UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

ANÁLISIS COMPARATIVO DE ALGORITMOS DE CLASIFICACIÓN PARA DIAGNOSTICAR TIPOS DE LEUCEMIA INFANTIL

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

Autor(a) (es):

Bach. Gonzalez Flores Paul Gustavo

ORCID: https://orcid.org/0000-0003-3457-2202

Asesor(a):

Dr. Ramos Moscol Mario Fernando

ORCID: https://orcid.org/0000-0003-3812-7384

Línea de Investigación:

Infraestructura, Tecnología y Medio Ambiente

Pimentel – Perú 2021

APROBACIÓN DEL JURADO

ANÁLISIS COMPARATIVO DE ALGORITMOS DE CLASIFICACIÓN PARA DIAGNOSTICAR TIPOS DE LEUCEMIA INFANTIL

_____

Bach. González Flores Paul Gustavo
Autor

_____

Dr. Ramos Moscol Mario Fernando
Asesor

_____ _____

Dr. Vásquez Leyva Oliver
Presidente de Jurado

Mg. Bravo Ruiz Jaime Arturo
Secretario de Jurado

_____

Mg. Atalaya Urrutia Carlos William
Vocal de Jurado

Dedicatorias

A mi madre Ermilda por su entereza, esfuerzo constante, dedicación, por sostenerme en los momentos más complicados y por ser mi mayor inspiración.

A mis hermanos Anderson y Fiorella por apoyarme y concederme parte de su tiempo para poder lograr este objetivo profesional.

A mi esposa Marilin y mi hija Emilia Sofía que son mi motor y motivo para salir adelante cada día.

Agradecimientos

Expreso mi gratitud de manera especial a mi Mamá Ermilda, mis hermanos Anderson y Fiorella por apoyarme y brindarme sus consejos.

A mi esposa Marilin, por su confianza, aliento y por cada palabra de ánimo.

A mis docentes y asesor, por la orientación prestada en esta investigación y por motivarme a culminar esta etapa exitosamente.

Resumen

Mundialmente el cáncer afecta en gran proporción a la niñez, en el 2020 se reportaron a nivel mundial más de 400 mil casos nuevos de leucemias, siendo el cáncer más frecuente en la niñez. En Perú, el Instituto Nacional de Enfermedades Neoplásicas, realizó un informe estadístico de los pacientes que residen en Lima Metropolitana entre los años 2010 y 2012, donde se reportaron 1604 pacientes diagnosticados con leucemia, representando más del 40% de todas las neoplasias. La influencia del aprendizaje automático en la medicina ha sido muy importante hoy en día, siendo aplicado para diagnosticar diferentes enfermedades, destacando entre ellas el diagnóstico de diferentes tipos de cáncer. La elección de los algoritmos de clasificación a implementar se realizó mediante una revisión de la literatura científica, donde se seleccionaron los algoritmos Regresión logística y Árboles de decisión por haber obtenido mejores resultados de exactitud. Los datos utilizados para desarrollar la presente investigación se obtuvieron del Hospital Regional Docente “Las Mercedes”, siguiendo criterios de inclusión y exclusión se recolectaron 75 datos de pacientes diagnosticados con tipos de leucemia infantil. Posteriormente, se consideró utilizar 60 datos de pacientes que representa el 80% para realizar el entrenamiento y 15 datos de pacientes que representa el 20% para las pruebas. La evaluación del desempeño de los algoritmos de clasificación se realizó mediante la matriz de confusión. Los resultados mostraron que el algoritmo de clasificación Árboles de decisión obtuvo una exactitud de 100%, precisión 100%, especificidad 100%, F1 Score 100% y tiempo de respuesta de 0.02 segundos, mientras que el algoritmo de clasificación Regresión logística obtuvo una exactitud de 93.3%, precisión 92.9%, sensibilidad 100%, F1 Score 96.3% y un tiempo de respuestas de 0.05 segundos. La comparación de los resultados obtenidos mostró que el algoritmo de clasificación Árboles de decisión, es el mejor para diagnosticar los tipos de leucemia infantil, considerando el desempeño obtenido al evaluarse todos los indicadores propuestos en esta investigación.

Palabras Clave: Aprendizaje supervisado, Leucemia infantil, Árboles de decisión, Regresión logística, Algoritmos de clasificación, Predicción, Diagnóstico.

Abstract

Worldwide, cancer affects a large proportion of children; in 2020, more than 400 thousand new cases of leukemia were reported worldwide, being the most frequent cancer in childhood. In Peru, the National Institute of Neoplastic Diseases, conducted a statistical report of patients residing in Metropolitan Lima between 2010 and 2012, where 1604 patients diagnosed with leukemia were reported, representing more than 40% of all neoplasms. The influence of machine learning in medicine has been very important nowadays, being applied to diagnose different diseases, highlighting among them the diagnosis of different types of cancer. The choice of the classification algorithms to be implemented was made through a review of the scientific literature, where the algorithms Logistic Regression and Decision Trees were selected for having obtained better accuracy results. The data used to develop the present research were obtained from the Hospital Regional Docente "Las Mercedes", following inclusion and exclusion criteria, 75 data were collected from patients diagnosed with types of childhood leukemia. Subsequently, it was considered to use 60 patient data representing 80% for training and 15 patient data representing 20% for testing. The evaluation of the performance of the classification algorithms was performed using the confusion matrix. The results showed that the classification algorithm Decision Trees obtained an accuracy of 100%, precision 100%, specificity 100%, F1 Score 100% and response time of 0.02 seconds, while the classification algorithm Logistic Regression obtained an accuracy of 93.3%, precision 92.9%, sensitivity 100%, F1 Score 96.3% and a response time of 0.05 seconds. The comparison of the results obtained showed that the classification algorithm Decision Trees is the best for diagnosing the types of childhood leukemia, considering the performance obtained when evaluating all the indicators proposed in this research.

Keywords: Supervised learning, Childhood leukemia, Decision trees, Logistic Regression, Classification Algorithms, Prediction, Diagnosis.

Índice

I. INTRODUCCIÓN 8

1.1. Realidad Problemática. 8

1.2. Trabajos previos. 10

1.3. Teorías relacionadas al tema. 19

1.4. Formulación del Problema. 32

1.5. Justificación e importancia del estudio. 32

1.6. Hipótesis. 33

1.7. Objetivos. 33

1.7.1. Objetivo general. 33

1.7.2. Objetivos específicos. 33

II. MATERIAL Y MÉTODO 33

2.1. Tipo y Diseño de Investigación. 33

2.2. Población y muestra. 34

2.3. Variables, Operacionalización. 34

2.4. Técnicas e instrumentos de recolección de datos, validez y confiabilidad. 35

2.5. Procedimiento de análisis de datos. 36

2.6. Criterios éticos. 36

2.7. Criterios de Rigor Científico. 37

III. RESULTADOS. 37

3.1. Resultados en Tablas y Figuras. 37

3.2. Discusión de resultados. 42

3.3. Aporte práctico. 44

IV. CONCLUSIONES Y RECOMENDACIONES 68

4.1. Conclusiones. 68

4.2. Recomendaciones. 68

REFERENCIAS. 70

ANEXOS. 75

I. INTRODUCCIÓN

1.1. Realidad Problemática.

Mundialmente el cáncer afecta en gran proporción a la niñez, en el 2020 se reportaron a nivel mundial más de 400 mil casos nuevos de leucemias, siendo el cáncer más frecuente en la niñez. Hoy en día existen diversos mecanismos que permiten mejorar la supervivencia de los pacientes, en un estado temprano de la enfermedad, siendo muy importante que sea detectada de manera correcta para brindar un tratamiento adecuado y lograr la curación de los pacientes. (Organización Mundial de la Salud, 2021)

En Perú, el Instituto Nacional de Enfermedades Neoplásicas (INEN) realizó un informe estadístico de las enfermedades con más alto índice de mortalidad de los pacientes que residen en Lima Metropolitana. Los periodos considerados son entre 2010 y 2012. Se detalla que en los años antes mencionados se reportaron 1604 pacientes diagnosticados con leucemia, donde se precisa que de todas las neoplasias que se diagnóstica en menores de edad, la leucemia representa más del 40% del diagnóstico general. (INEN, 2016)

La leucemia es el cáncer más común en los niños, siendo los tipos de leucemias Linfoblástica Aguda (LLA) y la leucemia mieloide aguda (LMA) las más frecuentes. Es de suma importancia brindar un buen diagnóstico en un estado temprano de la enfermedad para lograr altas tasas de curación. Según reportes del Ministerio de Salud, en el Perú, los pacientes tienen entre un 35% y 45% de probabilidad de tener un tratamiento temprano de la enfermedad. (Ministerio de Salud, 2017)

El aprendizaje automático es un aliado perfecto dentro de las organizaciones e instituciones innovadoras que sacan provecho de sus datos como fuente activa de conocimiento. Se pueden utilizar los datos para ser entrenados y evaluados por distintas técnicas, con la finalidad de evaluar cual técnica puede ser más precisa y eficiente evaluando los datos. (Hurwitz & Kirsch, 2018)

La influencia del aprendizaje automático en la medicina ha sido muy importante hoy en día, siendo aplicado en la predicción y diagnóstico de diferentes enfermedades, destacando entre ellas el diagnóstico de diferentes tipos de cáncer, siendo esta enfermedad una de las mayores causas de muerte en el mundo, donde el mayor interés es la detección en un estado inicial, siendo fundamental para que el paciente logre un tratamiento oportuno y adecuado. (Müller & Guido, 2017)

Álvarez Vega, Quirós Mora, & Cortés Badilla, (2020) precisaron que los temas asociados de medicina se han tornado aún más complejos en los últimos años, contemplando un alto volumen de datos para resolver distintos casos de alta complejidad. Siendo de mucha necesidad contar con una herramienta de apoyo para la decisión médica que excede la capacidad de la mente humana. Utilizar el aprendizaje automático es una alternativa de solución, ya que utilizando distintas técnicas se podrían construir herramientas de apoyo a la decisión médica y de esta manera disminuir el tiempo en el diagnóstico de enfermedades.

Existen varios factores que producen un retardo en la decisión médica para diagnosticar a un paciente con un tipo de leucemia, ante una sospecha de la enfermedad el médico debe en la mayoría de casos tener una amplia experiencia para acertar en un diagnóstico temprano de la enfermedad sin la necesidad de realizar otras pruebas adicionales, considerando que lo antes mencionado solo es una presunción, más no podría ser suficiente para dar un diagnóstico exacto y preciso de la enfermedad. Necesariamente el médico tiene que recurrir a otros tipos de exámenes que permiten brindar información profunda y detallada del paciente. Entre estas pruebas la más común es realizar un hemograma completo al paciente, de esta manera el médico podrá observar los detalles que le apoyaran a tomar una decisión final en el diagnóstico. En algunos casos se necesitará de pruebas más profundas como un aspirado de médula ósea. (Ministerio de Salud, 2017)

Se puede hacer uso de diversos algoritmos de clasificación utilizando el aprendizaje supervisado para diagnosticar el cáncer. En el aprendizaje automático se pueden aplicar algoritmos de clasificación de tipo aprendizaje supervisado. Estos algoritmos tienen una particularidad, utilizando un dataset son entrenados y evaluados, de esa manera se determina su desempeño y se puede elegir el mejor en el diagnóstico de una enfermedad. (Hurwitz & Kirsch, 2018)

1.2. Trabajos previos.

Ara, Das, & Dey, (2021), realizó la investigación, Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms, en Pakistán. Tuvo como problemática, la identificación temprana del tipo de tumor benigno o maligno del cáncer de mama, por lo cual tuvieron que evaluar el rendimiento de los algoritmos de clasificación para clasificar el cáncer maligno y benigno. Por esta razón utilizaron un dataset, que se recopiló del repositorio de UCI, para posteriormente analizar el dataset y evaluar el rendimiento con varios algoritmos de clasificación, con la finalidad de predecir el cáncer de mama. Utilizaron el algoritmo Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Decision Tree, Naive Bayes y Random Forest. Para clasificar el cáncer en benigno y maligno, utilizaron una matriz de confusión, de esta manera verificar el rendimiento y precisión de los algoritmos, posteriormente realizaron una comparación para encontrar el más adecuado. Los resultados obtenidos indicaron al algoritmo Random Forest y Support Vector Machine como los que superan a otros clasificadores con una precisión del 96.5%, mientras que Logistic Regression muestra un 94.4%, KNN obtuvo 95.8%, Decision Tree 95.1% y Naive Bayes 92.3%. Estos clasificadores se pueden utilizar para construir un sistema de diagnóstico automático, para el diagnóstico preliminar del cáncer de mama.

Hsu, Chen, Lin, Jiang, Zhang, Hao & Chung (2021), realizó la investigación, Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning, en Estados Unidos. Tuvieron como problemática,

reducir las características y seleccionar las más óptimas para el sistema propuesto. Por esta razón, propusieron un modelo de funciones que estuvo basado en el aprendizaje automático con el propósito de mejorar el rendimiento predictivo, utilizaron datos del repositorio de Irvine de la Universidad de California, donde accedieron a conjunto de datos de cáncer de mama, cuello uterino y pulmón, los que se utilizaron para el estudio experimental, utilizaron algoritmos de aprendizaje supervisado para realizar el entrenamiento y validación, usaron la validación cruzada de 10 veces evaluando la accuracy, f-score, precisión y recall. El algoritmo GA-CFS propuesto obtuvo una precisión de 99.62%, 96.88%, 98.21% respectivamente. Los métodos computacionales cada vez tienen más protagonismo en el campo de la medicina y pueden brindar soluciones para diversos sistemas complejos.

Patil Babaso, Mishra & Junnarkar (2020), realizaron la investigación, Leukemia Diagnosis Based on Machine Learning Algorithms, en la India. Tuvieron como problemática, diagnosticar la leucemia a partir de imágenes en color de frotis de sangre teñidos, donde tuvieron que procesar y segmentar las imágenes para la extracción de características y clasificar si el paciente esta afectado con leucemia, usaron las técnicas de extracción de características con fines de preprocesamiento. Por esta razón, utilizaron algoritmos de clasificación supervisados y no supervisados con la finalidad de obtener la mejor precisión, para ello utilizaron Support Vector Machines, k-Nearest Neighbour, Neural Networks, Naïve Bayes y Deep Learning. Se obtuvo como resultado de precisión que el algoritmo Support Vector Machines alcanzó el 92%, k-Nearest Neighbour 80%, Neural Networks 93.7%, Naïve Bayes 80.88% y Deep Learning obtuvo el mejor resultado con una precisión del 97.78%. Los algoritmos de Machine Learning están siendo muy utilizados para diversos casos de estudio con la finalidad de mejorar el tratamiento administrado a los pacientes.

El-Shair, Sánchez-Pérez, & Rawashdeh, (2020), realizó la investigación, Comparative Study of Machine Learning Algorithms using a Breast Cancer Dataset, en Estados Unidos. Tuvieron como problemática, evaluar el conjunto de datos utilizado ya que tenía un valor relativamente alto de características,

que puede resultar complicado en algunos modelos, para ello tuvieron que aplicar la selección de características que para estos casos puede resultar vital, dividieron los datos en entrenamiento y prueba. Por esta razón, utilizaron el dataset de diagnóstico de cáncer de mama de Wisconsin, para entrenar y probar los diferentes modelos de clasificación, luego se comparan entre sí, utilizando diferentes métricas de clasificación para identificar los modelos más sólidos y precisos, dividieron el dataset en 80% entrenamiento y 20% para prueba final, el dataset de entrenamiento utilizaron la validación cruzada para garantizar un modelo preciso y utilizaron varias métricas diferentes para analizar el rendimiento de los modelos. Se obtuvo como resultado que el algoritmo Logistic Regression obtuvo en precisión un 97% y Naive Bayes 93%. Para futuros trabajos se pueden implementar más algoritmos de clasificación y hacer uso de más métricas para evaluar su desempeño.

Preethi & Dharmarajan, (2020), realizó la investigación, Diagnosis of chronic disease in a predictive model using machine learning algorithm, en India. Tuvieron como problemática, seleccionar las características críticas para diagnosticar enfermedades crónicas, tuvieron que evaluar el rendimiento en función de varias métricas mediante una matriz de confusión, el criterio de AUC y tiempo en el procesamiento. Por esta razón seleccionaron las características críticas para diagnosticar enfermedades crónicas, utilizando el método de filtro como el coeficiente de correlación, chi-cuadrado fue utilizado en el dataset de la enfermedad renal crónica con base en el método de selección de características que se aplican para clasificar, para la predicción aplicaron la matriz de confusión utilizando diversas métricas como Precisión, Sensibilidad y Especificidad. Los resultados obtenidos mostraron que el algoritmo Support Vector Machine usando la selección chi-cuadrado, ofrece la mayor precisión con 98.3% en enfermedades crónicas de riñón, precisión de 98.7% en diabetes y precisión de 89.9% en enfermedades crónicas del corazón. Las enfermedades crónicas son consideradas de alto riesgo debido al desconocimiento en dicha enfermedad en una etapa temprana, es importante el diagnóstico temprano para tratar al paciente con los medicamentos adecuados, el clasificador SVM tuvo

mejor precisión en la predicción de las enfermedades crónicas en comparación de otros algoritmos.

Arora, Som, & Rana, (2020), realizó la investigación, Predictive Analysis of Machine Learning Algorithms for Breast Cancer Diagnosis, en India. Tuvieron como problemática, Comparar diversos algoritmos de aprendizaje supervisado donde se exigía que aplicaran ciertas bibliotecas de Python como numpy, pandas, matplotlib, csv y math, se usaron para entrenar el conjunto de datos, dividir los datos y luego comparar los resultados con los datos de prueba. Por esta razón utilizaron el dataset de Wisconsin y se dividió sobre la base del 80% al 20%, se utilizaron datos del 80% en los cinco algoritmos, estos fueron Random Forest, Naive Bayes, Support Vector Machine, Decision Tree y K Nearest Neighbor. Mientras que el 20% de las pruebas se utilizaron para comparar el dataset de entrenamiento, se compararon los algoritmos de acuerdo a su precisión y las características implementadas como parámetros. Los resultados mostraron que Random Forest obtuvo en precisión 96.49%, por lo tanto, se desempeñó mejor entre los otros algoritmos utilizados y K Nearest Neighbor fue el segundo algoritmo con mejor desempeño con 95.61% de precisión. También hay diferentes investigaciones donde sobresalen mejor otros algoritmos de aprendizaje supervisado, esto ayudará a analizar mejor estas técnicas.

Sengar, Gaikwad, & Nagdive, (2020), realizó la investigación, Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction, en India. Tuvieron como problemática, seleccionar solo 426 entradas de las 570 entradas del conjunto de datos para entrenar los modelos, también necesitaron los datos restantes para probar la predicción, los algoritmos fueron implementados en Python y utilizaron bibliotecas de aprendizaje automático. Por esta razón utilizo el conjunto de datos de alrededor de 570 entradas de datos y 32 atributos, estos datos fueron entrenados para predecir si el cáncer es maligno o benigno, se seleccionaron solo 550 muestras de entradas de datos para entrenar y se usó el resto para evaluar, el preprocesamiento de datos se realizó con la conversión de los datos, caracteres en datos enteros y la eliminación de datos innecesarios, seleccionaron funciones de definición de correlaciones de características con

funciones y figuras like pair plot y mapa de calor, utilizaron dos modelos propuestos el algoritmo de Logistic Regression y Decision Tree. Los resultados obtenidos indicaron que Decision Tree tiene en precisión 95.10% y Logistic Regression 94.40%. El algoritmo Decision Tree es el más adecuado entre los dos algoritmos comparados para predecir el cáncer de mama, sin embargo, se recomienda evaluar el conjunto de datos con otros algoritmos de aprendizaje automático para determinar cuál es el mejor en la predicción del cáncer de mama.

Sujatha & Mahalakshmi, (2020), realizó la investigación, Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease, en India. Tuvieron como problemática, comparar los algoritmos de clasificación supervisados con un dataset de pacientes diagnosticados con enfermedades cardíacas. Utilizando el dataset, evaluaron los algoritmos de clasificación supervisados de acuerdo a varias métricas. Por esta razón utilizaron el dataset de enfermedades cardíacas de la base de datos kaggle con python 3.7, evaluaron el rendimiento de los algoritmos en función de las métricas Exactitud, Precisión, puntuación F1 y AUC, su dataset tiene 303 registros con 14 atributos, dividiéndolos en datos de entrenamiento y prueba. El 30% de los registros, es decir, 91 registros, se utilizaron para pruebas y los 212 registros restantes para entrenamiento, utilizaron una matriz de confusión con la finalidad de describir que algoritmo de clasificación tiene más rendimiento. Los resultados mostraron que Random Forest obtuvo más precisión en comparación con otros algoritmos de aprendizaje automático supervisados con 88.89%, Support Vector Machine 85.42%, Naive Bayes 86.95%, Logistic Regression 82%, Decision Tree 82.97% y K Nearest Neighbors 75.51%. Los modelos no tuvieron problema para predecir las enfermedades cardíacas con un conjunto de datos pequeño, sin embargo, se pueden utilizar más datos y otros algoritmos para predecir enfermedades cardíacas, con la finalidad de construir un modelo de predicción confiable y preciso.

Mahmood, Shahid, Bakhshi, Riaz, Ghufran & Yaqoob, (2020), realizó la investigación, Identification of significant risks in pediatric acute lymphoblastic

leukemia (ALL) through machine learning (ML) approach, en Pakistán. Tuvo como problemática, determinar la importancia de las variables clínicas y fenotípicas, así como las condiciones ambientales que pueden identificar las causas subyacentes de la Leucemia linfoblástica aguda (LLA) infantil. Por lo cual incluyeron a cincuenta pacientes pediátricos quienes fueron diagnosticados con LLA siguiendo sus criterios de inclusión y exclusión, usaron también variables clínicas compuestas por los resultados de bioquímica sanguínea, donde utilizaron cuatro algoritmos de machine learning supervisados, Árboles de Clasificación, Random Forest, Gradient Boosted Machine, Árbol de Decisión C5.0 y realizaron la validación cruzada. Los resultados obtenidos mostraron que los árboles de clasificación y regresión (CART) proporcionó el mejor y más completo ajuste para el conjunto de datos, obteniendo una precisión de 99.83%, error de 0.17%, el algoritmo Árbol de Decisión C5.0 obtuvo una precisión de 98.6%, Random Forest obtuvo 94.44% y Gradient Boosted Machine obtuvo 95.61%. Los algoritmos utilizados fueron aplicados eficientemente para proporcionar el mejor pronóstico y brindar un mejor tratamiento.

Günaydin, Günay, & Şengel, (2019), realizó la investigación, Comparison of Lung Cancer Detection Algorithms, en Turquía. Tuvieron como problemática, Comparar diversos algoritmos de aprendizaje automático, tanto después del preprocesamiento como sin preprocesamiento de las imágenes, evaluaron el rendimiento en función a varias métricas utilizando una matriz de confusión. Por esta razón utilizaron la base de datos de imágenes digitales estándar de la Sociedad Japonesa de Tecnología Radiológica (JSRT). Las imágenes se etiquetan como ausencia o presencia de un nódulo en el pulmón, etiquetando 154 imágenes como con nódulo y 93 imágenes se etiquetan como sin nódulo, las imágenes se guardan como una matriz de 2048 * 2048 y cada dato es de 2 bytes, el dataset se divide en dos partes, dataset para entrenar y dataset para pruebas de la clasificación, el 70% de los datos se utilizó para entrenar y el 30% para la prueba, eligiendo aleatoriamente del conjunto de datos, los algoritmos utilizados fueron K-Nearest Neighbors, Support Vector Machines, Naive Bayes, Decision Trees y Artificial Neural Networks. Los resultados obtenidos mostraron que el algoritmo Artificial Neural Networks obtuvo mejor resultado con un

82,43% de precisión después del procesamiento de imágenes y Decision Tree proporciona el mejor resultado con un 93,24% de precisión sin procesamiento de imágenes. El método más utilizado para evaluar el rendimiento de modelos derivados en el dataset de aprendizaje automático, es una matriz de confusión, sin embargo, se podrían haber aplicado varios métodos de eliminación de ruido para obtener una mayor precisión.

Chand & Vishwakarma (2019), realizaron la investigación, Leukemia Diagnosis using Computational Intelligence, en la India. Tuvieron como problemática, diagnosticar la leucemia a partir de un preprocesamiento de imágenes, segmentación de imágenes, extracción de características y clasificación. Por esta razón, utilizaron la segmentación de imágenes con el método de eliminación de fondo con extracción de características de las imágenes segmentadas, luego procedieron a clasificar las imágenes usando el algoritmo Support Vector Machine (SVM) de aprendizaje automático y el algoritmo Máquina de aprendizaje extremo (ELM), las características extraídas se guardaron en un archivo de Excel, luego los datos los dividieron en dos conjuntos de entrenamiento y prueba en varias proporciones, para evitar el sobreajuste y el sobre entrenamiento, utilizaron la validación cruzada de k-fold, los datos que utilizaron consistió en 12 características, de las cuales una es la etiqueta de clasificación, las otras once características restantes se utilizaron para entrenar el modelo para la predicción. Los resultados obtenidos mostraron que el algoritmo Máquina de aprendizaje extremo (ELM) obtuvo una precisión de 92.24%, mientras que Support Vector Machine (SVM) alcanzó un 86.36%. La clasificación fue basada en la segmentación de imágenes de frotis de sangre del conjunto de datos ALL-IDB1 que está disponible públicamente.

Das Mou & Kumar Saha, (2019), realizó la investigación, A Comprehensive Study of Machine Learning algorithms for Predicting Leukemia Based on Biomedical Data, en Bangladesh. Tuvieron como problemática, predecir la leucemia aplicando diferentes algoritmos predictivos a los cuales se les asigno trece atributos. Por esta razón, utilizaron el 70% de datos para el entrenamiento y el 30% de datos para la prueba, utilizaron el entorno virtual Anaconda, con

Python 3.7 con varios paquetes de aprendizaje automático, para la evaluación utilizaron la matriz de confusión y calcular la precisión de cada algoritmo, de los cuales utilizaron Decision Tree, KNN, Naive Bayes y SVM, todos los algoritmos pasaron por el método de validación cruzada de diez veces. Los resultados obtenidos muestran que el algoritmo Decision Tree obtuvo una precisión de 100%, KNN 97.5%, Naive Bayes 91.5% y SVM 75%. En dicho estudio se enfocaron principalmente en detectar la leucemia utilizando algunas de las varias técnicas de aprendizaje automático, teniendo en cuenta que se pueden usar técnicas más avanzadas.

Amrane, Oukid, Gagaoua, & Ensarİ, (2018), realizó la investigación, Breast Cancer Classification Using Machine Learning, en Turquía. Tuvieron como problemática, construir el clasificador del cáncer de mama donde necesitaron nueve características que tuvieron que poner cada observación en una categoría a la que pertenece para evaluar la precisión de dos técnicas de aprendizaje automático. Por esta razón, utilizaron el método de clasificación binaria mediante dos algoritmos de aprendizaje automático, Naive Bayes (NB) y knearest neighbor (KNN), el dataset estuvo conformado con nueve características, el fin fue verificar y evaluar los algoritmos donde utilizaron la validación cruzada, dividiendo aleatoriamente ente en particiones de 60% para entrenar datos y 40% para probar los datos. Los resultados obtenidos mostraron que el algoritmo knearest neighbor (KNN) ofrece la mayor precisión (97.51%) y el algoritmo Naive Bayes (NB) obtuvo (96.19%). El objetivo y desafío de la clasificación del cáncer de mama es construir clasificadores que sean precisos y confiables, el algoritmo KNN ofrece la mayor precisión, sin embargo, si el conjunto de datos es más grande, el tiempo de ejecución de KNN aumentará.

Sharma, Aggarwal, & Choudhury, (2018), realizó la investigación, Breast Cancer Detection Using Machine Learning Algorithms, en India. Tuvieron como problemática, comparar diversas técnicas de aprendizaje automático para detectar el cáncer de mama, utilizando un dataset de Wisconsin Diagnosis Breast Cancer. Por esta razón utilizó el dataset que estaba definido con 569 instancias, 32 atributos y no tenía valores perdidos, luego se seleccionaron las

variables más influyentes, para medir el rendimiento se utilizó una matriz de confusión para la clase real, evaluando la Accuracy, Recall, Precisión y F1 Score. Los resultados obtenidos mostraron que el algoritmo Random Forest obtuvo Accuracy 94.74%, Recall 93.65%, Precisión 92.18% y F1 Score 92.90%, knearest neighbor obtuvo Accuracy 95.90%, Recall 90.47%, Precisión 98.27% y F1 Score 94.20%, Naive Bayes obtuvo Accuracy 94.47%, Recall 85.71%, Precisión 88.52% y F1 Score 87.09%. Cada uno de los algoritmos obtuvo una precisión de más del 94% para determinar si el cáncer de mama es benigno o maligno, donde el algoritmo knearest neighbor tuvo los mejores resultados.

Khuriwal & Mishra, (2018), realizó la investigación, Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm, en India. Tuvieron como problemática, seleccionar las características del conjunto de datos, también conocida como selección de atributos para su uso en la construcción de los modelos, aplicando el método de selección de características univariadas para examinar cada característica individualmente y poder determinar la fuerza de relación de la característica con la variable de respuesta. Por esta razón utilizaron un conjunto de datos BCI que tiene 569 filas y 30 columnas de conjunto de datos, para la parte de experimento, primero evaluaron las características del dataset predeterminado, para seleccionar las características, el método de selección de características recursivas y usaron también la validación cruzada, Implementaron los algoritmos de regresión logística y red neuronal individual, luego implementaron el algoritmo Voting Ensemble para calcular la precisión final y para evaluar el desempeño utilizaron una matriz de confusión. Los resultados obtenidos mostraron que el algoritmo Voting Ensemble utilizado para calcular la precisión final obtuvo como resultado una precisión de 98.50%. En esta investigación solo utilizaron 16 características del conjunto de datos para realizar el diagnóstico, se pueden aplicar otras pruebas con otros modelos utilizando otras características para determinar si se puede tener más precisión en el diagnóstico.

1.3. Teorías relacionadas al tema.

1.3.1. Leucemia

Es un cáncer de la sangre, por lo tanto, hace que la sangre se enferme. En la sangre hay tres células principales, eritrocitos, leucocitos y trombocitos. Los eritrocitos se encargan de transportar el oxígeno, los trombocitos se encargan de la coagulación y los leucocitos son las defensas que genera nuestro sistema inmune. La sangre transporta estos tres grupos celulares, de esta manera tenemos oxígeno en nuestros tejidos, si hay una herida se coagula y no se pierde tanta sangre y si hay alguna infección, las células de defensa son las que protegen. (American Cancer Society, 2019)

Estas células están en la médula ósea, hay una célula hemocitoblasto que es muy potencial y precursora de los glóbulos sanguíneos, esta célula da lugar a dos células muy importantes, las células mieloides y las células linfoides, son los dos grandes tipos de células que tenemos. Toda célula linfoide es aquella que genera los glóbulos blancos (leucocitos) muy importantes para el sistema inmune, mientras que las células mieloides van a dar lugar a los glóbulos rojos (eritrocitos) y plaquetas (trombocitos), muy importantes para el sistema inmune. Cuando hay alguna exposición como la radiación solar, algún químico, herbicidas o infecciones bacterianas generan que los cromosomas que están en las células muten. De esta manera se inicia una proliferación exponencial de células inmaduras que desplazan a las células hematopoyéticas normales, es ahí donde la sangre empieza a ser disfuncional causando leucemias del tipo mieloide y del tipo linfoide. (American Cancer Society, 2019)

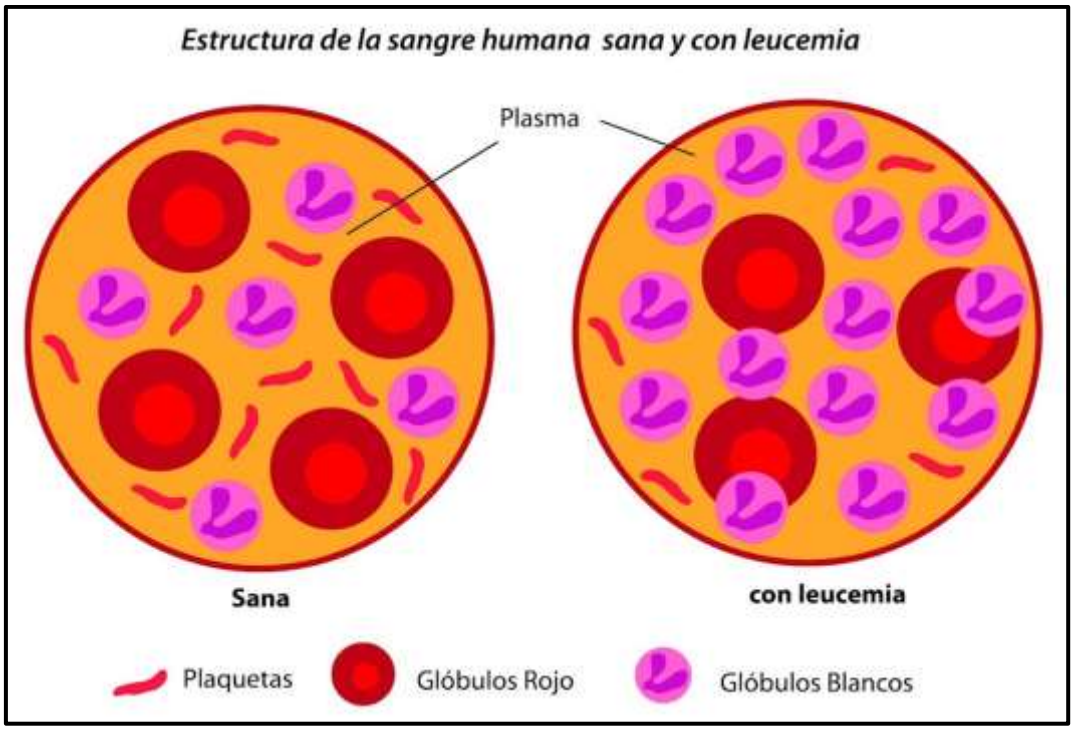

Figura 1. Estructura de la sangre humana sana y con leucemia. Fuente: (AEAL, 2014)

1.3.2. Tipos de leucemia

Existen cuatro principales tipos de leucemia que hay en el mundo, la Leucemia mieloide aguda (LMA), Leucemia mieloide crónica (LMC), Leucemia linfoblástica aguda (LLA) y Leucemia linfocítica crónica (LLC). Las leucemias atacan a todos los grupos de edad, pero hay patologías que afectan más a personas mayores y menores de edad. Los tipos de leucemia linfoblástica aguda y mieloide aguda son las más frecuentes en la infancia. (American Cancer Society, 2019)

1.3.2.1. Leucemia Infantil

Las leucemias LLA y LMA afectan principalmente a los menores de edad. Es muy común por que los genes se están dividiendo y pueden sufrir alteraciones. La leucemia en este grupo se presenta más en pacientes masculinos que femeninos, siendo de mucha importancia investigar porque sucede esto, pero aún no hay estudios concretos que respondan a esta interrogante. Los factores que llevan a contraer leucemia son varios, entre ellos, la exposición a químicos, insecticidas, herbicidas,

benceno y otros. También existen diversos síntomas como la anemia, pérdida de peso, pérdida de apetito, debilidad muscular y otros. Estos signos y síntomas son también comunes en otras patologías, por eso es necesario el apoyo de radiografías, electrocardiogramas, análisis de sangre y otros exámenes que puedan ayudar al médico a dar un diagnóstico presuntivo de la leucemia en los menores de edad. (Ministerio de Salud, 2017)

1.3.2.2. Leucemia Linfoblástica Aguda

Es muy común en el departamento de Pediatría. Consiste en la invasión de glóbulos blancos inmaduros, esta invasión se genera en la médula ósea y llega a alcanzar aproximadamente el 20% de las células totales. Este tipo de leucemia tiene la capacidad de invadir también otros órganos, como el hígado, ganglios linfáticos, bazo y otros. Generalmente el paciente puede presentar anemia, trombocitopenia y neutropenia. (AEAL, 2017)

Existen varios síntomas asociados, como la deficiencia respiratoria, fiebre muy alta, sangrados por efecto que no hay plaquetas en su médula ósea, porque están siendo ocupados por células malignas que son llamados blastos. También el paciente suele presentar anemia por el espacio ocupado por las células malignas. Se diagnóstica primero con el análisis de sangre, para ver alteraciones en el hemograma, donde se analiza Leucocitosis, anemia y trombopenia. (Ministerio de Salud, 2017)

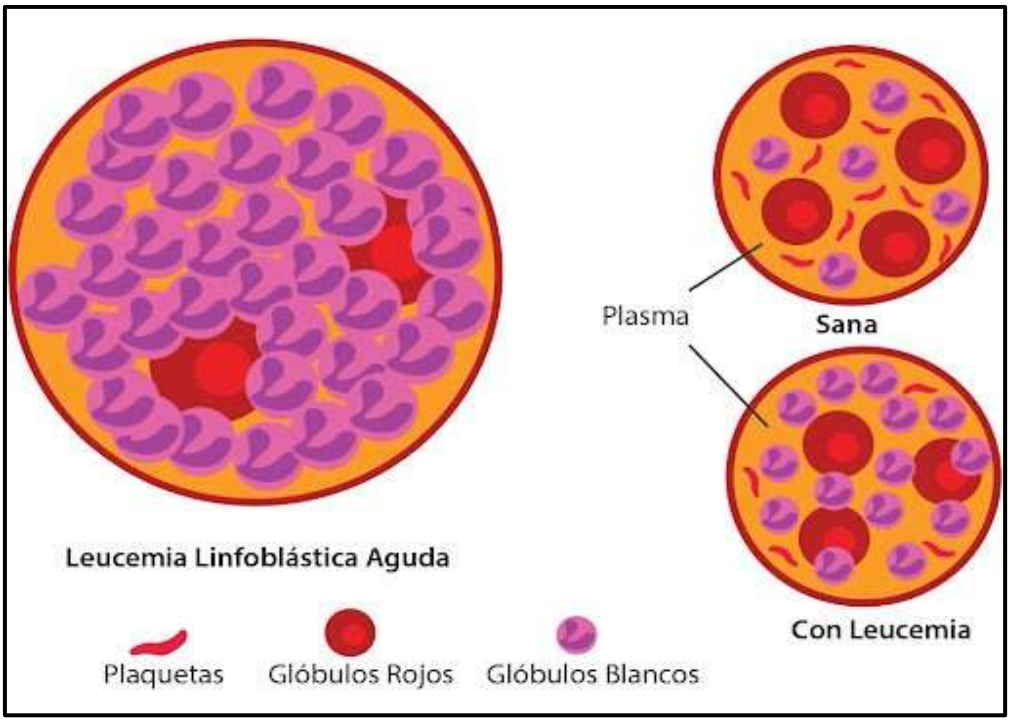

Figura 2. Estructura de sangre humana sana y con LLA. Fuente: (AEAL, 2017)

1.3.2.3. Leucemia Mieloide Aguda

Se manifiesta más en adultos, pero también hay presencia en la infancia. La célula madre inmadura que se produce puede convertirse en mieloide o linfoide. Cuando se producen de manera descontrolada los glóbulos blancos anómalos, llamados blastos es cuando se originan y superan en número a las células normales. No hay una investigación que esclarezca el origen de esta enfermedad. (AEAL, 2014)

Debido a la falta de los glóbulos rojos se produce algunos síntomas en los pacientes. Se encuentra también déficit de plaquetas que producen hematomas que pueden aparecer en cualquier parte del cuerpo. Para diagnosticar la leucemia mieloide aguda se tiene que hacer un análisis de la sangre de la médula ósea, posteriormente se analiza en el microscopio, donde se identifican las células tumorales, los blastos mieloides y con ello se identifica el tipo de leucemia. (American Cancer Society, 2019)

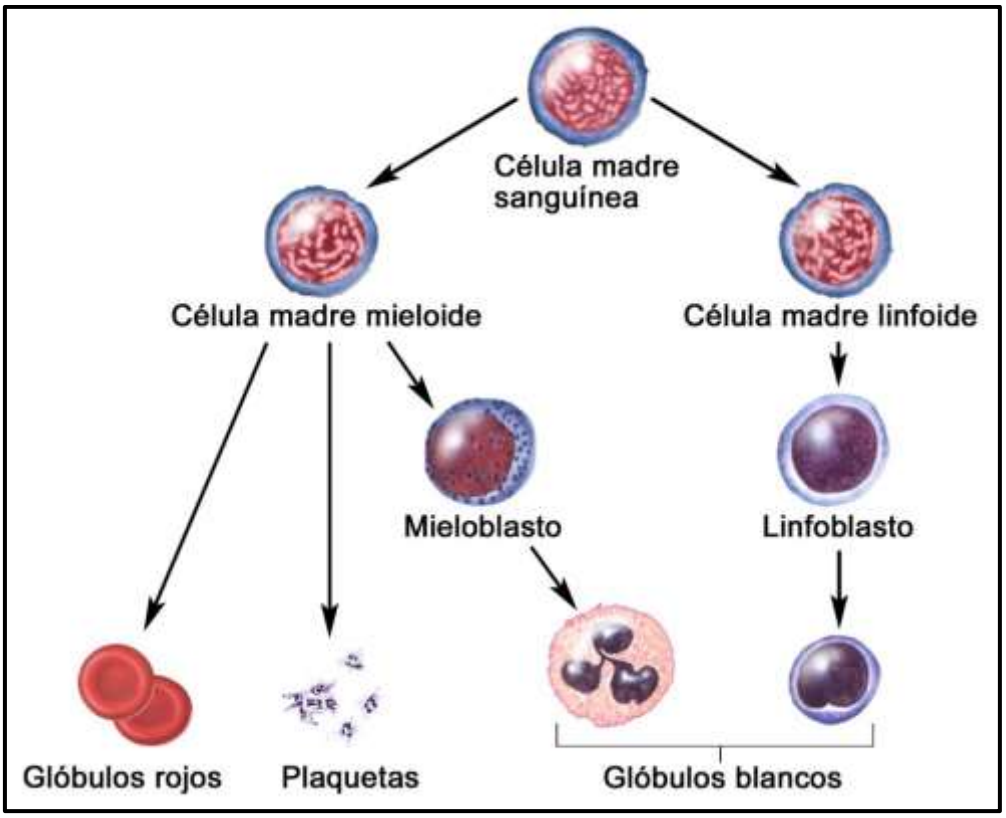

Figura 3. Árbol Hematopoyético de célula mieloide y linfoide. Fuente: (Terese Winslow LLC, 2007)

1.3.2.4. Diagnóstico de Leucemia Infantil

Para diagnosticar la leucemia en la infancia primero se evalúan diferentes síntomas que se presentan y manifiestan de manera severa en algunos casos, de esta manera el médico puede dar un diagnóstico presuntivo de la existencia de la enfermedad. Se considera que la leucemia reemplaza las células sanas por células malignas, si las células que más reemplaza son los glóbulos rojos se puede observar la presencia de anemia, al reemplazar los glóbulos blancos se producen infecciones a repetición y al reemplazar las plaquetas se produce una tendencia a los hematomas, moretones que aparecen en diferentes partes del cuerpo. Para diagnosticar el tipo de leucemia el médico necesita más detalles que solo los síntomas que presenta un paciente, por ello es necesario realizar al paciente un hemograma completo. (American Cancer Society, 2019)

1.3.2.5. Recuento sanguíneo completo

El recuento sanguíneo completo (CBC), se realiza mediante un Hemograma completo, el recuento sanguíneo puede ayudar al médico a determinar alguna enfermedad que este padeciendo una persona. Los recuentos de células en los medios líquidos como sangre, plasma, linfa o enjuague de laboratorio se expresan habitualmente como un número de células por unidad de volumen. Son numerosos los procedimientos en la biología y la medicina que requieren el recuento de células, en la medicina, la concentración de varios glóbulos rojos y glóbulos blancos pueden proporcionar información crucial relacionada con la situación de salud de una persona. (Kaisermann, Pawlowski, & Mendel, 2020)

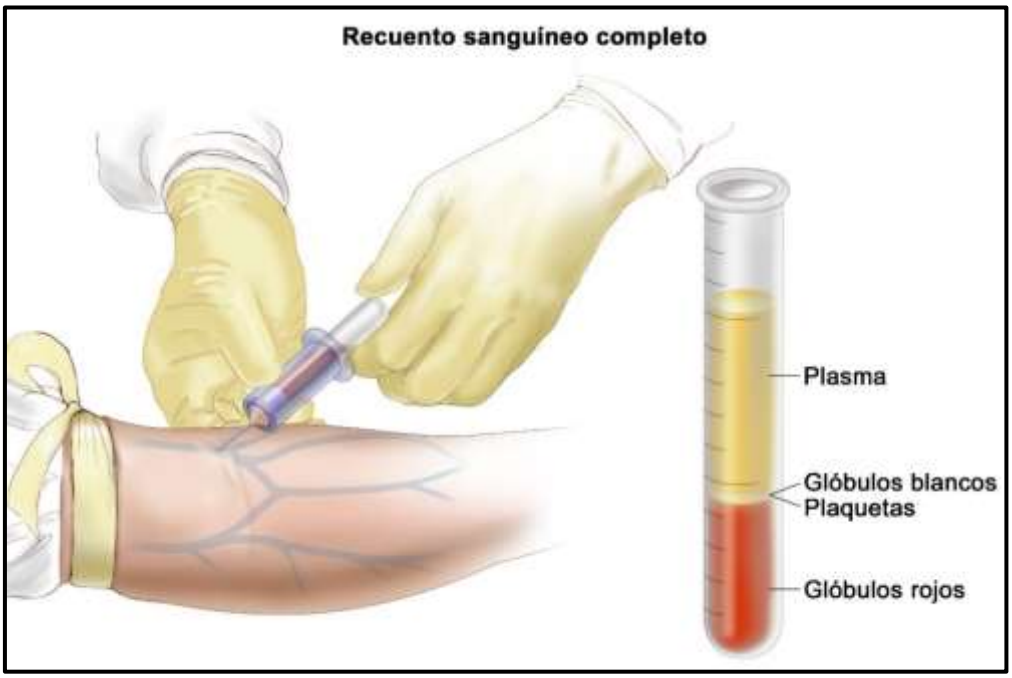

Figura 4. Recuento sanguíneo completo. Fuente: (Terese Winslow LLC, 2007)

1.3.3. Aprendizaje Automático

El Aprendizaje Automático hace que los dispositivos o las aplicaciones sean artificialmente inteligentes. Capaces de realizar cálculos y predecir una enfermedad, sin embargo, con el desarrollo del aprendizaje automático se busca automatizar y agilizar estos procesos. En el pasado las computadoras solo hacían aquello para lo cual habían sido programados, pero con el uso

del aprendizaje automático pueden adquirir conocimiento de acuerdo a experiencias, así como los humanos, de esta manera el aprendizaje automático les permite a los programas aprender e implementar mejoras en sí mismos. (Russell, 2018)

Podemos identificar la presencia del aprendizaje automático cuando las redes sociales nos notifican con publicaciones en base a nuestras preferencias personales o cuando los buscadores de los navegadores web mejoran la exactitud de los resultados de búsqueda. En la medicina también se aplica para predecir el promedio de vida de las personas, organizar la información del paciente e incluso diagnosticar diversas enfermedades. (Hurwitz & Kirsch, 2018)

Los algoritmos de aprendizaje automático se pueden clasificar como aprendizaje supervisado y no supervisado. Es importante conocer la función de cada algoritmo para ser utilizado en el tratamiento de los datos. Los datos constituyen una parte fundamental en el uso de estos algoritmos ya que se utilizan datos para realizar un entrenamiento y datos para realizar una prueba del rendimiento que tiene un algoritmo frente a un determinado problema de clasificación. (Aggarwal, 2014)

1.3.3.1. Aprendizaje Supervisado

Consiste en un tipo de aprendizaje que tiene definida una variable de entrada y salida. Los modelos se constituyen a partir de algoritmos de machine Learning que son entrenados con datos etiquetados para posteriormente ser comparados con datos de prueba. Los algoritmos son entrenados y posteriormente se evalúa su desempeño, Este aprendizaje surge de enseñarle a los algoritmos cual es el resultado que se desea obtener para un determinado valor. El entrenamiento es fundamental ya que si se dan las condiciones puede aprender muy bien, incluso cuando sea entrenado con datos que no haya visto antes. (Bing Liu, 2011)

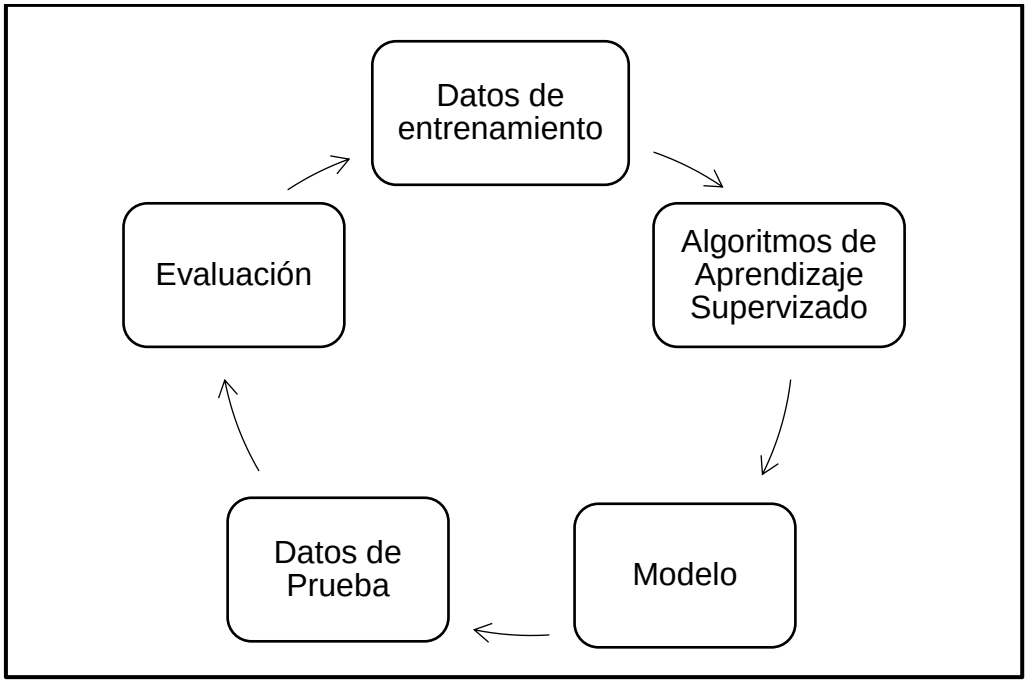

Figura 5. Proceso de Aprendizaje Supervisado. Fuente: (Bing Liu, 2011)

1.3.4. Algoritmos de Clasificación

1.3.4.1. Algoritmo Logistic Regression

Logistic Regression (en español, Regresión Logística) es un algoritmo estadístico utilizado para la clasificación de diversos problemas, se usa para encontrar la relación entre la variable dependiente que es de naturaleza binaria y una o más variables independientes. Su uso es apropiado cuando estamos seguros que la variable dependiente es de tipo binario. La regresión logística binaria se aplica cuando se pretende explicar una característica dicotómica y también se suele usar en un caso más general. Sus variables independientes pueden ser cualitativas o cuantitativas, tanto dicotómicas como politómicas. (Lemeshow, Hosmer Jr, & Sturdivant, 2013)

Este algoritmo estima la probabilidad, su representación matemática es la siguiente:

$$Pr(y = 1) = P$$
$$Pr(y = 0) = 1 - P$$

Donde la variable y representa a dos sucesos o fenómenos excluyentes y exhaustivos, siendo codificados con valores de 0 y 1. Considerando la probabilidad de que suceda uno de ellos es P, la probabilidad de que suceda lo otro es igual a $1 - P$.

La notación matemática del algoritmo de regresión logística es la siguiente:

$$y = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

1.3.4.2. Algoritmo Naive Bayes

Naive Bayes (en español, Bayesiano Ingenuo) es un algoritmo particular cuando la dimensionalidad de las entradas es alta. A pesar de su simplicidad, el clasificador Naive Bayes a menudo puede lograr un rendimiento comparable con algunos métodos de clasificación sofisticados. Los clasificadores Naive Bayes también han exhibido una alta precisión y velocidad cuando se aplica a grandes conjuntos de datos. Si bien es un clasificador bayesiano práctico o simple, se ha convertido en un modelo probabilístico importante y ha tenido un éxito notable en la práctica, siendo usado en el diagnóstico médico, Clasificación de textos, gestión del rendimiento informático, entre otras aplicaciones. (Aggarwal, 2014)

Naive Bayes pertenece también al grupo de clasificadores probabilísticos basados en la aplicación del teorema de bayes. Representa un modelo gráfico para las relaciones de probabilidad. En un conjunto de variables aleatorias, una característica en una clase condicionalmente independiente de cualquier otra característica dada la clase. (Thida, Yamamori, & Ma Ma, 2020)

Su representación matemática es la siguiente:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Donde, $P(c|x)$ es la posibilidad objetivo dado el atributo, $P(x|c)$ es la posibilidad del predictor, $P(c)$ es la posibilidad de la clase, $P(x)$ es la posibilidad previa del predictor.

1.3.4.3. Algoritmo Decision Tree

Decision Tree (en español, Árbol de Decisión) es un algoritmo no paramétrico eficiente que se puede aplicar a tareas de clasificación o regresión. Cuando se usa para la clasificación se denomina árbol de clasificación y cuando se utiliza para tareas de regresión, se denomina árbol de regresión. Se utilizan con frecuencia en campos aplicados como finanzas, ingeniería y medicina. Este algoritmo es útil como técnica exploratoria al igual que existen diversas técnicas para la clasificación, como SVM o las redes neuronales artificiales. (Rokach & Maimon, 2014)

Se usa el algoritmo árbol de decisión en situaciones que se tiene varias alternativas posibles con resultados inciertos. Los árboles de decisión se utilizan específicamente para igualar una táctica óptima y alcanzar una meta. (Boryczka & Kozak, 2010)

La función de evaluación de los árboles de decisión se calcula de acuerdo a la siguiente fórmula:

$$Q(T) = \emptyset . w(T) + \psi . a(T, P)$$

Donde $w(T)$ es el tamaño (número de nodos) del árbol de decisión T. $a(T, P)$ es la precisión del objeto de clasificación de un conjunto de prueba P por el árbol T. ϕ y ψ son las constantes que determinan la importancia relativa de $w(T)$ y $a(T, P)$.

1.3.4.4. Algoritmo K-Nearest Neighbors

K-Nearest Neighbors (en español, K-Vecinos más cercanos) es un algoritmo utilizado para la clasificación. Se logra la clasificación identificando los vecinos más cercanos a un ejemplo de consulta y utilizando esos vecinos para determinar la clase de la consulta. Debido a que se basa en instancias este algoritmo no requiere entrenamiento antes de realizar ediciones, el aprendizaje incremental se puede adoptar fácilmente. Por estas razones, K-Vecinos más cercanos se ha aplicado activamente en una amplia variedad de tareas de aprendizaje supervisado. (Seokho, 2021)

Este algoritmo calcula la distancia de cada ejemplo a clasificar con todos los ejemplos del conjunto de entrenamiento y clasifica al ejemplo de acuerdo a la clase de los ejemplos más cercanos. Para calcular la distancia se utilizan principalmente tres métodos, la distancia euclídea, la distancia manhattan y la distancia Minkowski. (Kramer, 2013)

Formula de la distancia euclídea:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Formula de la distancia manhattan:

$$D(x,y) = \sum_{i=1}^{k}|x_i - y_i|$$

Formula de la distancia Minkowski:

$$D(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

1.3.4.5. Algoritmo Support Vector Machine

Support Vector Machine (en español, Máquinas de vectores de soporte) es un algoritmo que ofrece una alta precisión en comparación con otras técnicas de aprendizaje automático. Varios investigadores han demostrado que este algoritmo es quizás el más preciso para la clasificación de texto, páginas web y diversas enfermedades que requieren un diagnóstico. Se le considera un clasificador discriminatorio definido por un hiperplano de separación. La idea básica es que puede encontrar el mejor hiperplano(s) para separar los datos de dos clases diferentes, de modo que la distancia entre las dos clases, es decir, el margen, se maximice. Según las características de los datos, este algoritmo puede realizar clasificaciones lineales o no lineales. (Chandra, Khemchandani, & Jayadeva, 2017)

Se usa el algoritmo Support Vector Machine para clasificar dos clases. A lo que se le conoce como la clasificación binaria, que consiste en entrenamiento y prueba. En la etapa de entrenamiento se extrae cada atributo y se procede a entrenar el modelo. (Murty & Raghava, 2016)

El Support Vector Machine usa el mejor hiperplano de separación, donde la posición se calcula entre varios hiperplanos y su representación matemática es la siguiente:

$$\{\vec{x}.\vec{y} + b = 0\}, \vec{w} \in \Re^n, \vec{x} \in \Re^n, b \in \Re$$

Cuya función de decisión concierne a:

$$f(\vec{x}) = sign(\vec{w}.\vec{x} + b)$$

1.3.4.6. Algoritmo Random Forest

Random Forest (en español, Bosque Aleatorio) es un algoritmo que se utiliza para categorizar algo en función de otros datos que tiene. El nombre bosque aleatorio proviene de la combinación de la aleatoriedad

que se usa para elegir el subconjunto de datos. Es posible utilizar este algoritmo para clasificar una enfermedad en función de los síntomas de una persona, además puede aplicarse a una amplia gama de problemas y ser bastante bueno en todos ellos. (Hartshorn, 2016)

Este algoritmo necesita datos etiquetados para aprender durante su entrenamiento, donde combina la simplicidad de los árboles de decisión con flexibilidad y agrega aleatoriedad a los datos objetivos. Para el entrenamiento utiliza datasets distintos y cada árbol genera un modelo diferente. (Rokach & Maimon, 2014)

En función de datos de clasificación Random Forest utiliza la formula del índice de Gini:

$$Gini = 1 - \sum_{i=1}^{C} (P_i)^2$$

Donde Pi es la frecuencia relativa de la clase del dataset y C es el número de clases. Se aplica la probabilidad para comprobar Gini de cada rama en un nodo.

1.3.5. Lenguajes de Programación

1.3.5.1. Python

Fue creado por Guido van Rossum. Permite crear aplicaciones para computadoras, aplicaciones web y se usa también en la ciencia de datos, siendo el lenguaje de programación más popular en este campo. Este compuesto por una sintaxis fácil de aprender permitiendo a personas sin experiencia en programación acelerar su aprendizaje a diferencia de otros lenguajes de programación. (Fernández Montoro, 2013)

Es también un lenguaje de programación de código abierto, cuenta con una importante cantidad masiva de desarrolladores que trabajan constantemente en agregar nuevas cosas, debido a ello se ha vuelto

cada vez más poderoso. Actualmente existen diversos sitios web muy famosos por excelencia y conocidos por muchos que están construidos propiamente en Python. Permite el manejo de varios procesadores haciendo mucho más fácil manejar procesadores a diferencia de otros lenguajes de programación. Haciendo uso de diversas librerías se puede dar diversas soluciones, debido al gran abanico de posibilidades es considerado el lenguaje de programación idóneo para aprender ciencia de datos. (González Duque, 2017)

1.3.5.2. R

Es un lenguaje pensado en la computación estadística que fue creado por Ross Ihaka y Robert Gentleman. Este lenguaje permite la creación de gráficos, permitiendo utilizar varios paquetes y librerías. Es de código abierto con una comunidad muy grande que facilita mucho la implementación del código ya que se cuenta con mucha información, paquetes y librerías que ya traen implementadas funciones. (García Montero, 2015)

La facilidad de uso es una de las principales motivaciones para utilizar el lenguaje R, tiene buena documentación contando con manuales que ayudan mucho en la programación. En cada versión de R aumenta el número se posibilidades para utilizar paquetes. Este lenguaje es excelente en el análisis estadístico, gráficos y reportes. Es utilizado también por los investigadores científicos en el área de ciencia de datos.

1.4. Formulación del Problema.

¿Cuál es el mejor algoritmo de clasificación para diagnosticar tipos de leucemia infantil?

1.5. Justificación e importancia del estudio.

Diagnosticar los tipos de leucemia infantil en una etapa temprana, es de suma importancia para brindar el tratamiento adecuado y oportuno al paciente, de esta manera tendrá más altas posibilidades de vencer a este cáncer.

En Perú se diagnostica cáncer a 1300 niños aproximadamente cada año, uno de cada tres pacientes de cáncer infantil es diagnosticado con leucemia, una enfermedad que suele identificarse en un estado muy avanzado, esto surge debido a los diversos análisis y pruebas que necesita un médico para determinar la enfermedad. (INEN, 2016)

La investigación con su propuesta permitirá expandir el conocimiento acerca del uso de diferentes algoritmos de clasificación de aprendizaje supervisado. Siendo importante su uso con el análisis de datos, de esta manera se puede utilizar la tecnología para proponer herramientas confiables en el diagnóstico de los tipos de leucemia infantil.

1.6. Hipótesis.

El algoritmo de clasificación Árbol de decisión tendrá los mejores resultados para diagnosticar los tipos de leucemia infantil.

1.7. Objetivos.

1.7.1. Objetivo general.

Comparar los algoritmos de clasificación para diagnosticar tipos de leucemia infantil.

1.7.2. Objetivos específicos.

a) Seleccionar los algoritmos de clasificación.
b) Estructurar el tratamiento del conjunto de datos.
c) Implementar los algoritmos de clasificación.
d) Evaluar el desempeño de los algoritmos de clasificación.

II. MATERIAL Y MÉTODO

2.1. Tipo y Diseño de Investigación

2.1.1. Tipo de Investigación

Derivada del conocimiento científico es una investigación cuantitativa, empieza con una idea, posteriormente se hace una revisión de artículos científicos, con la finalidad de tener un panorama amplio sobre la

investigación. Luego se plantea una pregunta donde se muestre el problema de investigación y su hipótesis. De esta manera se busca predecir y explicar un fenómeno en la cual se miden variables y para analizar estas medidas se pueden usar varios modelos estadísticos y matemáticos.

2.1.2. Diseño de investigación

Es de tipo cuasi experimental, permitiendo analizar el efecto que hay entre la variable independiente sobre las variables dependientes, siendo un diseño netamente descriptivo donde los datos generan resultados esperados. Este diseño se aplica en procesos y métodos para comprobar el grado de certeza del estudio.

2.2. Población y muestra.

2.2.1. Población

Fue establecida por seis algoritmos de clasificación, los cuales obtuvieron mejor exactitud en el diagnóstico de diferentes tipos de cáncer, los algoritmos seleccionados son: Regresión logística, Árbol de decisión, K-vecinos más cercanos, Random forest, Máquinas de vectores de soporte y Naive bayes.

2.2.2. Muestra

Se seleccionó por conveniencia a los dos algoritmos con mejor porcentaje de exactitud, estos son: Regresión logística y Árbol de decisión.

2.3. Variables, Operacionalización.

2.3.1. Variable independiente

Algoritmos de clasificación

2.3.2. Variable dependiente

Diagnóstico de los tipos de leucemia infantil

2.3.3. Operacionalización de variables

Tabla 1.

Operacionalización de las variables

Variables	Dimensión	Indicador	Ítem	Técnica e instrumentos de recolección de datos
Algoritmos de clasificación	Tiempo	Tiempo de respuesta de los algoritmos	$T = TF - TI$	Registro electrónico Entrevista Observación
Diagnóstico de los tipos de leucemia infantil	Matriz de Confusión	Exactitud	$\frac{VP + VN}{VP + VN + FP + FN}$	
		Precisión	$\frac{VP}{VP + FP}$	
		Sensibilidad	$\frac{VP}{VP + FN}$	
		F1 Score	$2 * \frac{Precisión * Recall}{Precisión + Recall}$	

Fuente: Elaboración propia

2.4. Técnicas e instrumentos de recolección de datos, validez y confiabilidad.

2.4.1. Observación

Está basada en los hechos o los casos que lo involucran, con el fin de registrar todo lo encontrado para un posterior análisis. Este procedo debe incluir un objetivo claro, preciso y definido.

2.4.2. Documentación

Consiste en recolectar información de distintas fuentes, entre ellas, libros, revistas científicas, ensayos y sitios web de investigación confiables, esto permitirá describir y explicar el fenómeno en estudio con base científica.

2.4.3. Entrevista

Consiste en interactuar de manera asertiva con el entrevistado, realizando consultas sobre un tema en específico, con la finalidad de obtener respuestas claras y precisas sobre un tema. Siendo muy importante para concretar ideas de manera eficaz.

2.5. Procedimiento de análisis de datos.

Para evaluar el desempeño de los modelos se utilizó la matriz de confusión de acuerdo a los indicadores propuestos.

2.6. Criterios éticos.

2.6.1. Confidencialidad

Los datos utilizados en la presente investigación se obtuvieron del Hospital Regional Docente “Las Mercedes” en el departamento de Pediatría, previa autorización para recolectar los datos. Obteniendo datos reales por lo que es importante establecer el criterio de confidencialidad y privacidad de la información, de esta manera dicha información no es divulgada a terceros.

2.6.2. Privacidad

La privacidad es un derecho fundamental de las personas a quedarse con la información para uno mismo, sin que otras personas tengan que decidir cuándo acceder a su información privada. Es importante que la persona solo brinde acceso a su información privada con previo consentimiento de la misma.

2.7. Criterios de Rigor Científico.

2.7.1. Fiabilidad

Es fiable cuando la información presenta seguridad y veracidad, siendo aplicada para recolectar los datos.

2.7.2. Validez

Se aplicó el criterio de validez en la presente investigación, donde la información fue analizada con apoyo de Hematólogos, siendo ellos los profesionales expertos en el diagnóstico de enfermedades de la sangre, de esta manera se pudo obtener la información estableciendo criterios de inclusión y exclusión.

III. RESULTADOS.

3.1. Resultados en Tablas y Figuras.

Los resultados fueron generados con la comparación de los dos algoritmos de clasificación, Regresión logística y Árboles de decisión que fueron implementados para diagnosticar los tipos de leucemia infantil.

En primer lugar, se comparó la Exactitud de los modelos, siendo el algoritmo de Árboles de decisión el que obtuvo el mejor desempeño con un 100.0%, demostrando ser el más exacto para diagnosticar los tipos de leucemia infantil, el algoritmo de Regresión logística tiene ligeramente un menor desempeño, obteniendo un 93.3% de exactitud. A continuación, se muestra la fórmula utilizada y la comparación gráfica de los resultados obtenidos.

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

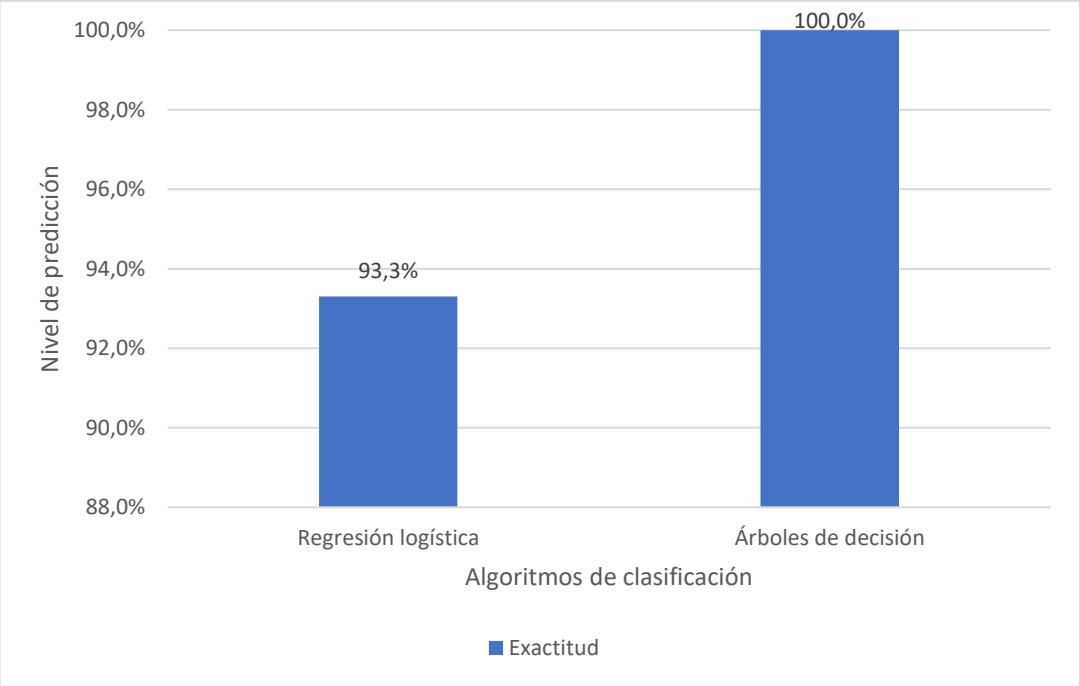

Figura 6. Comparación de los resultados de Exactitud de los modelos. Fuente: Elaboración propia.

La Figura 6 mostró que el algoritmo Árboles de decisión es el más exacto para diagnosticar los tipos de leucemia infantil, clasificando correctamente los pacientes diagnosticados con LLA y LMA respectivamente, siendo fundamental para brindar un correcto tratamiento al paciente.

En segundo lugar, se comparó la Precisión de los modelos, siendo el algoritmo de Árboles de decisión el que obtuvo el mejor desempeño con un 100.0%, demostrando ser el más preciso para diagnosticar los tipos de leucemia infantil, el algoritmo de Regresión logística tiene ligeramente un menor desempeño, obteniendo un 92.9% de Precisión. A continuación, se muestra la fórmula utilizada y la comparación gráfica de los resultados obtenidos.

$$Precisión = \frac{VP}{VP + FP}$$

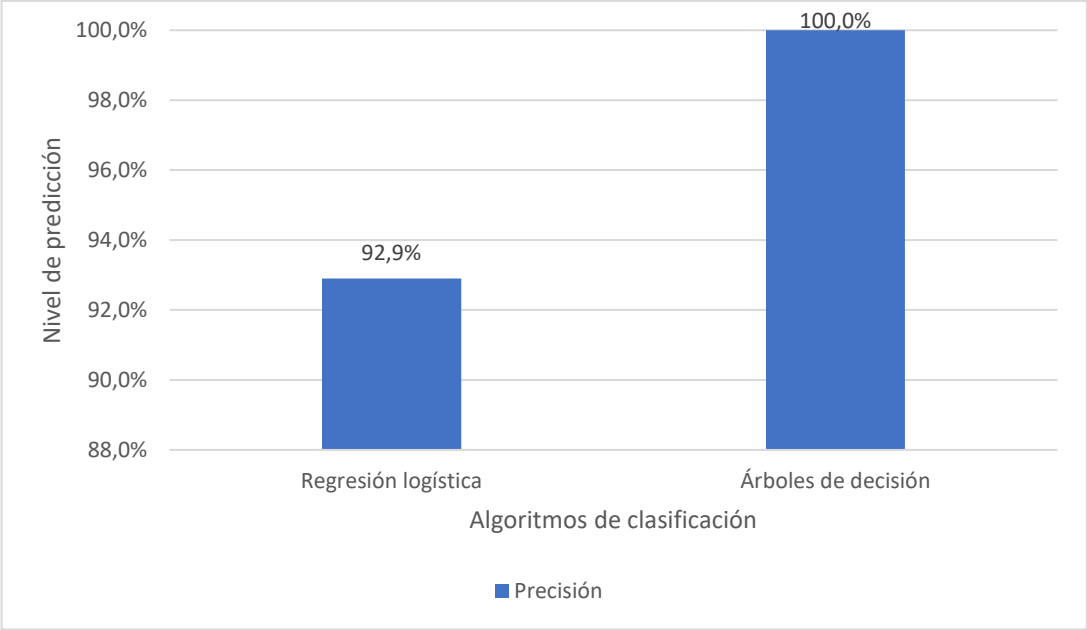

Figura 7. Comparación de los resultados de Precisión de los modelos. Fuente: Elaboración propia.

La Figura 7 mostró que el algoritmo Árboles de decisión es más Preciso que el algoritmo de Regresión logística para predecir los tipos de leucemia infantil, clasificando correctamente a los pacientes de acuerdo al tipo de leucemia que padecen, de esta manera el modelo no se equivoca al predecir el tipo de leucemia de un paciente.

En tercer lugar, se comparó la Sensibilidad de los modelos, obteniendo ambos algoritmos Árboles de decisión y Regresión logística el 100.0% de sensibilidad en el diagnóstico de los tipos de leucemia infantil. A continuación, se muestra la fórmula utilizada y la comparación gráfica de los resultados obtenidos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

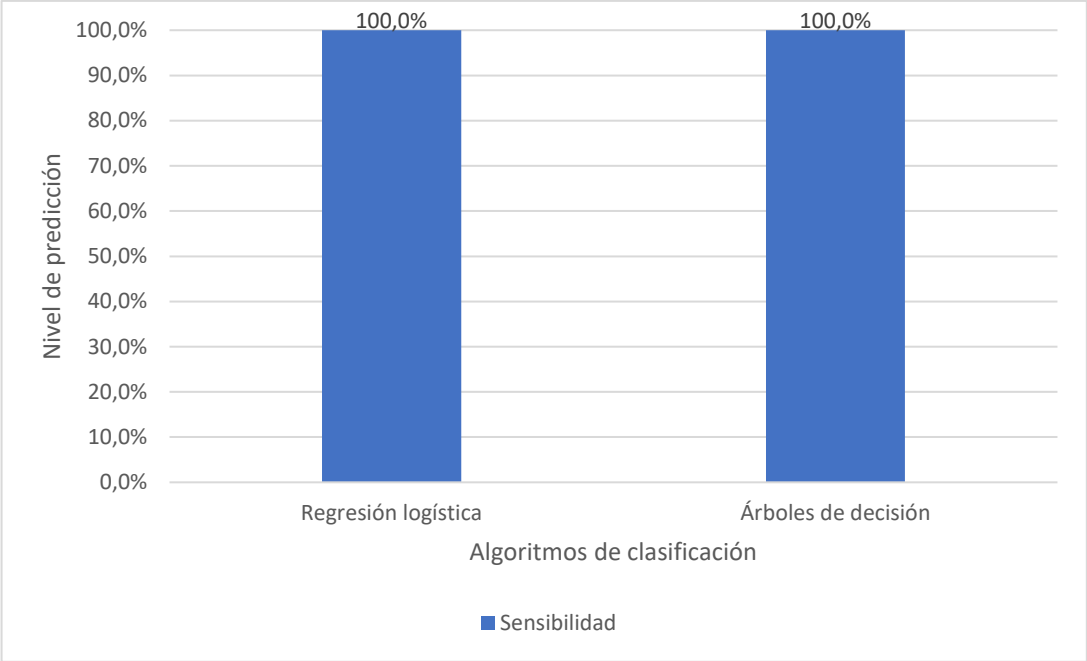

Figura 8. Comparación de los resultados de Sensibilidad de los modelos. Fuente: Elaboración propia.

La Figura 8 mostró que los algoritmos Árboles de decisión y Regresión logística tienen una sensibilidad del 100.0%, donde se tiene como prioridad identificar correctamente a los pacientes de acuerdo al tipo de leucemia que padecen, para poder brindarles un tratamiento correcto.

En cuarto lugar, se comparó el F1 Score de los modelos, siendo el algoritmo de Árboles de decisión el que obtuvo el mejor desempeño con un 100.0%, demostrando ser el que mejor puntuación tiene en el diagnóstico de los tipos de leucemia infantil, el algoritmo de Regresión logística obtuvo ligeramente una menor puntuación, obteniendo un 96.3%. A continuación, se muestra la fórmula utilizada y la comparación gráfica de los resultados obtenidos.

$$F1\ Score = 2 * \frac{Precisión * Recall}{Precisión + Recall}$$

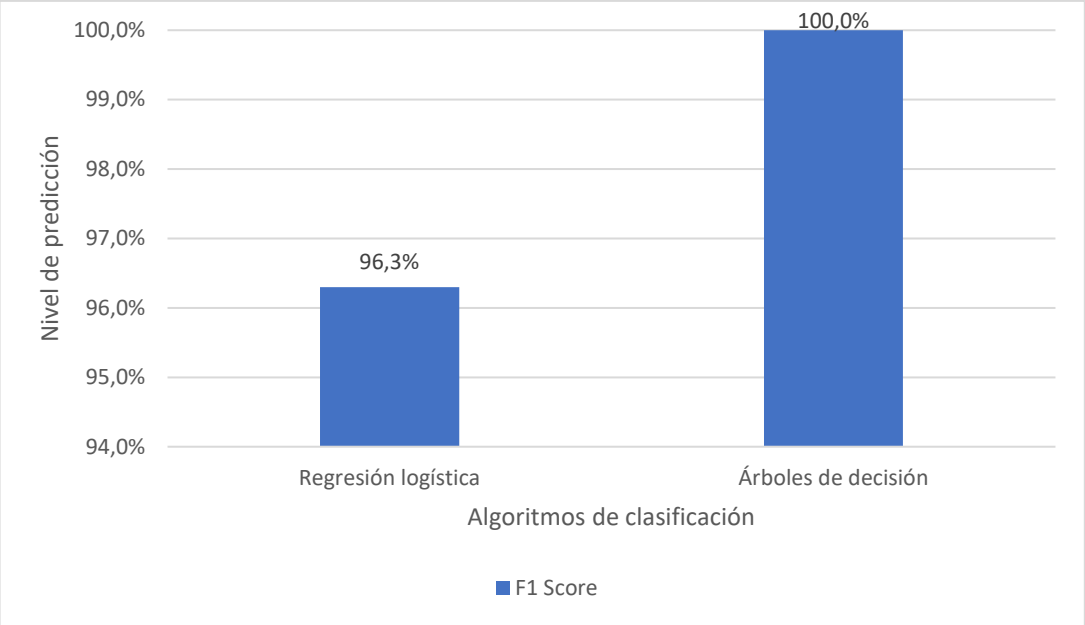

Figura 9. Comparación de los resultados de F1 Score de los modelos. Fuente: Elaboración propia.

La Figura 9 mostró que el algoritmo Árboles de decisión tiene la mejor puntuación de F1 Score con 100.0%, demostrando tener mejor rendimiento en el diagnóstico de los tipos de leucemia infantil.

Finalmente, se comparó el Tiempo de respuesta de los algoritmos, donde se observa que el algoritmo Árboles de decisión obtuvo el mejor desempeño en tiempo de respuesta con 0.02 segundos y el algoritmo de Regresión logística obtuvo ligeramente un mayor tiempo de respuesta con 0.05 segundos en el diagnóstico de los tipos de leucemia infantil. A continuación, se muestra la fórmula utilizada y la comparación gráfica de los resultados obtenidos.

$$Tiempo\ de\ respuesta\ de\ los\ algoritmos = TF - TI$$

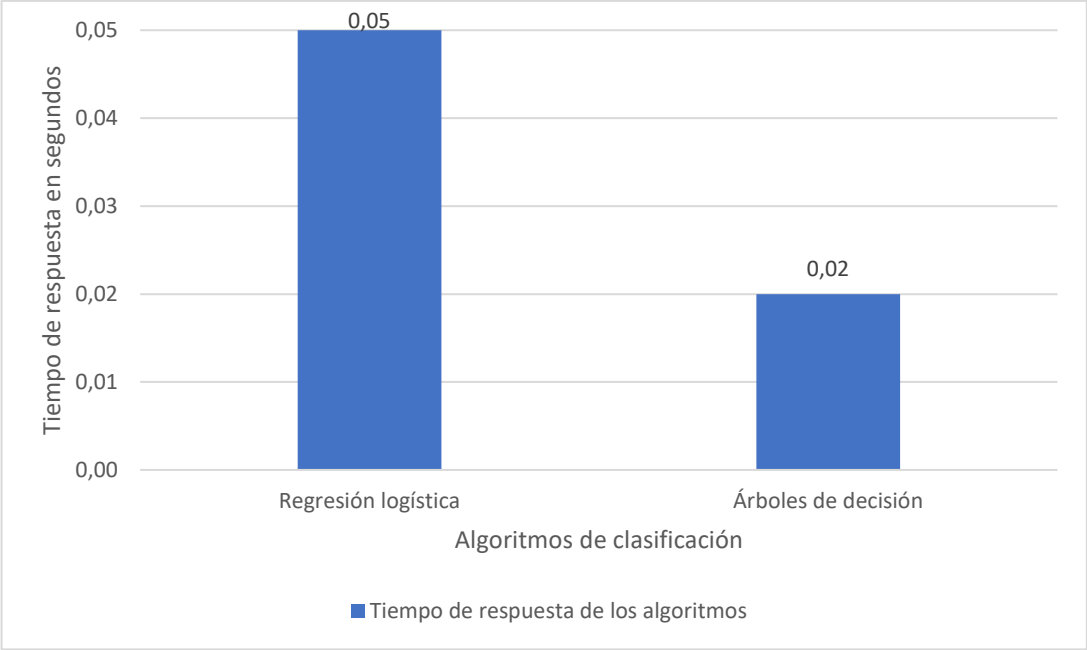

Figura 10. Comparación de los resultados del tiempo de respuesta de los modelos. Fuente: Elaboración propia.

La Figura 10 mostró que el algoritmo Árboles de decisión tiene el mejor tiempo de respuesta con 0.02 segundos, siendo de importancia que el modelo a utilizar, brinde un diagnóstico del tipo de leucemia, en el menor tiempo posible, para brindar una atención rápida y oportuna a los pacientes.

3.2. Discusión de resultados.

El algoritmo de clasificación Árboles de decisión es el mejor para diagnosticar los tipos de leucemia infantil de acuerdo a los resultados, obteniendo el mejor desempeño en todos los indicadores, mientras que el algoritmo de clasificación Regresión logística obtuvo ligeramente un menor desempeño. La investigación realizada por Roy, Pal, Das, & Huq, (2020), muestra que utilizó el algoritmo de Regresión logística en el diagnóstico del cáncer de mama, donde obtuvo una exactitud del 99.0%, siendo superior al resultado obtenido en esta investigación, donde se obtuvo un 93.3% de exactitud, considerando que hay diversas diferencias respecto a los datos utilizados. Los investigadores antes

mencionados utilizaron dos dataset para entrenar y probar sus modelos, donde obtuvieron entre ambos un total de 1268 datos crudos, los cuales fueron divididos en 80.0% para el entrenamiento y el 20.0% para realizar las pruebas. Por contrario en esta investigación se contó con un dataset de 75 datos de pacientes, de los cuales 60 datos que representan el 80% se usó para entrenar y 15 datos que representan el 20% para la prueba, considerando que estas diferencias pueden influir al momento de obtener los resultados, ya que mientras más datos se brinden para entrenar y evaluar el modelo, mejor es la predicción.

En algoritmo que obtuvo la mejor exactitud para diagnosticar los tipos de leucemia infantil, fue el algoritmo de Árboles de decisión que obtuvo un 100.0%, siendo superior al resultado obtenido por Setiawan, Rustam, Hartini, Laeli, & Wirasati, (2020), quienes obtuvieron una Exactitud de 98.8.%, utilizando el algoritmo Árboles de decisión en el diagnóstico del cáncer de carcinoma hepatocelular (CHC), donde utilizaron un dataset con 90 datos, de los cuales destinaron el 90.0% de datos para entrenamiento y el 10.0% de datos para realizar las pruebas. por contrario en esta investigación se utilizó un dataset con 75 datos de pacientes diagnosticados con tipos de leucemia infantil, que son LLA y LMA, dividiendo el dataset en 80.0% de datos para entrenar y el 20.0% para la prueba.

Finalmente, los indicadores evaluados mostraron que el algoritmo Árboles de decisión, obtuvo un 100.0% de exactitud y un tiempo de respuesta con 0.02 segundos, siendo el mejor y más rápido para diagnosticar los tipos de leucemia infantil, sin embargo, el algoritmo de Regresión logística ligeramente obtuvo un resultado menor con una Exactitud del 93.3% y un tiempo de respuesta de 0.05 segundos.

3.3. Aporte práctico.

En la presente investigación se realizó un análisis comparativo de algoritmos de clasificación para diagnosticar tipos de leucemia infantil siguiendo la metodología propuesta en la Figura 11.

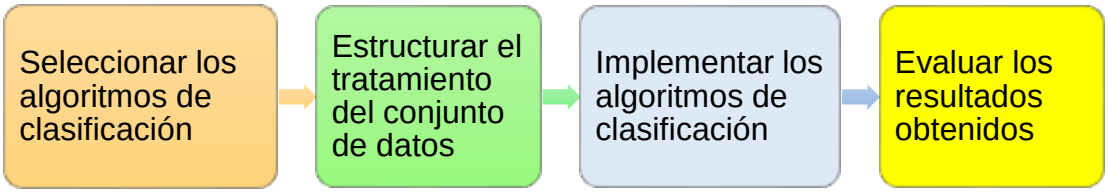

Figura 11. Metodología propuesta. Fuente: Elaboración propia.

Seleccionar los algoritmos de clasificación.

Para seleccionar los algoritmos de clasificación se realizó un proceso de revisión de artículos científicos en las bases de datos científicas IEEE Xplore, Scopus y ScienceDirect. Primero se planteó como objetivo de búsqueda Identificar los algoritmos de clasificación utilizados en el diagnóstico de diferentes tipos de cáncer, donde se establecieron palabras clave como, “Machine Learning” “Classification Algorithms”, “Cancer Diagnosis” y “Supervised learning”, obteniendo la siguiente cadena de búsqueda ("Machine Learning") AND ("Classification Algorithms") AND ("Cancer Diagnosis") AND ("Supervised learning").

La base de datos científica Scopus encontró 241 publicaciones, IEEE Xplore mostró 69 publicaciones y ScienceDirect mostró 137 publicaciones, todas estas publicaciones encontradas en las tres bases de datos científicas guardaban relación con el objetivo de búsqueda que se planteó. Luego se establecieron criterios de búsqueda de los artículos científicos, tales como, considerar solo artículos científicos publicados en los años (2016 - 2021), que se hayan utilizado en las investigaciones solo algoritmos de clasificación para diagnosticar diferentes tipos de cáncer y que los algoritmos de clasificación utilizados en las investigaciones sean de aprendizaje supervisado. Obteniendo de la base de

datos científica Scopus un total de 40 artículos, IEEE Xplore mostró 13 artículos y ScienceDirect mostró 16 artículos. Los artículos científicos fueron analizados aplicando criterios de calidad, descartando aquellos que sean duplicados, no sustenten mediciones, no hayan usado algoritmos de clasificación y no hayan usado aprendizaje supervisado. Obteniendo como resultado un total de 15 investigaciones, las mismas que ayudaron a obtener la población de los algoritmos de clasificación, como se muestra en la Tabla 2.

Tabla 2.

Investigaciones seleccionadas de las bases de datos científicas

Investigadores y año de publicación	**Escenario donde fue utilizado**	**Algoritmos de clasificación**	**Métricas utilizadas**	**Resultados obtenidos**
(Ara, Das, & Dey, 2021)	Clasificación del cáncer de mama maligno y benigno mediante algoritmos de aprendizaje automático	Regresión logística, K-vecinos más cercanos, Árbol de decisión, Naive bayes, Random forest y Máquinas de vectores de soporte	Exactitud	LR: 94.4% KNN: 95.8% DT: 95.1% NB: 92.3% RF: 96.5% SVM: 96.5%
(Naji, El Filali, Aarika, Abdelouhahid, & Debauche, 2021)	Algoritmos de aprendizaje automático para la predicción y el diagnóstico del cáncer de mama	Máquinas de vectores de soporte, Random forest, Regresión logística, Árbol de decisión y K-vecinos más cercanos	Precisión, Sensibilidad, F1-Score	SVM: 98%, 94%, 96% RF: 96%, 94%, 95% LR: 98%, 91%, 94% DT: 94%, 92%, 93% KNN: 92%, 91%, 91%
(Arora, Som, & Rana, 2020)	Análisis predictivo de algoritmos de aprendizaje automático para el	Máquinas de vectores de soporte, Random forest,	Exactitud	SVM: 94.7% RF: 96.4%

	diagnóstico del cáncer de mama	Naive bayes, K-vecinos más cercanos y Árbol de decisión		NB: 92.1% KNN: 95.6% DT: 93.8%
(El-Shair, Sánchez-Pérez, & Rawashdeh, 2020)	Estudio comparativo de algoritmos de aprendizaje automático utilizando un conjunto de datos de cáncer de mama	Naive bayes, Regresión logística	Exactitud, Precisión, Recall, F1-Score	NB: 91%, 89%, 85%, 87% LR: 97%, 100%, 92%, 96%
(Sujatha & Mahalakshmi, 2020)	Evaluación del rendimiento de algoritmos de aprendizaje automático supervisados en la predicción de enfermedades cardíacas	Random forest, Máquinas de vectores de soporte, Naive bayes, Regresión logística, Árbol de decisión y K-vecinos más cercanos	Exactitud, Precisión, AUC, F1-Score	RF: 83.5%, 88.8%, 88.2%, 84.2% SVM: 82.4%, 85.4%, 86.2%, 83.6% NB: 82.4%, 86.9%, 86%, 83.3% LR: 80.2%, 82%, 85.5%, 82% DT: 79.1%, 82.9%, 78.9%, 80.4% KNN: 72.5%, 75.5%, 76.8%, 74.7%

(Roy, Pal, Das, & Huq, 2020)	Estudio comparativo de enfoques de aprendizaje automático para el diagnóstico del cáncer de mama para dos conjuntos de datos diferentes	Naive bayes, Máquinas de vectores de soporte, K-vecinos más cercanos y Regresión logística	Exactitud, Sensibilidad, Precisión, F1-Score	NB: 92%, 92%, 93%, 91% SVM: 96%, 97%, 97%, 97% KNN: 97%, 97%, 97%, 97% LR: 99%, 99%, 99%, 99%
(Setiawan, Rustam, Hartini, Laeli, & Wirasati, 2020)	Comparación de Naive bayes y árbol de decisión para clasificar el carcinoma hepatocelular (CHC)	Naive bayes y Árbol de decisión	Exactitud, Precisión, Recall, F1-Score	NB: 94.1%, 100%, 83.2%, 93.2% DT: 98.8%, 96.8%, 100%, 98.72%
(Sengar, Gaikwad, & Nagdive, 2020)	Estudio comparativo de algoritmos de aprendizaje automático para la predicción del cáncer de mama	Regresión logística, Árbol de decisión	Exactitud	LR: 94.4% DT: 95.1%
(Al Helal, Islam Chowdhury, Islam, Ahmed, & Hossain, 2019)	Un enfoque de optimización para mejorar el rendimiento de la clasificación en la predicción del cáncer y la diabetes	K-vecinos más cercanos, Naive bayes y Random forest	Exactitud, Precisión, Recall, F1-Score	KNN: 94.4%, 94.5%, 94.4%, 94.4% NB: 96%, 96.4%, 96.4%, 96.4%

				RF: 96.9%, 97%, 97%, 97%
(Günaydin, Günay, & Şengel, 2019)	Comparación de algoritmos de detección de cáncer de pulmón	K-vecinos más cercanos, Máquinas de vectores de soporte, Naive bayes y Árbol de decisión	Exactitud, Precisión, Recall, F1-Score	KNN: 74.3%, 74.5%, 86.3%, 80% SVM: 50%, 33.3%, 85%, 47.8% DT: 93.24%, 94.12%, 96%, 95%
(Harshitha, Chaitanya, Killedar, Revankar, & Pushpa , 2019)	Reconocimiento y predicción del cáncer de mama mediante diagnóstico supervisado	Máquinas de vectores de soporte y K-vecinos más cercanos	Exactitud	SVM: 95% KNN: 97%
(Amrane, Oukid, Gagaoua, & Ensari, 2018)	Clasificación del cáncer de mama mediante aprendizaje automático	K-vecinos más cercanos y Naive bayes	Exactitud	KNN: 97.5% NB: 96.1%
(Sharma, Aggarwal, & Choudhury, 2018)	Detección del cáncer de mama mediante algoritmos de aprendizaje automático	Random forest, K-vecinos más cercanos y Naive bayes	Exactitud, Precisión, Recall y F1-Score	RF: 94.7%, 92.1%, 93.6%, 92.9% KNN: 95.9%, 98.2%, 90.4%, 94.2% NB: 94.4%, 88.5%, 85.7%, 87%

(Khuriwal & Mishra, 2018)	Diagnóstico de cáncer de mama mediante el algoritmo de aprendizaje automático del conjunto de votación adaptativa	Regresión logística	Precisión, Recall, F1-Score	LR: 98%, 98%, 98%
(Wen, Li, Li, Li, & Yin, 2018)	Comparación de cuatro técnicas de aprendizaje automático para la predicción de la supervivencia del cáncer de próstata	K-vecinos más cercanos, Árbol de decisión, Naive bayes y Máquinas de vectores de soporte	Exactitud, Precisión, Recall y F1-Score	KNN: 85.6%, 87%, 98%, 992% DT: 84.9%, 87%, 97%, 91% NB: 71%, 92%, 72%, 81% SVM: 85.4%, 87%, 98%, 92%

Fuente: Las Investigaciones se seleccionaron de las bases de datos científicas Scopus, IEEE Xplore y ScienceDirect.

De acuerdo a los resultados de las investigaciones se elaboró un top de los algoritmos de clasificación más utilizados y con los mejores resultados basándose en la métrica de Exactitud, esta métrica es una de las más importantes ya que permite mostrar las predicciones correctas. El top de algoritmos sirvió para conformar la población de algoritmos de la presente investigación.

Tabla 3.

Top de los algoritmos de clasificación con mejor exactitud

N°	Algoritmos de clasificación	Exactitud
1	Regresión logística [a]	99%
2	Árbol de decisión [b]	98.8%
3	K-vecinos más cercanos [c]	97.5%
4	Random forest [d]	96.9%
5	Máquinas de vectores de soporte [e]	96.5%
6	Naive bayes [f]	94.4%

Nota: Tomado de [a]Roy, Pal, Das, & Huq (2020, pág. 32). [b]Setiawan, Rustam, Hartini, Laeli, & Wirasati (2020, pág. 4). [c]Amrane, Oukid, Gagaoua, & Ensari (2018, pág. 3). [d]Al Helal, Islam Chowdhury, Islam, Ahmed, & Hossain (2019, pág. 4). [e]Ara, Das, & Dey (2021, pág. 100). [f]Sharma, Aggarwal, & Choudhury (2018, pág. 117).

Finalmente, de acuerdo a los resultados mostrados en el top de los algoritmos de clasificación con mejor exactitud, se seleccionó los algoritmos de Regresión logística y Árbol de decisión por tener los mejores resultados, dichos algoritmos fueron seleccionados para el desarrollo de la presente investigación.

Estructurar el tratamiento del conjunto de datos.

Los datos fueron obtenidos del departamento de Pediatría del Hospital Regional Docente "Las Mercedes" de Chiclayo con previa autorización, la misma que se puede visualizar en el Anexo 2. Con dicha autorización se pudo acceder a los libros de registros de los pacientes dados de alta durante los años 2016 - 2021. En los libros se observaron diferentes datos personales y de diagnóstico de los pacientes, los datos a recolectar solo serán utilizados para el desarrollo de la investigación, respetando la privacidad y confidencialidad de la información.

Tabla 4.

Datos de los libros de registros de pacientes del departamento de Pediatría

Nombre del Campo	**Descripción**
Paciente	Por confidencialidad se asignó un alias
Sexo	Sexo del paciente
Edad	Edad del paciente
Cama	Cama asignada al paciente
Dirección	Dirección domiciliaria del paciente
Diagnóstico de Ingreso	Diagnóstico de ingreso del paciente
Fecha de Ingreso	Fecha de ingreso del paciente
Hora	Hora de ingreso del paciente
Fecha de Egreso	Fecha de Alta Hospitalaria
Estancia	Cantidad de días de permanencia del paciente
Diagnóstico Final	Diagnóstico final del paciente
SIS	Número de seguro del paciente
Historia Clínica	Número de historia clínica del paciente
Teléfono	Número de teléfono del paciente o apoderado(a)
Firma	Firma del paciente o apoderado(a)

Fuente: Elaboración propia.

En los libros de registros de los pacientes dados de alta se identificó un total 75 pacientes diagnosticados con leucemia, de los cuales 68 fueron diagnosticados

con el tipo de Leucemia Linfoblástica Aguda o LLA y 7 pacientes fueron diagnosticados con Leucemia Mieloide Aguda o LMA. Con los números de las historias clínicas de los pacientes se accedió al departamento de Archivo para localizar las historias clínicas de cada paciente que fue diagnosticado con leucemia de tipo LLA y LMA.

Se realizó una entrevista con la especialista en hematología del Hospital, quien es la encargada de evaluar los exámenes aplicados a los pacientes del departamento de Pediatría y diagnosticar los tipos de leucemia. Las preguntas planteadas y las respuestas se muestran en el Anexo 3. De esta manera se aplicaron criterios de inclusión y exclusión para recolectar la información.

Tabla 5.

Criterios de Inclusión y Exclusión de la información

Criterios	**Detalle**
Inclusión	Exámenes de diagnóstico de leucemia aplicados a pacientes con edades de 0 a 14 años.
	Exámenes de diagnóstico de leucemia aplicado en los años 2016-2021.
	Otros exámenes complementarios para el diagnóstico de los tipos de leucemia.
	Pacientes con diagnóstico final de LLA y LMA.
Exclusión	Exámenes de diagnóstico de leucemia incompletos.
	Exámenes de diagnóstico de leucemia realizados fuera del periodo establecido.
	Exámenes de diagnóstico de leucemia ausente.
	Pacientes sin diagnóstico final.

Fuente: Elaboración propia.

Posteriormente se construyó el dataset con los datos recolectados, teniendo un total de 75 pacientes de los cuales 68 tienen diagnóstico de Leucemia Linfoblástica Aguda y 7 pacientes tienen Leucemia Mieloide Aguda. Las primeras y últimas filas que contiene el dataset se visualizan en la Tabla 6.

Tabla 6.

Datos que contiene el dataset

Paciente	Sexo	Edad	Leucocitos	Eritrocitos	Plaquetas	Hemoglobina	Mieloperoxidasa	Fosfatasa_ácida	Diagnóstico
1	Masculino	6	26000	2100000	18000	8.5	Negativo	Positivo	LLA
2	Masculino	3	21100	1320000	26000	7.5	Negativo	Positivo	LLA
3	Masculino	5	21800	1450000	47000	10.5	Negativo	Positivo	LLA
4	Masculino	2	19100	2350000	11800	6.9	Negativo	Positivo	LLA
5	Femenino	11	20200	1350000	26000	6.8	Negativo	Positivo	LLA
.	.	.	.	.	.	.	.	.	.
72	Femenino	10	29100	2600000	224000	9.7	Positivo	Negativo	LMA
73	Masculino	5	10790	1400000	60000	8.4	Negativo	Positivo	LLA
74	Masculino	12	29800	2240000	221000	10.0	Positivo	Negativo	LMA
75	Masculino	2	19800	1400000	98000	7.5	Negativo	Positivo	LLA

```
# Cantidad de valores de la columna Diagnóstico
df["Diagnóstico"].value_counts()
LLA    68
LMA     7
Name: Diagnóstico, dtype: int64
```

Fuente: Elaboración propia.

Se utilizó la función value_counts para contar cuantos pacientes fueron diagnosticados con el tipo de leucemia LLA y LMA, dando como resultado 68 pacientes con LLA y 7 pacientes con LMA, luego se procedió a identificar el tipo de dato de las variables que contiene el dataset, siendo necesario identificar si existen variables categóricas para poder brindar al algoritmo solo variables cuantitativas, para ello se utilizó el comando df.dtypes.

Tabla 7.

Tipo de dato de las variables del dataset

Columna	**Tipo de dato**
dataf.dtypes	
Sexo	object
Edad	int64
Leucocitos	int64
Eritrocitos	int64
Plaquetas	int64
Hemoglobina	float64
Mieloperoxidasa	object
Fosfatasa_ácida	object
Diagnóstico	object

Fuente: Elaboración propia.

Se identificó 4 variables de tipo categóricas, las cuales fueron seleccionadas para convertirlas en cuantitativas, para este proceso se utilizó la clase LabelEncoder de la librería preprocessing de sklearn, que permitió codificar los datos categóricos a numéricos, haciendo uso de la función fit_transform, se cambió los datos categóricos a numéricos entre 0 y 1.

Tabla 8.

Convertir las variables categóricas en cuantitativas

Código

```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
laen = LabelEncoder()
dataf.Sexo = laen.fit_transform(dataf.Sexo)
dataf.Mieloperoxidasa =
laen.fit_transform(dataf.Mieloperoxidasa)
dataf.Fosfatasa_ácida =
laen.fit_transform(dataf.Fosfatasa_ácida)
labelencoder_V = LabelEncoder()
dataf.iloc[:,9]=
labelencoder_V.fit_transform(dataf.iloc[:,9].values)
print(labelencoder_V.fit_transform(dataf.iloc[:,9].values))
dataf.head()
```

Fuente: Elaboración propia.

Se obtuvo que en la variable Sexo, el valor Masculino fue sustituido por 1 y Femenino por 0. En la variable Mieloperoxidasa, el valor Positivo fue sustituido por 1 y Negativo por 0. En la variable Fosfatasa_ácida, el valor Positivo fue sustituido por 1 y Negativo por 0 y en la variable Diagnóstico, el valor LLA fue sustituido por 1 y LMA por 0. Los nuevos valores del dataset se visualizan en la Tabla 9.

Tabla 9.

Nuevos valores del dataset

Paciente	Sexo	Edad	Leucocitos	Eritrocitos	Plaquetas	Hemoglobina	Mieloperoxidasa	Fosfatasa_ácida	Diagnóstico
1	1	6	26000	2100000	18000	8.5	0	1	1
2	1	3	21100	1320000	26000	7.5	0	1	1
3	1	5	21800	1450000	47000	10.5	0	1	1
4	1	2	19100	2350000	11800	6.9	0	1	1
5	0	11	20200	1350000	26000	6.8	0	1	1
.	.	.	.	.	.	.	.	.	.
72	0	10	29100	2600000	224000	9.7	1	0	0
73	1	5	10790	1400000	60000	8.4	0	1	1
74	1	12	29800	2240000	221000	10.0	1	0	0
75	1	2	19800	1400000	98000	7.5	0	1	1

```
# Cantidad de valores de la columna Diagnóstico
df["Diagnóstico"].value_counts()
1    68
0     7
Name: Diagnóstico, dtype: int64
```

Fuente: Elaboración propia.

Posteriormente se realizó la comprobación de datos faltantes, para esto se utilizó la función isna(), que verifica si hay valores nulos en el dataset, corroborando de esta manera que el dataset no contenía ningún valor nulo.

Tabla 10.

Comprobación de datos faltantes en el dataset

Código

```
dataf.isna().sum()
Paciente           0
Sexo               0
Edad               0
Leucocitos         0
Eritrocitos        0
Plaquetas          0
Hemoglobina        0
Mieloperoxidasa    0
Fosfatasa_ácida    0
Diagnóstico        0
dtype: int64
```

Fuente: Elaboración propia.

Implementar los algoritmos de clasificación.

Con el dataset definido, se procedió a implementar los algoritmos de clasificación seleccionados en la muestra del presente trabajo de investigación, que son el algoritmo de Regresión logística y Árbol de decisión. Para ello se instaló el ambiente de trabajo Anaconda Navigator que proporciona el uso de la aplicación de código abierto Jupyter Notebook, la cual se utilizó para aplicar el código en Python.

En la aplicación Jupyter Notebook se importó la librería pandas, esta librería brinda herramientas para el análisis de los datos. También se importó la librería Numpy, esta librería sirve para generar vectores y matrices y permite realizar diversas operaciones matemáticas. Posteriormente se importó el dataset en

formato de Excel con nombre LeucemiaInfantil.xlsx, para ello se utilizó la función read_excel, que permite leer este tipo de archivos.

Tabla 11.

Importar el dataset

Código

```
dataf = pd.read_excel("LeucemiaInfantil.xlsx")
dataf.head()
```

Fuente: Elaboración propia.

Los algoritmos de clasificación utilizados para diagnosticar tipos de leucemia infantil son de tipo Aprendizaje Supervisado, para ello se tienen definidos los valores de salida para nuestras muestras. Para comprobar que los modelos a implementar funcionan correctamente, se consideró realizar una prueba previa con los datos de dos pacientes. Se tomó primero los datos del paciente número 5, el cual se conoce que fue diagnosticado con LLA. Para ello se tomaron los valores de la columna 1 que es Sexo, a la columna 8 que es Fosfatasa_ácida. No se consideró la columna 9 que es diagnóstico, por ser la columna que contiene los valores a predecir y también no se consideró la primera columna 0, que es Paciente, por tener datos unique.

Tabla 12.

Selección de pacientes para probar los modelos

Código

```
# Paciente 5, diagnóstico = LLA
paciente_5 = dataf.iloc[5, 1:8].values
print("datos paciente 5 : \n", paciente_5)
# paciente 8, diagnóstico = LMA
paciente_8 = dataf.iloc[8, 1:8].values
print("datos paciente 8 : \n", paciente_8)
```

Fuente: Elaboración propia.

Posteriormente se definió las variables de análisis, siendo E, las variables predictoras, se consideró de la columna 1 a la 8, que son, Sexo, Edad, Leucocitos, Eritrocitos, Plaquetas, Hemoglobina, Mieloperoxidasa y Fosfatasa ácida. Mientras que P, representa la variable a predecir, por lo cual se le asigno la columna 9, que es Diagnóstico. Luego haciendo uso de la función train_test_split, se dividió el conjunto de datos en entrenamiento y prueba, donde se le asignó el parámetro test_size=0.2, que representa al 20% de datos para prueba y el 80% de los datos se asignaron para el entrenamiento. El parámetro random_state=20 representa la semilla utilizada para la aleatoriedad de los datos.

Tabla 13.

Definir las variables y dividir el dataset

Código

```
from sklearn.model_selection import train_test_split
E = dataf.iloc[:, 1:8].values
P = dataf.iloc[:, 9].values
E_train, E_test, P_train, P_test = train_test_split(E, P,
test_size =0.2, random_state = 20)
```

Fuente: Elaboración propia.

Algoritmo de Regresión logística

El algoritmo de Regresión logística es usado para encontrar la relación entre la variable dependiente que es de naturaleza binaria y una o más variables independientes, para su implementación se utilizó la función LogisticRegression de la librería linear_model de sklearn, donde se entrenó el algoritmo para realizar el diagnóstico de los tipos de leucemia infantil a los dos pacientes seleccionados, para probar el modelo con las variables (E_train, P_train), donde se le asignó 60 datos de pacientes para el entrenamiento que representa el 80% y 15 datos de pacientes para las pruebas que representa el 20% del dataset total. Luego de entrenar el modelo, se procedió a realizar el diagnóstico con los datos del paciente 5 y el paciente 8 respectivamente.

Tabla 14.

Implementación del algoritmo de Regresión logística

Código

```
import time
inicio = time.time()
from sklearn.linear_model import LogisticRegression
logisticRegression = LogisticRegression(random_state = 20)
logisticRegression.fit(E_train, P_train)
logisticRegression.score(E_train,P_train)
#Prueba del modelo con el Paciente 5, diagnóstico = LLA
pred_1 = logisticRegression.predict([paciente_5])
pred_proba_1 = logisticRegression.predict_proba([paciente_5])
print("Diagnóstico: ", pred_1)
print("Probabilidad LLA: ", pred_proba_1[0][1])
print("Probabilidad LMA: ", pred_proba_1[0][0])
Diagnóstico:  [1]
Probabilidad LLA:  0.9973756502856075
Probabilidad LMA:  0.0026243497143925154
#Prueba del modelo con el Paciente 8, diagnóstico = LMA
pred_2 = logisticRegression.predict([paciente_8])
pred_proba_2 = logisticRegression.predict_proba([paciente_8])
print("Diagnóstico: ", pred_2)
print("Probabilidad LLA: ", pred_proba_2[0][1])
print("Probabilidad LMA: ", pred_proba_2[0][0])
Diagnóstico:  [0]
Probabilidad LLA:  0.0697282266888823
Probabilidad LMA:  0.9302717733111177
```

Fuente: Elaboración propia.

La Tabla 14 mostró que el diagnóstico del paciente 5 obtuvo como resultado LLA, con una probabilidad del 0.9973756502856075 y una probabilidad de 0.0026243497143925154 para LMA. El paciente 8 obtuvo como resultado LMA, con una probabilidad del 0.9302717733111177 y una probabilidad de 0.0697282266888823 para LLA. Siendo de esta manera correcto el diagnóstico de los tipos de leucemia de ambos pacientes.

Algoritmo Árboles de decisión

Este algoritmo es usado para predecir el valor de una variable en situaciones donde en resultado es incierto, los árboles se componen de nodos, los cuales buscan la división más optima de las características, para su implementación se utilizó la función DecisionTreeClassifier de la librería tree de sklearn, donde se le asignó max_depth=3 de profundidad del árbol. Luego se entrenó el algoritmo para realizar el diagnóstico de los tipos de leucemia infantil con las variables (E_train, P_train), donde se le asignó 60 datos de pacientes para el entrenamiento que representa el 80% y 15 datos de pacientes para las pruebas que representa el 20% del dataset total. Luego de entrenar el modelo, se procedió a realizar el diagnóstico con los datos del paciente 5 y el paciente 8 respectivamente.

Tabla 15.

Implementación del algoritmo de Árboles de decisión

Código

```
import time
inicio = time.time()
from sklearn.tree import DecisionTreeClassifier
decisionTreeClassifier = DecisionTreeClassifier(criterion =
"entropy", max_depth=3,random_state = 20)
decisionTreeClassifier.fit(E_train, P_train)
#Prueba del modelo con el Paciente 5, diagnóstico = LLA
pred_1 = decisionTreeClassifier.predict([paciente_5])
pred_proba_1 =
decisionTreeClassifier.predict_proba([paciente_5])
print("Diagnóstico: ", pred_1)
print("Probabilidad LLA: ", pred_proba_1[0][1])
print("Probabilidad LMA: ", pred_proba_1[0][0])
Diagnóstico:  [1]
Probabilidad LLA:  1.0
Probabilidad LMA:  0.0
#Prueba del modelo con el Paciente 8, diagnóstico = LMA
pred_2 = decisionTreeClassifier.predict([paciente_8])
pred_proba_2 =
decisionTreeClassifier.predict_proba([paciente_8])
print("Diagnóstico: ", pred_2)
print("Probabilidad LLA: ", pred_proba_2[0][1])
```

```
print("Probabilidad LMA: ", pred_proba_2[0][0])
Diagnóstico:  [0]
Probabilidad LLA:  0.0
Probabilidad LMA:  1.0
```

Fuente: Elaboración propia.

La Tabla 15 mostró que el diagnóstico del paciente 5 obtuvo como resultado LLA, con una probabilidad del 1.0 y una probabilidad de 0.0 para LMA. El paciente 8 obtuvo como resultado LMA, con una probabilidad del 1.0 y una probabilidad de 0.0 para LLA. Siendo de esta manera correcto el diagnóstico de los tipos de leucemia de ambos pacientes.

Evaluar el desempeño de los algoritmos de clasificación.

El desempeño de los algoritmos en el diagnóstico de los tipos de leucemia infantil fue evaluado mediante los indicadores de Exactitud, Precisión, Sensibilidad y F1 Score, para ello se utilizó la matriz de confusión. Otro de los indicadores evaluados también fue el tiempo de respuesta de los algoritmos también, de esta manera obtener un desempeño óptimo en cuanto a un correcto y rápido diagnóstico del tipo de leucemia infantil.

Para aplicar la matriz de confusión a los dos modelos a evaluar, primero se importaron las funciones confusion_matrix de la Matriz de confusión, accuracy_score para calcular la Exactitud, precision_score para calcular la Precisión, recall_score para calcular la Sensibilidad, f1_score para calcular el F1 Score y plot_confusion_matrix para generar una gráfica de la matriz de confusión. Todas las funciones pertenecen a la librería metrics de sklearn.

Tabla 16.

Matriz de confusión aplicada al algoritmo de Regresión logística

Código

```
#Medir el modelo de regresión logística
P_pred = logisticRegression.predict(E_test)
confusion_matrix
confusion_matrix = confusion_matrix(P_test,P_pred)
confusion_matrix
array([[ 1,  1],
       [ 0, 13]], dtype=int64)
```

Fuente: Elaboración propia.

Para visualizar de manera gráfica la matriz de confusión aplicada al modelo de Regresión logística, se utilizó la función plot_confusion_matrix, como se observa en la Figura 12.

```
plot_confusion_matrix(logisticRegression, E_test, P_test,
values_format='3g')
```

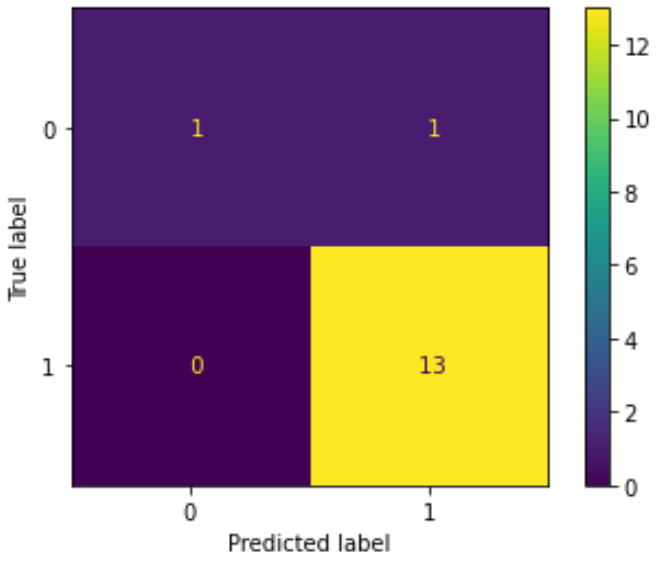

Figura 12. Matriz de confusión del modelo de Regresión logística. Fuente: Elaboración propia.

Los resultados mostraron, que el modelo obtuvo 13 predicciones correctas como verdaderos positivos, 1 como verdadero negativo, 1 falso positivo y 0 falsos

negativos. Luego se procedió a aplicar las métricas para evaluar el desempeño del modelo de Regresión logística.

Tabla 17.

Desempeño del modelo de Regresión logística

Código

```
#Aplicar las métricas
Exactitud = accuracy_score(P_test,P_pred)
Precisión = precision_score(P_test,P_pred)
Sensibilidad = recall_score(P_test,P_pred)
F1_Score = f1_score(P_test,P_pred)
#Imprimir los resultados
print("Total de Exactitud:", round(Exactitud*100, 1),'%')
print("Total de Precisión:", round(Precisión*100, 1),'%')
print("Total de Sensibilidad:", round(Sensibilidad*100,
1),'%')
print("Total de F1_Score:", round(F1_Score*100, 1),'%')
#Imprimir resultado del tiempo de respuesta del algoritmo
fin = time.time()
print('Tiempo de respuesta del Algoritmo: ', (fin-inicio))
Total de Exactitud: 93.3%
Total de Precisión: 92.9%
Total de Sensibilidad: 100.0%
Total de F1_Score: 96.3%
Tiempo de respuesta del Algoritmo: 0.052268028259277344
```

Fuente: Elaboración propia.

La Tabla 17 mostró que el algoritmo de clasificación Regresión logística obtuvo un 93.3% de exactitud para diagnosticar tipos de leucemia infantil. También se evaluó el también el tiempo de respuesta del algoritmo de Regresión logística, teniendo en cuenta como inició del tiempo a calcular, desde la implementación del algoritmo hasta la impresión de los resultados de las métricas. Para ello se utilizó la función time, indicando el inicio del cálculo con la variable inicio = time.time(), como se observa en la Tabla 13 y para el fin del tiempo a calcular, se utilizó la variable fin = time.time(), como se muestra en la Tabla 17, obteniendo un tiempo de respuesta de 0.05 segundos.

Luego se evaluó el desempeño del algoritmo de Árboles de decisión. El código utilizado se visualiza en la Tabla 18.

Tabla 18.

Matriz de confusión aplicada al algoritmo de Árboles de decisión

Código

```
#Medir el modelo de Árboles de decisión
P_pred = decisionTreeClassifier.predict(E_test)
confusion_matrix
confusion_matrix = confusion_matrix(P_test,P_pred)
confusion_matrix
array([[ 2,  0],
       [ 0, 13]], dtype=int64)
```

Fuente: Elaboración propia.

Para visualizar de manera gráfica la matriz de confusión aplicada al modelo Árboles de decisión, se utilizó la función plot_confusion_matrix, como se observa en la Figura 13.

```
plot_confusion_matrix(decisionTreeClassifier, E_test, P_test,
values_format='3g')
```

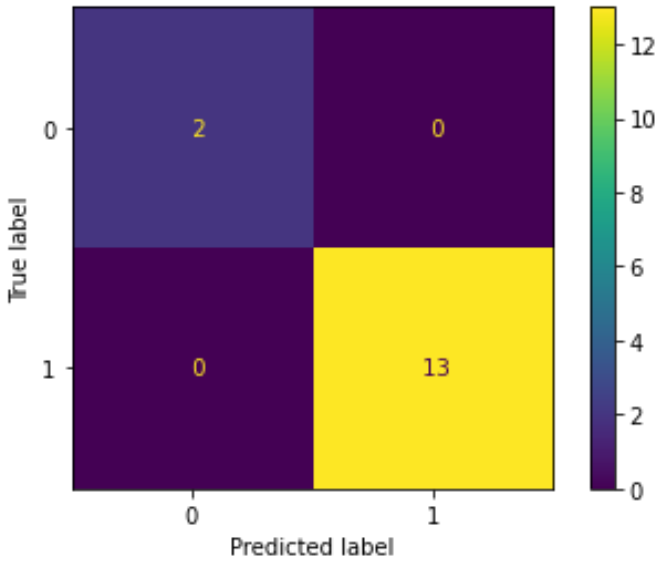

Figura 13. Matriz de confusión del modelo de Árboles de decisión. Fuente: Elaboración propia.

Los resultados mostraron, que el modelo obtuvo 13 predicciones correctas como verdaderos positivos, 2 como verdaderos negativos, 0 falsos positivos y 0 falsos negativos. Luego se procedió a aplicar las métricas para evaluar el desempeño del modelo de Árboles de decisión.

Tabla 19.

Desempeño del modelo de Árboles de decisión

Código

```
#Aplicar las métricas
Exactitud = accuracy_score(P_test,P_pred)
Precisión = precision_score(P_test,P_pred)
Sensibilidad = recall_score(P_test,P_pred)
F1_Score = f1_score(P_test,P_pred)
#Imprimir los resultados
print("Total de Exactitud:", round(Exactitud*100, 1),'%')
print("Total de Precisión:", round(Precisión*100, 1),'%')
print("Total de Sensibilidad:", round(Sensibilidad*100,
1),'%')
print("Total de F1_Score:", round(F1_Score*100, 1),'%')
#Imprimir resultado del tiempo de respuesta del Algoritmo
fin = time.time()
print('Tiempo de respuesta del Algoritmo: ', (fin-inicio))
Total de Exactitud: 100.0%
Total de Precisión: 100.0%
Total de Sensibilidad: 100.0%
Total de F1_Score: 100.0%
Tiempo de respuesta del Algoritmo:  0.024228334426879883
```

Fuente: Elaboración propia.

La Tabla 19 mostró que el algoritmo de clasificación de Árboles de decisión obtuvo un 100.0% de exactitud para diagnosticar tipos de leucemia infantil. También se evaluó el tiempo de respuesta del algoritmo, teniendo en cuenta como inició del tiempo a calcular, desde la implementación del algoritmo hasta la impresión de los resultados de las métricas. Para ello se utilizó la función time, indicando el inicio del cálculo con la variable inicio = time.time(), como se observa en la Tabla 15 y para el fin del tiempo a calcular, se utilizó la variable

fin = time.time(), como se muestra en la Tabla 19, obteniendo un tiempo de respuesta de 0.02 segundos.

IV. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones.

a) La revisión de la literatura científica permitió obtener los algoritmos de clasificación que mejor desempeño tuvieron en el diagnóstico de diferentes tipos de cáncer de acuerdo a la métrica de exactitud, de esta manera se elaboró un top para determinar los dos algoritmos que mejores resultados obtuvieron, siendo seleccionados los dos primeros para ser implementados en la presente investigación.
b) Para estructurar correctamente los datos del dataset, primero se identificó los tipos de datos de las variables, donde se identificó que el dataset tenía datos categóricos que posteriormente fueron convertidos a numéricos, siendo este proceso necesario para que los modelos puedan interpretar de manera correcta los datos a predecir.
c) Con la finalidad de entrenar con la mayor cantidad de datos posibles los modelos, se dividió el dataset que contenía 75 datos de pacientes en total, de los cuales se seleccionaron 60 datos de pacientes que representa el 80% del dataset para realizar el entrenamiento y 15 datos de pacientes que representa el 20% del dataset se utilizaron para realizar las pruebas, considerando que los modelos necesitan más datos asignados al entrenamiento para obtener una predicción correcta.
d) La comparación de los resultados obtenidos mostró que el algoritmo de clasificación Árboles de decisión, es el mejor para diagnosticar los tipos de leucemia infantil, considerando el desempeño obtenido al evaluarse todos los indicadores propuestos en esta investigación.

4.2. Recomendaciones.

a) Para realizar la selección de los algoritmos de clasificación a implementar, se puede hacer una revisión de otras bases de datos

científicas y considerar hacer un top de los algoritmos con el mejor desempeño de otras métricas.

b) Para convertir las variables categóricas a numéricas se recomienda utilizar diferentes criterios que permitan hacer una correcta identificación de tal manera que solo se utilicen los datos necesarios para implementar los modelos correctamente.
c) Los resultados obtenidos en esta investigación pueden variar si se le asignan otros porcentajes de datos para entrenar y probar los modelos. Así también como hacer modificaciones en la aleatoriedad de los datos.
d) Para obtener resultados semejantes o mejores en el diagnóstico de los tipos de leucemia infantil, se recomienda utilizar las características de otros tipos de exámenes complementarios en el diagnóstico de los tipos de leucemia.

REFERENCIAS.

AEAL. (2014). *Leucemia Mieloide Aguda.* España: AEAL, Asociación Española de Afectados por Linfoma, Mieloma y Leucemia.

AEAL. (2017). *Leucemia Linfoblástica Aguda.* España: Aeal. Asociación Española de Afectados por Linfoma, Mieloma y Leucemia.

Aggarwal, C. (2014). *Data Classification Algorithms and Applications.* Estados Unidos: Chapman and Hall/CRC.

Al Helal, M., Islam Chowdhury, A., Islam, A., Ahmed, E., & Hossain, S. (2019). An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction. *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1-5.

Álvarez Vega, M., Quirós Mora, L. M., & Cortés Badilla, M. V. (2020). Inteligencia artificial y aprendizaje automático en Medicina. *Revista Médica Sinergia*, 2-3.

American Cancer Society. (12 de Febrero de 2019). *cancer.org*. Obtenido de https://www.cancer.org/es/cancer/leucemia-en-ninos/deteccion-diagnostico-clasificacion-por-etapas.html

American Cancer Society. (12 de Febrero de 2019). *Detección temprana de leucemia en niños*. Obtenido de https://www.cancer.org/es/cancer/leucemia-en-ninos/deteccion-diagnostico-clasificacion-por-etapas/como-se-diagnostica.html

Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 1-4.

Ara, S., Das, A., & Dey, A. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. *International Conference on Artificial Intelligence (ICAI)*, Pakistán.

Arora, M., Som, S., & Rana, A. (2020). Predictive Analysis of Machine Learning Algorithms for Breast Cancer Diagnosis. *8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1-5.

Bing Liu. (2011). *Web Data Mining.* Estados Unidos: Springer.

Boryczka, U., & Kozak, J. (2010). Ant Colony Decision Trees – A New Method for Constructing Decision Trees Based on Ant Colony Optimization. *Computational Collective Intelligence. Technologies and Applications*, 376-378.

Chand, S., & Vishwakarma, V. (2019). Leukemia Diagnosis using Computational Intelligence. *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 1-6.

Chandra, S., Khemchandani, R., & Jayadeva. (2017). *Twin Support Vector Machines: Models, Extensions and Applications.* Suiza: Springer.

Das Mou, A., & Kumar Saha, P. (2019). A Comprehensive Study of Machine Learning algorithms for Predicting Leukemia Based on Biomedical Data. *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, 1-5.

El-Shair, Z., Sánchez-Pérez, L., & Rawashdeh, S. (2020). Comparative Study of Machine Learning Algorithms using a Breast Cancer Dataset. *IEEE International Conference on Electro Information Technology (EIT)*, 1-9.

Fernández Montoro, A. (2013). *Python 3 al descubierto - 2a ed.* México: Alfaomega.

García Montero, P. (2015). *Aprende a Programar en R: 2ª Edición.* España: IT Campus Academy.

González Duque, R. (2017). *Python para todos.* España: Creative Commons.

Günaydin, Ö., Günay, M., & Şengel, Ö. (2019). Comparison of Lung Cancer Detection Algorithms. *Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 1-4.

Harshitha, Chaitanya, V., Killedar, S., Revankar, D., & Pushpa , M. (2019). Reconocimiento y predicción del cáncer de mama mediante diagnóstico supervisado. *4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 1-6.

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners.* Estados Unidos: Kindle.

Hsu, C.-H., Chen, X., Lin, W., Jiang, C., Zhang, Y., Hao, Z., & Chung, Y. (2021). Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning. *Measurement: Journal of the International Measurement Confederation*, 2-7.

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning for dummies.* Estados Unidos: John Wiley & Sons.

INEN. (2016). Registro de Cáncer de Lima Metropolitana: Incidencia y Mortalidad 2010-2012. *Registro de cáncer de Lima Metropolitana*, 9-177. Obtenido de https://portal.inen.sld.pe/registro-de-cancer-en-lima-metropolitana/

Kaisermann, J., Pawlowski, M., & Mendel, Y. (2020). *Técnicas de biología molecular I.* Estados Unidos: Cambridge Stanford Books.

Khuriwal, N., & Mishra, N. (2018). Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. *IEEMA Engineer Infinite Conference (eTechNxT)*, 1-5.

Kramer, O. (2013). *Dimensionality Reduction with Unsupervised Nearest Neighbors.* Alemania: Springer.

Lemeshow, S., Hosmer Jr, D., & Sturdivant, R. (2013). *Applied Logistic Regression Third Edition.* Estados Unidos: Wiley.

Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufran, H., & Yaqoob, M. (2020). Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. *Medical & Biological Engineering & Computing*, 2-8.

Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufran, H., & Yaqoob, M. (2020). Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. *Medical & Biological Engineering & Computing*, 2-8.

Ministerio de Salud. (2017). *Plan Nacional Para la Atención Integral de la Leucemia Linfática Aguda en Pacientes de 1 a 21 años.* Lima: MINSA.

Müller, A., & Guido, S. (2017). *Introduction to Machine Learning with Python.* Estados Unidos: O'Reilly.

Murty, M., & Raghava, R. (2016). *Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks.* Suiza: Springer.

Naji, M., El Filali, S., Aarika, K., Abdelouhahid, R., & Debauche, O. (2021). Algoritmos de aprendizaje automático para la predicción y el diagnóstico del cáncer de mama. *International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT)*, 487-492.

Organización Mundial de la Salud. (2021). Cáncer en la Niñez y la Adolescencia. *El cáncer infantil*, 17-21.

Patil Babaso, S., Mishra, S. K., & Junnarkar, A. (2020). Leukemia Diagnosis Based on Machine Learning Algorithms. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 1-4.

Preethi, I., & Dharmarajan, K. (2020). Diagnosis of chronic disease in a predictive model using machine learning algorithm. *International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 1-6.

Rokach, L., & Maimon, O. (2014). *Data Mining With Decision Trees: Theory And Applications (2nd Edition).* Estados Unidos: World Scientific.

Roy, B., Pal, M., Das, S., & Huq, A. (2020). Comparative Study of Machine Learning Approaches on Diagnosing Breast Cancer for Two Different Dataset. *2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, 32.

Russell, R. (2018). *Machine Learning.* Estados Unidos: CreateSpace Independent Publishing Platform.

Salah, H. T., Muhsen, I. N., Salama, M. E., Owaidah, T., & Hashmi, S. K. (2019). Statistical morphological analysis based supervised classification algorithm for diagnosing acute lymphoblastic Leukemia. *International Journal of Laboratory Hematology*, 2-8.

Sengar, P., Gaikwad, M., & Nagdive, A. (2020). Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. *Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1-6.

Seokho, K. (2021). k-Nearest Neighbor Learning with Graph Neural Networks. *Mathematics*, 1-4.

Setiawan, Q., Rustam, Z., Hartini, S., Laeli, A., & Wirasati, I. (2020). Comparison of Naive Bayes and Decision Tree for Classifying Hepatocellular Carcinoma (HCC). *International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, 1-5.

Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast Cancer Detection Using Machine Learning Algorithms. *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 1-5.

Sujatha, P., & Mahalakshmi, K. (2020). Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease. *IEEE International Conference for Innovation in Technology (INOCON)*, 1-7.

Terese Winslow LLC. (2007). *Terese Winslow*. Obtenido de https://www.teresewinslow.com/#/circulatory/

Thida, A., Yamamori, K., & Ma Ma, T. (2020). A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification. *IEEE 9th Global Conference on Consumer Electronics (GCCE)*, 1-3.

Wen, H., Li, S., Li, W., Li, J., & Yin, C. (2018). Comparision of Four Machine Learning Techniques for the Prediction of Prostate Cancer Survivability. *15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 114-115.

ANEXOS.

Anexo 1. Resolución de aprobación del proyecto de investigación.

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO
RESOLUCIÓN N°1000-A-2021/FIAU-USS

Pimentel, 11 de noviembre de 2021

VISTO:
El oficio N° 0359-2021/FIAU-IS-USS de fecha 14 de octubre de 2021, de la Dirección de Escuela profesional de INGENIERÍA DE SISTEMAS con el que remite el Acta de reunión N°0610-2021 del Comité de investigación de la referida Escuela profesional, acerca de la Tesis presentada por estudiantes del Programa de estudios de INGENIERÍA DE SISTEMAS, y;

CONSIDERANDO:

Que, de conformidad con la Ley Universitaria N° 30220 en su artículo 48° que a letra dice: "La investigación constituye una función esencial y obligatoria de la universidad, que la fomenta y realiza, respondiendo a través de la producción de conocimiento y desarrollo de tecnologías a las necesidades de la sociedad, con especial énfasis en la realidad nacional. Los docentes, estudiantes y graduados participan en la actividad investigadora en su propia institución o en redes de investigación nacional o internacional, creadas por las instituciones universitarias públicas o privadas.";

Que, de conformidad con el Reglamento de grados y títulos en su artículo 21° señala: "Los temas de trabajo de investigación, trabajo académico y *tesis* son *aprobados por el Comité de Investigación* y derivados a la facultad o Escuela de Posgrado, según corresponda, para la emisión de la resolución respectiva. El *periodo de vigencia de los mismos será de dos años*, a partir de su aprobación. En caso un tema perdiera vigencia, el Comité de Investigación evaluará la ampliación de la misma.

Que, de conformidad con el Reglamento de grados y títulos en su artículo 24° señala: La tesis es un estudio que debe denotar rigurosidad metodológica, originalidad, relevancia social, utilidad teórica y/o práctica en el ámbito de la escuela profesional. Para el grado de doctor se requiere una tesis de máxima rigurosidad académica y de carácter original. Es individual para la obtención de un grado; *es individual o en pares para obtener un título profesional*. Asimismo, en su artículo 25° señala: "El tema debe responder a alguna de las líneas de investigación institucionales de la USS S.A.C.".

Que, mediante documentos de vistos, el Comité de investigación de la referida Escuela profesional acordó aprobar la ampliación de la vigencia de las tesis que se detallan en el Acta de reunión N°0610-2021, a cargo de estudiantes del Programa de estudios INGENIERÍA DE SISTEMAS, hasta el 6 de octubre de 2023.

Estando a lo expuesto, y en uso de las atribuciones conferidas y de conformidad con las normas y reglamentos vigentes;

SE RESUELVE:

ARTÍCULO ÚNICO: AMPLIAR VIGENCIA, de la Tesis a cargo de los estudiantes del Programa de estudios de **INGENIERÍA DE SISTEMAS** que se detallan en el anexo de la presente Resolución, hasta el 6 de octubre de 2023.

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE

Cc: Interesado, Archivo

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO
RESOLUCIÓN N°1000-A-2021/FIAU-USS

Pimentel, 11 de noviembre de 2021

ANEXO

N°	APELLIDOS Y NOMBRES	TEMA DE TESIS	FECHA RESOLUCIÓN DE APROBACIÓN / MODIFICACIÓN TEMA DE TESIS / AMPLIACIÓN DE VIGENCIA
1	DIAZ CARRASCO NATIVIDAD ALEJANDRO	EVALUACIÓN DE ALGORITMOS PARA LA DETECCIÓN DE HUELLAS DACTILARES ALTERADAS	13-05-2016
2	MENDOZA LINARES JERSSON GERMAN	METODOLOGÍA DE CONVERSIÓN DE APLICACIONES MONOLÍTICAS A MICROSERVICIOS DESPLEGABLE EN LA NUBE PARA PEQUEÑAS EMPRESAS	22-07-2019
3	GONZALEZ FLORES PAUL GUSTAVO	ANÁLISIS COMPARATIVO DE ALGORITMOS DE CLASIFICACIÓN PARA DIAGNOSTICAR TIPOS DE LEUCEMIA INFANTIL	17-11-2020
4	SOPLOPUCO MONJA BRAYAN ALONSO	COMPARACIÓN DE TÉCNICAS CONSTRUCCIÓN DE PROTOTIPOS Y REVISIÓN DE REQUERIMIENTOS PARA REALIZAR UNA CORRECTA VALIDACIÓN DE REQUERIMIENTOS DE SOFTWARE	Año 2018
5	MONTENEGRO GUERRERO VICTOR AGUSTIN	ANALISIS COMPARATIVO DE ALGORITMOS DE MACHINE LEARNING PARA DETECCION DE MALWARE EN APLICACIONES ANDROID	Año 2019

Facultad de Ingeniería, Arquitectura y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN SAC

Anexo 2. Carta de aceptación de la institución para la recolección de datos.

Ministerio de Salud
Personas que atendemos Personas

HOSPITAL REGIONAL DOCENTE "LAS MERCEDES"

Nº 048/ 21

AUTORIZACIÓN

El Director y el Jefe de la Unidad de Apoyo a la Docencia e Investigación del Hospital "Las Mercedes" Chiclayo, Autoriza a:

GONZALEZ FLORES
PAUL GUSTAVO

Estudiante de la Universidad Particular Señor de Sipán; Para que realice la Ejecución del Proyecto de Tesis Titulado: "*Análisis Comparativo de Algoritmos de Clasificación para Diagnosticar Tipos de Leucemia Infantil*" en los Servicios del Departamento de Pediatría de este nosocomio, debiendo al término remitir las conclusiones respectivas.

Chiclayo, Setiembre 2021.

GOBIERNO REGIONAL LAMBAYEQUE
GERENCIA REGIONAL DE SALUD LAMBAYEQUE
HOSPITAL "LAS MERCEDES" - CHICLAYO
Dr. Pirilo Junior Muro Solano
DIRECTOR EJECUTIVO
CMP. 64281 - RNE 32154

GOBIERNO REGIONAL LAMBAYEQUE
GERENCIA REGIONAL DE SALUD
HOSP. REG. DOC. "LAS MERCEDES"
Mag. Isabel G. Lizarraga de
C.S.P. 4078
JEFA DE LA UNIDAD DE APOYO A LA DOCENCIA E INVESTIGACIÓN

Anexo 3. Instrumentos de recolección de datos.

Guía de entrevista con la especialista en hematología	
Aspectos previos	
Objetivo: Establecer criterios de inclusión y exclusión de los datos a recolectar. Entrevistado(a): Dra. Zarela Ivonne Lamas Ramirez Especialidad: Hematología. Ambiente: Presencial. Fecha: 20/09/2021. Hora: 11:30 A.M.	
Preguntas	Respuestas
¿Qué es la leucemia?	La leucemia es un Cáncer de la sangre que se caracteriza por el aumento anormal y desordenado del número de leucocitos, lo que da lugar a una invasión de la medula ósea e impide a su vez el desarrollo normal de las células progenitoras de la sangre.
¿Cuáles son los tipos de leucemia más frecuentes en la infancia?	Tenemos la leucemia linfoblástica aguda (LLA) en un 70% y la leucemia mieloide aguda (LMA) con un 30% de casos.
¿Cuáles son los síntomas y signos que presenta un paciente con leucemia?	Un paciente con alta sospecha de leucemia manifiesta tener pérdida de peso, sudoración nocturna, fiebre, hiporexia, aparición de equimosis y petequias, cansancio, fatiga, dolor óseo, epistaxis, presencia de adenopatías. Cabe señalar que además cursa con alteración en su analítica del hemograma, en donde empieza la principal sospecha.
¿Qué exámenes se realiza para hacer el diagnostico de leucemia?	Se deben realizar los siguientes estudios: Una citometría de flujo, un aspirado de medula ósea, una biopsia de medula ósea, estudios de panel molecular para leucemia linfática y mieloide, asimismo estudios de citogenética.
¿Qué otros exámenes complementarios se llevan a cabo para el estudio de la leucemia?	Está sujeto del debut de la enfermedad y como se llegue a presentar, teniendo así el estudio de imágenes: Ecografía abdominal, Tomografía de tórax, abdomen y pelvis sin contraste, biopsia de las adenopatías con alta sospecha de malignidad.

Tabla 20.

Registro de resultados del modelo Regresión logística

Registro de resultados Regresión logística		
División del conjunto de datos	Entrenamiento 80%, prueba 20%	
Tiempo de respuesta del algoritmo:	0.05 segundos	
Matriz de confusión:	VN = 1	FP = 1
	FN = 1	VP = 13
Exactitud:	93.3%	
Precisión:	92.9%	
Sensibilidad:	100.0%	
F1 Score:	96.3%	

Fuente: Elaboración propia.

Tabla 21.

Registro de resultados del modelo Árboles de decisión

Registro de resultados Árboles de decisión		
División del conjunto de datos	Entrenamiento 80%, prueba 20%	
Tiempo de respuesta del algoritmo:	0.02 segundos	
Matriz de confusión:	VN = 2	FP = 0
	FN = 0	VP = 13
Exactitud:	100.0%	
Precisión:	100.0%	
Sensibilidad:	100.0%	
F1 Score:	100.0%	

Fuente: Elaboración propia.

Anexo 4. Código fuente de la implementación de los algoritmos

Tabla 22.

Implementación del algoritmo Regresión logística

Código

```
#Importar las librerías
import pandas as pd
import numpy as np
#Importar el dataset
dataf = pd.read_excel("LeucemiaInfantil.xlsx")
dataf.head()
#Importar la función
from sklearn.linear_model import LogisticRegression
logisticRegression = LogisticRegression(random_state = 20)
logisticRegression.fit(E_train, P_train)
logisticRegression.score(E_train,P_train)
#Prueba del modelo con el Paciente 5, diagnóstico = LLA
pred_1 = logisticRegression.predict([paciente_5])
pred_proba_1 = logisticRegression.predict_proba([paciente_5])
print("Diagnóstico: ", pred_1)
print("Probabilidad LLA: ", pred_proba_1[0][1])
print("Probabilidad LMA: ", pred_proba_1[0][0])
Diagnóstico:  [1]
Probabilidad LLA:  0.9973756502856075
Probabilidad LMA:  0.0026243497143925154
#Prueba del modelo con el Paciente 8, diagnóstico = LMA
pred_2 = logisticRegression.predict([paciente_8])
pred_proba_2 = logisticRegression.predict_proba([paciente_8])
print("Diagnóstico: ", pred_2)
print("Probabilidad LLA: ", pred_proba_2[0][1])
print("Probabilidad LMA: ", pred_proba_2[0][0])
Diagnóstico:  [0]
Probabilidad LLA:  0.0697282266888823
Probabilidad LMA:  0.9302717733111177
```

Fuente: Elaboración propia

Tabla 23.

Implementación del algoritmo Árboles de decisión

Código

```
#Importar las librerías
import pandas as pd
import numpy as np
#Importar el dataset
dataf = pd.read_excel("LeucemiaInfantil.xlsx")
dataf.head()
#Importar la función
from sklearn.tree import DecisionTreeClassifier
decisionTreeClassifier = DecisionTreeClassifier(criterion = "entropy",
max_depth=3,random_state = 20)
decisionTreeClassifier.fit(E_train, P_train)
#Prueba del modelo con el Paciente 5, diagnóstico = LLA
pred_1 = decisionTreeClassifier.predict([paciente_5])
pred_proba_1 = decisionTreeClassifier.predict_proba([paciente_5])
print("Diagnóstico: ", pred_1)
print("Probabilidad LLA: ", pred_proba_1[0][1])
print("Probabilidad LMA: ", pred_proba_1[0][0])
Diagnóstico:  [1]
Probabilidad LLA:  1.0
Probabilidad LMA:  0.0
#Prueba del modelo con el Paciente 8, diagnóstico = LMA
pred_2 = decisionTreeClassifier.predict([paciente_8])
pred_proba_2 = decisionTreeClassifier.predict_proba([paciente_8])
print("Diagnóstico: ", pred_2)
print("Probabilidad LLA: ", pred_proba_2[0][1])
print("Probabilidad LMA: ", pred_proba_2[0][0])
Diagnóstico:  [0]
Probabilidad LLA:  0.0
Probabilidad LMA:  1.0
```

Fuente: Elaboración propia