Universidad Internacional de La Rioja (UNIR)

ESIT

Máster Universitario en Inteligencia Artificial

Modelo piloto de aprendizaje automático para la predicción de mortalidad por Leucemia Mieloide Aguda

Trabajo Fin de Máster

Presentado por: Hernández Martínez, Ricardo

Directora: Guerrero García, Josefina

Codirectora: Sossa Melo, Claudia

Ciudad: Bucaramanga
Fecha: 22 de septiembre de 2021

Resumen

La Leucemia Mieloide Aguda (LMA) es un tipo de cáncer de sangre que se encuentra entre los primeros con mayor cantidad de casos nuevos reportados en Colombia. El uso de aplicaciones basadas en Inteligencia Artificial ha aumentado en los últimos años, utilizadas como herramientas de apoyo al diagnóstico, pronóstico y tratamiento de enfermedades. En este trabajo se aplicaron algoritmos de aprendizaje supervisado para la creación de un modelo piloto predictivo de mortalidad de pacientes que padecen de LMA, siguiendo la metodología tradicional de proyectos de analítica de datos. Se utilizó para su desarrollo información recopilada en tratamientos bajo el protocolo terapéutico del Programa Español de Tratamientos en Hematología (PETHEMA). El modelo desarrollado alcanzó 0.88 de exactitud, con un AUC de 0.92. Los resultados obtenidos son un indicio favorable de la posibilidad de aplicar la Inteligencia Artificial en el campo de la medicina para el apoyo de la práctica clínica tradicional.

Palabras Clave: Inteligencia artificial, Leucemia mieloide aguda, Modelo predictivo.

Abstract

Acute Myeloid Leukemia (AML) is a type of blood cancer among the ones with the highest number of new cases reported in Colombia. Due to this, in recent years the use of applications based on Artificial Intelligence to support the diagnosis, prognosis and treatment of this disease has increased. In this work, the application of supervised learning algorithms is tested for the creation of a predictive pilot model of mortality in patients suffering from AML, following the traditional methodology of data analytics projects. For its development, information collected in treatments under the therapeutic protocol of the Spanish Program for Hematology Treatments (PETHEMA) was used. The developed model reached 0.88 accuracy, with an AUC of 0.92. The results obtained are a favorable indication of the possibility of applying Artificial Intelligence in the field of medicine to support traditional clinical practice.

Keywords: Acute myeloid leukemia, Artificial intelligence, Predictive model.

Índice de contenidos

Índice de tablas

Índice de figuras

Abreviaturas

aGVHD: *acute Graft-versus-Host Disease*. Enfermedad de injerto contra huésped aguda.

AUC: *Area Under Curve*. Área bajo la curva.

AUROC: *Area Under the Receiver Operating Characteristic*. Área bajo la característica operativa del receptor.

CAC: Cuenta de Alto Costo.

CNN: *Convolutional Neural Network*. Red Neuronal Convolucional.

CONPES: Consejo Nacional de Política Económica y Social.

DNN: *Deep Neural Network*. Red Neuronal Profunda.

EPV Eventos por variable.

HLA: *Human leukocyte antigen.* Antígeno leucocitario humano.

HSCT: *Hematopoietic Stem Cell Transplant*. Trasplante hematopoyético de células madre.

HTML: *Hypertext Markup Language*. Lenguaje de marcado de hipertexto.

LMA: Leucemia Mieloide Aguda.

PCNR: Proporción de Casos Nuevos Reportados.

PETHEMA: Programa Español de Tratamiento en Hematología.

PROBAST: *Prediction model Risk Of Bias Assessment Tool*. Herramienta de evaluación de riesgo de sesgo de modelos predictivos.

RC: Remisión completa.

SISMED: Sistema de Información de Precios de Medicamentos.

SG: Supervivencia global.

SVM: *Support Vector Machine*. Máquina de Vector de Soporte.

1. Introducción

La leucemia es un tipo de cáncer de sangre en el que se observa un crecimiento anómalo de células inmaduras que saturan la médula ósea y el torrente sanguíneo, lo cual produce en la persona que padece la enfermedad una serie de complicaciones que pueden llevarla a la muerte. En Colombia, la Leucemia Mieloide Aguda (LMA) se encuentra entre los 10 tipos de cáncer con mayor cantidad de casos nuevos reportados, con una prevalencia que se ha visto en aumento durante los últimos años (Cuenta de Alto Costo, 2019). Por situaciones externas a la enfermedad, al paciente y al tratamiento, en el país existen condiciones adicionales relacionadas con el acceso y la organización del sistema de salud, que suponen una dificultad adicional que impacta de manera negativa en los resultados en la lucha contra esta enfermedad.

Realizar el pronóstico de enfermedades o futuras condiciones de los pacientes, si bien es algo imperativo en muchas circunstancias clínicas y de investigación, es una tarea compleja que en muchas ocasiones se encuentra por encima de la capacidad humana debido a la gran cantidad de factores y variables que influyen en el proceso. Las herramientas digitales, y el análisis de la información disponible haciendo uso de tecnología de vanguardia, ofrecen una valiosa ayuda en el campo de la medicina poniendo al alcance del personal de salud los medios para realizar diagnósticos más rápidos y acertados, y para el análisis de información que permita predecir y crear políticas en beneficio de la comunidad. Consciente de esta situación, el Gobierno de Colombia tiene dentro de sus políticas el apoyo para la transformación digital e Inteligencia Artificial del Estado, con el fin de potenciar la generación de valor económico y social, favoreciendo la productividad y el bienestar de los ciudadanos, según lo expresó el Consejo Nacional de Política Económica y Social (CONPES) en el año 2019 (Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia, 2019).

Teniendo en cuenta lo mencionado anteriormente, y los avances tecnológicos actuales, en este trabajo se aplicaron algoritmos de Inteligencia Artificial para la creación de un modelo piloto predictivo de mortalidad para personas diagnosticadas con LMA que son atendidas en un hospital universitario ubicado en la ciudad de Floridablanca, Colombia, como un paso inicial para considerar la viabilidad de incluir análisis de este mismo estilo en la práctica médica regional y en las instituciones que aplican el protocolo del Programa Español de Tratamiento en Hematología (PETHEMA) como guía de tratamiento de esta enfermedad.

Se planteó el uso de una metodología propia de los proyectos de Inteligencia Artificial y *data science* para el desarrollo de este trabajo, planteando requisitos enfocados en cumplir con los

objetivos trazados. El resultado es la obtención de un conjunto de datos ajustado a las condiciones encontradas y de un modelo predictivo que alcanzó 0.88 de exactitud, 1 de sensibilidad y *recall*, y un AUC de 0.92.

Un trabajo como el presentado en este documento tiene justificación si se analiza la situación de Colombia, no sólo si se tiene en cuenta la etiología particular que pueden tener ciertas enfermedades como la LMA en la población, sino además por la aún incipiente aplicación de los modelos de Inteligencia Artificial en la escena nacional. Los resultados obtenidos en este trabajo muestran las capacidades que la aplicación de algoritmos de aprendizaje supervisado pueden ofrecer al campo de la medicina a nivel local y también pueden ser tomados como un paso inicial que abre la puerta a un nuevo enfoque que apoye la práctica médica.

1.1 Motivación

La LMA es un cáncer de la sangre y la médula ósea, que suele empeorar rápidamente si no se recibe tratamiento, siendo el tipo más común de leucemia aguda en adultos. La LMA también puede ser nombrada como leucemia mielógena aguda, leucemia mieloblástica aguda, leucemia granulocítica aguda y leucemia no linfocítica aguda (Instituto Nacional del Cáncer NCI, 2020).

En Instituto Nacional del Cáncer (2020), sitio web de la dependencia principal del gobierno de los Estados Unidos para la investigación del cáncer, se muestra información resumida de manera sencilla acerca de la LMA. En una persona sana, la médula ósea produce células madre que maduran y se convierten en las células presentes en el torrente sanguíneo. Estas células madre sanguíneas pueden ser de tipo mieloide o linfoide. Las células madre linfoides dan lugar a un tipo de glóbulos blancos, y por su parte, las células madre mieloides al madurar se convierten en glóbulos rojos, plaquetas y otro tipo de glóbulos blancos.

Cuando una persona padece LMA, las células madre mieloides no terminan de madurar y se convierten en mieloblastos o, en algunos casos, en demasiados glóbulos rojos o plaquetas anormales. A estas células anormales se les conoce como blastocitos leucémicos. Cuando hay proliferación de estas células leucémicas, y se acumulan en la médula ósea y la sangre, hay menos espacio para células sanguíneas saludables. Esto da lugar a complicaciones de salud como infecciones, anemia o sangrados fáciles. Es posible que los blastocitos leucémicos se diseminen fuera de la sangre a otras partes del cuerpo como el encéfalo, la medula espinal, la piel y las encías.

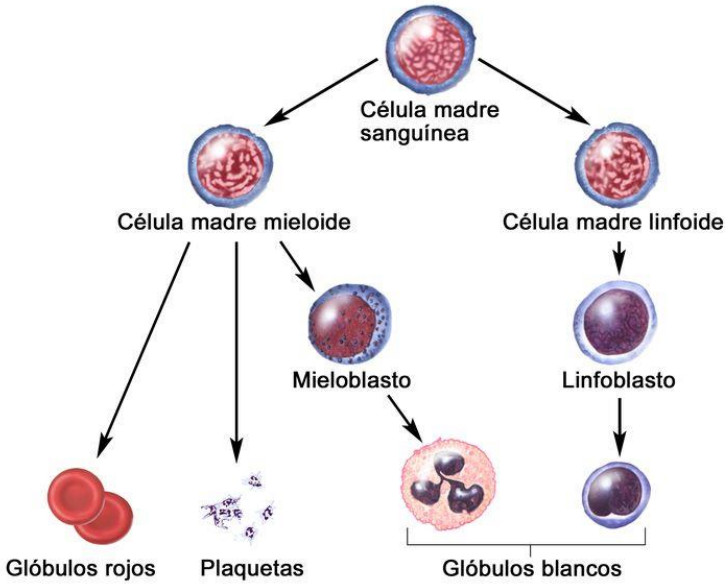

Figura 1. Evolución de una célula madre sanguínea
(Instituto Nacional del Cáncer NCI, 2020)

Situación en Colombia

En el año 2020, la leucemia ocupó en Colombia el noveno puesto entre los tipos de cáncer con mayor cantidad de nuevos casos, con un 3% de estos. Si se revisa la escala de mortalidad, ocupó el octavo puesto entre todos los cánceres con un 4.4% del total de las muertes reportadas, dos puestos por encima del décimo lugar que ocupa esta enfermedad a nivel mundial si se analiza este mismo rubro (The Global Cancer Observatory, 2020; WHO, 2020). Dentro de los diferentes tipos de leucemia, la Cuenta de Alto Costo (CAC) reportó que en Colombia la LMA ocupó el décimo lugar entre hombres y mujeres, en términos de mayor Proporción de Casos Nuevos Reportados (PCNR) entre todos los tipos de cáncer en el año 2019 (Figura 2).

Los casos totales de LMA en la población adulta fueron 1,047, de los cuales 158 se diagnosticaron durante el periodo y se reportaron 185 fallecidos (Cuenta de Alto Costo, 2019). La prevalencia de la LMA en la población adulta se ha incrementado a través de los últimos años, con una pequeña excepción entre los años de 2018 y 2019, al igual que la PCNR y la mortalidad (Figura 3).

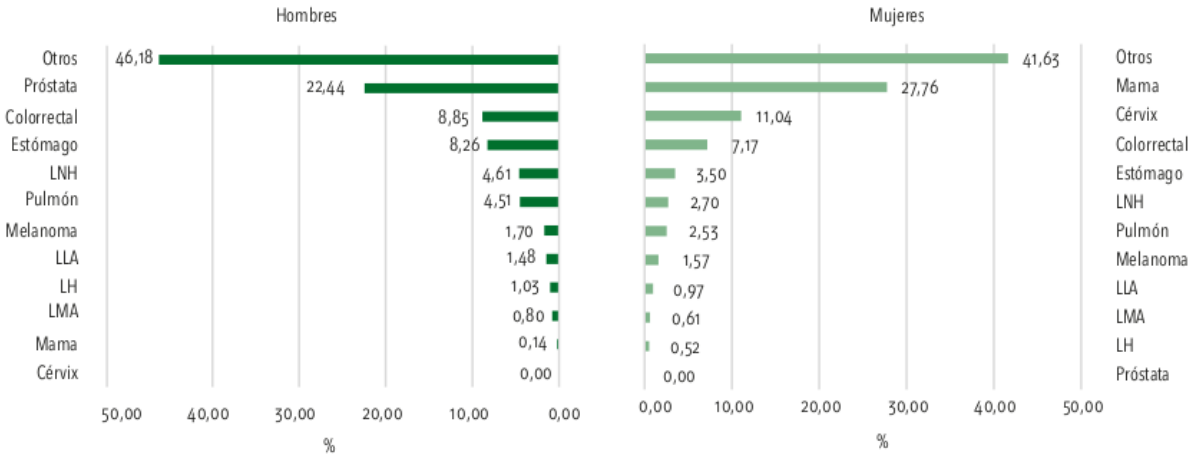

Figura 2. Distribución porcentual de casos nuevos por tipo de cáncer priorizado según sexo en Colombia 2019 (Cuenta de Alto Costo, 2019)

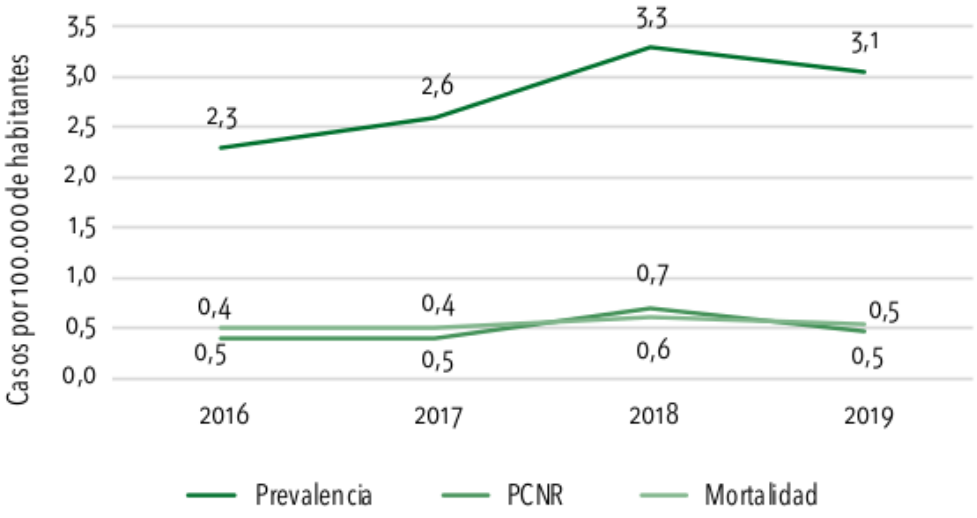

Figura 3. Medidas de frecuencia para LMA en adultos por cada 100.000 habitantes. Colombia 2015-2019. (Cuenta de Alto Costo, 2019)

El mayor porcentaje de casos revisados por la CAC son clasificados como de riesgo alto o desfavorable (Cuenta de Alto Costo, 2019). Los avances son precarios si se analiza la supervivencia de los pacientes a largo plazo en Colombia. Es complejo analizar las causas de este resultado pues no se debe únicamente a las condiciones de salud del paciente y al tratamiento. En Colombia, el acceso de los pacientes a los servicios de salud y la organización de esta última juegan un papel de vital relevancia. No hay datos específicos actualizados, pero los últimos señalan que los resultados clínicos de sobrevida en el país son muy bajos (menores del 20%) (Centro Nacional de Investigación en evidencia y Tecnologías en Salud, 2013).

Clasificación de riesgo	n (%)	Edad mediana (RIQ)
Bajo, estándar o favorable	17(15,89)	54(31-58)
Intermedio	17(15,89)	43(35-58)
Alto o desfavorable	55(55,40)	60(40-69)
Otro	16(14,95)	56,5(32-64)

Figura 4. Distribución de casos nuevos de LMA según clasificación del riesgo y edad en Colombia 2019. (Cuenta de Alto Costo, 2019)

1.2 Planteamiento del trabajo

Pronosticar la probabilidad que un paciente desarrolle una enfermedad, o calcular cuál va a ser su curso una vez esta es adquirida, es una tarea que puede estar por encima de la capacidad de la mente humana en muchas ocasiones, debido a la cantidad de alternativas y variables de importancia que hay que tener cuenta, así como el desconocimiento de las relaciones entre las mismas. Sin embargo, en muchas circunstancias clínicas y de investigación, es valioso que el personal de salud obtenga un resultado estimativo lo más preciso posible del pronóstico de un paciente.

El uso de herramientas que permitan identificar de manera temprana el riesgo y la probabilidad que un paciente pueda desarrollar una condición crítica es de gran importancia. Identificar la mortalidad en cierto estadio de la enfermedad puede influir en la selección del tratamiento más adecuado, el cual se ajuste a la condición del paciente, buscando obtener el mayor beneficio para este y la optimización de los recursos disponibles del sistema de salud. Además, al realizar análisis de riesgo y mortalidad se pueden llegar a determinar otros factores que inciden en el pronóstico, permitiendo sentar las bases para posteriores estudios que deriven en la planificación de mejores políticas de salud para la comunidad.

Como se ve en este trabajo, la utilidad de la Inteligencia Artificial se ha probado con éxito a través de diferentes estudios y trabajos desarrollados en el transcurso de los últimos años, y aunque tiene aún algunos retos por resolver, se ha convertido poco a poco en un aliado importante en diferentes ámbitos entre los que se encuentra la medicina. Teniendo en cuenta lo anterior, el Gobierno de Colombia tiene una política nacional para la transformación digital e Inteligencia Artificial del Estado. Esta política tiene como objetivo potenciar la generación de valor social y económico en el país a través del uso estratégico de tecnologías digitales, para

impulsar la productividad y favorecer el bienestar de los ciudadanos (Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia, 2019).

Teniendo en cuenta lo mencionado anteriormente, y con base en los métodos y resultados obtenidos en trabajos afines revisados en el estado del arte, en este trabajo se planteó la aplicación de algoritmos de aprendizaje automático para la creación de un modelo piloto de predicción de mortalidad en pacientes que padecen LMA, analizando los datos clínicos de pacientes atendidos entre los años de 2009 y 2021 en un hospital universitario ubicado en Colombia. El resultado obtenido es un paso inicial que sirve de piloto para la aplicación de un modelo de Inteligencia Artificial para la predicción de mortalidad que puede ser replicado en las instituciones o centros médicos que usan el protocolo especificado por PETHEMA para el tratamiento de enfermedades hematológicas a nivel nacional, y que muestra la viabilidad de la aplicación de este tipo de modelos.

1.3 Estructura de la memoria

En el documento se inicia mencionando algunas generalidades de la LMA y planteando la relevancia de la enfermedad, mostrando las estadísticas a nivel nacional de la incidencia poblacional que esta tiene en Colombia. Se plantea la importancia del pronóstico en el tratamiento de enfermedades haciendo hincapié en el uso de herramientas que sirvan de apoyo para este proceso.

Posteriormente, se hace un análisis del estado del arte iniciando por la revisión de algunas de las escalas utilizadas por el personal médico para la predicción de futuras condiciones de los pacientes. Luego, se revisan algunas aplicaciones de la Inteligencia Artificial utilizadas para el desarrollo de herramientas y modelos en las áreas de pronóstico, diagnóstico y tratamiento clínico. Al final de este análisis se revisan algunos trabajos e investigaciones en los que se utilizó la Inteligencia Artificial en diferentes etapas del tratamiento y pronóstico de la LMA, y qué tipo de investigaciones se están realizando a nivel nacional referentes a esta enfermedad.

A continuación de esto, se enuncian los objetivos que se plantean con la realización de este trabajo y la metodología seguida en la ejecución de las actividades, identificando los requisitos necesarios para cumplir con la ruta inicialmente planteada.

En el capítulo 5 se realiza una descripción detallada de las actividades realizadas, explicando los pasos seguidos desde la exploración, análisis y limpieza de los datos presentes en el

conjunto de datos utilizado, hasta la implementación de una primera interfaz con funciones básicas para acceder al modelo piloto predictivo desarrollado.

El documento finaliza con la evaluación de los resultados obtenidos y con la realización de un análisis usando la lista de verificación PROBAST (*Prediction model Risk Of Bias Assessment Tool*) para determinar el riesgo de sesgo del modelo generado. Después de esto se detallan las conclusiones obtenidas con la realización del trabajo, y se visualizan posibles acciones que sirvan en la definición de líneas de trabajo futuras.

2. Contexto y estado del arte

2.1 Métodos actuales de predicción utilizados en medicina

En el área de medicina se realizan evaluaciones que permiten determinar el estado de salud de los pacientes, teniendo en ocasiones la capacidad de predecir futuras consecuencias derivadas. El uso de escalas está estandarizado en la comunidad médica. Se trata de algoritmos definidos tras estudios médicos que permiten pronosticar ciertas variables (Fernández, 2019). En unidades de cuidados críticos, se aplican escalas como *Acute Physiology and Chronic Health Evaluation II* (APACHE II) o *Sequential Organ Failure Assessment* (SOFA), con el fin de calcular la probabilidad de muerte de un paciente. Cabe señalar que las escalas no son usadas únicamente para predecir la probabilidad de muerte. También son utilizadas para evaluar el riesgo de ocurrencia de eventos futuros asociados a morbilidades, y en algunos casos, es posible aplicar ciertas escalas para calcular la probabilidad de diagnósticos secundarios asociados a una condición de salud.

El uso de escalas requiere de conocimiento y atención específica de profesionales especializados. Esto hace que su aplicación se realice de forma netamente manual, que conlleva tiempo de procesamiento por parte del personal de salud, y la posibilidad de error implícita en cualquier tipo de actividad humana. El manejo de pacientes es complejo, un diagnóstico acertado puede requerir tener en cuenta un gran número de variables al mismo tiempo. Estudios demuestran que el límite de capacidad de procesamiento del ser humano es de 7 elementos de manera simultánea (Saaty & Ozdemir, 2003). Este límite puede afectar en el diagnóstico final, dejando fuera variables que pueden ayudar en la toma de decisiones.

Pero no todo es bueno y los modelos de predicción y escalas desarrollados no son infalibles. Si bien su uso se ha popularizado en medicina, desafortunadamente la mayoría de modelos no han sido implementados en la práctica clínica a gran escala. Los modelos publicados a menudo se desarrollan utilizando métodos inapropiados y están pobremente reportados, lo que hace difícil o incluso imposible juzgar la calidad metodológica. Además, rara vez se han informado estudios de impacto que muestren mejores resultados clínicos cuando se utiliza un modelo de predicción en la práctica clínica habitual (Dekker et al., 2017). El cuidado para crear estos modelos deber ser máximo, invirtiendo el tiempo suficiente para la validación y evaluación de los mismos, analizando su impacto. Una buena implementación influye positivamente en los resultados de los pacientes.

Otros métodos se han estructurado integrando nuevos recursos disponibles, aprovechando el avance tecnológico de los dispositivos de monitoreo y la integración a nivel de datos existente producto de la digitalización de la información. Si bien estos dispositivos de monitoreo de pacientes pueden generar falsas alarmas, si a esta información se incorporan otros tipos de datos, es posible realizar una predicción estableciendo previamente patrones. Un estudio publicado en *Journal of Biomedical Informatics* muestra las conclusiones obtenidas después de integrar los datos de resultados de prueba de laboratorio con señales de alarma emitidas por los dispositivos de monitoreo de los pacientes. En este trabajo se tomó como base un estudio anterior en el que se analizaron combinaciones de alarmas de monitores de pacientes para predecir eventos de código azul en hospitales, patrones a los que denominaron *SuperAlarm*. Al realizar la integración de datos señalada anteriormente, se mejora la predicción de eventos de código azul en una ventana de tiempo de 1 hora previa al incidente (Bai et al., 2015).

2.1.1 Escalas aplicadas en LMA

Para la LMA se han creado escalas con el fin de evaluar las condiciones de los pacientes o para realizar predicciones sobre su estado futuro. Dentro de este último grupo de escalas se puede encontrar la escala CIBMTR, que predice la supervivencia general en pacientes con LMA activa en recaída, para quienes se está considerando el trasplante de células madre hematopoyéticas (*Hematopoietic Stem Cell Transplant*, siglas en inglés HSCT). Para su cálculo se toman 5 variables que influyen significativamente en la supervivencia: duración de la primera remisión completa menor a 6 meses, blastocitos circulantes, donante distinto del hermano HLA (*Human leukocyte antigen*) idéntico, puntuación en la escala de Karnofsky o Lansky inferior a 90 y citogenética de bajo riesgo. Para su obtención se realizó un cálculo multivariable de las variables previas al trasplante y se desarrolló el sistema de puntuación con la información de 2,255 pacientes que se sometieron a un trasplante por leucemia aguda en recaída o con falla de inducción primaria después de un régimen de acondicionamiento mieloablativo entre 1995 y 2004 (Duval et al., 2010).

Otra escala fue creada para medir el tratamiento post remisión óptimo de la LMA. La escala *Post-Remision Treatment* (PRT) fue desarrollada sobre la base de 586 pacientes con LMA entre los 15 y 60 años, tratados en el ensayo prospectivo AML96 del *Study Alliance Leukemia* (SAL). Todos los pacientes habían logrado una remisión completa después de haber recibido una terapia de doble inducción. La puntuación de la escala proporciona 3 grupos de riesgo diferentes (favorable, intermedio y no favorable) con respecto a la supervivencia general

después de la remisión completa. Se evaluó la asociación entre variables potencialmente pronósticas y la supervivencia general después de la remisión completa, usando el análisis de regresión Cox estratificado. Con las variables significativas obtenidas se desarrolló la escala y se validó con los datos de 407 pacientes pertenecientes a la prueba AML2003. Los grupos de puntuación de esta escala podrían ayudar a los médicos a adaptar el tratamiento para los pacientes con LMA (Pfirrmann et al., 2012).

También existen escalas para la predicción de la supervivencia de un paciente a una de las alternativas de tratamiento para la LMA como lo es el HSCT. Se pueden encontrar escalas usadas ampliamente como la HCT-CI (*Hematopoietic Cell Transplantation - Comorbidity Index*) y la EBMT (*European Group for Blood and Marrow Transplantation*). La escala HCT-CI fue desarrollada para permitir la evaluación de riesgo antes de la realización de un trasplante alogénico. Esta escala incorpora medidas específicas de funcionalidad de órgano y comorbilidades relevantes. Fue creada analizando datos de 1,055 pacientes entre los años de 1997 y 2004, tratados en el *Seattle Cancer Care Alliance* (Sorror et al., 2005). Por su parte, la escala EBMT ofrece una manera sencilla para evaluar los riesgos de un paciente que va a ser sometido a un HSCT. Toma 5 factores para calcular el riesgo asociado para el paciente, entre los que se encuentran edad del paciente, etapa de la enfermedad, tiempo desde el diagnóstico, tipo de donante y combinación de género del donante-receptor (Gratwohl, 2012).

2.1.2 Sistemas y aplicaciones

Las escalas y modelos de predicción han sido tenidos en cuenta para la creación de aplicaciones para teléfonos inteligentes. Estas aplicaciones son creadas para facilitar el acceso a las escalas más utilizadas por el personal médico, debido a la penetración de esta tecnología en el mundo, permitiendo diligenciar información para realizar la valoración de los pacientes. Se cuenta con aplicaciones como *Medical Scales*[1] que incluye escalas como la Escala de Coma de Glasgow (*Glasgow Coma Scale*, siglas en inglés GCS), de aplicación neurológica, que permite medir el nivel de conciencia de una persona, utilizada en casos de pacientes con traumatismo craneoencefálico; o la escala de agitación y sedación Richmond (*Richmond Agitation-Sedation Scale*, siglas en inglés RASS), la cual mide el estado de sedación de los pacientes, utilizada en unidades de cuidados intensivos en pacientes bajo sedación médica.

[1] URL: https://play.google.com/store/apps/details?id=com.vicentesg.escalasmedicas.free&hl=en

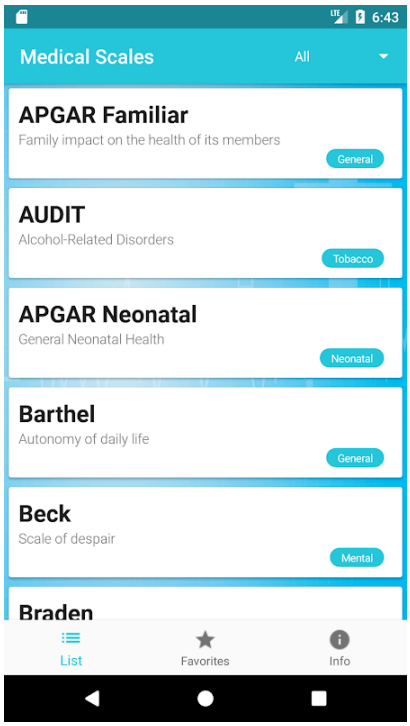

Figura 5. Aplicación Medical Scales
(Fuente: https://play.google.com/store/apps/details?id=com.vicentesg.escalasmedicas.free&hl=en)

Existen otros sistemas que, además de poner a disposición las escalas más utilizadas al personal de salud para su consulta en teléfonos inteligentes, también permiten realizar consultas desde el navegador de Internet de cualquier equipo o computador. Tal es el caso de QxMD[2], que ofrece un sitio web fácil de usar para buscar escalas por especialidad o de manera alfabética, cuya vista de su interfaz se puede observar en la figura 6. En complemento, este sitio ofrece la posibilidad a los profesionales de salud de hacer parte de su comunidad de colaboradores para validar y mejorar las escalas existentes.

En la línea de los sistemas de información, existen algunos disponibles a través de Internet, en los que puede ser buscada la escala por especialidad o nombre, y en ciertos casos, permite revisar el trabajo a partir de cual fue creada. Dentro de estos, se pueden encontrar sistemas como *Rapid Critical Care Consult*[3] y *Medscape*[4]. Esta última, además de consolidar algunas de las escalas más utilizadas, es un sitio de referencia y capacitación para la comunidad médica y profesionales de la salud, con noticias, perspectivas, opiniones actualizadas de los expertos, e información sobre medicamentos y enfermedades en función del paciente.

[2] URL: https://qxmd.com/calculate
[3] URL: https://www.rccc.eu/
[4] URL: https://reference.medscape.com/guide/medical-calculators

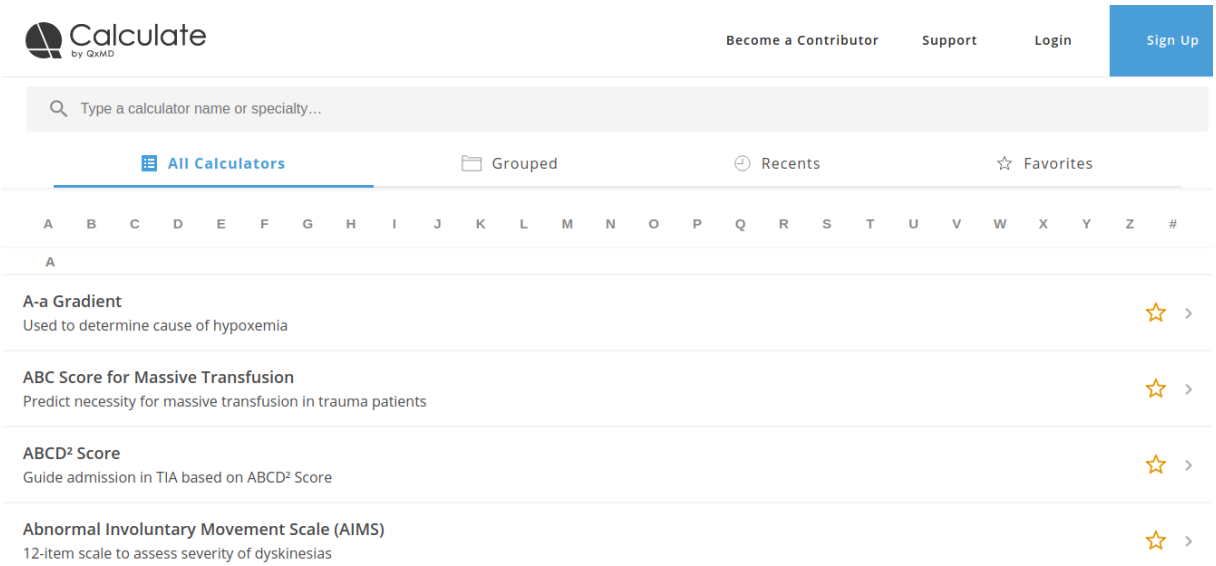

Figura 6. Aplicación QxMD
(Fuente: https://qxmd.com/calculate)

2.2 Inteligencia artificial y medicina

En el transcurso de los últimos años se ha probado la utilidad del uso de la Inteligencia Artificial en diferentes campos de la medicina. En el área de tratamiento de datos, el aprendizaje automático tiene un gran potencial en la identificación de patrones que pueden parecer ocultos al razonamiento común, ocultos al ser humano que no tiene la capacidad de detectarlos de manera eficiente, rápida y precisa. La aplicación de la Inteligencia Artificial juega además un papel importante teniendo en cuenta la cantidad, cada vez más grande y compleja, de datos que se genera cada día por todos los actores del sistema de salud, y la dificultad de poder extraer información de utilidad con los métodos tradicionales de análisis de datos. La Inteligencia Artificial aplicada en el campo de la medicina tiene la capacidad de ayudar a mejorar la calidad de vida de las personas y, haciendo uso de sus capacidades predictivas, podría salvar vidas al identificar posibles problemas que podrían afectar a un paciente en un futuro.

Con la notable mejora de los modelos de Inteligencia Artificial ha aumentado a su vez su aplicación en estudios médicos, como lo refleja el aumento de trabajos de investigación que aplican estos algoritmos para la solución de diferentes problemas. Como prueba de ello, en la figura 7 se evidencia la cantidad de artículos publicados durante los últimos 30 años en MEDLINE, una de las bases de datos más reconocidas a nivel mundial, y clasificados bajo sus términos normalizados como “Inteligencia Artificial”.

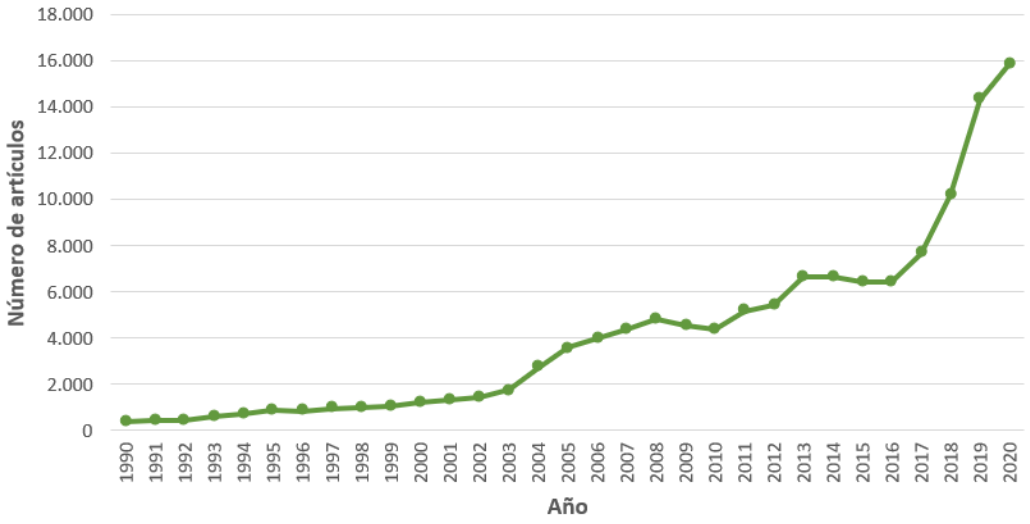

Figura 7. Aumento de estudios de Inteligencia Artificial aplicados al campo de medicina
Fuente (https://pubmed.ncbi.nlm.nih.gov). Elaboración propia.

A continuación, se muestra un panorama general de la aplicación de la Inteligencia Artificial en diferentes campos de la medicina, con algunos ejemplos de trabajos y estudios realizados en los últimos años.

2.2.1 Pronóstico y prevención

Recientemente, se han realizado investigaciones sobre la enfermedad cardiovascular aplicando modelos de aprendizaje automático y aprendizaje profundo en la predicción de la mortalidad de pacientes con disección espontánea de arterias coronarias (*Spontaneus Coronary Artery Disection*, siglas en inglés SCAD). A pesar que el curso clínico de la SCAD es variable, y que actualmente no existen métodos disponibles para realizar este tipo de predicción, se aplicaron los modelos mencionados a variables de registros de salud electrónicos de pacientes intrahospitalarios con SCAD para predecir su mortalidad. El índice de mortalidad durante la hospitalización de los casos analizados fue de 11.5%. Se compararon diversos modelos de aprendizaje automático a casos con información clínica completa, en los cuales los modelos de aprendizaje profundo mostraron los mejores resultados con alta precisión predictiva con AUC (*Area Under Curve*) de 0.98 (95% CI 0.97-0.99) (Krittanawong et al., 2021).

Siguiendo la misma especialidad de la enfermedad cardiovascular, se han propuesto muchas herramientas para evaluación y prevención del riesgo como la *Framingham risk score* (escala de riesgo Framingham), *Pooled Cohort Equation* (PCE), *Systematic Coronary Risk Evaluation* (SCORE) y QRISK3, las cuales son ampliamente usadas, y se han conducido estudios para la verificación y comparación de diferentes *scores*. En Corea del Sur, con datos de 222,998 personas entre los 40 y 79 años, sin historia de enfermedad cardiovascular, se realizó una comparación de modelos y escalas predictivas preexistentes de riesgo cardiovascular, como las mencionadas anteriormente, con algoritmos basados en aprendizaje automático. Los modelos previos mostraron moderada a buena discriminación para predecir futuros eventos cardiovasculares (C-statistics 0.70–0.80). Usando un modelo de red neuronal se alcanzó la mayor C-statistic (0.751), la cual fue significativamente más alta entre las comparadas. Este estudio muestra que los algoritmos de aprendizaje automático pueden aumentar la capacidad de prevención de riesgo cardiovascular sobre otros modelos o escalas existentes, y la capacidad de estos modelos para ser adoptados en otras áreas para evaluación de riesgo y toma de decisiones clínicas (Cho et al., 2021).

2.2.2 Diagnóstico

La Inteligencia Artificial ofrece una gran capacidad de reconocimiento de patrones complejos a pesar del ruido que pueda contener la fuente de datos analizada. Esto es especialmente útil en el área de análisis de imágenes diagnósticas como las que se procesan en el campo de la radiología, en donde su aplicación puede permitir la generación de modelos tridimensionales a partir de imágenes de pacientes concretos (Hosny et al., 2018). Hay un gran debate acerca de la prontitud con la que se implementarán los nuevos métodos de aprendizaje profundo en la práctica clínica radiológica. Se habla de unos pocos años hasta décadas, donde el desarrollo de soluciones automatizadas tomará la mayoría de problemas clínicos donde se cuente con suficiente cantidad de datos disponible. Aún hay un camino por recorrer debido a que con muchos algoritmos de segmentación automáticos no se tienen resultados óptimos, y en el caso de enfermedades poco comunes, no existen.

En la edición 22 de la Revista Colombiana de Reumatología del año 2015 (González, 2015), se incluyó un artículo de investigación en el que se menciona la utilización de modelos de aprendizaje computacional para la clasificación de enfermedades haciendo uso de datos clínicos, genéticos y serológicos. En esta investigación los autores hicieron uso algoritmos de aprendizaje supervisado, como redes bayesianas y redes neuronales, que obtuvieron buenos resultados en la identificación de enfermedades como la artritis reumatoide, alcanzando una

sensitividad y especificidad por encima del 92% en el proceso de clasificación (Morales Muñoz et al., 2015).

2.2.3 Tratamiento

Se está usando el aprendizaje automático para analizar la efectividad de los tratamientos médicos. Un estudio publicado en 2021, menciona la creación de la clasificación de medicamentos usando aprendizaje automático (*Drug Ranking Using Machine Learning,* siglas en inglés DRUML). En el DRUML se tienen en cuenta medicamentos para el tratamiento del cáncer, y se usaron modelos de aprendizaje automático entrenados con los datos de respuesta de células a más de 400 medicamentos, clasificándolos en función de su eficacia prevista para reducir la proliferación de una determinada población de células cancerosas. Inicialmente se evaluaron métodos de aprendizaje como estimación bayesiana de modelos lineales generalizados, mínimos cuadrados parciales, *Random Forest*, máquinas de vector de soporte (*Support Vector Machine*, siglas en inglés SVM), redes neuronales y modelos de aprendizaje profundo, los cuales mostraron las mejores métricas. Los resultados mostraron que DRUML clasifica los fármacos de diferente modo de acción en función de su eficacia prevista en diferentes tipos de cáncer con un error razonablemente bajo. En última instancia, DRUML podría ayudar en la priorización de fármacos al complementar la información obtenida de los parámetros clínico-patológicos y el análisis mutacional (Gerdes et al., 2021).

2.3 Inteligencia Artificial y LMA

Los avances realizados hasta este momento para comprender completamente la enfermedad no han sido perfectos. Se han realizado diferentes estudios en los últimos años aprovechando los diferentes campos de acción de la Inteligencia Artificial, obteniendo resultados prometedores en el análisis de esta y otras enfermedades, que suponen un mejor porvenir para la humanidad.

En el área de visión computacional, se han adelantado estudios como el especificado por (Kimura et al., 2019), en el que se desarrolló un sistema automático de apoyo al diagnóstico, que combina un sistema de reconocimiento de imágenes de células sanguíneas usando un modelo de aprendizaje profundo impulsado por redes neuronales convolucionales (*Convolutional Neural Network*, siglas en inglés CNN), y un sistema de decisión XGBoost. Se utilizó un conjunto de datos con imágenes de 695,030 células sanguíneas, obtenidas de 3,261

frotis de sangre periférica, que logra diferenciar Síndromes Mielodisplásicos (*Myelodysplastic syndromes,* siglas en ingles MDS) de la anemia aplásica (AA) con un 96.2% de sensibilidad y 100% de especificidad (AUC 0.99). En esta misma línea, se entrenó un modelo de CNN con muestras de 10,000 células que fueron anotadas manualmente, provenientes de frotis de aspirado de médula ósea (*Bone Marrow Aspirate*, siglas en inglés BMA), con el fin de realizar el análisis para la clasificación de trastornos hematológicos usando tejido no neoplásico. Se obtuvo como resultado que el modelo realiza la detección y clasificación con un AUROC (*Area Under the Receiver Operating Characteristic*) de hasta 0.98, mostrando niveles similares al utilizar muestras de LMA (Chandradevan et al., 2020).

También se ha aprovechado la gran capacidad de análisis de volúmenes de datos que los algoritmos de Inteligencia Artificial poseen. En esta área se pueden encontrar trabajos como el publicado en la revista *Nature Communications* en el año 2018 (Lee et al., 2018). En este estudio se analizó el efecto de los medicamentos en los pacientes, creando un nuevo método computacional basado en aprendizaje automático que identifica marcadores de expresión genética fiables para la sensibilidad a los fármacos, integrando información previa multiómica relevante para los procesos de la enfermedad. Se tomó como base los datos de 30 pacientes con LMA incluidos los perfiles de expresión genética de todo el genoma y sensibilidad in vitro a 160 medicamentos utilizados en quimioterapia.

De igual manera en esta área, se construyó un nuevo modelo conjunto optimizado, combinando un modelo DNN (*Deep Neural Network*) con dos modelos de aprendizaje automático, para la predicción de enfermedades utilizando como conjunto de datos los resultados de pruebas de laboratorio, incluyendo resultados de pruebas de sangre y de orina, relacionados con el diagnóstico final de cada paciente al momento del alta. Se tomó una muestra de 5,145 casos con 88 atributos, incluyendo el sexo y la edad del paciente, y se realizó la investigación sobre un total de 39 enfermedades específicas. Para el caso de la LMA, el modelo entrenado obtuvo excelentes resultados, realizando una predicción con un AUC de 0.99 (Park et al., 2021).

Del mismo modo, en el área de análisis de datos, se encuentran estudios asociados al riesgo de desarrollar enfermedad de injerto contra huésped aguda (*acute Graft-versus-Host Disease*, siglas en inglés aGVHD) después de HSCT en el tratamiento de la LMA. En Japón, con datos de 26,695 pacientes, se entrenó un modelo *Alternating Decision Tree* (ADTree) que predice el riesgo de desarrollar aGVHD grado II-IV y III-IV, con una AUC de 0.616 y 0.622 respectivamente (Arai et al., 2019). En Estados Unidos, se usaron datos de 324 pacientes, entre los que se incluyeron las mediciones de signos vitales tomadas desde el día de la

infusión (día 0) hasta el día 9 después del trasplante. Con esta información se creó un conjunto de datos para realizar el entrenamiento de un modelo *L2-regularized logistic regression* que puede realizar la predicción de desarrollar aGVHD grado II-IV al día 100 posterior al trasplante, con un AUC de 0.659 (P=0.019) (Tang et al., 2020). En Irán, se desarrolló un sistema de ayuda para toma de decisiones que puede ser consultado a través de Internet[5]. Para su desarrollo, se creó un conjunto de datos que contaba con la información de 182 pacientes y 31 variables, el cual fue utilizado para entrenar modelos de clasificación entre los que se destacó el *Extreme Gradient Boosting Classifier* (XGBCClassifier), que puede predecir, el día del trasplante, el riesgo del paciente de desarrollar la aGVHD con una precisión de 90.7%, sensibilidad de 92.5% y especificidad de 89.13% (Salehnasab, 2021).

2.4 Estudios de LMA en Colombia

En Colombia se han realizado estudios destinados a la investigación de diferentes tipos de enfermedades y cáncer hematológico. Entre los temas de los artículos asociados a la leucemia se pueden encontrar los estudios de cohortes, las series de casos, las pruebas diagnósticas en las principales ciudades, y el análisis de mortalidad infantil a nivel nacional. La mayoría de las investigaciones realizadas se centran en estudios epidemiológicos y clínicos, dejando un poco de lado el análisis biológico y molecular de la enfermedad, lo cual debería ser un aspecto a considerar si se tiene en cuenta la particular etiología que parece tener esta enfermedad en el país, comparado con otros, debido posiblemente al perfil genómico de la población (Gacha Garay et al., 2017).

Recientemente fue publicado un estudio sobre la implementación del protocolo PETHEMA LPA 99 en el tratamiento de niños con leucemia promielocítica aguda (LPA) en Bogotá D.C. Este tipo de leucemia es un subtipo de la LMA. La motivación de este estudio es la mayor incidencia de este subtipo de leucemia en Latinoamérica comparado con otros lugares como Estados Unidos. En él se recopila la experiencia en el tratamiento de la LPA durante 7 años, con seguimiento hasta de 11.4 años. Se aplicó el protocolo PETHEMA LPA 99, diseñado para adultos, pues ha mostrado una supervivencia global (SG) mayor a 80%. Como parte de las limitaciones del estudio, se puede encontrar que se trata de una cohorte retrospectiva y que no fue realizado un análisis molecular a todos los pacientes. De la misma manera, en el estudio se analizó un número limitado de pacientes exclusivamente menores de 18 años. En el trabajo se concluye que, en términos generales, realizar la implementación del protocolo

[5] Clinical Decision Support System. URL: https://agprcdss.ir/dappx/about

PETHEMA LPA 99 en el tratamiento de la población de estudio arroja resultados satisfactorios. Aunque los valores de supervivencia global reportados fueron menores a los descritos en poblaciones pediátricas de otros países, es posible recomendar el uso del protocolo PETHEMA LPA 99 en población pediátrica una vez ajustado a las características de este grupo etario según las recomendaciones del mismo (Pardo-Gonzalez et al., 2020).

En otro estudio publicado en el año de 2020, teniendo en cuenta que la leucemia es el cáncer más común en la niñez y que su tasa de incidencia en Colombia es una de las más altas de América, se propusieron conocer la distribución espacio-temporal de esta enfermedad (Rodriguez-Villamizar et al., 2020). En este trabajo se analizaron 3,846 casos de niños menores de 15 años con diagnóstico confirmado de leucemia aguda entre 2009 y 2017. Se usaron estadísticas de escaneo espacio-temporal de Kulldorff, incluyendo el municipio y año de diagnóstico en el análisis. En este estudio se identificaron cinco grupos espaciales de leucemia infantil en diferentes regiones del país y grupos de tiempo específicos durante el período de estudio. Este resultado permite llegar a la conclusión que existen factores etiológicos o condiciones comunes por región asociados a la enfermedad que deberían ser estudiados.

A nivel preclínico, se han desarrollado en el mundo modelos para el estudio del cáncer con el fin de realizar una clasificación y caracterización de las leucemias, y la determinación de posibles factores causales. Siguiendo esta línea, se han realizado análisis en peces transgénicos para modelar la leucemia, desarrollando líneas que buscan establecer similitudes genéticas con esta patología entre estas especies y el ser humano (Gacha Garay et al., 2017). Con base en lo anterior, en el país se adelantó un estudio en el que se obtuvieron resultados prometedores usando al pez cebra para el estudio preclínico de la leucemia, realizando experimentos de xenoinjertos de células Jurkat.

En el año 2015 se realizó un estudio donde se evaluó el costo-efectividad de las alternativas de tratamiento de consolidación en niños con LMA. Por una parte, el trasplante alogénico con progenitores hematopoyéticos, y por la otra el tratamiento con quimioterapia. Los datos para su elaboración fueron extraídos de estudios y reportes encontrados en la literatura científica, al igual que del Sistema de Información de Precios de Medicamentos (SISMED) del Ministerio de Salud y Protección Social para establecer el precio de los medicamentos teniendo en cuenta el año de referencia. El análisis realizado en el estudio se basó tomando como resultado los años de vida ganados después del tratamiento. Realizando cálculos de sensibilidad y probabilísticos, se concluyó que el trasplante resulta ser costo-efectivo en Colombia frente al tratamiento con quimioterapia (García et al., 2015).

2.5 Conclusiones de la revisión

Tras la revisión documental de trabajos con relevancia técnica y médica para la realización del presente trabajo, se verificó cómo en el ámbito mundial se realizan esfuerzos e investigaciones enfocadas a la aplicación de la Inteligencia Artificial, extrapolando las lecciones aprendidas en otras áreas de la sociedad y la industria, en busca de apoyar y mejorar el desempeño médico en los campos del pronóstico, diagnóstico y tratamiento.

La selección de la opción adecuada, dentro del abanico de posibilidades ofrecidas por los algoritmos de aprendizaje supervisado, varía dependiendo del origen de los datos analizados y del problema abordado. Se observaron modelos de gran utilidad en el tratamiento de imágenes, modelos basados en análisis probabilístico, geométrico o en árboles de decisión, que ofrecen cada uno diferentes bondades para alcanzar los objetivos buscados. El *Deep Learning*, apalancado en el aumento de la capacidad de almacenamiento y cómputo de las últimas décadas, es una opción de gran relevancia en los bancos de pruebas que buscan un modelo óptimo en diferentes tipos de aplicaciones.

Los modelos y sistemas predictivos o de diagnóstico pueden llegar a ser de gran utilidad en la práctica médica, apoyando al personal y siendo de ayuda para fortalecer las limitaciones humanas como la capacidad máxima de procesamiento simultáneo y el cansancio físico y mental. También han demostrado que mejoran ciertas tareas, pues se vio en trabajos como el presentado por Park, publicado en la revista *Scientific Reports* (Park et al., 2021), que los modelos generados superaron en ocasiones la capacidad humana. Cabe señalar que se debe ser prudente aún con las herramientas generadas, teniendo en cuenta que los modelos predictivos desarrollados no son infalibles, y que en algunos casos en su generación se incurre en errores que derivan al final en su poca o nula popularización y utilización en la práctica clínica a gran escala.

En Colombia, en el área en particular, son pocas las publicaciones asociadas a la aplicación de la Inteligencia Artificial para el pronóstico o diagnóstico de enfermedades, situación que es diferente si se compara con la producción de la comunidad científica internacional. Sería importante abordar el tema buscando ofrecer una verdadera utilidad, y posteriormente avanzar en otros campos de manera similar a como se hace en otras partes del mundo, manteniendo así al país a la vanguardia del desarrollo y de la tecnología. Esto puede también llevar a tomar conciencia de la importancia de administrar de manera seria y correcta la información para su posterior análisis, lo que propende hacia el mejoramiento de los procedimientos médicos y tratamientos disponibles para los pacientes.

La Inteligencia Artificial aplicada al campo de medicina no se encuentra totalmente explorada, pudiendo ser de gran utilidad para la humanidad pues su aporte ha sido verificado en diferentes áreas. Teniendo en cuenta los algoritmos, herramientas utilizadas y recomendaciones leídas en los trabajos del estado del arte, es posible realizar un trabajo de calidad que sirva de base para el desarrollo de una herramienta útil, con la posibilidad de continuar avanzando en la consecución de una aplicación útil y de extender las lecciones aprendidas a otros ámbitos de la medicina.

3. Objetivos y metodología de trabajo

3.1. Objetivo general

Crear un modelo piloto de aprendizaje supervisado que permita predecir la mortalidad en pacientes con LMA, basado en información clínica de pacientes tratados en un hospital universitario de la ciudad de Floridablanca, Colombia.

3.2. Objetivos específicos

- Determinar las variables más relevantes para la creación y entrenamiento del modelo de aprendizaje supervisado, realizando una exploración de los registros de la base de datos de pacientes con LMA analizada, bajo la supervisión de profesionales médicos expertos.
- Crear un conjunto de datos, a partir de los registros de pacientes con LMA, que sirva de base para el entrenamiento de modelos de aprendizaje supervisado.
- Aplicar algoritmos de aprendizaje supervisado para la creación de modelos predictivos de mortalidad, utilizando el conjunto de datos obtenido a partir de la base de datos de pacientes con LMA.
- Seleccionar el modelo de aprendizaje supervisado que realice las mejores predicciones de mortalidad, después del análisis y evaluación de algunos de los algoritmos más utilizados en problemas de clasificación en Inteligencia Artificial.
- Crear un prototipo de interfaz de usuario que permita acceder al modelo piloto predictivo desarrollado.
- Crear un prototipo de interfaz de usuario que permita realizar el entrenamiento del modelo piloto desarrollado con nuevos conjuntos de datos basados en el protocolo PETHEMA.

3.3. Metodología del trabajo

Este trabajo es un análisis anidado al proyecto “Registro epidemiológico de pacientes adultos con Leucemia Mieloide Aguda" que se desarrolla en un hospital universitario del oriente colombiano, en cabeza de la Jefe del Servicio de Hematología y Unidad de Trasplante y Terapia Celular de dicha institución. Para la conformación de la base de datos se tomaron las variables recopiladas en el registro LMA, consignadas de manera anonimizada. Tomando como base este registro de pacientes atendidos, se planteó la realización de un estudio poblacional retrospectivo en el que se tuvo en cuenta la información clínica de pacientes

mayores de 18 años diagnosticados con LMA según los criterios de la Organización Mundial de la Salud (OMS). Con corte 30 de junio de 2021, se cuenta con el registro de 169 pacientes, con información de variables que recopilan datos de diferentes momentos de la atención, evaluación y tratamiento de los pacientes, tales como:

- Datos sociodemográficos: edad, tipo seguridad social, estado civil, nivel educativo, área de procedencia, entre otros.
- Datos clínicos al diagnóstico: fecha de diagnóstico, pruebas iniciales, parámetros laboratorio, inmunofenotipo, citogenética.
- Datos de tratamiento: tipo de tratamiento, fechas inicio/fin, modificación/retraso/ suspensión de dosis, toxicidades.
- Evaluación de la respuesta: respuesta, fecha de evaluación, técnica.
- Seguimiento: recaída, fecha de recaída, estado en el último control, fecha último control.

El esquema de la metodología general que se siguió en la realización de este trabajo (Figura 8) es característico en proyectos de minería de datos, *data science* o Inteligencia Artificial. Se inició con la exploración y análisis del registro LMA. Se parte del origen de los datos, entendiendo cómo fue adquirida la información y su relación con la enfermedad analizada. Este proceso de familiarización permitió entender la información disponible y su significado. En este punto se analizó la distribución de los datos y se tomaron decisiones dependiendo de los vacíos encontrados, del balance y de la correlación de la información contenida.

Se determinaron las variables más relevantes para la elaboración del modelo predictivo, lo cual fue realizado con el acompañamiento de la Jefe del Servicio de Hematología y Unidad de Trasplante y Terapia Celular. Al final de esta etapa se obtuvo el conjunto de datos con el que se realiza el entrenamiento de los posibles modelos. Cabe señalar que como resultado de este proceso se obtuvieron diferentes conjuntos de datos cuya utilidad real se comprobó después de realizar las predicciones y evaluaciones de los modelos obtenidos a partir de aquellos.

Después de realizar la depuración del conjunto de datos, se realizó la división de la información para obtener datos de entrenamiento y datos de prueba, con el fin de validar los resultados. Con los primeros se procedió a realizar el proceso de entrenamiento aplicando diferentes algoritmos de aprendizaje automático, dando lugar a diferentes modelos. Con cada uno de estos se realizaron predicciones que fueron comparadas con los datos de prueba, separados inicialmente, para realizar la evaluación de las métricas obtenidas en cada caso. El siguiente proceso se realizó de manera iterativa con cada algoritmo utilizado para la

generación de los modelos a analizar: ejecución de entrenamiento, realización de predicciones, evaluación del modelo obtenido, y ajuste de los parámetros del mismo. Posterior a la realización de las iteraciones de entrenamiento, predicción, evaluación y afinación de todos los modelos utilizados, se verificó con cuál de ellos se obtienen los mejores resultados a la luz de la información disponible.

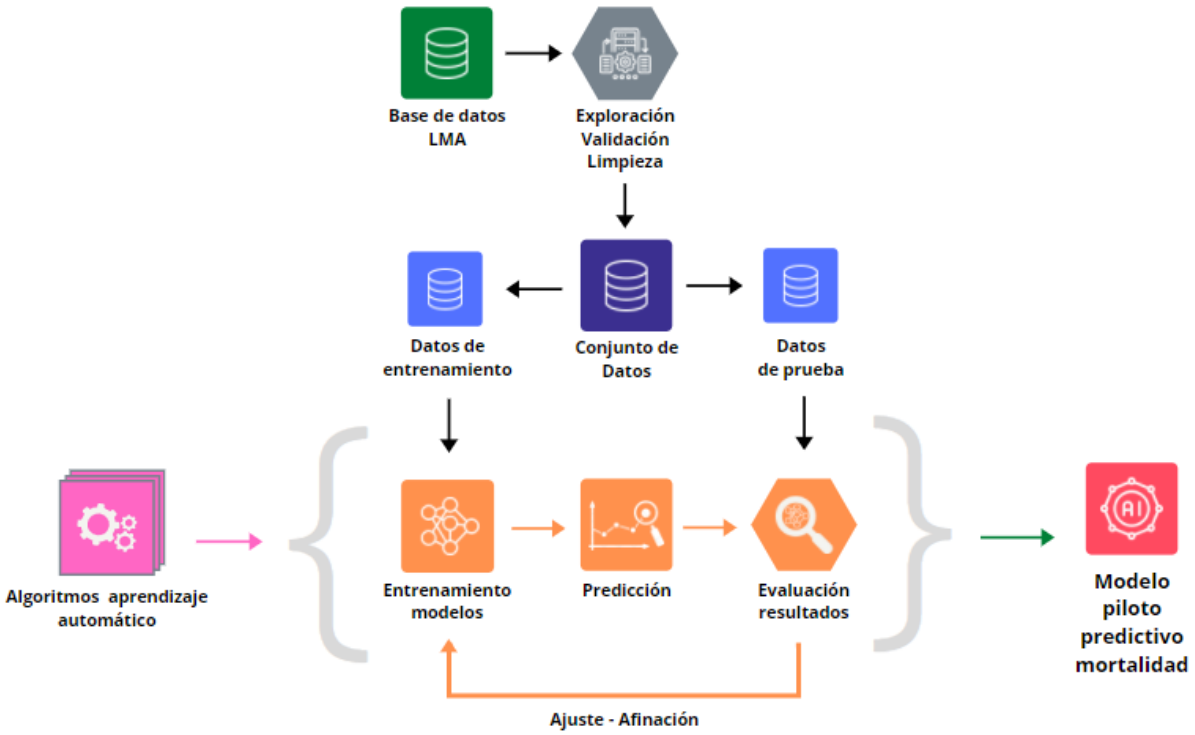

Figura 8. Metodología general

Debido a la cantidad de registros disponible, en el proceso de entrenamiento de los modelos se tuvo en cuenta el riesgo de obtener sobreajuste (*overfitting*), fenómeno que se presenta en los modelos entrenados que se puede resumir en que estos aprenden muy bien las características del conjunto de datos trabajado, perdiendo su capacidad de generalización posterior. Como medida para prevenir esto, se hizo uso de la técnica de validación cruzada usando los datos de entrenamiento, algo que también sirvió para tener una idea de la capacidad predictiva de los modelos.

Finalmente, después de la selección del modelo con mejores predicciones, se implementó la interfaz de usuario para acceder al modelo piloto predictivo. Esta interfaz se propone como punto de partida para continuar aplicando el análisis realizado en este trabajo de manera posterior, utilizando nuevos conjuntos de datos con más observaciones, extrapolando este análisis a una población mayor que sea tratada según el protocolo PETHEMA.

Con el fin de realizar un análisis de resultados, se realizó una evaluación crítica del modelo desarrollado haciendo uso de la lista de verificación PROBAST. Esta es una herramienta en cuyo desarrollo participaron revisores expertos en los campos de la investigación de modelos predictivos y el desarrollo de herramientas de evaluación de calidad, que consiste de 20 preguntas guía distribuidas en 4 dominios (participantes, predictores, resultado y análisis). Fue diseñada para revisiones sistemáticas, y puede ser usada para realizar una evaluación crítica de modelos de predicción ofreciendo una guía para analizar el riesgo de sesgo y la aplicabilidad a la población y entorno previstos. Todos los conceptos de esta lista son ilustrados con ejemplos publicados en diferentes temas que pueden consultados en su sitio web[6] (Wolff et al., 2019).

[6] Prediction Model Risk of Bias Assessment Tool: https://www.probast.org/

4. Identificación de requisitos

En este trabajo se realizó un análisis de datos piloto usando algoritmos de Inteligencia Artificial en el área de Hematología y Unidad de Trasplante y Terapia Celular. Este análisis busca aprovechar los beneficios que la Inteligencia Artificial puede llegar a ofrecer, beneficios que son una realidad en el mundo en diferentes ámbitos, incluyendo en medicina, motivo por el cual la comunidad científica internacional ha integrado esfuerzos alrededor del tema.

Se proyectó crear una interfaz de usuario con funciones básicas que permita implementar de manera inicial el modelo piloto de predicción obtenido, para que pueda ser utilizada posteriormente en análisis del mismo tipo que el presentado en este trabajo. Teniendo en cuenta esto, y el procedimiento a seguir para el cumplimiento de los objetivos de este trabajo, se definieron los siguientes requisitos de alto nivel:

Código	Requisito	Entregable
RE01	Exploración y análisis de datos de archivos.	Resultados de Análisis de datos recibidos.
RE02	Creación de conjunto de datos para entrenamiento.	Conjunto de datos con variables características.
RE03	Entrenamiento de modelos de predicción usando los algoritmos de aprendizaje automático seleccionados después de la revisión del estado del arte.	Scripts de código con el proceso de cada algoritmo.
RE04	Determinación de hiperparámetros óptimos de entrenamiento de los algoritmos de aprendizaje automático utilizados, ajustados al conjunto de datos procesado.	- Scripts del proceso con la implementación del algoritmo de exploración de parámetros. - Hiperparámetros con los que se obtienen los mejores resultados en el entrenamiento de algoritmos.
RE05	Selección del mejor modelo predictivo según las métricas obtenidas en el proceso de entrenamiento.	- Métricas de modelos predictivos. - Gráficas comparativas. - Cuadros comparativos.
RE06	Implementación de interfaz de usuario para el acceso al modelo de entrenamiento y predicción.	Interfaz implementada y funcional.
RE07	Función de carga de archivos en la interfaz de usuario para entrenamiento del modelo. La estructura de los archivos debe ser acorde a la especificación de PETHEMA.	Funcionalidad implementada en la interfaz de usuario.
RE08	La interfaz de usuario debe tener módulo para realización de predicciones de un paciente.	Funcionalidad implementada en la interfaz de usuario.

Tabla 1. Requisitos de alto nivel del desarrollo

5. Descripción de la herramienta software desarrollada

5.1 Modelo predictivo

5.1.1 Exploración de datos

En este trabajo se utilizaron datos que fueron recopilados durante el tratamiento de pacientes con LMA bajo el protocolo terapéutico establecido por PETHEMA. En términos generales, este protocolo consta de un tratamiento inicial de inducción buscando llegar a remisión completa (RC), para posteriormente aplicar un tratamiento de consolidación dependiendo de la edad del paciente (Onecha, 2019).

En su tesis doctoral, Onecha (2019) menciona algunos datos estadísticos obtenidos del tratamiento de la LMA:

> Aproximadamente el 20% de los pacientes con LMA presentan refractariedad primaria al tratamiento de inducción y no llegan a alcanzar RC. Y en torno al 50% de los casos la enfermedad resurge y el paciente recae. Ambos escenarios, refractariedad primaria y recaída, suponen fracaso terapéutico asociado a pronóstico adverso y menos del 30% de los pacientes sobreviven 12 meses después de una recaída. (p. 46)

De la misma manera, se menciona que “varios factores condicionan el pronóstico adverso en pacientes que sufren recaída, incluyendo citogenética adversa detectada en el diagnóstico, estado en RC menor a 12 meses, edad avanzada, y recaídas posteriores a un trasplante hematopoyético.” (Onecha, 2019, p. 46)

En el conjunto de datos recibido se encuentra información de las diferentes fases del tratamiento de los pacientes, desde la etapa de inducción, pasando por las diferentes etapas de consolidación según el desarrollo y respuesta obtenidos. Los pacientes no pasan por todas las etapas del tratamiento, algo que se puede inferir de lo citado anteriormente de Onecha (2019). Por esta razón se definió un punto de corte que sirviera de referencia, buscando que las condiciones de los pacientes fueran los más similares posibles y que se contara con la misma información de cada uno de ellos. Por este motivo, y teniendo en cuenta el protocolo terapéutico, se eligió trabajar con la información de los pacientes hasta la etapa de inducción,

donde se cuenta con información de pruebas de laboratorio, datos clínicos y algunos análisis genéticos.

La fundación PETHEMA pone a disposición una plataforma (PLATAFO-LMA), que permite el diligenciamiento de la información generada durante el tratamiento de los pacientes. Los diferentes centros médicos afiliados, que siguen el esquema propuesto por este programa español, reciben acceso a esta plataforma para llevar registro y seguimiento.

Figura 9. Pantalla de inicio de sesión de la Plataforma de PETHEMA

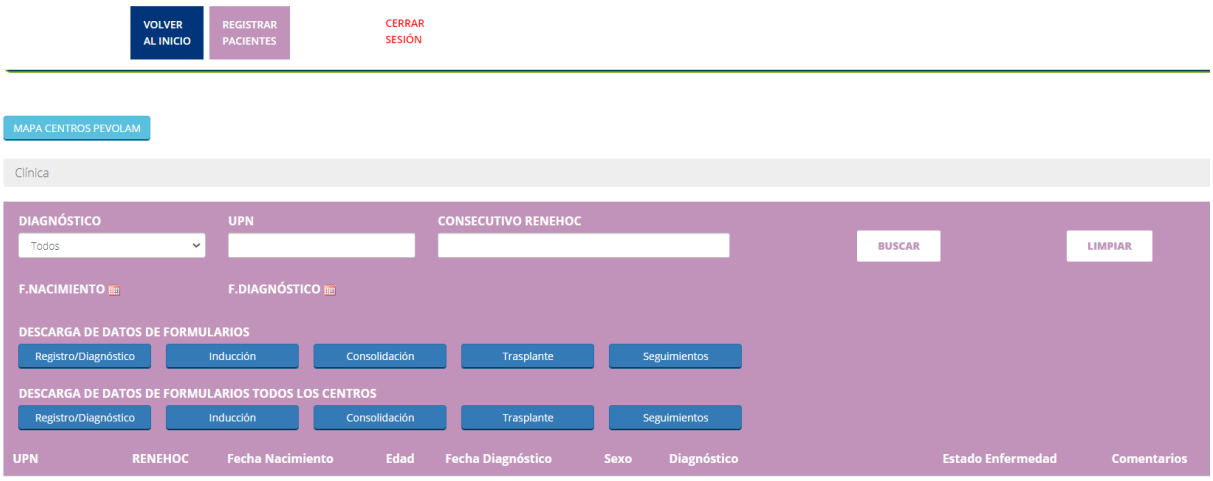

Figura 10. Imagen plataforma PETHEMA

Esta plataforma permite la exportación de la información de los pacientes del centro médico registrados en formato de archivo plano con caracteres de separación, en cuya primera línea se encuentra el nombre de los campos contenidos (Figura 11). Es posible descargar archivos consolidados que contienen información de las etapas del tratamiento. Al tratarse de información estandarizada que sigue el protocolo PETHEMA, la información registrada allí tendrá siempre el mismo formato y estructura. Por este motivo, el prototipo presentado en este trabajo sirve de punto inicial para extender este análisis a los demás pacientes de los centros médicos que siguen las mismas directrices. Esto tendría evidentes beneficios que impactarían positivamente en la calidad de las predicciones realizadas al contar con una mayor cantidad de registros disponibles para realizar el entrenamiento del modelo.

```
upn_number;finida;indesquema;nindesquema;ifpmn5;ifpmn1;ifplt5;ifplt100;ingini;fingini;faltaini;idiahosp;
570100001;5/09/2011;20;FLUGA;10/09/2011;10/09/2011;;;Si;25/08/2012;22/09/2012;29;2;Remisión completa inc
570100002;1/02/2012;8;Dauno+Ara-C (3+7);24/02/2012;27/02/2012;1/03/2012;3/03/2012;Si;1/02/2012;27/03/201
570100003;14/03/2012;20;FLUGA;2/04/2012;2/04/2012;;;Si;2/03/2012;3/04/2012;33;4;Resistencia;;;3/04/2012;
570100004;28/11/2012;20;FLUGA;27/12/2012;31/12/2012;14/03/2013;;Si;14/11/2012;9/01/2013;57;3;Remisión pa
570100005;13/05/2012;20;FLUGA;4/06/2012;5/06/2012;4/06/2012;8/06/2012;Si;10/05/2012;10/06/2012;32;1;Remi
570100006;21/01/2014;20;FLUGA;4/02/2014;4/02/2014;7/02/2014;13/02/2014;Si;15/01/2014;7/02/2014;24;1;Remi
570100007;16/02/2013;1;IDA+Ara-C (3+7);10/03/2013;13/03/2012;12/03/2013;16/03/2013;Si;8/01/2013;12/03/20
570100008;30/01/2014;20;FLUGA;30/01/2014;;;;Si;22/01/2014;10/03/2014;48;5;Muerte en inducción;10/03/2014
570100009;1/09/2013;20;FLUGA;25/09/2013;25/09/2013;20/09/2013;;Si;19/08/2013;26/09/2013;39;3;Remisión pa
570100010;19/09/2008;1;IDA+Ara-C (3+7);14/10/2008;14/10/2008;14/10/2008;14/10/2008;Si;15/09/2008;16/10/2
570100011;7/05/2010;21;MTZ+Ara-C (LA FE OLD);;;;;;;;;3;Remisión parcial (RP);;;;;;;;;;;;;;;;;;;;;;;;;;;
570100012;15/11/2009;1;IDA+Ara-C (3+7);7/12/2009;12/12/2009;4/12/2009;7/12/2009;Si;;;;1;Remisión complet
570100013;;0;Soporte/no tratado/oral (describir);;;;;;;;;0;No aplicable;;;15/06/2011;;;;;;;;;;;;;;;;;;;
```

Figura 11. Muestra de archivo de pacientes

Después de realizar una revisión del contenido de los archivos, se decidió trabajar con la información de los archivos de pacientes, inducción y seguimiento, teniendo en cuenta lo expuesto inicialmente en este numeral. Adicionalmente, se tuvo en cuenta un archivo de datos trabajado por la unidad de hematología del centro médico con el cual se trabajó, el cual cumple con el mismo formato y se encuentra relacionado por el identificador con los demás archivos obtenidos de PLATAFO-LMA.

Haciendo uso de las herramientas del lenguaje de programación Python versión 3.8.6, se contó la cantidad de registros y variables disponibles, y se realizó una exploración inicial para analizar los diferentes tipos de datos presentes dentro de los archivos analizados. Esto se debe tener en cuenta por las restricciones en materia de datos impuestas por los algoritmos de aprendizaje automático.

```
Conteo de registros y variables:

       Descripcion Cantidad
1  Total registros      169
2  Total variables      393

Tipos de datos encontrados:

                0
0           int64
1          object
2  datetime64[ns]
3         float64
```

Figura 12. Cantidad de registro y variables. Tipos de datos encontrados

El origen de datos cuenta con 169 registros con 393 variables. En el siguiente punto se relaciona el procedimiento de análisis, validación y limpieza de datos realizado.

5.1.2 Análisis, validación y limpieza de datos

A continuación se relaciona el procedimiento efectuado con la información del origen de datos paso a paso:

1. Se consultaron los registros que contaran con menos del 50% de variables pues implica que no cuentan con la información mínima del tratamiento dentro del registro y no deben ser tenidos en cuenta. Esto derivó en la eliminación de 4 registros. Resultado: cantidad de registros en el conjunto de datos = 165.

2. Se analizó el porcentaje de observaciones vacías (incompletas) dentro de las variables obtenidas de los archivos del origen de datos. Se consultaron las variables que tengan un porcentaje vacío mayor que 5%.

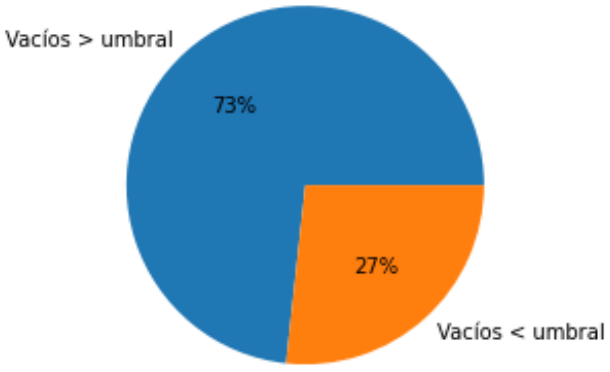

Figura 13. Porcentaje de columnas con cantidad de registros vacíos mayor al 5%

Se observó un porcentaje elevado de variables con observaciones incompletas. Esto es acorde a lo señalado anteriormente, debido a que todos los pacientes no han pasado por todas las etapas del tratamiento y en los archivos se reservan los campos para el registro de esta información. Se determinó que no se van a tener en cuenta variables con un porcentaje de datos vacíos mayor al valor del 5% que fue el consultado anteriormente. Resultado: cantidad de variables del conjunto de datos = 104.

3. Las variables de tipo fecha no se tuvieron en cuenta dentro del análisis pues no es posible incluirlas en el entrenamiento de los algoritmos. Dentro del origen de datos, las variables de este tipo incluyen información como la fecha de nacimiento, fecha de diagnóstico y fecha de muerte. En su lugar se generaron nuevos campos a partir de estas, para obtener valores numéricos que expresen la edad y el tiempo transcurrido hasta la aparición del evento muerte. La variable que contiene la fecha de muerte no se tuvo en cuenta en el análisis posterior por tratarse de la variable a partir de la cual se obtuvo la variable objetivo para el entrenamiento de los modelos. Resultado: cantidad de variables del conjunto de datos = 103.

4. Se revisaron las distribuciones de los valores de las variables disponibles. Este proceso es de utilidad porque permite explorar los valores de las variables en el origen de datos para toma de decisiones, o para validar si en el momento de cargar la información la herramienta utilizada asignó de manera incorrecta el tipo de datos de alguna variable. El proceso se inició con el conjunto de variables categóricas que, al momento de consultar, cuenta con 52 unidades. Por la extensión del listado de variables de este tipo se omite su presentación en este trabajo, pero para entrar en detalle del proceso realizado con ellas en este punto, se muestran algunos casos particulares. En la figura 14 se pueden observar 2 tipos de casos encontrados en las variables analizadas. En los casos señalados en la figura, los posibles valores de las variables se restringen a un solo valor (figura 14a) o al dominio *casi* completo (figura 14b) de un valor dentro de dos posibilidades. Este tipo de variables no aportan información relevante al modelo, motivo por el cual no se deben tener en cuenta en el entrenamiento. Las variables con un único valor son descartadas, al igual que las variables que cuentan con hasta 2 observaciones (1.2 %) del valor menos predominante según los casos ejemplificados en la figura 14b. Resultado: cantidad de variables del conjunto de datos = 95.

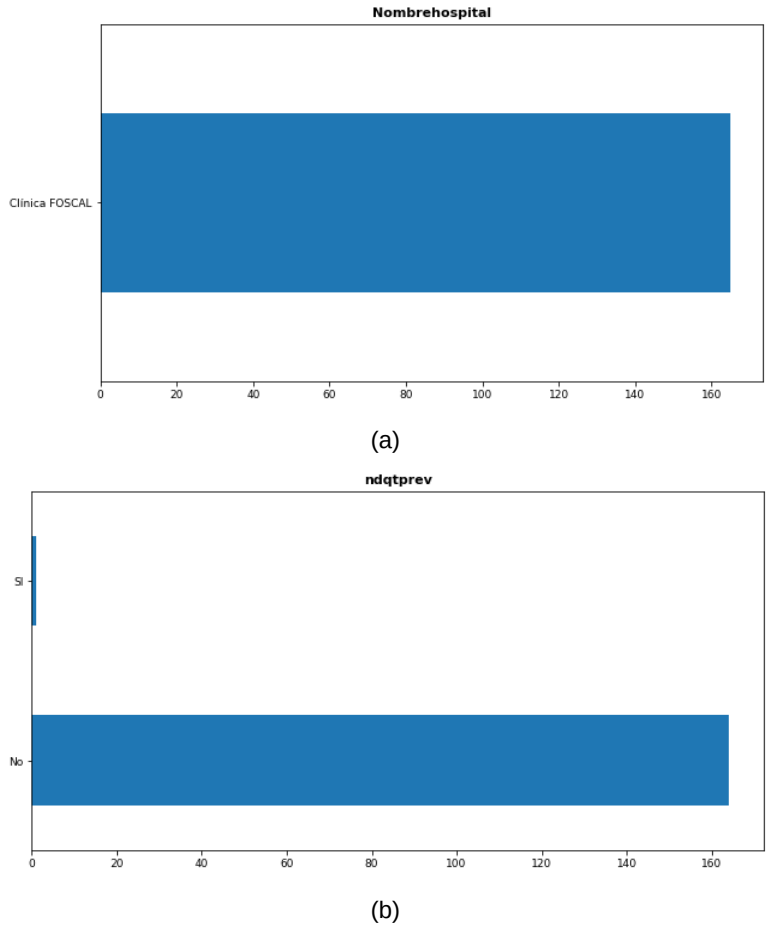

Figura 14. Ejemplos de distribuciones de posibles valores en variables categóricas

5. Se realizó el mismo procedimiento realizado con las variables categóricas, pero esta vez con las variables numéricas. El conjunto de variables numéricas cuenta con 45 unidades. Por la extensión del listado de variables de este tipo, se omite su presentación en este trabajo, pero se muestran algunos casos particulares. En las figuras 15 y 16 se pueden observar 2 casos encontrados en las variables analizadas. Al momento de cargar el origen de datos para análisis, se asignaron de manera errónea algunas variables como las ejemplificadas en la figura 15. Son variables que tienen un dominio discreto y deberían ser tratadas como variables categóricas. Se realizó la revisión del conjunto de las gráficas obtenidas y el conjunto variables numéricas para cambiar el tipo de datos de las variables como corresponde, y se les realizó a estas variables, asignadas de manera errónea, el tratamiento dado anteriormente a las variables de tipo categórico. Resultado: cantidad de variables del conjunto de datos = 90.

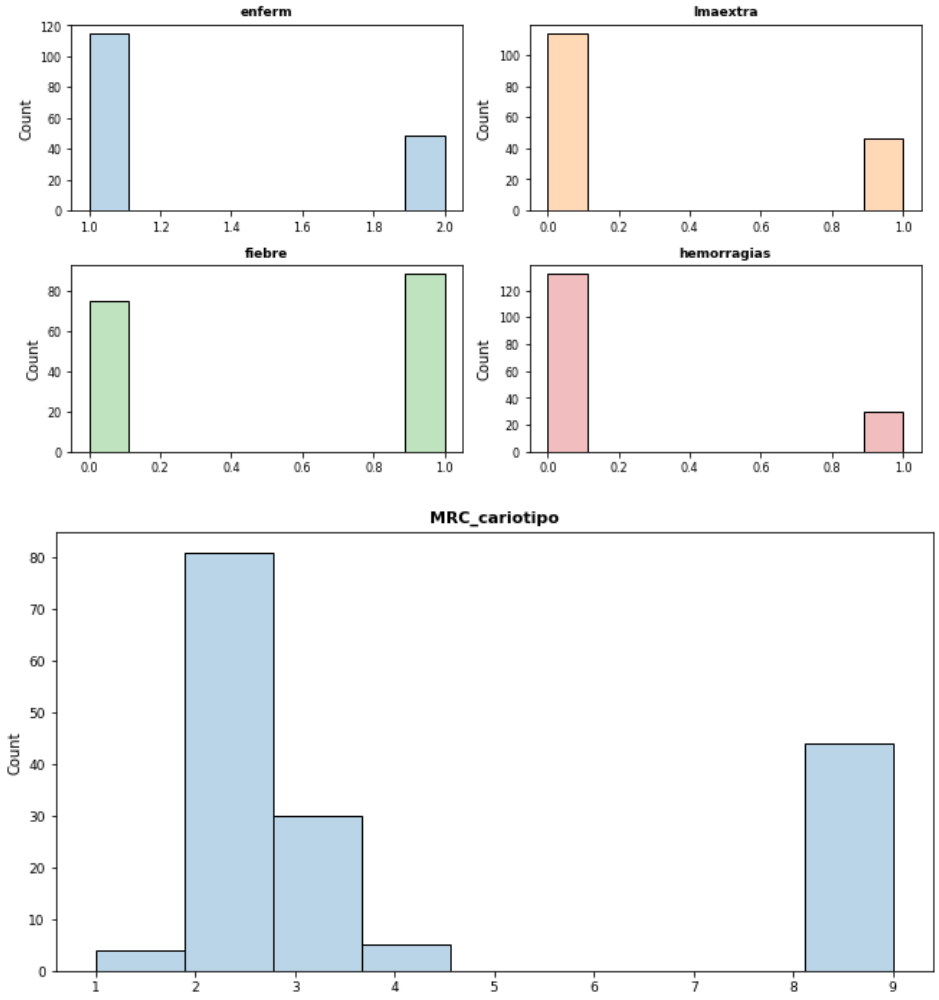

Figura 15. Primer caso encontrado en el conjunto de variables numéricas

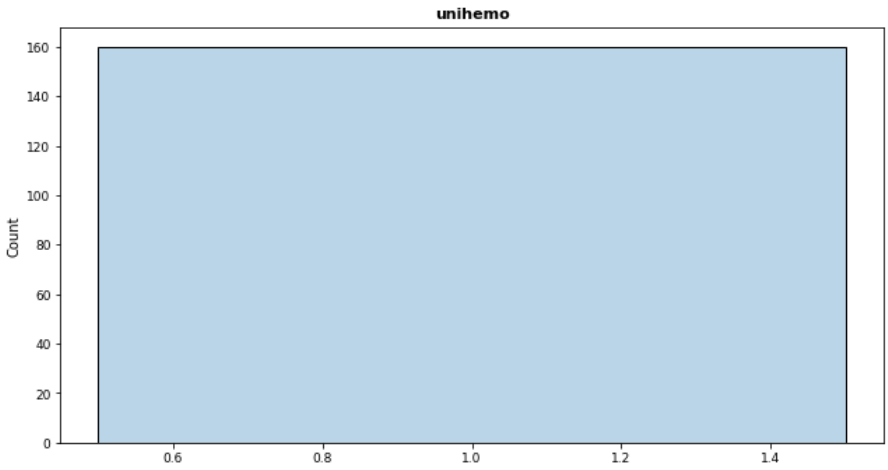

Figura 16. Segundo caso encontrado en el conjunto de variables numéricas

6. Después de la revisión anterior, se obtuvo como resultado que las variables de tipo numérico se reducen a 4. En la figura 17 se muestra la distribución inicial de cada una de ellas:

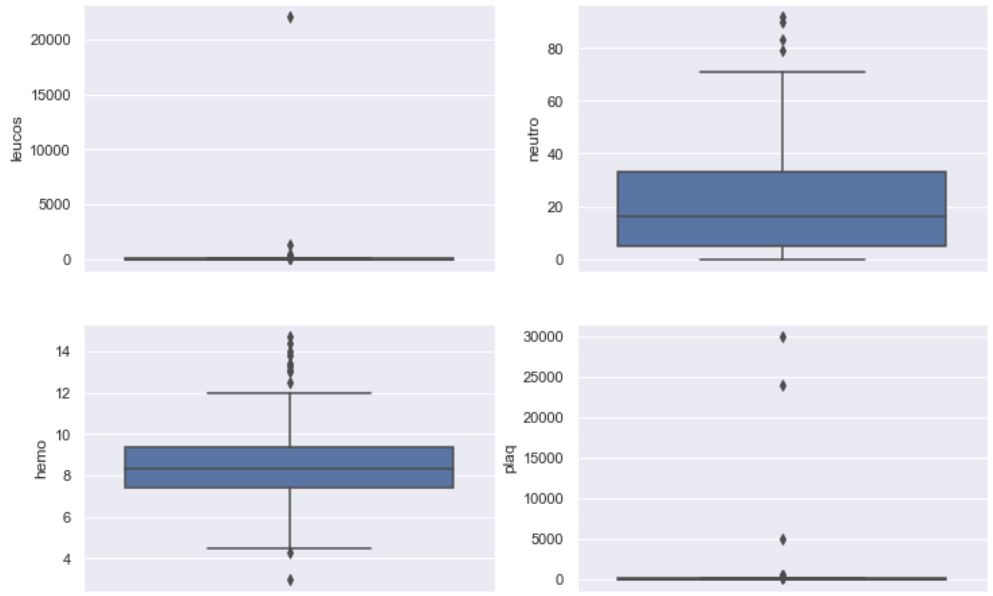

Figura 17. Diagramas de cajas de variables numéricas

Las variables *leucos* y *plaq* tienen observaciones anormales (*outliers*) que se encuentran muy por encima del rango de las demás. Estas observaciones se deben a errores de digitación en el momento del registro de la información en la plataforma. Estos valores fueron corregidos según la directriz recibida.

7. Se analizó la correlación de las variables numéricas del conjunto de datos. En la figura 18 se presenta la matriz de correlación de las variables:

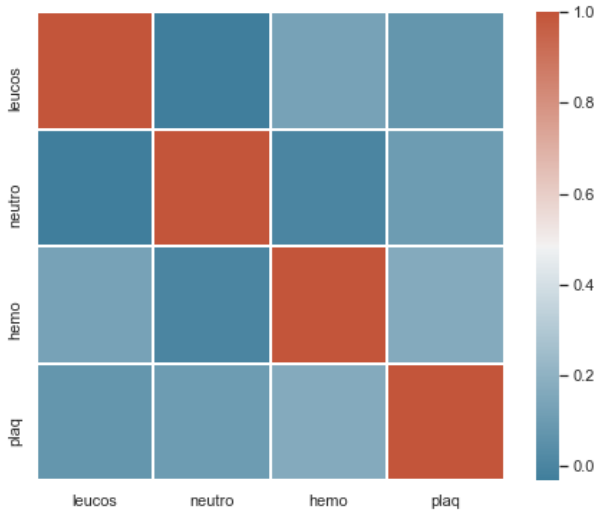

Figura 18. Matriz de correlación de variables numéricas

Se observa que no hay correlación relevante entre las variables analizadas, motivo por el cual se mantienen todas dentro del conjunto de datos.

8. Aunque se descartaron las variables con gran porcentaje de datos vacíos, era evidente que se encontrarían variables con datos faltantes. También se contó en este punto con una cantidad de variables a las que se podría realizar algún tratamiento adicional para unir valores buscando disminuir su cantidad. No se realizó la imputación de los valores ausentes ni procedimientos adicionales para tener en cuenta el concepto médico experto de manera posterior en la organización final del conjunto de datos de entrenamiento.

5.1.3 Creación de conjunto de datos para entrenamiento

La selección del conjunto de datos con el que será realizado el entrenamiento, es uno de los hitos más importantes en el desarrollo de cualquier proyecto de Inteligencia Artificial. No es para menos pues una correcta selección de este influye directamente en los resultados finales. Uno de los riesgos siempre latentes en el desarrollo de cualquier modelo es la aparición del sobreajuste (*overfitting*), el cual es un fenómeno que se presenta cuando el modelo aprende muy bien las características de los datos utilizados, prediciendo de manera correcta los resultados de los datos de entrenamiento, pero perdiendo su capacidad de generalizar estas predicciones con otros datos no procesados.

El sobreajuste aumenta cuando el número de observaciones del conjunto de datos es relativamente pequeño si se compara con el número de variables del mismo (Moons et al., 2014). El número de observaciones o eventos por variable (EPV) es comúnmente usado para calcular el tamaño de la muestra recomendada para realizar el entrenamiento. Un tamaño de muestra con un EPV de 10 o más veces es recomendado con frecuencia para evitar el sobreajuste en modelos predictivos (Vergouwe et al., 2005; Vittinghoff & McCulloch, 2007).

Si bien en los pasos previos la cantidad de variables del conjunto de datos fue disminuida a menos de una cuarta parte de su tamaño inicial, aún era necesario realizar una depuración de las variables remanentes para llegar al EPV recomendado en estos casos. Del conjunto de datos que se tenía en este punto, se eligieron variables de relevancia epidemiológica, como la edad y el género, así como otras que según lo encontrado en la literatura han demostrado tener un impacto en el pronóstico, la toma de decisiones para el tratamiento, o en la sobrevida de los pacientes. Se decidió trabajar con las 13 variables especificadas en la tabla 2.

Variables	
Mayor de 65 años	Tipo de LMA
Género	Escala ECOG
Clasificación de cariotipo MRC	Presenta diabetes
Enfermedad mínima residual (EMR)	Presenta hipertensión arterial (HTA)
Intensidad del tratamiento	Cantidad de hemoglobina
Conteo de leucocitos mayores a 50.000 / mm^3	Conteo de plaquetas
Índice de masa corporal (IMC) mayor que 25	

Tabla 2. Variables seleccionadas para la creación del conjunto de datos

Se aplicó una serie de transformaciones para obtener el conjunto de datos que se va a usar en el entrenamiento, teniendo en cuenta que las variables cumplieran con ciertos parámetros que garanticen el buen funcionamiento de los algoritmos. Se realizaron reclasificaciones, normalizaciones y codificaciones según el caso presentado. En la tabla 3 se especifica de manera general las variables procesadas y los procesos realizados para la obtención del conjunto de datos final.

El conjunto de datos para iniciar el entrenamiento finaliza con 165 registros y 17 variables. La cantidad seleccionada de variables permite que el EPV de la muestra se aproxime a la recomendación (proporción de 10 a 1) según la cantidad de observaciones.

Variable	Variables procesadas	Proceso realizado para obtención	Resultado
Mayor de 65 años	Fecha de nacimiento Fecha de diagnóstico	• Cálculo de edad del paciente a partir de las variables procesadas. • Binarización de la variable según el valor límite.	Variable con dos posibles valores: 0 si la edad del paciente es menor o igual a 65. 1 si es mayor de 65 años.
Género	Sexo		Variable con dos posibles valores: 1 Masculino. 2 Femenino.
MRC	MRC	• Imputación de valores vacíos asignando categoría correspondiente. • Reclasificación de valores.	Variable con 4 posibles valores: 0 Riesgo no valorable o sin datos disponibles. 1 Riesgo bajo. 2 Riesgo intermedio. 3 Riesgo alto.

Tabla 3. Variables finales para el entrenamiento del modelo predictivo

Variable	Variables procesadas	Proceso realizado para obtención	Resultado
EMR	EMR EMR > 0.01 EMR < 0.01	• Imputación de valores vacíos asignando categoría correspondiente. • Reclasificación de valores. • Creación de una columna con valores reclasificados.	Variable con 3 posibles valores: 0 Sin dato. 1 EMR negativo. 2 EMR positivo.
Intensidad de tratamiento	Esquema	• Imputación de valores vacíos asignando categoría correspondiente. • Reclasificación de valores.	Variable con 3 posibles valores: 0 Sin dato o sin tratamiento. 1 Esquema no intensivo. 2 Esquema intensivo.
Leucocitos mayores de 50,000/mm^3	Leucocitos	• Imputación de valores vacíos asignando la media de los datos presentes. • Binarización de la variable según el valor límite.	Variable con dos posibles valores: 0 Menor que el valor límite. 1 Mayor que el valor límite.
IMC mayor que 25	IMC	• Imputación de valores vacíos asignando categoría correspondiente. En este campo podría considerarse el valor de la media colombiana. Como este valor es menor que el límite propuesto, se asigna el valor correspondiente a menor que 25. • Binarización de la variable según el valor límite.	Variable con dos posibles valores: 0 Menor que el valor límite. 1 Mayor que el valor límite.
Tipo LMA	LMA OMS 2016	• Agrupación por frecuencia de aparición. Se tomaron los 5 tipos más frecuentes. • Binarización *one-hot encoding* teniendo en cuenta las variables agrupadas.	5 variables que indican el tipo de LMA según los valores agrupados: 0 No es del tipo. 1 Es del tipo.
ECOG	ECOG	Imputación de valores vacíos.	Valores según clasificación del protocolo.
Diabetes	Diabetes		Valores según clasificación del protocolo.
HTA	HTA		Valores según clasificación del protocolo.
Hemoglobina	Hemoglobina	• Imputación de valores vacíos asignando la media de los datos presentes. • Normalización de valores.	Valores entre 0 y 1.
Plaquetas	Plaquetas	• Imputación de valores vacíos asignando la media de los datos presentes. • Normalización de valores.	Valores entre 0 y 1.

Tabla 3. Variables finales para el entrenamiento del modelo predictivo (Continuación)

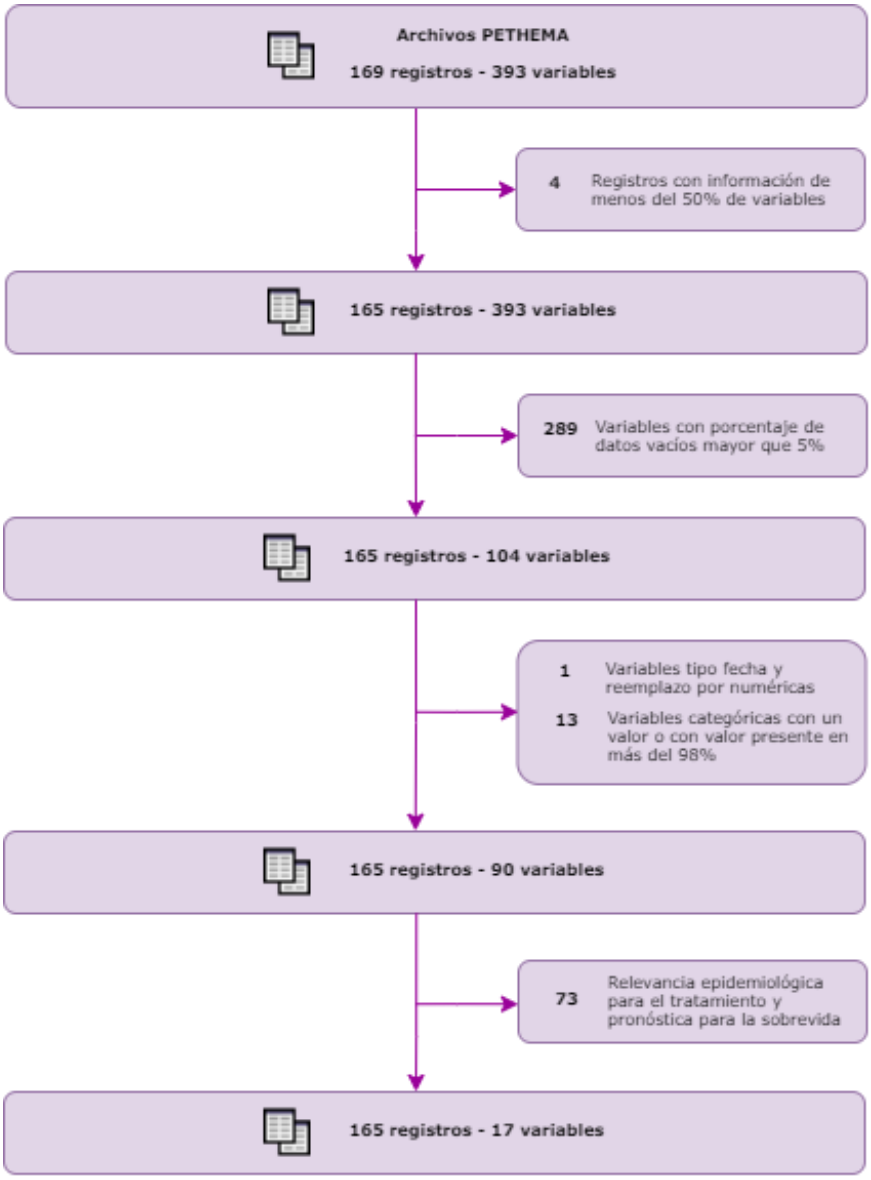

Figura 19. Pasos del proceso de revisión del conjunto de datos

5.1.4 Entrenamiento de modelos de aprendizaje automático

Después de la revisión de la literatura de trabajos previos del estado del arte en los que se usaron conjuntos de datos similares al conjunto utilizado en este trabajo para la creación de modelos predictivos, se decidió trabajar con los algoritmos *XGBoost*, *Random Forest*, SVM y redes neuronales, algoritmos estos de aprendizaje supervisado que se basan cada uno en conceptos y principios diferentes para resolver problemas de clasificación binaria y multivariable.

La variable a predecir (dependiente) fue el evento muerte – tiempo transcurrido entre diagnóstico y fallecimiento – de los pacientes tratados en diferentes periodos de tiempo. Se muestran en la figura 20 en el tiempo entre diagnóstico y la aparición del evento para los pacientes del conjunto de datos:

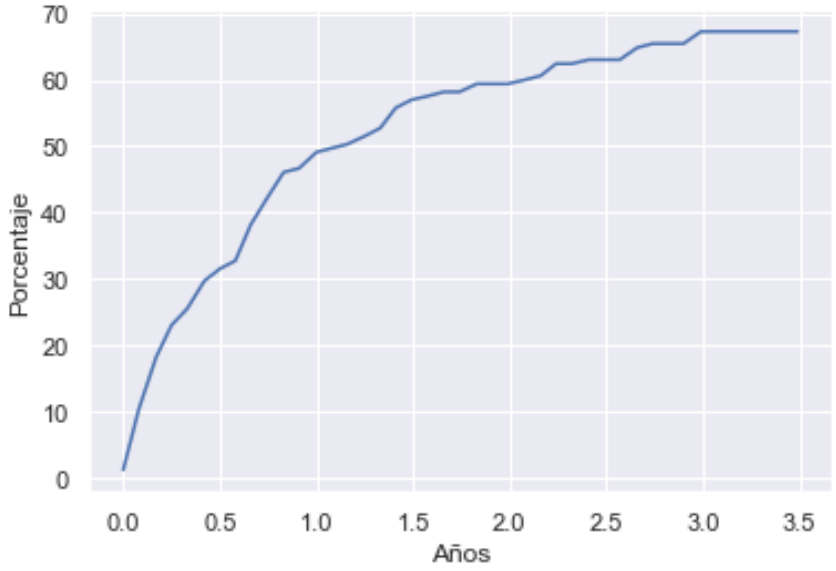

Figura 20. Tiempo entre diagnóstico y muerte en pacientes del conjunto de datos

Se observa que la cantidad de muertes es mayor a medida que transcurre tiempo desde el diagnóstico, teniendo en este conjunto de datos la aparición del evento muerte en la mitad de los pacientes en el primer año de tratamiento.

Se enfocó el análisis en predecir la mortalidad de los pacientes para los periodos de tiempo de 30 días, 6 meses y 2 años. Estos intervalos de tiempo son periodos de observación tomados como referencia, usados habitualmente en el tratamiento de la LMA. Adicionalmente se incluyó en el análisis el periodo de 1 año desde el diagnóstico, teniendo en cuenta que, como fue mencionado anteriormente, para este lapso de tiempo la cantidad de pacientes fallecidos equivale a la mitad de las muestras disponibles. Lo anterior para realizar el entrenamiento de los modelos con la variable objetivo lo más balanceada posible, lo cual según la teoría incide en la obtención de mejores resultados de predicción.

Se realizaron varias series de entrenamiento variando los parámetros de los algoritmos, también llamados hiperparámetros, para analizar los resultados de las predicciones. Para este fin, se generaron objetos en el lenguaje de programación Python que permitieron la variación de hiperparámetros de cada algoritmo, procurando conseguir los mejores resultados con el conjunto de datos disponible. Al final se realizó la comparación de los resultados, a partir de

las métricas obtenidas, para seleccionar el modelo de mejor desempeño a la hora de realizar las predicciones requeridas.

Se dividió el conjunto de datos disponible en 2 conjuntos: entrenamiento y pruebas. Con el primero se realizó el entrenamiento del modelo propiamente dicho, y con el segundo se realizaron las validaciones que permiten verificar la capacidad de predicción del modelo y el cálculo de métricas para medir el desempeño del mismo (exactitud, precisión, *recall*, *F1 score*, etc.). La cantidad de observaciones disponibles para los conjuntos de entrenamiento y pruebas depende en gran medida de la cantidad de datos disponibles, buscando tener lo necesario para el cálculo de las métricas. Como regla general, se acepta una división 80% - 20%, entrenamiento y pruebas respectivamente, haciendo un reparto de los datos entre los conjuntos de manera aleatoria y estratificada.

5.1.4.1 Determinación de parámetros de los modelos

Se formaron conjunto de datos para el entrenamiento de los modelos en cada periodo de tiempo de análisis (30 días, 6 meses, 1 año y 2 años), particionando los datos en la distribución 80% - 20% señalada en el numeral anterior. En cada conjunto de datos se tenían las mismas variables predictoras, modificando la variable objetivo dependiendo de la aparición del evento en ese intervalo de tiempo (caso positivo). En la tabla 4 se puede observar el balanceo de clases en los conjuntos de entrenamiento y pruebas para cada periodo de análisis, después de la división de los datos:

	Entrenamiento				Pruebas			
	Vivo		Muerto		Vivo		Muerto	
Periodo	**Cantidad**	**Porcentaje**	**Cantidad**	**Porcentaje**	**Cantidad**	**Porcentaje**	**Cantidad**	**Porcentaje**
30 días	125	94.7%	7	5.3%	32	94.12%	2	5.88%
180 días	90	68.18%	42	31.82%	23	67.65%	11	32.35%
1 año	67	50.76%	65	49.24%	17	50%	17	50%
2 años	53	40.15%	79	59.85%	14	41.18%	20	58.82%

Tabla 4. Balanceo de clases de la variable a predecir en los conjuntos de datos

Con esto se comprobó que la división de los datos se realizó de manera estratificada, conservando en la medida de lo posible la misma proporción de datos por clase en los conjuntos de entrenamiento y pruebas para cada periodo de análisis.

Con los conjuntos de datos depurados se realizó el entrenamiento de los algoritmos seleccionados, variando los hiperparámetros de los mismos para determinar la combinación

que produjera las mejores predicciones. Esta no es una tarea trivial pues la selección de estos valores impacta de manera significativa el comportamiento del modelo y altera los resultados obtenidos.

La variación de hiperparámetros para la evaluación del modelo se ejecutó mediante el uso de la función *GridSearchCV* de la librería *scikit-learn*. Esta función permite evaluar y seleccionar de forma sistemática los hiperparámetros de un modelo. Se especifica a la función el algoritmo o modelo a analizar y una serie de valores posibles para los hiperparámetros del mismo. Luego, la función evalúa todas las combinaciones posibles de hiperparámetros definidos y retorna como resultado el modelo con la mejor combinación, dentro de las opciones suministradas, junto con el resultado de la métrica analizada. En este proceso de variación de hiperparámetros se hizo uso de la técnica de validación cruzada con 5 *folds* con el fin de no disminuir drásticamente los datos disponibles para entrenamiento y validación en cada iteración. En la figura 21 se observan algunos resultados obtenidos en una de las iteraciones realizadas durante el entrenamiento en el proceso de variación de parámetros para el algoritmo *Random Forest* con el conjunto de datos para la predicción del evento muerte a 1 año:

```
Mejores parámetros
{'criterion': 'entropy', 'max_depth': 20, 'n_estimators': 500}
0.652 (+/-0.134) for {'criterion': 'gini', 'max_depth': None, 'n_estimators': 10}
0.696 (+/-0.211) for {'criterion': 'gini', 'max_depth': None, 'n_estimators': 100}
0.666 (+/-0.200) for {'criterion': 'gini', 'max_depth': None, 'n_estimators': 300}
0.679 (+/-0.208) for {'criterion': 'gini', 'max_depth': None, 'n_estimators': 500}
0.685 (+/-0.234) for {'criterion': 'gini', 'max_depth': 5, 'n_estimators': 10}
0.650 (+/-0.188) for {'criterion': 'gini', 'max_depth': 5, 'n_estimators': 100}
0.666 (+/-0.246) for {'criterion': 'gini', 'max_depth': 5, 'n_estimators': 300}
0.693 (+/-0.222) for {'criterion': 'gini', 'max_depth': 5, 'n_estimators': 500}
0.670 (+/-0.092) for {'criterion': 'gini', 'max_depth': 10, 'n_estimators': 10}
0.649 (+/-0.229) for {'criterion': 'gini', 'max_depth': 10, 'n_estimators': 100}
0.688 (+/-0.221) for {'criterion': 'gini', 'max_depth': 10, 'n_estimators': 300}
0.687 (+/-0.221) for {'criterion': 'gini', 'max_depth': 10, 'n_estimators': 500}
0.632 (+/-0.219) for {'criterion': 'gini', 'max_depth': 20, 'n_estimators': 10}
0.631 (+/-0.214) for {'criterion': 'gini', 'max_depth': 20, 'n_estimators': 100}
0.697 (+/-0.160) for {'criterion': 'gini', 'max_depth': 20, 'n_estimators': 300}
0.686 (+/-0.232) for {'criterion': 'gini', 'max_depth': 20, 'n_estimators': 500}
0.604 (+/-0.133) for {'criterion': 'gini', 'max_depth': 50, 'n_estimators': 10}
0.675 (+/-0.155) for {'criterion': 'gini', 'max_depth': 50, 'n_estimators': 100}
0.660 (+/-0.194) for {'criterion': 'gini', 'max_depth': 50, 'n_estimators': 300}
0.661 (+/-0.238) for {'criterion': 'gini', 'max_depth': 50, 'n_estimators': 500}
0.621 (+/-0.180) for {'criterion': 'entropy', 'max_depth': None, 'n_estimators': 10}
0.712 (+/-0.147) for {'criterion': 'entropy', 'max_depth': None, 'n_estimators': 100}
0.663 (+/-0.197) for {'criterion': 'entropy', 'max_depth': None, 'n_estimators': 300}
0.680 (+/-0.148) for {'criterion': 'entropy', 'max_depth': None, 'n_estimators': 500}
0.648 (+/-0.127) for {'criterion': 'entropy', 'max_depth': 5, 'n_estimators': 10}
0.666 (+/-0.199) for {'criterion': 'entropy', 'max_depth': 5, 'n_estimators': 100}
0.649 (+/-0.226) for {'criterion': 'entropy', 'max_depth': 5, 'n_estimators': 300}
0.678 (+/-0.228) for {'criterion': 'entropy', 'max_depth': 5, 'n_estimators': 500}
0.637 (+/-0.210) for {'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 10}
0.682 (+/-0.164) for {'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 100}
0.679 (+/-0.208) for {'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 300}
0.654 (+/-0.199) for {'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 500}
0.698 (+/-0.159) for {'criterion': 'entropy', 'max_depth': 20, 'n_estimators': 10}
0.708 (+/-0.173) for {'criterion': 'entropy', 'max_depth': 20, 'n_estimators': 100}
0.691 (+/-0.164) for {'criterion': 'entropy', 'max_depth': 20, 'n_estimators': 300}
0.719 (+/-0.208) for {'criterion': 'entropy', 'max_depth': 20, 'n_estimators': 500}
0.648 (+/-0.158) for {'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 10}
0.687 (+/-0.184) for {'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 100}
0.703 (+/-0.144) for {'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 300}
0.684 (+/-0.172) for {'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 500}
```

Figura 21. Métricas calculadas en el proceso de variación de hiperparámetros (Random Forest para predicción de evento muerte en 1 año)

En la red neuronal se implementó una arquitectura de 4 capas. Una capa de entrada que recibe las variables del conjunto de datos, una capa oculta *fully connected* de 256 neuronas a la que se aplica la técnica de regularización *batch normalization*, una capa oculta *fully connected* con 512 neuronas a la que se aplica la técnica de regularización *dropout*, y una capa de salida con 1 neurona que representa el evento (clase) predicho. Esta arquitectura se implementa tomando como guía lo revisado en la literatura referente a la implementación de modelos de clasificación de la misma naturaleza.

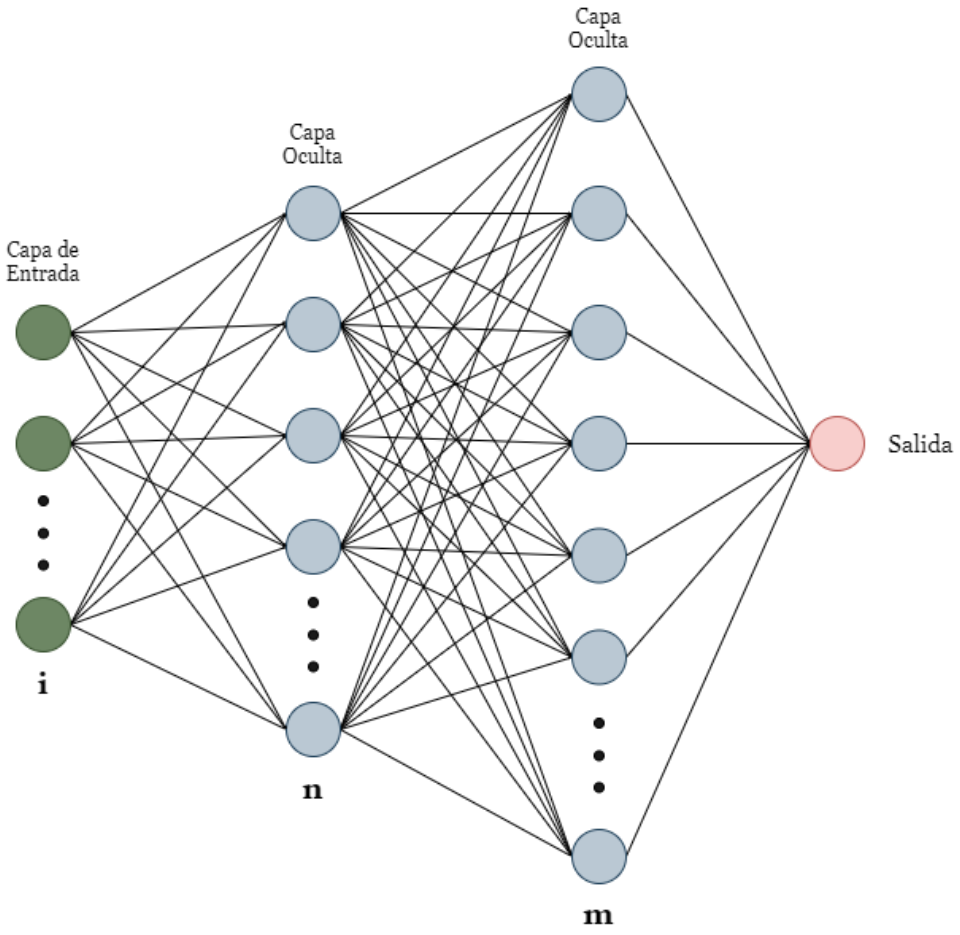

Figura 22. Esquema de red neuronal implementado para el entrenamiento

En la figura 23 se muestra la precisión y pérdida obtenidas en el entrenamiento realizado durante el proceso de variación de parámetros para una red neuronal con el conjunto de datos para la predicción del evento muerte a 2 años:

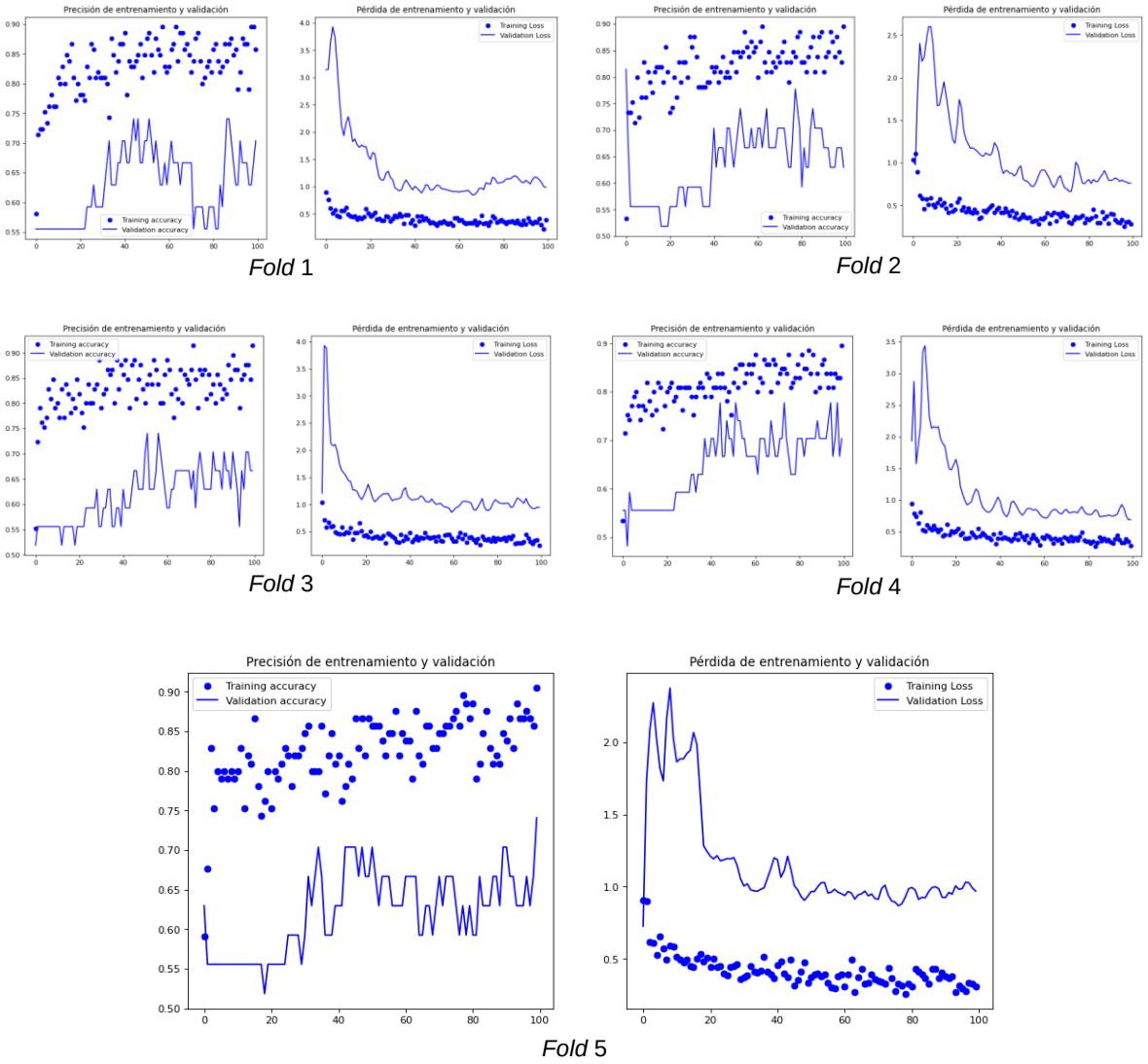

Figura 23. Resultados de entrenamiento de red neuronal en el proceso de variación de hiperparámetros para predicción de evento muerte en 2 años

En la tabla 5 se observan los resultados finales de la búsqueda de hiperparámetros obtenidos luego de la ejecución de entrenamientos previos teniendo en cuenta el funcionamiento de los algoritmos analizados. Los hiperparámetros se encuentran agrupados por cada conjunto de datos de los periodos analizados. El proceso fue realizado tomando como referencia la métrica F1, la cual combina en una medida las métricas de precisión y *recall* utilizando la media harmónica. Con un valor máximo de 1 – para precisión y *recall* perfectos – la métrica F1 ofrece una buena medida del comportamiento de los modelos en problemas de clasificación. Posterior a este proceso se analizan los resultados con métricas adicionales y matrices de confusión para medir el desempeño obtenido y tomar decisiones.

		30 días	180 días	1 año	2 años
XBGoost	H.	{'learning_rate': 1, 'max_depth': 5, 'n_estimators': 50}	{'learning_rate': 1e-3, 'max_depth': 5, 'n_estimators': 500}	{'learning_rate': 1e-4, 'max_depth': 5, 'n_estimators': 50}	{'learning_rate': 1, 'max_depth': 5, 'n_estimators': 50}
	F1	0.257	0.625	0.815	0.739
Random Forest	H.	{'criterion': 'entropy', 'max_depth': None, 'n_estimators': 10}	{'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 100}	{'criterion': 'gini', 'max_depth': None, 'n_estimators': 100}	{'criterion': 'gini', 'max_depth': None, 'n_estimators': 100}
	F1	0.2	0.623	0.835	0.787
SVM	H.	{'C': 20, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}	{'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	{'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}	{'C': 1, 'degree': 2, 'gamma': 'auto', 'kernel': 'rbf'}
	F1	0.38	0.593	0.825	0.773
Red neuronal Adaptativo	H.	{'activation': 'relu', 'dropout_rate': 0.7}	{'activation': 'relu', 'dropout_rate': 0.5}	{'activation':'sigmoid', 'dropout_rate': 0.5}	{'activation': 'sigmoid', 'dropout_rate': 0.7}
	F1	0.358	0.597	0.758	0.757
Red Neuronal Gradiente	H.	{'activation': 'relu', 'dropout_rate': 0.5}	{'activation': 'relu', 'dropout_rate': 0.2}	{'activation': 'relu', 'dropout_rate': 0.5}	{'activation': 'sigmoid', 'dropout_rate': 0.2}
	F1	0.358	0.651	0.763	0.751

H.: Hiperparámetros

Tabla 5. Resultado de búsqueda de hiperparámetros

Se visualizan en la tabla dos resultados de redes neuronales porque se ejecutó la variación de hiperparámetros separando los entrenamientos dependiendo del optimizador utilizado: basados en gradiente descendente o adaptativos.

5.1.4.2 Selección de modelo con mejores resultados

Después de la ejecución de la variación de hiperparámetros, se realizaron las siguientes observaciones:

- Los hiperparámetros de los algoritmos cambian dependiendo de la predicción que se requiere realizar. Es una muestra de cómo es afectado el modelo dependiendo del conjunto de datos utilizado y de la distribución de la variable a predecir.
- Las predicciones mejoran cuando las clases de la variable a predecir se encuentran más balanceadas (cantidad similar de observaciones de cada clase presente).
- La red neuronal sufre sobreajuste aun usando técnicas de regularización como *dropout* y *batch normalization*. Este es un comportamiento que se esperaba teniendo en cuenta la cantidad de observaciones del conjunto de datos trabajado.

Se tomó como parámetro de aceptación la obtención de una métrica mayor o igual a 0.8 en el proceso de variación realizado. Se puede determinar que con los modelos entrenados para la predicción del evento muerte a los 30 días, 180 días y 2 años no se alcanza a cumplir con este requisito. Por otro lado, algunos de los modelos obtenidos para la predicción del evento en el primer año cumplen con el criterio de aceptación propuesto.

Se revisaron las matrices de confusión, algunas métricas adicionales propias del análisis de problemas de clasificación, y curvas ROC para los modelos *XGBoost*, *Random Forest y* SVM*,* los cuales cumplen con el criterio de aceptación planteado:

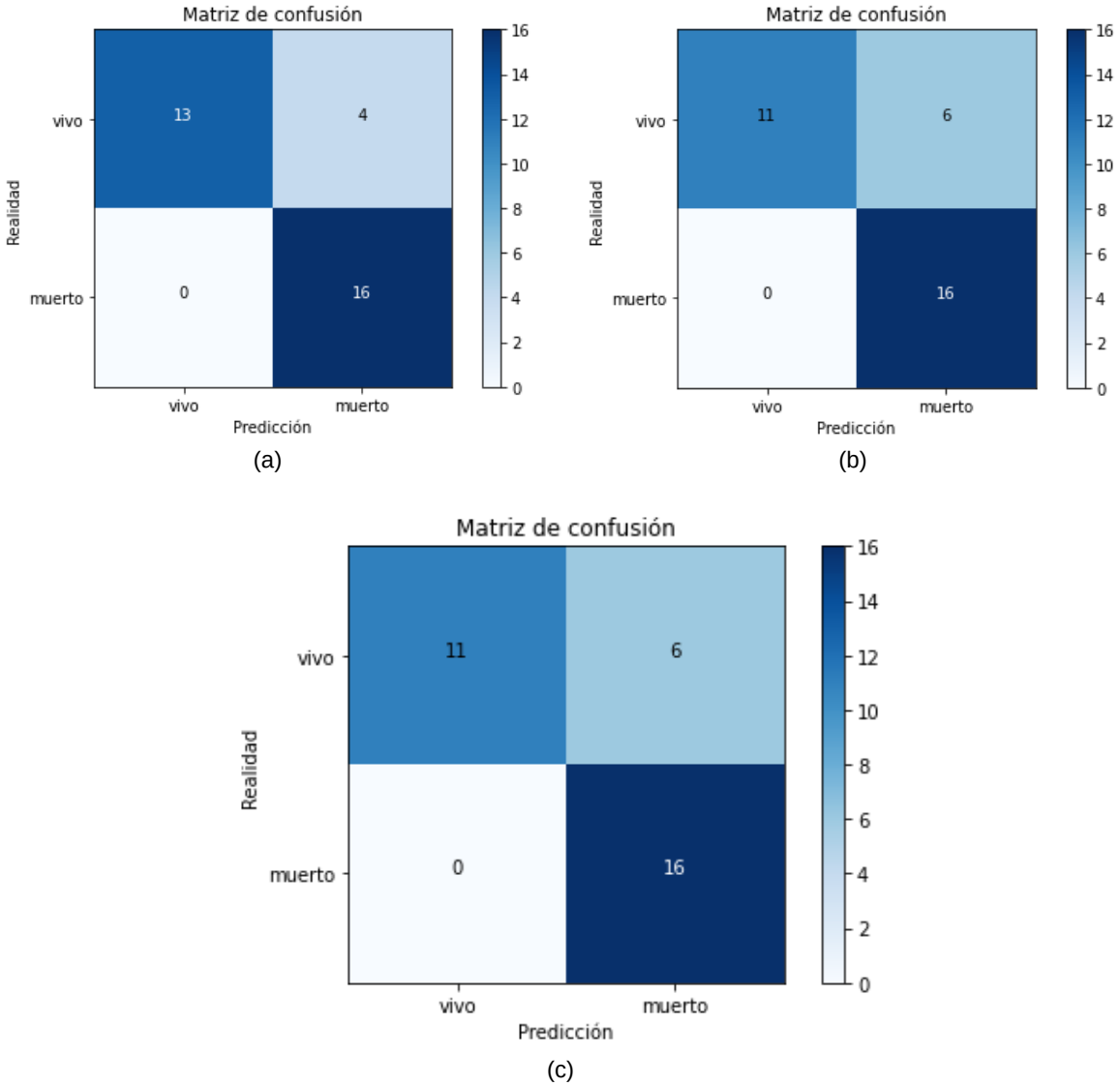

Figura 24. Matrices de confusión de predicciones de evento muerte 1 año
(a) XGBoost (b) Random Forest (c) SVM

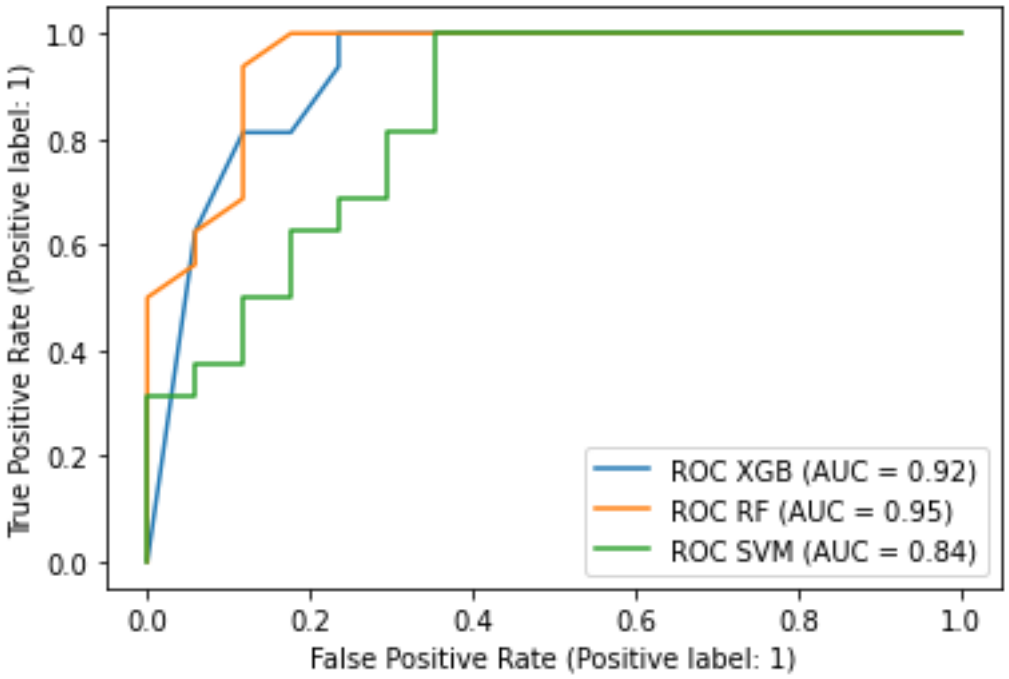

Figura 25. AUROC de las predicciones de evento de muerte 1 año

En la tabla 6 se pueden observar métricas derivadas de la matriz de confusión como la exactitud, sensibilidad, especificidad, precisión y *recall* que, junto con la ya utilizada F1, sirven de utilidad para conocer el comportamiento de un modelo de clasificación. Se puede observar que el modelo *XGBoost* obtuvo los mejores resultados para el caso analizado, superando a los demás en las métricas calculadas.

Modelo	Exactitud	Sensibilidad	Especificidad	Precisión	Recall	F1
XGBoost	0.88	1	0.76	0.8	1	0.89
Random Forest	0.82	1	0.65	0.73	1	0.84
SVM	0.82	1	0.65	0.73	1	0.84

Tabla 6. Métricas de modelos predictivos analizados para el evento de muerte en 1 año

Se observó una tendencia del mejor modelo a clasificar de manera más adecuada los casos positivos que los negativos, algo que se ve reflejado en la obtención de una mayor sensibilidad

que especificidad. Esto se evidenció de igual manera al analizar la matriz de confusión obtenida en la predicción.

El modelo *XGBoost* mostró mejor desempeño en la predicción. Fue superado por poco en el AUC por el modelo *Random Forest*, al cual mejora con mayor cantidad de casos verdaderos negativos predichos. El modelo *XGBoost* fue elegido como el mejor en la predicción del evento muerte en el primer año de tratamiento.

5.2 Implementación de interfaz de usuario

La interfaz de usuario fue desarrollada para ser consultada en un navegador de Internet. Se utilizó HTML con *javascript* para su despliegue. Según los requisitos planteados, se hizo necesaria la creación de 3 vistas para poder cargar datos para realizar un nuevo entrenamiento, visualizar los resultados obtenidos de dicho entrenamiento, y poder realizar una predicción individual a partir de los datos de las variables predictoras especificadas en la tabla 2.

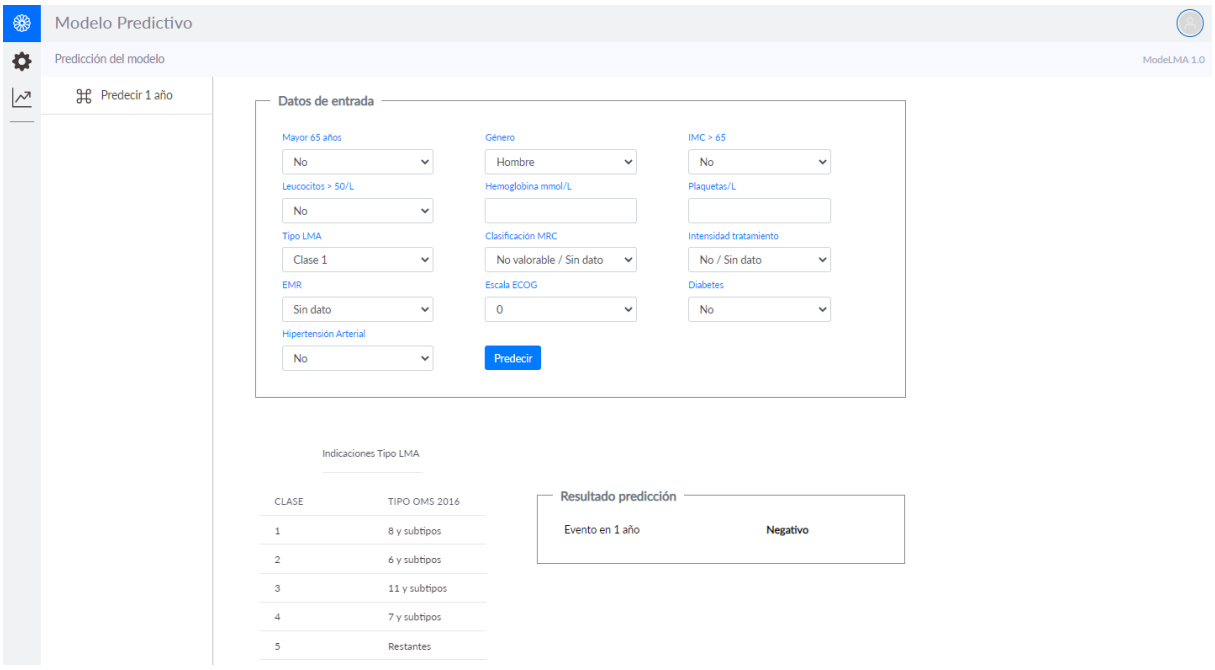

Figura 26. Vista de recepción de datos para predicciones

En cada vista la tarea de recolección de datos es realizada haciendo uso de *javascript*. La información es recolectada y enviada a un servicio el cual ejecuta una rutina desarrollada en Python. El proceso es síncrono, esperando que la tarea finalice para notificar a la interfaz.

Cuando se finaliza el proceso se envía la información a la interfaz de usuario y se carga de manera automática la información.

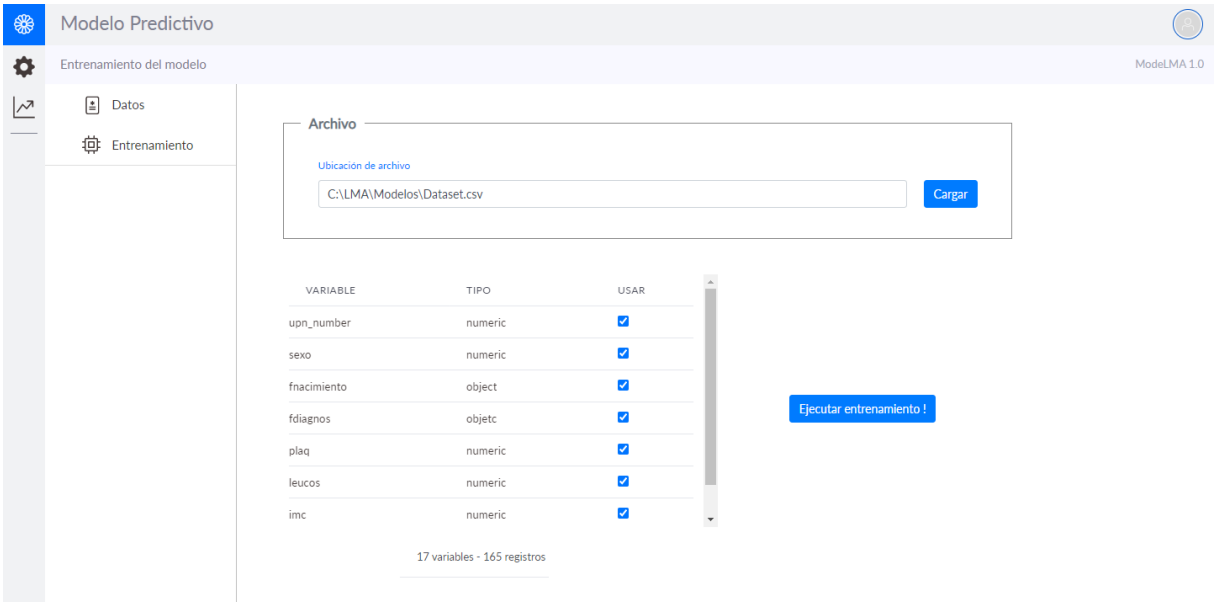

Figura 27. Vista para cargar la información de un nuevo entrenamiento

Para el entrenamiento se especifican las variables seleccionadas del protocolo PETHEMA, a partir de las cuales se realiza el tratamiento de datos para obtener la información en el formato esperado.

Figura 28. Vista de visualización de resultados de entrenamiento

6. Evaluación

6.1 Análisis de resultados del modelo predictivo

La generación del modelo predictivo fue realizada para diferentes periodos de tiempo de aparición del evento muerte del paciente. Revisando los resultados obtenidos, se consideró excluir las predicciones realizadas para los periodos de 30 días, 180 días y 2 años. La predicción es muy baja en los 2 primeros casos, y aunque en el tercero mejoró la situación, no fue suficiente para cumplir con el criterio de aceptación planteado inicialmente.

Por otro lado, el modelo para la predicción del evento muerte del paciente en el primer año de tratamiento alcanzó unos resultados de 0.88 de exactitud, 1 de sensibilidad y *recall*, y AUC de 0.92. Este es un resultado comparable con otros obtenidos en algunos de los trabajos revisados en el estado del arte que, teniendo en cuenta que se alcanza con un modelo piloto entrenado con una muestra pequeña de datos, hace presagiar la obtención de mejores resultados al incluir más información que aumente la cantidad de observaciones disponible para realizar entrenamientos en el futuro. Es inevitable pensar que las posibilidades de éxito del tipo de análisis realizado en este trabajo son altas, pero es necesaria la realización de una validación externa que permita analizar el modelo de manera exhaustiva, de cara a tener en cuenta más factores y conceptos en su realización para conseguir mejores resultados.

El modelo desarrollado predice de manera correcta el 88% de los casos totales presentados. Tiene una tendencia a clasificar de manera más adecuada los casos positivos (evento) que los negativos (sin evento), y realiza una clasificación correcta de la ocurrencia del evento muerte en el primer año con una precisión del 80%. La inclusión de características adicionales puede mejorar los resultados que obtenidos, algo que no pudo ser realizado en este trabajo pues el conjunto de observaciones era limitado y se debía cumplir con el EPV para obtener un resultado aceptable que permitiera analizar si la aplicación de estos modelos puede traer beneficios en este campo.

Algo importante que vale la pena señalar son las facilidades ofrecidas por las herramientas del lenguaje Python. Cuenta con un kit de trabajo para proyectos de Inteligencia Artificial que dan agilidad a la hora de trabajar con volúmenes de datos, y permiten la implementación de diferentes algoritmos para el entrenamiento de modelos de manera rápida y sencilla.

Además del análisis de la capacidad predictiva del modelo, se realizó una evaluación final para determinar la relevancia de las variables utilizadas por el mismo. Para este fin, se utilizó el método SHAP (*Shapley Additive Explanations*), desarrollado por Lundberg y Lee (Lundberg

et al., 2017). Este método se basa en un concepto en teoría de juegos, *Shapley Value*, que indica cuánto contribuye cada variable en la obtención de un resultado (Park et al., 2021). SHAP asigna a cada variable o característica un valor de importancia para predicción particular. En la figura 29 se puede observar el resultado del análisis realizado con este método:

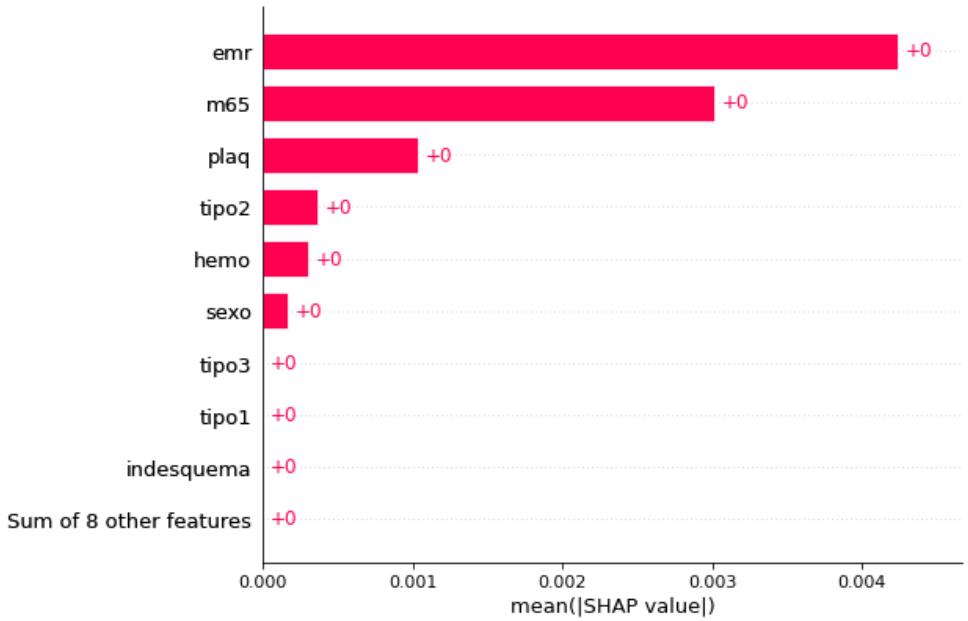

Figura 29. Variables relevantes según método SHAP

Este orden se ajusta al criterio de relevancia en el pronóstico de muerte en el tratamiento de LMA. Este es un buen punto comparativo que indica que los resultados del modelo pueden ser validados con el conocimiento previo de años de experiencia en el pronóstico de la condición analizada, sin importar que se trate de un piloto que aún tiene condiciones por mejorar.

Para finalizar la evaluación, se realiza la verificación del modelo utilizando la lista de verificación PROBAST la cual puede ser consultada en el anexo 2. Después de diligenciar esta lista, se obtiene un juicio general que indica que existe un riesgo de sesgo alto debido a que no se ha realizado un proceso de validación externa y el modelo no fue desarrollado con un gran conjunto de datos.

6.2 Evaluación de cumplimiento de requisitos

En la tabla 7 se referencian los requisitos definidos en el trabajo. En la columna entregable se especifica la sección del trabajo donde se puede encontrar el entregable para el cumplimiento del mismo.

Código	Requisito	Entregable	Entregable
RE01	Exploración y análisis de datos de archivos.	Resultados de Análisis de datos recibidos.	Sección 5.1.1 Sección 5.1.2
RE02	Creación de conjunto de datos para entrenamiento.	Conjunto de datos con variables características.	Sección 5.1.3
RE03	Entrenamiento de modelos de predicción usando los algoritmos de aprendizaje automático seleccionados después de la revisión del estado del arte.	Scripts del proceso con la implementación del algoritmo de exploración de parámetros.	Sección 5.1.4 Anexo 1
RE04	Determinación de hiperparámetros óptimos de entrenamiento de los algoritmos de aprendizaje automático utilizados, ajustados al conjunto de datos procesado.	Hiperparámetros con los que se obtienen los mejores resultados en el entrenamiento de algoritmos.	Sección 5.1.4.1
RE05	Selección del mejor modelo predictivo según las métricas obtenidas en el proceso de entrenamiento.	- Métricas de modelos predictivos. - Gráficas comparativas. - Cuadros comparativos.	Sección 5.1.4.2
RE06	Implementación de interfaz de usuario para el acceso al modelo de entrenamiento y predicción.	Interfaz implementada y funcional.	Sección 5.2
RE07	Función de carga de archivos en la interfaz de usuario para entrenamiento del modelo. La estructura de los archivos debe ser acorde a la especificación de PETHEMA.	Funcionalidad implementada en la interfaz de usuario.	Figura 27
RE08	La interfaz de usuario debe tener un módulo para realización de predicciones de un paciente.	Funcionalidad implementada en la interfaz de usuario.	Figura 26

Tabla 7. Verificación de cumplimiento de requisitos

7. Conclusiones y trabajo futuro

7.1. Conclusiones

Durante el proceso de desarrollo del presente trabajo se presentaron diversas limitaciones en el proceso de obtención de la información. Es necesario pasar por una serie de pasos para poder acceder a las bases de datos, esperar por la aprobación de Comités de Ética de las instituciones a la luz de la Guía de Buenas Prácticas Clínicas, algo que puede ser entendido sobre la base de la importancia de cuidar los datos para garantizar la privacidad de los pacientes y la protección de sus derechos.

La información se encuentra en gran medida atomizada, con diferentes centros sirviendo de custodios de fragmentos de la información, lo cual dificulta el poder reunir volúmenes de datos que permitan la realización de análisis y estudios que beneficien la práctica clínica y en últimas el acceso a una medicina cada vez más personalizada. Es necesaria la creación de convenios y acuerdos que permitan la correcta integración de la información con los diferentes actores del sistema de salud aportando su experiencia traducida en forma de datos. Esto puede también llevar a tomar conciencia de la importancia de administrar de manera seria y correcta la información para su posterior análisis, y a trabajar de manera conjunta siguiendo un objetivo común que beneficie a los pacientes y al sistema de salud.

En la búsqueda de conjuntos de datos para la realización de este trabajo, se tomó nota que existen algunos esfuerzos para la integración de información de tratamientos médicos como lo adelantado por el Ministerio de Salud con la creación y reglamentación del RIPS (Registro Individual de Prestaciones de Servicios de Salud), registro donde se centralizan datos a nivel nacional de procedimientos, consultas y diagnósticos. Se necesita una política clara enfocada al aprovechamiento de la información recopilada pues esta base de datos es usada principalmente para facturación de procedimientos y la obtención de datos estadísticos. Prueba de ello es que no existen muchos trabajos que usen esta información con fines investigativos o de análisis.

Aunque la comunidad internacional produce cada vez más trabajos apoyándose en la Inteligencia Artificial, haciendo estudios de diversos tipos para el tratamiento de imágenes diagnósticas, el análisis de información clínica o genética, el seguimiento a tratamientos para estudiar el impacto y efectividad de ciertos medicamentos según se vio en la literatura analizada en el estado del arte, en Colombia esta área se encuentra rezagada y no ha despertado el interés que sí despierta en otros lugares del mundo.

Se evidencia la evolución de la situación en el área de la Inteligencia Artificial y analítica de datos. El impulso del avance tecnológico de las últimas décadas y el interés de la comunidad de desarrollo que ha fijado su vista en lo que se ha llamado *data science*, ha permitido la aparición de herramientas que han facilitado la implementación de algoritmos que pueden ser ejecutados sin la necesidad de tener grandes centros de cómputo, lo cual ha repercutido de manera positiva en los tiempos de desarrollo de este tipo de proyectos.

Los resultados obtenidos en este trabajo indican que es posible aplicar la Inteligencia Artificial en el campo de la medicina a nivel local, lo cual se encuentra acorde al consenso obtenido en trabajos de la misma área. En Colombia, es posible iniciar aprovechando los avances obtenidos por la comunidad científica internacional para aplicarlos a nivel regional, ajustando las lecciones aprendidas y generando nuevos resultados que se adapten a la realidad demográfica, social y económica del país.

7.2. Líneas de trabajo futuro

Se recomienda continuar con el análisis realizado en este trabajo, incluyendo otros conjuntos de datos al entrenamiento del modelo de Inteligencia Artificial. Por orden natural, se deberían incluir los registros de pacientes tratados siguiendo el protocolo PETHEMA aprovechando que la estructura de datos es la misma que la tratada en el desarrollo realizado.

Pueden ser desarrollados otros trabajos en los que se tengan en cuenta factores específicos para realizar análisis por raza, género, análisis etarios o geográficos, de manera que se pueda personalizar cada vez más el tratamiento en medicina. Se han adelantado esfuerzos como el Atlas del Genoma del Cáncer, un programa iniciado en 2006 en el que se han caracterizado molecularmente 20.000 cánceres primarios emparejando muestras de 33 tipos de cáncer (National Cancer Institute, n.d.). Este proyecto ha generado más de 2.5 petabytes de información y ofrece estos datos para que puedan ser accedidos y utilizados de manera pública[7]. Se podría analizar si esta información puede ser utilizada en el ámbito local. Con bases de datos de perfiles genéticos, se pueden plantear investigaciones que permitan caracterizar las enfermedades ajustándolas a la realidad colombiana.

Con el aumento en la cantidad de registros disponibles en el conjunto de datos a analizar, sería importante profundizar en el uso de las redes neuronales aplicadas al análisis y generación de modelos para la predicción, diagnóstico y tratamiento. Este tipo de algoritmos

[7] Genomic Data Commons Data Portal: https://portal.gdc.cancer.gov/

ha probado su utilidad en diferentes ámbitos, y explota su potencial a medida que procesa grandes volúmenes de datos.

Es posible continuar agregando funcionalidades al sistema, creando módulos que permitan la realización de diferentes tareas propias del ámbito de la Inteligencia artificial como por ejemplo cargar conjuntos de datos de diferentes fuentes permitiendo la selección de variables, visualizar un análisis del conjunto de datos a procesar, realizar el entrenamiento de nuevos modelos seleccionando diferentes algoritmos y recibiendo la evaluación de resultados después de los análisis realizados.

Existe una línea de trabajo amplia que beneficiaría diferentes líneas de proyectos, no sólo proyectos de Inteligencia Artificial, y es la de la integración de la información para crear una o varias bases de datos especializadas que puedan ser utilizadas y sirvan como base para la realización de investigaciones. Para lograr esto se debe crear una política dirigida que permita realizar esta labor de manera juiciosa, con pautas claras, que no sólo tome información existente, sino que siente las bases para empezar a generar valor en los lugares donde aún no se realiza, soportada sobre una infraestructura tecnológica que permita el acceso seguro y la disponibilidad de los datos.

8. Bibliografía

Arai, Y., Kondo, T., Fuse, K., Shibasaki, Y., Masuko, M., Sugita, J., Teshima, T., Uchida, N., Fukuda, T., Kakihana, K., Ozawa, Y., Eto, T., Tanaka, M., Ikegame, K., Mori, T., Iwato, K., Ichinohe, T., Kanda, Y., & Atsuta, Y. (2019). Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Advances*, *3*(22), 3626–3634. https://doi.org/10.1182/bloodadvances.2019000934

Bai, Y., Do, D. H., Harris, P. R. E., Schindler, D., Boyle, N. G., Drew, B. J., & Hu, X. (2015). Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction. *Journal of Biomedical Informatics*, *53*, 81–92. https://doi.org/10.1016/j.jbi.2014.09.006

Centro Nacional de Investigación en evidencia y Tecnologías en Salud, C. (2013). Guia Colombiana de Leucemia en Niños 2013. In *Guía Colombiana de Leucemia en niños, niñas y adolescentes.* (Vol. 9, Issue 9). Ministerio de Salud y Protección Social.

Chandradevan, R., Aljudi, A. A., Drumheller, B. R., Kunananthaseelan, N., Amgad, M., Gutman, D. A., Cooper, L. A. D., & Jaye, D. L. (2020). Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Laboratory Investigation*, *100*(1), 98–109. https://doi.org/10.1038/s41374-019-0325-7

Cho, S.-Y., Kim, S.-H., Kang, S.-H., Lee, K. J., Choi, D., Kang, S., Park, S. J., Kim, T., Yoon, C.-H., Youn, T.-J., & Chae, I.-H. (2021). Pre-existing and machine learning-based models for cardiovascular risk prediction. *Scientific Reports*, *11*(1), 1–10. https://doi.org/10.1038/s41598-021-88257-w

Cuenta de Alto Costo. (2019). Situación del cáncer en población adulta en el SGSSS de Colombia, 2019. *Fondo Colombiano de Enfermedades de Alto Costo*, 336.

Dekker, F. W., Ramspek, C. L., & Van Diepen, M. (2017). Con: Most clinical risk scores are useless. *Nephrology Dialysis Transplantation*, *32*(5), 752–755. https://doi.org/10.1093/ndt/gfx073

Duval, M., Klein, J. P., He, W., Cahn, J. Y., Cairo, M., Camitta, B. M., Kamble, R., Copelan, E., De Lima, M., Gupta, V., Keating, A., Lazarus, H. M., Litzow, M. R., Marks, D. I., Maziarz, R. T., Rizzieri, D. A., Schiller, G., Schultz, K. R., Tallman, M. S., & Weisdorf, D. (2010). Hematopoietic stem-cell transplantation for acute leukemia in relapse or primary induction

failure. *Journal of Clinical Oncology*, *28*(23), 3730–3738. https://doi.org/10.1200/JCO.2010.28.8852

Fernández, P. (2019). *Desarrollo entorno visualización de datos médicos* [Universidad Internacional de La Rioja]. https://reunir.unir.net/handle/123456789/9502

Gacha Garay, M. J., Akle, V., Enciso, L., & Garavito Aguilar, Z. V. (2017). La leucemia linfoblástica aguda y modelos animales alternativos para su estudio en Colombia. *Revista Colombiana de Cancerología*, *21*(4), 212–224. https://doi.org/10.1016/j.rccan.2016.10.001

García, M., Chicaíza, L. A., Quitián, H., Linares, A., & Ramírez, Ó. (2015). Costo-efectividad de los tratamientos de consolidación para la leucemia mieloide aguda en niños en riesgo alto en el sistema de salud colombiano. *Biomedica*, *35*(4), 549–556. https://doi.org/10.7705/biomedica.v35i4.2563

Gerdes, H., Casado, P., Dokal, A., Hijazi, M., Akhtar, N., Osuntola, R., Rajeeve, V., Fitzgibbon, J., Travers, J., Britton, D., Khorsandi, S., & Cutillas, P. R. (2021). Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nature Communications*, *12*(1), 1–15. https://doi.org/10.1038/s41467-021-22170-8

González, F. (2015). Modelos de aprendizaje computacional en reumatología TT - Machine learning models in rheumatology. *Rev. Colomb. Reumatol*, *22*(2), 77–78. https://doi.org/https://doi.org/10.1016/j.rcreu.2015.06.001

Gratwohl, A. (2012). The EBMT risk score. *Bone Marrow Transplantation*, *47*(6), 749–756. https://doi.org/10.1038/bmt.2011.110

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

Instituto Nacional del Cáncer NCI. (2020). *Tratamiento de la leucemia mieloide aguda en adultos (PDQ®)–Versión para pacientes*. https://www.cancer.gov/espanol/tipos/leucemia/paciente/tratamiento-lma-adultos-pdq

Kimura, K., Tabe, Y., Ai, T., Takehara, I., Fukuda, H., Takahashi, H., Naito, T., Komatsu, N., Uchihashi, K., & Ohsaka, A. (2019). A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-49942-z

Krittanawong, C., Virk, H. U. H., Kumar, A., Aydar, M., Wang, Z., Stewart, M. P., & Halperin, J. L. (2021). Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection. *Scientific Reports*, *11*(1), 8992. https://doi.org/10.1038/s41598-021-88172-0

Lee, S.-I., Celik, S., Logsdon, B. A., Lundberg, S. M., Martins, T. J., Oehler, V. G., Estey, E. H., Miller, C. P., Chien, S., Dai, J., Saxena, A., Blau, C. A., & Becker, P. S. (2018). A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nature Communications*, *9*(1), 42. https://doi.org/10.1038/s41467-017-02465-5

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia. (2019). CONPES 3975 - Politica Nacional Para La Transformacion Digital e Inteligencia Artificial. In *Consejo Nacional de Politica Economica y Social - Republica de Colombia* (p. 115). https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3975.pdf

Moons, K. G. M., de Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Medicine*, *11*(10). https://doi.org/10.1371/journal.pmed.1001744

Morales Muñoz, L., Quintana, G., & Niño, L. F. (2015). Modelo computacional para la identificación de endofenotipos y clasificación de pacientes con artritis reumatoide a partir de datos genéticos, serológicos y clínicos, utilizando técnicas de inteligencia computacional. *Revista Colombiana de Reumatología*, *22*(2), 90–103. https://doi.org/10.1016/j.rcreu.2015.05.005

National Cancer Institute. (n.d.). *The Cancer Genome Atlas Program*. Retrieved September 3, 2021, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Onecha, E. M. (2019). *Implicaciones clínicas de la detección de mutaciones recurrentes mediante secuenciación masiva en leucemia mieloide aguda y correlación con la sensibilidad a fármacos antileucemia* [Universidad Complutense de Madrid].

https://eprints.ucm.es/id/eprint/51644/

Pardo-Gonzalez, C. A., Lagos-Ibarra, J. J., Linares-Ballesteros, A., Sarmiento-Urbina, I. C., Contreras-Acosta, A. D., Cabrera-Bernal, E. V., Uribe-Botero, G. I., Barros-García, G., & Aponte-Barrios, N. H. (2020). Resultados de la implementación del protocolo PETHEMA LPA 99 en el tratamiento niños con leucemia promielocitica aguda en Bogotá, Colombia. *Revista de La Facultad de Medicina*, *69*(2). https://doi.org/10.15446/revfacmed.v69n2.80152

Park, D. J., Park, M. W., Lee, H., Kim, Y. J., Kim, Y., & Park, Y. H. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-87171-5

Pfirrmann, M., Ehninger, G., Thiede, C., Bornhäuser, M., Kramer, M., Röllig, C., Hasford, J., & Schaich, M. (2012). Prediction of post-remission survival in acute myeloid leukaemia: a post-hoc analysis of the AML96 trial. *The Lancet Oncology*, *13*(2), 207–214. https://doi.org/10.1016/S1470-2045(11)70326-6

PROBAST group Julius Center for Health Sciences and Primary Care UMC Utrecht. (2019). *Probast*. https://www.probast.org/downloads/

Rodriguez-Villamizar, L. A., Rojas Díaz, M. P., Acuña Merchán, L. A., Moreno-Corzo, F. E., & Ramírez-Barbosa, P. (2020). Space-time clustering of childhood leukemia in Colombia: A nationwide study. *BMC Cancer*, *20*(1), 1–10. https://doi.org/10.1186/s12885-020-6531-2

Saaty, T. L., & Ozdemir, M. S. (2003). Why the magic number seven plus or minus two. *Mathematical and Computer Modelling*, *38*(3–4), 233–244. https://doi.org/10.1016/S0895-7177(03)90083-5

Salehnasab, C. (2021). An Intelligent Clinical Decision Support System for Predicting Acute Graft-versus-host Disease (aGvHD) following Allogeneic Hematopoietic Stem Cell Transplantation. *Journal of Biomedical Physics and Engineering*, *11*(3), 345–356. https://doi.org/10.31661/jbpe.v0i0.2012-1244

Sorror, M. L., Maris, M. B., Storb, R., Baron, F., Sandmaier, B. M., Maloney, D. G., & Storer, B. (2005). Hematopoietic cell transplantation (HCT)-specific comorbidity index: A new tool for risk assessment before allogeneic HCT. *Blood*, *106*(8), 2912–2919. https://doi.org/10.1182/blood-2005-05-2004

Tang, S., Chappell, G. T., Mazzoli, A., Tewari, M., Choi, S. W., & Wiens, J. (2020). Predicting Acute Graft-Versus-Host Disease Using Machine Learning and Longitudinal Vital Sign Data From Electronic Health Records. *JCO Clinical Cancer Informatics*, *4*, 128–135. https://doi.org/10.1200/cci.19.00105

The Global Cancer Observatory. (2020). *GLOBOCAN 2020: International Agency Research on Cancer*. *509*, 1–2.

Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*, *58*(5), 475–483. https://doi.org/10.1016/J.JCLINEPI.2004.06.017

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, *165*(6), 710–718. https://doi.org/10.1093/AJE/KWK052

WHO. (2020). Globocan 2020: Leukaemia. *International Agency for Research on Cancer*, *419*, 3–4. https://ascopost.com/news/december-2020/globocan-2020-database-provides-latest-global-data-on-cancer-burden-cancer-deaths/#:~:text=Female breast cancer has now,with 685%2C000 deaths in 2020.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/M18-1376

Anexos

Anexo I. Código de clase creada para realizar el proceso de entrenamiento

```
"""
@author: kricher
"""
import matplotlib.pyplot as plt
import numpy as np
import itertools
from time import time
import xgboost as xgb # Instalación: pip install xgboost
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report, confusion_matrix,
roc_auc_score, plot_roc_curve
from tensorflow import keras
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.wrappers.scikit_learn import KerasClassifier
from keras.callbacks import EarlyStopping, ReduceLROnPlateau,
LearningRateScheduler


def funcion_variacion_parametros(modelo, caracteristicas, valores,
parametros, scoring, refit):
    grid_search = GridSearchCV(modelo, parametros, n_jobs=-1,
scoring=scoring, refit=refit, verbose=1)

    strPrint = 'Ejecutando cálculos\n'
    # Inicio
    t0 = time()
    grid_search.fit(caracteristicas, valores)
    strPrint += 'Finalizado en %0.3fs' % (time() - t0)
    strPrint += '\nMejor puntaje: %0.3f' % grid_search.best_score_

    strPrint += '\nParámetros:'
    best_parameters = grid_search.best_estimator_.get_params()
    for param_name in sorted(parametros.keys()):
        strPrint += '\n%s: %r' % (param_name, best_parameters[param_name])

    return grid_search, strPrint

class ModeloIA:
    def __init__(self):
        self.modelo = None
        self.historia = None
        self.val_prediccion = None
        self.caracteristicas_pruebas = None
        self.valores_pruebas = None
        self.mc = None
        self.reporte_clas = None
        self.roc = None
```

```
    def entrenar(self, caracteristicas, valores):
        pass

    def predecir(self, caracteristicas, valores):
        self.caracteristicas_pruebas = caracteristicas
        self.valores_pruebas = valores
        self.val_prediccion = self.modelo.predict(caracteristicas)
        self.mc = confusion_matrix(self.valores_pruebas,
self.val_prediccion, normalize='true')
        # Redondear los valores de la matriz al normalizar
        self.mc = np.vectorize(lambda x: round(x, 2))(self.mc)
        self.reporte_clas = classification_report(self.valores_pruebas,
self.val_prediccion, output_dict=True, zero_division=0)
        self.roc = roc_auc_score(self.valores_pruebas,
self.modelo.predict_proba(caracteristicas)[:, 1])

    def matriz_confusion(self, modelo, caracteristicas, valores):
        self.modelo = modelo
        self.caracteristicas_pruebas = caracteristicas
        self.valores_pruebas = valores
        self.val_prediccion =
self.modelo.predict(self.caracteristicas_pruebas)
        self.mc = confusion_matrix(self.valores_pruebas,
self.val_prediccion, normalize='true')
        # Redondear los valores de la matriz al normalizar
        self.mc = np.vectorize(lambda x: round(x, 2))(self.mc)
        self.reporte_clas = classification_report(self.valores_pruebas,
self.val_prediccion, output_dict=False, zero_division=0)

    def graficar_matriz_confusion(self, nombres_clases):
        plt.imshow(self.mc, interpolation='nearest', cmap=plt.cm.Blues)
        plt.title('Matriz de confusión')
        plt.colorbar()
        tick_marks = np.arange(len(nombres_clases))
        plt.xticks(tick_marks, nombres_clases, rotation=0)
        plt.yticks(tick_marks, nombres_clases)

        thresh = (self.mc.max() - self.mc.min()) * (0.7) + self.mc.min()
        for i, j in itertools.product(range(self.mc.shape[0]),
range(self.mc.shape[1])):
            plt.text(j, i, self.mc[i, j], horizontalalignment="center",
color="white" if self.mc[i, j] > thresh else "black")

        plt.tight_layout()
        plt.ylabel('Realidad')
        plt.xlabel('Predicción')
        plt.show()

    def graficar_roc(self, nombre, ax):
        return plot_roc_curve(self.modelo, self.caracteristicas_pruebas,
self.valores_pruebas,
                              name=nombre, ax=ax)

    def ejecutar_variacion_parametros(self, caracteristicas, valores,
parametros, scoring, refit):
        pass

    def validacion_cruzada(self, caracteristicas, valores, folds, scoring):
        pass
```

```
class ModeloXGB(ModeloIA):
    def entrenar(self, caracteristicas, valores, l_rate, m_depth,
n_estimators):
        self.modelo = xgb.XGBClassifier(learning_rate=l_rate,
max_depth=m_depth, n_estimators=n_estimators,
                                        objective='binary:logistic',
eval_metric='logloss',
                                        use_label_encoder=False)
        self.modelo.fit(caracteristicas, valores)

    def ejecutar_variacion_parametros(self, caracteristicas, valores,
parametros, scoring, refit):
        modelo = xgb.XGBClassifier(objective='binary:logistic',
eval_metric='logloss',
                                   use_label_encoder=False)

        return funcion_variacion_parametros(modelo, caracteristicas,
valores, parametros, scoring, refit)

    def validacion_cruzada(self, caracteristicas, valores, folds, scoring):
        modelo = xgb.XGBClassifier(learning_rate=0.1, max_depth = 10,
n_estimators = 500)
        puntuaciones = cross_val_score(modelo, X = caracteristicas, y =
valores, cv=folds, scoring=scoring)
        return puntuaciones, puntuaciones.mean(), puntuaciones.std()


class ModeloRF(ModeloIA):
    def entrenar(self, caracteristicas, valores, criterion, m_depth,
n_estimators):
        self.modelo = RandomForestClassifier(n_estimators=n_estimators,
criterion=criterion, max_depth=m_depth)
        self.modelo.fit(caracteristicas, valores)

    def ejecutar_variacion_parametros(self, caracteristicas, valores,
parametros, scoring, refit):
        modelo = RandomForestClassifier()
        return funcion_variacion_parametros(modelo, caracteristicas,
valores, parametros, scoring, refit)

    def validacion_cruzada(self, caracteristicas, valores, folds, scoring):
        modelo = RandomForestClassifier(n_estimators=300, criterion='gini')
        puntuaciones = cross_val_score(modelo, X = caracteristicas, y =
valores, cv=folds, scoring=scoring)
        return puntuaciones, puntuaciones.mean(), puntuaciones.std()


class ModeloSVM(ModeloIA):
    def entrenar(self, caracteristicas, valores, C, gamma, kernel, degree):
        self.modelo = SVC(C=C, degree=degree, gamma=gamma, kernel=kernel,
probability=True)
        self.modelo.fit(caracteristicas, valores)

    def ejecutar_variacion_parametros(self, caracteristicas, valores,
parametros, scoring, refit):
        modelo = SVC(probability=True)
        return funcion_variacion_parametros(modelo, caracteristicas,
valores, parametros, scoring, refit)

    def validacion_cruzada(self, caracteristicas, valores, folds, scoring):
        modelo = SVC(degree=4, probability=True)
```

```
        puntuaciones = cross_val_score(modelo, X = caracteristicas, y =
valores, cv=folds, scoring=scoring)
        return puntuaciones, puntuaciones.mean(), puntuaciones.std()


class ModeloNN(ModeloIA):
    def __init__(self, entradas):
        self.entradas = entradas
        ModeloIA.__init__(self)

    def crear_red(self, dropout_rate=0.5, activation='relu'):
        modelo = keras.models.Sequential([
            keras.layers.Dense(256, input_dim=self.entradas,
activation=activation),
            keras.layers.BatchNormalization(),
            keras.layers.Dense(512, activation=activation),
            keras.layers.Dropout(dropout_rate),
            keras.layers.Dense(1, activation="sigmoid")
        ])

        if (self.toptimizer == 'a'):
            opt = Adam()
        else:
            opt = keras.optimizers.SGD(momentum=0.9)

        modelo.compile(loss="binary_crossentropy", optimizer=opt, metrics
=["accuracy"])
        return modelo

    def entrenar(self, caracteristicas, valores, drate, act, vs, opt='a'):
        self.toptimizer = opt
        self.modelo = KerasClassifier(build_fn= lambda:
self.crear_red(drate, act),
                                      epochs=100, verbose=2,
validation_split=vs)
        self.modelo._estimator_type = "classifier"

        m_callbacks = [
            LearningRateScheduler(lambda epoch: 1e-5 * 10**(epoch / 20))
            ]

        if (self.toptimizer == 'a'):
            self.historia = self.modelo.fit(caracteristicas, valores,
epochs=100, validation_split=vs,
                                            verbose = 2)
        else:
            self.historia = self.modelo.fit(caracteristicas, valores,
epochs=100, validation_split=vs,
                                            verbose = 2,
callbacks=m_callbacks)

    def ejecutar_variacion_parametros(self, caracteristicas, valores,
parametros, scoring, refit, opt='a'):
        self.toptimizer = opt
        self.modelo = KerasClassifier(build_fn=self.crear_red, epochs=100,
verbose=1)

        grid_search = GridSearchCV(estimator=self.modelo,
param_grid=parametros, scoring=scoring, refit=refit)
        grid_search.fit(caracteristicas, valores)
```

```
        return grid_search

    def predecir(self, caracteristicas, valores):
        self.caracteristicas_pruebas = caracteristicas
        self.valores_pruebas = valores
        # Predicciones
        yhat_probs = self.modelo.predict(caracteristicas, verbose=0)
        self.val_prediccion = (self.modelo.predict(caracteristicas) >
0.5).astype("int32")

        # Tratamiento de variables de predicciones para cálculo posterior y
presentación
        yhat_probs = yhat_probs[:, 0]
        self.val_prediccion = self.val_prediccion[:, 0]
        # Matriz de confusión normalizada
        self.mc = confusion_matrix(valores, self.val_prediccion,
normalize='true')
        self.mc = np.vectorize(lambda x: round(x, 2))(self.mc)
        self.reporte_clas = classification_report(valores,
self.val_prediccion, output_dict=True, zero_division=0)
        self.roc = roc_auc_score(valores, yhat_probs)

    def graficar_entreno(self):
        acc = self.historia.history['accuracy']
        val_acc = self.historia.history['val_accuracy']
        loss = self.historia.history['loss']
        val_loss = self.historia.history['val_loss']

        epochs = range(len(acc))

        fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(14, 6), dpi=80)
        #-----------------------------------------------
        # Imprimir la precisión del entrenamiento y validación por epoch
        #-----------------------------------------------
        ax[0].plot(epochs, acc, 'bo', label='Training accuracy')
        ax[0].plot(epochs, val_acc, 'b', label='Validation accuracy')
        ax[0].set_title('Precisión de entrenamiento y validación')
        ax[0].legend()

        #-----------------------------------------------
        # Imprimir la perdida de entrenamiento y validación por epoch
        #-----------------------------------------------
        ax[1].plot(epochs, loss, 'bo', label='Training Loss')
        ax[1].plot(epochs, val_loss, 'b', label='Validation Loss')
        ax[1].set_title('Pérdida de entrenamiento y validación')
        ax[1].legend()
        plt.show()
```

Anexo II. Lista de verificación PROBAST

Extraída de (PROBAST group Julius Center for Health Sciences and Primary Care UMC Utrecht, 2019)

Criteria	**Specify your systematic review question**
Intended use of model:	Predecir la mortalidad debida a la Leucemia Mieloide Aguda (LMA) en personas mayores de 18 años
Participants	Pacientes con LMA tratados bajo el esquema terapéutico especificado por el Programa Español de Tratamientos en Hematología (PETHEMA)
Predictors	Edad Género Clasificación de cariotipo MRC Especificación de Enfermedad mínima residual (EMR) Intensidad del tratamiento recibido Conteo de leucocitos Índice de masa corporal (IMC) Tipo de LMA Valoración del paciente por escala ECOG Diagnóstico de diabetes Diagnóstico de Hipertensión arterial (HTA) Cantidad de hemoglobina Conteo de Plaquetas
Outcome to be predicted:	Mortalidad de pacientes en un año

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development	SI	Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	NO	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation	NO	External validation of existing (previously developed) model in other participants.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Datos recopilados de pacientes diagnosticados con LMA según los criterios de la Organización Mundial de la Salud (OMS), tratados entre los años 2009 y 2021. Con corte 30 de junio de 2021, se cuentan 169 registros, con más de 300 variables que recopilan datos de diferentes momentos de la atención, evaluación y tratamiento de los pacientes.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	
1.2 Were all inclusions and exclusions of participants appropriate?		Y	
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**Low**	
Rationale of bias rating: Pacientes con diagnóstico confirmado, sin problemas de elegibilidad para selección, información registrada durante el tratamiento recibido por una institución avalada por el Ministerio de Salud.			
B. Applicability			
Describe included participants, setting and dates: Grupo de 169 pacientes tratados entre los años 2009 y 2021, diagnosticados con LMA según los criterios de la OMS.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**Low**	
Rationale of applicability rating: Tanto los pacientes como el entorno se encuentran bajo las condiciones del tratamiento médico específico de la enfermedad.			

DOMAIN 2: Predictors

A. Risk of Bias

List and describe predictors included in the final model, e.g. definition and timing of assessment:

Las variables predictoras fueron seleccionadas teniendo en cuenta criterios de relevancia epidemiológica, como la edad y el género, o según lo encontrado en la literatura, aceptado por la comunidad médica, teniendo en cuenta predictores que han demostrado tener un impacto en el pronóstico, la toma de decisiones para el tratamiento, o en la sobrevida de los pacientes.

Se realiza la toma de diferentes exámenes de laboratorio y pruebas diagnósticas durante el tiempo del tratamiento, que sirven para la definición del esquema a aplicar y la medición de los resultados obtenidos. Se encuentran dentro de este grupo exámenes para la determinación de la clasificación de cariotipo según el Medical Research Council (MRC), determinación de tipo de LMA y de padecimiento de Enfermedad mínima residual (EMR), diabetes e hipertensión arterial, conteo de leucocitos, plaquetas y cantidad de hemoglobina. Además, se realiza el registro de la intensidad de tratamiento recibido, el índice de masa corporal y la valoración del paciente utilizando la escala ECOG.

		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		PY	
2.2 Were predictor assessments made without knowledge of outcome data?		PY	
2.3 Are all predictors available at the time the model is intended to be used?		Y	
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	**Low**	

Rationale of bias rating:

La información obtenida fue tomada en un escenario controlado, durante el desarrollo del tratamiento médico aplicado. La información utilizada en el momento de la generación del modelo predictivo se encuentra disponible por tratarse de datos que se recopilan al inicio del esquema terapéutico.

B. Applicability

Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**Low**	

Rationale of applicability rating:

Los predictores considerados para el desarrollo del modelo fueron tomados durante el tratamiento de la enfermedad que se está analizando. No hay preocupación por una definición errónea o porque la información se haya recopilado fuera del momento del tratamiento desde el que se pretende realizar el análisis.

DOMAIN 3: Outcome

A. Risk of Bias

Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:

El resultado esperado fue determinado teniendo en cuenta los periodos de referencia tomados durante el desarrollo del tratamiento de la LMA. Se realizaron predicciones para la aparición del evento de muerte del paciente en diferentes periodos de tiempo, y se decidió incluir en la revisión el periodo de 1 año por las características de la información presentada en el conjunto de datos, atendiendo a contar con una cantidad balanceada de casos a predecir (aparición del evento y no aparición del evento).

		Dev	Val
3.1 Was the outcome determined appropriately?		Y	
3.2 Was a pre-specified or standard outcome definition used?		Y	
3.3 Were predictors excluded from the outcome definition?		Y	
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	
3.5 Was the outcome determined without knowledge of predictor information?		Y	
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	**Low**	

Rationale of bias rating:

B. Applicability

At what time point was the outcome determined:
Mortalidad después de 1 año de tratamiento.

If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:
No fue utilizado ningún resultado compuesto.

Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**Low**	

Rationale of applicability rating:
El resultado definido concuerda completamente con la pregunta de revisión.

DOMAIN 4: Analysis		
Risk of Bias		
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor: Cantidad de participantes: 165 pacientes Número de predictores candidatos: 17 Cantidad de Eventos: 67 Eventos por cantidad de predictores: 3.94		
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition): Se seleccionan algoritmos de Inteligencia artificial encontrados en el estado del arte en trabajos del mismo tipo, desarrollados para la predicción de diferentes condiciones y enfermedades. Basado en esto, se decide utilizar los algoritmos XGBoost, Random Forest, SVM y redes neuronales. El desarrollo del modelo se realizó teniendo en cuenta la cantidad de datos disponibles. Se realizó la separación de los datos en conjuntos de entrenamiento y pruebas para realizar un procedimiento de búsqueda de parámetros utilizando validación cruzada con el fin de determinar la configuración con la que se obtienen los mejores resultados de cada uno de los algoritmos teniendo en cuenta la métrica F1. Posteriormente se tomaron los modelos con los mejores resultados según un criterio de aceptación, y se entrenaron los mismos para medir su capacidad predictiva, analizando nuevas métricas, la matriz de confusión y AUC.		
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants): El modelo fue validado internamente aplicando el método de validación cruzada con los datos de entrenamiento, y posteriormente se validaron las predicciones comparando los resultados con el conjunto de pruebas analizando y comparando las métricas obtenidas.		
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism: El modelo predictivo alcanzó unos resultados de 0.88 de exactitud, 1 de sensibilidad y recall, y AUC de 0.92.		
Describe any participants who were excluded from the analysis: Se excluyeron variables del conjunto de datos inicial por contener gran cantidad de datos vacíos, debido a que se trataban de datos recopilados en otras etapas del tratamiento que no se incluyeron en el análisis.		
Describe missing data on predictors and outcomes as well as methods used for missing data: Los datos vacíos en el conjunto de datos utilizado en el desarrollo del modelo se debían a mal reporte de datos en el momento de la codificación en variables, y falta de información del tratamiento. Dependiendo del caso, se diligenció el campo con la categoría “Sin Dato”, y para el caso de las variables numéricas, se registraron pocos casos en los que se asignó la media de los valores presentes en el predictor. Datos vacíos en las salidas no fueron reportados.		
	Dev	Val
4.1 Were there a reasonable number of participants with the outcome?	PN	
4.2 Were continuous and categorical predictors handled appropriately?	Y	
4.3 Were all enrolled participants included in the analysis?	N	

4.4 Were participants with missing data handled appropriately?		Y	
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		PY	
4.7 Were relevant model performance measures evaluated appropriately?		Y	
4.8 Were model overfitting and optimism in model performance accounted for?		PY	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		Y	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	**High**	
Rationale of bias rating: La cantidad de datos utilizada en el desarrollo del modelo puede influir de manera significativa en los resultados obtenidos. Es recomendado incluir una cantidad mayor de datos y realizar validaciones externas.			

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	**High**
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	**Low**
Summary of applicability concerns:		

Anexo. Artículo de investigación

Modelo piloto de aprendizaje automático para la predicción de mortalidad por Leucemia Mieloide Aguda

Ricardo Hernández Martínez

Universidad Internacional de la Rioja, Logroño (España)

22 de septiembre de 2021

Resumen

La Leucemia Mieloide Aguda (LMA) es un tipo de cáncer de sangre que se encuentra entre los primeros con mayor cantidad de casos nuevos reportados en Colombia. El uso de aplicaciones basadas en Inteligencia Artificial ha aumentado en los últimos años, utilizadas como herramientas de apoyo al diagnóstico, pronóstico y tratamiento de enfermedades. En este trabajo se aplicaron algoritmos de aprendizaje supervisado para la creación de un modelo piloto predictivo de mortalidad de pacientes que padecen de LMA, siguiendo la metodología tradicional de proyectos de analítica de datos. Se utilizó para su desarrollo información recopilada en tratamientos bajo el protocolo terapéutico del Programa Español de Tratamientos en Hematología (PETHEMA). El modelo desarrollado alcanzó 0.88 de exactitud, con un AUC de 0.92. Los resultados obtenidos son un indicio favorable de la posibilidad de aplicar la Inteligencia Artificial en el campo de la medicina para el apoyo de la práctica clínica tradicional.

Palabras Clave

Inteligencia artificial, Leucemia mieloide aguda, Modelo predictivo.

I. Introducción

La leucemia es un tipo de cáncer de sangre en el que se observa un crecimiento anómalo de células inmaduras que saturan la médula ósea y el torrente sanguíneo, lo cual produce en la persona que padece la enfermedad una serie de complicaciones que pueden llevarla a la muerte. En Colombia, la Leucemia Mieloide Aguda (LMA) se encuentra entre los 10 tipos de cáncer con mayor cantidad de casos nuevos reportados, con una prevalencia que se ha visto en aumento durante los últimos años [1].

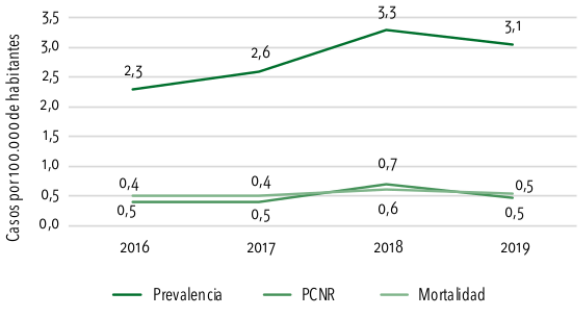

Figura 1. Medidas de frecuencia para LMA en adultos por cada 100.000 habitantes. Colombia 2015-2019 [1].

Por situaciones externas a la enfermedad, al paciente y al tratamiento, en el país existen condiciones adicionales relacionadas con el acceso y la organización del sistema de salud, que suponen una dificultad adicional que impacta de manera negativa los resultados en la lucha contra esta enfermedad.

Las herramientas digitales, y el análisis de la información disponible haciendo uso de tecnología de vanguardia, ofrecen una valiosa ayuda en el campo de la medicina poniendo al alcance del personal de salud los medios para realizar diagnósticos más rápidos y acertados, y para el análisis de información que permita predecir y crear políticas en beneficio de la comunidad. Consciente de esta situación, el Gobierno de Colombia tiene dentro de sus políticas el apoyo para la transformación digital e Inteligencia Artificial del Estado, con el fin de potenciar la generación de valor económico y social, favoreciendo la productividad y el bienestar de los ciudadanos, según lo expresó el Consejo Nacional de Política Económica y Social (CONPES) en el año 2019 [2].

Teniendo en cuenta lo mencionado anteriormente, y los avances tecnológicos actuales, en este trabajo se aplicaron algoritmos de Inteligencia Artificial para la creación de un modelo piloto predictivo de mortalidad para personas diagnosticadas con LMA que son atendidas en un hospital universitario ubicado en la ciudad de Floridablanca, Colombia, como un paso inicial para considerar la viabilidad de incluir análisis de este mismo estilo en la práctica médica regional y en las instituciones que aplican el protocolo del Programa Español de Tratamiento en Hematología (PETHEMA) como guía de tratamiento de esta enfermedad.

Se planteó el uso de una metodología propia de los proyectos de Inteligencia Artificial y *data science* para el desarrollo de este trabajo, planteando requisitos enfocados en cumplir con los objetivos trazados. El resultado es la obtención de un conjunto de datos ajustado a las condiciones encontradas y de un modelo predictivo que alcanzó 0.88 de exactitud, 1 de sensibilidad y recall, y un AUC de 0.92.

Un trabajo como el presentado en este documento tiene justificación si se analiza la situación de Colombia, no sólo si se tiene en cuenta la etiología particular que pueden tener ciertas enfermedades como la LMA en la población, sino además por la aún incipiente aplicación de los modelos de Inteligencia Artificial en la escena nacional. Los resultados obtenidos en este trabajo muestran las capacidades que la aplicación de algoritmos de aprendizaje supervisado pueden ofrecer al campo de la medicina a nivel local y también pueden ser tomados como un paso inicial que abre la puerta a un nuevo enfoque que apoye la práctica médica.

II. Estado del Arte

Métodos actuales de predicción utilizados en medicina

En el área de medicina se realizan evaluaciones que permiten determinar el estado de salud de los pacientes, teniendo en ocasiones la capacidad de predecir futuras consecuencias derivadas. En unidades de cuidados críticos, se aplican escalas como *Acute Physiology and Chronic Health Evaluation II* (APACHE II) o *Sequential Organ Failure Assessment* (SOFA), con el fin de calcular la probabilidad de muerte de un paciente. También son utilizadas para evaluar el riesgo de ocurrencia de eventos futuros asociados a morbilidades, y en algunos casos, es posible aplicar ciertas escalas para calcular la probabilidad de diagnósticos secundarios asociados a una condición de salud.

El uso de escalas requiere de conocimiento y atención específica de profesionales especializados. Esto hace que su aplicación se realice de forma netamente manual, que conlleva tiempo de procesamiento por parte del personal de salud, y la posibilidad de error implícita en cualquier tipo de actividad humana. Adicionalmente, estos modelos de predicción y escalas no son infalibles pues a menudo se desarrollan utilizando métodos inapropiados y están pobremente reportados, lo que hace difícil o incluso imposible juzgar la calidad metodológica [3].

Para la LMA se han creado escalas con el fin de evaluar las condiciones de los pacientes o para realizar predicciones sobre su estado futuro. Dentro de este último grupo de escalas se puede encontrar la escala CIBMTR, que predice la supervivencia general en pacientes con LMA activa en recaída, para quienes se está considerando el trasplante de células madre hematopoyéticas (*Hematopoietic Stem Cell Transplant*, siglas en inglés HSCT). [4]. Para medir el tratamiento post remisión óptimo de la LMA, existe la escala *Post-Remision Treatment* (PRT) fue desarrollada sobre la base de 586 pacientes con LMA entre los 15 y 60 años, tratados en el ensayo prospectivo AML96 del *Study Alliance Leukemia* (SAL) [5]. Para la predicción de la supervivencia de un paciente a una de las alternativas de tratamiento para la LMA, el HSCT, se encuentran escalas usadas ampliamente como la HCT-CI (*Hematopoietic Cell Transplantation – Comorbidity Index*) y la EBMT (*European Group for Blood and Marrow Transplantation*). La escala HCT-CI fue desarrollada para permitir la evaluación de riesgo antes de la realización de un trasplante alogénico [6]. Por su parte, la escala EBMT ofrece una manera sencilla para evaluar los riesgos de un paciente que va a ser sometido a un HSCT [7].

Las escalas y modelos de predicción han sido tenidos en cuenta para la creación de sistemas de información y aplicaciones para teléfonos inteligentes facilitando el acceso a las escalas más utilizadas por el personal médico, permitiendo diligenciar información para realizar la valoración de los pacientes.

Inteligencia Artificial y medicina

En el transcurso de los últimos años se ha probado la utilidad del uso de la Inteligencia Artificial en diferentes campos de la medicina. En el área de tratamiento de datos, el aprendizaje automático tiene un gran potencial en la identificación de patrones que pueden parecer ocultos al razonamiento común, ocultos al ser humano que no tiene la capacidad de detectarlos de manera eficiente, rápida y precisa. La aplicación de la Inteligencia Artificial juega además un papel importante teniendo en cuenta la cantidad, cada vez más grande y compleja, de datos que se genera cada día por todos los actores del sistema de salud, y la dificultad de poder extraer información de utilidad con los métodos tradicionales de análisis de datos. La Inteligencia Artificial aplicada en el campo de la medicina

Con la notable mejora de los modelos de Inteligencia Artificial ha aumentado a su vez su aplicación en estudios médicos, como lo refleja el aumento de trabajos de investigación que aplican estos algoritmos para la solución de diferentes problemas. Como prueba de ello, en la figura 2 se evidencia la cantidad de artículos publicados durante los últimos 30 años en MEDLINE, una de las bases de datos más reconocidas a nivel mundial, y clasificados bajo sus términos normalizados como “Inteligencia Artificial”.

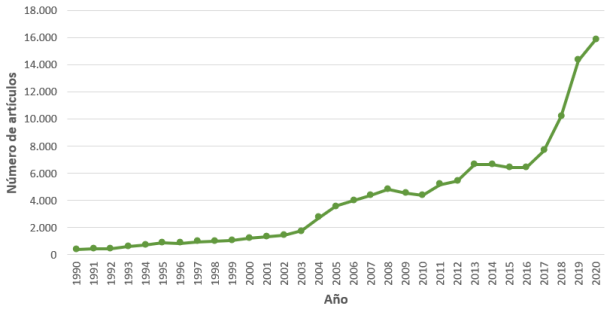

Figura 2. Aumento de estudios de Inteligencia Artificial aplicados al campo de medicina Fuente (https://pubmed.ncbi.nlm.nih.gov). Elaboración propia.

En el área de pronóstico y prevención, se han realizado investigaciones sobre la enfermedad cardiovascular aplicando modelos de aprendizaje automático y aprendizaje profundo en la predicción de la mortalidad de pacientes con disección espontánea de arterias coronarias (*Spontaneus Coronary Artery Disection*, siglas en inglés SCAD). A pesar que el curso clínico de la SCAD es variable, y que actualmente no existen métodos disponibles para realizar este tipo de predicción, se aplicaron los modelos mencionados a variables de registros de salud electrónicos de pacientes intrahospitalarios con SCAD para predecir su mortalidad. El índice de mortalidad durante la hospitalización de los casos analizados fue de 11.5%. Se compararon diversos modelos de aprendizaje automático a casos con información clínica completa, en los cuales los modelos de aprendizaje profundo mostraron los mejores resultados con alta precisión predictiva con AUC (*Area Under Curve*) de 0.98 (95% CI 0.97-0.99) [8].

Siguiendo la misma especialidad de la enfermedad cardiovascular, en Corea del Sur se realizó una comparación de modelos y escalas predictivas preexistentes de riesgo cardiovascular con algoritmos basados en aprendizaje automático. Los modelos previos mostraron moderada a buena discriminación para predecir futuros eventos cardiovasculares (C-statistics 0.70–0.80). Usando un modelo de red neuronal, se alcanzó la mayor C-statistic (0.751), la cual fue significativamente más alta entre las comparadas [9].

En el área de diagnóstico, La Inteligencia Artificial ofrece una gran capacidad de reconocimiento de patrones complejos a pesar

del ruido que pueda contener la fuente de datos analizada. Esto es especialmente útil en el área de análisis de imágenes diagnósticas como las que se procesan en el campo de la radiología, en donde su aplicación puede permitir la generación de modelos tridimensionales a partir de imágenes de pacientes concretos [10].

En la edición 22 de la Revista Colombiana de Reumatología del año 2015, se incluyó un artículo de investigación en el que se menciona la utilización de modelos de aprendizaje computacional para la clasificación de enfermedades haciendo uso de datos clínicos, genéticos y serológicos (González, 2015). En esta investigación los autores hicieron uso algoritmos de aprendizaje supervisado, como redes bayesianas y redes neuronales, que obtuvieron buenos resultados en la identificación de enfermedades como la artritis reumatoide, alcanzando una sensitividad y especificidad por encima del 92% en el proceso de clasificación

En el área de tratamiento, se está usando el aprendizaje automático para analizar la efectividad de los tratamientos médicos. Un estudio publicado en 2021, menciona la creación de la clasificación de medicamentos usando aprendizaje automático (*Drug Ranking Using Machine Learning*, siglas en inglés DRUML). En el DRUML se tienen en cuenta medicamentos para el tratamiento del cáncer, y se usaron modelos de aprendizaje automático entrenados con los datos de respuesta de células a más de 400 medicamentos, clasificándolos en función de su eficacia prevista para reducir la proliferación de una determinada población de células cancerosas. Inicialmente se evaluaron métodos de aprendizaje como estimación bayesiana de modelos lineales generalizados, mínimos cuadrados parciales, *Random Forest*, máquinas de vector de soporte (*Support Vector Machine*, siglas en inglés SVM), redes neuronales y modelos de aprendizaje profundo, los cuales mostraron las mejores métricas. Los resultados mostraron que DRUML clasifica los fármacos de diferente modo de acción en función de su eficacia prevista en diferentes tipos de cáncer con un error razonablemente bajo. En última instancia, DRUML podría ayudar en la priorización de fármacos al complementar la información obtenida de los parámetros clínico-patológicos y el análisis mutacional [13].

Inteligencia Artificial y LMA

En el área de visión computacional, se han adelantado estudios como el especificado por [14], en el que se desarrolló un sistema automático de apoyo al diagnóstico, que combina un sistema de reconocimiento de imágenes de células sanguíneas usando un modelo de aprendizaje profundo impulsado por redes neuronales convolucionales (*Convolutional Neural Network*, siglas en inglés CNN), y un sistema de decisión XGBoost. Se utilizó un conjunto de datos con imágenes de 695,030 células sanguíneas, obtenidas de 3,261 frotis de sangre periférica, que logra diferenciar Síndromes Mielodisplásicos (*Myelodysplastic syndromes*, siglas en ingles MDS) de la anemia aplásica (AA) con un 96.2% de sensibilidad y 100% de especificidad (AUC 0.99). En esta misma línea, se entrenó un modelo de CNN con muestras de 10,000 células que fueron anotadas manualmente, provenientes de frotis de aspirado de médula ósea (*Bone Marrow Aspirate*, siglas en inglés BMA), con el fin de realizar el análisis para la clasificación de trastornos hematológicos usando tejido no neoplásico. Se obtuvo como resultado que el modelo realiza la detección y clasificación con un AUROC (*Area Under the Receiver Operating Characteristic*) de hasta 0.98, mostrando niveles similares al utilizar muestras de LMA [15].

También se ha aprovechado la gran capacidad de análisis de volúmenes de datos que los algoritmos de Inteligencia Artificial poseen. En esta área se pueden encontrar trabajos como el publicado en la revista *Nature Communications* en el año 2018 [16]. En este estudio se analizó el efecto de los medicamentos en los pacientes, creando un nuevo método computacional basado en aprendizaje automático que identifica marcadores de expresión genética fiables para la sensibilidad a los fármacos, integrando información previa multiómica relevante para los procesos de la enfermedad. Se tomó como base los datos de 30 pacientes con LMA incluidos los perfiles de expresión genética de todo el genoma y sensibilidad in vitro a 160 medicamentos utilizados en quimioterapia.

De igual manera en esta área, se construyó un nuevo modelo conjunto optimizado, combinando un modelo DNN (*Deep Neural Network*) con dos modelos de aprendizaje automático, para la predicción de enfermedades utilizando como conjunto de datos los resultados de pruebas de laboratorio, incluyendo resultados de pruebas de sangre y de orina, relacionados con el diagnóstico final de cada paciente al momento del alta. Se tomó una muestra de 5,145 casos con 88 atributos, incluyendo el sexo y la edad del paciente, y se realizó la investigación sobre un total de 39 enfermedades específicas. Para el caso de la LMA, el modelo entrenado obtuvo excelentes resultados, realizando una predicción con un AUC de 0.99 [17].

Del mismo modo, en el área de análisis de datos, se encuentran estudios asociados al riesgo de desarrollar enfermedad de injerto contra huésped aguda (*acute Graft-versus-Host Disease*, siglas en inglés aGVHD) después de HSCT en el tratamiento de la LMA. En Japón, con datos de 26,695 pacientes, se entrenó un modelo *Alternating Decision Tree* (ADTree) que predice el riesgo de desarrollar aGVHD grado II-IV y III-IV, con una AUC de 0.616 y 0.622 respectivamente [18]. En Estados Unidos, se usaron datos de 324 pacientes, entre los que se incluyeron las mediciones de signos vitales tomadas desde el día de la infusión (día 0) hasta el día 9 después del trasplante. Con esta información se creó un conjunto de datos para realizar el entrenamiento de un modelo L2-*regularized logistic regression* que puede realizar la predicción de desarrollar aGVHD grado II-IV al día 100 posterior al trasplante, con un AUC de 0.659 (P=0.019) [19]. En Irán, se desarrolló un sistema de ayuda para toma de decisiones que puede ser consultado a través de Internet. Para su desarrollo, se creó un conjunto de datos que contaba con la información de 182 pacientes y 31 variables, el cual fue utilizado para entrenar modelos de clasificación entre los que se destacó el *Extreme Gradient Boosting Classifier* (XGBCClassifier), que puede predecir, el día del trasplante, el riesgo del paciente de desarrollar la aGVHD con una precisión de 90.7%, sensibilidad de 92.5% y especificidad de 89.13% [20].

Estudios de LMA en Colombia

En Colombia se han realizado estudios destinados a la investigación de diferentes tipos de enfermedades y cáncer hematológico. Entre los temas de los artículos asociados a la leucemia se pueden encontrar los estudios de cohortes, las series de casos, las pruebas diagnósticas en las principales ciudades, y el análisis de mortalidad infantil a nivel nacional. La mayoría de las investigaciones realizadas se centran en estudios epidemiológicos y clínicos, dejando un poco de lado el análisis biológico y molecular de la enfermedad, lo cual debería ser un aspecto a considerar si se tiene en cuenta la particular etiología que parece tener esta enfermedad en el país, comparado con otros, debido posiblemente al perfil genómico de la población [21].

Recientemente fue publicado un estudio sobre la implementación del protocolo PETHEMA LPA 99 en el tratamiento de niños con leucemia promielocítica aguda (LPA) en Bogotá D.C. Este tipo de leucemia es un subtipo de la LMA. La

motivación de este estudio es la mayor incidencia de este subtipo de leucemia en Latinoamérica comparado con otros lugares como Estados Unidos. En él se recopila la experiencia en el tratamiento de la LPA durante 7 años, con seguimiento hasta de 11.4 años. Se aplicó el protocolo PETHEMA LPA 99, diseñado para adultos, pues ha mostrado una supervivencia global (SG) mayor a 80%. Como parte de las limitaciones del estudio, se puede encontrar que se trata de una cohorte retrospectiva y que no fue realizado un análisis molecular a todos los pacientes. De la misma manera, en el estudio se analizó un número limitado de pacientes exclusivamente menores de 18 años. En el trabajo se concluye que, en términos generales, realizar la implementación del protocolo PETHEMA LPA 99 en el tratamiento de la población de estudio arroja resultados satisfactorios. Aunque los valores de supervivencia global reportados fueron menores a los descritos en poblaciones pediátricas de otros países, es posible recomendar el uso del protocolo PETHEMA LPA 99 en población pediátrica una vez ajustado a las características de este grupo etario según las recomendaciones del mismo [22].

En otro estudio publicado en el año de 2020, teniendo en cuenta que la leucemia es el cáncer más común en la niñez y que su tasa de incidencia en Colombia es una de las más altas de América, se propusieron conocer la distribución espacio-temporal de esta enfermedad [23]. Se analizaron 3,846 casos de niños menores de 15 años con diagnóstico confirmado de leucemia aguda entre 2009 y 2017. Se usaron estadísticas de escaneo espacio-temporal de Kulldorff, incluyendo el municipio y año de diagnóstico en el análisis. En este estudio se identificaron cinco grupos espaciales de leucemia infantil en diferentes regiones del país y grupos de tiempo específicos durante el período de estudio. Este resultado permite llegar a la conclusión que existen factores etiológicos o condiciones comunes por región asociados a la enfermedad que deberían ser estudiados.

A nivel preclínico, se han desarrollado en el mundo modelos para el estudio del cáncer con el fin de realizar una clasificación y caracterización de las leucemias, y la determinación de posibles factores causales. Siguiendo esta línea, se han realizado análisis en peces transgénicos para modelar la leucemia, desarrollando líneas que buscan establecer similitudes genéticas con esta patología entre estas especies y el ser humano [21]. Con base en lo anterior, en el país se adelantó un estudio en el que se obtuvieron resultados prometedores usando al pez cebra para el estudio preclínico de la leucemia, realizando experimentos de xenoinjertos de células Jurkat.

En el año 2015 se realizó un estudio donde se evaluó el costo-efectividad de las alternativas de tratamiento de consolidación en niños con LMA. Por una parte, el trasplante alogénico con progenitores hematopoyéticos, y por la otra el tratamiento con quimioterapia. Los datos para su elaboración fueron extraídos de estudios y reportes encontrados en la literatura científica, al igual que del Sistema de Información de Precios de Medicamentos (SISMED) del Ministerio de Salud y Protección Social para establecer el precio de los medicamentos teniendo en cuenta el año de referencia. El análisis realizado en el estudio se basó tomando como resultado los años de vida ganados después del tratamiento. Realizando cálculos de sensibilidad y probabilísticos, se concluyó que el trasplante resulta ser costo-efectivo en Colombia frente al tratamiento con quimioterapia [24].

Conclusiones de la revisión

Tras la revisión documental de trabajos con relevancia técnica y médica para la realización del presente trabajo, se verificó cómo en el ámbito mundial se realizan esfuerzos e investigaciones enfocadas a la aplicación de la Inteligencia Artificial, extrapolando las lecciones aprendidas en otras áreas de la sociedad y la industria, en busca de apoyar y mejorar el desempeño médico en los campos del pronóstico, diagnóstico y tratamiento.

La selección de la opción adecuada, dentro del abanico de posibilidades ofrecidas por los algoritmos de aprendizaje supervisado, varía dependiendo del origen de los datos analizados y del problema abordado. Se observaron modelos de gran utilidad en el tratamiento de imágenes, modelos basados en análisis probabilístico, geométrico o en árboles de decisión, que ofrecen cada uno diferentes bondades para alcanzar los objetivos buscados. El *Deep Learning*, apalancado en el aumento de la capacidad de almacenamiento y cómputo de las últimas décadas, es una opción de gran relevancia en los bancos de pruebas que buscan un modelo óptimo en diferentes tipos de aplicaciones.

Los modelos y sistemas predictivos o de diagnóstico pueden llegar a ser de gran utilidad en la práctica médica, apoyando al personal y siendo de ayuda para fortalecer las limitaciones humanas como la capacidad máxima de procesamiento simultáneo y el cansancio físico y mental. También han demostrado que mejoran ciertas tareas, pues se vio en trabajos como el presentado por Park, publicado en la revista *Scientific Reports* [17], que los modelos generados superaron en ocasiones la capacidad humana. Cabe señalar que se debe ser prudente aún con las herramientas generadas, teniendo en cuenta que los modelos predictivos desarrollados no son infalibles, y que en algunos casos en su generación se incurre en errores que derivan al final en su poca o nula popularización y utilización en la práctica clínica a gran escala.

En Colombia, en el área en particular, son pocas las publicaciones asociadas a la aplicación de la Inteligencia Artificial para el pronóstico o diagnóstico de enfermedades, situación que es diferente si se compara con la producción de la comunidad científica internacional. Sería importante abordar el tema buscando ofrecer una verdadera utilidad, y posteriormente avanzar en otros campos de manera similar a como se hace en otras partes del mundo, manteniendo así al país a la vanguardia del desarrollo y de la tecnología. Esto puede también llevar a tomar conciencia de la importancia de administrar de manera seria y correcta la información para su posterior análisis, lo que propende hacia el mejoramiento de los procedimientos médicos y tratamientos disponibles para los pacientes.

La Inteligencia Artificial aplicada al campo de medicina no se encuentra totalmente explorada, pudiendo ser de gran utilidad para la humanidad pues su aporte ha sido verificado en diferentes áreas. Teniendo en cuenta los algoritmos, herramientas utilizadas y recomendaciones leídas en los trabajos del estado del arte, es posible realizar un trabajo de calidad que sirva de base para el desarrollo de una herramienta útil, con la posibilidad de continuar avanzando en la consecución de una aplicación útil y de extender las lecciones aprendidas a otros ámbitos de la medicina.

III. Objetivos y Metodología

El objetivo general del trabajo fue crear un modelo piloto de aprendizaje supervisado que permita predecir la mortalidad en pacientes con LMA, basado en información clínica de pacientes tratados en un hospital universitario de la ciudad de Floridablanca, Colombia.

Este trabajo es un análisis anidado al proyecto “Registro epidemiológico de pacientes adultos con Leucemia Mieloide Aguda" que se desarrolla en un hospital universitario del oriente colombiano, en cabeza de la Jefe del Servicio de Hematología y Unidad de Trasplante y Terapia Celular de dicha institución. Para

la conformación de la base de datos se tomaron las variables recopiladas en el registro LMA, consignadas de manera anonimizada. Tomando como base este registro de pacientes atendidos, se planteó la realización de un estudio poblacional retrospectivo en el que se tuvo en cuenta la información clínica de pacientes mayores de 18 años diagnosticados con LMA según los criterios de la Organización Mundial de la Salud (OMS). Con corte 30 de junio de 2021, se cuenta con el registro de 169 pacientes, con información de variables que recopilan datos de diferentes momentos de la atención, evaluación y tratamiento de los pacientes

Se siguió una metodología de trabajo característica de proyectos de minería de datos, *data science* o Inteligencia Artificial.

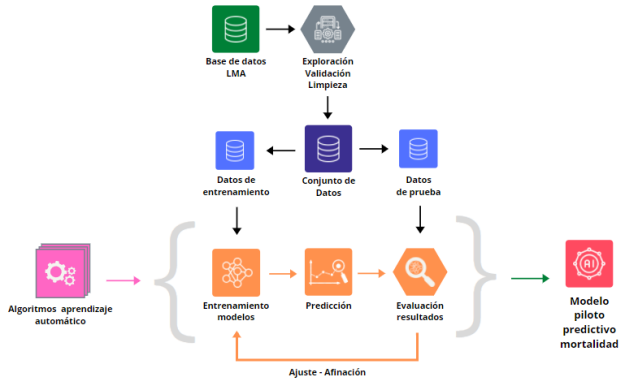

Figura 3. Metodología general. Elaboración propia

Debido a la cantidad de registros disponible, en el proceso de entrenamiento de los modelos se tuvo en cuenta el riesgo de obtener sobreajuste (*overfitting*), fenómeno que se presenta en los modelos entrenados que se puede resumir en que estos aprenden muy bien las características del conjunto de datos trabajado, perdiendo su capacidad de generalización posterior. Como medida para prevenir esto, se hizo uso de la técnica de validación cruzada usando los datos de entrenamiento, algo que también sirvió para tener una idea de la capacidad predictiva de los modelos.

IV. Contribución

Exploración, análisis y creación del conjunto de datos

En este trabajo se trabaja con datos que fueron recopilados durante el tratamiento de pacientes con LMA bajo el protocolo terapéutico establecido por PETHEMA. En términos generales, este protocolo consta de un tratamiento inicial de inducción buscando llegar a remisión completa (RC), para posteriormente aplicar un tratamiento de consolidación dependiendo de la edad del paciente [25].

En su tesis doctoral, Onecha menciona algunos datos estadísticos obtenidos del tratamiento de la LMA:

> Aproximadamente el 20% de los pacientes con LMA presentan refractariedad primaria al tratamiento de inducción y no llegan a alcanzar RC. Y en torno al 50% de los casos la enfermedad resurge y el paciente recae. Ambos escenarios, refractariedad primaria y recaída, suponen fracaso terapéutico asociado a pronóstico adverso y menos del 30% de los pacientes sobreviven 12 meses después de una recaída. [25].

De la misma manera, se menciona que "varios factores condicionan el pronóstico adverso en pacientes que sufren recaída, incluyendo citogenética adversa detectada en el diagnóstico, estado en RC menor a 12 meses, edad avanzada, y recaídas posteriores a un trasplante hematopoyético." [25]

En el conjunto de datos recibido se encuentra información de las diferentes fases del tratamiento de los pacientes, desde la etapa de inducción, pasando por las diferentes etapas de consolidación según el desarrollo y respuesta obtenidos. Los pacientes no pasan por todas las etapas del tratamiento, algo que puede ser inferido de lo citado anteriormente de Onecha [25]. Por esta razón se definió un punto de corte que sirviera de referencia, buscando que las condiciones de los pacientes fueran los más similares posibles y que se contara con la misma información de cada uno de ellos. Por este motivo, y teniendo en cuenta el protocolo terapéutico, se eligió trabajar con la información de los pacientes hasta la etapa de inducción, donde se cuenta con información de pruebas de laboratorio, datos clínicos y algunos análisis genéticos.

Para obtener el conjunto de datos necesario para efectuar el desarrollo del modelo, se realizó un proceso de análisis, validación y limpieza de datos, en el que se eliminaron registros incompletos de pacientes, se eliminaron variables que no contaban con la información suficiente o con información relevante que aportar al proceso, se realizó la revisión de *outliers* y se analizó la correlación de las variables numéricas. Para prevenir el sobreajuste, se tuvo en cuenta el número de observaciones o eventos por variable (EPV) para calcular el tamaño de la muestra recomendada para realizar el entrenamiento del modelo. Un tamaño de muestra con un EPV de 10 o más veces es recomendado con frecuencia para evitar el sobreajuste en modelos predictivos [26][27].

Para llegar al EPV recomendado se eligieron variables de relevancia epidemiológica, como la edad y el género, así como otras que según lo encontrado en la literatura han demostrado tener un impacto en el pronóstico, la toma de decisiones para el tratamiento, o en la sobrevida de los pacientes. Se decidió trabajar con las 13 variables especificadas en la tabla I.

Tabla I

Variables seleccionadas para el conjunto de datos de entrenamiento

Mayor de 65 años	Tipo de LMA
Género	Escala ECOG
Clasificación de cariotipo MRC	Diabetes
Enfermedad mínima residual (EMR)	Hipertensión arterial (HTA)
Intensidad del tratamiento	Cantidad de hemoglobina
Conteo de Leucocitos mayor que 50.000/mm^3	Conteo de plaquetas
Índice de masa corporal (IMC) mayor que 25	

Se aplicaron una serie de transformaciones para obtener el conjunto de datos que se va a usar en el entrenamiento, teniendo en cuenta que las variables cumplieran con ciertos parámetros que garanticen el buen funcionamiento de los algoritmos. Se realizaron reclasificaciones, normalizaciones y codificaciones según el caso presentado. El conjunto de datos se constituye con 165 registros y 17 variables. La cantidad seleccionada de variables permite que el EPV de la muestra se aproxime a la

recomendación (proporción de 10 a 1) según la cantidad de observaciones.

Entrenamiento de modelos de aprendizaje automático

Después de la revisión de la literatura de trabajos previos del estado del arte en los que se usaron conjuntos de datos similares al conjunto utilizado en este trabajo para la creación de modelos predictivos, se decidió trabajar con los algoritmos *XGBoost*, *Random Forest*, SVM y redes neuronales, algoritmos estos de aprendizaje supervisado que se basan cada uno en conceptos y principios diferentes para resolver problemas de clasificación binaria y multivariable. La variable a predecir (dependiente) fue el evento muerte – tiempo transcurrido entre diagnóstico y fallecimiento – de los pacientes tratados en diferentes periodos de tiempo. Se muestran en la figura 4 en el tiempo entre diagnóstico y la aparición del evento para los pacientes del conjunto de datos.

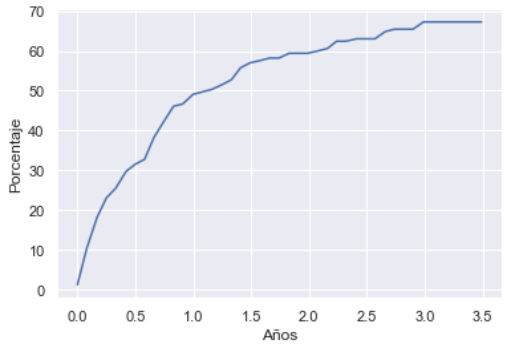

Figura 4. Tiempo entre diagnóstico y muerte en pacientes del conjunto de datos

Se observa que la cantidad de muertes es mayor a medida que transcurre tiempo desde el diagnóstico, teniendo en este conjunto de datos la aparición del evento muerte en la mitad de los pacientes el primer año de tratamiento.

Se enfocó el análisis en predecir la mortalidad de los pacientes para los periodos de tiempo de 30 días, 6 meses y 2 años. Estos intervalos de tiempo son periodos de observación tomados como referencia, usados habitualmente en el tratamiento de la LMA. Adicionalmente se incluyó en el análisis el periodo de 1 año desde el diagnóstico, teniendo en cuenta que, como fue mencionado anteriormente, para este lapso de tiempo la cantidad de pacientes fallecidos equivale a la mitad de las muestras disponibles. Lo anterior para realizar el entrenamiento de los modelos con la variable objetivo lo más balanceada posible, lo cual según la teoría incide en la obtención de mejores resultados de predicción.

Se dividió el conjunto de datos disponible en 2 conjuntos: entrenamiento y pruebas. Se formaron conjunto de datos para cada periodo de tiempo de análisis (30 días, 6 meses, 1 año y 2 años), particionando los datos en una distribución 80% - 20% y haciendo un reparto de información entre los conjuntos de manera aleatoria y estratificada. En cada conjunto de datos se tenían las mismas variables predictoras, modificando la variable objetivo dependiendo de la aparición del evento en ese intervalo de tiempo (caso positivo). Con el conjunto de datos de entrenamiento se realiza el entrenamiento de los modelos propiamente dicho, y con el de pruebas se realizaron las validaciones que permiten verificar la capacidad de predicción del modelo y el cálculo de métricas para medir el desempeño del mismo (exactitud, precisión, recall, F1 score, etc.). Se realizaron entrenamientos variando los hiperparámetros de los modelos, para buscar la mejor configuración de cada uno en cada periodo de tiempo analizado. En la tabla 2 se pueden observar los resultados de los entrenamientos mencionados.

		30 días	180 días	1 año	2 años
XBGoost	H.	{'learning_rate': 1, 'max_depth': 5, 'n_estimators': 50}	{'learning_rate': 1e-3, 'max_depth': 5, 'n_estimators': 500}	{'learning_rate': 1e-4, 'max_depth': 5, 'n_estimators': 50}	{'learning_rate': 1, 'max_depth': 5, 'n_estimators': 50}
	F1	0.257	0.625	0.815	0.739
Random Forest	H.	{'criterion': 'entropy', 'max_depth': None, 'n_estimators': 10}	{'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 100}	{'criterion': 'gini', 'max_depth': None, 'n_estimators': 100}	{'criterion': 'gini', 'max_depth': None, 'n_estimators': 100}
	F1	0.2	0.623	0.835	0.787
SVM	H.	{'C': 20, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}	{'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}	{'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}	{'C': 1, 'degree': 2, 'gamma': 'auto', 'kernel': 'rbf'}
	F1	0.38	0.593	0.825	0.773
Red neuronal Adaptativo	H.	{'activation': 'relu', 'dropout_rate': 0.7}	{'activation': 'relu', 'dropout_rate': 0.5}	{'activation':'sigmoid', 'dropout_rate': 0.5}	{'activation': 'sigmoid', 'dropout_rate': 0.7}
	F1	0.358	0.597	0.758	0.757
Red Neuronal Gradiente	H.	{'activation': 'relu', 'dropout_rate': 0.5}	{'activation': 'relu', 'dropout_rate': 0.2}	{'activation': 'relu', 'dropout_rate': 0.5}	{'activation': 'sigmoid', 'dropout_rate': 0.2}
	F1	0.358	0.651	0.763	0.751

H.: Hiperparámetros

Tabla 2. Resultado de búsqueda de hiperparámetros

V. Evaluación y Resultados

Se tomó como parámetro de aceptación la obtención de una métrica mayor o igual a 0.8 en el proceso de variación realizado. Se puede determinar que con los modelos entrenados para la predicción del evento muerte a los 30 días, 180 días y 2 años no se alcanza a cumplir con este requisito. Por otro lado, algunos de los modelos obtenidos para la predicción del evento en el primer año cumplen con el criterio de aceptación propuesto.

Se revisaron las matrices de confusión, algunas métricas adicionales propias del análisis de problemas de clasificación, y curvas ROC para los modelos *XGBoost*, *Random Forest* y SVM, los cuales cumplen con el criterio de aceptación planteado (figuras 5 y 6).

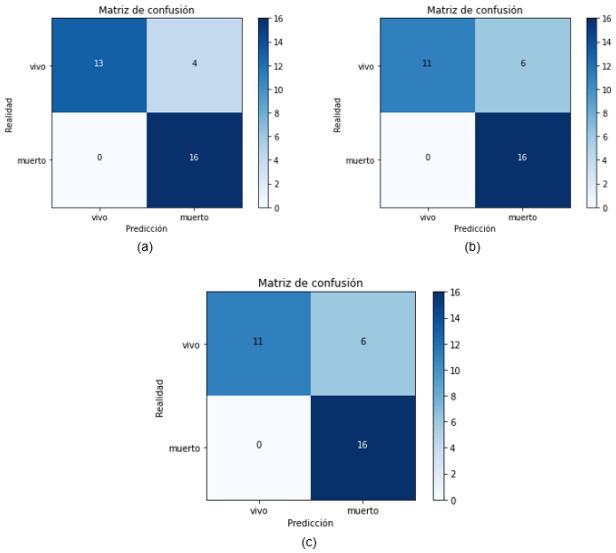

Figura 5. Matrices de confusión de predicciones de evento muerte 1 año (a) XGBoost (b) Random Forest (c) SVM

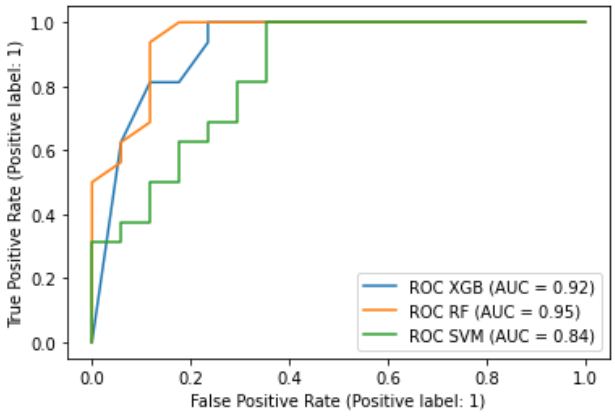

Figura 6. AUROC de las predicciones de evento de muerte 1 año

En la tabla 3 se pueden observar métricas derivadas de la matriz de confusión como la exactitud, sensibilidad, especificidad, precisión, recall y F1, las cuales sirven de utilidad para conocer el comportamiento de un modelo de clasificación. Se puede observar que el modelo *XGBoost* obtuvo los mejores resultados para el caso analizado, superando a los demás en las métricas calculadas.

Modelo	Exactitud	Sensibilidad	Especificidad	Precisión	Recall	F1
XGBoost	0.88	1	0.76	0.8	1	0.89
Random Forest	0.82	1	0.65	0.73	1	0.84
SVM	0.82	1	0.65	0.73	1	0.84

Tabla 3. Métricas de modelos predictivos analizados para el evento de muerte en 1 año

Se observó una tendencia del mejor modelo a clasificar de manera más adecuada los casos positivos que los negativos, algo que se ve reflejado en la obtención de una mayor sensibilidad que especificidad. Esto se evidenció de igual manera al analizar la matriz de confusión obtenida en la predicción.

VI. Discusión

El modelo para la predicción del evento muerte del paciente en el primer año de tratamiento alcanzó unos resultados de 0.88 de exactitud, 1 de sensibilidad y recall, y AUC de 0.92. Este es un resultado comparable con otros obtenidos en algunos de los trabajos revisados en el estado del arte que, teniendo en cuenta que se alcanza con un modelo piloto entrenado con una muestra pequeña de datos, hace presagiar la obtención de mejores resultados al incluir más información que aumente la cantidad de observaciones disponible para realizar entrenamientos en el futuro. Es inevitable pensar que las posibilidades de éxito del tipo de análisis realizado en este trabajo son altas, pero es necesaria la realización de una validación externa que permita analizar el modelo de manera exhaustiva, de cara a tener en cuenta más factores y conceptos en su realización para conseguir mejores resultados.

El modelo desarrollado predice de manera correcta el 88% de los casos totales presentados. Tiene una tendencia a clasificar de manera más adecuada los casos positivos (evento) que los negativos (sin evento), y realiza una clasificación correcta de la ocurrencia del evento muerte en el primer año con una precisión del 80%. La inclusión de características adicionales puede mejorar los resultados que obtenidos, algo que no pudo ser realizado en este trabajo pues el conjunto de observaciones era limitado y se debía cumplir el EPV para obtener un resultado aceptable que permitiera analizar si la aplicación de estos modelos puede traer beneficios en este campo.

Además del análisis de la capacidad predictiva del modelo, se realizó una evaluación final para determinar la relevancia de las variables utilizadas por el mismo. Para este fin, se utilizó el método SHAP (*Shapley Additive Explanations*), desarrollado por Lundberg y Lee [28]. Este método se basa en un concepto en teoría de juegos, *Shapley Value*, que indica cuánto contribuye cada variable en la obtención de un resultado [17]. SHAP asigna a cada variable o característica un valor de importancia para predicción particular. En la figura 7 se puede observar el resultado del análisis realizado con este método.

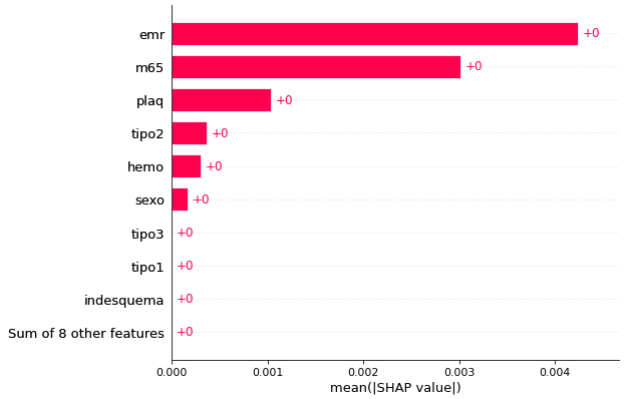

Figura 7. Variables relevantes según método SHAP

Este orden se ajusta al criterio de relevancia en el pronóstico de muerte en el tratamiento de LMA. Este es un buen punto comparativo que indica que los resultados del modelo pueden ser validados con el conocimiento previo de años de experiencia en el pronóstico de la condición analizada, sin importar que se trate de un piloto que aún tiene condiciones por mejorar.

VII. Conclusiones

Durante el proceso de desarrollo del presente trabajo se presentaron diversas limitaciones en el proceso de obtención de la información. Es necesario pasar por una serie de pasos para poder acceder a las bases de datos, esperar por la aprobación de Comités de Ética de las instituciones a la luz de la Guía de Buenas Prácticas Clínicas, algo que puede ser entendido sobre la base de la importancia de cuidar los datos para garantizar la privacidad de los pacientes y la protección de sus derechos.

La información se encuentra en gran medida atomizada, con diferentes centros sirviendo de custodios de fragmentos de la información, lo cual dificulta el poder reunir volúmenes de datos que permitan la realización de análisis y estudios que beneficien la práctica clínica y en últimas el acceso a una medicina cada vez más personalizada. Es necesaria la creación de convenios y acuerdos que permitan la correcta integración de la información con los diferentes actores del sistema de salud aportando su experiencia traducida en forma de datos. Esto puede también llevar a tomar conciencia de la importancia de administrar de manera seria y correcta la información para su posterior análisis, y a trabajar de manera conjunta siguiendo un objetivo común que

beneficie a los pacientes y al sistema de salud.

En la búsqueda de conjuntos de datos para la realización de este trabajo, se tomó nota que existen algunos esfuerzos para la integración de información de tratamientos médicos como lo adelantado por el Ministerio de Salud con la creación y reglamentación del RIPS (Registro Individual de Prestaciones de Servicios de Salud), registro donde se centralizan datos a nivel nacional de procedimientos, consultas y diagnósticos. Se necesita una política clara enfocada al aprovechamiento de la información recopilada pues esta base de datos es usada principalmente para facturación de procedimientos y la obtención de datos estadísticos. Prueba de ello es que no existen muchos trabajos que usen esta información con fines investigativos o de análisis.

Aunque la comunidad internacional produce cada vez más trabajos apoyándose en la Inteligencia Artificial, haciendo estudios de diversos tipos para el tratamiento de imágenes diagnósticas, el análisis de información clínica o genética, el seguimiento a tratamientos para estudiar el impacto y efectividad de ciertos medicamentos según se vio en la literatura analizada en el estado del arte, en Colombia esta área se encuentra rezagada y no ha despertado el interés que sí despierta en otros lugares del mundo.

Se evidencia la evolución de la situación en el área de la Inteligencia Artificial y analítica de datos. El impulso del avance tecnológico de las últimas décadas y el interés de la comunidad de desarrollo que ha fijado su vista en lo que se ha llamado *data science*, ha permitido la aparición de herramientas que han facilitado la implementación de algoritmos que pueden ser ejecutados sin la necesidad de tener grandes centros de cómputo, lo cual ha repercutido de manera positiva en los tiempos de desarrollo de este tipo de proyectos.

Los resultados obtenidos en este trabajo indican que es posible aplicar la Inteligencia Artificial en el campo de la medicina a nivel local, lo cual se encuentra acorde al consenso obtenido en trabajos de la misma área. En Colombia, es posible iniciar aprovechando los avances obtenidos por la comunidad científica internacional para aplicarlos a nivel regional, ajustando las lecciones aprendidas y generando nuevos resultados que se adapten a la realidad demográfica, social y económica del país.

Se recomienda continuar con el análisis realizado en este trabajo, incluyendo otros conjuntos de datos al entrenamiento del modelo de Inteligencia Artificial. Por orden natural, se deberían incluir los registros de pacientes tratados siguiendo el protocolo PETHEMA aprovechando que la estructura de datos es la misma que la tratada en el desarrollo realizado.

Pueden ser desarrollados otros trabajos en los que se tengan en cuenta factores específicos para realizar análisis por raza, género, análisis etarios o geográficos, de manera que se pueda personalizar cada vez más el tratamiento en medicina. Se han adelantado esfuerzos como el Atlas del Genoma del Cáncer, un programa iniciado en 2006 en el que se han caracterizado molecularmente 20.000 cánceres primarios emparejando muestras de 33 tipos de cáncer [29]. Este proyecto ha generado más de 2.5 petabytes de información y ofrece estos datos para que puedan ser accedidos y utilizados de manera pública. Se podría analizar si esta información puede ser utilizada en el ámbito local. Con bases de datos de perfiles genéticos, se pueden plantear investigaciones que permitan caracterizar las enfermedades ajustándolas a la realidad colombiana.

Con el aumento en la cantidad de registros disponibles en el conjunto de datos a analizar, sería importante profundizar en el uso de las redes neuronales aplicadas al análisis y generación de modelos para la predicción, diagnóstico y tratamiento. Este tipo de algoritmos ha probado su utilidad en diferentes ámbitos, y explota su potencial a medida que procesa grandes volúmenes de datos.

Existe una línea de trabajo amplia que beneficiaría diferentes líneas de proyectos, no sólo proyectos de Inteligencia Artificial, y es la de la integración de la información para crear una o varias bases de datos especializadas que puedan ser utilizadas y sirvan como base para la realización de investigaciones. Para lograr esto se debe crear una política dirigida que permita realizar esta labor de manera juiciosa, con pautas claras, que no sólo tome información existente, sino que siente las bases para empezar a generar valor en los lugares donde aún no se realiza, soportada sobre una infraestructura tecnológica que permita el acceso seguro y la disponibilidad de los datos.

REFERENCIAS

[1] Cuenta de Alto Costo. (2019). Situación del cáncer en población adulta en el SGSSS de Colombia, 2019. *Fondo Colombiano de Enfermedades de Alto Costo*, 336.

[2] Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia. (2019). CONPES 3975 - Politica Nacional Para La Transformacion Digital e Inteligencia Artificial. In *Consejo Nacional de Politica Economica y Social - Republica de Colombia* (p. 115). https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3975.pdf

[3] Dekker, F. W., Ramspek, C. L., & Van Diepen, M. (2017). Con: Most clinical risk scores are useless. *Nephrology Dialysis Transplantation*, *32*(5), 752–755. https://doi.org/10.1093/ndt/gfx073

[4] Duval, M., Klein, J. P., He, W., Cahn, J. Y., Cairo, M., Camitta, B. M., Kamble, R., Copelan, E., De Lima, M., Gupta, V., Keating, A., Lazarus, H. M., Litzow, M. R., Marks, D. I., Maziarz, R. T., Rizzieri, D. A., Schiller, G., Schultz, K. R., Tallman, M. S., & Weisdorf, D. (2010). Hematopoietic stem-cell transplantation for acute leukemia in relapse or primary induction failure. *Journal of Clinical Oncology*, *28*(23), 3730–3738. https://doi.org/10.1200/JCO.2010.28.8852

[5] Pfirrmann, M., Ehninger, G., Thiede, C., Bornhäuser, M., Kramer, M., Röllig, C., Hasford, J., & Schaich, M. (2012). Prediction of post-remission survival in acute myeloid leukaemia: a post-hoc analysis of the AML96 trial. *The Lancet Oncology*, *13*(2), 207–214. https://doi.org/10.1016/S1470-2045(11)70326-6

[6] Sorror, M. L., Maris, M. B., Storb, R., Baron, F., Sandmaier, B. M., Maloney, D. G., & Storer, B. (2005). Hematopoietic cell transplantation (HCT)-specific comorbidity index: A new tool for risk assessment before allogeneic HCT. *Blood*, *106*(8), 2912–2919. https://doi.org/10.1182/blood-2005-05-2004

[7] Gratwohl, A. (2012). The EBMT risk score. *Bone Marrow Transplantation*, *47*(6), 749–756. https://doi.org/10.1038/bmt.2011.110

[8] Krittanawong, C., Virk, H. U. H., Kumar, A., Aydar, M., Wang, Z., Stewart, M. P., & Halperin, J. L. (2021). Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection. *Scientific Reports*, *11*(1), 8992. https://doi.org/10.1038/s41598-021-88172-0

[9] Cho, S.-Y., Kim, S.-H., Kang, S.-H., Lee, K. J., Choi, D., Kang, S., Park, S. J., Kim, T., Yoon, C.-H., Youn, T.-J., & Chae, I.-H. (2021). Pre-existing and machine learning-based models for cardiovascular risk prediction. *Scientific Reports*, *11*(1), 1–10. https://doi.org/10.1038/s41598-021-88257-w

[10] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

[11] González, F. (2015). Modelos de aprendizaje computacional en reumatología TT - Machine learning models in rheumatology. *Rev. Colomb. Reumatol*, *22*(2), 77–78. https://doi.org/https://doi.org/10.1016/j.rcreu.2015.06.001

[12] Morales Muñoz, L., Quintana, G., & Niño, L. F. (2015). Modelo computacional para la identificación de endofenotipos y clasificación de pacientes con artritis reumatoide a partir de datos

genéticos, serológicos y clínicos, utilizando técnicas de inteligencia computacional. *Revista Colombiana de Reumatología*, *22*(2), 90–103. https://doi.org/10.1016/j.rcreu.2015.05.005

[13] Gerdes, H., Casado, P., Dokal, A., Hijazi, M., Akhtar, N., Osuntola, R., Rajeeve, V., Fitzgibbon, J., Travers, J., Britton, D., Khorsandi, S., & Cutillas, P. R. (2021). Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nature Communications*, *12*(1), 1–15. https://doi.org/10.1038/s41467-021-22170-8

[14] Kimura, K., Tabe, Y., Ai, T., Takehara, I., Fukuda, H., Takahashi, H., Naito, T., Komatsu, N., Uchihashi, K., & Ohsaka, A. (2019). A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-49942-z

[15] Chandradevan, R., Aljudi, A. A., Drumheller, B. R., Kunananthaseelan, N., Amgad, M., Gutman, D. A., Cooper, L. A. D., & Jaye, D. L. (2020). Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Laboratory Investigation*, *100*(1), 98–109. https://doi.org/10.1038/s41374-019-0325-7

[16] Lee, S.-I., Celik, S., Logsdon, B. A., Lundberg, S. M., Martins, T. J., Oehler, V. G., Estey, E. H., Miller, C. P., Chien, S., Dai, J., Saxena, A., Blau, C. A., & Becker, P. S. (2018). A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nature Communications*, *9*(1), 42. https://doi.org/10.1038/s41467-017-02465-5

[17] Park, D. J., Park, M. W., Lee, H., Kim, Y. J., Kim, Y., & Park, Y. H. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-87171-5

[18] Arai, Y., Kondo, T., Fuse, K., Shibasaki, Y., Masuko, M., Sugita, J., Teshima, T., Uchida, N., Fukuda, T., Kakihana, K., Ozawa, Y., Eto, T., Tanaka, M., Ikegame, K., Mori, T., Iwato, K., Ichinohe, T., Kanda, Y., & Atsuta, Y. (2019). Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Advances*, *3*(22), 3626–3634. https://doi.org/10.1182/bloodadvances.2019000934

[19] Tang, S., Chappell, G. T., Mazzoli, A., Tewari, M., Choi, S. W., & Wiens, J. (2020). Predicting Acute Graft-Versus-Host Disease Using Machine Learning and Longitudinal Vital Sign Data From Electronic Health Records. *JCO Clinical Cancer Informatics*, *4*, 128–135. https://doi.org/10.1200/cci.19.00105

[20] Salehnasab, C. (2021). An Intelligent Clinical Decision Support System for Predicting Acute Graft-versus-host Disease (aGvHD) following Allogeneic Hematopoietic Stem Cell Transplantation. *Journal of Biomedical Physics and Engineering*, *11*(3), 345–356. https://doi.org/10.31661/jbpe.v0i0.2012-1244

[21] Gacha Garay, M. J., Akle, V., Enciso, L., & Garavito Aguilar, Z. V. (2017). La leucemia linfoblástica aguda y modelos animales alternativos para su estudio en Colombia. *Revista Colombiana de Cancerología*, *21*(4), 212–224. https://doi.org/10.1016/j.rccan.2016.10.001

[22] Pardo-Gonzalez, C. A., Lagos-Ibarra, J. J., Linares-Ballesteros, A., Sarmiento-Urbina, I. C., Contreras-Acosta, A. D., Cabrera-Bernal, E. V., Uribe-Botero, G. I., Barros-García, G., & Aponte-Barrios, N. H. (2020). Resultados de la implementación del protocolo PETHEMA LPA 99 en el tratamiento niños con leucemia promielocitica aguda en Bogotá, Colombia. *Revista de La Facultad de Medicina*, *69*(2). https://doi.org/10.15446/revfacmed.v69n2.80152

[23] Rodriguez-Villamizar, L. A., Rojas Díaz, M. P., Acuña Merchán, L. A., Moreno-Corzo, F. E., & Ramírez-Barbosa, P. (2020). Space-time clustering of childhood leukemia in Colombia: A nationwide study. *BMC Cancer*, *20*(1), 1–10. https://doi.org/10.1186/s12885-020-6531-2

[24] García, M., Chicaíza, L. A., Quitián, H., Linares, A., & Ramírez, Ó. (2015). Costo-efectividad de los tratamientos de consolidación para la leucemia mieloide aguda en niños en riesgo alto en el sistema de salud colombiano. *Biomedica*, *35*(4), 549–556. https://doi.org/10.7705/biomedica.v35i4.2563

[25] Onecha, E. M. (2019). *Implicaciones clínicas de la detección de mutaciones recurrentes mediante secuenciación masiva en leucemia mieloide aguda y correlación con la sensibilidad a fármacos antileucemia* [Universidad Complutense de Madrid]. https://eprints.ucm.es/id/eprint/51644/

[26] Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*, *58*(5), 475–483. https://doi.org/10.1016/J.JCLINEPI.2004.06.017

[27] Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, *165*(6), 710–718. https://doi.org/10.1093/AJE/KWK052

[28] Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[29] National Cancer Institute. (n.d.). *The Cancer Genome Atlas Program*. Retrieved September 3, 2021, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga