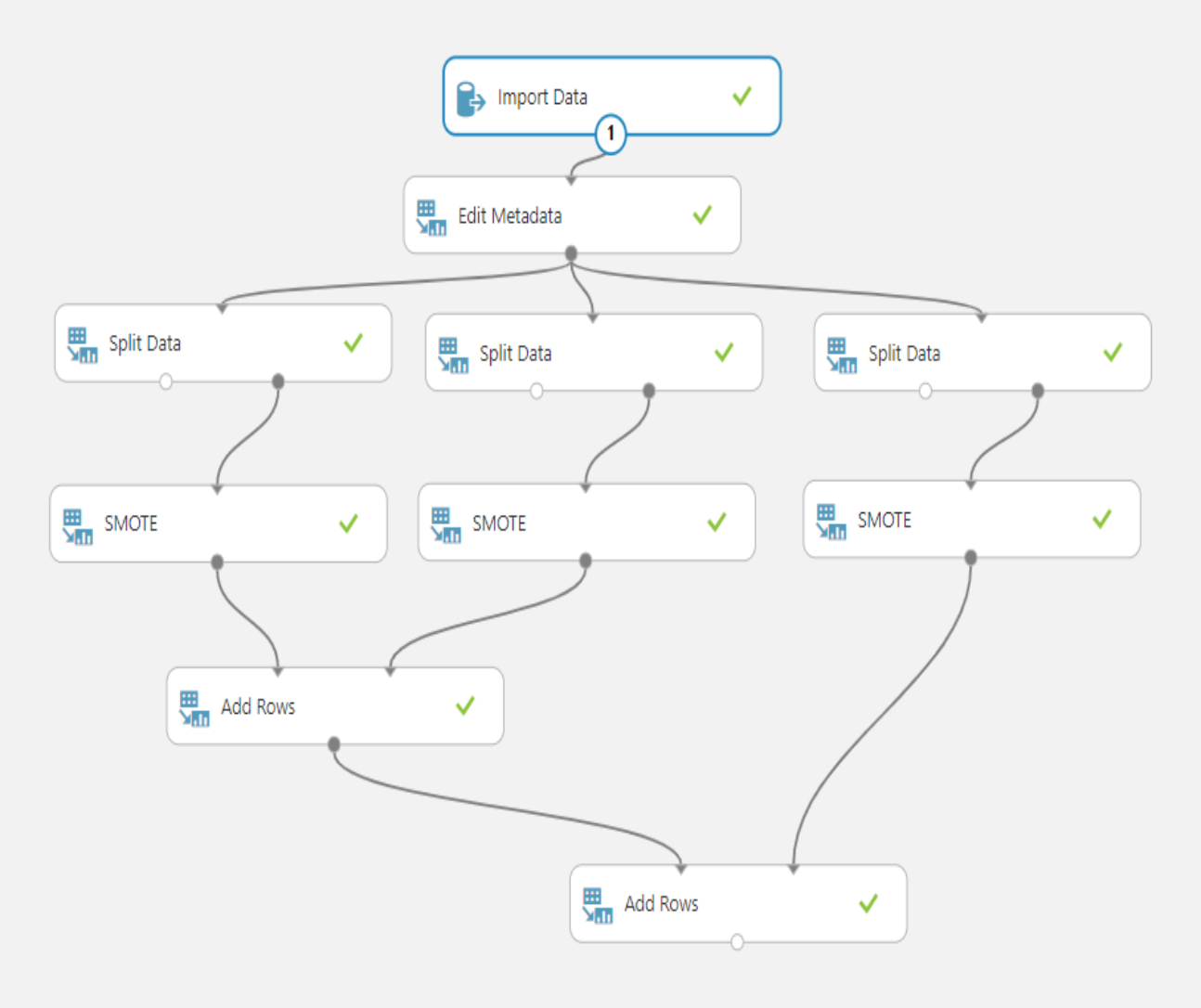

Fig. 15. Imagen Experimento de entrenamiento de balanceo.

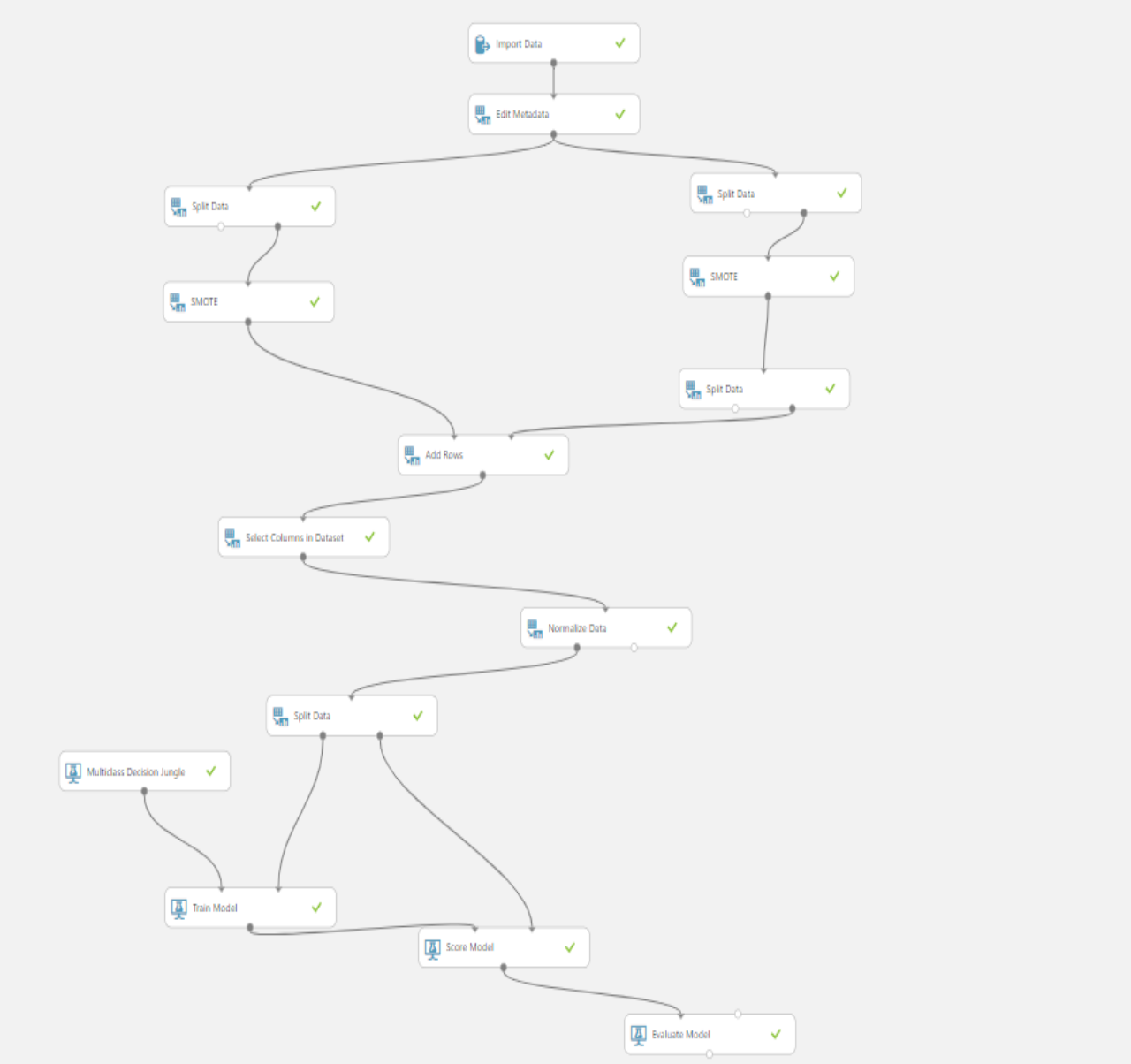

Fig. 16. Imagen Experimento de entrenamiento de los modelos balanceado.

Al realizar el balanceo de los datos Fig.16. Se realiza de nuevo la evaluación del modelo donde se obtienen mejores resultados como se indica en la Tabla 11.

Tabla 11. Comparación de rendimiento entre los modelos balanceados

	Multiclase Red Neuronal	**Multiclase Bosques de Decision**	**Multiclase Regresión Logística**	**Multiclase Jungla de Decisiones**
Precisión General	0.78591	0.925283	0.627158	0.750807
Precisión Media	0.857273	0.950189	0.751439	0.833871
Precisión Micro-Promediada	0.78591	0.925283	0.627158	0.750807
Precisión Macro-Promediada	0.788227	0.926279	0.624196	0.755939
Sensibilidad Micro-Promediado	0.78591	0.925283	0.627158	0.750807
Sensibilidad Macro-Promediado	0.786377	0.925339	0.627571	0.75134

Se puede observar que el rendimiento del Bosques de decisión aumentó en comparación al modelo desbalanceado, sigue siendo mejor que los modelos de contraste con la mayor precisión como se evidencia en la matriz de confusión.

Tabla 12. Matriz de Redes Neuronales Artificiales balanceadas

		3	**4**	**5**
Clase Actual	**3**	88.3%	6.7%	5.0%
	4	16.9%	67.6%	15.5%
	5	10.4%	9.6%	80.0%

Tabla 13. Matriz de Bosques de Decisión balanceada

		3	4	5
Clase Actual	3	97.5%	1.2%	1.2%
	4	6.7%	88.9%	4.4%
	5	3.8%	5.0%	91.2%

Tabla 14. Matriz de Regresión Logística balanceada

		3	4	5
Clase Actual	3	44.8%	39.6%	15.6%
	4	19.7%	68.9%	11.4%
	5	17.1%	8.3%	74.5%

Tabla 15. Matriz de Jungla de Decisiones balanceada

		3	4	5
Clase Actual	3	81.5%	11.0%	7.5%
	4	21.9%	65.4%	12.8%
	5	13.9%	7.6%	78.5%

Una vez balanceados los modelos, el siguiente paso es ajustar sus hiperparámetros para obtener el mayor poder predictivo posible Fig.17.

Para este estudio se utilizará el módulo Ajuste de hiperparámetros del Modelo. El cual logra determinar los hiperparámetros óptimos para un mejor modelo de Aprendizaje Automático determinado, ayudando a obtener unos resultados óptimos.

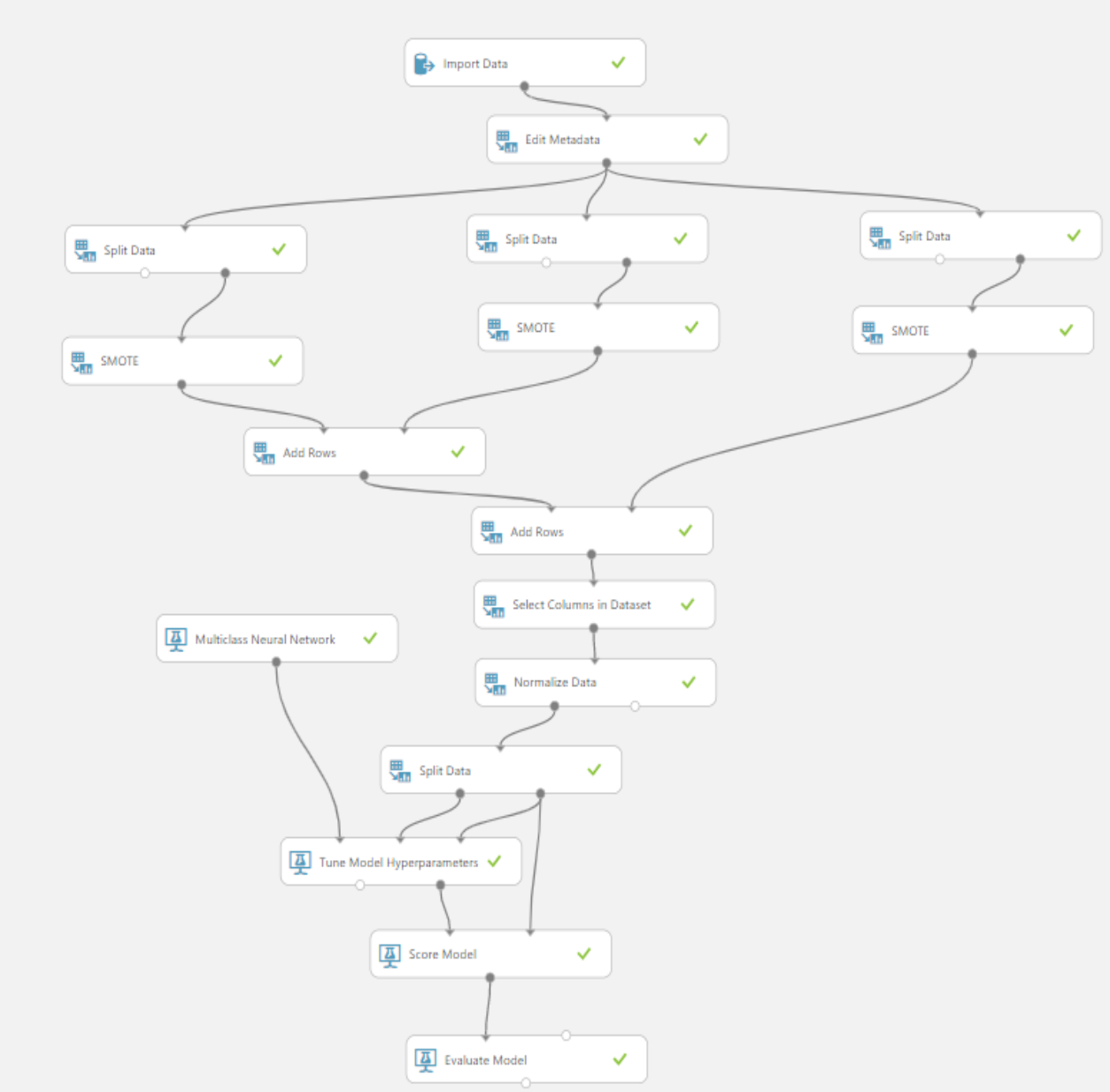

Fig. 17. Imagen Experimento de entrenamiento balanceado y ajustados a hiperparámetros.

Una vez que a los modelos se le ajustaron los hiperparámetros, a partir de los parámetros seleccionados, se obtuvieron las métricas de rendimiento que se muestran en la Tabla 16, y la matriz de confusión para todos los algoritmos.

Tabla 16. Métricas de rendimiento modelos con hiperparámetros

	Multiclase Redes Neuronales	**Multiclase Bosques de Decision**	**Multiclase Regresión Logística**	**Multiclase Jungla de Decisiones**
Precisión General	0.802177	0.931849	0.628873	0.828425
Precisión Media	0.868118	0.954566	0.752582	0.885617
Precisión Micro-Promediada	0.802177	0.931849	0.628873	0.828425
Precisión Macro-Promediada	0.80768	0.93219	0.625872	0.833657
Sensibilidad Micro-Promediado	0.802177	0.931849	0.628873	0.828425
Sensibilidad Macro-Promediado	0.802698	0.931894	0.629325	0.828786

Tabla 17. Matriz de confusión de Redes Neuronales Artificiales con hiperparámetros de los estadios 3-4-5

		3	**4**	**5**
Clase Actual	**3**	93.4%	2.8%	3.8%
	4	17.9%	66.5%	15.6%
	5	9.9%	9.3%	80.8%

Los parámetros seleccionados para la obtención de estos resultados fueron los siguientes:

Tabla 18. Parámetros seleccionados para Redes Neuronales

Número de capas ocultas	100
Tasa de aprendizaje	0.1
Número de iteraciones de aprendizaje	100
Pesos de inicialización	0.1
Normalización	Min-Máx

Tabla 19. Matriz de confusión de Bosques de Decisión con hiperparámetros de los estadios 3-4-5

		3	**4**	**5**
Clase Actual	**3**	96.6%	2.0%	1.3%
	4	5.3%	90.6%	4.1%
	5	3.0%	4.7%	92.3%

Los parámetros seleccionados para la obtención de estos resultados fueron los siguientes:

Tabla 20. Parámetros seleccionados para Bosques de Decisión

Método de Re-muestreo	Bagging
Número de árboles de decisión	8
Profundidad máxima de los árboles de decisión	32
Número de divisiones aleatorias por nodo	128
Número mínimo de muestras por hoja	1

Tabla 21. Matriz De Confusión De Regresión Logística Con Hiperparámetros De Los Estadios 3-4-5

		3	**4**	**5**
Clase Actual	**3**	46.2%	38.0%	15.8%
	4	20.6%	67.9%	11.5%
	5	17.3%	8.1%	74.6%

Los parámetros seleccionados para la obtención de estos resultados fueron los siguientes:

Tabla 22. Parámetros seleccionados para Regresión Logística

Tolerancia de optimización	1e-7
Regularización de peso L1	1
Regularización de peso L2	1
Tamaño de memoria para L-BFGS	20

Tabla 23. Matriz De Confusión De Jungla de Decisiones Con Hiperparámetros De Los Estadios 3-4-5

		3	**4**	**5**
Clase Actual	**3**	94.5%	1.5%	3.9%
	4	16.0%	71.8%	12.1%
	5	8.9%	8.8%	82.2%

Los parámetros seleccionados para la obtención de resultados fueron los siguientes:

Tabla 24 .Parámetros seleccionados para Jungla de Decisiones

Método de Re-muestreo	Bagging
Número de DAGs de decisión	8
Profundidad máxima de los DAGs de decisión	32
Amplitud máxima de los DAGs de decisión	128
Número de pasos de optimización por cada capa de los DAGs de decisión	2048

7.5 Fase V: Evaluación

Fig. 18. Evaluación (obtención de resultados).

La medida más importante de los algoritmos de clasificación es su exactitud. Como se observa en los resultados comparativos basados en la precisión Tabla 25, se encuentra que entre los cuatros algoritmos de clasificación comparados. El Regresión Logística Multiclase (LR) logra la exactitud más baja 68%, lo que implica que es un clasificador de resultados deficiente.

Sin embargo, los algoritmos Jungla de Decisión Multiclase (DJ) y Multiclase Red Neuronal (ANN) funcionan bien y muestran un rendimiento competitivo entre sí.

Aunque han alcanzado una exactitud de 75% y 80% respectivamente, no logran mostrar un rendimiento superior al algoritmo Multiclase Bosques de Decisión (RF) que alcanzo una exactitud del 92% que indica el rendimiento efectivo en la clasificación del conjunto de datos ERC empleado.

Tabla 25. Valores finales de Exactitud y Exhaustividad

	EXACTITUD	EXHAUSTIVIDAD
Bosques de Decisión	0.92226	0.921469
Redes Neuronales	0.805562	0.799806
Regresión Logística	0.688943	0.688542
Jungla de Decisión	0.754054	0.749988

8. CONCLUSIONES

Este estudio exploró como se puede utilizar un modelo de Aprendizaje para clasificar la posibilidad de un diagnóstico de la ERC.

En consecuencia, y en concordancia directa con los objetivos del proyecto se adaptó la metodología CRISP-DM al contexto del problema de manera que se surtieron diferentes etapas lógicamente organizadas; Recolección de datos, Preprocesamiento, Aprendizaje, Evaluación y Selección, las cuales permitieron la construcción de un modelo capaz de clasificar la posibilidad de un diagnóstico de la ERC con una precisión del 93%.

Como se evidencia en los resultados, el algoritmo de Bosques de Decisión ha obtenido unos resultados bastante óptimos, donde se han obtenidos predicciones del 93%. La preparación de los datos es un paso fundamental en el proceso y la precisión final del modelo tiene una dependencia directa de esta fase.

Gracias a los modelos podemos ver cómo afecta el cambio de las características a la búsqueda del valor objetivo con un simple cambio de selección de columnas, o mejoras en la data.

La innovación de este trabajo resulta del diseño ajustado al entorno del sistema de salud en Colombia y a la patología el ERC en nuestro país, con una metodología adaptada al caso de estudio y una propuesta de arquitectura de producción para el modelo con herramientas de Microsoft Azure de forma que, permita satisfacer en un futuro la escalabilidad de la solución. Además, esta metodología podría ser aplicable a los datos clínicos de otras enfermedades y patologías en el diagnóstico médico real.

El desarrollo de este proyecto permitió al autor adquirir un mayor conocimiento a través de trabajo tanto practico como teórico acerca de las técnicas actuales para el desarrollo del Aprendizaje Automático.

Este estudio tiene limitaciones por lo que hay margen para futuras investigaciones. El estudio no conto con una muestra de datos significativa, por las restricciones de los datos médicos y sus afectaciones legales en Colombia. Continuar con la expansión de la base de datos (incrementando la cantidad de ejemplos por cada variable) disminuiría el error de generalización limitado, para el modelo y al mismo tiempo permita detectar la gravedad de la enfermedad. Este modelo puede perfeccionarse con el aumento de tamaño y calidad de los datos.

También se abre espacio para una variedad de estudios de otras disciplinas, como estudios económicos alrededor del impacto que tiene obtener un diagnóstico en menor tiempo para poder dar tratamiento a la enfermedad en etapas tempranas, reduciendo sobre costos en el sistema de salud. Además, de una variedad de estudios sociológicos y clínicos sobre las consecuencias del manejo de la ERC tempranamente trae sobre la calidad de vida de los pacientes y sus familias.

Si bien la validez de esta investigación es interna, por cuanto el corpus de documento es privado y no puede ser publicado para otros trabajos, ayudará a profesionales interesados con de Aprendizaje Automático, como base para realizar sus estudios en el área de clasificación.

REFERENCIAS

[1] J. V. Román, «Sngula,» 02 Agosto 2016. [En línea]. Available: https://www.sngular.com/es/data-science-crisp-dm-metodologia/. [Último acceso: 15 08 2020].

[2] Institute for Health Metrics and Evaluation, «Institute for Health Metrics and Evaluation. Global Burden of Disease (GBD),» 2019. [En línea]. Available: http://www.healthdata.org/gbd. [Último acceso: 08 2020].

[3] Cuenta de Alto Costo/ Situación de la Enfermedad Renal Crónica en Colombia, «MINISTERIO DE SALUD COLOMBIA,» 2019. [En línea]. Available: https://acortar.link/B7iXG. [Último acceso: 09 2020].

[4] Fresenius Medical Care Colombia S.A., «Enfermedad Renal: causas y prevención - Fresenius Medical Care,» 2019. [En línea]. Available: https://acortar.link/Z2gKf. [Último acceso: 09 2020].

[5] Alan S. Go, MD, Glenn M. Chertow, MD, MPH, Fan de Dongjie, MSPH, Charles E. McCulloch, Ph.D., y Chi-yuan Hsu, MD, «Chronic Kidney Disease and the Risks of Death, Cardiovascular Events, and Hospitalization,» *The new England Journal of Medicine,* vol. 1, pp. 1296-1305, 2004.

[6] SITUACIÓN DE LA ERC EN COLOMBIA, «FONDO COLOMBIANO DE ENFERMEDADES DE ALTO COSTO,» 2019. [En línea]. Available: https://cuentadealtocosto.org/site/erc/. [Último acceso: 09 2020].

[7] Glassock, R. J. , Warnock, D. G. & Delanaye, P., «The global burden of chronic kidney disease: estimates, variability and pitfalls. Nature Reviews Nephrology, 13(2), 104–114. doi: 10.1038/nrneph.2016.163.,» 12 diciembre 2016. [En línea]. Available: https://pubmed.ncbi.nlm.nih.gov/27941934/. [Último acceso: 07 2020].

[8] Juan Ignacio Bagnato, «Crea un Arbol de Decisión en Python | Aprende Machine Learning,» 13 04 2018. [En línea]. Available: https://cutt.ly/Lg2EB9V. [Último acceso: 07 2020].

[9] Juan Ignacio Bagnato, «Crea un Arbol de Decisión en Python | Aprende Machine Learning,» 27 08 2017. [En línea]. Available: https://cutt.ly/pg2Riwx. [Último acceso: 07 2020].

[10] Jeff Hawkins, «"Machine Learning": definición, tipos y aplicaciones prácticas - Iberdrola,» 2004. [En línea]. Available: https://cutt.ly/jg2ROWJ. [Último acceso: 07 2020].

[11] Jonathan H. Chen, M.D., Ph.D. and Steven M. Asch, M.D., M.P.H, «Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations,» 16 mayo 2018. [En línea]. Available: https://cutt.ly/ng2RZQ4. [Último acceso: 07 2020].

[12] A. Niknejad , D. Petrovic., «Introduction to computational intelligence techniques and areas of their applications in medicine. Med Appl Artif Intell, 51,» 2013. [En línea]. Available: https://cutt.ly/7g2R8Jo. [Último acceso: 08 2020].

[13] Fondo Nacional de Enfermedades de alto costo, Cuenta de Alto Costo[CAC], «Situación de la enfermedad renal crónica, la hipertensión arterial y la diabetes mellitus en Colombia.,» 2016. [En línea]. Available: https://cuentadealtocosto.org/site/general/cuenta-de-alto-costo-nuestra-linea-del-tiempo/. [Último acceso: 06 2020].

[14] Microsoft, «Azure,» 11 04 2020. [En línea]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml. [Último acceso: 20 09 2020].

[15] MICROSOFT, «POWERBI,» Gartner , 2020. [En línea]. Available: https://powerbi.microsoft.com/es-es/what-is-power-bi/. [Último acceso: 18 09 2020].

[16] Xiao, J., Ding, R., Xu, X. et al, «Comparison and development of machine learning tools in the prediction of chronic kidney disease progression,» *Revista de Medicina Traslacional volumen,* vol. 17, nº 119, 2019.

[17] Polat, H., Danaei Mehr, H. & Cetin, «A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. J Med Syst 41, 55,» *Revista de sistemas médicos volumen,* nº 55, 2017.

[18] APA Li, Qi1; Fan, Qiu-Ling2; Han, Qiu-Xia1; Geng, Wen-Jia3; Zhao, Huan-Huan1; Ding, Xiao-Nan1; Yan, Jing-Yao1; Zhu, Han-Yu1, «Machine learning in nephrology: scratching the surface, Chinese Medical Journal: March 20, 2020 - Volume - Issue 6 - p 687-698,» *Chinese Medical Journal,* vol. 1, nº 6, pp. 687-698, 2020.

[19] D. K. ,. R. ,. E. ,. E. ,. M. ,. L. ,. T. ,. K. y. K. Zvi Segal, «Algoritmo de aprendizaje automático para la detección temprana de la enfermedad renal en etapa terminal,» *Nefrología BMC,* vol. 21, nº 518, 2020 .

[20] L. C. Y. L. C. F. y. B. C. J. Qin, Una metologia de aprendizaje automatico para diagnosticar la enfermedad renal cronica, vol. 8, China: Instituto de Ingenieros Electricos y Electronicos, 2020, pp. 20991-21002.

[21] A. G. L. U. A. N. K. A.-C. L. S.-A. Erik Dovgan, Uso de modelos de aprendizaje automático para predecir el inicio de la terapia de reemplazo renal en pacientes con enfermedad renal crónica, U. d. S. M. G. d. C. Giuseppe Coppolino, Ed., ITALIA: Dovgan et al, 2020.

[22] Li, Qi1; Fan, Qiu-Ling2; Han, Qiu-Xia1; Geng, Wen-Jia3; Zhao, Huan-Huan1; Ding, Xiao-Nan1; Yan, Jing-Yao1; Zhu, Han-Yu1, «Machine learning in nephrology: scratching the surface, Chinese Medical Journal:,» *Chinese Medical Journal*, nº 6, pp. 687-698, 2020.

[23] Tim Newman, «MEDICAL NEWS TODAY,» San francisco/ Estados Unidos, 2017.

CRONOGRAMA

ACTIVIDADES	2020																																														
	Meses	1				2				3				4				5				6				7				8				9				10				11				12			
	Semanas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Elección del tema																																																
Planteamiento del problema																																																
Justificación																																															
Objetivos																																															
Estado de Arte																																															
Marco Conceptual																																															
Metodología																																															
Fase I																																															
Fase II																																															
Fase III																																															
Fase IV																																															
Fase V																																															
Resultado																																															
Conclusiones																																															
Diseño de informe final																																															