PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MÁSTER EN SISTEMAS DE INFORMACIÓN, MENCIÓN DATA SCIENCE.

TEMA:

"APLICACIÓN DE INTELIGENCIA ARTIFICIAL MEDIANTE EL USO DE MACHINE LEARNING PARA EL PROCESO DE CLASIFICACIÓN DE DATOS ASOCIADOS AL CENTRO MÉDICO CMC."

AUTOR:

DEYSI MAGALY ESPIN ESPIN

TUTOR: MSC. OSWALDO ESPINOSA

QUITO, junio 2023

DEDICATORIA

Se lo dedico a Dios porque me dio vida y sabiduría, para mi familia, quienes siempre estuvieron a mi lado, brindándome su apoyo incondicional y motivándome a seguir adelante en todo momento. A mis amigos, por ser mi soporte emocional y por ser una fuente de inspiración constante. Gracias a mis maestros por compartir sus conocimientos y experiencias conmigo y por ser un modelo que seguir en mi carrera profesional. Y finalmente, a todos que me han ayudado en el camino, gracias por creer en mí y por ser parte de este logro.

AGRADECIMIENTO

A Dios por guía y darme la oportunidad para cumplir una nueva meta. A mis padres por ser el pilar fundamental en el transcurso de mi vida y brindarme su amor, compresión y por siempre estar a mi lado en cada paso que doy. A mi tutor por darme la oportunidad de usar su conocimiento científico y tenerme la paciencia para guiarme durante el desarrollo de mi trabajo de titulación. A la universidad que me ayudo a formarme como profesional. Y sin todos ellos, no hubiera sido posible llegar hasta aquí.

Tema: "Aplicación de inteligencia artificial mediante el uso de machine learning para el proceso de clasificación de datos asociados al Centro Médico CMC."

Autor: Espín Espín Deysi Magaly

RESUMEN

La inteligencia artificial y el aprendizaje automático han revolucionado muchos campos, incluido el ámbito médico. El Centro Médico CMC ha reconocido el potencial de estas tecnologías y se ha permitido aplicar la inteligencia artificial en la información que posee para clasificar los datos asociados.

El proceso de clasificación de datos en un Centro Médico puede ser complejo debido a la gran cantidad de información generada. El uso de machine learning permite que el sistema aprenda automáticamente a partir de los datos existentes y genere modelos predictivos. En el caso del Centro Médico CMC, se utilizan algoritmos de aprendizaje automático para analizar y clasificar los datos de manera precisa, eficiente y estos pueden contribuir significativamente a mejorar la atención prestada a los pacientes.

El primer paso en el proceso está en la preparación de los datos. Esto implica recopilar y limpiar los datos, eliminando cualquier información redundante o ruidosa. A continuación, se seleccionan las características relevantes. Una vez que los datos están preparados, se utilizan el algoritmo árbol de decisión y regresión logística múltiple. Durante el entrenamiento, el modelo aprende a reconocer patrones y relaciones en los datos para realizar predicciones precisas. Después del entrenamiento, el modelo se evalúa utilizando datos de prueba para medir su rendimiento y precisión.

Los beneficios de la aplicación de inteligencia artificial y machine learning en el proceso de clasificación de datos asociados al Centro Médico CMC son diversos. Permite una clasificación más rápida y precisa de los datos, lo que mejora la eficiencia y la toma de decisiones del centro médico.

Palabras claves: Machine Learning, Aprendizaje automático, Algoritmos, árbol de decisión y Regresión logística múltiple.

Theme: "Application of Artificial Intelligence through Machine Learning for the Classification Process of Data Associated with the CMC Medical Center."

Author: Espín Espín Deysi Magaly

ABSTRACT

Artificial intelligence and automatic learning have revolutionized many fields, including the medicine field. The CMC Medical Center has recognized these technologies potential and it has been allowed to apply artificial intelligence to the information, which it has to classify the associated data.

The data classification process into a medical center can be complex, due to the generated information large amount. The automatic learning use allows the system for automatically learn as by existing data and generating predictive models. In the CMC Medical Center case, they used learning automatic algorithms to analyze and classify data accurately, efficiently and these can significantly contribute to improve the provided attention to patients.

The first step in the process is data preparation. This involves collecting and cleaning the data, by removing any redundant or noisy information. Next, they are selected the relevant features. Once, it is prepared the data, they are used the tree decision algorithm and multiple logistic regression. During training, the model learns to recognize patterns and relationships into data to make accurate predictions. After training, the model is assessed using test data to measure its performance and accuracy.

The artificial intelligence application benefits and machine learning in the process by classifying associated data with the CMC Medical Center are diverse. It enables faster and more accurate data classification, what improves efficiency and Medical Center decision making.

Keywords: Machine Learning, Artificial intelligence, Automatic learning, Algorithms, tree decision, Multiple logistic regression.

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I

1. DESCRIPCIÓN EL PROBLEMA

1.1. Introducción

En estos días, hay mucha información disponible en varias organizaciones, esto se debe al aumento de productos relacionados con Internet y al aumento del poder de las computadoras. Desde la informática, el aprendizaje automático ha adquirido muchas aplicaciones, incluido el procesamiento de señales en general y el procesamiento de imágenes y videos en particular. Aunque el aprendizaje automático suele asociarse a métodos con gran potencia de cálculo, debido a su complejo rendimiento informático, en la actualidad existe un gran interés por incorporar este algoritmo a dispositivos y sensores conectados a Internet de las Cosas. (Estrella, 2020).

"La clasificación en el aprendizaje automático se refiere a entrenar un modelo en un conjunto de datos etiquetados para categorizar puntos de datos existentes o para clasificar nuevos puntos de datos. Se utilizan técnicas de modelado predictivo como la clasificación" (AEFOL, 2022).

Se presenta la necesidad de usar los modelos de lógica de regresión múltiple y árbol de decisión de Machine Learning para la clasificación de la información más importante y necesaria del Centro Médico CMC para poder predecir o clasificar y saber el porcentaje de demanda al Centro Médico, en citas médicas, tratamiento de fisioterapia y signos vitales con los resultados será de utilidad para el gerente para que tomar unas buenas decisiones y conocer su giro de negocio.

1.2. Justificación

La inteligencia artificial en la actualidad es una de las áreas de conocimiento que cada día se va innova, pero, sin embargo, su aceptación a manos de las empresas va a menor velocidad debido a su componente o la necesidad de adaptarse a cada sector.

Su alta determinación y el requerimiento de personal muy cualificado, favorable para asesorar y desplegar el entorno de tecnologías basadas en inteligencia artificial que mejor se acomoda a la situación concreta del negocio, para que este pueda crecer con base en ahorrar tiempo, recursos y consiguiendo la información más precisa.

Esta aplicación de inteligencia artificial va a permitir tener una información de las variables más usadas para tener un resultado y así la persona responsable pueda analizar la situación y el comportamiento del giro del negocio y así conseguir el cambio o mejorar el negocio.

La aplicación de inteligencia artificial en el trabajo de titulación va a permitir tener una información de las variables más usadas y tener los resultados que permitirá a la persona

responsable, pueda analizar la situación y el comportamiento del giro del negocio, y conseguir el cambio o mejorar el negocio dentro del Centro Médico CMC.

1.3. Planteamiento el problema

¿Mediante la aplicación de Machine Learning a una data set que contiene información histórica del Centro Médico CMC, se puede predecir y mejorar el giro del negocio?

1.4. Contextualización del tema u objeto

El objetivo de la propuesta es la aplicación de inteligencia artificial con el uso de Machine Learning en la acción de clasificación de los datos históricos que posee él Centro Médico CMC es posible realizar la aplicación del análisis porque existe la facilidad y disponibilidad de recurso humano y materiales para llegar a cumplir el objetivo de **clasificación o predecir** para determinar la concurrencia, diagnóstico y grupos de pacientes que más asisten al establecimiento y así mejorar tanto en su atención y en los recursos que demande.

1.5. Objetivo

1.5.1. Objetivo general

Aplicar inteligencia artificial mediante el uso de Machine Learning para el proceso de **clasificación o predicción** de datos asociados al Centro Médico CMC.

1.5.2. Objetivo específico

- Analizar las variables que proporciona el Centro Médico CMC para luego desarrollar la clasificación.
- **Predecir** cuáles son los valores más influyentes dentro del Centro Médico CMC.
- Observar los cambios que existen dependiendo su tiempo.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. Antecedentes

La inteligencia artificial entiende un extenso conjunto de métodos y normas, en particular, los medios de visión, percepción, habla y diálogo, toma de decisiones y organización, resolución de inconvenientes, igualmente como otros tipos de aplicaciones de aprendizaje autónomo.

Es necesario lograr un ambiente propicio y eficaz para promover la innovación y la utilización honesta y eficaz de las tecnologías y la inteligencia artificial. La formulación de políticas debe tener en cuenta al avance las necesidades de grupos de usuarios específicos, a fin de evitar la discriminación y velar porque todo el mundo pueda beneficiarse con la práctica de entrenar el modelo y coger los mejores resultados (unesco, 2023).

La automatización de tareas de trabajo ayuda a planificar, diagnosticar y predecir pacientes, hacer más eficiente la atención médica y también permite el análisis remoto de resultados, lo que permite una mejor distribución de los servicios médicos. Además de automatizar tareas, el uso de inteligencia artificial ayuda a planificar, diagnosticar y predecir pacientes y mejorar la atención médica, porque no solo reduce costos, sino que analiza los resultados a distancia, lo que asegura mejores resultados en los tratamientos. Compartir los servicios de salud. La telemedicina permite a los pacientes interactuar virtualmente con los proveedores de atención médica y ofrece una alternativa a una visita en persona. Ofrece una alternativa a la medicina de entrada para pacientes que se encuentran en instalaciones médicas costosas o tienen que viajar largas distancias para llegar a un centro de salud (Cecco, 2021).

2.2. Inteligencia artificial

Está compuesta de diferentes algoritmos existentes o planteados con el propósito de la creación de máquinas que tratan de presentar las mismas habilidades que el ser humano. Es una tecnología de cada día que pasa mucho puede ser un resultado lejano y misterioso, pero ya está presente en todo momento y cada hora (Abello, 2019).

2.2.1. Machine Learning

El aprendizaje automático básicamente implica identificar automáticamente patrones o tendencias "ocultos" en los datos utilizando varios algoritmos. Por lo tanto, es muy importante no solo elegir el algoritmo más adecuado (y su posterior parametrización para cada problema específico), sino también tener una cantidad de datos grande y de suficiente calidad. En los últimos años, la importancia del aprendizaje automático en la vida empresarial ha aumentado considerablemente, porque el uso inteligente del análisis de datos es la clave del éxito empresarial (Recuero, 2021).

Machine Learning tradicional o no profundo es más dependiente de la participación humana para su aprendizaje, donde los técnicos humanos establecen la estructura innata necesarias para percibir las diferencias entre los ingresos de datos.

Machine Learning profundo beneficia los conjuntos de datos marcados también es conocido como aprendizaje supervisado, para comunicar al algoritmo no es necesario un grupo de datos etiquetados y también se puede recibir datos no estructurados en formato bruto.

2.2.2. Aplicaciones y utilidades en la salud

El aprendizaje automático en medicina juega un papel central en la formación y el desarrollo de todos los profesionales de la salud. Gracias a esta tecnología, disponen de información suficiente y precisa en el momento exacto en que la necesitan. En consecuencia, el número de aplicaciones y utilidades que ofrece esta tecnología en el campo de la salud es cada vez mayor (Uniteco, 2022). Algunas aplicaciones comunes incluyen:

- Proporciona herramientas avanzadas para confirmar el proceso de toma de decisiones tanto para los profesionales de la salud como para otros Centros Médicos.
- Proporciona información actualizada y actualizada sobre determinados procedimientos y soluciones a dudas o problemas que surgen en la práctica diaria.
- Influye en los procesos centrales de gestión de la salud tanto en asuntos de atención al paciente como en la gestión de la salud. Contribuye al diagnóstico precoz de muchas enfermedades, especialmente enfermedades oncológicas.
- Estas patologías se verán muy beneficiadas con esta técnica. Diagnóstico no invasivo de determinados problemas de salud.
- Esto facilita la adquisición de biomarcadores para enfermedades específicas al predecir con precisión el riesgo existente de ciertas enfermedades.
- Primero evalúe el efecto del tratamiento en cada persona.
- Esto permite lograr una atención médica y una medicina cada vez más individualizadas.
- Incrementar el rendimiento de los profesionales.
- Ahorre costos.
- Facilita la innovación y el incremento de nuevos productos y servicios (Uniteco, 2022).

2.2.3. Utilidad de Machine Learning en la gestión de la salud

Los algoritmos de aprendizaje automático se pueden aplicar además de hacer predicciones clínicas y epidemiológicas, también se pueden aplicar en la diligencia de servicios de salud. En esta área tienen el potencial de contribuir al análisis de los procesos clínicos y administrativos en una organización de salud. Un proceso puede definirse como una secuencia de pasos interrelacionados en el contexto de una organización cuyo propósito es crear bienes o servicios.

Ejemplos de procesos son la atención al paciente en la sala de emergencia, la gestión de suministros, la distribución de los trabajadores de la salud (Pedrero, 2021).

Con la aplicación de Machine Learning ha explicado que la integración de servicios es un aspecto necesario para brindar atención médica de calidad. Esto significa integrar diferentes procesos clínicos, como la atención de salud brindada por diferentes especialistas, así como procesos administrativos, como la provisión de insumos o la gestión de costos. Por lo general, estos procesos no ocurren linealmente en una organización de atención médica, sino que ocurren en paralelo e interactúan entre sí. Por tanto, la intervención de un determinado proceso afecta no sólo a su propio funcionamiento, sino también al funcionamiento de otros relacionados con él. La integración presenta desafíos en la planificación, implementación y especialmente en la evaluación de los procesos de organización de la salud. En todos estos aspectos, es importante analizar los datos obtenidos de tales procesos.

El aprendizaje organizacional se refiere al proceso mediante el cual una organización considera la información interna y externa al tomar una decisión. Hoy en día, este aspecto se considera clave para alcanzar las metas, asegurar la estabilidad en el mercado y superarlo. Las estrategias de Machine Learning podrían proporcionar retroalimentación a la organización al integrar tanto su propio procesamiento de información como la información del entorno.

2.2.4. Temperatura corporal

"La fiebre provoca una disminución del gasto cardíaco y un aumento de las demandas metabólicas, el consumo de oxígeno y los niveles de lactato sérico. También conduce a la vasoconstricción, daño tisular, inquietud y colapso" (Santiago A. , 2022).

Tabla 1. **Cuadro de rango de temperatura corporal escala térmica**

Rango	**Tipo**	**Atención**
17°C – 35.5°C	Hipotermia	Emergencia
35.6 °C – 37.8 °C	Normal	No Emergencia
37.9 °C – 42 °C	Fiebre	Emergencia

Fuente: (Santiago A. , 2022)

2.3. Proceso de distribución de datos

Le permite organizar sus datos en categorías para que sea más fácil analizar cada uno, guardarlo y proteger la integridad de los datos que las empresas brindan para el análisis diferente, donde los científicos de datos pueden usar diferentes métodos para clasificar los datos. Por lo general, los datos en una distribución se ordenan de menor a mayor, los gráficos y cuadros facilitan la observación tanto los valores como la repetición con la que ocurren (Data, 2023) .

2.3.1. Aprendizaje supervisado

El aprendizaje supervisado enseña máquinas y modelos. Por lo tanto, el controlador proporciona al algoritmo de aprendizaje automático un conjunto de datos conocido que contiene las entradas y salidas necesarias, y el algoritmo debe encontrar una manera de encontrar la entrada y la salida. Una vez que el usuario sabe la respuesta correcta a una pregunta, el algoritmo identifica patrones en los datos, aprende de las observaciones y hace predicciones. El algoritmo hace predicciones y el usuario hace ajustes, y este proceso continúa hasta que el algoritmo alcanza un alto nivel de precisión y rendimiento (Apd, 2019).

- **Regresión logística múltiple**

  El uso de la regresión logística múltiple amplía el modelo básico para predecir respuestas binarias utilizando predictores múltiples, que pueden ser continuos o categóricos.

Figura 1. **Ecuación para obtener la predicción.**

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p$$

Fuente: (Martínez, 2018).

donde X = (X1,...,Xp) es la proyección de p.

También utilizamos el método de máxima verosimilitud para estimar los coeficientes β0,β1,...,βp. Cada coeficiente se define como una constante.

Al igual con la regresión lineal, los resultados obtenidos con un solo predictor pueden diferir de los resultados obtenidos con varios predictores, especialmente si existe una correlación entre ellos. Este fenómeno se conoce como confusión (Martínez, 2018).

- **Entornos del modelo logístico**

  La regresión logística no requiere condiciones específicas como la linealidad, la normalidad y la homocedasticidad de los residuos que requiere la regresión lineal. Las principales condiciones requeridas para este modelo son:

  - **Respuesta Binaria:** La variable dependiente debe ser binaria.
  - **Independencia:** Las observaciones deben ser independientes.

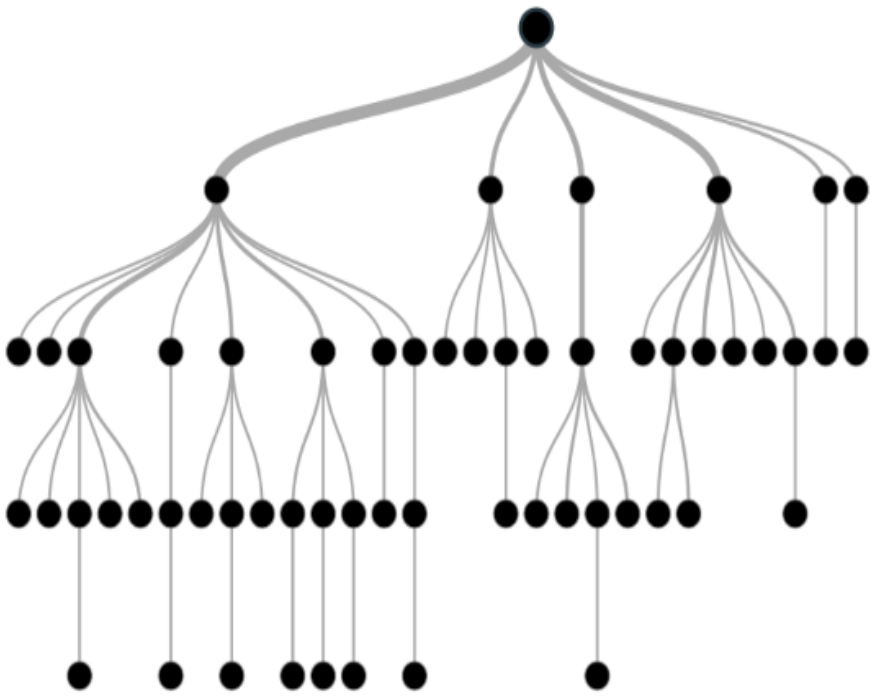
- **Multicolinealidad:** Se requiere poca o ninguna multicolinealidad entre predictores (para regresión logística múltiple).
- Linealidad entre la variable independiente y el logaritmo natural de las probabilidades.
- **Tamaño de la muestra:** como regla general, debe tener al menos 10 casos con el resultado menos común para cada variable independiente en su modelo (Martínez, 2018).

- **Árbol de decisión**

El árbol de decisión forma un algoritmo de aprendizaje automático popular que se puede usar tanto para trabajos de regresión como de clasificación. Son fáciles de entender, interpretar e implementar, esto lo convierte en una opción ideal para principiantes en aprendizaje automático.

Es una herramienta con aplicaciones que cubren varias áreas diferentes. Los árboles de decisión se pueden utilizar para problemas de clasificación y regresión. El propio nombre sugiere que el diagrama de flujo se utiliza como una estructura de árbol que muestra las predicciones obtenidas a partir de una serie de subdivisiones basadas en características. Comienza en el nodo raíz y termina con decisiones hoja (Saini, 2021).

Figura 2. **Diseño de árbol de decisión.**



Fuente: (Saini, 2021).

Antes de aprender más sobre los árboles de decisión, familiaricémonos con algunas de las terminologías:

- **Nodos raíz:** es el nodo presente al comienzo de un árbol de decisión a partir de este nodo, la población comienza a dividirse según varias características.
- **Nodos de decisión:** los nodos que obtenemos después de dividir los nodos raíz se denominan nodos de decisión.
- **Nodos de hoja:** los nodos en los que no es posible una mayor división se denominan nodos de hoja o nodos terminales
- **Subárbol:** al igual que una pequeña porción de un gráfico se denomina subgráfico, de manera similar, una subsección de este árbol de decisión se denomina subárbol.
- **Poda:** no es más que cortar algunos nodos para evitar el sobreajuste (Saini, 2021).

2.3.2. Aprendizaje sin supervisión

En el aprendizaje no supervisado, los algoritmos de aprendizaje automático son responsables de interpretar grandes conjuntos de datos y administrarlos en consecuencia. Luego, el algoritmo intenta organizar estos datos de una manera que describa su estructura. Esto significa agrupar u organizar los datos para que se vean más organizados. Cuantos más datos evalúe, mejor y más precisa será su capacidad para tomar decisiones (Apd, 2019).

2.3.3. Aprendizaje por refuerzo

El aprendizaje por refuerzo se refiere a un proceso de aprendizaje estructurado en el que un algoritmo de aprendizaje automático recibe un conjunto de funciones, parámetros y valores finales. Al definir reglas, los algoritmos de aprendizaje automático intentan evaluar diferentes opciones y posibilidades al observar y evaluar el resultado de cada una para determinar cuál es la mejor (Apd, 2019).

2.4. Científico de datos

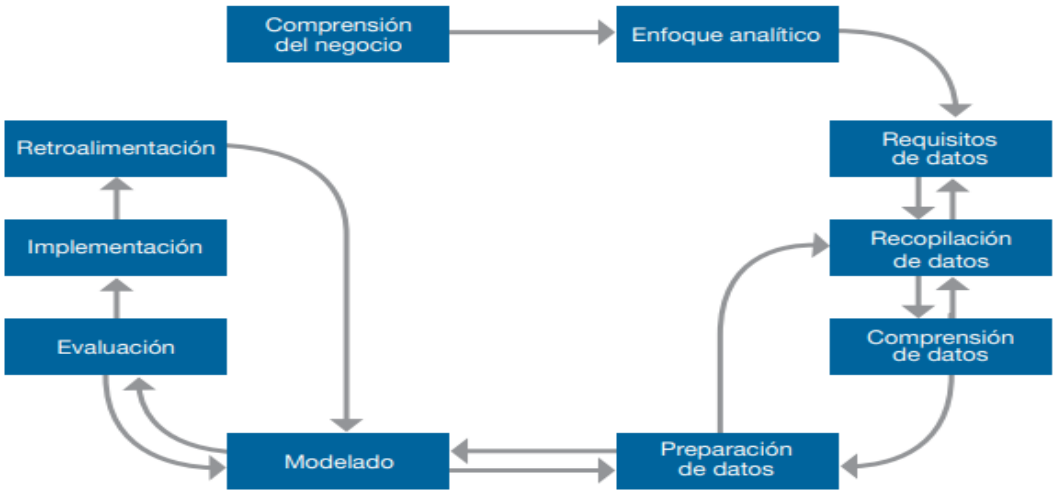
Examina datos y extrae información que es importante para una empresa. También es interdisciplinario, combinando los principios y prácticas de las matemáticas, la estadística, la computación y la propia inteligencia artificial para analizar grandes cantidades de datos. Este análisis permite al científico de datos diseñar y responder algunas preguntas. Qué sucedió, por qué sucedió, qué sucederá y qué se puede hacer al respecto (AWS, 2022).

2.4.1. Metodologías de ciencia de datos

Una metodología es una estrategia general que actúa como guía de técnicas y actividades en un área determinada. Una metodología no está relacionada con una técnica o herramienta en

particular, ni es un conjunto de técnicas o métodos. Más bien, la metodología proporciona a los científicos de datos en la Figura 3 ilustrar cómo manipular las técnicas, procesos y explicaciones que se manejarán para llegar a una respuesta o resultado (Rollins, 2015).

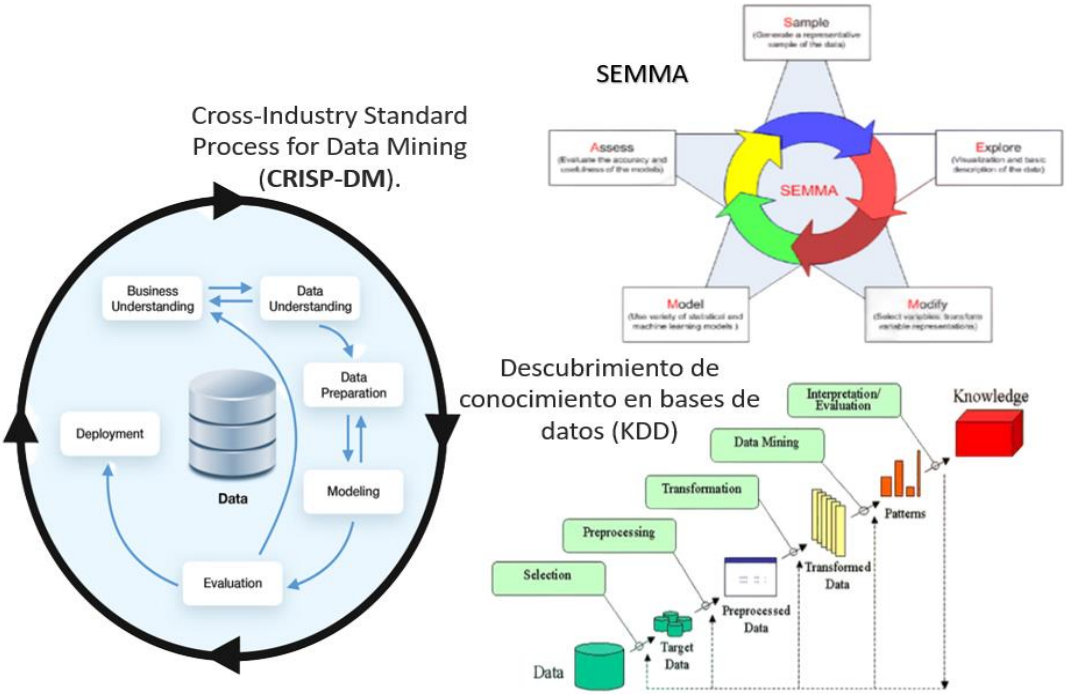
Figura 3. **Metodología fundamental para la ciencia de datos.**



Fuente: (Rollins, 2015).

En la Figura 3 indica el siclo de la ciencia de datos para tener un conocimiento de cómo llegar a dar una resolución al trabajo de titulación.

Figura 4. **Metodología para extraer información de los datos incluyente en KDD, SEMMA y CRISP-DM.**



Fuente: (Quantum, 2019).

La Figura 4 permite tener la idea de cómo esta parametrizada, cada una de las metodologías para ser ejecutadas en cada una de sus etapas con sus respectivos componentes.

- **KDD**

  Esto permite una serie de pasos para extraer patrones y conocimiento de grandes cantidades de datos.

- **SEMMA**

  Se define como un método de selección, análisis y modelado de grandes cantidades de datos para descubrir tendencias comerciales desconocidas.

- **CRISP_DM**

  Esta metodología proporciona dos documentos separados, un modelo de revisión y una guía del usuario, para especificar cada paso del análisis de datos históricos como herramientas para ayudarlo a progresar en su proyecto de minería de datos.

Tabla 2. **Cuadro comparativo de metodologías de ciencia de datos**

KDD	**SEMMA**	**CRISP-DM**
		Compresión del negocio
Selección	Muestra	Compresión de los datos
Preprocesamiento	Exploración	
Transformación	Modificación	Preparación de los datos
Procesamiento de datos	Modelo	Modelo
Interpretación /evaluación	Evaluación	Evaluación
		Despliegue

Fuente: (Luther, 2020).

La Tabla 2 y la Figura 4 muestra la comparación de las metodologías en sus diferentes fases, para escoger la más adecuada, para el desarrollo de trabajo de titulación, la metodología escogida es CRISP-DM porque permite realizar todos los pasos y el análisis del desarrollo de datos y de una forma más adecuada entendiendo el giro de negocio en este caso del Centro Médico CMC.

2.5. Herramientas

2.5.1. Python

Python es un gran lenguaje de programación de propósito general y está creciendo en popularidad para la ciencia de datos. La demanda de estas habilidades está aumentando porque las empresas quieren obtener información de sus datos (School, 2022).

2.5.2. Jupyter notebook

Es una herramienta web de código abierto que nos permite crear y compartir código y documentación. Es un entorno informático colaborativo que permite a los usuarios revisar y compartir código. Jupyter es la abreviatura de Julia, Python y R, estos son los tres lenguajes de

programación con los que comencé Jupyter, aunque hoy en día admite una gran cantidad de lenguajes(Figueiras, 2021).

2.5.3. Sql server 2012

Esta herramienta se utilizará para la recopilación de la información del histórico del Centro Médico CMC que se encuentra estructurada en una base de datos estructurada.

2.5.4. Numpy

Esta es una biblioteca del lenguaje de programación Python que admite la creación de grandes vectores y matrices multidimensionales, y una gran colección de funciones matemáticas complejas para manipularlos.

2.5.5. Panda

Esta es una de las bibliotecas de Python más útiles para los científicos de datos. Las principales estructuras de datos en pandas son cadenas para datos unidimensionales y marcos de datos para datos bidimensionales. Se puede utilizar en muchos campos, como finanzas, estadísticas, ciencias sociales y muchos campos de la ingeniería (Martinez, 2020).

2.5.6. Matplotlib

Se puede utilizar para producir los gráficos de calidad necesarios para la publicación en papel y digital y puede mostrar gráficos como series temporales, histogramas, espectros de potencia, gráficos de barras y gráficos de error (Martinez, 2020).

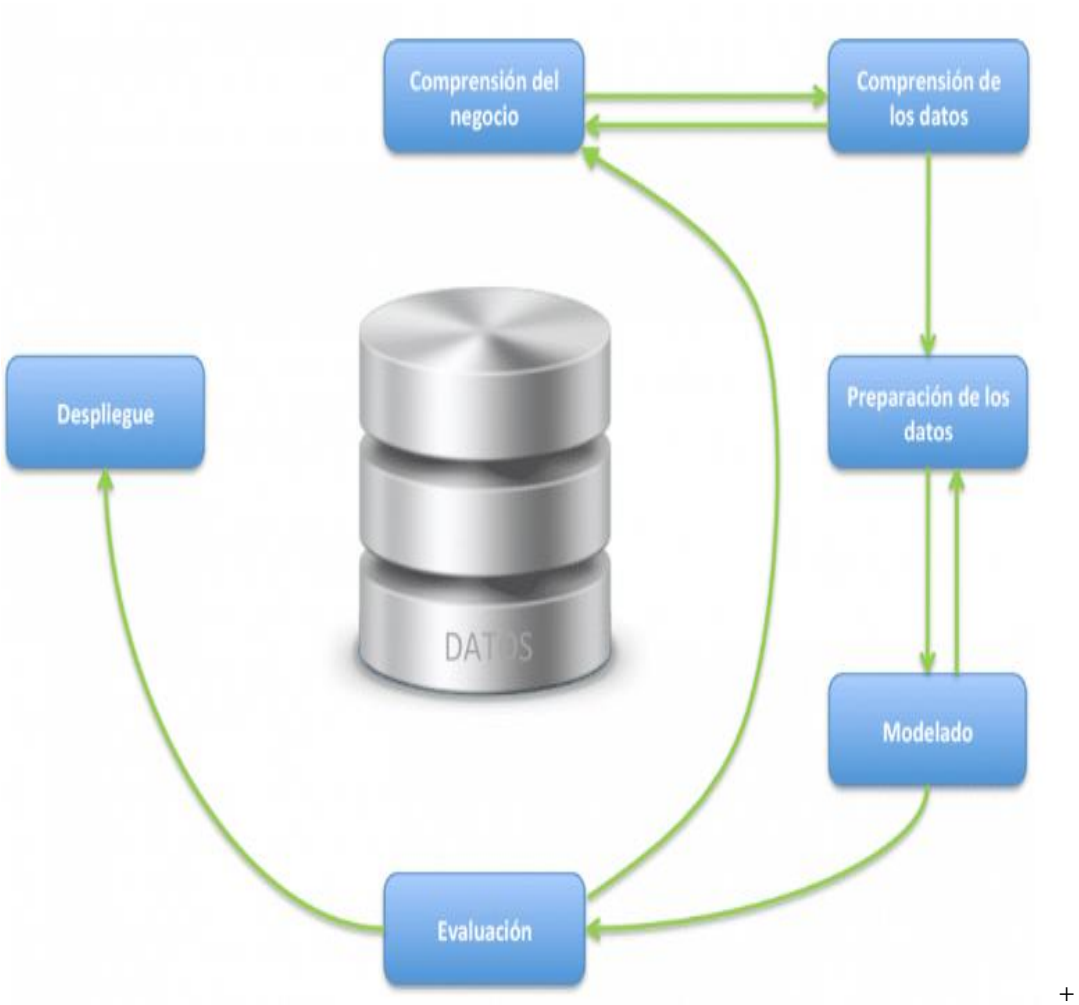
CAPÍTULO III

3. METODOLOGÍA

3.1. Metodología CRISP-DM

Al proporcionar una descripción general del ciclo de vida de un proyecto de minería de datos, el modelo de referencia CRISP-DM incluye las fases del proyecto, sus respectivas tareas y las relaciones entre tareas.

Figura 5. **Fases del modelo de referencia CRISP-DM.**



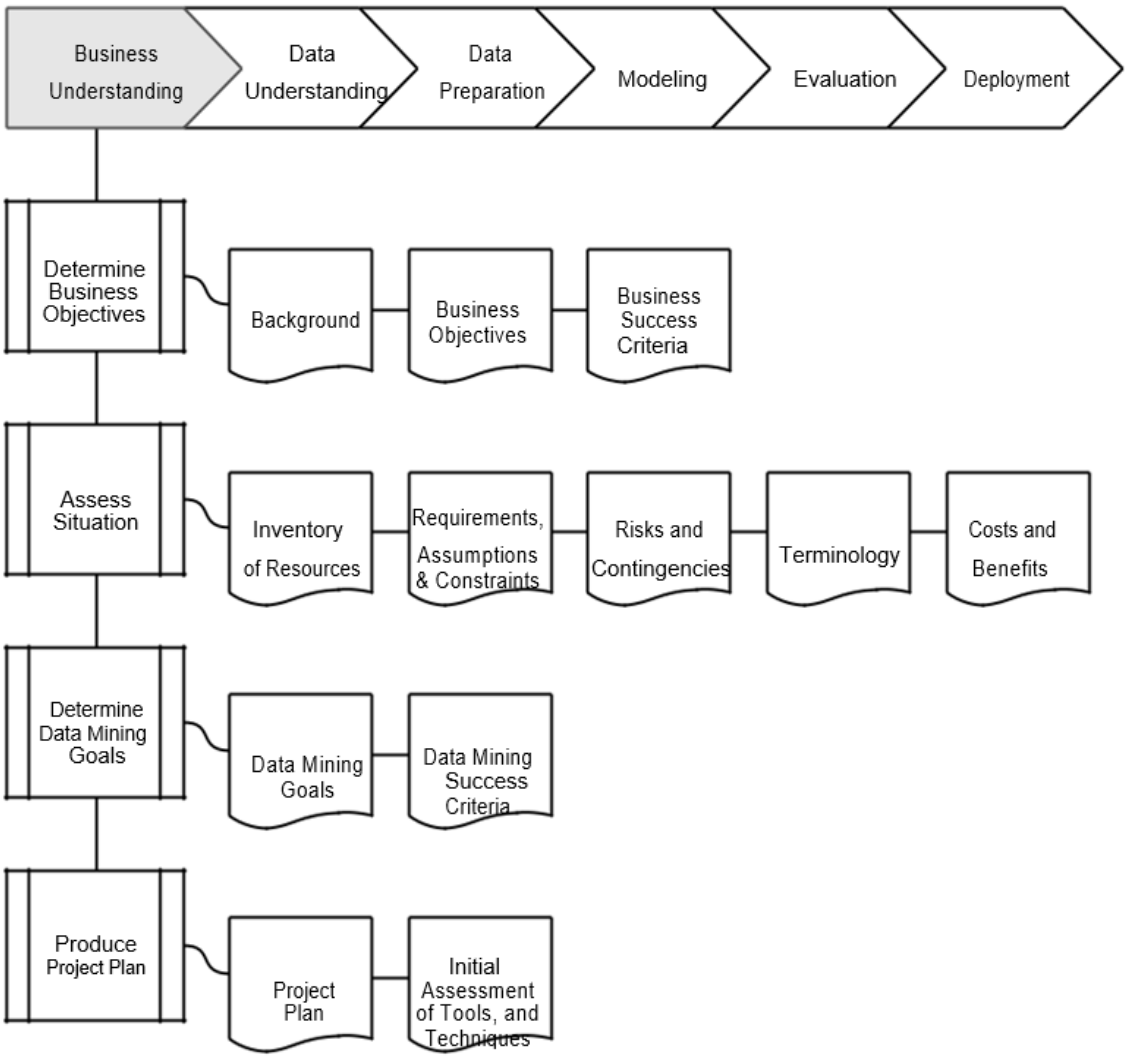
Fuente: (Gil, 2021).

Estas etapas de modelado pueden guiar los esfuerzos de minería de datos al proporcionar una descripción estandarizada del ciclo de vida de un proyecto típico de análisis de datos.

La metodología CRISP-DM forma parte de una secuencia de fases para un plan de minería de datos que son:

3.1.1. Comprensión del negocio

Figura 6. **Compresión del negocio con sus tareas y actividades.**



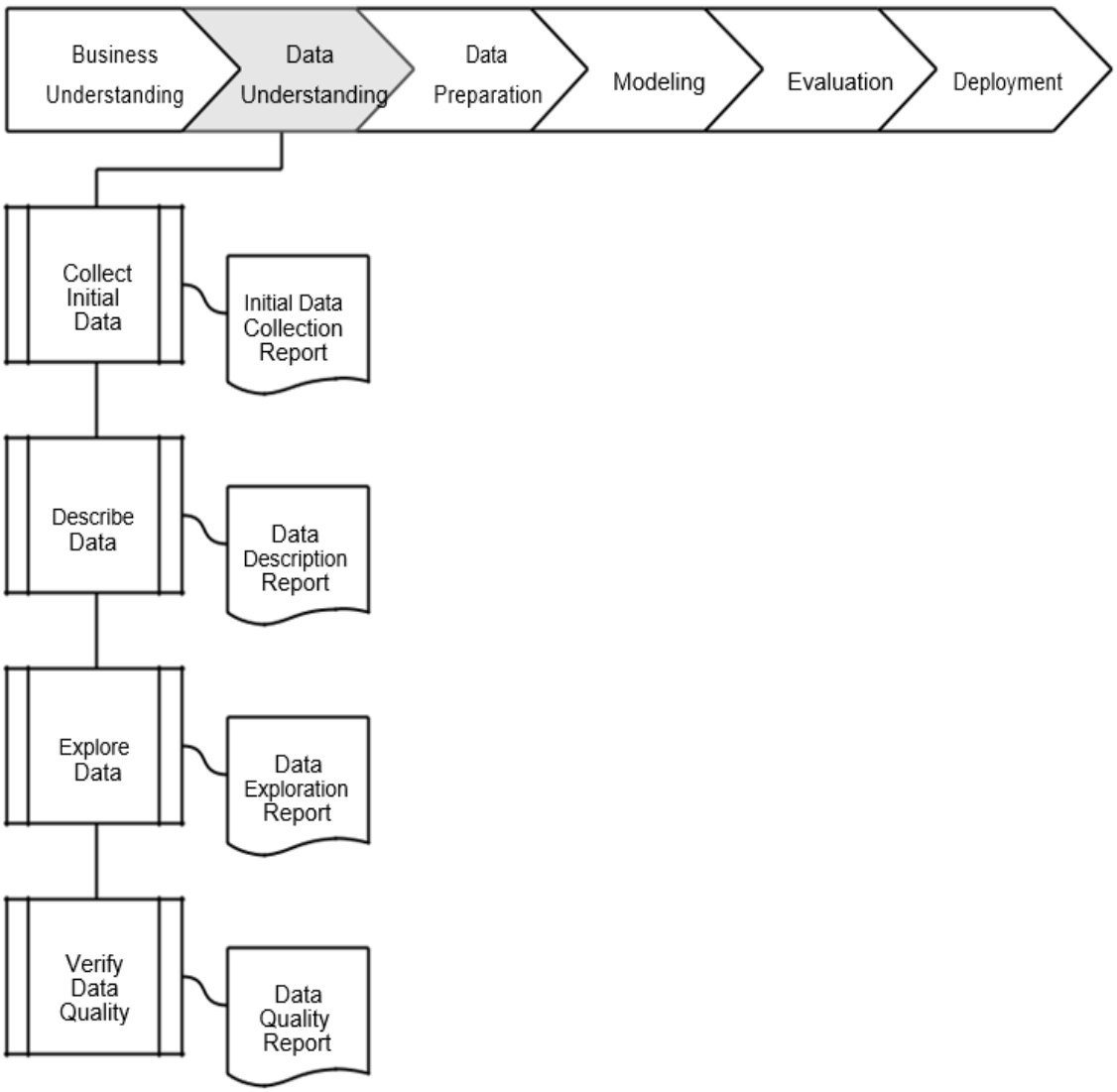
Fuente: (BizMetriks, 2013)**.**

Es la que nos permite entender el giro de negocio para ello se necesita realizar las siguientes actividades o tareas**:**

- **Identificación de problemas:** puede comprender y detallar problemas e identificar los requisitos, prerrequisitos, prohibiciones y beneficios del proyecto.
- **Determinación de objetivos:** Especifica las metas a alcanzar al representar una solución basada en un modelo de minería de datos.
- **Evaluación del estado actual:** Describe el estado actual antes de efectuar la implementación a la solución de minería de datos propuesta (B, 2015).

3.1.2. Compresión de los datos

Figura 7. **Compresión de los datos con sus tareas y actividades.**



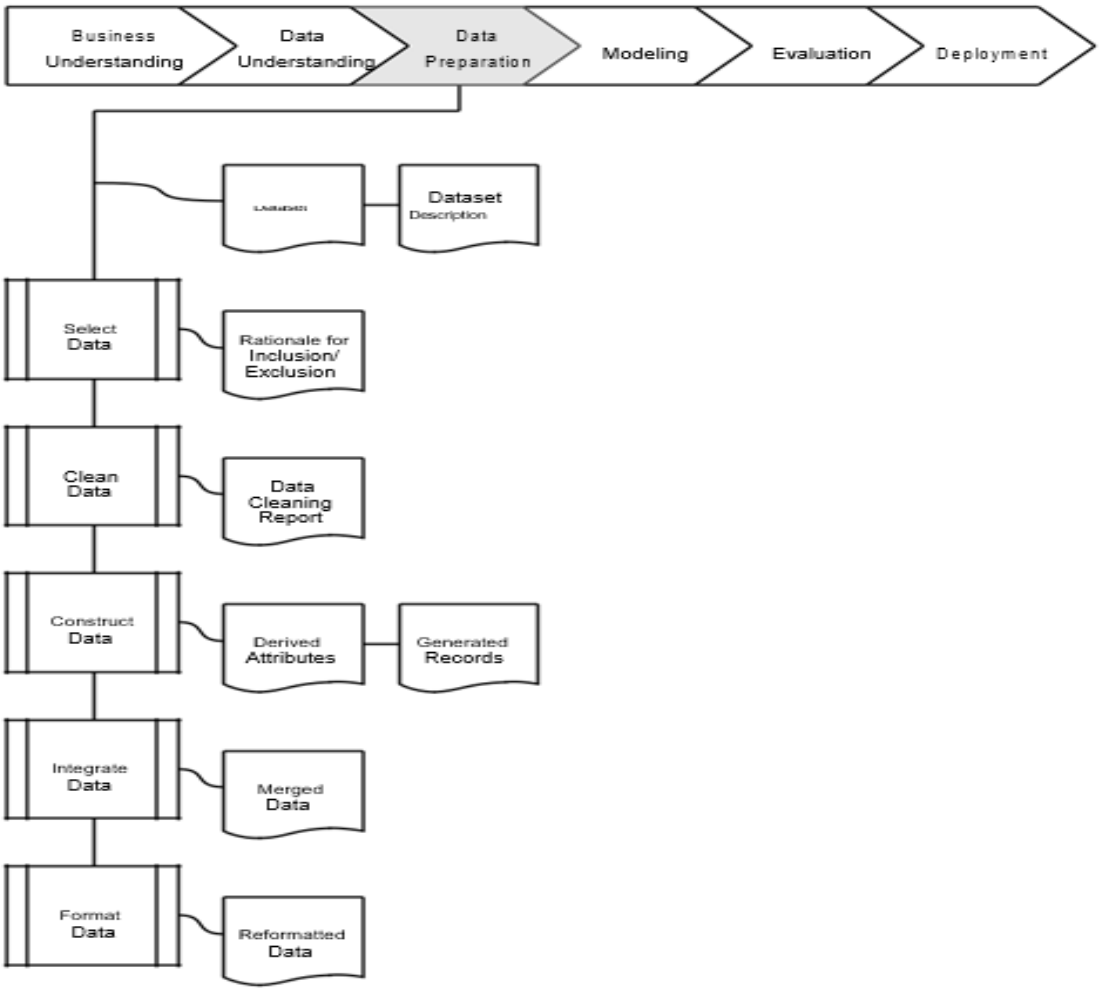
Fuente: (BizMetriks, 2013)**.**

Esta fase permite comprender como esta los datos para ser analizados para ello se tiene las siguientes actividades o tareas:

- **Recolección de datos:** Permite obtener e identificar de los datos que pueden ser de diferentes fuentes e implementar la técnica para la recolección de una forma más adecuada.
- **Descripción de datos:** Reconocer el tipo, formato, y considerado de cada dato que se encuentra dentro de la información obtenida del Centro Médico CMC.
- **Exploración de datos:** Permite realizar pruebas de estadísticas básica para conocer las propiedades de los datos y el fin es entender de mejor manera posible la información histórica (IBM, 2021).

3.1.3. Preparación de los datos.

Figura 8. **Preparación de los datos con sus tareas y actividades.**



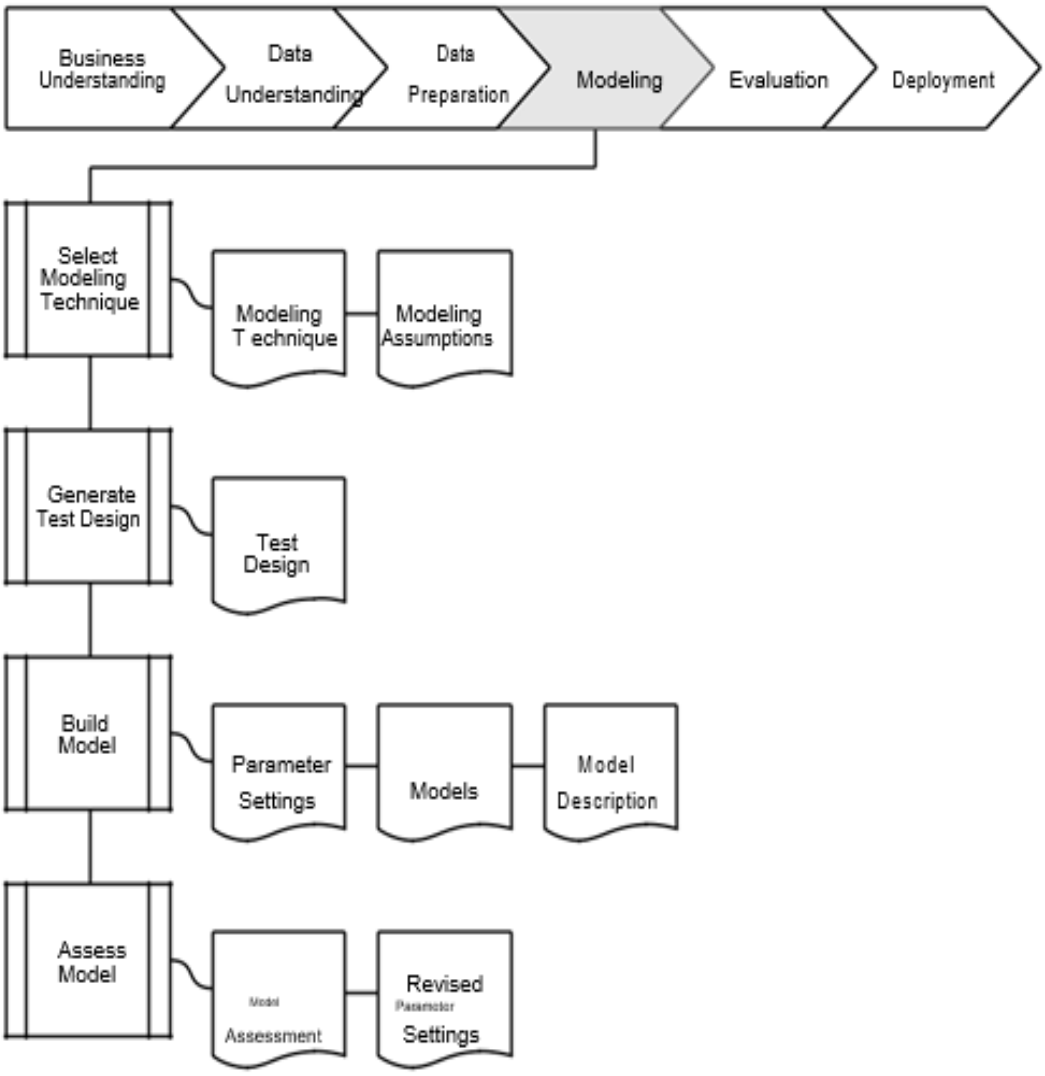
Fuente: (BizMetriks, 2013)**.**

Esta fase es la que consume más tiempo, ya que requiere la selección de datos que se transformarán en función de los resultados de las fases anteriores y se utilizarán adecuadamente en la fase de modelado. Para ello, realice las siguientes actividades o tareas:

- **Limpieza de datos:** se aplica las técnicas aprendidas en estadística, como normalización de datos, discretización de campos numéricos, manejo de valores perdidos y manejado de duplicados.
- **Crear indicadores:** Puede crear o construir indicadores a partir de datos existentes para mejorar el poder predictivo de sus datos y ayudar a identificar comportamientos interesantes para el modelado.
- **Transformación de datos:** aquí ayuda al cambio de formato o de tipo de dato sin modificar el significado principal del dato (IBM, 2021).

3.1.4. Modelado

Figura 9. **Modelado con sus tareas y actividades.**



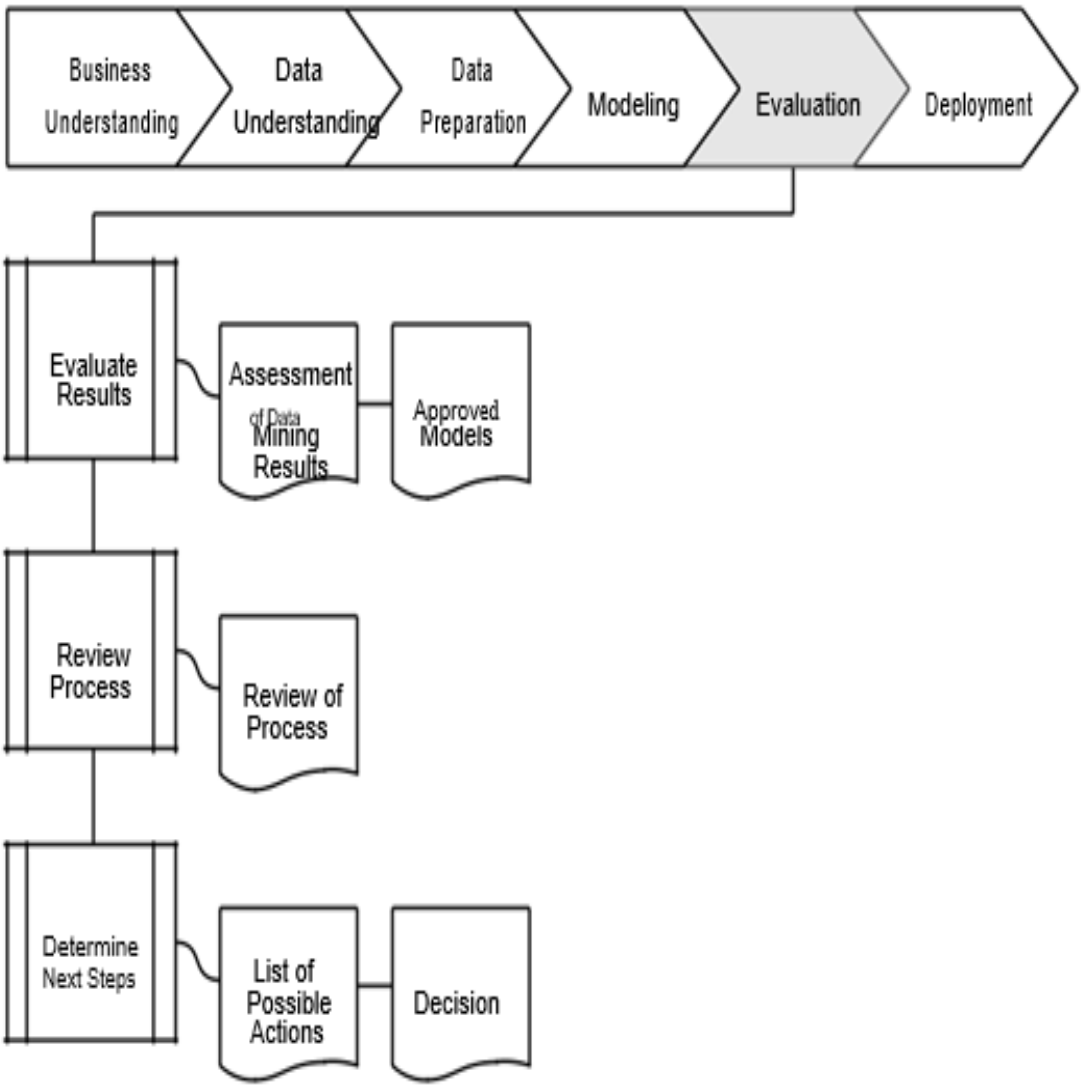
Fuente: (BizMetriks, 2013)**.**

Esta etapa es propiamente la minería de datos para ello se requiere las siguientes actividades o técnicas.

- **Selección del método para el modelado:** Permite elegir un método según el problema que se desee resolver utilizando los datos que se tiene disponible.
- **Selección de datos de prueba:** La información se puede dividir entre datos de entrenamiento y validación. Puede ser 80 y 20 % de datos de prueba.
- **Obtención del modelo:** Aplicar el modelo más sobresaliente según la clasificación (B, 2015).

3.1.5. Evaluación

Figura 10. **Evaluación con sus tareas y actividades.**

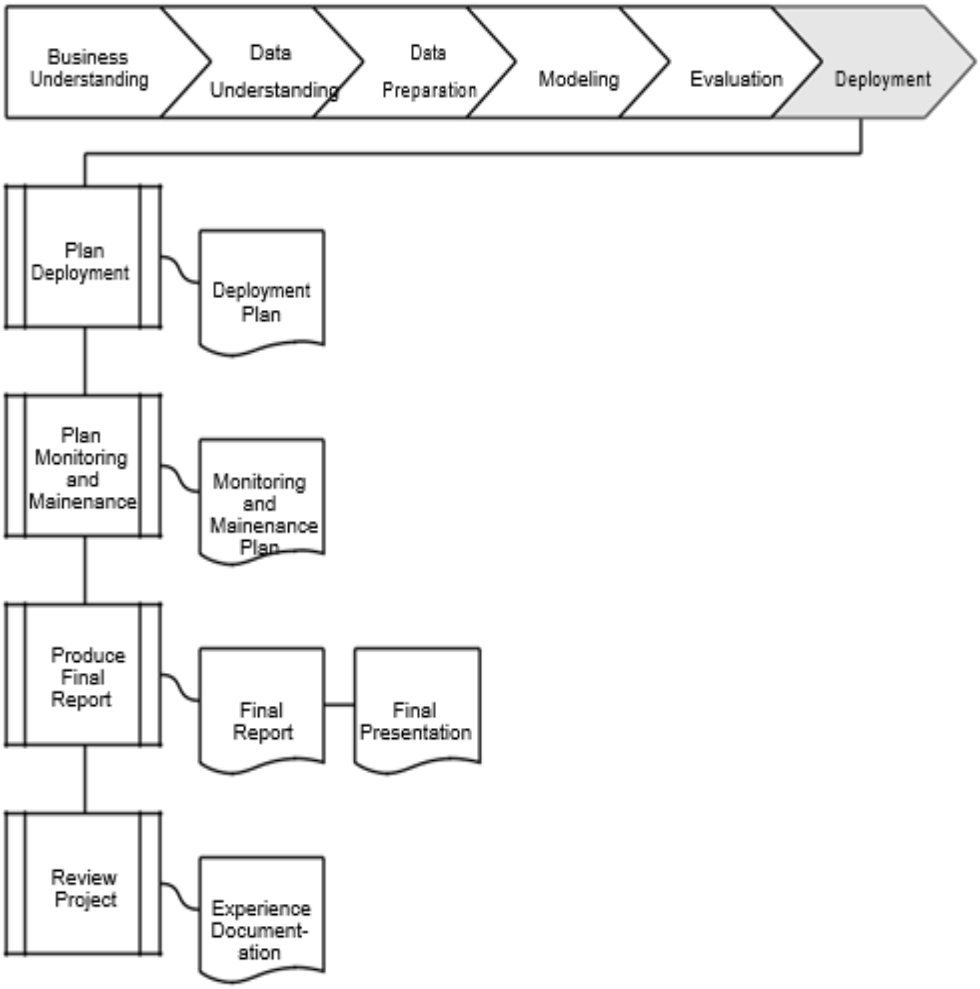


Fuente: (BizMetriks, 2013)**.**

En esta fase permite evaluará la calidad del modelo en base al análisis de métricas y estadísticas y comparará los resultados con los resultados anteriores, todo dependiendo de los resultados, y esta vez se determina o de acuerdo con la última parte del proceso. o Queda por volver a los pasos anteriores o volver a empezar desde el primer paso (B, 2015).

3.1.6. Despliegue.

Figura 11. **Despliegue con sus tareas y actividades.**



Fuente: (BizMetriks, 2013)**.**

Es valioso documentar los resultados obtenidos en las fases para la persona encargada pueda leer con claridad y también todas las fases sean documentadas con eso permitirá realizar una revisión final de todos los procesos que se obtiene o se seleccionó que se aprendió durante el desarrollo y con ellos pueden tomar una acción y aprovechar la oportunidad que puede dar el resultado (IBM, 2021).

3.2. Tipo de investigación.

3.2.1. Métodos de investigación

El Método Inductivo será de ayuda para sacar conclusiones generales a partir del conocimiento previos de la información histórica del Centro Médico CMC y usar la técnica de la observación para el análisis (José, 2015).

CAPÍTULO IV

4. Desarrollo de la Metodología CRISP-DM

La aplicación de la técnica se aplicará la metodología CRISP-DM para cuál se desarrolla todas las fases para la compresión de negocio, compresión de la información (datos), modelado, evaluación y despliegue.

Primero especificaremos la misión y visión del Centro Médico CMC.

- **Misión del Centro Médico CMC**

  "Servir con atención médica de calidad, en la búsqueda incondicional del bien pleno de nuestros pacientes" (Celina, 2006).

- **Visión del Centro Médico CMC**

  "Mejorar continuamente nuestro servicio para que sientan toda la confianza de poner su salud en nuestras manos (Celina, 2006).

4.1. COMPRESIÓN DEL NEGOCIO

4.1.1. DETERMINAR OBJETIVOS DE NEGOCIO

a) Antecedentes

"El Centro Médico CMC, ve a cada uno de los pacientes como miembros de su propia familia; y cómo tal, el personal entero vela con esmero y empatía por la salud de todos sus integrantes.

Lo que en el año 2006 empezó como una pequeña farmacia, con esfuerzo y gracias a ustedes, hoy en día se ha edificado como un Centro Médico con más de 15 especialidades.

Las puertas siempre estarán abiertas para usted y sus seres queridos. gracias por formar parte de esta familia" (Celina, 2006).

Ya que el Centro Médico CMC brinda los servicios en diferentes especialidades a cada uno de los pacientes que se acerca a lugar se ve la necesidad de predecir cual es el grupo de citas médicas de mayor demanda en laboratorio o para consulta médica.

En lo que respecta en fisioterapia se requiere identificar cuál de las alternativas son más frecuentemente usadas, si es la fisioterapia manual o la fisioterapia tecnológica

En cuanto a el monitoreo signos vitales se requiere identificar si asisten pacientes con signos vitales que presenta una alerta de emergencia o son paciente que no presenta alerta de emergencia con los resultados se podría cambiar el giro de negocio o mejorar la misma.

b) Objetivos de negocios

- **Predecir** la demanda dentro de citas médicas, tratamiento de fisioterapia y signos vitales con el fin de mejorar la atención para los pacientes.
- Conocer cuál fue el comportamiento de los pacientes en el año 2021 y 2022 en citas médicas y signos vitales; y en lo que respecta a tratamiento de fisioterapia durante 2019,2021 y 2022.
- Aumentar el nivel de satisfacción del paciente en base a los resultados que se obtengan del análisis de datos que se realice en este trabajo de titulación.

c) Criterios de éxito empresarial

- Aumentar el uso en citas médicas, tratamiento de fisioterapia y signos vitales dentro del Centro Médico CMC.
- Lograr fidelidad del paciente.
- Mejorar la satisfacción del paciente.

4.1.2. EVALUAR LA SITUACIÓN

a) Inventario de recursos

En la información proporcionada por el Centro Médico CMC se encontró un abanico de datos dentro del data set PATRONATO.xlsx que consta de 13 subDataset de información del cual se va a utilizar las siguientes subDatset, CITAS_MEDICAS, TRATAMIENTO_FISIOTERAPIA y SIGNOS_VITALES con sus respectivos atributos que lleva información adecuada de cada paciente que se acercan al Centro Médico y con esta información se puede aplicar la metodología CRISP-DM con sus respectivas etapas para identificar si el Centro Médico requiere implementar más recursos humanos o tecnológicos y así brindar una atención más adecuado según como arroje los resultados.

b) Requisitos, suposiciones y restricciones

- No se conoce el tipo de género en las consultas
- No se especifica si es niño o adulto
- Descomponer la fecha para obtener el año 2019, 2021 al 2022.
- No se conoce el rango de edad del paciente que más acuden al centro médico.
- Existe columnas concatenadas por coma.

c) Riesgos y contingencias

- Se desconoce la política del Centro Médico en citas médicas.
- Implementar un control de llenados de los campos en secuencias.
- Se desconoce los nombres de los doctores, auxiliares tratantes.

d) Terminología

Data set CITAS_MEDICAS con los siguientes atributos.

- Id_c = Código.
- Fecha_c = Fecha que realizó el trámite el paciente
- Hora = Hora que se realizó el trámite el paciente
- Detalle = Detalla un breve diagnóstico de cita médica.
- Tipo_Cita = Área de la cita médica.
- HISTORIA = Número de historia del paciente.
- id_doc = Código del doctor.
- Estado = Es si está pendiente o no de atención.
- Abono_c = Es el anticipo de su cita médica.
- Saldo_c = Es el saldo de la cita médica.
- Total_c = Total del valor de la cita médica.
- Fecha_ci = Fecha de transacción de la cita médica.

Data set TRATAMIENTO_FISIOTERAPIA con los siguientes atributos.

- Id_exa = Codigo.
- HISTORIA = Número de historia del paciente.
- Id_es = Codigo del especialista
- Id_grado = Grado de esguince.
- CI = Cédula del paciente.
- Nom_doc = Nombre del doctor.
- Diag_doc= Diagnostico del doctor.
- Diag_tera = Diagnóstico del tratamiento.
- Exa_comple = Exámenes solicitados o solicita.
- Deporte = Si debe realizar deporte o realiza deporte.
- Mati_cons = Descripción si hay presencia de dolor.
- Exa_fisico = Exploración al paciente.
- Trata_fisio = Descripción del tratamiento.
- Num_seciones = Cuantas secciones debe usar para la rehabilitación.
- Fecha_tera = Fecha que realizó el trámite el paciente.

- Hora_tera = Hora que realizo el trámite el paciente.

Data set SIGNOS_VITALES con los siguientes atributos.

- ID_APAR2 = Codigo.
- HISTORIA = Número de historia del paciente.
- CI = Cédula del paciente.
- TEMP = Temperatura tomada al paciente se clasificada en 1 como alerta de emergencia y 0 como no alerta de emergencia.
- TEM = Temperatura real tomada al paciente.
- PESO = Peso tomado al paciente.
- FRECU_CARDIA = Frecuencia cardiaca tomado al paciente.
- TENSION_ARTERIAL = Tensión arterial tomado al paciente
- FREC_RESP = Frecuencia respiratoria tomado al paciente.
- TALLA = Talla tomado al paciente.
- FECHA2 = Fecha que realizó el trámite el paciente
- HORA2 = Hora que le tomaron al paciente todos los signos vitales.
- SATURA = saturación tomada al paciente.

e) Costos y beneficios

El Centro Médico podrá mejor su atención al paciente y sobre todo tener prioridad en el uso de equipos que posee en las diferentes áreas que más demanda tenga.

Beneficiará a los pacientes con la atención más oportuna o rápida en las áreas que más demanda tenga.

4.1.3. DETERMINAR OBJETIVOS DE MINERÍA DE DATOS

a) Objetivos de minería de datos

- Obtener información valiosa que nos permita realizar un análisis comparativo dentro de los data set de citas médicas, tratamiento de fisio terapia y signos vitales para verificar cuál de los grupos es más frecuente.
- La base obtenida es de los años 2019, 2021 y 2022.
- Los tipos de problemas de minería de datos que se pueden analizar son predicciones tanto en citas médicas, tratamiento de fisioterapia y signos vitales.

b) Criterios de éxito de la minería de datos

Se deben considerar los siguientes criterios:

- Cumplir con los objetivos planteados.

- ❖ Alcanzar los mejores resultados de predicción entre la afluencia de personas que visitan al Centro Médico en el área de citas médicas, tratamiento de fisioterapia y signos vitales.
- ❖ Tener un porcentaje alto de clasificación de acuerdo el 20 % de prueba y el 80 % de entrenamiento.
- ❖ Realizar la limpieza adecuada y con coherencia de la información proporcionada por el Centro Médico CMC.
- ❖ Escoger un rango de años para tener una información proporcional a lo que se va a analizar.

4.1.4. PRODUCIR EL PLAN DE PROYECTO

a) Plan de proyecto

Tabla 3**. Plan de trabajo en la aplicación de la metodología CRISP-DM**

Fase	**Tiempo**	**Recursos**
Compresión del negocio	1 semanas	Analista
Compresión de los datos	3 semanas	Analista
Preparación de datos	5 semanas	Ejecución de minería de datos, análisis base de datos.
Modelado	2 semanas	Ejecución de minería de datos, análisis base de datos.
Evaluación	1 semanas	Analista
Despliegue	1 semanas	Consultoría de minería de datos análisis base de datos.

Fuente: Elaboración propia.

b) Evaluación inicial de herramientas y métodos.

Herramienta es Python y Jupyter Notebook y las librerías para analizar y realizar el modelo de predicción que sea más adecuado al desarrollo del trabajo de titulación.

4.2. COMPRESIÓN DE LOS DATOS

4.2.1. RECOGER DATOS INICIALES

a) Informe inicial de recopilación de datos

❖ Data set CITAS_MEDICAS

Figura 12. **Presentación de datos citas médicas.**

	Id_c	Fecha_c	Hora	Detalle	Tipo_Cita	HISTORIA	id_doc	Estado
0	1.0	2022-07-21	14:00:00	Cita médica : CONSULTAS : TRAUMATOLOGIA	TRAUMATOLOGIA	15070.0	19.0	Atendido - Cancelado
1	2.0	2022-07-21	17:00:00	Cita médica : CONSULTAS : MEDICINA INTERNA	MEDICINA INTERNA	3398.0	16.0	Atendido - Cancelado
2	3.0	2022-07-21	17:30:00	Cita médica : CONSULTAS : MEDICINA INTERNA	MEDICINA INTERNA	15061.0	16.0	Atendido - Cancelado
3	4.0	2022-07-22	15:00:00	Cita médica : CONSULTAS : TRAUMATOLOGIA	TRAUMATOLOGIA	15080.0	19.0	Atendido - Cancelado
4	5.0	2022-07-22	17:00:00	Cita médica : CONSULTAS : MEDICINA INTERNA CAN...	MEDICINA INTERNA	15084.0	16.0	Atendido - Cancelado
...	...	...	...	...	...	...	...	...
4717	4719.0	2023-08-25	18:34:32	ACNE SEVERO (CIE 10; L701)	NaN	8659.0	NaN	ATENDIDO

Fuente: Elaboración propia.

❖ Data set TRATAMIENTO_FISIOTERAPIA

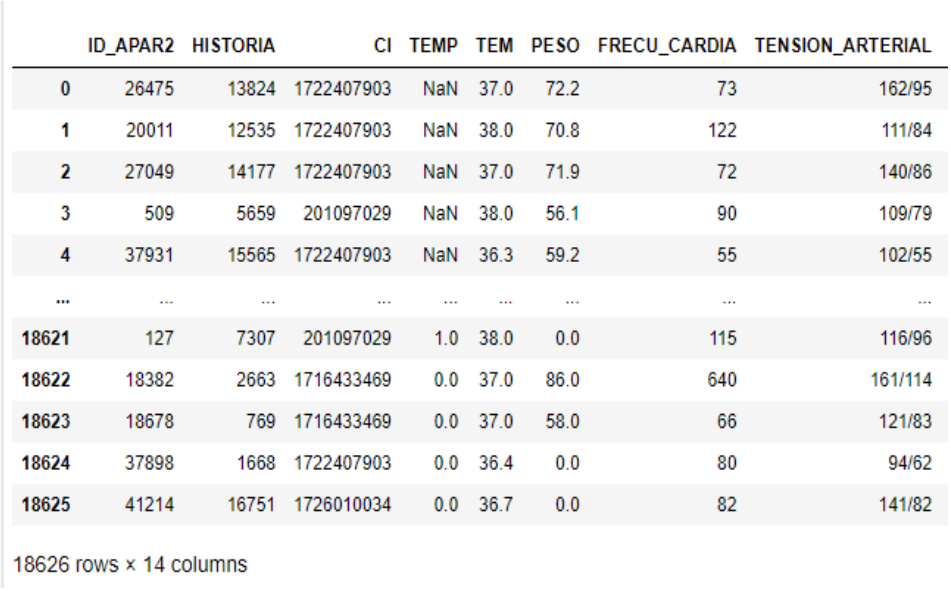
Figura 13. **Presentación de datos exámenes fisioterapia.**

	Id_exa	HISTORIA	Id_es	Id_grado	CI	Nom_doc	Diag_doc	Diag_tera	Exa_comple	Deporte	Mati_cons
0	228	11194	10	2	1721921300	RUTH MORENO	FRACTURA DE FEMUR PARTE SUPERIOS COXOARTROSIS	COXOARTROSIS OSTEOARTROSIS EDEMA DE MIEMBROS I...	RX	NINGUNO	DOLOR INTENSO EN EL AREA LUMBAR Y GLUTEA DE LA...
1	229	11194	6	4	1721921300	RUTH MORENO	FRACTURA DE FEMUR PARTE SUPERIOS COXOARTROSIS	COXOARTROSIS OSTEOARTROSIS EDEMA DE MIEMBROS I...	RX	NINGUNO	DOLOR INTENSO EN EL AREA LUMBAR Y GLUTEA DE LA...
2	239	11194	8	4	1721921300	RUTH MORENO	FRACTURA DE FEMUR PARTE SUPERIOS COXOARTROSIS	COXOARTROSIS OSTEOARTROSIS EDEMA DE MIEMBROS I...	RX	NINGUNO	DOLOR INTENSO EN EL AREA LUMBAR Y GLUTEA DE LA...
3	308	11294	10	3	1721921300	RUTH MORENO	ESGUINCE RESIDIVANTE	ESGUINCE RESIDIVANTE Y FASCITIS PLANTAR DEL PI...	RX	NINGUNO	DOLOR EN EL PIE IZQ
4	310	11294	7	3	1721921300	RUTH MORENO	ESGUINCE RESIDIVANTE	ESGUINCE RESIDIVANTE Y FASCITIS PLANTAR DEL PI...	RX	NINGUNO	DOLOR EN EL PIE IZQ

Fuente: Elaboración propia.

❖ Data set SIGNO_VITALES

Figura 14. **Presentación de datos de signo vitales.**



	ID_APAR2	HISTORIA	CI	TEMP	TEM	PESO	FRECU_CARDIA	TENSION_ARTERIAL
0	26475	13824	1722407903	NaN	37.0	72.2	73	162/95
1	20011	12535	1722407903	NaN	38.0	70.8	122	111/84
2	27049	14177	1722407903	NaN	37.0	71.9	72	140/86
3	509	5659	201097029	NaN	38.0	56.1	90	109/79
4	37931	15565	1722407903	NaN	36.3	59.2	55	102/55
...	...	...	...	...	...	...	...	...
18621	127	7307	201097029	1.0	38.0	0.0	115	116/96
18622	18382	2663	1716433469	0.0	37.0	86.0	640	161/114
18623	18678	769	1716433469	0.0	37.0	58.0	66	121/83
18624	37898	1668	1722407903	0.0	36.4	0.0	80	94/62
18625	41214	16751	1726010034	0.0	36.7	0.0	82	141/82

18626 rows × 14 columns

Fuente: Elaboración propia.

4.2.2. DESCRIBIR DATOS

b) Informe de descripción de datos

La información proporcionada por el Centro Médico se carga en el Júpiter Notebook que es una herramienta de uso libre que se maneja con el lenguaje Python para el análisis de datos. Los data set que se cargaran son los siguientes:

❖ Data set CITAS_MEDICAS

Se cuenta con 4722 columnas y 12 filas.

❖ Data set TRATAMIENTO_FISIOTERAPIA

Se cuenta con 1198 columnas y 16 filas

❖ Data set SIGNOS VITALES

Se cuenta con 18626 columnas y 13 filas

4.2.3. VERIFICAR LA CALIDAD DE LOS DATOS

a) Informe de calidad de datos

❖ Existe gran cantidad de datos no validos que contienen valores vacíos o están agrupados como sin respuesta, por ejemplo $null$.

❖ No se encontraron errores de datos de tipo tipográfico.

❖ No se encontraron errores e incoherencias de codificación.

❖ Se pierde datos por la falta de información en los tres datas set.

- Se presenta un solo registro de datos que son concatenados en una misma columna.

4.3. PREPARACIÓN DE LOS DATOS

a) Conjunto de datos

Los conjuntos de datos de las datas set se puede visualizar en la figura 41, 42 y 43

b) Descripción de conjunto de datos

En él data set citas médicas encontramos datos de tipo datatime de dos variables, float 6 variables y object 4 variables, existe una gran cantidad de valores nulos, y la mayoría son de datos float para realizar el análisis se va a descomponer la fecha en año, mes, días y algunas columnas de transformar de tipo object a numérico las variables cuantitativas.

En el data set de tratamiento de fisioterapia encontramos datos de tipo int con 6 variables y object 10 variables, existe una gran cantidad de valores nulos, y la mayoría son de datos object para realizar el análisis se va a descomponer la fecha en año, mes, día y transformar algunas variables de tipo object a numéricas y separar columnas que se encuentra concatenada que es de utilidad para el análisis.

En el data set de signos vitales encontramos datos de tipo float 5 variables, int 8 variables y object 3 variables, existe una gran cantidad de valores nulos, y la mayoría de los datos son de tipo int para realizar el análisis se va a descomponer la fecha en año, mes y día.

4.3.1. SELECCIONAR DATOS

- Para **predecir** la concurrencia de los tipos de citas médicas que se agrupa en laboratorio y consulta médica en el periodo 2021 al 2022, se utilizará las variables Tipo_cita, HISTORIA, Detalle, Estado y AÑO.
- Para la **predicción** según su tratamiento de fisioterapia en la alternativa de tratamiento manual o tratamiento tecnológico se usará las variables Trata_fisio1, Diag_doc, Diag_tera, Num_seciones y AÑO, de los años 2019, 2021 y 2022 para visualizar cual puede ser la alternativa más usada.
- Para la **predicción** según los signos vitales, al dato de la temperatura se lo clasificará en alerta de emergencia y alerta no emergencia de los años 2021 y 2022, para ello se utilizará las siguientes variables TEMP, FRECU_CARDIA, FREC_RESP, SATURA y AÑO.

4.3.2. LIMPIAR DATOS

a) Informe de limpieza de datos

Para proceder a realizar la limpieza de los datos del data set de citas médicas, se realizan las siguientes tareas:

- Colocar el rango de fecha del 2021 al 2022 Figura 49.
- Verificar los datos duplicados Figura 50.
- Eliminación de datos duplicados figura 51.
- Verificar valores únicos Figura 52.
- Cambiar al mismo nombre por minúsculas a mayúsculas y por faltas de ortografía Figura 53 y 54.
- Asignar con el valor de **0** al dato Tipo_cita, cuando la cita se solicita al área de laboratorio y el valor de **1** cuando la cita es para el área consulta médica se puede observar en la Figura 44.
- Transformar las variables object a numerico Figura 43.
- Diagrama de los mínimos y máximos para tener una idea cómo se comporta los datos en la Figura 55.
- Verificar valores nulos Figura 56.
- Diagrama de bigotes para ver si presenta outliers se puede evidenciar en la Figura 57 y 58.

Para proceder a realizar la limpieza de los datos del data set de tratamiento de fisioterapia, se realizan las siguientes tareas:

- Descomposición de la fecha en mes día y hora Figura 58.
- Colocar el rango de fecha del 2021 al 2022 Figura 59.
- Verificar valores nulos Figura 60.
- Eliminar valores nulos Figura 61.
- Asignar la columna a predecir Trata_fisio1 con el 0 a todos los tratamientos que se realiza con el apoyo tecnológico y el valor **1** si el tratamiento es manual como muestra la Figura 67.
- Tipo de datos Figura 62.
- Transformar los tipos de datos object a numérico Figura 63.
- Diagramar los mínimos y máximos para tener una idea cómo se comporta los datos en la Figura 64.
- Realizar un diagrama de bigotes para ver si presenta outliers se puede evidenciar en la Figura 65 y 66.

Para proceder a realizar la limpieza de los datos del data set de signos vitales, se realizan las siguientes tareas:

- Descomposición de la fecha en mes día y año Figura 68.
- Colocar el rango de fecha del 2021 al 2022 Figura 69.
- Tipo de variable Figura 70
- Remplazo de valores de **NAN** con la media Figura 71.
- Valores nulos en la Figura 72.
- Valores duplicados Figura 73.
- Verificar que los rangos presentados en la Tabla 1 presente con el valor 1 como alerta de emergencia y el valor 0 como no alerta de emergencia como se muestra en la Figura 77,78 y 79.
- Diagrama de los mínimos y máximos para tener una idea cómo se comporta los datos en la Figura 74.
- Diagrama de bigotes para ver si presenta outliers se puede evidenciar en la Figura 75 y 76.

4.3.3. CONSTRUIR DATOS

a) Atributos derivados

Los atributos derivados en citas médicas, tratamiento de fisioterapia y signos vitales se realiza remplazando en algunos casos los valores NAN por la media se presenta en las figuras 71.

b) Registros generados

Con los registros nuevos en los tres datas set se podrá realizar el análisis de predicción de los años 2019, 2021 y 2022.

4.4. MODELADO

4.4.1. SELECCIONA TÉCNICA DE MODELADO

a) Técnica de modelo

Objetivo de negocio

Predecir mediante la clasificación del modelo de datos, cuál de las siguientes áreas: citas médicas, tratamiento de fisioterapia y signos vitales tiene más afluencia de pacientes que asisten al Centro Médico. CMC.

En el modelo vamos a dividir el dataset en “entrenamiento” 80% y el dataset “Pruebas” 20% y comparamos con los datos para probar la precisión de nuestros modelos.

Técnica seleccionada

Para ello se realizan regresión logística múltiple y árbol de decisión. Aquí hay algunas condiciones que pueden afectar a sus opciones:

¿" Se requiere que los datos para el modelo se dividan en grupo de entrenamiento y prueba" (IBM, 2021)**?**

Se necesita que los datos de entrenamiento se deben encontrar en este análisis del 80%, se utilizara para entrenar cada modelo. La calidad de nuestro modelo de aprendizaje automático dependerá de la calidad de los datos.

Los datos de prueba es el 20 % restante, que son los datos guardados para probar, cuando se crea el modelo a partir de los datos de entrenamiento.

¿" Dispone de datos suficientes para traer resultados fiables para un modelo determinado" (IBM, 2021)**?**

Se dispone de 1741 entradas de citas médicas, de 451 entrada en el data set de tratamiento de fisioterapia y 3231 datos de entrada en el data set signos vítale. Son datos limpios para aplicar los dos algoritmos mencionados anteriormente y por ende se garantiza que los resultados son totalmente fiables para el modelo que se construye.

¿" Solicita el modelo un cierto nivel de calidad de datos? ¿Logra alcanzar este nivel con los datos que dispone" (IBM, 2021)**?**

Para que el teorema central comience a funcionar y las estimaciones sean estables, se necesita un mínimo de 30 datos. En los algoritmos escogidos para la clasificación también necesitaremos una pequeña cantidad de casos dependiendo de la diferencia en la entrada. Se dice que más de 30 casos, se requieren al menos 10 casos para cada cambio adicional.

- La data PATRONATO con sus hojas de CITAS_MEDICA, TRATAMIENTO_FISIOTERAPIA Y SIGNOS_VITALES cuenta con este nivel de calidad de datos por lo cual si se puede realizar la regresión logística múltiple y árbol de decisión.

  ¿" Los datos son de tipo correcto para un modelo concreto? En caso contrario, ¿consigue realizar las conversiones necesarias utilizando nodos de manejo de datos" (IBM, 2021)**?**

  Se tuvo que realizar algunas acciones sobre la base de datos entre las cuales tenemos:

  - Transformamos varias variables de tipo object a numérico.
  - Eliminamos las variables que no vamos a usar.

- Descomponer columnas concatenadas.
- Cambio de valores a 0 y 1.

b) Supuestos de modelado

En el data set de citas médicas describe 12 características, las variables que se utilizará son:

- Detalle: se detalla el contenido de la cita
- Tipo_cita: clasificación de consulta médica y laboratorio.
- HISTORIA = Número de historia del paciente.
- Estado = Es si está pendiente o no de atención.
- AÑO: año que es gestionado la cita.

Para predecir se asigna la variable Tipo_cita, que está representada con **1** cuando la cita es solicitada en el área de consultas médicas, y con el valor de **0** cuando la cita es para el área de laboratorio, con la descripción de las siguientes variables: Detalle, HISTORIA, Estado y AÑO.

En el data set tratamiento de fisioterapia describe 16 características, las variables que se utilizará son:

- Diag_doc: es el diagnóstico del doctor frente al paciente.
- Diag_tera = diagnóstico del tratamiento.
- Trata_fisio1: tratamiento a realizar
- Num_Seciones: cuantas secciones debe usar para la rehabilitación.
- AÑO: que gestiona el tratamiento.

Para predecir se asignará la variable Trata_fisio1, donde se incorpora el valor 1 para el tratamiento manual y el valor 0 para el tratamiento que se realiza mediante el apoyo tecnológico, con las variables: Diag_doc, Diag_tera, Num_seciones y AÑO.

En el data set signos vitales describe 14 características, las variables que se utilizará son:

- TEMP = Temperatura tomada al paciente en 0 y 1.
- FRECU_CARDIA: Frecuencia cardiaca tomado al paciente.
- FREC_RESP = Frecuencia respiratoria tomado al paciente.
- SATURA = saturación tomada al paciente.
- AÑO: Año que el paciente es atendido.

Para lo cual se puede predecir la variable TEMP, que consta de **1** cuando tiene una alerta de emergencia y el valor **0** cuando no tiene la alerta de

emergencia, con las variables: FRECU_CARDIA, FREC_RESP, SATURA y AÑO.

4.4.2. GENERAR DISEÑO DE PRUEBA

a) **Diseño de prueba**

Para crear el diseño de comprobación, se considera lo siguiente:

¿Cuál es el porcentaje que se utilizó para comprobar los modelos?

Se utilizará el 20% para comprobar la precisión de cada uno de los modelos.

¿Ocurre partición de datos en conjuntos de entrenamiento y prueba?

Los datos se van a dividir en 80% para entrenamiento y 20% para pruebas en cada uno de los data set.

¿Cómo puede medir el rendimiento de modelos de regresión logística múltiple y árbol de decisión?

Para medir el rendimiento del modelo usaremos las siguientes medidas con los criterios establecidos a continuación:

Accuracy: La precisión es una de las métricas que se utiliza para evaluar el rendimiento de un modelo de clasificación.

Se determina con el número de elementos pronosticados correctamente obtenidos del número total de elementos, en el conjunto de datos y Matriz de confusión

¿Cuál es el número de veces que piensa ejecutar un modelo con los valores ajustados antes de ejecutar otro tipo de modelo?

Los datos que serán utilizados para el modelo podrán ser ajustados tres veces antes de intentar con otro modelo diferente, la elección de las tres veces se realiza por la siguiente razón:

Primer ajuste: se elimina los outliers de las variables independientes.

Segundo ajuste: se realiza un escalamiento de la variable, para observar si en la correlación entre variables, la variable dependiente aumenta.

Tercer ajuste: se realiza un escalamiento a la variable independiente, para observar si en la correlación de la variable dependiente aumenta.

Una vez realizados estos tres ajustes a los datos, si no se llega a alcanzar los criterios se deberá usar otro grupo de variables, para un nuevo modelo o cambiar aquella que tenga la menor correlación y todo esto se repite para cada uno de los data set.

4.4.3. CONSTRUIR MODELO

a) Configuración de parámetros

Después de la limpieza de los datos se escogen las siguientes variables en citas médicas, en el eje **x** refiere: HISTORIA, Detalle, Estado y AÑO y en eje de las **y** Tipo_Cita, con esta distribución permitirá aplicar el proceso de los modelos de regresión logística múltiple y árbol de decisión.

Figura 15. **Presenta la distribución para la aplicación de los dos modelos en citas médica.**

```
# split data
from sklearn.model_selection import train_test_split
x= df[['HISTORIA','Detalle', 'Estado','AÑO']]
y = df.Tipo_Cita
print(x)
print(y)
```

```
      HISTORIA  Detalle  Estado   AÑO
206     3002.0      231       0  2021
222       20.0     1219       0  2021
234     2892.0     1203       0  2021
235    10223.0      170       0  2021
236    10243.0      170       0  2021
...        ...      ...     ...   ...
4233    8617.0     1083       0  2021
4236   18451.0     1216       0  2021
4662    6272.0      856       0  2021
4666   18754.0      238       0  2021
4676   17157.0       47       0  2021

[1741 rows x 4 columns]
206     1
222     1
234     0
235     0
236     0
       ..
4233    1
4236    1
4662    1
4666    1
4676    1
Name: Tipo_Cita, Length: 1741, dtype: object
```

Fuente: Elaboración propia.

Después de limpiar los datos se escogen las siguientes variables en tratamiento de fisioterapia, en el eje **x** se describe: Diag_doc, Diad_tera,

Num_Seciones y AÑO y en eje de las **y** Trata_fisio1, con esta distribución permitirá aplicar el proceso de los modelos de regresión logística múltiple y árbol de decisión.

Figura 16. **Presenta la distribución para la aplicación de los dos modelos en tratamiento de fisioterapia.**

```
# split data
from sklearn.model_selection import train_test_split
x= df[['Diag_doc','Diag_tera','Num_seciones','AÑO']]
y = df['Trata_fisio1']
print(x)
print(y)
```

```
      Diag_doc  Diag_tera  Num_seciones   AÑO
0          109        137            10  2022
1          109        137            10  2022
2          109        137            10  2022
3           98        212            10  2022
4           98        212            10  2022
...        ...        ...           ...   ...
1049       217        214            10  2019
1050       217        214            10  2019
1051       217        214            10  2019
1052       217        214            10  2019
1053       217        214            10  2019

[451 rows x 4 columns]
0       1
1       1
2       1
3       0
4       0
       ..
1049    1
1050    1
1051    1
1052    1
1053    1
Name: Trata_fisio1, Length: 451, dtype: object
```

Fuente: Elaboración propia.

Después de limpiar los datos se escogen las siguientes variables en signos vitales, en el eje **x** se especifica: FRECU_CARDIA, FREC_RESP, SATURA Y AÑO y en eje de las **y** TEMP, con esta distribución permitirá

aplicar el proceso de los modelos de regresión logística múltiple y árbol de decisión.

Figura 17. **Presenta la distribución para la aplicación de los dos modelos en signo vitales.**

```
# split data
from sklearn.model_selection import train_test_split
x= df[['FRECU_CARDIA','FREC_RESP','SATURA','AÑO']]
y = df['TEMP']
print(x)
print(y)
```

```
       FRECU_CARDIA  FREC_RESP  SATURA   AÑO
23              104         20    91.0  2021
25               88         95    95.0  2021
27               92         78    91.0  2021
36               91          0    97.0  2021
43               63         19    91.0  2022
...             ...        ...     ...   ...
16034           100         20    89.0  2022
16035           105         20    90.0  2022
16036            78         20    94.0  2022
18612           101          0    87.0  2022
18625            82          0    90.0  2021

[3231 rows x 4 columns]
23       0.0
25       1.0
27       0.0
36       1.0
43       1.0
        ...
16034    0.0
16035    0.0
16036    0.0
18612    0.0
18625    0.0
Name: TEMP, Length: 3231, dtype: float64
```

Fuente: Elaboración propia.

b) Modelo

Se aplicará el modelo de regresión logística múltiple y el modelo árbol de decisión, donde se puede observar los resultados de la aplicación de cada uno de los modelos y la gráfica del modelo de árbol de decisión. Se muestra en la Figura 46 de citas médicas, Figura 47 de tratamiento de fisioterapia, Figura 48 de signos vitales.

Figura 18. **Resultado regresión logística múltiple en citas médica.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train.head())
```

```
      HISTORIA  Detalle  Estado   AÑO
1133    5027.0      310       0  2021
1087   10056.0     1041       0  2021
4059   18415.0      661       0  2021
875    16956.0      962       0  2021
532    12330.0     1002       0  2021
```

```
from sklearn import linear_model
Classifier = linear_model.LogisticRegression()
Classifier.fit(X_train, y_train)
# make predictions for test data
y_pred = Classifier.predict(X_test)

print(list(y_test))
print(list(y_pred))
```

```
['1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '0', '1', '1', '1', '1', '1', '1', '1', '1
'1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1', '1', '1', '0', '1', '1'
'0', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1'
'1', '1', '1', '1', '0', '0', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1'
'1', '0', '1', '1', '1', '0', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1', '0', '1', '1', '1'
'1', '0', '0', '1', '1', '0', '1', '1', '1', '0', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1'
'1', '1', '0', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1'
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1'
['1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1
'1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1'
```

Fuente: Elaboración propia.

Figura 19. **Resultado de árbol de decisión en citas médica.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train.head())
```

```
      HISTORIA  Detalle  Estado   AÑO
1133    5027.0      310       0  2021
1087   10056.0     1041       0  2021
4059   18415.0      661       0  2021
875    16956.0      962       0  2021
532    12330.0     1002       0  2021
```

```
from sklearn.tree import DecisionTreeClassifier
Classifier = DecisionTreeClassifier(max_depth=8)
#Classifier = DecisionTreeClassifier()
Classifier_arbol=Classifier.fit(X_train, y_train)
# make predictions for test data
confusion_matrix(y_test, y_pred)
y_pred1 = Classifier_arbol.predict(X_test)
print(y_pred1)
```

```
['1' '1' '1' '1' '1' '1' '1' '1' '0' '1' '0' '1' '1' '1' '1' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1']
```

Fuente: Elaboración propia.

Figura 20. **Resultados de regresión logística múltiple en tratamiento de fisioterapia.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train.head())
```

```
     Diag_doc  Diag_tera  Num_seciones       AÑO
52  -0.897124   0.225100      0.077123  0.737485
101 -0.897124   0.225100      0.077123 -1.535447
26  -0.104862   0.696967      0.077123  0.737485
45   0.917411  -0.158292      0.077123  0.737485
151  1.083530  -1.170840      0.077123  0.737485
```

```
from sklearn import linear_model
Classifier = linear_model.LogisticRegression()
Classifier.fit(X_train, y_train)
# make predictions for test data
y_pred = Classifier.predict(X_test)

print(list(y_test))
print(list(y_pred))
```

```
['0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0', '1
0', '0', '0', '0', '0', '0', '1', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '1',
', '0', '0', '0', '1', '0', '1', '1', '0', '0', '0', '1', '0', '1', '0', '1', '0', '0',
'0', '0', '1', '1', '1', '0', '0', '0', '1', '1', '1', '0', '0', '0', '1', '1', '0', '0'
['0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0
0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0',
```

Fuente: Elaboración propia.

Figura 21. **Resultado de árbol de decisión en tratamiento de fisioterapia.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train.head())
```

```
     Diag_doc  Diag_tera  Num_seciones       AÑO
52  -0.897124   0.225100      0.077123  0.737485
101 -0.897124   0.225100      0.077123 -1.535447
26  -0.104862   0.696967      0.077123  0.737485
45   0.917411  -0.158292      0.077123  0.737485
151  1.083530  -1.170840      0.077123  0.737485
```

```
from sklearn.tree import DecisionTreeClassifier
Classifier = DecisionTreeClassifier(max_depth=8)
#Classifier = DecisionTreeClassifier()
Classifier_arbol=Classifier.fit(X_train, y_train)
# make predictions for test data
y_pred1 = Classifier_arbol.predict(X_test)
confusion_matrix(y_test, y_pred1)
print(y_pred1)
```

```
['0' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1'
 '0' '0' '0' '0' '1' '0' '1' '0' '0' '0' '0' '0' '1' '0' '0' '0' '0' '0'
 '0' '0' '0' '0' '0' '0' '1' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1' '0'
 '0' '0' '0' '0' '0' '0' '0' '0' '0' '0' '0' '0' '0' '0' '1' '0' '0' '0'
```

Fuente: Elaboración propia.

Figura 22. **Resultados de regresión logística múltiple en signos vitales.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train.head())
```

```
       FRECU_CARDIA  FREC_RESP  SATURA   AÑO
1093             76          0    87.0  2022
271              71         20    92.0  2021
15184            77          0    92.0  2022
13623            88         20    93.0  2021
7629             85          0    91.0  2022
```

```
from sklearn import linear_model
Classifier = linear_model.LogisticRegression()
Classifier.fit(X_train, y_train)
# make predictions for test data
y_pred = Classifier.predict(X_test)

print(list(y_test))
print(list(y_pred))
```

```
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0
0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0,
0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0
```

Fuente: Elaboración propia.

Figura 23. **Resultado de árbol de decisión en signos vitales.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train.head())
```

```
       FRECU_CARDIA  FREC_RESP  SATURA   AÑO
1093             76          0    87.0  2022
271              71         20    92.0  2021
15184            77          0    92.0  2022
13623            88         20    93.0  2021
7629             85          0    91.0  2022
```

```
from sklearn.tree import DecisionTreeClassifier
#Classifier = DecisionTreeClassifier()
Classifier = DecisionTreeClassifier(max_depth=8)
Classifier.fit(X_train, y_train)
# make predictions for test data
y_pred1 = Classifier.predict(X_test)
print(y_pred1)
```

```
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
```

Fuente: Elaboración propia.

c) Descripción del modelo

El modelo de regresión logística múltiple y el modelo de árbol de decisión permiten predecir valores futuros basados en información histórica, para lo cual se utilizará el lenguaje Python que facilitará el uso de varias librerías para el desarrollo del trabajo de titulación.

Las variables que se utilizó para este análisis son: para citas médicas: HISTORIA, Detalle, Estado, AÑO y Tipo_Cita. Para tratamiento de fisioterapia: Diag_doc,diag_tera, Num_Seciones, AÑO y Trata_fisio1, y finalmente en signos vitales se utilizó: FRECU_CARDIA, FREC_RESP, SATURA, AÑO y TEMP, mismas que permiten entrenar y probar los modelos. La información se encuentra en el siguiente archivo “PATRONATO.xlsx".

4.4.4. MODELO DE EVALUACIÓN

a) Evaluación

La evaluación de los modelos generado está más orientada hacia los objetivos de minería de datos, pero en el siguiente paso de la metodología, la evaluación está más orientada hacia los objetivos comerciales. Ambos

objetivos están estrechamente relacionados entre sí. en este trabajo de titulación.

Desde una perspectiva de minería de datos, el uso de predicciones de modelos de regresión logística múltiple y árboles de decisión es una buena forma de evaluar la efectividad del modelo. Se refleja que el modelo que da como resultado el valor más alto, es el árbol de decisión, con el resultado de predicción de un 93.70% para las citas médicas. El otro modelo presento un valor menor lo cual se refleja en la Tabla 4.

Figura 24. **Evaluación del modelo matriz de confusión en citas médicas.**

```
# evaluamos el modelo con la matriz de confusión
cm = confusion_matrix(y_test, y_pred1)
print(cm)
accuracy=accuracy_score(y_test,y_pred1)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
[[  7  21]
 [  1 320]]
Accuracy: 93.70%
```

Fuente: Elaboración propia.

El modelo más alto en tratamiento de fisioterapia es el modelo árbol de decisión, con un resultado de 83.52% de predicción. El otro modelo presento un valor menor lo cual se refleja en la Tabla 4.

Figura 25. **Evaluación del modelo matriz de confusión en tratamiento de fisioterapia.**

```
# evaluamos el modelo con la matriz de confusión
cm = confusion_matrix(y_test, y_pred1)
print(cm)
accuracy=accuracy_score(y_test,y_pred1)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
[[65  3]
 [12 11]]
Accuracy: 83.52%
```

Fuente: Elaboración propia.

El modelo más alto en signos vitales es el modelo de regresión logística múltiple, con un resultado del 89.95% de predicción. El otro modelo presento un valor menor lo cual se refleja en la Tabla 4.

Figura 26. **Evaluación del modelo matriz de confusión en signos vitales**.

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score

# Evaluamos con la matriz de confusión
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy=accuracy_score(y_test,y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
[[582   0]
 [ 65   0]]
Accuracy: 89.95%
```

Fuente: Elaboración propia.

b) Revisado / configuración

Para elaborar de manera correcta la regresión logística múltiple y árbol de decisión, es ideal la distribución o configuración total del dataset, el 80% de entrenamiento y el 20% que se usa para pruebas, con ello se puede tener los resultados de predicción del modelo.

Figura 27. **Partir los datos en grupos de entrenamiento y prueba de citas médicas.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train)
```

```
      HISTORIA  Detalle  Estado   AÑO
1133    5027.0      310       0  2021
1087   10056.0     1041       0  2021
4059   18415.0      661       0  2021
875    16956.0      962       0  2021
532    12330.0     1002       0  2021
...        ...      ...     ...   ...
1414    6515.0     1046       0  2021
3707   17259.0     1221       0  2021
4146   16697.0      833       0  2021
1135    8027.0      561       0  2021
1262   18479.0      404       0  2021

[1392 rows x 4 columns]
```

Fuente: Elaboración propia.

Figura 28. **Partir los datos en grupo de entrenamiento y prueba de tratamiento de fisioterapia.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train)
```

```
     Diag_doc  Diag_tera  Num_seciones       AÑO
52  -0.897124   0.225100      0.077123  0.737485
101 -0.897124   0.225100      0.077123 -1.535447
26  -0.104862   0.696967      0.077123  0.737485
45   0.917411  -0.158292      0.077123  0.737485
151  1.083530  -1.170840      0.077123  0.737485
..        ...        ...           ...       ...
888 -0.092084  -1.013551      0.077123 -1.535447
226  0.252933  -0.816939      0.077123  0.737485
146  0.201820  -1.200331      0.077123  0.737485
47  -0.897124   0.225100      0.077123  0.737485
206 -1.753278  -1.161009      0.077123  0.737485
```

Fuente: Elaboración propia.

Figura 29. **Partir los datos en grupo de entrenamiento y prueba de signos vitales.**

```
# split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)
print(X_train)
```

```
       FRECU_CARDIA  FREC_RESP  SATURA   AÑO
1093             76          0    87.0  2022
271              71         20    92.0  2021
15184            77          0    92.0  2022
13623            88         20    93.0  2021
7629             85          0    91.0  2022
...             ...        ...     ...   ...
2163             60         20    89.0  2022
2409             54         18    90.0  2022
14015            93         18    91.0  2021
15242            62          0    96.0  2022
15402            78          0    94.0  2022

[2584 rows x 4 columns]
```

Fuente: Elaboración propia.

4.5. EVALUACIÓN

4.5.1. EVALUAR RESULTADOS

La evaluación del modelo de clasificación del data set de citas médicas que está dividido en consultas médicas y laboratorio, dio como resultado una precisión del 91.98% con el modelo de regresión logística múltiple y 93.70% con el modelo de árbol de decisión, de esta forma el segundo modelo obtuvo mejor resultado de predicción.

Como resultado de la evaluación del modelo de clasificación para tratamiento de fisioterapia en tratamiento manual y tecnológico, se obtuvo una precisión del 74.73% con el modelo de regresión logística múltiple y 83.52% con el modelo de árbol de decisión, de esta forma el segundo modelo obtuvo mejor resultado de predicción.

Como resultado de la evaluación del modelo de clasificación para signos vitales clasificado en emergencia y no emergencia, se obtuvo una precisión del 89.95% con el modelo de regresión logística múltiple y 89.34% el modelo de árbol de decisión, de esta forma el primer modelo obtuvo mejor resultado de predicción.

a) Evaluación de los resultados de la minería de datos

Para evaluar la clasificación, se consideran los siguientes parámetros: **Precisión de clasificación:** determina el porcentaje de ejemplos correctamente clasificados.

Precisión: "Usando métricas de precisión, podemos medir la calidad de los modelos de aprendizaje automático en tareas de clasificación" (Heras, 2020).

Figura 30. **Fórmula para el cálculo de la precisión.**

$$precision = \frac{TP}{TP + FP}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Precisión (precision)

Fuente: (Heras, 2020)

Figura 31. **Cálculo de accuracy.**

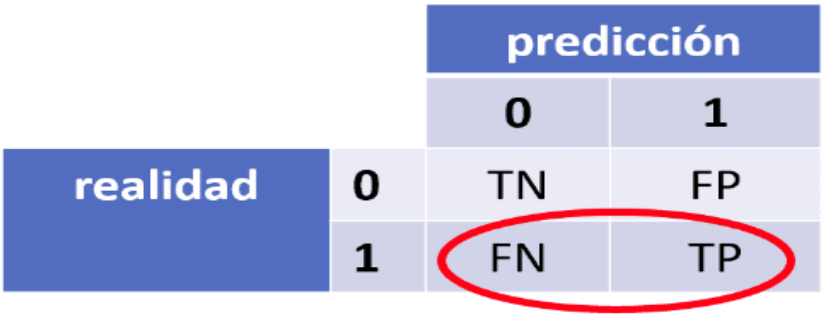
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Fuente: (Heras, 2020)

Recordar (sensibilidad o exhaustividad): Esta es la igualdad de verdaderos positivos a todo positivo. Si su puntaje es 1, entonces se encontró un verdadero positivo en el conjunto de datos, por lo que no hay ruido. Por el contrario, si su valor es cero, los datos no están correlacionados (Heras, 2020).

Figura 32. **Fórmula para el cálculo de exhaustividad (recall).**

$$recall = \frac{TP}{TP + FN}$$



Exhaustividad (recall)

Fuente: (Heras, 2020)

Tabla 4. **Resumen entre regresión logística múltiple y árbol de decisión de evaluación de matriz de confusión.**

Data set	**Regresión logística múltiple**	**Árbol de decisión**
Citas médicas	[[0 28] [0 321]] Accuracy: 91.98%	[[7 21] [1 320]] Accuracy: 93.70%
Tratamiento de fisioterapia	[[68 0] [23 0]] Accuracy: 74.73%	[[65 3] [12 11]] Accuracy: 83.52%
Signos vitales	[[582 0] [65 0]] Accuracy: 89.95%	[[575 7] [62 3]] Accuracy: 89.34%

Fuente: Elaboración propia.

Matriz de confusión

Las matrices de confusión nos permiten observar el rendimiento de los algoritmos utilizados en el entrenamiento supervisado.

Figura 33. **Matriz de confusión en citas médicas.**

	precision	recall	f1-score	support
0	0.00	0.00	0.00	28
1	0.92	1.00	0.96	321
accuracy			0.92	349
macro avg	0.46	0.50	0.48	349
weighted avg	0.85	0.92	0.88	349

Fuente: Elaboración propia.

Figura 34. **Matriz de confusión en tratamiento de fisioterapia.**

	precision	recall	f1-score	support
0	0.75	1.00	0.86	68
1	0.00	0.00	0.00	23
accuracy			0.75	91
macro avg	0.37	0.50	0.43	91
weighted avg	0.56	0.75	0.64	91

Fuente: Elaboración propia.

Figura 35. **Matriz de confusión en signos vitales.**

	precision	recall	f1-score	support
0.0	0.90	1.00	0.95	582
1.0	0.00	0.00	0.00	65
accuracy			0.90	647
macro avg	0.45	0.50	0.47	647
weighted avg	0.81	0.90	0.85	647

Fuente: Elaboración propia.

4.5.2. PROCESO DE REVISIÓN

a) Revisión del proceso

- Continuar con la fase de despliegue
- Volver atrás y mejorar o sustituir el modelo.

Una vez que los resultados obtenidos en la definición y construcción del modelo son satisfactorios, el investigador confía en la precisión y relevancia de los resultados del proyecto y pasan a la fase de implementación. Ahora esperan los resultados del informe final y el visto bueno de las políticas del Centro Médico CMC.

4.6. DESPLIEGUE

4.6.1. IMPLEMENTACIÓN DEL PLAN

a) Plan de empleo

- Este paso es importante para garantizar que la aplicación esté completamente integrada y tenga una interrupción mínima en el sistema actual. Es necesario saber cómo se ampliarán los sistemas existentes para incluir la nueva información, cómo se cambiarán los procesos de acuerdo con la nueva información del negocio, si se controlarán los recursos humanos necesarios para implementar los cambios.
- Para crear un plan de implementación, es importante resumir primero los resultados. Luego se crea un modelo de dispersión para cada modelo de minería de datos aplicado, si se considera útil. También es práctico desarrollar un plan para publicar los resultados de cada nuevo descubrimiento.

  Para aprovechar con éxito los resultados de la minería de datos del investigador, la información adecuada debe llegar a las personas adecuadas.

- **Gerentes.** Los gerentes deben recibir información sobre las recomendaciones, los cambios propuestos y una breve explicación de cómo afectarán estos cambios. Si acepta los resultados del estudio, debe notificarlo a la persona responsable de implementar estos cambios.
- **Analistas.** Las personas responsables de proteger la información del paciente deben considerar nuevas políticas o procedimientos para el centro médico. Se le debe informar de los posibles cambios relacionados con la formación adicional, para que pueda empezar a trabajar lo antes posible.
- **Experto en bases de datos.** Los administradores de la base de datos están familiarizados con el entorno de entrega de especialidad del paciente, cómo se usa la información en la base de datos y qué atributos se pueden agregar a la base de datos en proyectos futuros que se necesiten. Lo que es más importante, el equipo del proyecto necesita el aporte de todos estos grupos

para coordinar la implementación de los resultados y planes futuros para el proyecto.

4.6.2. SEGUIMIENTO Y MANTENIMIENTO DEL PLAN

b) Monitoreo y mantenimiento

Además de la planificación, los procesos de gestión y mantenimiento también deben planificarse en función de los resultados esperados, especialmente en términos de ahorro de costes y sostenibilidad. A medida que los proyectos futuros conduzcan a modelos más complejos, aumentará la necesidad de control. Siempre que sea posible, las actividades de seguimiento deben automatizarse con informes estructurados que puedan revisarse periódicamente. La construcción de modelos predictivos también puede ser la dirección que una empresa quiera tomar. Esto requiere herramientas más avanzadas que un proyecto básico de minería de datos. El propósito de este trabajo de titulación es demostrar la efectividad de la minería de datos como una forma de comprender mejor las grandes bases de datos relacionadas con actividades del centro médico.

4.6.3. PRODUCIR INFORME FINAL

a) Reporte final

- Actualmente, el mercado de la medicina se enfrenta a una fuerte competencia. La predicción de la asistencia de pacientes se ha transformado en un tema importante de la gestión de relaciones con el paciente, para que el paciente tenga la confianza de acercarse al Centro Médico CMC. Por lo tanto, al realizar una investigación, se comprenderán bien los factores clave de que especialidades tiene más acogidas, si es en consultas o laboratorio para poder derivar los pacientes de forma más precisa a las demás especialidades.
- La gestión adecuada de la asistencia de los pacientes permitirá ser un Centro Médico que pueda prestar de mejor manera la atención en cada una de las especialidades que se ofrece.
- Los pacientes satisfechos pueden traer nuevos pacientes.
- Los pacientes que mayor tiempo tiene en el Centro Médico no se irán con facilidad, porque no se dejarán influenciar mucho por la competencia.
- Los pacientes de mayor tiempo que se hacen atender, pueden acudir con mayor seguridad a las demás especialidades que se oferta en el Centro Médico CMC.
- El Centro Médico puede centrarse en satisfacer las necesidades de los pacientes existentes y los nuevos.

- ❖ Los pacientes que se van pueden compartir experiencias negativas, por lo tanto, tendrán una influencia negativa en la imagen del centro médico.
- ❖ La retención de paciente en función de, por ejemplo, (precio, oferta de especialidades y satisfacción del paciente) podría conducir a una mejor lealtad del paciente.
- ❖ En estudios futuros, se probarán diferentes métodos para comparar diferentes algoritmos y la precisión de su modelo.

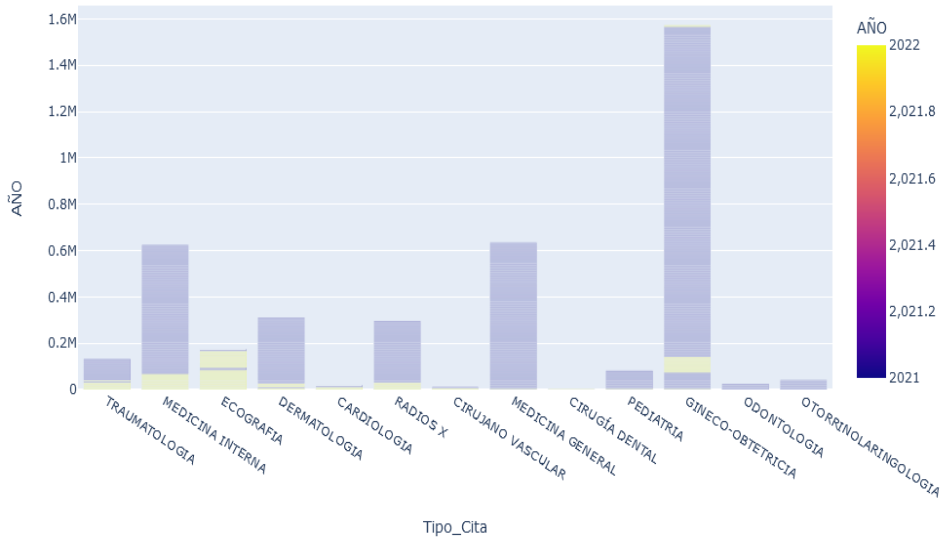
b) Presentación final

Este estudio muestra, los diferentes análisis realizados tanto en consultas médicas-laboratorio, tratamiento manual, tratamiento tecnológico y finalmente signos vitales de emergencia y no emergencia, se observa que en citas médicas donde tiene más acogida es en GINECOLOGÍA-OBSTETRICIA.

En el segundo data set de tratamiento de fisioterapia se escoge el total de datos que tiene el data set, porque si se tomaba solo datos del periodo 2021 al 2022 representaba muy poca información, para lo cual se incorporó el año 2019, dando como resultado, la alternativa terapia tecnológica en el área de MAGNETOTERAPIA.

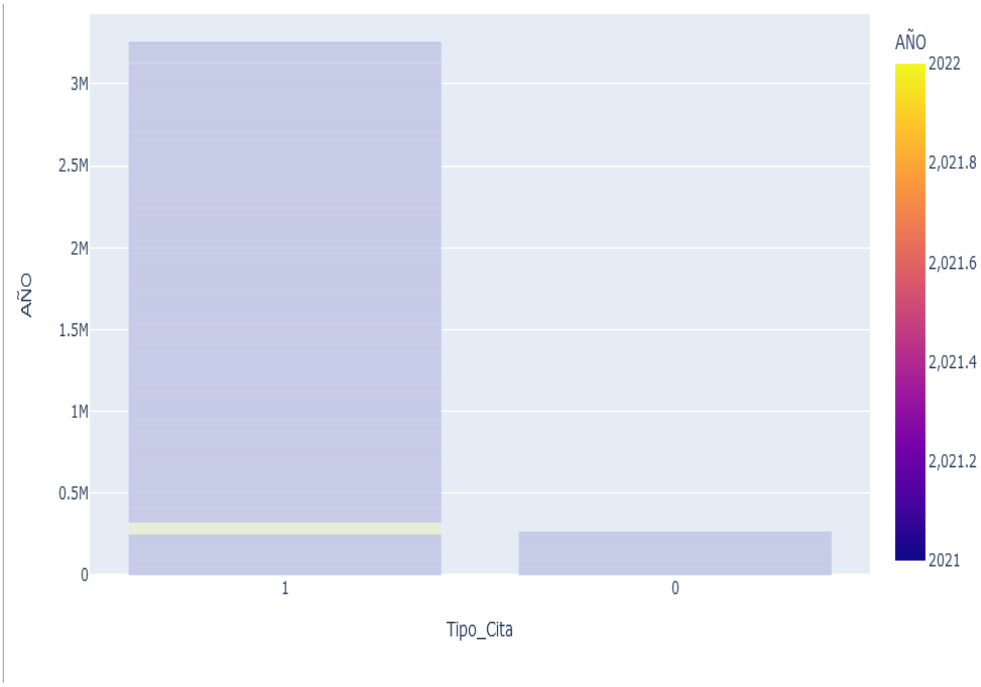
Y en el último data set en el periodo 2021 al 2022, se obtuvo más información en la temperatura de alerta de emergencia como resultado de la predicción.

Figura 36. **Presenta la afluencia en las especialidades que oferta el centro médico.**



Fuente: Elaboración propia.

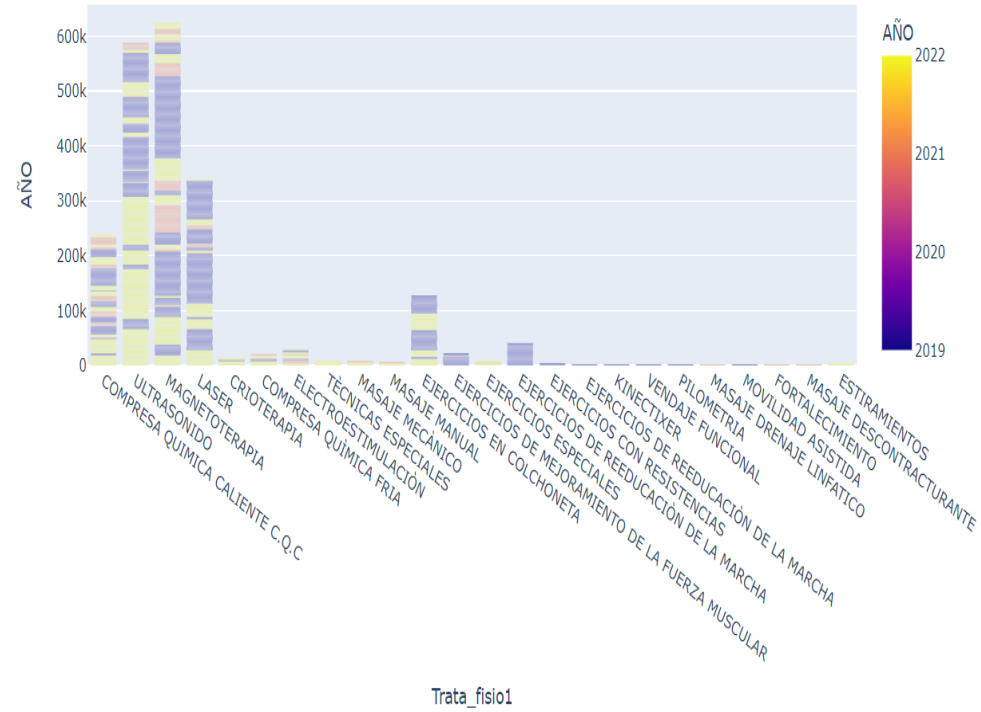
Figura 37. **Presenta el resultado de 0 y 1 de citas médicas.**



Fuente: Elaboración propia.

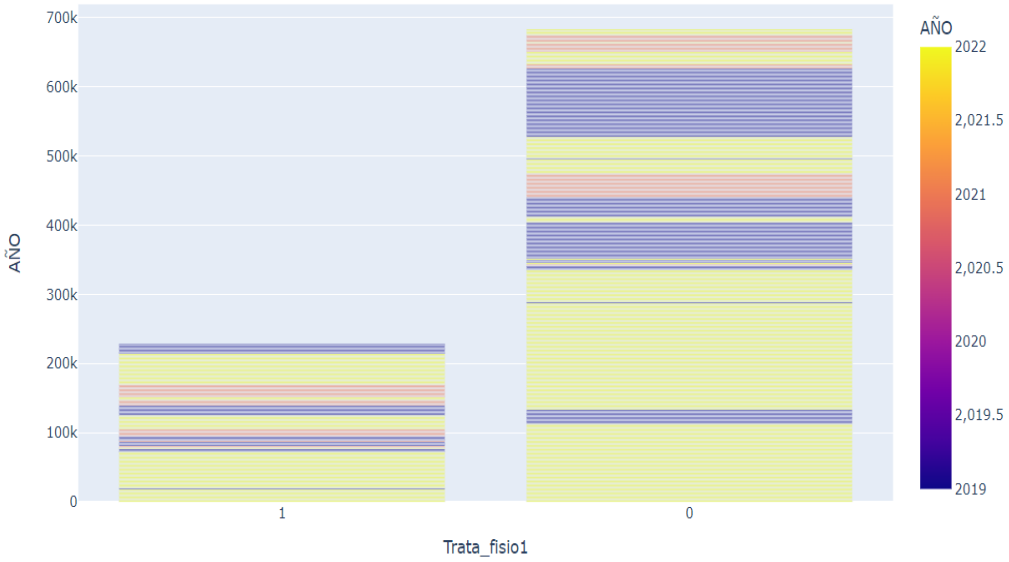
Se puede mencionar que las consultas médicas son las que tiene mayor afluencia de pacientes en relación a la atención en laboratorio y la consulta médica que tiene más acogida es ginecología-obstetricia.

Figura 38. **Presenta el tratamiento de fisioterapia que cada paciente, que es atendido.**



Fuente: Elaboración propia.

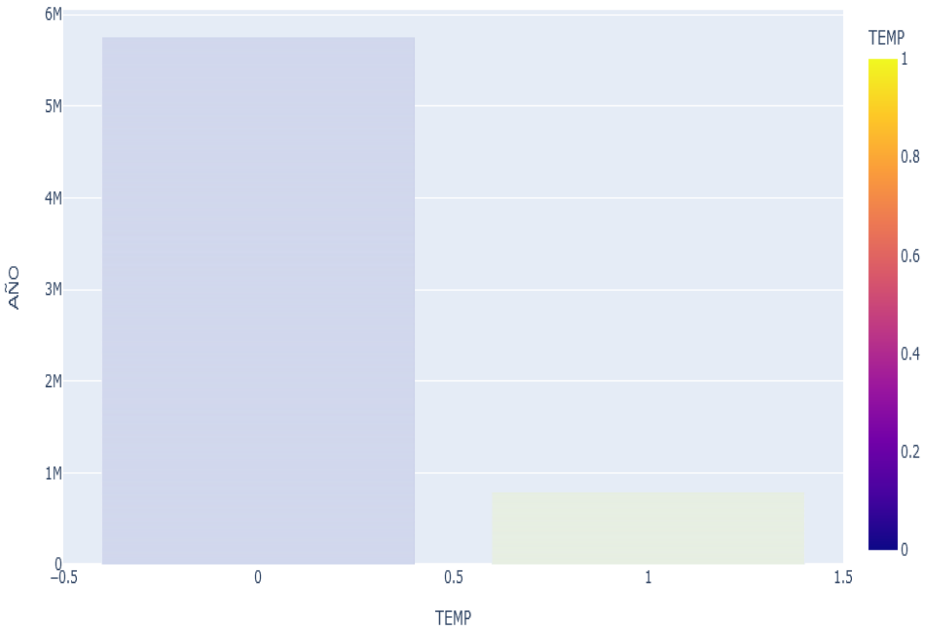
Figura 39. **Presenta el tratamiento de fisioterapia que cada paciente, que es atendido en el grupo de tratamiento manual que es 1 y 0 tratamiento tecnológico.**



Fuente: Elaboración propia.

En el estudio de tratamiento de fisioterapia, refleja que el tratamiento tecnológico supera al tratamiento manual y esto se puede concluir que existe mayor acogida en el tratamiento anteriormente mencionado.

Figura 40. **Presenta los resultados de los signos vitales de pacientes que fueron atendidos en el año 2021 al 2022, con temperatura de alerta de emergencia es representa con el valor 1 y 0 con no alerta de emergencia.**



Fuente: Elaboración propia.

En el estudio de signos vitales, refleja que en los años 2021 y 2022 hay pacientes que asisten al Centro Médico para ser atendidos normalmente porque no presenta una alerta de emergencia, tal como muestra en la Tabla 1, en los dos casos es clasificado como un signo de no emergencia, pero no obstante se puede dejar de lado la alerta de emergencia, si se presenta este caso el Centro Médico debe estar preparado y dispuesto atender de manera rápida y eficaz.

CAPÍTULO V

5. CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

A partir del trabajo de titulación del análisis de datos y la ejecución de la metodología CRISP-DM se puede concluir lo siguiente:

- Se concluye que el proceso de clasificación y análisis de las grandes cantidades de datos que maneja el Centro Médico CMC, relacionados con pacientes, historiales médicos, diagnósticos, tratamientos y más, pueden mejorarse mediante el uso de inteligencia artificial y especialmente mediante el uso de métodos de aprendizaje automático con ayuda de la metodología CRISP-DM.
- Se concluye que, en el caso del Centro Médico CMC, si se pueden utilizar técnicas de machine learning para clasificar los datos de los pacientes específicamente en las variables citas médicas, tratamiento de fisioterapia y signos vitales. Esto permitió tener una mejor comprensión de los patrones y tendencias en los datos, a su vez ayudará a mejorar la calidad y eficiencia de las atenciones médicas.
- Con los resultados de este trabajo de titulación, se puede predecir que los valores más influyentes del centro médico CMC se relaciona con la variable citas, que se clasificó en consultas médica y de laboratorio. Con los resultados se puede observar que existe mayor demanda en las consultas médicas, en la especialidad ginecología-obstetricia. Respecto a la variable de tratamiento de fisioterapia que se clasifica en fisioterapia manual y tecnológica una vez corrido el modelo, se observa que la mayor demanda de pacientes requiere de fisioterapia tecnológica y específicamente en magnetoterapia y ultrasonidos. Y en relación a signos vitales se clasifica en alerta de emergencia y no emergencia y luego de correr el modelo, la mayor demanda es en pacientes que asisten con alerta de no emergencia.
- Con relación al tiempo se puede concluir que las citas médicas tienen mayor demanda en el año 2021 que en el año 2022. Con respecto a la variable signos vitales se identifica que en el año 2021 se cuenta con un mayor número de pacientes que han presentado una temperatura de no emergencia, en relación al año 2022. Y con la variable de fisioterapia se identifica que en los años 2019 y 2021 existe mayor demanda en fisioterapia tecnológica por parte de los pacientes, en relación al año 2022.

5.2. RECOMENDACIONES

A partir del trabajo de titulación del análisis de datos y la aplicación de la metodología CRISP-DM se puede recomendar lo siguiente:

- Al Centro Médico, que desarrolle políticas de análisis de datos de forma continua o al menos anual, con el fin de comprender el comportamiento de sus líneas de negocio, con la ayudad de inteligencia artificial, clasificación y predicción de datos y que con los resultados que se obtengan se tomen mejores decisiones y acciones oportunas para mejorar los servicios.
- Que se designe en el Centro Médico a una persona o a un equipo de trabajo que se encargue de mantener actualizado el modelo generado en este trabajo de titulación, y que se definan la periodicidad de aplicar éste para ayudar a optimizar la gestión.
- Que se agreguen otras variables al modelo, de importancia para identificar comportamientos que pueden estar afectando la calidad de los servicios del Centro Médico.
- Que se defina un estándar para el manejo y almacenamiento de los datos e información que se generan de forma diaria, de tal forma que sean fáciles de utilizarlos en el modelo,

BIBLIOGRAFÍA

Abello, J. (2019). *La Inteligencia Artificial Es La Combinación De Algoritmos Planteados Con El Propósito De Crear Máquinas Que Presenten Las Mismas Capacidades Que El Ser Humano*. Obtenido de https://www.calameo.com/books/00586323924bced8a5411

AEFOL, E. (18 de 07 de 2022). *4 TIPOS DE TAREAS DE CLASIFICACIÓN EN EL MACHINE LEARNING*. Obtenido de https://elearningactual.com/4-tipos-de-tareas-de-clasificacion-en-el-machine-learning/

Apd. (04 de 04 de 2019). *¿Cuáles son los tipos de algoritmos del machine learning?* Obtenido de https://www.apd.es/algoritmos-del-machine-learning/

AWS. (2022). *¿Qué es la ciencia de datos?* Obtenido de https://acortar.link/zLTe55

B, J. (2015). *Metodología Fundamental IBM.* Obtenido de https://www.ibm.com/downloads/cas/6RZMKDN8

BizMetriks. (2013). *Los datos se están convirtiendo en la nueva materia prima de los negocios.* Obtenido de http://www.bizmetriks.com/metodologia.html

Cecco, C. N. (02 de 09 de 2021). *¿Cómo puede la inteligencia artificial mejorar la salud de los latinoamericanos?* Obtenido de https://acortar.link/k70wKt

Celina. (2006). *Centro Médico Celina*. Obtenido de https://medicelina.com/nosotros#contenido

Data, B. (12 de 06 de 2023). *Procesos de Análisis de Datos*. Obtenido de https://acortar.link/ADo0FG

Estrella, À. (2020). *Aplicaciones basadas en aprendizaje automatico (machine learnin) en plataforma de bajo consumo.* Obtenido de https://oa.upm.es/66520/1/TFG_ALVARO_ESTRELLA_OLIVA.pdf

Figueiras, S. (20 de 09 de 2021). *¿CONOCES JUPYTER NOTEBOOK?* Obtenido de https://www.ceupe.mx/blog/conoces-jupyter-notebook.html

Gil, D. Á. (14 de 01 de 2021). *Metodología CRISP-DM.* Obtenido de https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/

Heras, J. M. (10 de 09 de 2020). *Precision, Recall, F1, Accuracy en clasificación*. Obtenido de https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

IBM. (18 de 08 de 2021). *CRISP-DM en IBM SPSS Modeler*. Obtenido de https://acortar.link/kZ9yl4

José, M. (21 de 09 de 2015). *La metodología de la investigación*. Obtenido de https://www.gestiopolis.com/la-metodologia-de-la-investigacion/

Luther, M. (16 de 01 de 2020). *4 Metodologías para proyectos de Data Science – INVESTIGACIÓN DATLAS.* . Obtenido de https://acortar.link/vrDVok

Martínez, C. G. (5 de 2018). *REGRESIÓN LOGÍSTICA (Simple y Múltiple)*. Obtenido de https://rstudio-pubs-

static.s3.amazonaws.com/388799_ac1988bada0143d4a4cef0847e3605f8.html#regresi%C3%B3n_log%C3%ADstica_m%C3%BAltiple

Martinez, J. (10 de 10 de 2020). *Librerías de Python para Machine Learning*. Obtenido de 10: https://www.iartificial.net/librerias-de-python-para-machine-learning/

Pedrero, V. (1 de 2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. Obtenido de https://www.scielo.cl/scielo.php?pid=S0034-98872021000200248&script=sci_arttext

Quantum. (20 de 08 de 2019). *Metodologías de gestión de proyectos de Data Science*. Obtenido de https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb

Recuero, P. (2 de 12 de 2021). *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*. Obtenido de https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

Rollins, I. J. (06 de 2015). *Metodología Fundamental* . Obtenido de https://www.ibm.com/downloads/cas/6RZMKDN8

Saini, A. (29 de 08 de 2021). *Algoritmo de árbol de decisión: una guía completa*. Obtenido de https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

Santiago. (02 de 2021). *Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia*. Obtenido de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-98872021000200248

Santiago, A. (16 de 06 de 2022). *Control de temperatura por medios físicos – Enfermería*. Obtenido de https://yoamoenfermeriablog.com/2020/02/12/control-de-temperatura/

School, T. (11 de 10 de 2022). *¿Qué es Python? Te contamos todo sobre este popular lenguaje*. Obtenido de https://www.tokioschool.com/noticias/que-es-python/

unesco. (14 de 03 de 2023). *Recomendación sobre la ética de la inteligencia artificial*. Obtenido de https://eduteka.icesi.edu.co/articulos/unesco-etica-de-la-inteligencia-artificial

Uniteco. (6 de 4 de 2022). *EL MACHINE LEARNING EN MEDICINA, EL FUTURO DE LA PROFESIÓN*. Obtenido de https://acortar.link/8vdXYz

ANEXOS

Anexo de citas medicas

Figura 41. **Información de grupos de datos de citas médicas.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4722 entries, 0 to 4721
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Id_c        4719 non-null   float64
 1   Fecha_c     4719 non-null   datetime64[ns]
 2   Hora        4719 non-null   object
 3   Detalle     4715 non-null   object
 4   Tipo_Cita   4189 non-null   object
 5   HISTORIA    4719 non-null   float64
 6   id_doc      195 non-null    float64
 7   Estado      4719 non-null   object
 8   Abono_c     195 non-null    float64
 9   Saldo_c     195 non-null    float64
 10  Total_c     481 non-null    float64
 11  Fecha_ci    166 non-null    datetime64[ns]
dtypes: datetime64[ns](2), float64(6), object(4)
memory usage: 442.8+ KB
```

Fuente: Elaboración propia.

Figura 42. **Información de conjunto de datos de tratamiento de fisioterapia.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1198 entries, 0 to 1197
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Id_exa          1198 non-null   int64
 1   HISTORIA        1198 non-null   int64
 2   Id_es           1198 non-null   int64
 3   Id_grado        1198 non-null   int64
 4   CI              1198 non-null   int64
 5   Nom_doc         1058 non-null   object
 6   Diag_doc        1059 non-null   object
 7   Diag_tera       1058 non-null   object
 8   Exa_comple      1059 non-null   object
 9   Deporte         1058 non-null   object
 10  Mati_cons       1058 non-null   object
 11  Exa_fisico      1059 non-null   object
 12  Trata_fisio     1053 non-null   object
 13  Num_seciones    1198 non-null   int64
 14  Fecha_tera      1198 non-null   object
 15  Hora_tera       1198 non-null   object
dtypes: int64(6), object(10)
memory usage: 149.9+ KB
```

Fuente: Elaboración propia.

Figura 43. **Información de conjunto de datos de signos vitales.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18626 entries, 0 to 18625
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID_APAR2           18626 non-null  int64
 1   HISTORIA           18626 non-null  int64
 2   CI                 18626 non-null  int64
 3   TEMP               18512 non-null  float64
 4   TEM                18626 non-null  float64
 5   PESO               18626 non-null  float64
 6   FRECU_CARDIA       18626 non-null  int64
 7   TENSION_ARTERIAL   18509 non-null  object
 8   PERIME_FALICO      18626 non-null  float64
 9   FREC_RESP          18626 non-null  int64
 10  TALLA              18626 non-null  float64
 11  FECHA2             18626 non-null  object
 12  HORA2              18626 non-null  object
 13  SATURA             5203 non-null   float64
dtypes: float64(6), int64(5), object(3)
memory usage: 2.0+ MB
```

Fuente: Elaboración propia.

Figura 44. **Transformar de tipo object a numérica.**

```
#1
from sklearn.preprocessing import LabelEncoder
l=LabelEncoder()
for col in df.Estado:
    if df['Estado'].dtype=='object':
        df['Estado']=l.fit_transform(df['Estado'])
```

```
#tranformarmos de tipo object a numerico las variables cuantitativas
from sklearn.preprocessing import LabelEncoder
l=LabelEncoder()
for col in df.Detalle:
    if df['Detalle'].dtype=='object':
        df['Detalle']=l.fit_transform(df['Detalle'])
```

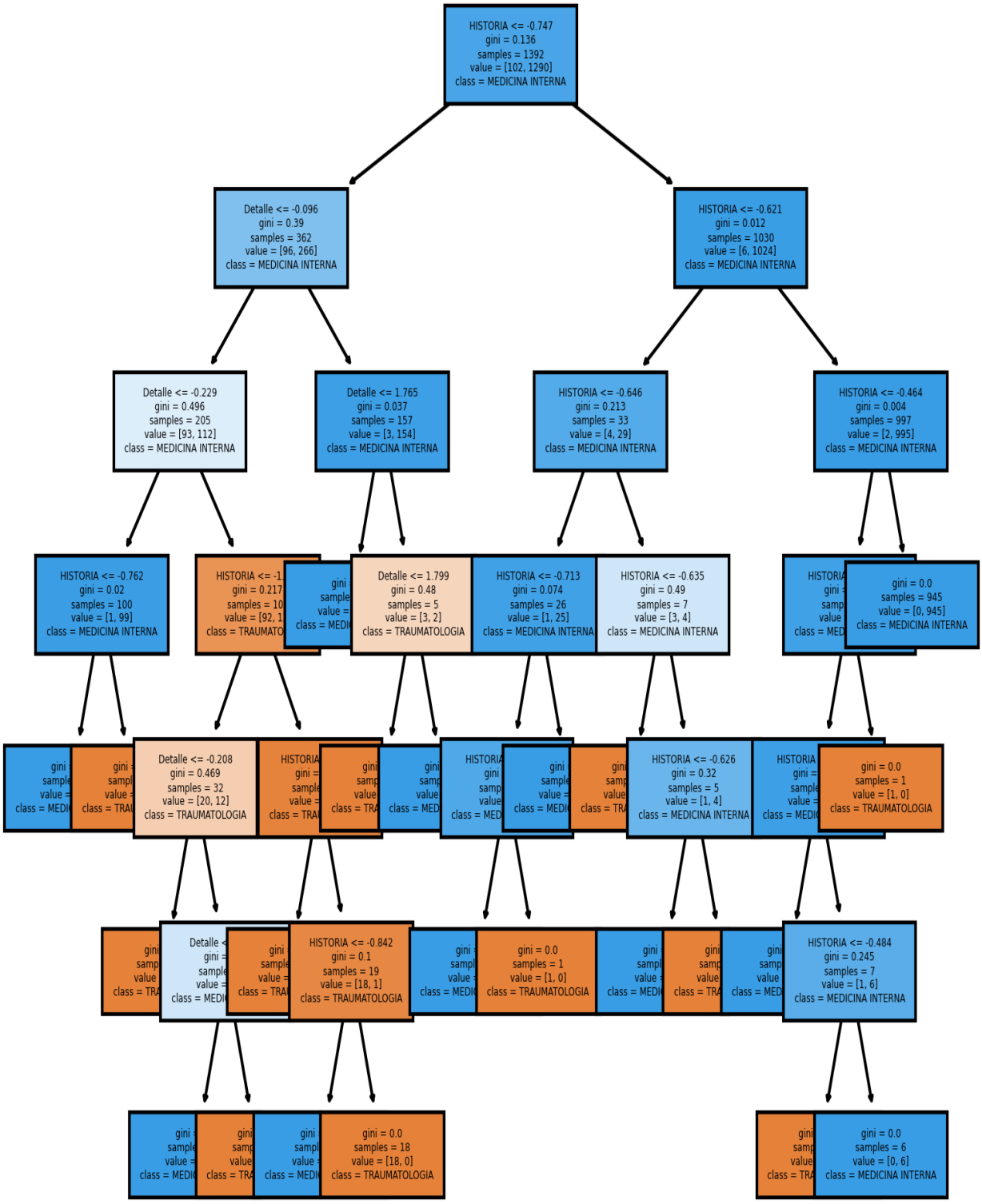
Fuente: Elaboración propia.

Figura 45. **Transformar de tipo object a numérica.**

```
df=df.replace('TRAUMATOLOGIA','1')
df=df.replace('MEDICINA INTERNA','1')
df=df.replace('ECOGRAFIA','1')
df=df.replace('DERMATOLOGIA','1')
df=df.replace('CARDIOLOGIA','1')
df=df.replace('MEDICINA GENERAL','1')
df=df.replace('PEDIATRIA','1')
df=df.replace('GINECO-OBTETRICIA','1')
df=df.replace('ODONTOLOGIA','1')
df=df.replace('OTORRINOLARINGOLOGIA','1')
df=df.replace('PEDRIATIA','1')
df=df.replace('RADIOS X','0')
df=df.replace('CIRUGÍA DENTAL','0')
df=df.replace('CIRUJANO VASCULAR','0')
```

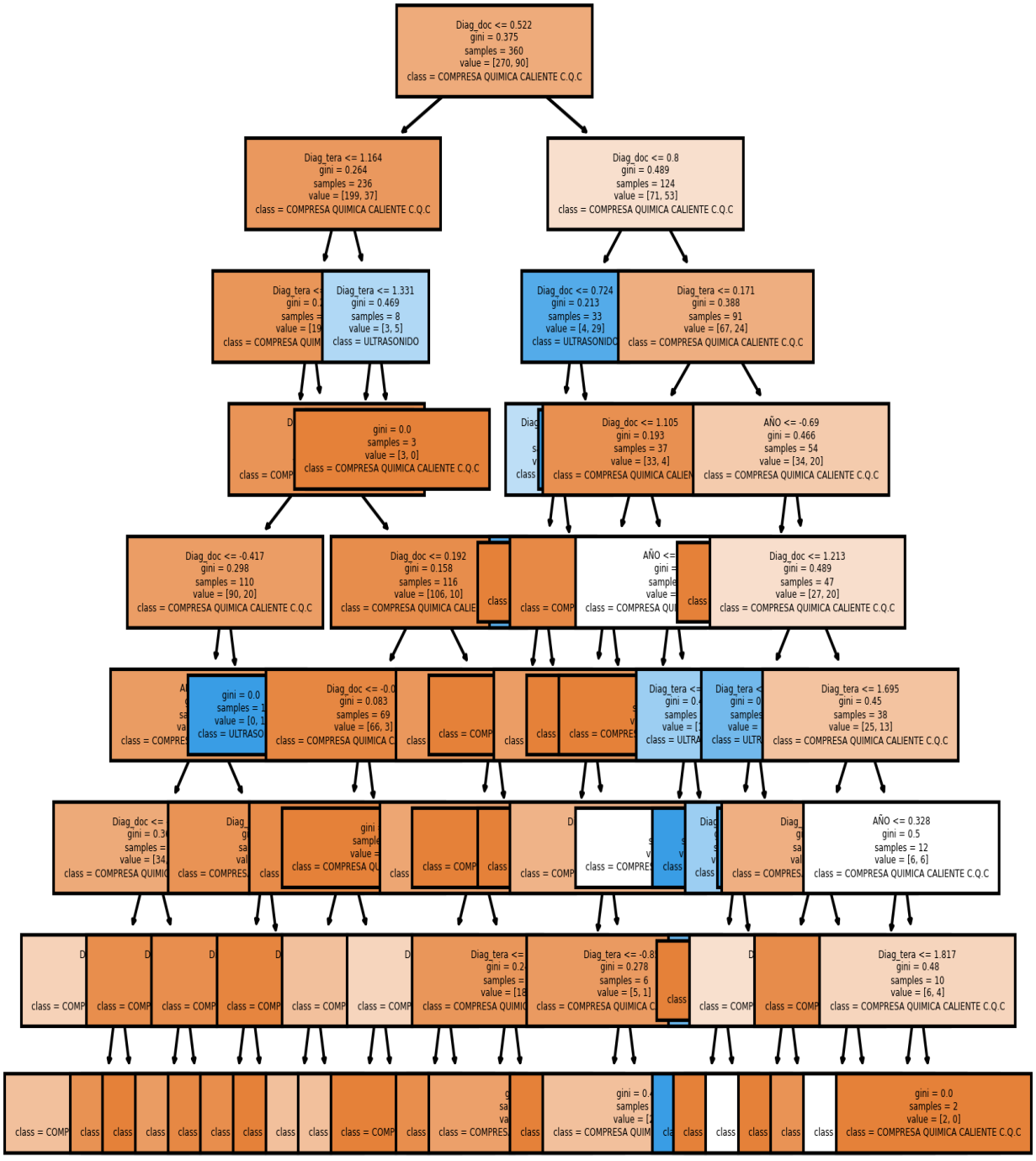
Fuente: Elaboración propia.

Figura 46. **Representación gráfica del modelo árbol de decisión en citas médica.**



Fuente: Elaboración propia.

Figura 47. **Representación gráfica del modelo árbol de decisión de tratamiento de fisioterapia.**



Fuente: Elaboración propia.

Figura 48. **Representación gráfica del modelo árbol de decisión en signos vitales.**

Fuente: Elaboración propia.

Figura 49. **Muestra las nuevas columnas de la descomposición de la fecha en citas médicas.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4722 entries, 0 to 4721
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Fecha_c    4719 non-null   datetime64[ns]
 1   Detalle    4715 non-null   object
 2   Tipo_Cita  4189 non-null   object
 3   HISTORIA   4719 non-null   float64
 4   Estado     4719 non-null   object
 5   AÑO        4719 non-null   float64
 6   MES        4719 non-null   float64
 7   DIA        4719 non-null   float64
dtypes: datetime64[ns](1), float64(4), object(3)
memory usage: 295.2+ KB
```

Fuente: Elaboración propia.

Figura 50. **Muestra los datos duplicados.**

	Fecha_c	Detalle	Tipo_Cita	HISTORIA	Estado	AÑO	MES	DIA
56	2022-09-26	Cita médica : CONSULTAS : DERMATOLOGIA	DERMATOLOGIA	14442.0	Pendiente - cancelado	2022	9	26
80	2022-11-27	Cita médica : CONSULTAS : TRAUMATOLOGIA	TRAUMATOLOGIA	12115.0	Pendiente - cancelado	2022	11	27
313	2021-12-23	EXODONCIA	RADIOS X	9960.0	ATENDIDO	2021	12	23
529	2021-09-25	RINOFARINGITIS AGUDA	MEDICINA GENERAL	10374.0	ATENDIDO	2021	9	25
2281	2021-12-18	LUMOCIATALGIA _x000D_\nD/C INFECCION DE VIAS U...	GINECO-OBTETRICIA	9538.0	ATENDIDO	2021	12	18

Fuente: Elaboración propia.

Figura 51. **Muestra la eliminación de los datos duplicados.**

```
#eliminamos los duplicados
df.drop_duplicates(keep=False,inplace=True)
```

Fuente: Elaboración propia.

Figura 52. **Muestra los valores únicos del tipo de cita médica.**

```
Tipo_Cita
['TRAUMATOLOGIA' 'MEDICINA INTERNA' 'MEDICINA INTERNA ' 'ECOGRAFIA'
 'DERMATOLOGIA' 'CARDIOLOGIA' 'RADIOS X' 'CIRUJANO VASCULAR'
 'MEDICINA GENERAL' 'CIRUGÍA DENTAL' 'PEDIATRIA' 'OBSTETRICIA'
 'ODONTOLOGIA' 'GINECO-OBTETRICIA' 'DERMADERMATOLOGIA'
 'OTORRINOLARINGOLOGIA' 'MEDICINA GENERALES' 'TRAMATOLOGIA']
18

Estado
['Atendido - Cancelado' 'Pendiente - cancelado' 'Pendiente - saldo'
 'PENDIENTE' 'Atendido' 'ATENDIDO']
6
```

Fuente: Elaboración propia.

Figura 53. **Muestra el cambio de nombre y los espacio.**

```
df=df.replace('DERMADERMATOLOGIA','DERMATOLOGIA')
df=df.replace('TRAMATOLOGIA','TRAUMATOLOGIA')
df=df.replace('MEDICINA INTERNA ','MEDICINA INTERNA') #espacio
df=df.replace('MEDICINA GENERALES','MEDICINA GENERAL') #espacio
df= df.replace('OBSTETRICIA','GINECO-OBTETRICIA')
```

Fuente: Elaboración propia.

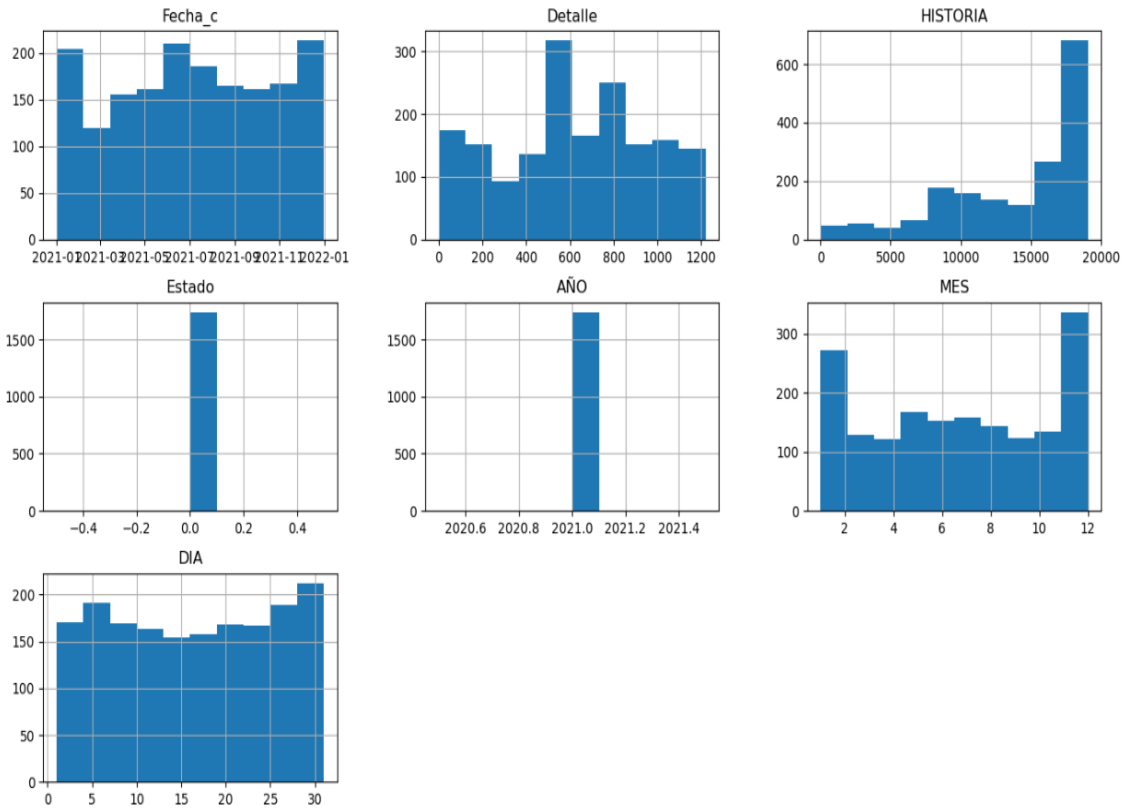
Figura 54. **Muestra el valor final únicos de tipo de cita médica.**

```
Tipo_Cita
['TRAUMATOLOGIA' 'MEDICINA INTERNA' 'ECOGRAFIA' 'DERMATOLOGIA'
 'CARDIOLOGIA' 'RADIOS X' 'CIRUJANO VASCULAR' 'MEDICINA GENERAL'
 'CIRUGÍA DENTAL' 'PEDIATRIA' 'GINECO-OBTETRICIA' 'ODONTOLOGIA'
 'OTORRINOLARINGOLOGIA']
13

Estado
['Atendido - Cancelado' 'Pendiente - cancelado' 'Pendiente - saldo'
 'PENDIENTE' 'Atendido' 'ATENDIDO']
6
```

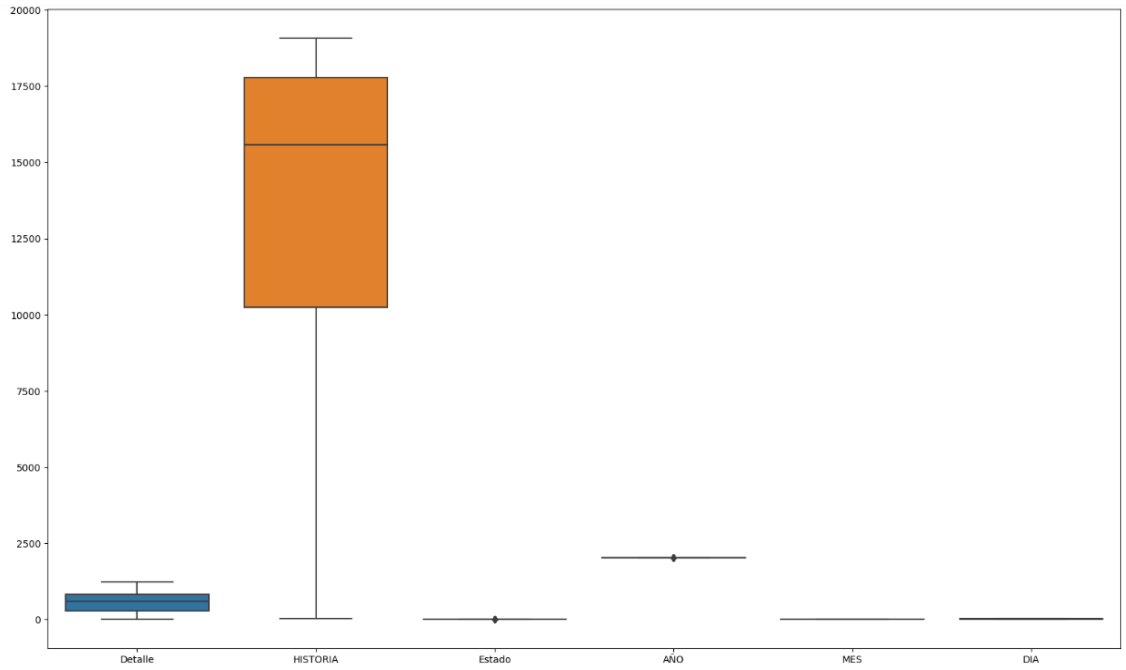
Fuente: Elaboración propia.

Figura 55. **Muestra los mínimos y máximos de las variables en citas médicas.**



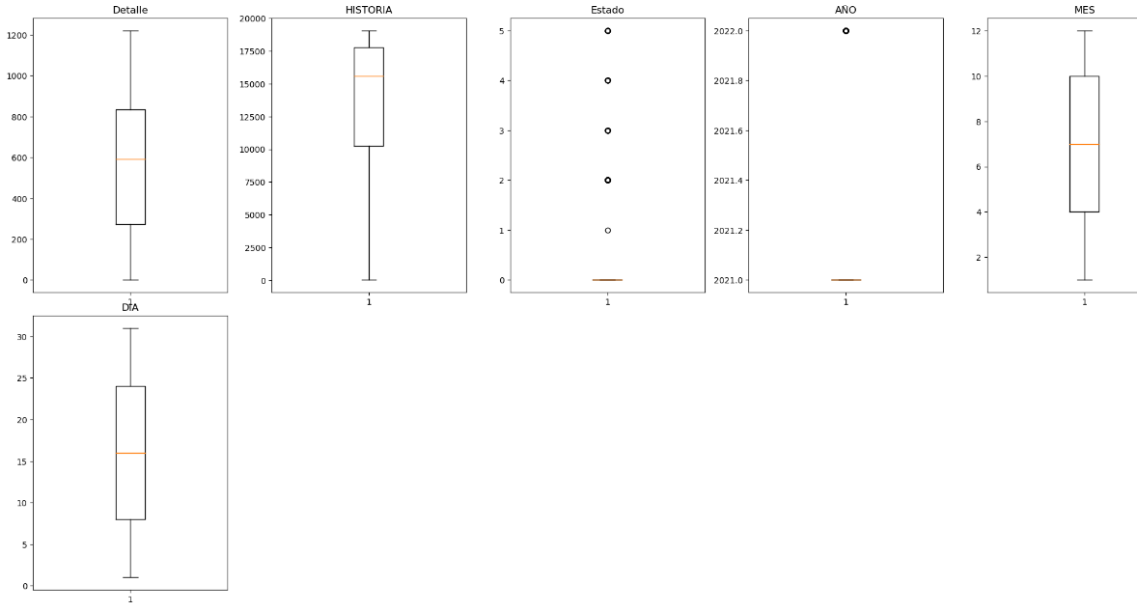
Fuente: Elaboración propia.

Figura 56. **Muestra diagrama de bigotes.**



Fuente: Elaboración propia.

Figura 57. **Presenta caja de bigotes o los valores atípicos (outliers).**



Fuente: Elaboración propia.

Anexo de tratamiento de fisioterapia.

Figura 58. **Muestra la descomposición de año mes y día de la fecha de tratamiento de fisioterapia.**

```
#df['años'] = (df['FECHA2']).year
df['AÑO'] = pd.DatetimeIndex(df['Fecha_tera']).year
df['MES'] = pd.DatetimeIndex(df['Fecha_tera']).month
df['DIA'] = pd.DatetimeIndex(df['Fecha_tera']).day
```

Fuente: Elaboración propia.

Figura 59. **Muestra el rango de fechas tratamiento de fisioterapia.**

```
df=df[(df['AÑO']>=2021)& (df['AÑO'] <=2022)]
```

Fuente: Elaboración propia.

Figura 60. **Muestra la suma de valores nulos en tratamiento de fisioterapia.**

```
Id_exa              0
HISTORIA            0
Id_es               0
Id_grado            0
CI                  0
Nom_doc           140
Diag_doc          139
Diag_tera         140
Exa_comple        139
Deporte           140
Mati_cons         140
Exa_fisico        139
Num_seciones        0
Fecha_tera          0
Hora_tera           0
Trata_fisio1      145
Trata_fisio 2     155
AÑO                 0
MES                 0
DIA                 0
dtype: int64
```

Fuente: Elaboración propia.

Figura 61. **Muestra la eliminación de los valores nulos en tratamiento de fisioterapia.**

```
Id_exa             0
HISTORIA           0
Id_es              0
Id_grado           0
CI                 0
Nom_doc            0
Diag_doc           0
Diag_tera          0
Exa_comple         0
Deporte            0
Mati_cons          0
Exa_fisico         0
Num_seciones       0
Fecha_tera         0
Hora_tera          0
Trata_fisio1       0
Trata_fisio 2      0
AÑO                0
MES                0
DIA                0
dtype: int64
```

Fuente: Elaboración propia.

Figura 62. **Muestra el tipo de datos que tiene el data set de tratamiento de fisioterapia.**

```
Id_grado           int64
Diag_doc           int32
Diag_tera          int32
Num_seciones       int64
Hora_tera         object
Trata_fisio1      object
Trata_fisio 2     object
AÑO                int64
MES                int64
DIA                int64
dtype: object
```

Fuente: Elaboración propia.

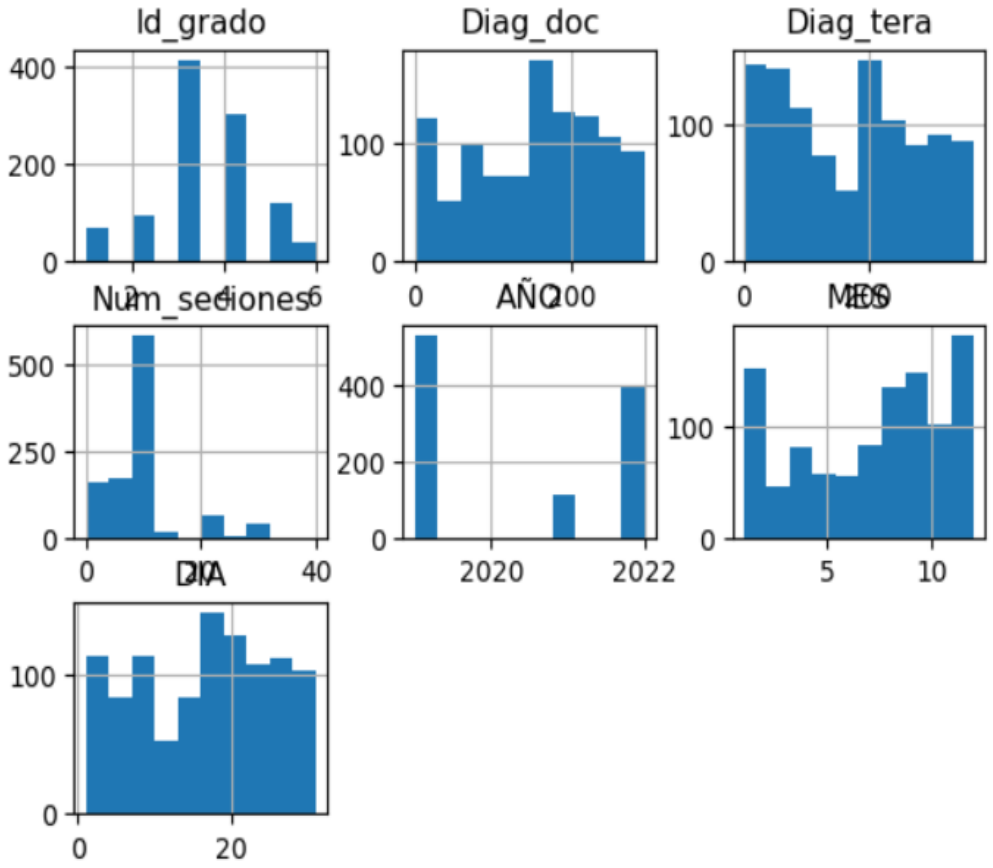
Figura 63. **Muestra la transformación de los valores object a valores de tipo numérico en tratamiento de fisioterapia.**

```
#tranformarmos de tipo object a numerico las variables cuantitativas
from sklearn.preprocessing import LabelEncoder
l=LabelEncoder()
for col in df.Diag_doc:
    if df['Diag_doc'].dtype=='object':
        df['Diag_doc']=l.fit_transform(df['Diag_doc'])
```

```
 from sklearn.preprocessing import LabelEncoder
l=LabelEncoder()
for col in df.Diag_tera:
    if df['Diag_tera'].dtype=='object':
        df['Diag_tera']=l.fit_transform(df['Diag_tera'])
```

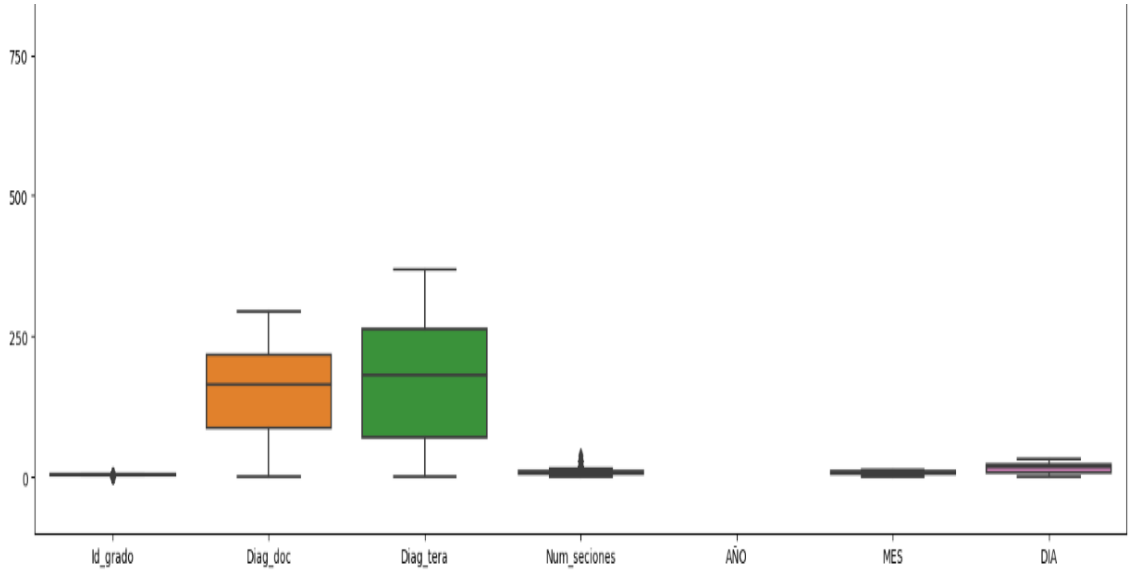
Fuente: Elaboración propia.

Figura 64. **Muestra los mínimos y máximo de las variables de tratamiento de fisioterapia.**



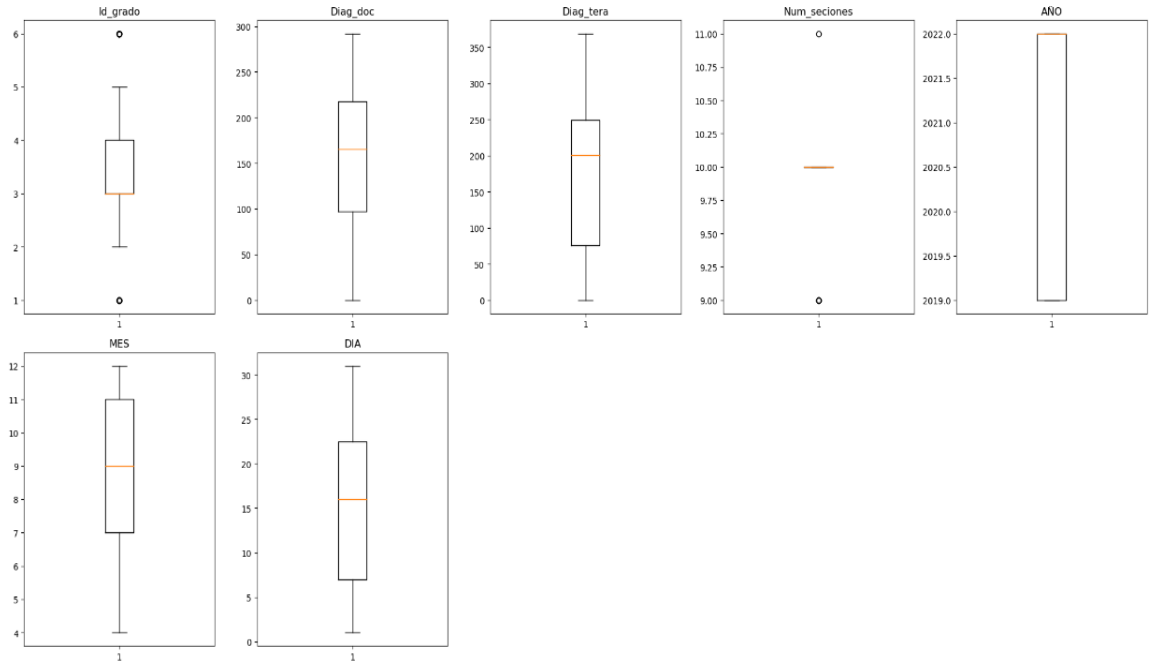
Fuente: Elaboración propia.

Figura 65. **Muestra los valores atípicos (outliers) de tratamiento de fisioterapia.**



Fuente: Elaboración propia.

Figura 66. **Presenta caja de bigotes a los valores atípicos (outliers) de tratamiento de fisioterapia.**



Fuente: Elaboración propia.

Figura 67. **Muestra el cambio de 0 y 1 en tratamiento de fisioterapia.**

```
df=df.replace('COMPRESA QUIMICA CALIENTE C.Q.C','1')
df=df.replace('COMPRESA QUÌMICA FRIA','1')
df=df.replace('TÈCNICAS ESPECIALES','1')
df=df.replace('MASAJE MECÀNICO','1')
df=df.replace('MASAJE MANUAL','1')
df=df.replace('EJERCICIOS EN COLCHONETA','1')
df=df.replace('EJERCICIOS DE MEJORAMIENTO DE LA FUERZA MUSCULAR','1')
df=df.replace('EJERCICIOS ESPECIALES','1')
df=df.replace('EJERCICIOS DE REEDUCACIÒN DE LA MARCHA','1')
df=df.replace('EJERCICIOS CON RESISTENCIAS','1')
df=df.replace('MASAJE DRENAJE LINFATICO','1')
df=df.replace('MOVILIDAD ASISTIDA','1')
df=df.replace('FORTALECIMIENTO','1')
df=df.replace('MASAJE DESCONTRACTURANTE','1')
df=df.replace('ESTIRAMIENTOS','1')
df=df.replace('VENDAJE FUNCIONAL','1')
df=df.replace('PILOMETRIA','1')
df=df.replace('ULTRASONIDO','0')
df=df.replace('MAGNETOTERAPIA','0')
df=df.replace('LASER','0')
df=df.replace('CRIOTERAPIA','0')
df=df.replace('ELECTROESTIMULACIÒN','0')
df=df.replace('KINECTIXER','0')
```

Fuente: Elaboración propia.

Anexo de signos vitales.

Figura 68. **Muestra la descomposición de año mes y día de la fecha de signos vitales.**

```
#df['años'] = (df['FECHA2']).year
df['AÑO'] = pd.DatetimeIndex(df['FECHA2']).year
df['MES'] = pd.DatetimeIndex(df['FECHA2']).month
df['DIA'] = pd.DatetimeIndex(df['FECHA2']).day
```

Fuente: Elaboración propia.

Figura 69. **Muestra el rango de fechas en signos vitales.**

```
df=df[(df['AÑO']>=2021)& (df['AÑO'] <=2022)]
```

Fuente: Elaboración propia.

Figura 70. **Muestra el tipo de variables que tiene el data set de signos vitales.**

```
df.info()
```

```
 0   ID_APAR2           4364 non-null   int64
 1   HISTORIA           4364 non-null   int64
 2   CI                 4364 non-null   int64
 3   TEMP               4364 non-null   float64
 4   TEM                4364 non-null   float64
 5   PESO               4364 non-null   float64
 6   FRECU_CARDIA       4364 non-null   int64
 7   TENSION_ARTERIAL   4364 non-null   object
 8   PERIME_FALICO      4364 non-null   float64
 9   FREC_RESP          4364 non-null   int64
 10  TALLA              4364 non-null   float64
 11  FECHA2             4364 non-null   object
 12  HORA2              4364 non-null   object
 13  SATURA             4364 non-null   float64
 14  AÑO                4364 non-null   int64
 15  MES                4364 non-null   int64
 16  DIA                4364 non-null   int64
dtypes: float64(6), int64(8), object(3)
memory usage: 613.7+ KB
```

Fuente: Elaboración propia.

Figura 71. **Muestra el remplazo de valores de la media en los valores de nan en signos vitales.**

```
avg_Toma = df["SATURA"].astype("float").mean(axis=0)
df['SATURA'].replace(np.nan,avg_Toma, inplace=True)
```

Fuente: Elaboración propia.

Figura 72. **Muestra si existe valores nulos en signos vitales.**

```
ID_APAR2           0
HISTORIA           0
CI                 0
TEMP               0
TEM                0
PESO               0
FRECU_CARDIA       0
TENSION_ARTERIAL   0
PERIME_FALICO      0
FREC_RESP          0
TALLA              0
FECHA2             0
HORA2              0
SATURA             0
AÑO                0
MES                0
DIA                0
dtype: int64
```

Fuente: Elaboración propia.

Figura 73. **Muestra valores duplicados en signos vitales.**

```
df.duplicated().sum()
```

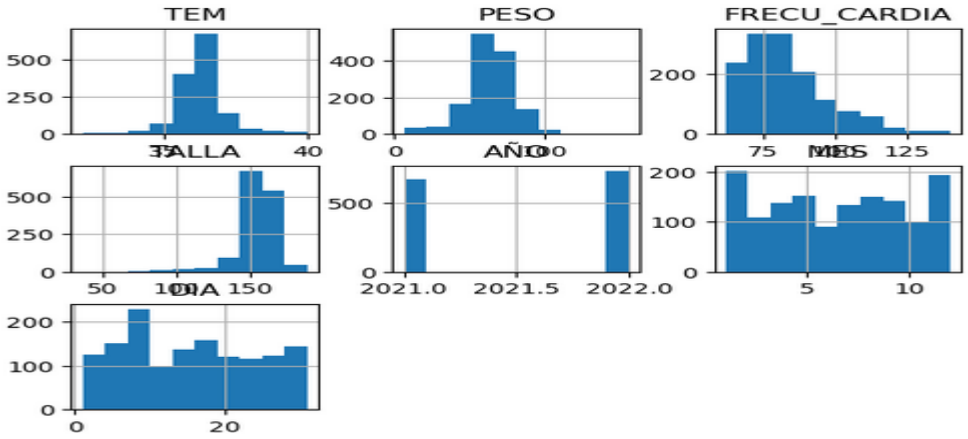
```
10
```

```
#eliminamos los duplicados
df.drop_duplicates(keep=False,inplace=True)
```

```
df.describe()
```

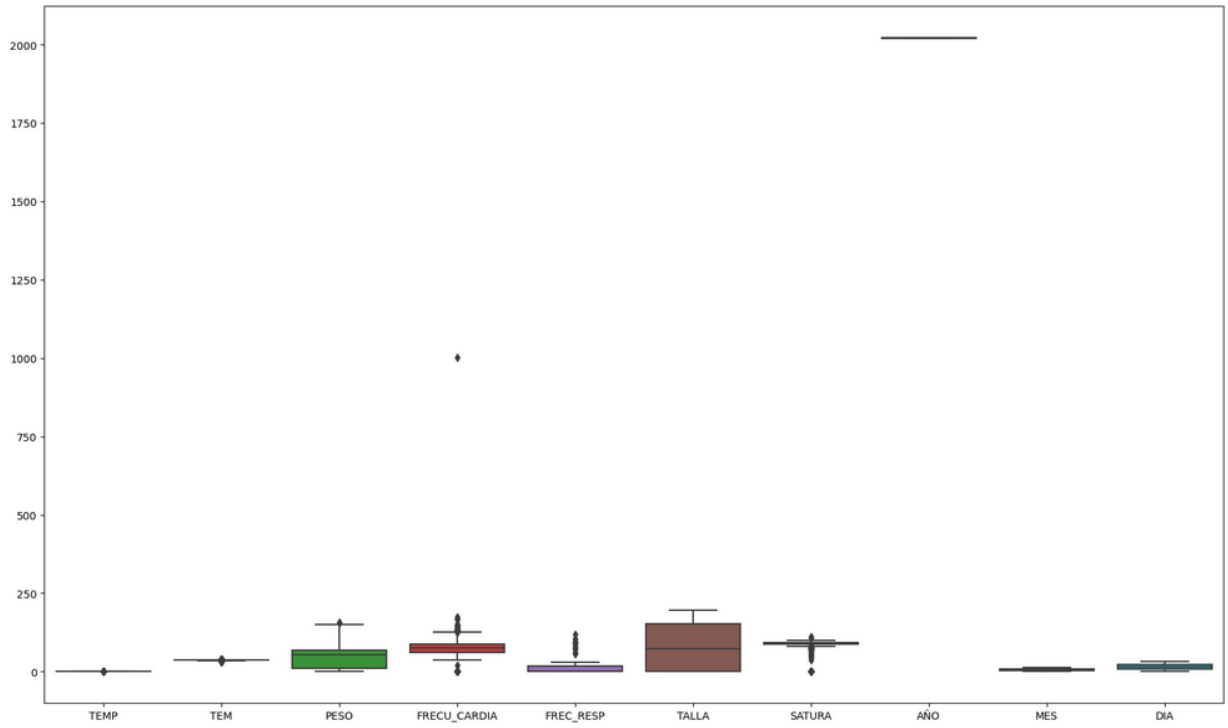
Fuente: Elaboración propia.

Figura 74. **Muestra los mínimos y máximo de las variables de signos vitales.**



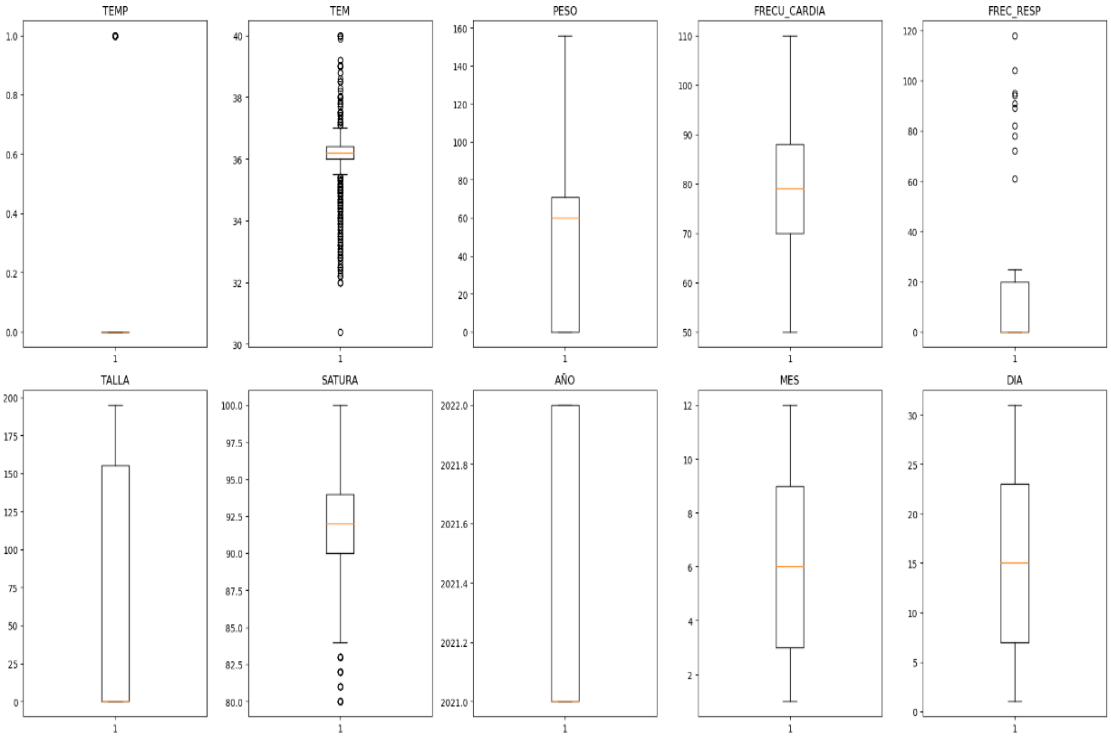
Fuente: Elaboración propia.

Figura 75. **Muestra los valores atípicos (outliers) en signos vitales.**



Fuente: Elaboración propia.

Figura 76. **Presenta caja de bigotes a los valores atípicos (outliers) en signos vitales.**



Fuente: Elaboración propia.

Figura 77. **Muestra los valores únicos en los rangos estableciods en la variable TEM y el resultado de la variable TEMP.**

```
df1=df[(df['TEM']>=17)& (df['TEM'] <=35.5)]
#queremos saber cuáles son los registros de las variables cualitativas
col=['TEM']
for i in col:
    print(i)
    print(df[i].unique())
    print(df[i].nunique())
    print()
```

```
TEM
[37.3  35.4  36.3  35.   35.2  36.2  36.5  36.1  35.7  36.   37.   37.5
 32.2  36.6  39.2  35.6  38.5  36.4  35.3  35.8  36.8  35.5  40.   35.9
 34.6  38.   36.32 36.7  38.8  39.   36.9  34.4  33.6  33.9  32.5  37.4
 32.9  34.1  33.   33.1  34.7  33.8  34.8  34.5  32.8  38.3  37.7  33.7
 34.9  33.2  37.8  34.   34.3  34.2  33.5  34.06 35.98 33.3  35.1  37.1
 30.4  37.2  38.6  32.4  39.9  36.21 32.   33.4  37.9  38.2  32.7  32.3
 32.6  36.28 36.29]
75
```

```
df1[['TEM','TEMP']].head()
```

	TEM	TEMP
25	35.4	1.0
36	35.0	1.0
43	35.2	1.0
100	32.2	1.0
175	35.3	1.0

Fuente: Elaboración propia.

Figura 78. **Muestra los valores únicos en los rangos estableciods en la variable TEM y el resultado de la variable TEMP.**

```
df2=df[(df['TEM']>=35.6)& (df['TEM'] <=37.8)]
#queremos saber cuáles son los registros de las variables cualitativas
col=['TEM']
for i in col:
    print(i)
    print(df[i].unique())
    print(df[i].nunique())
    print()
```

```
TEM
[37.3  35.4  36.3  35.   35.2  36.2  36.5  36.1  35.7  36.   37.   37.5
 32.2  36.6  39.2  35.6  38.5  36.4  35.3  35.8  36.8  35.5  40.   35.9
 34.6  38.   36.32 36.7  38.8  39.   36.9  34.4  33.6  33.9  32.5  37.4
 32.9  34.1  33.   33.1  34.7  33.8  34.8  34.5  32.8  38.3  37.7  33.7
 34.9  33.2  37.8  34.   34.3  34.2  33.5  34.06 35.98 33.3  35.1  37.1
 30.4  37.2  38.6  32.4  39.9  36.21 32.   33.4  37.9  38.2  32.7  32.3
 32.6  36.28 36.29]
75
```

```
df2[['TEM','TEMP']].head()
```

	TEM	TEMP
23	37.3	0.0
27	36.3	0.0
44	36.2	0.0
48	36.5	0.0
54	36.1	0.0

Fuente: Elaboración propia.

Figura 79. **Muestra los valores únicos en los rangos establecidos en la variable TEM y el resultado de la variable TEMP.**

```
df3=df[(df['TEM']>=37.9)& (df['TEM'] <=42)]
#queremos saber cuáles son los registros de las variables cualitativas
col=['TEM']
for i in col:
    print(i)
    print(df[i].unique())
    print(df[i].nunique())
    print()
```

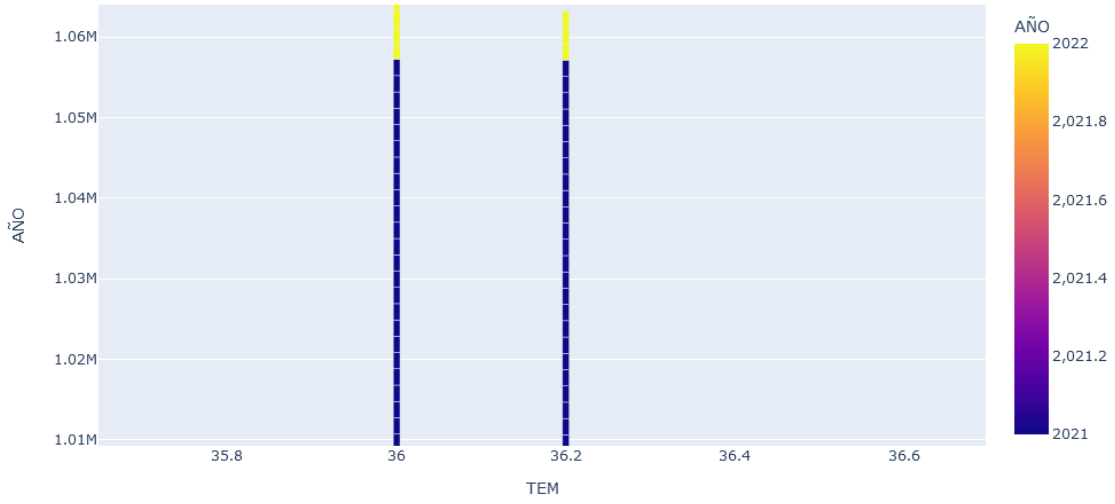
```
TEM
[37.3  35.4  36.3  35.   35.2  36.2  36.5  36.1  35.7  36.   37.   37.5
 32.2  36.6  39.2  35.6  38.5  36.4  35.3  35.8  36.8  35.5  40.   35.9
 34.6  38.   36.32 36.7  38.8  39.   36.9  34.4  33.6  33.9  32.5  37.4
 32.9  34.1  33.   33.1  34.7  33.8  34.8  34.5  32.8  38.3  37.7  33.7
 34.9  33.2  37.8  34.   34.3  34.2  33.5  34.06 35.98 33.3  35.1  37.1
 30.4  37.2  38.6  32.4  39.9  36.21 32.   33.4  37.9  38.2  32.7  32.3
 32.6  36.28 36.29]
75
```

```
df3[['TEM','TEMP']].head()
```

	TEM	TEMP
117	39.2	1.0
134	38.5	1.0
225	40.0	1.0
265	38.0	1.0
312	38.5	1.0

Fuente: Elaboración propia.

Figura 80. **Muestra el comportamiento de signos vitales durante los años.**



Fuente: Elaboración propia.