PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

CARRERA

INGENIERÍA EN SISTEMAS DE INFORMACIÓN

Trabajo de integración curricular

Tema: Aplicación de Técnicas de Ciencia de Datos en Resultados de Biometría Hemática obtenidos de un Laboratorio Clínico de la ciudad de Quito

AUTOR:

Paúl Alexander Campaña García

DIRECTOR

Miguel Ortiz Navarrete Mtr.

DEDICATORIA

Dedico este trabajo de integración curricular a mis padres Paúl y Vanessa, han sido el motor que me han permitido seguir adelante sin rendirme; gracias a ustedes he conseguido superar las adversidades y lograr muchas cosas a lo largo de mi vida.

Por ello, dedico este trabajo a ustedes que me criaron y guiaron por el camino del bien, permitiéndome ser la persona en la que hoy me he convertido.

ÍNDICE DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

RESUMEN

El presente trabajo de integración curricular se enfoca en el análisis de los resultados del examen de biometría hemática realizados a pacientes de un laboratorio clínico en Quito; la infomación obtenida del sistema especializado de laboratorio clínico fue anonimizada para garantizar la privacidad de los pacientes.

En el primer capítulo se evidencian los objetivos, la problemática y el alcance que tendrá este trabajo de integración curricular.

En el segundo capítulo se presenta todo el marco teórico que sustenta la ciencia de datos, las metodologías cómo: CRISP DM, MLOps y de investigación aplicada; así como los modelos de: agrupamiento, clasificación y redes neuronales simples.

En el tercer capítulo se describe el estado del arte de este trabajo de integración curricular, especificando el proceso metodológico que se siguió para la consecución de los objetivos planteados.

En el cuarto capítulo se especifican las actividades y los resultados obtenidos en cada una de las fases establecidas por la MLOps; definida cómo la metodología base para este proyecto.

Finalmente se presentan conclusiones, recomendaciones y bibliografía de este trabajo de integración curricular.

1. CAPÍTULO 1: INTRODUCCIÓN

1.1 Introducción

En el campo de la medicina y la salud, la aplicación de herramientas tecnológicas ha tenido gran impacto en la forma en la que se puede entender y solucionar los problemas médicos de los pacientes. Entre estas herramientas, los exámenes de laboratorio se destacan como una solución que ofrece un punto de vista más detallada del estado del paciente, lo que permite a los profesionales médicos reconocer, diagnosticar y dar seguimiento a una amplia variedad de malestares y padecimientos de forma óptima.

El desarrollo tecnológico ha llevado al desarrollo de distintas soluciones de analítica de datos progresivamente más refinados, lo que va a permitir el procesamiento y análisis de grandes repositorios de información de forma eficiente y sistémica. Entre estas soluciones, se encuentran sistemas de análisis hematológico, que permiten recopilar los datos de las distintas muestras sanguíneas de los pacientes, así también se tienen soluciones informáticas expertas en el almacenamiento, procesamiento e interpretación de datos, que permitirán facilitar la toma de decisiones al personal médico.

El siguiente proyecto de integración curricular propone investigar y reconocer el impacto de las herramientas de análisis de datos de resultados de exámenes de laboratorio (biometrías hemáticas) en la ciudad de Quito, con la finalidad de apoyar al personal médico en las tomas de decisiones respecto al diagnóstico de enfermedades. A lo largo de este estudio, se busca identificar las distintas tendencias crecientes y la mejor metodología de análisis de datos para datos hematológicos, para lograr contribuir de mejor manera al avance médico basado en investigación y a la optimización de diagnósticos.

1.2 Planteamiento del problema

El análisis de datos basado en resultados de laboratorio de biometrías hemáticas en la ciudad de Quito presenta cierta cantidad de desafíos que podrían afectar su adecuada interpretación.

Para empezar, la complejidad de los datos obtenidos de resultados hemáticos implica la clasificación y análisis de la variedad de parámetros que incluye, como por ejemplo la concentración de glóbulos rojos que tiene el paciente, lo que permitirá analizar de manera detallada la relación entre estas variables y las distintas afecciones que el paciente puede tener. La diversidad de variables genera conjuntos de datos con mayor complejidad en el análisis e interpretación de estos.

Por otro lado, a pesar de que hoy en día se cuenta con distintas soluciones tecnológicas especializadas, el entendimiento de los resultados de análisis hemáticos va a depender del criterio y conocimiento del profesional de la salud, debido a que únicamente son herramientas de apoyo. Produce subjetividad en distintos diagnósticos y decisiones dependiendo del personal médico.

Además, la variabilidad particular de cada paciente producida por distintos parámetros físicos como la edad, el género o la estatura puede influir en los resultados de las biometrías hemáticas. Dificulta la estandarización de los datos, debido a que se pueden producir valores inusualmente altos o bajos en los valores. Requiriendo así un mayor nivel de preprocesado de los datos.

Las dificultades presentadas permiten señalar el impacto de investigar y aplicar técnicas de analítica de datos para mejorar la interpretación de resultados de laboratorio de biometrías hemáticas. Con la finalidad de apoyar al personal médico a mejorar la interpretación y generación de diagnósticos.

1.3 **Objetivos**

1.3.1 Objetivo General

Aplicar técnicas de ciencia de datos con la finalidad de identificar patrones, clasificarlos, relacionarlos o asociarlos, para aportar con información inteligente en el diagnóstico de enfermedades relacionadas a una biometría hemática de un laboratorio de la ciudad de Quito.

1.3.2 Objetivos Específicos

1. Analizar las metodologías CRISP DM y MLOps para identificar la metodología más adecuada para el análisis de los resultados de exámenes de biometría hemática.
2. Aplicar el proceso ETL para adecuar y garantizar la estructura y calidad de los datos obtenidos de exámenes de biometría hemática de un laboratorio de la ciudad de Quito.
3. Construir los modelos de análisis de datos más adecuados a la data obtenida y estructurada para este proyecto.
4. Evaluar los modelos elaborados para avalar los resultados obtenidos

1.4 **Justificación**

El análisis de datos procedente de los resultados de exámenes de biometría hemática tiene un alto impacto y vital importancia dentro del área de la medicina y la salud, ya que permite conocer información específica y amplia acerca del funcionamiento de la sangre y su respectiva estructura, facilitando la interpretación y generación de diagnósticos sobre las distintas afecciones derivadas de esta. Por esta razón, un análisis de datos de resultados de exámenes de biometría hemática en un laboratorio de la ciudad de Quito es justificable por motivos de alto impacto.

Para empezar, la importancia para el personal médico de un examen de biometría hemática se basa en su utilidad para proporcionar indicadores e información detallada de la salud del paciente. Dicha información es utilizada por los profesionales médicos para monitorear la existencia de irregularidades en la composición sanguínea del paciente, además es de utilidad para diagnosticar y evaluar la presencia de afecciones tanto leves o graves, además va a permitir realizar un seguimiento del avance del paciente sobre el tratamiento establecido.

Por otro lado, en el escenario de un constante desarrollo de soluciones tecnológicas para distintas ramas, el análisis de datos de exámenes de laboratorio de biometrías hemáticas tiene ventajas gracias al desarrollos de herramientas y metodologías más refinadas. Mediante la aplicación de software dedicado al análisis y metodologías de ciencias de datos se ha optimizado

la forma en la que se podía extraer información para convertirla en conocimiento, esto nos permite identificar patrones de comportamiento, tendencias estadísticas, entre otros.

Para finalizar, la importancia del siguiente proyecto de integración curricular se encuentra en la generación de conocimiento por medio de la evidencia estadística, facilitando la interpretación y comprensión de la información del paciente. De esta manera, se busca que el personal docente pueda optimizar la generación de diagnósticos y el seguimiento de tratamientos a raíz de este.

En resumen, el siguiente proyecto busca profundizar en las distintas herramientas y metodologías de ciencias de datos, con la finalidad de contribuir en una mejor toma de decisiones clínicas y mejores prácticas del personal médico.

1.5 **Alcance**

El presente trabajo de integración curricular tiene como objetivo principal la aplicación de tecnologías de análisis de datos a los resultados de exámenes de biometría hemática en un laboratorio de la ciudad de Quito.

Se busca entender de forma detallada la interrelación entre las distintas variables presentes en estos resultados dentro del contexto de diversas afecciones. Esto permitirá desarrollar un análisis de datos que facilite la identificación de afecciones médicas y contribuya a una toma de decisiones referente a diagnósticos clínicos, centrándose exclusivamente en la información derivada de los resultados de exámenes de laboratorio de biometría hemática.

Por otro lado, se busca e identificar la aplicación adecuada de metodologías de ciencias de datos con el fin de elegir aquella que se acomode de mejor manera, estas metodologías permitirán explorar patrones, tendencias y correlaciones en los datos hematológicos. Con la finalidad de facilitar la toma de decisiones al personal médico.

En resumen, el presente trabajo de integración curricular busca aplicar tecnologías de análisis de datos en el contexto específico de las biometrías hemáticas, con la meta de mejorar la comprensión y el manejo de enfermedades a través de un enfoque más preciso.

2. CAPÍTULO 2: MARCO TEÓRICO

2.1 **Metodología de Investigación**

Conocer la metodología de investigación del presente trabajo de integración curricular es de suma relevancia para el entendimiento y ejecución de un adecuado análisis de datos de resultados de biometrías hemáticas. En este capítulo se busca profundizar la teoría de las distintas metodologías, lo que permitirá destacar un enfoque investigativo, por lo que se usarán técnicas y metodologías afines al tipo de estudio planteado anteriormente. Por otro lado, otra finalidad del capítulo es entregar un punto de vista general del seguimiento y observación de un proceso investigativo, estableciendo una base metodológica lógica que garantice obtener resultados óptimos respecto a los objetivos planteados.

Resulta importante para una comprensión adecuada del proyecto definir el concepto de metodología de investigación. (Alban et al., 2020) afirma que “Los métodos de investigación localizan y delimitan un problema, permiten recolectar datos importantes para generar hipótesis que posteriormente sean probadas o respaldadas.” Mediante este concepto, se infiere que una adecuada aplicación de metodologías de investigación, además de una recolección óptima de datos de valor, se llega a la obtención de resultados óptimos y deseados en relación con los objetivos definidos con anterioridad.

Dentro del presente trabajo de integración curricular se usarán dos metodologías distintas: la metodología descriptiva y la metodología aplicada. La elección de estas metodologías se debe a la naturaleza de la investigación abordada en este estudio.

2.1.1 Metodología descriptiva

Inicialmente, el presente trabajo de integración curricular utilizará una metodología descriptiva, esto debido a que se desea recopilar información y documentarse adecuadamente respecto a los resultados obtenidos por medio de exámenes de biometrías hemáticas en un laboratorio de Quito.

Respecto a dicho tipo de investigación, (Alban et al., 2020) nos dice que “La información suministrada por la investigación descriptiva debe ser verídica, precisa y sistemática”. Entendiendo que este tipo de investigación debe ofrecer información detallada de lo que se desea estudiar, no se realizará ninguna inferencia debido a que rompe con los principios de la aplicación de esta metodología.

Para implementarla existen múltiples herramientas que facilitarán su uso. Sobre dicho tema, (Alban et al., 2020) dice que “Los métodos de recolección de datos empleados son la observación, encuesta y estudio de casos.” Siendo muy útiles para aplicar adecuadamente la metodología, en el presente trabajo de integración curricular se documentará adecuadamente para ampliar el conocimiento respecto a biometrías hemáticas con la información obtenida por medio de los datos del sistema de información del laboratorio.

Aplicando metodología descriptiva se espera mejorar el entendimiento de los resultados obtenidos en biometrías hemáticas en un laboratorio de la ciudad de Quito, lo que permitirá aplicar adecuadamente metodologías de ciencias de datos y conseguir un resultado final acorde a los objetivos establecidos.

2.1.2 Metodología aplicada

La metodología aplicada permite usar el conocimiento adquirido con la investigación descriptiva en ambientes más realistas. Dentro del presente proyecto, se usarán los conceptos adquiridos para aplicar metodologías y técnicas de ciencias de datos.

Para comprender de forma óptima dicha metodología de investigación, (Tres, 2008) nos dice que “La investigación aplicada busca el conocer para hacer, para actuar, para construir, para modificar.” La define como una metodología que permite actuar gracias al conocimiento adquirido, debido a que de esta manera se evita la experimentación sin guía previa.

Se busca utilizar la metodología para aplicar los conceptos abordados dentro del marco teórico, con la finalidad de obtener resultados acordes a los parámetros definidos.

2.2 **Ciencia de datos**

Dentro del presente capítulo se busca entender la importancia de la ciencia de datos dentro del caso de estudio, con la finalidad de reconocer patrones de comportamiento en los datos entregados. También se busca realizar una comparativa entre metodologías de ciencias de datos y encontrar la mejor adaptada a la problemática.

(Lemus-Delgado & Pérez Navarro, 2020) afirman lo siguiente:

> La ciencia de datos se compone de tres áreas. La primera es el big data, que se emplea para procesar los datos. La segunda es la minería de datos, cuya finalidad es encontrar patrones, incluso sin que estos fueran antes imaginados. Por último, la visualización de los datos, cuyo propósito es facilitar la comprensión de la información de manera clara y propiciar su socialización.

Define a la ciencia de datos como una disciplina que puede ser tratada desde múltiples perspectivas, logrando así un análisis completo de los resultados en biometrías hemáticas usando técnicas que faciliten su manipulación.

2.2.1 Metodologías de ciencias de datos

Aplicar metodologías de ciencia de datos dentro del caso de estudio permitirá la observación y seguimiento de la aplicación de conceptos y técnicas de ciencias de datos. Permitirá el acercamiento a los resultados esperados por medio del seguimiento idóneo de pasos metodológicos.

Dentro del presente trabajo de integración curricular se buscará realizar un análisis comparativo de las metodologías CRISP DM y MLOPS, con la finalidad de encontrar la más idónea al caso de estudio.

2.2.1.1 Metodología CRISP DM

(Moine et al., 2011) afirma que “CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining.” Se establece que se trata de una guía referencial y

orienta al lector respecto al desarrollo de proyectos de minería de datos por medio de una serie de fases.

(Moine et al., 2011) respecto a estas fases nos dice que "La sucesión de fases, no es necesariamente rígida. Cada fase es descompuesta en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, pero en ningún momento se propone como realizarlas." Las fases establecidas dentro de la guía referencial no disponen de un orden obligatorio, pero es recomendable seguirlas en ese orden. Por otro lado, dichas fases están sujetas a la visión del lector y es quien define como aplicarlas.

(IBM, 2002) define las siguientes fases dentro de la guía referencial CRISP DM.

2.2.1.1.1 Comprensión del negocio

Es el paso preliminar e inicial de la guía referencial CRISP DM. Sobre la comprensión del negocio IBM (2002) define que "Se debe dedicar tiempo a explorar las expectativas de su organización con respecto a la minería de datos." Se analiza el contexto en el que están definidos los datos, de esta manera la comprensión de las variables y sus relaciones serán más sencillas, brindando una visión más general del caso de estudio.

Determinar la visión de la empresa en relación a la minería de datos llega a ser tardío si no se tiene una orientación y documentación adecuada. (IBM, 2002) define varios puntos para determinar objetivos comerciales:

- Inicie la recolección de datos sobre el estado actual de la actividad comercial.
- Documente los objetivos comerciales específicos establecidos por los directivos.
- Acuerde los criterios que se emplearán para evaluar el rendimiento del proceso de minería de datos desde una óptica comercial.

2.2.1.1.2 Comprensión de los datos

En esta fase se realiza una comprensión de los datos de forma más técnica y específica. Respecto a este paso (IBM, 2002) nos dice que "Implica estudiar más de cerca los datos disponibles de minería. Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos) que suele ser la fase más larga de un proyecto." Analizar los datos es importante para explorar los distintos problemas que pueden ocasionar estos datos en futuras fases, conocer el estado de los datos permitirá realizar un preprocesamiento completo.

(IBM, 2002) proporciona una lista de tareas a cumplir:

- Recolección inicial de datos.
- Descripción de los datos.
- Exploración de los datos.
- Evaluación de la calidad de los datos.

2.2.1.1.3 Preparación de los datos

La preparación de los datos es clave, ya que de esta manera se evitará que el conocimiento adquirido posteriormente por el proceso de ciencia de datos sea inexacto. Respecto a esta fase, (IBM, 2002) presenta una serie de tareas que la preparación de datos suele abordar. Algunas de estas son:

- Integración de conjuntos de datos y/o registro de datos.
- Elección de una muestra de datos de un subconjunto.
- Agrupación de registros.
- Generación de nuevos atributos.
- Categorización de los datos para su modelado.
- Tratamiento de valores faltantes o nulos mediante eliminación o imputación.

- Separación en conjuntos de datos de entrenamiento y prueba.

Luego de seleccionar y aplicar las tareas que el caso de estudio requiera, se puede continuar con el análisis y generación de modelos a partir de los datos refinados.

2.2.1.1.4 Modelado

En esta fase principalmente se buscará aplicar distintas técnicas y algoritmos con la finalidad de conseguir conocimiento. Respecto a esta fase, (IBM, 2002) nos dice que "Los analistas de datos ejecutan varios modelos utilizando los parámetros predeterminados y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias por su modelo." Por medio de varios modelos de ciencia de datos que varían entre lógica y funcionalidad, se busca llegar a resultados que permitan aportar a la toma de decisiones del usuario final.

(IBM, 2002) define las siguientes tareas para la fase de modelado:

- Elección de técnicas de modelado.
- Desarrollo de un plan de validación.
- Construcción de los modelos.

2.2.1.1.5 Evaluación

En esta fase se aplican las técnicas y métricas de evaluación definidas dentro de cada modelo con la finalidad de medir su rendimiento. (IBM, 2002) resalta la importancia de "evaluar los resultados de sus esfuerzos utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto." Alinear los criterios de rendimiento inicial con los resultados y conocimiento obtenidos por el modelo resulta fundamental, de esta manera el usuario final puede tomar decisiones con mayor exactitud.

(IBM, 2002) define las siguientes tareas en la fase de evaluación:

- Documentar la evaluación para determinar si los resultados de la minería de datos satisfacen los criterios de desempeño empresarial.
- Enumerar los modelos aprobados.

2.2.1.1.6 Despliegue

En esta fase se entregan los modelos resultantes al usuario final, con la finalidad de cumplir los requerimientos definidos. Respecto a esta fase, (IBM, 2002) nos dice que "Es el proceso que consiste en utilizar sus nuevos conocimientos para implementar las mejoras en su organización." Define esta fase como la implementación de los modelos aprobados dentro del ambiente del usuario final.

(IBM, 2002) define las siguientes tareas en la fase de despliegue:

- Elaboración de estrategias para implementar el sistema.
- Organización de la supervisión y mantenimiento continuo.
- Preparación del informe final.

2.2.1.2 Metodología MLOPS

(Google Cloud, 2023) define a MLOps como:

MLOps es una práctica y cultura de la ingeniería de AA, cuyo fin es unificar el desarrollo (Dev) y las operaciones (Ops) del sistema de AA. La práctica de MLOps implica abogar por la automatización y la supervisión en todos los pasos de la construcción del sistema de AA, incluida la integración, las pruebas, el lanzamiento, la implementación y la administración de la infraestructura.

Define a MLOps cómo el conjunto de prácticas y recomendaciones con la finalidad de optimizar el proceso de desarrollo de soluciones de aprendizaje automático. Se enfoca en la

optimización de los distintos procesos de aprendizaje automático, con la finalidad de supervisar los procesos de construcción y evaluación de modelos.

(Google Cloud, 2023) presenta un ciclo de vida basado en niveles a seguir para desarrollar un proyecto de aprendizaje automático.

2.2.1.2.1 Nivel 0: Proceso manual

(Google Cloud, 2023) nos dice que "Muchos equipos tienen investigadores de aprendizaje automático y científicos de datos que pueden compilar modelos de vanguardia, pero su proceso para compilar y, luego, implementar modelos de aprendizaje automático es manual en su totalidad." El proceso de desarrollo de modelos para aprendizaje automático es manual, por lo que en este nivel no se presenta automatización de ningún tipo.

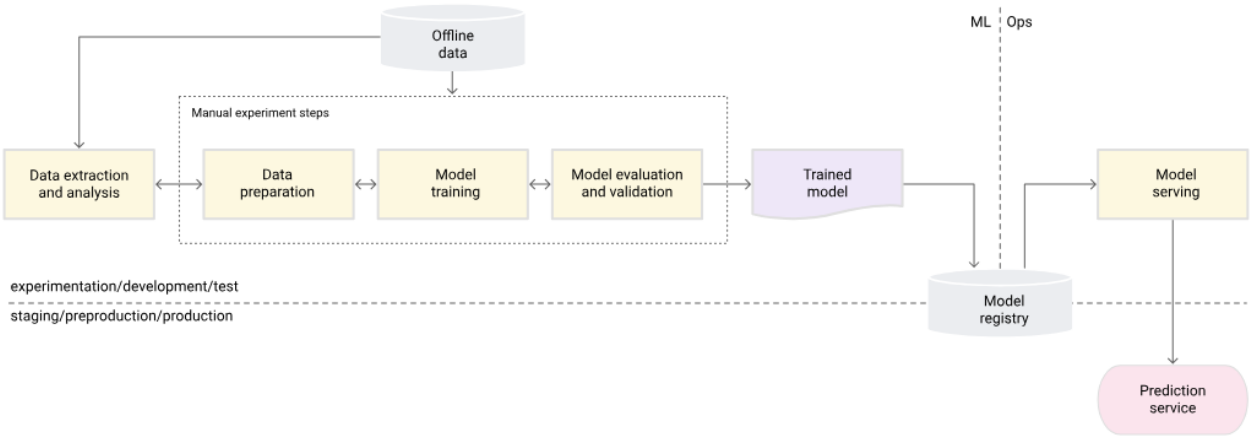

Ilustración 1: Esquema de nivel 0.
Fuente: Google Cloud, 2023

En la ilustración 1 se muestra el esquema de nivel 0 de MLOps. Cada uno de los pasos y sus transiciones son manuales, por medio de comandos y ejecución de código, los científicos de datos siguen la serie de pasos definidas hasta terminar con un modelo factible.

2.2.1.2.2 Nivel 1: Automatización de canalización de aprendizaje automático

(Google Cloud, 2023) nos dice que “El objetivo del nivel 1 es realizar un entrenamiento continuo del modelo mediante la automatización de la canalización de aprendizaje automático, lo que te permite lograr una entrega continua del servicio de predicción del modelo.” La información resultante de los modelos será entregada de manera automatizada, permitiendo que el usuario final disponga de resultados con mayor velocidad.

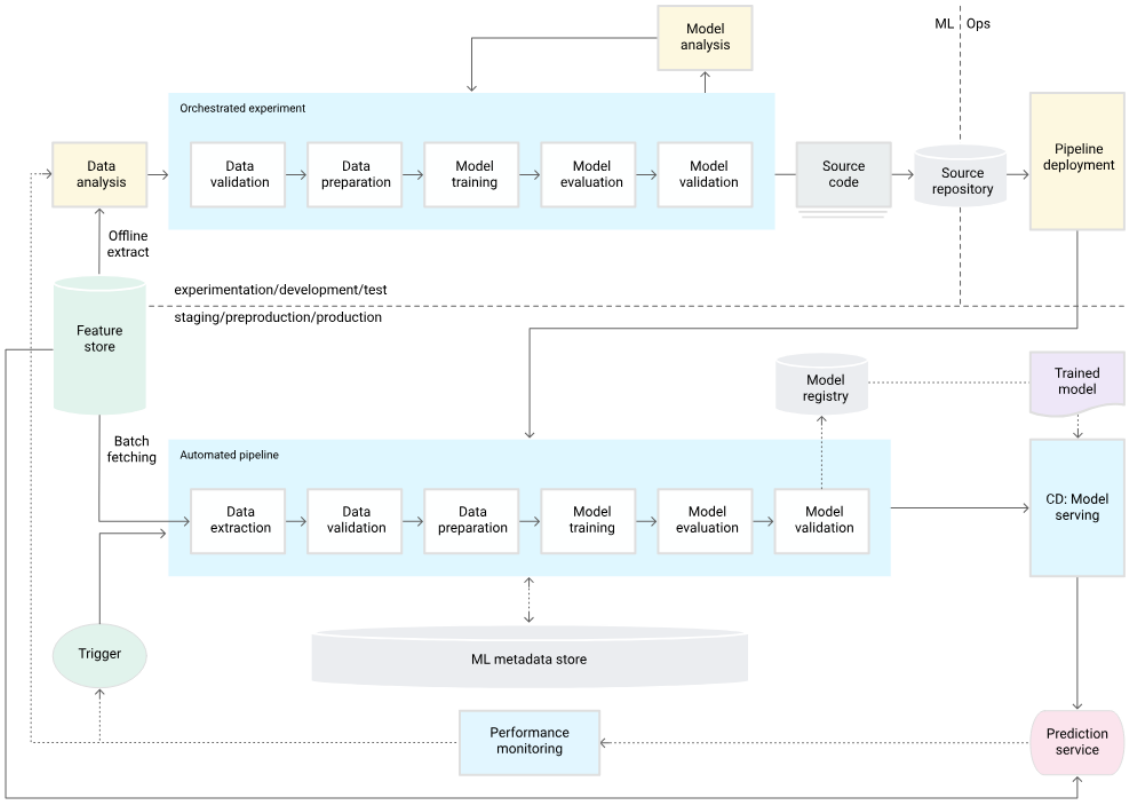

Ilustración 2: Esquema de nivel 1 de MLOps.
Fuente: Google Cloud, 2023

En la ilustración 2 se muestra el nivel 1 de MLOps. En este nivel los pasos están más organizados, además de que la línea de pasos del proceso de generación de modelos de aprendizaje automático está automatizada.

Además de que se divide dos canales principales: El canal experimental y el canal automatizado, permitiendo mucho más orden y una sucesión de pasos clara.

2.2.1.2.3 Nivel 2: Automatización de la canalización de CI/CD

(Google Cloud, 2023) especifica que “Permite que los científicos de datos exploren con rapidez ideas nuevas en torno a la ingeniería de atributos, la arquitectura de modelos y los hiperparámetros.” Por medio de atributos, características o hiperparámetros de los modelos definidos, se puede lograr personalización de las soluciones de aprendizaje automático. Esto permite dar mayor precisión a los resultados finales.

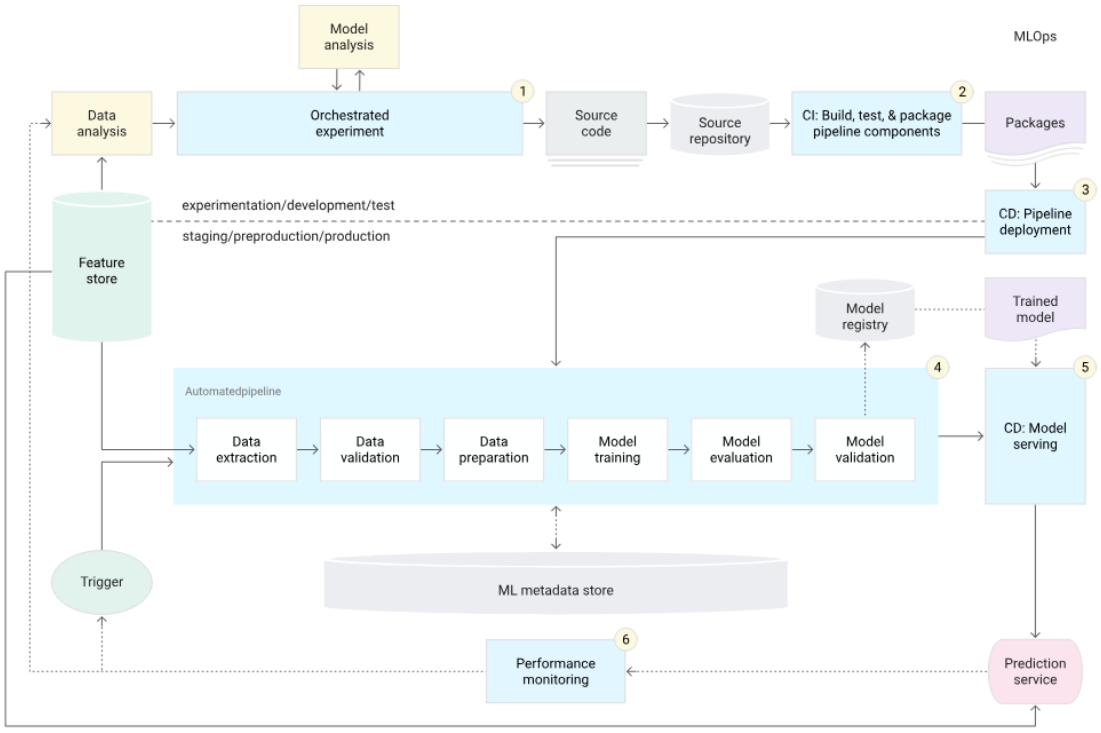

Ilustración 3: Nivel 3 de MLOps.
Fuente: Google Cloud, 2023

En la ilustración 3 se presenta el nivel 2 de MLOps. La automatización de los canales de aprendizaje automático aumenta, en esta etapa se presenta una estructura de implementación de modelos de aprendizaje automático desarrollando 6 nuevas etapas con la finalidad de automatizar y personalizar el modelo. El cual empieza con un modelo experimental y finaliza con el monitoreo de rendimiento.

2.2.1.2.4 Pasos de la ciencia de datos para el aprendizaje automático

Dentro de la metodología MLOPS se define una serie de pasos para su aplicación en el proceso de implementación de modelos de aprendizajes automáticos. (Google Cloud, 2023) define los pasos que serán definidos a continuación.

2.2.1.2.4.1 Extracción de datos

Acerca de este paso, (Google Cloud, 2023) afirma que "Selecciona y, luego, integra los datos relevantes de fuentes de datos para la tarea de aprendizaje automático." Define a la extracción de datos como el proceso de elección de fuentes y la construcción de una arquitectura adecuada de los datos, con la finalidad de que sea simple y entendible.

2.2.1.2.4.2 Análisis de datos

En este paso, se busca entender y documentar el contexto de dónde se originan los datos, además de analizar su estructura. Respecto a esta fase, Google Cloud (2023) nos dice que "Busca la compresión de las características y el esquema de datos que espera el modelo." Dar seguimiento a la estructura de los datos y los distintos parámetros a analizar, nos va a permitir identificar variables y sus relaciones.

(Google Cloud, 2023) define las siguientes tareas para la fase de análisis de datos:

- Entender las características y la estructura de los datos requeridos por el modelo.
- Reconocer las acciones necesarias de preparación y mejora de datos para satisfacer las necesidades del modelo.

2.2.1.2.4.3 Preparación de los datos

El objetivo del paso es aplicar técnicas de limpieza y preprocesamiento a los datos obtenidos en fases anteriores. (Google Cloud, 2023) nos dice que “Esta preparación incluye la limpieza de datos, en la que se dividen los datos en conjuntos de entrenamiento, validación y pruebas.” Se preparan los datos con la finalidad de obtener un formato adecuado que acepten los modelos.

2.2.1.2.4.4 Entrenamiento de modelos

En este paso se utilizan los datos con la finalidad de conseguir conocimiento nuevo, esto se consigue utilizando distintos modelos de ciencia de datos. (Google Cloud, 2023) nos dice que “el científico de datos implementa algoritmos diferentes con los datos preparados para entrenar varios modelos de AA.” Probar distintos modelos va a permitir la comparación de resultados, de esta manera se elegirá aquel que nos permita acercarnos más al resultado óptimo.

2.2.1.2.4.5 Evaluación de modelos

En este paso se realizarán evaluaciones mediante métricas específicas para cada modelo, buscando comprender si es adecuado para el caso de estudio o, por otro lado, el enfoque que se busca es incorrecto. (Google Cloud, 2023) nos dice que “El modelo se analiza en un conjunto de pruebas de exclusión para evaluar la calidad del modelo.” Cada modelo se debe analizar de forma independiente y únicamente usando los valores de prueba definidos en la fase de preparación de los datos.

2.2.1.2.4.6 Validación de modelos

En este paso se tiene como objetivo concluir si los modelos son adecuados o no para el caso de estudio definido. (Google Cloud, 2023) nos dice que “Se confirma que el modelo es adecuado para la implementación si su rendimiento predictivo es mejor que un modelo de referencia determinado.” Define una comparativa entre un marco de referencia manual, como

puede ser documentación tradicional y el modelo predictivo que entregamos, buscando concluir si el modelo de ciencia de datos es más eficiente que la solución anterior.

2.2.1.2.4.7 Entrega del modelo

En este paso se entrega el modelo ya revisado al usuario final. Google Cloud (2023) nos dice que "Se implementa el modelo validado en un entorno de destino a fin de entregar predicciones." De esta manera, el usuario final ya puede utilizar el modelo para conseguir nuevo conocimiento.

(Google Cloud, 2023) define implementaciones del modelo revisado:

- Microservicios con una APIs de tipo REST que proporcionan predicciones en tiempo real.
- Un modelo integrado en un dispositivo móvil.
- Integración en un sistema de predicción a través de procesamiento por lotes.

2.2.1.2.4.8 Supervisión del modelo

Este es el paso final de la metodología MLOPS, se realizan auditorías a la calidad de la implementación, debido a que pueden existir modificaciones o correcciones de posibles valores inexactos. (Google Cloud, 2023) nos dice que "Se supervisa el rendimiento predictivo del modelo para invocar, de manera potencial, una iteración nueva en el proceso de AA." Se pueden realizar múltiples iteraciones de la metodología MLOps con el objetivo de realizar cambios en los modelos de aprendizaje automático.

2.2.2 Análisis comparativo de las metodologías de ciencias de datos

Comparar los marcos referenciales CRISP DM y MLOps un punto fundamental de este capítulo. De esta manera, se puede seleccionar aquella metodología que se adapte al caso de estudio.

CRISP DM es un marco de referencia basado en etapas que permite gestionar adecuadamente proyectos de minería de datos. Además de esto, dispone de 6 fases metodológicas iniciando desde el entendimiento del negocio y finalizando en el despliegue del resultado del proceso.

Por otro lado, MLOps es un marco de referencia que dispone de niveles que tienen como finalidad automatizar el despliegue de los modelos de aprendizaje automático. MLOps define pasos a seguir para dar seguimiento a la generación, evaluación y supervisión de modelos de aprendizaje automático.

Las principales diferencias de ambos marcos referenciales se presentan en la estructura de fases. En las primeras fases de la metodología CRISP DM se separa la obtención de los datos en dos fases iniciales siendo el entendimiento del negocio y el entendimiento de los datos. Por otro lado, MLOps condensa dichas tareas en una fase denominada extracción de los datos. En el caso de la fase de preparación de los datos, en ambas metodologías disponen del mismo nombre.

CRISP DM define una fase de modelado dónde se definen tareas para diseño e implementación de modelos de ciencia de datos. En el caso de MLOps, es definida una fase de entrenamiento de modelos.

Otra diferencia se presenta en la fase de evaluación en CRISP DM, en dónde se analiza el rendimiento de los modelos generados. MlOps define dos fases que cumplen la tarea de evaluar el rendimiento de los modelos generados, siendo: Evaluación de modelos y validación de modelos.

CRISP DM define una fase de despliegue en dónde se entrega el modelo final a un ambiente realista. MLOps comprende una fase de entrega del modelo, además de definir una fase final en dónde se supervisa el rendimiento de dicho modelo.

2.2.3 Modelos de ciencias de datos

Los modelos de ciencias de datos permiten que los datos extraídos y refinados con anterioridad puedan ser usado para conseguir nueva información. Existen modelos con diferente lógica y finalidad.

Resulta importante presentar un concepto concreto y adecuado respecto a los modelos de ciencias de datos. Asencios (2004) nos dice que los modelos de ciencia de datos “entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos”. Por medio de múltiples técnicas de aprendizaje automático se puede identificar relaciones entre las variables de los datos extraídos y generar nuevo conocimiento.

Dentro del presente trabajo de integración curricular se compararán modelos de: clasificación, predicción y agrupamiento.

2.2.3.1 Modelos de clasificación

Asencios (2004) respecto a los modelos de clasificación nos dice que “Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto redefinido de clases)”. Su objetivo es separar cada uno de los registros de los datos en clases, por medio de los valores y relaciones entre las variables.

2.2.3.1.1 Árboles de decisión

(Martínez et al., 2009)afirma que “Un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas” Basado en la información adquirida en base a los datos de entrenamiento ingresado, el modelo aprende gradualmente los patrones que relacionan las variables.

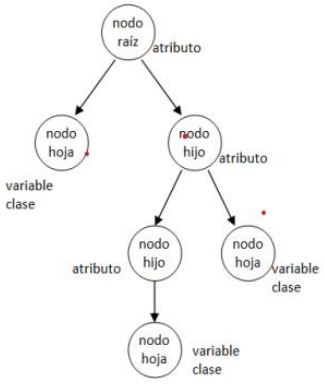

Ilustración 4: Estructura de un árbol de decisión.
Fuente: Martínez et al., 2009

En la ilustración 4 se presenta la estructura de un árbol de decisión, cada uno de los posibles casos se encuentran representados en los nodos hijos. La decisión que concluye en la clasificación del registro, se visualiza en los nodos hojas. De esta manera, los nodos hojas representan todas las clasificaciones del conjunto de datos.

2.2.3.2 Modelos de agrupamiento

Asencios (2004) respecto a los modelos de predicción nos dice que "Hace corresponder cada caso a una clase, con la peculiaridad de que las clases se obtienen directamente de los datos de entrada utilizando medidas de similaridad." Tienen como finalidad agrupar conjuntos de datos basados en características similares entre sí, separando estas en distintos grupos cuyo comportamiento se define por los valores de sus variables.

2.2.3.2.1 K-Means

(Pascual et al., 2007) nos dice que:

> La idea principal es definir k centroides (uno para cada grupo) y luego tomar cada punto de la base de datos y situarlo en la clase de su centroide más cercano. El próximo paso es recalcular el centroide de cada grupo y volver a distribuir todos los objetos según el centroide más cercano. El proceso se repite hasta que ya no hay cambio en los grupos de un paso al siguiente

De esta manera se van agrupando los datos en distintas clases. Los centroides pueden ser modificados por el usuario, es importante entender la cantidad de grupos de datos que consideremos necesarios para lograr un buen resultado.

2.2.3.2.2 DBSCAN

Se trata de un algoritmo de agrupamiento basado en la densidad. (Pascual et al., 2007) nos dice que "se definen los conceptos de punto central (puntos que tienen en su vecindad una cantidad de puntos mayor o igual que un umbral especificado), borde y ruido." Estas variables

permiten modificar valores para generar más grupos basados en la cantidad de datos que planeemos asignar en cada grupo, en este caso la cantidad de grupos que generará al final dicho algoritmo.

(Pascual et al., 2007) respecto al funcionamiento de DBSCAN, afirma que:

> El algoritmo comienza seleccionando un punto p arbitrario, si p es un punto central, se comienza a construir un grupo y se ubican en su grupo todos los objetos denso-alcanzables desde p. Si p no es un punto central se visita otro objeto del conjunto de datos. El proceso continúa hasta que todos los objetos han sido procesados. Los puntos que quedan fuera de los grupos formados se llaman puntos ruido, los puntos que no son ni ruido ni centrales se llaman puntos bordes.

De esta manera DBSCAN puede generar grupos a partir de los parámetros definidos en el algoritmo. De esta manera se agruparán los datos en los grupos que el algoritmo considere necesario y no finalizará hasta que termine de procesar todos los datos.

2.2.3.3 Redes neuronales artificiales

(Izaurieta & Saavedra, 2000) nos dice que "una neurored es un procesador de información, de distribución altamente paralela, constituido por muchas unidades sencillas de procesamiento llamadas neuronas." Una red neuronal artificial está definida por varias estructuras de procesamiento independientes, se organizan de tal manera que permiten generar información de salida de la red, a partir de datos de entrada. Se trata de un modelo tanto de predicción como de clasificación

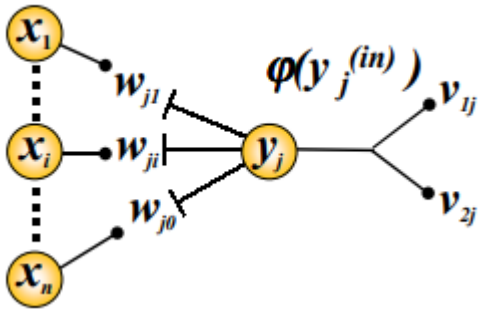

Ilustración 5: Esquema de una neurona.
Fuente: Izaurieta & Saavedra, 2000

En la ilustración 5 se representa un esquema estructural de una neurona artificial. Por medio de valores x de entrada, se puede realizar operaciones matemáticas definidas por una función y, cuyo valor procesado da un resultado v. El conjunto de dichas neuronas interconectadas mediante niveles se conoce cómo red neuronal.

3. CAPÍTULO 3: MARCO METODOLÓGICO

3.1 Materiales

Los materiales utilizados en este proyecto de integración curricular están definidos en base a los requerimientos del mismo. Se usarán los siguientes materiales:

3.1.1. SQL Server

Es un sistema de administración de bases de datos, en dicho sistema se almacena la información del laboratorio clínico.

3.1.2. Sistema especializado del laboratorio clínico

Se trata de un sistema especializado en auditar y administrar las tareas principales del laboratorio clínico, cuya información está almacenada en un repositorio de SQL Server.

3.1.3. Oracle Database

Sistema de gestión de base de datos robusta, en dónde se almacenarán los datos del repositorio del sistema especializado del laboratorio clínico en la plataforma SQL Server. Además de esto, facilitará el procesamiento de datos requerido para aplicar un formato adecuado para los modelos de ciencias de datos. La distribución de Oracle Database que se usará es XE, ya que se trata de una distribución gratuita.

3.1.4. Linux-Centos

Sistema operativo de código abierto y distribuido por Linux, destacada por su facilidad de uso y estabilidad en ejecución de tareas. Dentro de este sistema operativo se instalará Oracle XE para el almacenamiento y limpieza de los datos. Por otro lado, también se usará dicho sistema operativo para ejecutar los distintos modelos de ciencias de datos.

3.1.5. Python

Lenguaje de programación caracterizado por su sintaxis simple y la amplia variedad de bibliotecas que posee. Se usará para la implementación y ejecución de los modelos de ciencia de datos.

3.1.6. PL-SQL

Lenguaje de programación que nos permite realizar procedimientos en los cuales intervengan sentencias SQL de una o varias tablas de una base de datos relacional.

3.1.7. Scikit-learn

Librería de modelos de aprendizaje automático implementada en el lenguaje de programación python. Permite automatizar y simplificar el proceso de implementación de los modelos.

3.2 Metodología

Para el caso de estudio definido en puntos anteriores, se usará la metodología MLOps, la cual brinda un enfoque estructurado para el desarrollo e implementación de proyectos de minería de datos y aprendizaje automático.

De acuerdo a los pasos establecidos por el marco de referencia MLOps, se inicia con la extracción de los datos. Se extraerán los datos del repositorio en SQL Server, que están conectados a la aplicación del laboratorio clínico. Se extraerán datos de los últimos 4 años para posteriormente realizar el análisis y limpieza de los datos.

En el segundo paso, MLOps define el proceso de Análisis de datos. Se debe entender de dónde provienen los datos, en este caso se consultará a médicos de diversas especialidades para aclarar y conceptualizar las variables definidas dentro de los datos obtenidos. Además de esto, se deberá documentar y conceptualizar cada una de las variables extraídas para tener un enfoque más amplio de los requerimientos definidos.

En el tercer paso, MLOps define el proceso de Preparación de los datos. Se realizarán procesos de limpieza de datos en el ambiente de ORACLE XE. Tales cómo: Eliminación de datos inconsistentes, corrección de valores nulos o vacíos, creación de nuevas variables, separación en datos de entrenamiento y prueba, y eliminación de outliers.

En el cuarto paso, MLOps define el proceso de Modelado. Se implementarán distintos modelos de ciencia de datos, tales como: Modelos de agrupamiento, modelos de clasificación y modelos de predicción en el lenguaje de programación Python. Con el objetivo de alimentarlos con los

En el quinto paso, MLOps define el proceso de Entrenamiento de Modelos. Dados los distintos modelos implementados en el lenguaje de programación Python en el paso anterior, se debe entrenar con los datos de entrenamiento, que inicialmente supondrán el 70% de los datos totales.

En el sexto paso, MLOps define el proceso de Evaluación de Modelos. Con los modelos previamente entrenados, se implementarán parámetros de evaluación matemáticos, que permitan analizar a detalle el rendimiento de cada modelo. Dichos parámetros de evaluación están definidos en la documentación de cada modelo.

En el séptimo paso, MLOps define el proceso de Validación de Modelos. Con los resultados obtenidos en el paso anterior, se separará aquellos modelos que permitan acercarnos a los objetivos definidos, además de realizar rectificaciones que mejoren el rendimiento de los modelos elegidos.

Si bien el marco referencial MLOps comprende hasta los pasos restantes de Entrega de Modelos y Supervisión de Modelos, dentro de este trabajo de integración académica solo se llegará a las previamente definidas.

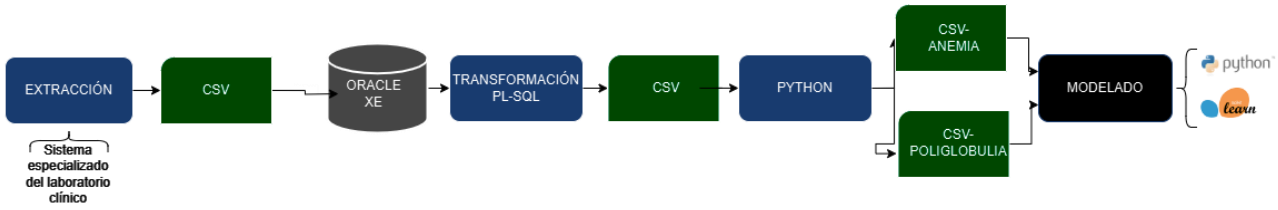

Ilustración 6: Diagrama de flujo de datos del proyecto.
Fuente: autor del documento

En la ilustración 6 se observa el diagrama de flujo de datos del proyecto. Inicialmente se realiza la extracción de los datos del sistema especializado del laboratorio clínico, el cual está almacenado en una base de datos SQL Server.

Dicho sistema exporta sus registros en formato *.csv* y se procede a la fase de transformación o preparación de datos, se almacenan los registros en una base de datos ORACLE XE y por medio de algoritmos en los lenguajes de programación PL-SQL y Python se generará un nuevo *.csv* con los datos limpios.

Posteriormente se realiza el diseño de los modelos de ciencia de datos en el lenguaje de programación Python y haciendo uso de la librería scikit-learn.

4. CAPÍTULO 4: RESULTADOS

4.1. Extracción de los datos

Se selecciona la fuente de extracción de los datos, en este caso de estudio la fuente de los datos es una base de datos relacional administrada por el gestor de bases de datos SQL Server, dicha base de datos almacena los registros de resultados históricos desde 2021 a 2023 de análisis de biometrías hemáticas del sistema especializado del laboratorio clínico.

El sistema del laboratorio clínico dispone de una funcionalidad de extracción de registros en formato CSV. Se va a extraer dicha información para su posterior transformación en siguientes fases.

	A	B	C	D	E	F	G	H	I
1	orden	edad	sexo	fecha_ingreso	examen	resultado	fecha_resultado	servicio	observacion
2	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	HEMATOCRITO: 49.3	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
3	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	HEMOGLOBINA: 15.3	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
4	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	PLAQUETAS: 150	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
5	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	GLOBULOS BLANCOS: 7.85	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
6	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	NEUTROFILOS: 5.19	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
7	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	LINFOCITOS: 1.98	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
8	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	NEUTROFILOS %: 66.2	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
9	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	LINFOCITOS %: 25.1	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
10	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	RECUENTO DE GLOBULOS ROJOS: 5.19	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
11	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	VOL. CORPUSCULAR MEDIO: 95.1	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
12	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	HGB CORPUSCULAR MEDIA: 29.6	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
13	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	CONC. HGB CORPUSCULAR MEDIA: 31.1	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
14	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	RDW CV: 12.8	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
15	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	MID %: 8.70	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
16	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	MID: 0.68	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
17	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	MPV: 11.4	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
18	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	PDW: 17.9	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
19	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	PCT: 0.171	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019
20	72981	57 AÑOS	F	02/01/2019	BIOMETRIA HEMATICA	RDW SD: 51	03/01/2019	LABORATORIO CLINICO	F. consulta: 02/01/2019

Ilustración 7: Registro original de biometría hemática.
Fuente: Sistema especializado de laboratorio clínica, 2024

En la ilustración 7 se muestra una serie de registros en formato .csv, cada uno de estos datos corresponden a una misma orden o un mismo paciente. Cada una de las variables de la biometría hemática está separada en distintos registros. Dentro de la variable resultado se incluye el nombre de dicha variable y el valor que corresponde a esta, además de definir fechas, género y el examen dado en el laboratorio.

4.2. Análisis de los datos

Se busca entender los datos extraídos en la fase anterior para manipularlos con criterio y de forma adecuada. Resulta importante contextualizar acerca del negocio del cual se extrajeron los datos.

4.2.1. Laboratorios clínicos

Acerca de los laboratorios clínicos, nos dice que (Instituto Europeo de Química, 2023) “es un tipo de instalación con finalidades médicas, pues se analizan muestras y se realizan pruebas para contribuir en el diagnóstico, tratamiento y prevención de enfermedades.” Se define como un lugar utilizado para realizar análisis médico con todo tipo de pruebas, con la finalidad de facilitar el diagnóstico médico.

(Ministerio de Salud Pública de Ecuador, 2012) define la siguiente tipificación de los laboratorios clínicos: Laboratorio clínico general y Laboratorio clínico especializado

4.2.1.1. Laboratorio clínico general

Respecto a la tipificación de laboratorio clínico general, (Ministerio de Salud Pública de Ecuador, 2012) afirma que:

> Es aquel servicio de salud al que le compete analizar cualitativa y cuantitativamente muestras biológicas, provenientes de individuos sanos o enfermos, que incluya las siguientes áreas básicas de baja complejidad: hematología, bioquímica, inmunología, uroanálisis y coproanálisis.

Cumple con la generalidad del análisis de muestras con el objetivo de facilitar el diagnóstico médico, aunque en este caso se dedica únicamente a análisis generales sin ningún tipo de especialización en algún área médica.

4.2.1.2. Laboratorio clínico especializado

Según (Ministerio de Salud Pública de Ecuador, 2012) un laboratorio clínico especializado:

> Es aquel servicio de salud en el que se realizan análisis clínicos generales de baja complejidad y especializados en una o más áreas de mediana o alta complejidad en: hematología, bioquímica, inmunología, uroanálisis y coproanálisis; microbiología, biología molecular, toxicología y genética.

Este tipo de laboratorio clínico es más específico, busca encontrar elementos exactos que un examen general no podría encontrar. Esto permite tener más detalle de las afecciones del paciente.

4.2.2. Biometrías hemáticas

Es un tipo de prueba basada en el análisis de una muestra de sangre. (Gaona, 2003) afirma que:

> Es el primer examen al que se enfrenta el clínico en la valoración diagnóstica de un paciente, y aunque se considera como un solo examen de laboratorio, en realidad, valora el estudio de tres líneas celulares, cada una con funciones diferentes entre sí, pero que tienen en común que las produce la médula ósea: eritrocitos, leucocitos y plaquetas.

Es definido como un examen cuyo estudio sanguíneo se divide en tres ramas siendo los elementos más principales de la estructura de la sangre. A raíz de conocer estas tres ramas y las relaciones que tienen entre sí, se pueden observar patrones de comportamiento de afecciones médicas.

Dentro de la estructura de los datos tenemos 19 variables que corresponden a estos 3 parámetros. En base al cuestionamiento que se realizó a los médicos de distintas especialidades se definieron rangos aceptables dependiendo de la edad y género del paciente.

4.2.2.1. Hematocrito

(Gaona, 2003) afirma que “El hematocrito es la porción de volumen total de la sangre ocupada por la masa de eritrocitos o glóbulos rojos; representa, entonces, el porcentaje de la masa de eritrocitos en la sangre total” Define que los hematocritos es la cantidad de glóbulos rojos existentes en el torrente sanguíneo.

4.2.2.2. Hemoglobina

(Gaona, 2003) nos dice que “La Hemoglobina, componente principal de los eritrocitos, representa el 32 % de la masa total del glóbulo rojo y es el mejor índice para medir la capacidad

de transporte de gases de la sangre” Es un componente que forma parte de la estructura de los glóbulos rojos y se encarga de mover los gases principales en todo el torrente sanguíneo.

4.2.2.3. Plaquetas

(National Cancer Institute, 2024) respecto al concepto de plaquetas, nos dice que “son fragmentos de células muy grandes de la médula ósea que se llaman megacariocitos. Ayudan a producir coágulos sanguíneos para hacer más lento el sangrado o frenarlo y para facilitar la cicatrización de las heridas.” Es la tercera variable del conjunto de datos y una de las tres líneas celulares estudiadas en las biometrías hemáticas. Se tratan de partes de célula proveniente del material óseo que permite generar coágulos, así evitando sangrado en las heridas.

4.2.2.4. Glóbulos blancos

(National Cancer Institute, 2024) afirma que los glóbulos blancos son un:

> Tipo de glóbulo sanguíneo (célula de la sangre) que se produce en la médula ósea y se encuentra en la sangre y el tejido linfático. Los glóbulos blancos son parte del sistema inmunitario del cuerpo y ayudan a combatir infecciones y otras enfermedades.

Este tipo de glóbulo se encuentra en mucha menos cantidad en el torrente sanguíneo comparado a los glóbulos rojos. Pero su importancia radica a la defensa del sistema inmunitario contra múltiples afecciones.

4.2.2.5. Neutrófilos

(National Cancer Institute, 2024) respecto a los neutrófilos, afirma que son un “tipo de glóbulo blanco (célula sanguínea) que cumple una función importante en el sistema inmunitario y ayuda a combatir las infecciones en el cuerpo. Son una de las primeras células inmunitarias que reaccionan cuando entran al cuerpo microorganismos.” Se encuentran muy presentes en el torrente sanguíneo, ya que son la primera línea de defensa ante virus y bacterias.

4.2.2.6. Linfocitos

(National Human Genome Research Institute, 2024)afirma que el linfocito “es un tipo de glóbulo blanco que es parte del sistema inmune”. Se trata de otra variante de glóbulos blancos que permiten proteger el sistema inmunitario del individuo, se encuentra en menos presencia que los neutrófilos. Existe tipificación de linfocitos, (National Human Genome Research Institute, 2024) nos dice que:

> Hay dos tipos principales de linfocitos: las células B y las células T. Las células B elaboran los anticuerpos para luchar contra bacterias, virus y toxinas invasoras. Las células T destruyen las propias células del cuerpo que han sido infectadas por virus o que se han vuelto cancerosas.

Dicha tipificación cumple distintas funciones. Los linfocitos B, generan anticuerpos que refuerzan las defensas del sistema inmune. Por otro lado, los linfocitos T eliminan aquellas células que el resto de glóbulos blancos no han sido capaces de proteger, son de gran importancia para evitar la dispersión de enfermedades.

4.2.2.7. Porcentaje de neutrófilos

Define la cantidad de neutrófilos dado un volumen de sangre. Está dado en formato porcentual y se trata de la cantidad de neutrófilos presente en la sangre.

4.2.2.8. Porcentaje de linfocitos

Se trata del porcentaje de linfocitos B y linfocitos T en el torrente sanguíneo. Se calcula la sumatoria de ambas células de linfocito para calcular el porcentaje presente en el volumen sanguíneo.

4.2.2.9. Conteo de glóbulos rojos

Los glóbulos rojos o eritrocitos son una de las líneas celulares principales de estudio dentro de una biometría hemática. (National Cancer Institute, 2024) define a los glóbulos rojos cómo “Tipo de glóbulo sanguíneo (célula de la sangre) que se produce en la médula ósea y se encuentra en la sangre. Contienen una proteína llamada hemoglobina, que transporta oxígeno desde los pulmones a todas las partes del cuerpo.” Esta célula sanguínea tiene como principal función el transporte de distintos gases en el torrente sanguíneo.

(UNAM, 2012) afirma que “el recuento de eritrocitos o glóbulos rojos, consiste en contar el número de ellos en un milímetro cúbico de sangre”. Calcula la cantidad de glóbulos rojos que existen en un volumen específico de torrente sanguíneo.

El conteo de glóbulos rojos nos permite detectar múltiples enfermedades. (National Cancer Institute, 2024)menciona que “se usa para determinar la presencia de afecciones como la anemia, la deshidratación, la desnutrición y la leucemia.” La importancia de dicho conteo radica en la detección y prevención de dichas enfermedades o afecciones.

4.2.2.10. Volumen corpuscular medio

(MedlinePlus, 2022) afirma que “La prueba de sangre de VCM mide el tamaño promedio de los glóbulos rojos.” Esto nos permite conocer las capacidades de transporte de gases que puede tener el torrente sanguíneo del paciente. Esto va a ayudar a determinar distintas enfermedades, respecto a esto (MedlinePlus, 2022) nos dice que “Una prueba de VCM puede ayudar a diagnosticar qué tipo de anemia tiene.” Además de determinar si el paciente tiene anemia, el VCM permite tipificar dicha patología.

4.2.2.11. Hemoglobina corpuscular media

(MedlinePlus, 2022) nos dice que la hemoglobina corpuscular media “mide la cantidad promedio de hemoglobina en un solo glóbulo rojo”. Es una variable que se encarga de calcular el promedio de hemoglobina en una unidad de eritrocito, esto permite medir la capacidad de mover gases del mismo.

4.2.2.12. Concentración de hemoglobina corpuscular media

(MedlinePlus, 2022) afirma que la concentración de hemoglobina corpuscular media “también mide la hemoglobina en los glóbulos rojos. Además, incluye un cálculo del tamaño y el volumen de los glóbulos rojos”. Se trata de una variable muy parecida a la anterior, debido a que calcula la hemoglobina en un volumen definido de torrente sanguíneo.

4.2.2.13. RDW CV

(MedlinePlus, 2022) afirma que “la prueba de amplitud de distribución eritrocitaria (RDW, por sus siglas en inglés) es un análisis que mide la variación en el volumen y el tamaño de los glóbulos rojos (eritrocitos)”. También denominado coeficiente de variación de distribución eritrocitaria nos permite conocer el espacio y densidad que ocupan los glóbulos rojos dentro del torrente sanguíneo.

(MedlinePlus, 2022) afirma que dicha variable se usa para “ayudar a diagnosticar la anemia, una enfermedad en la que los glóbulos rojos no pueden llevar suficiente oxígeno al resto del cuerpo”. Esto va a permitir el diagnóstico de anemia y otras afecciones como poliglobulia, junto al resto de variables mencionadas anteriormente.

4.2.2.14. MID

(MedlinePlus, 2022) nos dice que las células MID o también llamada fórmula leucocitaria es “un análisis de sangre que mide la cantidad de cada tipo de glóbulo blanco que hay en el cuerpo.” Dentro de dicho análisis se contemplan glóbulos blancos de tamaño y características similares. Se listan los siguientes tipos de glóbulos blancos: Monocitos, eosinófilos y basófilos.

4.2.2.15. Porcentaje de células MID

Se realiza un conteo acumulativo de los tres tipos de glóbulos blancos incluidos en células MID, esto con la finalidad de definir un valor porcentual de dichas células por volumen sanguíneo.

4.2.2.16. MPV

(MedlinePlus, 2022) afirma que “una prueba de sangre de VPM mide el tamaño promedio de las plaquetas”. Se trata de la cantidad de plaquetas promedio que tiene el individuo dado un volumen sanguíneo, conociendo la capacidad de coagulación del individuo.

El VPM nos va a permitir detectar múltiples enfermedades, (MedlinePlus, 2022) nos dice que “puede ayudar a diagnosticar enfermedades hemorrágicas y de la médula ósea”. Facilita la búsqueda de lesiones internas del paciente.

4.2.2.17. PDW

(MedlinePlus, 2022) nos dice que “el ancho de la distribución plaquetaria hace referencia a la variabilidad del tamaño de las plaquetas, es decir la dispersión del tamaño de la plaqueta con respecto al volumen plaquetario medio”. Esta variabilidad en el tamaño de las plaquetas nos permite la detección de posibles anomalías y afecciones.

(MedlinePlus, 2022) afirma sobre PDW, que es importante gracias a que “en el caso de las enfermedades con trombocitosis asociadas a procesos mieloproliferativos y trombocitopenias autoinmunes, empleándose como un elemento más de diagnóstico”. Permitiendo el diagnóstico de dichas enfermedades cuando sus valores varían en gran medida relacionados a los resultados obtenidos en la variable MPV.

4.2.2.18. PCT

(MedlinePlus, 2022) sobre el PCT o recuento plaquetario nos dice que “mide el número de plaquetas en la sangre”. Cuenta las unidades de plaquetas en un volumen específico de sangre.

Si la cantidad de plaquetas resultante del conteo plaquetario es muy alta o muy baja, permite conocer las patologías que tiene el paciente. Según (MedlinePlus, 2022) cuando el conteo de plaquetas es bajo, el paciente tiene trombocitopenia y cuando el conteo de plaquetas es alto, el paciente tiene trombocitosis.

4.2.2.19. RDW-SD

Representa el valor de desviación estándar de la variable RDW o amplitud de distribución eritrocitaria. Esto va a permitir identificar anomalías cuando dichos valores se separan de gran manera de la normalidad.

Se generarán estadísticas básicas con la finalidad de entender el estado original de los datos, esto va a permitir visualizar si la data tiene un comportamiento normal o si existe algún tipo de manipulación en ella.

	ORDEN	EDAD	SEXO	FECHA_INGRESO
count	6550.000000	6550.000000	6550.000000	6550
mean	90761.434198	41.749033	0.491603	2021-09-21 04:24:02.198473216
min	77871.000000	0.167000	0.000000	2020-01-02 00:00:00
25%	83471.250000	28.000000	0.000000	2020-12-05 00:00:00
50%	89734.500000	38.000000	0.000000	2021-08-16 00:00:00
75%	100700.750000	53.000000	1.000000	2022-07-05 00:00:00
max	104842.000000	100.000000	1.000000	2023-10-24 00:00:00
std	8664.519932	19.604923	0.499968	NaN

	HEMATOCRITO	HEMOGLOBINA	PLAQUETAS	GLOB_BLANCOS	NEUTROFILOS
count	6550.000000	6550.000000	6550.000000	6550.000000	6550.000000
mean	46.329145	14.577307	297.851908	6.653954	3.912136
min	14.900000	4.800000	26.000000	2.000000	0.370000
25%	43.000000	13.600000	249.000000	5.260000	2.740000
50%	46.400000	14.600000	292.000000	6.260000	3.500000
75%	50.100000	15.800000	338.000000	7.540000	4.500000
max	69.300000	21.100000	1044.000000	41.200000	25.040000
std	5.391739	1.723696	77.268392	2.303898	2.015657

	LINFOCITOS	...	C_HGB_CORPUSCULAR_M	RDW_CV	MID_PRC
count	6550.000000	...	6550.000000	6550.000000	6550.000000
mean	2.178374	...	31.480693	12.546840	8.600855
min	0.000000	...	25.400000	10.700000	0.000000
25%	1.710000	...	30.900000	12.000000	6.800000
50%	2.100000	...	31.400000	12.400000	8.200000
75%	2.530000	...	32.000000	12.800000	9.900000
max	33.400000	...	40.900000	23.900000	100.000000
std	0.872953	...	1.054059	0.983961	3.140614

	MID	MPV	PDW	PCT	RDW_SD
count	6550.000000	6550.000000	6550.000000	6550.000000	6550.000000
mean	0.556997	9.005237	15.944748	0.264703	48.157313
min	0.000000	6.300000	11.800000	0.026000	28.100000
25%	0.410000	8.400000	15.700000	0.228000	45.800000
50%	0.510000	8.900000	15.900000	0.261000	47.800000
75%	0.650000	9.500000	16.200000	0.297000	49.900000
max	10.650000	13.200000	18.300000	0.701000	97.700000
std	0.286478	0.865299	0.376739	0.058514	3.789900

Ilustración 8: Estadísticas descriptivas de variables del dataset.
Fuente: autor del documento.

En la ilustración 8 se observa la generación de valores de estadísticas descriptivas. Siendo estos los valores de conteo de registros, media aritmética, valores mínimos, primer cuartil, segundo cuartil o mediana, tercer cuartil, valor máximo y la desviación estándar entre las variables. Esto permite dar un vistazo general del estado base de las variables del dataset.

Con respecto a la edad, la edad promedio de los pacientes es de 41 años. Esta edad está distribuida en el primer cuartil desde menores de 1 año hasta 28 años, en el segundo cuartil desde 28 años hasta 38 años, en el tercer cuartil desde 38 años a 53 años y en el cuarto cuartil desde 53 años hasta mayores de 100 años. La desviación estándar es de 19.6 en la misma escala que la variable edad, lo que indica que hay alta variación entre los datos.

4.3.Preparación de los datos

Se preparan los datos para que los modelos de aprendizaje automático puedan procesarlos adecuadamente y conseguir generar conocimiento.

Se almacenan los registros u órdenes de biometría hemática en una base de datos en ORACLE XE, de esta manera se puede procesar la información por medio de algoritmos basados en el lenguaje de programación PL-SQL. Dicho algoritmo busca juntar todos los registros correspondientes a una misma orden en un nuevo registro que contiene en columnas la información de cada variable de biometría hemática.

Los pacientes que realizan la prueba de biometría hemática tienen un rango de edad bastante amplio, llegando a tener inclusive solo meses. Esto se visualiza dentro de la columna edad, debido a que se identifica la unidad de medida de dicha variable. Se realiza una condición, si se encuentra la palabra meses se va a dividir la edad del individuo para 12.

Se recorren todos los registros y extrae los parámetros biométricos que correspondan a una misma orden, para posteriormente colocarlo en columnas que correspondan a dichos parámetros. Cada valor después del carácter separador (:) es agregado en la columna correspondiente, por medio de condicionales basados en el término que les precede.

Para realizar modelos de clasificación y regresión es necesario definir una variable predictora. Para este proyecto de integración curricular se buscará predecir las patologías de anemia e poliglobulia. Por medio de experiencias y conocimientos de los médicos de distintas especialidades que fueron consultados, se definieron un grupo de 6 variables que son usadas para diagnosticar anemia e hipoglobulia y sus respectivos valores. Dichas variables son: Hematocrito, hemoglobina, recuento de glóbulos rojos, volumen corpuscular medio, hemoglobina corpuscular media, concentración de hemoglobina corpuscular media.

Parámetro	Rango Mujeres	Rango Hombres	Rango Niños 0-11 Años	Enfermedad
HEMATOCRITO	38-49	43-54	35-45	Menor al rango normal: Anemia. Mayor al rango normal: Poliglobulia.
HEMOGLOBINA	12.6-15.4	13.6-17.4	11.5-15.5	Menor al rango normal: Anemia. Mayor al rango normal: Poliglobulia.
RECUENTO DE GLOBULOS ROJOS	4-5.9	4.5-6.5	3.8-5.3	Menor al rango normal: Anemia. Mayor al rango normal: Poliglobulia.
VOL. CORPUSCULAR MEDIO	81-99	80-94	72-90	Menor al rango normal: Anemia. Mayor al rango normal: Poliglobulia.
HGB CORPUSCULAR MEDIA	27-31	27-31	24-32	Menor al rango normal: Anemia. Mayor al rango normal: Poliglobulia.
CONC. HGB CORPUSCULAR MEDIA	32-36	32-36	28-36	Menor al rango normal: Anemia. Mayor al rango normal: Poliglobulia.

Tabla 1: Tabla de rangos normales en variables relacionadas a anemia y poliglobulia.
Fuente: autor del documento

En la tabla 1 se puede observar los valores normales para adultos de género masculino, femenino y niños de 1 a 11 años dependiendo de la variable. Se observa de igual manera la enfermedad que tiene si el valor se sale de dicho rango superior o inferior, de esta manera se identifican las patologías que se busca predecir, basado en los reactivos usados en el laboratorio clínico.

Con los rangos de valores ya definidos se puede desarrollar una variable predictora que permita clasificar dichos valores. Por medio de un algoritmo basado en PL-SQL, se iteran los nuevos registros ya transformados y en cada una de las variables hematológicas se analizan los rangos de valores aceptados con estructuras condicionales.

Esto dará como resultado una variable clasificadora y una variable contadora para anemia y otra para poliglobulia, la primera se encarga de verificar que por lo menos una de las 6 variables tienda a dichas patologías, y la segunda se encarga de ver cuántas variables de las designadas tienden a valores fuera del rango normal. Según la experiencia de los médicos que fueron cuestionados, si el contador de la variable predictora tiene un valor de dos o más, ya puede ser considerado que el paciente dispone de una de estas dos patologías

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	ORDEN	EDAD	SEXO	FECHA_IN	HEMATOC	HEMOGLO	PLAQUETA	GLOB_BLA	NEUTROFI	LINFOCITO	NEUTROFI	LINFOCITO	GLOB_RO	VOL_CORF	HGB_CORI	C_HGB_C	RDW_CV	MID_PRC	MID	MPV	PDW	PCT	RDW_SD	SEDIMENT	ANEMIA	POLIGLOBI	NIV_CONF_	NIV_CONF_POL	
2	93537	28	0	25/2/2022	48,2	14,9	254	7,04	5,18	1,49	73,5	21,2	5,17	93,1	28,9	31	12,6	5,3	0,37	10,1	16,2	0,256	49,2		1	0	1	0	
3	93538	30	1	25/2/2022	49,8	16,2	259	6,62	3,15	2,96	47,6	44,6	5,42	91,9	30	32,6	12,7	7,8	0,51	8,1	15,5	0,209	48,9		1	0	1	0	
4	93547	51	1	25/2/2022	46,3	14,7	299	5,29	3,38	1,51	63,9	28,5	4,63	100	31,8	31,7	12,6	7,6	0,4	8,9	15,9	0,267	53,4		1	1	2	1	
5	93548	46	1	25/2/2022	49	15,8	291	7,09	3,96	2,57	55,8	36,3	5,36	91,4	29,4	32,2	12,2	7,9	0,56	8,9	16,5	0,261	47,1		1	0	1	0	
6	93549	49	0	25/2/2022	43,9	14	320	6,24	3,82	1,92	61,2	30,7	4,79	91,7	29,2	31,8	12,3	8,1	0,5	9,6	16,1	0,307	47,6		1	0	1	0	
7	93550	50	0	25/2/2022	44,6	14,5	232	7,74	4,91	2,43	63,4	31,3	4,94	90,2	29,3	32,4	12	5,3	0,4	8,2	15,9	0,191	45,8		0	0	0	0	
8	93571	37	0	2/3/2022	45,7	14,6	250	6,34	3,4	2,33	53,6	36,7	4,55	100,5	32,1	32	11,4	9,7	0,61	9,2	15,7	0,229	48,6		0	1	0	2	
9	93572	85	0	2/3/2022	38,1	12	312	4,63	3,09	1,27	66,6	27,4	4,11	92,8	29,1	31,4	12	6	0,27	6,9	15,3	0,216	47		1	0	2	0	
10	93577	13	1	2/3/2022	43,7	13,7	334	5,1	1,68	2,82	32,9	55,3	4,76	91,9	28,8	31,4	12,4	11,8	0,6	8,6	15,7	0,289	48,2		1	0	2	0	

Ilustración 9: Transformacíon de registros de biometría hemática.
Fuente: autor del documento

En la ilustración 9 se observa la transformación de los registros de biometría hemática, ahora las variables de biometría hemática están organizadas por orden, a diferencia de los registros vistos en la figura 6, dónde se repetía la orden en más de un registro. Además, se visualiza las nuevas variables predictoras generadas.

Estos nuevos registros también deben sufrir un procedimiento de limpieza ETL. Dicho procedimiento se va a realizar en el lenguaje de programación Python, por medio de un cuaderno en el ambiente de desarrollo jupyter. Empezando con la variable sedimentación, que no se tomará en cuenta debido a que no presenta valor en ningún registro.

```
ORDEN                  int64
EDAD                  object
SEXO                   int64
FECHA_INGRESO         object
HEMATOCRITO           object
HEMOGLOBINA           object
PLAQUETAS            float64
GLOB_BLANCOS          object
NEUTROFILOS           object
LINFOCITOS            object
NEUTROFILOS_PRC       object
LINFOCITOS_PRC        object
GLOB_ROJOS            object
VOL_CORPUSCULAR_M     object
HGB_CORPUSCULAR_M     object
C_HGB_CORPUSCULAR_M   object
RDW_CV                object
MID_PRC               object
MID                   object
MPV                   object
PDW                   object
PCT                   object
RDW_SD                object
ANEMIA                 int64
POLIGLOBULIA           int64
NIV_CONF_ANE           int64
NIV_CONF_POL           int64
dtype: object
```

Ilustración 10: Tipos de datos en dataset.
Fuente: autor del documento

En la ilustración 10 se observa que muchos de las variables del dataset generado no tienen un tipo de dato concreto. Esto se debe a que los registros de múltiples variables vistos en

la ilustración 8 no tienen un formato adecuado, siendo valores numéricos cuyo separador decimal es una coma en lugar de un punto. Se intercambiarán todos los puntos por comas y se le dará tipo de dato float a cada una de las variables que no disponen de un tipo de dato definido, exceptuando la variable fecha_ingreso. Esta variable únicamente será tranformado a tipo de dato datetime.

Verificar los outliers o ruido dentro del dataset es importante, ya que de esta manera los valores van a mantener una normalidad estadística y nos permite asegurar la calidad de los datos para posteriormente ser ingresados en los distintos modelos. Para esto se debe verificar si en el datase existe ruido o por el otro lado, no es necesaria la eliminación de datos.

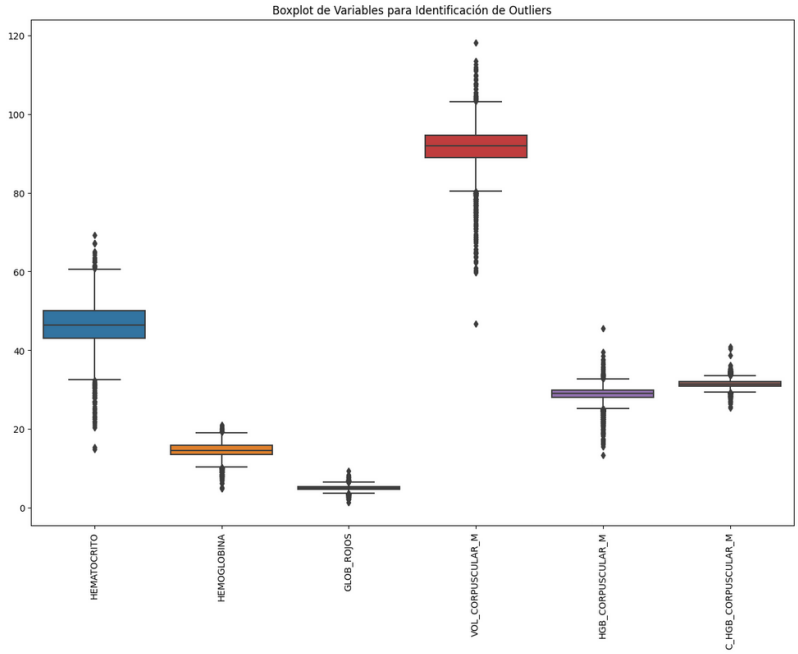

Ilustración 11: Gráfico de bigote para visualización de outliers.
Fuente: autor del documento

En la ilustración 11 se observa que existe ruido que escapa de la estructura del bigote observado en cada una de las variables, siendo el más notorio el mostrado en la variable volumen corpuscular medio. Es necesario eliminar dicho ruido para asegurar la precisión de los modelos.

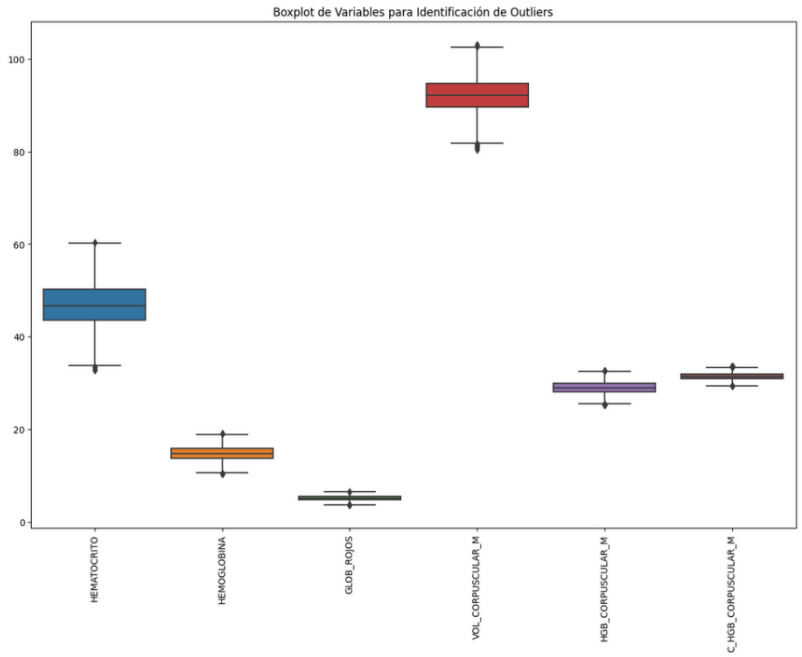

Ilustración 12: Gráfico de bigote de outliers, posterior a proceso de limpieza.
Fuente: autor del documento

En la ilustración 12 se observa el gráfico de bigote posterior al proceso de limpieza de ruido. Observamos que no existe el ruido que se observaba en la ilustración 10, la estructura de bigote en las variables está más condensada en los valores estadísticamente normales.

Posteriormente, se define una variable target para cada patología, cuya finalidad es generar un valor entre 0 o 1 basado en una condición de que la variable contadora de anemia o poliglobulia sea mayor a dos, de esta manera podemos identificar que paciente tiene cada patología.

4.4. Modelado

Es el cuarto paso de la metodología MLOps, dónde se desarrollan modelos de ciencia de datos con la finalidad de generar conocimiento que nos sirva para entender las relaciones y patrones que tiene el dataset, se implementó el coeficiente de correlación de Pearson para identificar las relaciones que podrían tener las diferentes variables del dataset.

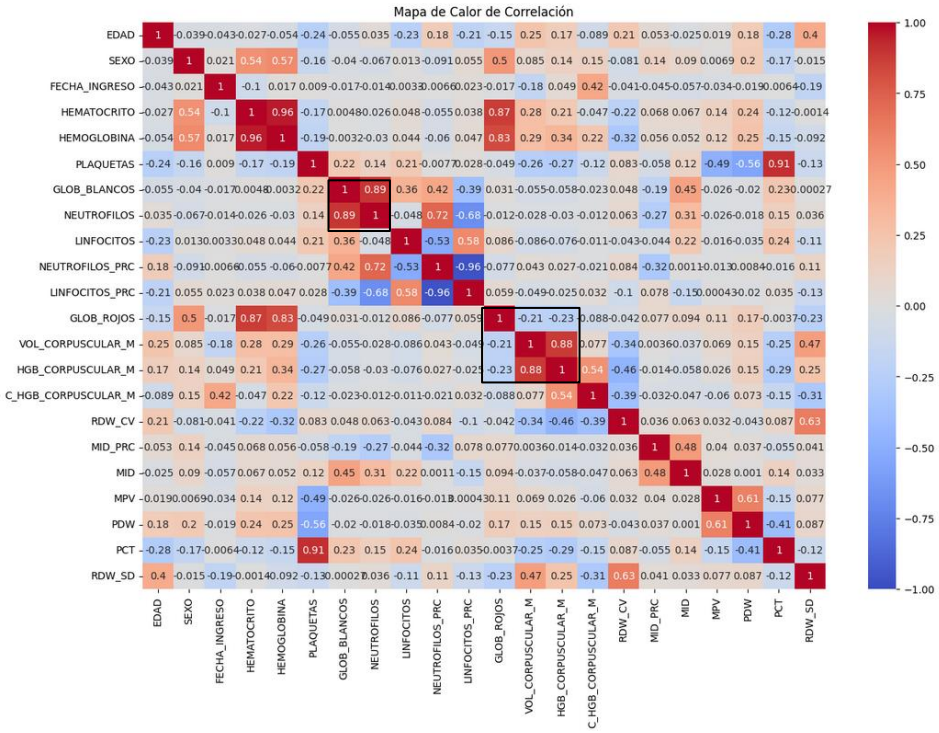

Ilustración 13: Mapa de calor de correlación entre variables del dataset.
Fuente: autor del documento

En la ilustración 13 se observa el mapa de calor de correlaciones basado en las variables del dataset. Basado en las 6 variables que son consideradas en este proyecto y que se visualizan en la tabla1, podemos identificar la alta correlación entre estas variables que garantizan su afectación tanto a anemia como a poliglobulia.

Existe alta correlación entre las variables de hematocrito y hemoglobina, dichas variables también están correlacionadas con glóbulos rojos. Por otro lado, existe correlación entre volumen corpuscular medio y hemoglobina corpuscular media.

No existen variables adicionales que afecten o aporten con alguna incidencia a las 6 variables que se relacionan a las patologías que son analizadas en este proyecto; con esto confirmamos la decisión acertada de no considerarlas.

4.4.1. Modelos de clasificación

4.4.1.1. Árbol de decisiones

El objetivo de este modelo es la clasificación de casos de anemia y poliglobuila, tanto negativos cómo positivos por medio de características guardadas en nodos del árbol. La librería scikit-learn tiene una funcionalidad de árboles de decisiones, dónde inicialmente se realizó la normalización del dataset y se separá en una proporción de 70% los datos de entrenamiento y 30 % los datos de prueba.

Para este proyecto se clasificará anemia y poliglobulia por separado. En ambos casos se definió la profundidad de los nodos del árbol de decisión en un máximo de 5.

4.4.2. Modelos de agrupamiento

4.4.2.1. K-means

El objetivo del modelo de agrupamiento k-means es reunir datos de características similares en clústeres que tienen su propio centroide basado en estos datos. La librería scikit-learn dispone de una funcionalidad para aplicar dicho modelo. Inicialmente se va a normalizar el dataset con la finalidad de que todos los datos mantengan la misma escala.

Para definir la cantidad óptima de clústeres que se usarán en el modelo, se utilizará un gráfico de codo.

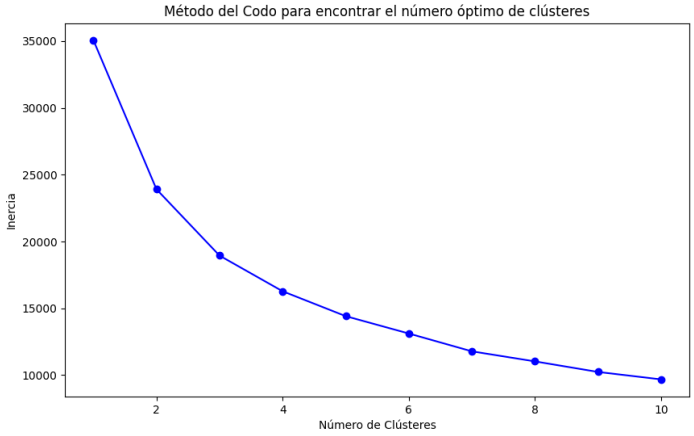

Ilustración 14: Gráfico de codo.
Fuente: autor del documento

En la ilustración 14 se observa el gráfico de codo basado en el dataset. A partir de 3 clústeres se empieza a perder inercia dentro del modelo de agrupamiento k-means, por lo que el valor óptimo que se debería usar es 3.

K-means es un modelo de agrupamiento basado en aprendizaje no supervisado, por lo que no existe fase de entrenamiento posterior. En este caso se mostrará a continuación el resultado del agrupamiento de los datos en los clústeres definidos.

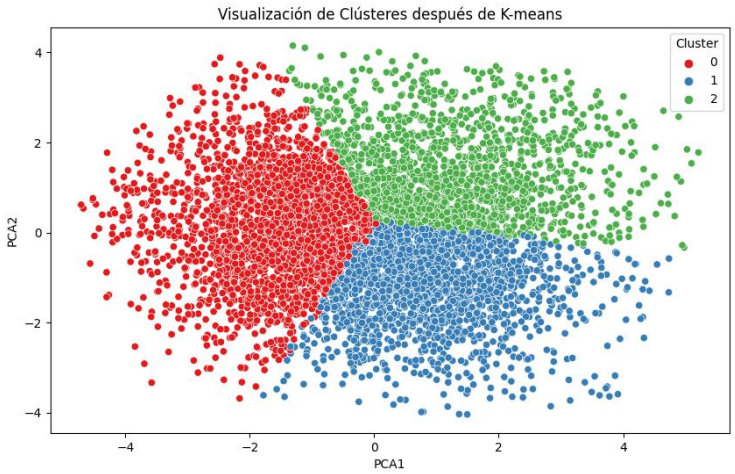

Ilustración 15: Visualización de clústeres en modelo de agrupamiento K-means.
Fuente: autor del documento

En la ilustración 15 se observan los 3 clústeres generados por el modelo. Los datos tienen un comportamiento específico, los centroides en los 3 clústeres están bastante cercanos entre sí, un centroide es el punto medio dentro de dicho clúster. Se debe a que los valores de las variables se encuentran en rangos similares, siendo aquellos que se alejan a dichos centroides los que tienden a tener anemia y poliglobulia.

4.4.2.2. DB SCAN

El objetivo del modelo de agrupamiento DB SCAN es reunir los datos de características similares en clústeres, pero a diferencia del modelo anterior no se define por un número de clústeres fijos. En este caso genera dichos clústeres en base a la profundidad que se define en el modelo, en este caso se va a definir dicha profundidad por medio de un gráfico de distancia k-ésima que nos permite visualizar la distancia de un punto del dataset con un punto cuya posición está definida por una variable k.

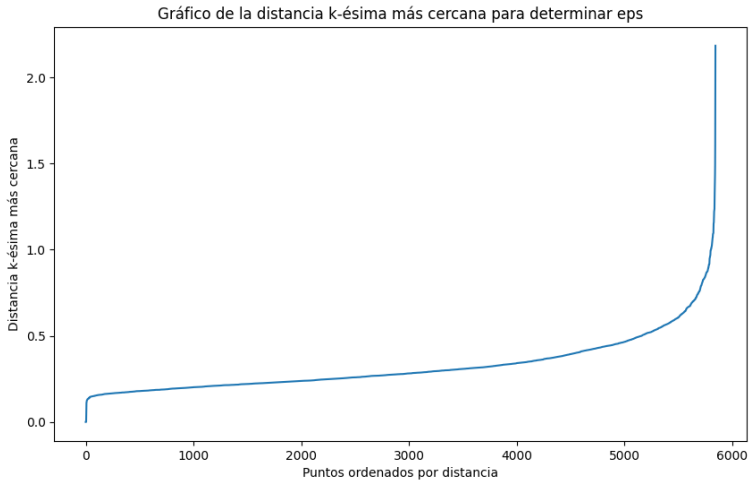

Ilustración 16: Gráfico de la distancia k-ésima.
Fuente: autor del documento

En la ilustración 16 se observa una gráfica que nos permite ver el valor de distancia k-ésima basado en relación a la cantidad de puntos ordenados por la distancia que tienen con otro punto. Se observa que dicha gráfica empieza a crecer de forma exponencial a partir del valor 0.6 de la distancia k-ésima, siendo este valor el punto óptimo para aplicar el modelo DB-SCAN.

Debido a que DB SCAN es un modelo de aprendizaje no supervisado, no requiere que dicho modelo tenga una fase de aprendizaje posterior. A continuación, se mostrarán los resultados del agrupamiento de datos usando este modelo.

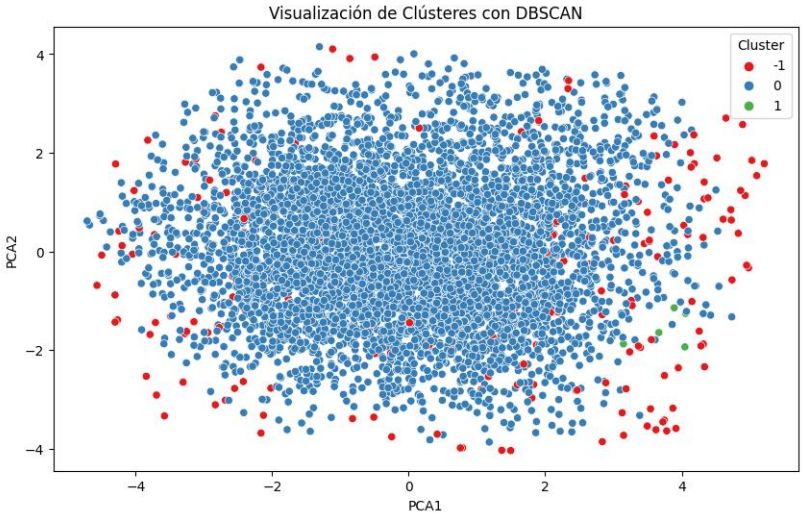

Ilustración 17: Resultado de agrupamiento de datos en DB SCAN.
Fuente: autor del documento

En la ilustración 17 se observa el agrupamiento de datos del dataset en el modelo de agrupamiento DB SCAN. Los datos están reunidos en 3 clústeres, aunque el más marcado es el clúster de color azul que corresponde a los registros que cumplen con el intervalo normal de las variables de biometría hemática. Y los otros dos clústeres que van a corresponder a los datos de anemia y poliglobulia, siendo el rojo y el verde respectivamente.

4.4.3. Redes neuronales simples

El objetivo de un modelo de red neuronal simple es clasificar los registros por medio de relaciones entre variables y patrones de comportamiento que posean. Se trata de un modelo de aprendizaje supervisado y se clasificará anemia y poliglobulia por separado.

Ambas redes neuronales inicialmente tendrán dos capas ocultas de 10 neuronas cada una para el posterior entrenamiento con el dataset. También es importante normalizar los datos y separarlos en una proporción de 70% de datos de entrenamiento y 30% de datos de prueba.

4.5.Entrenamiento de modelos

Es el quinto paso de la metodología MLOps cuyo objetivo principal es el entrenamiento de los distintos modelos y la visualización de los resultados tanto de clasificación y predicción que estos entregan.

4.5.1. Modelos de clasificación

4.5.1.1. Árboles de decisiones

En esta fase se entrenará el modelo de árbol de decisiones de la librería scikit-learn con la proporción de 70% de datos de prueba. Los resultados serán mostrados por medio de un gráfico de árbol de decisión y una gráfica de dispersión entre pares de variables.

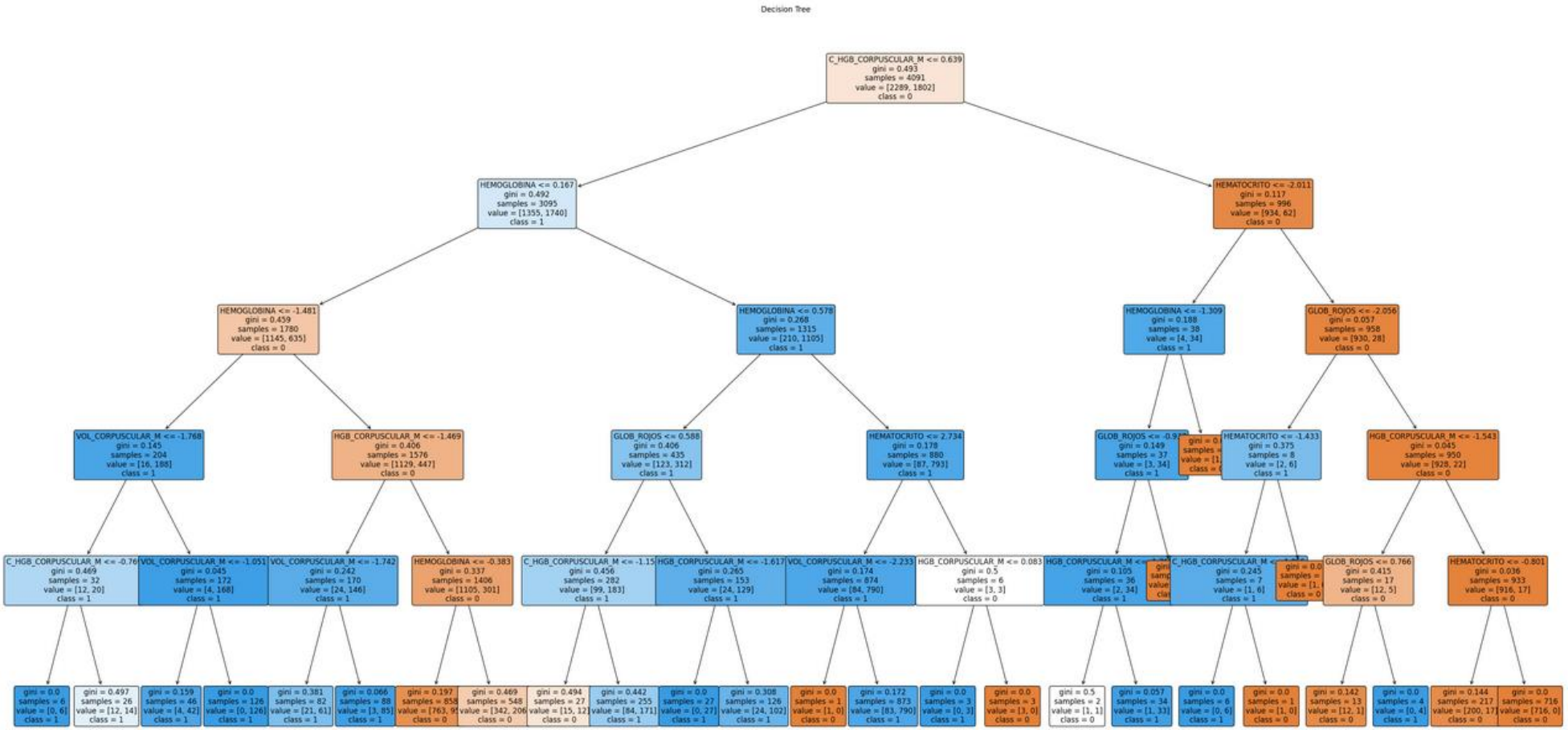

Ilustración 18: Gráfico de árbol de decisión de anemia.
Fuente: autor del documento

En la ilustración 18 se observa el árbol de decisión encargado de clasificar anemia. En cada nodo del árbol se incluyen 3 variables: La condición de ese nodo del árbol, el índice de ginis que define la pureza o precisión de dicho nodo mientras más pequeña sea y el número de ejemplos de dicho nodo.

En el nodo inicial la variable predominante es la concentración de hemoglobina corpuscular media, aunque no es determinante debido a que el índice de ginis es alto. Se generan dos nodos hijos dónde las variables determinantes son hemoglobina y hematocrito. Del lado izquierdo dónde la variable determinante es hemoglobina, el índice de ginis es alto a comparación del lado derecho, le toma más pasos asegurar que un paciente tenga o no anemia. Este árbol de decisión nos permite identificar que las variables principales para identificar anemia son: concentración de hemoglobina corpuscular media y hemoglobina corpuscular media.

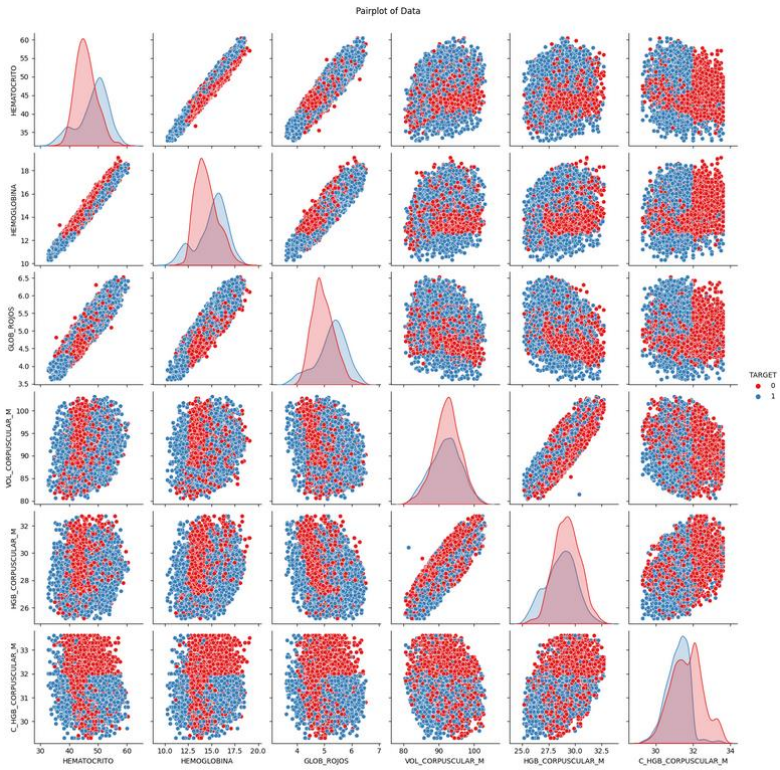

Ilustración 19: Gráfica de dispersión entre pares de variables de clasificación de anemia.
Fuente: autor del documento

En la ilustración número 19 se observa una matriz 6x6 de dispersión entre pares de variables, se observa la comparativa de una variable con las otras cinco dónde se evidencia los

casos de no anemia (0) y si anemia (1). En la comparación de la diagonal univariable tenemos que normalmente los valores de no anemia tienen volumen más alto que los valores de si anemia.

Por otro lado, en las variables de hematocrito, hemoglobina y concentración de hemoglobina corpuscular media se identifican picos claros que denotan bimodalidad y sugieren diferencia significativa en el grupo.

En la comparación bivariante entre hemoglobina, hematocrito y glóbulos rojos se presenta una correlación lineal, sucede lo mismo entre concentración de hemoglobina corpuscular media y hemoglobina corpuscular media. Esto confirma la correlación vista en la ilustración 14 entre dichas variables.

Existen varios puntos con una separación clara entre las diferentes clases, esto sugiere que estas variables podrían ser buenas discriminantes para el target.

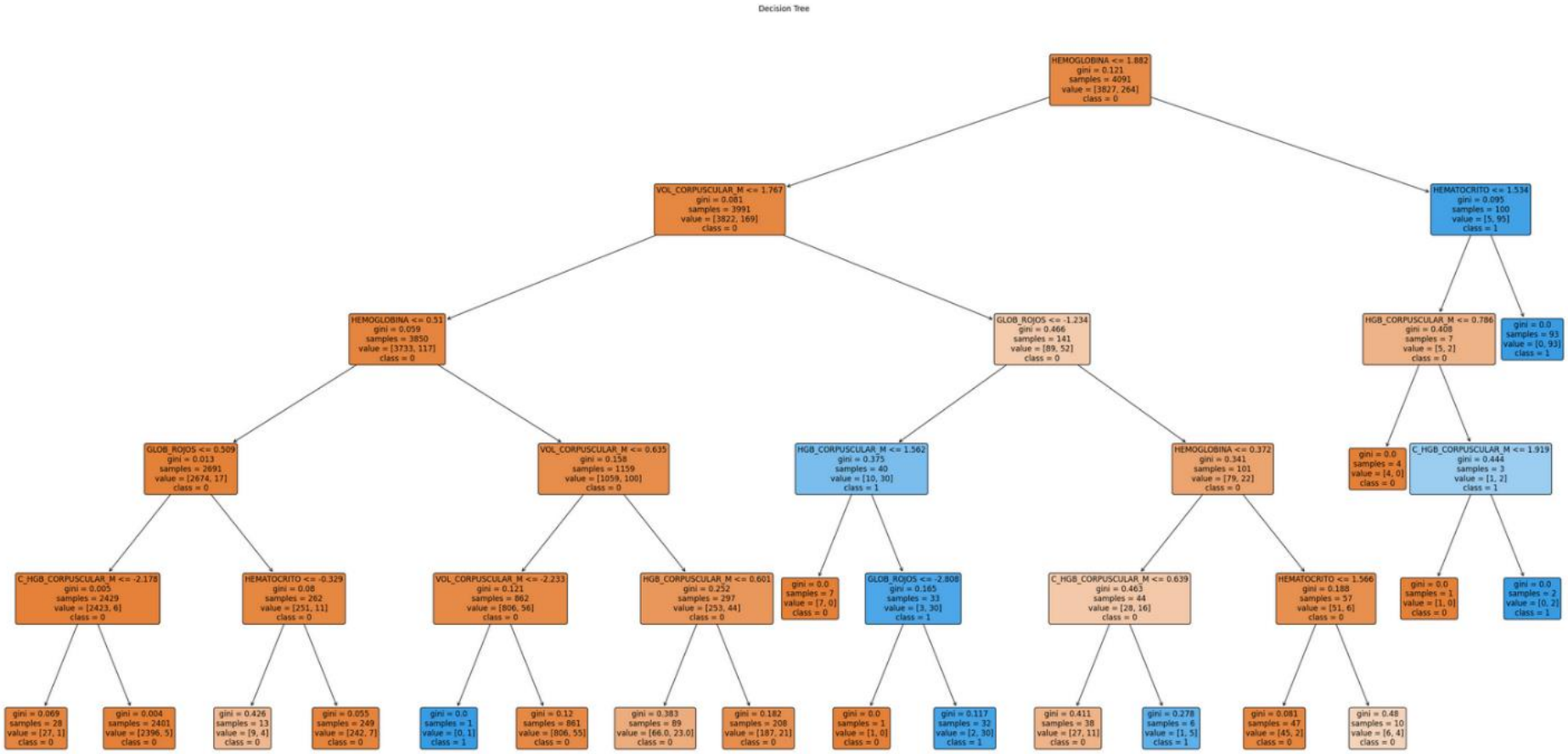

Ilustración 20: Gráfico de árbol de decisión de poliglobulia.
Fuente: autor del documento

En la ilustración 20 se observa el árbol de decisión que clasifica poliglobulia, dentro de dicho árbol se observa la variable que condiciona el nodo, el índice de gini y la cantidad de registros en dicho nodo.

En el nodo padre se presenta hemoglobina como variable predominante, en este caso el índice de Gini es bastante bajo lo que asegura la precisión. Se generan dos nodos hijos cuyas variables dominantes son volumen corpuscular medio y hematocrito; de acuerdo al índice Gini bajo demuestra que estas son dos variables predominantes para identificar poliglobulia. En nodos finales se observa que la variable predominante para determinar poliglobulia es la concentración de hemoglobina corpuscular media.

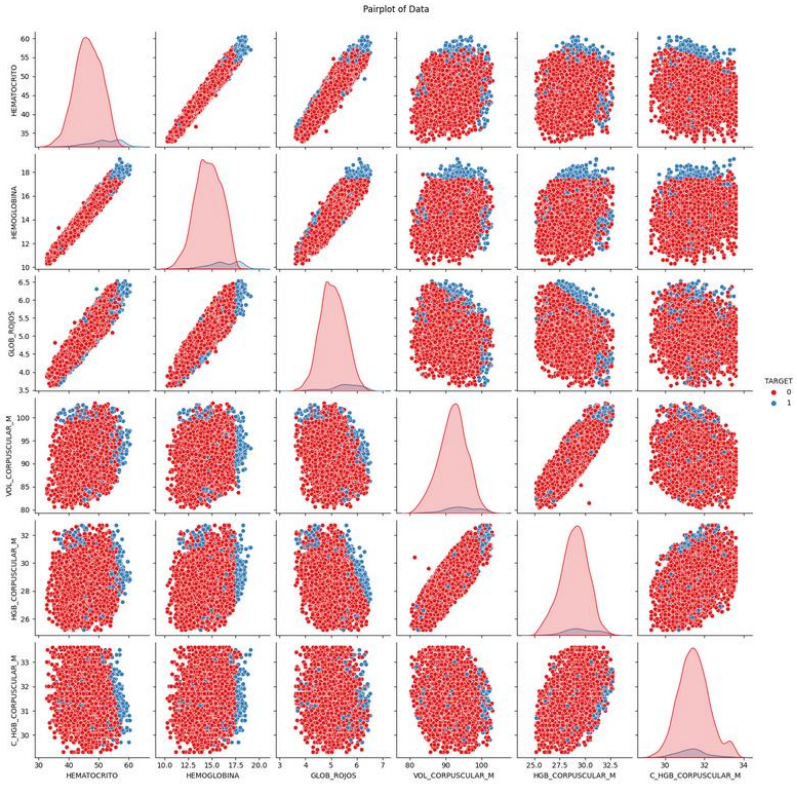

Ilustración 21: Gráfica de dispersión entre pares de variables de poliglobulia.
Fuente: autor del documento

En la ilustración 21 se observan una matriz 6x6 de dispersión de pares de variables en relación a la clasificación de poliglobulia del árbol de decisiones. Se observan los casos de no anemia (0) y si anemia (1), en colores rojos y azules respectivamente.

En la diagonal se comparan las gráficas univariables dónde los casos de no poliglobulia tienen mayor volumen. Se observa que en las gráficas de si anemia de las mayorías de variable

tienden a separarse por grupos, da a entender que dentro de la misma clasificación hay rangos de valores, por ello es importante medirlo en más de una variable.

Por otro lado, en la comparación entre las variables correlacionadas a la variable predictora observadas en la figura 14, se observa que mantienen un orden lineal de la dispersión.

4.5.2. Redes neuronales simples

El objetivo de los modelos de redes neuronales simples implementados en el paso anterior es clasificar los casos de anemia y poliglobulia, dicho modelo será entrenado con una proporción del 70% del total del dataset. Para visualizar los resultados del rendimiento se usarán gráficas de densidad entre valores reales y predictivos, gráficas de pasteles para comparar el porcentaje de valores si y no clasificados en comparación a los valores reales.

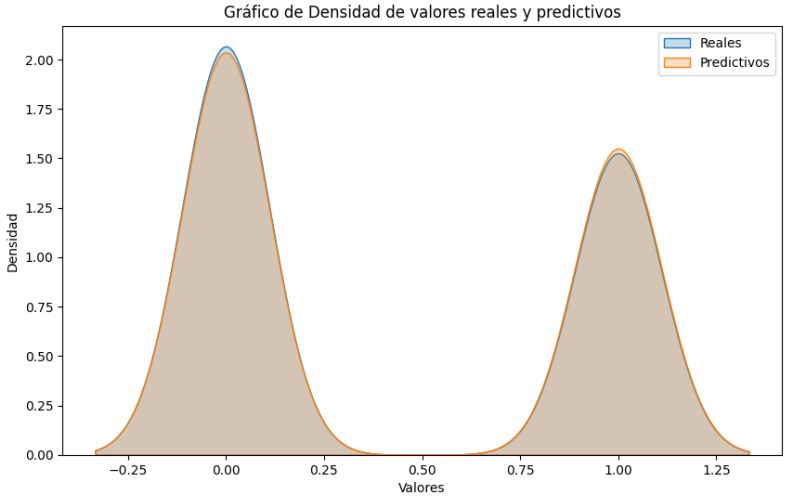

Ilustración 22: Gráfico de densidad de valores reales y predictivos de anemia.
Fuente: autor del documento

En la ilustración 22 se observa el gráfico de densidad entre valores reales y predictivos en resultados de anemia. Para pacientes sin anemia (0), la densidad de la clasificación es ligeramente inferior a los valores reales. Por otro lado, en pacientes que si tienen anemia (1), los datos predictivos son ligeramente superiores a los valores reales.

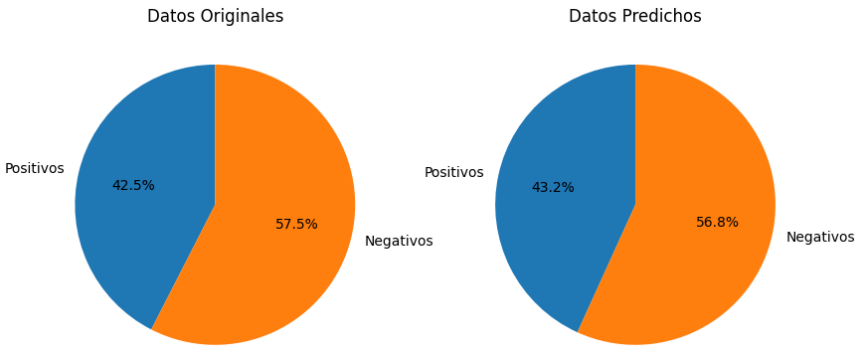

Ilustración 23: Gráfico de pasteles de datos originales y predictivos de anemia.
Fuente: autor del documento

En la ilustración 23 se observa un gráfico de pasteles de datos originales y datos predictivos de valores de anemia. En los casos positivos se observa que existe un porcentaje de 43.2 % en los datos de predicción, siendo mayor en comparación a los datos originales. Se equivoca en un 0.7%, por lo que tiende a dar falsos positivos.

Por otro lado, en los datos negativos existe un porcentaje 56.8 % en los datos de predicción, siendo menor comparado con los datos reales. Se equivoca en un 0.7%, tomando en cuenta el caso anterior, dichos valores que sobran en la predicción de casos positivos son los que faltan en los casos negativos.

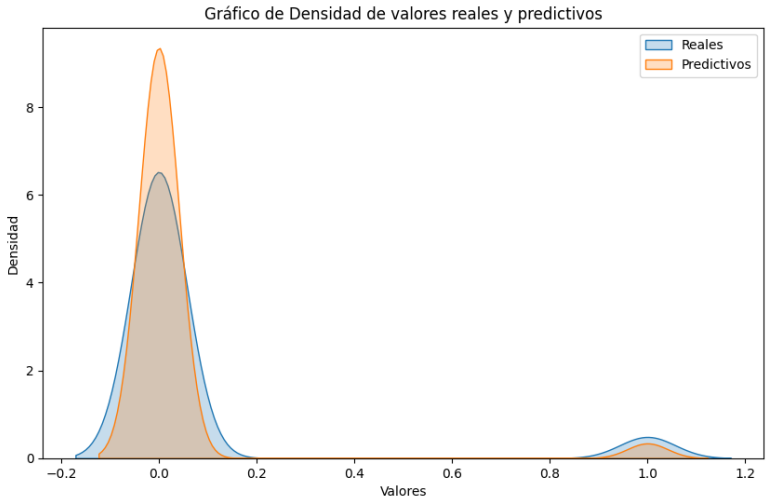

Ilustración 24: Gráfico de densidad de valores reales y predictivos de poliglobulia.
Fuente: autor del documento

En la ilustración 24 se observa el gráfico de densidad de valor reales y predictivos de poliglobulia. En los casos negativos (0), la densidad de valores predictivos es mayor que los

valores reales. Por otro lado, los casos positivos (1), la densidad de valores predictivos es menor que los valores reales.

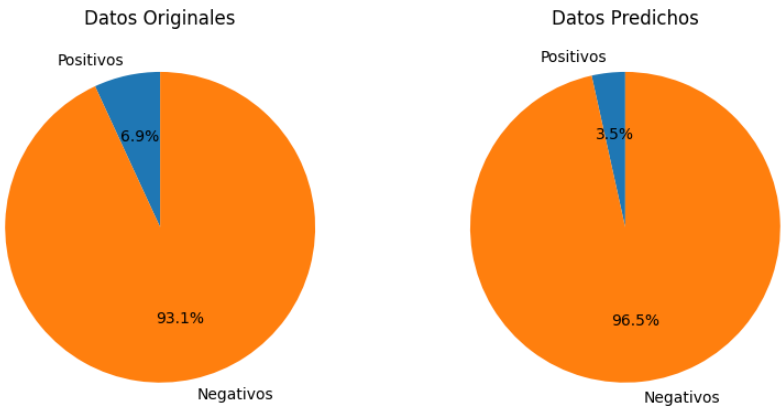

Ilustración 25: Gráfico de pasteles de valores reales y predictivos de poliglobulia.
Fuente: autor del documento

En la ilustración 25 se observa el gráfico de pasteles de valores reales y valores predictivos de poliglobulia. En los casos positivos se obtiene un porcentaje de 96.5 % en los datos de predicción, siendo mayor en relación a los datos reales. Se equivoca en 3.4% de valores positivos de poliglobulia, siendo mayor porcentaje que la red neuronal simple de anemia.

Por otro lado, en los casos negativos da un porcentaje de 3.5 % en los datos de predicción, siendo menor porcentaje en comparación a los datos originales. Se equivoca en 3.4%, estos datos faltantes son los falsos positivos del modelo.

4.6. Evaluación de modelos

Los modelos de aprendizaje supervisado cómo: clasificación y redes neuronales son medidos mediante funciones de *scikit-learn*.

4.6.1. Modelos de clasificación

4.6.1.1. Árbol de decisiones

La precisión del modelo de árbol de decisión en anemia es de 0.86, con 5 niveles en el árbol. Al ser un valor cercano al 1 demuestra que el modelo es adecuado (ilustración 26).

```
Rendimiento: 0.86
```

Ilustración 26: Rendimiento del árbol de decisión de anemia.
Fuente: autor del documento

En la ilustración 27 se observa la matriz de confusión de árbol de decisión en valores de anemia; presenta 104 registros cómo falsos positivos que equivalen al 14.6% de los registros positivos. Respecto a los falsos negativos se identifican 137 registros que equivalen al 13.14% de los registros negativos en anemia. Para los datos de prueba estos porcentajes son mínimos.

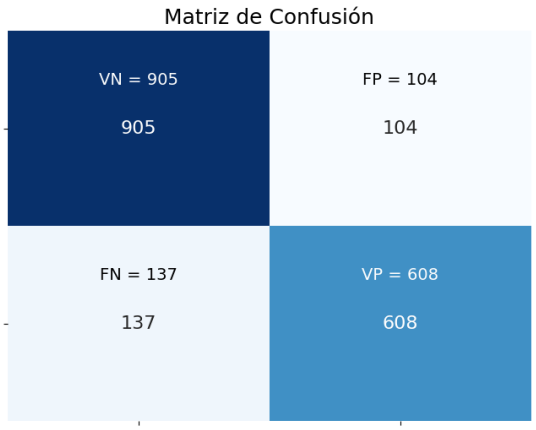

Ilustración 27: Matriz de confusión de árbol de decisión de anemia.
Fuente: autor del documento

La precisión del modelo de árbol de decisión en poliglobulia es de 0.96, con 5 niveles en el árbol. Al ser un valor cercano al 1 demuestra que el modelo es adecuado (ilustración 28).

```
Rendimiento: 0.96
```

Ilustración 28: Rendimiento del árbol de decisión de poliglobulia.
Fuente: autor del documento

En la ilustración 29 se observa la matriz de confusión de árbol de decisión en valores de poliglobulia; presenta 6 registros cómo falsos positivos que equivalen al 10.52% de los registros positivos. Respecto a los falsos negativos se identifican 70 registros que equivalen al 4.12% de los registros negativos en poliglobulia. Para los datos de prueba estos porcentajes son mínimos.

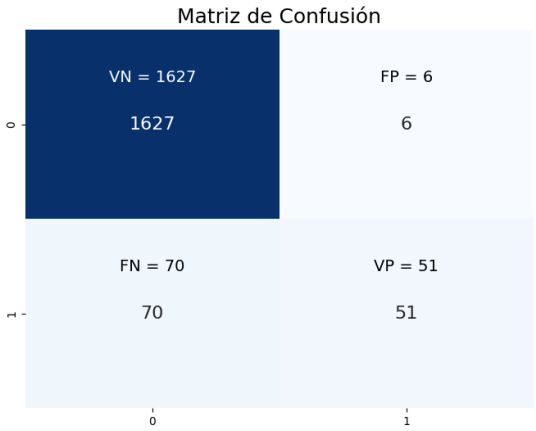

Ilustración 29: Matriz de confusión de árbol de decisión de poliglobulia.
Fuente: autor de documento

4.6.2. Redes neuronales simples

La precisión del modelo de redes neuronales simples en anemia es de 0.85, con 2 capas internas de 10 neuronas cada una. Al ser un valor cercano al 1 demuestra que el modelo es adecuado (ilustración 30).

```
Rendimiento: 0.85
```

Ilustración 30: Medida de rendimiento de red neuronal en anemia.
Fuente: autor de documento

En la ilustración 31 se observa la matriz de confusión de la red neuronal en valores de anemia; presenta 140 registros cómo falsos positivos que equivalen al 18.46% de los registros positivos. Respecto a los falsos negativos se identifican 127 registros que equivalen al 12.75% de los registros negativos en anemia. Para los datos de prueba estos porcentajes son medianamente altos.

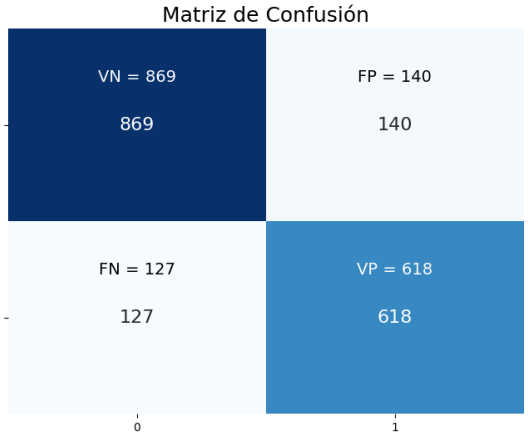

Ilustración 31: Matriz de confusión de red neuronal en anemia.
Fuente: autor del documento

La precisión del modelo de redes neuronales simples en poliglobulia es de 0.96, con 2 capas internas de 10 neuronas cada una. Al ser un valor cercano al 1 demuestra que el modelo es adecuado (ilustración 33).

```
Rendimiento: 0.96
```

Ilustración 32: Medida de rendimiento de red neuronal en poliglobulia.
Fuente: autor del documento

En la ilustración 34 se observa la matriz de confusión de la red neuronal en valores de poliglobulia; presenta 9 registros cómo falsos positivos que equivalen al 14.75% de los registros positivos. Respecto a los falsos negativos se identifican 69 registros que equivalen al 4.07% de los registros negativos en anemia. Para los datos de prueba estos porcentajes son bajos.

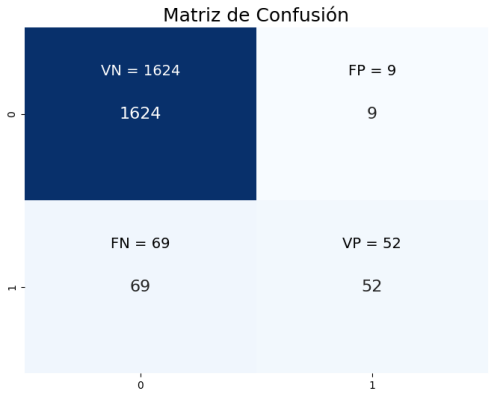

Ilustración 33: Matriz de confusión de red neuronal simple en poliglobulia.
Fuente: autor del documento.

4.7. Validación de modelos

MLOps es una metodología cuyo enfoque principal son los modelos de aprendizaje supervisado, no se tomará en cuenta los modelos de agrupamiento para este paso.

4.7.1. Modelos de clasificación

4.7.1.1. Árboles de decisión

Cómo se pudo observar en fases anteriores, los árboles de decisión nos permiten visualizar las relaciones y patrones de comportamiento del dataset por medio de los caminos

que siguen sus nodos. Además de esto, sus valores de precisión fueron ligeramente mejores en comparación a los modelos de red neuronal.

4.7.2. Redes neuronales simples

En el caso de las redes neuronales simples, su función principal de clasificar los casos de anemia y poliglobulia se cumplió con un rendimiento de 0.85 en anemia y 0.96 en poliglobulia en escala de 0 a 1. Se probaron múltiples estructuras de redes neuronales sin encontrar un cambio significativo, por lo que se mantiene la misma estructura de dos capas internas de 10 neuronas en ambos casos.

5. CAPÍTULO 5: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- Se aplicaron técnicas de ciencias de datos, orientados al uso de la metodología MLOps. Por medio de los modelos de clasificación de redes neuronales simples y árboles de decisión se consiguió clasificar los casos de anemia y poliglobulia con un rendimiento superior al 84% en resultados de biometrías hemáticas de un laboratorio clínico de la ciudad de Quito. Por otro lado, los modelos de árboles de decisión nos permitieron visualizar las relaciones y patrones de comportamiento que los resultados de biometría hemática tenían.
- La metodología MLOps aporta un enfoque más completo al que aporta CRISP DM, agregando más fases consiguiendo un seguimiento más preciso del proyecto de ciencia de datos. Además de esto, aporta un enfoque centrado en aprendizaje supervisado, apoyando al diseño de modelos de árboles de decisión y redes neuronales simples. Para modelos que no son de aprendizaje supervisado no se aplican las fases de entrenamiento, evaluación y validación de modelos.
- Se aplicaron procedimientos ETL con la finalidad de adecuar los datos extraídos del sistema de gestión de laboratorios clínicos a una estructura y formato que los modelos de aprendizaje automático puedan entender. Se realizaron procesos de transformación de registros y tipos de datos, eliminación de ruido y verificación de relaciones.
- Los modelos de aprendizaje automático que se adecuaron de mejor manera al dataset fueron modelos de clasificación, siendo estos árboles de decisiones y redes neuronales simples, esto debido a que la variable predictora solo incluía valores entre 0 y 1.
- Únicamente se evaluaron los modelos de aprendizaje supervisado, esto se hizo por medio de funciones incluídas en la librería *scikit-learn*. Los modelos que se aceptaron por su rendimiento fueron los modelos de árboles de decisión y redes neuronales simples.

- Por medio de los resultados presentados en el árbol de decisión con respecto a la variable raíz, se podría establecer un sistema de diagnóstico de anemia y poliglobulia para proyectos posteriores.

5.2.Recomendaciones

- En este proyecto de integración curricular se trabajaron con 6.000 registros de Resultados de biometría hemática de un laboratorio clínico. Es recomendable trabajar con más cantidad de registros históricos, esto va a permitir mejorar el rendimiento de los modelos de ciencias de datos.
- MLOps se centra en modelos de aprendizaje automáticos supervisados, es recomendable que su uso se base en modelos de aprendizaje supervisado como modelos de predicción y clasificación.
- Realizar un adecuado proceso ETL es de gran importancia, esto debido a que de esta manera se va a ingresar datos de calidad a los modelos de ciencia de datos. Si se desea tener un buen rendimiento en dichos modelos, se le debe poner especial atención a este paso.
- Verificar varios modelos de ciencia de datos nos permite tener una visión amplia de los patrones de comportamiento de los datos, de esta manera se puede elegir el modelo que se adapte de mejor manera a los datos. Aunque tengamos una preselección de ciertos modelos, es recomendable variar en los tipos de modelos que usamos.
- Es recomendable usar varios métodos para medir el rendimiento del modelo, esto va a permitir ver de forma mucho más precisa si el modelo funciona adecuadamente.
- En base a los resultados de este proyecto se recomienda la elaboración de un sistema de diagnóstico de anemia y poliglobulia.

6. BIBLIOGRAFÍA

Alban, G. P. G., Arguello, A. E. V., & Molina, N. E. C. (2020). Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). *Recimundo*, *4*(3), 163–173.

Gaona, C. A. (2003). Interpretación clínica de la biometría hemática. *Medicina Universitaria*, *5*(18), 35.

Google Cloud. (2023). *MLOps: canalizaciones de automatización y entrega continua en el aprendizaje automático*. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=es-419

IBM. (2002). *CRISP-DM 1.0: Metodología para el desarrollo de modelos de minería de datos*. https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

Instituto Europeo de Química, F. y B. (2023). *¿Qué es y qué se hace en un laboratorio clínico?* https://ieqfb.com/laboratorio-clinico-que-se-hace/

Izaurieta, F., & Saavedra, C. (2000). Redes neuronales artificiales. *Departamento de F\'\isica, Universidad de Concepción Chile*.

Lemus-Delgado, D., & Pérez Navarro, R. (2020). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. *Colombia Internacional*, *102*, 41–62.

Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de La Universidad Veracruzana*, *9*(2), 19–24.

MedlinePlus. (2022). *Biblioteca nacional de Medicina USA*. https://medlineplus.gov/spanish/pruebas-de-laboratorio

Ministerio de Salud Pública de Ecuador. (2012). *Reglamento para el funcionamiento de los laboratorios clínicos*.

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores En Ciencias de La Computación*.

National Cancer Institute. (2024). *Diccionario de cáncer del NCI*. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer

National Human Genome Research Institute. (2024). *Linfocito - Glosario parlante de términos genómicos y genéticos*. https://www.genome.gov/es/genetics-glossary/Linfocito

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. *Método Informáticos Avanzados*, 164–174.

Tres, S. (2008). Metodología de la Investigación. *Obtenido de Http://Www. Ceavirtual. Ceauniversidad. Com/Material/3/Metod1/353. Pdf*.

UNAM. (2012). *Biometría hemática – Fórmula roja*. http://www.telematica.ccadet.unam.mx/dentizta.ni/html/MATERIAL-DENTIZTA/instrumenta/practicas/biometriahematica/recuentoeritrocitos.htm