Pontificia Universidad Católica del Ecuador

Facultad De Ingeniería

Escuela de Sistemas

TEMA:

ANÁLISIS DE DATOS UTILIZANDO EL MODELO CRISP-DM, CASO DE ESTUDIO FACTORES QUE AFECTAN LA SALUD PEDIÁTRICA

AUTOR:

LEONARDO DAVID CHAMORRO VILLOTA

QUITO DM, JUNIO DE 2023

DEDICATORIA

Dedico este trabajo, producto de mi esfuerzo y dedicación, a mi amada familia y a mi querida novia. Su constante amor y apoyo incondicional han sido el faro que me ha guiado a través de cada desafío y obstáculo. Cada logro que obtengo es también un tributo a su fe inquebrantable en mí.

También, dedico este trabajo a mis mentores, quienes, con su sabiduría, me han enseñado más de lo que las palabras pueden expresar. Han moldeado mi camino y han inculcado en mí una profunda pasión por la ciencia. Su influencia ha sido determinante en cada paso que he dado.

Este trabajo es una manifestación de su amor, apoyo y guía.

A todos ustedes, les dedico este logro.

AGRADECIMIENTO

Mi más sincero reconocimiento y gratitud a todos aquellos que han estado conmigo, caminando a mi lado en este viaje, y que me han brindado su apoyo incondicional para alcanzar mis sueños. Cada paso que he dado ha sido posible gracias a ustedes.

Quiero expresar un profundo agradecimiento a todos los amigos, con quienes he compartido este emocionante viaje de descubrimiento y aprendizaje. Sus contribuciones, camaradería y apoyo han sido invaluables.

Asimismo, no puedo dejar de reconocer a mis maestros, cuya paciencia y orientación me han guiado en cada etapa de este proceso. Su dedicación y compromiso han enriquecido mi experiencia y han jugado un papel crucial en mi formación.

Gracias por ser parte de mi viaje y contribuir a la formación de la persona que soy hoy.

ABSTRACT

Español

Este estudio se centra en la aplicación de técnicas de aprendizaje automático para el diagnóstico de enfermedades neuropsicológicas en pacientes pediátricos en Ecuador. Se adoptó el modelo de Minería de Datos CRISP-DM para guiar el análisis. Se utilizó un conjunto de datos recopilados de diversas clínicas de Ecuador que abarcaba una amplia gama de características, como la edad, el sexo, la nacionalidad, el lugar de nacimiento, el historial médico, entre otros, junto con la inclusión de nuevos parámetros relevantes al estudio.

Se exploraron diversos modelos de aprendizaje automático, incluyendo la Regresión con Vectores de Soporte (SVR), los Árboles de Decisión, los Bosques Aleatorios y los k-Vecinos más Cercanos (k-NN), cada uno de los cuales se configuró y optimizó con hiperparámetros para mejorar su rendimiento.

El desempeño de cada modelo se evaluó y comparó en función de la precisión y el ajuste a los objetivos del negocio. Se implementó un tablero de mando interactivo utilizando el lenguaje de programación R para visualizar y comunicar los resultados del modelo, proporcionando una herramienta útil para los interesados.

Este estudio concluye que el aprendizaje automático puede desempeñar un papel significativo en la mejora del diagnóstico y la comprensión de las enfermedades neuropsicológicas pediátricas.

English

This study centers on the application of machine learning techniques for the diagnosis of neuropsychological diseases in pediatric patients in Ecuador. The CRISP-DM Data Mining model was adopted to guide the analysis. A dataset collected from various clinics in Ecuador was utilized, encompassing a broad range of features such as age, sex, nationality, birthplace, medical history, among others, alongside the inclusion of new parameters pertinent to the study.

Various machine learning models were explored, including Support Vector Regression (SVR), Decision Trees, Random Forests, and k-Nearest Neighbors (k-NN), each of which was configured and optimized with hyperparameters to enhance their performance.

The performance of each model was evaluated and compared based on accuracy and fit to business objectives. An interactive dashboard was developed using the R programming language to visualize and communicate the model's results, providing a useful tool for stakeholders.

This study concludes that machine learning can play a significant role in improving the diagnosis and understanding of pediatric neuropsychological diseases.

ÍNDICE

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS, GRÁFICOS Y TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO I – INTRODUCCIÓN

1. Introducción

1.1. Justificación

Un estudio de análisis de datos en el ámbito médico pediátrico permitirá identificar patrones y tendencias en los datos recopilados. Esto puede ser especialmente valioso para las empresas del sector de la salud, ya que les permitiría tomar decisiones más informadas y mejorar la calidad de sus servicios.

Además, en este caso particular, el proyecto se enfoca en datos pediátricos que no se limitan al Distrito Metropolitano de Quito. Esto permitirá una visión más amplia y representativa de la situación en todo el país, lo que a su vez permitirá una toma de decisiones más informada y estratégica por parte de las clínicas.

1.2. Planteamiento del Problema

El problema se enfoca en la extracción de conocimiento a partir de un gran volumen de datos privados que provienen del diagnóstico de varios médicos pediatras. La cantidad de datos recolectados exige el uso de herramientas informáticas para su almacenamiento y gestión, y la implementación de técnicas de análisis de datos que permitan aprovechar la información disponible para apoyar la toma de decisiones y la generación de alertas tempranas.

Por otro lado, el proceso de limpieza, filtrado y transformación de los datos registrados de estaciones de monitoreo implica problemas ampliamente conocidos en la comunidad, lo que afecta la eficiencia de los modelos y la disponibilidad para la aplicación de técnicas de análisis de datos (Bustamante Martínez, Galvis Lista, & Gómez Flórez, 2013). Además, se debe considerar la sensibilidad y privacidad de los datos médicos, lo que requiere medidas de seguridad y privacidad adecuadas. Por lo tanto, el objetivo de este trabajo es proponer y desarrollar técnicas efectivas y seguras para el análisis de datos médicos pediátricos, con el fin de extraer conocimiento útil para apoyar la toma de decisiones en la atención médica.

1.3. Objetivos

1.3.1. General

Análisis de datos utilizando el modelo CRISP-DM, caso de estudio Factores que Afectan la Salud Pediátrica.

1.3.2. Específicos

- Identificar variables relevantes para el análisis, la comprensión de la distribución de los datos y la identificación de valores faltantes.
- Limpieza de los datos, eliminación de variables irrelevantes o redundantes y la transformación de variables para que sean compatibles con los modelos de análisis.
- Construir modelos de regresión para predecir la probabilidad de una enfermedad en base a los datos de salud del paciente y la construcción de modelos de clasificación para identificar patrones en los datos.
- Evaluar la precisión y confiabilidad de los modelos, y realizar la validación de los modelos utilizando datos que no se utilizaron en la construcción de este.
- Implementar sistemas de alerta temprana para prevenir enfermedades pediátricas, mediante la identificación de factores de riesgo que se deben monitorear de manera regular.

1.4. Alcance

El objetivo principal de este estudio es analizar la información proporcionada por los registros médicos, que se encuentran en un repositorio de datos, para identificar patrones y tendencias en las enfermedades y afecciones pediátricas. Este análisis tiene como objetivo identificar los factores de riesgo que pueden afectar la salud de pacientes pediátricos. Se utilizarán técnicas de minería de datos, como reglas de asociación, para identificar patrones en los datos que puedan revelar la aparición de escenarios específicos y proporcionar información sobre métricas y mejoras relevantes en el ámbito de la salud pediátrica.

Para el desarrollo se abarcará estudios descriptivos que respalden la investigación, considerando los conceptos fundamentales de la salud pediátrica, los determinantes sociales de la salud y los enfoques de prevención y promoción de la salud. De esta forma, se podrá contextualizar el análisis de los datos médicos pediátricos y su relación con la salud infantil.

Además, se describirá en detalle el modelo CRISP-DM y su aplicación en el análisis de datos médicos pediátricos. Se deben considerar las etapas del modelo, que incluyen la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y la implementación. Cada una de estas etapas debe explicarse y justificarse en relación con el análisis de datos médicos pediátricos y su utilidad para la identificación de factores que afectan la salud pediátrica.

CAPÍTULO II – MARCO TEÓRICO

2. MARCO TEÓRICO

El modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) es un marco de trabajo que ofrece una estructura metodológica para llevar a cabo proyectos de minería de datos y análisis de información (Chapman et al., 2000). Este modelo ha sido ampliamente utilizado en diversas aplicaciones, incluyendo el análisis de datos médicos y de salud (Cios & Moore, 2002). En el contexto de la salud pediátrica, investigaciones previas han aplicado el CRISP-DM para identificar factores de riesgo y predecir el desarrollo de enfermedades en niños (Smith et al., 2018).

Uno de los primeros pasos en el análisis de datos médicos pediátricos es la identificación de variables relevantes para el análisis. La comprensión de la distribución de los datos y la identificación de valores faltantes son cruciales para garantizar la calidad de los datos en la etapa de preparación (Witten et al., 2011). La selección de variables se basa en criterios como la relevancia clínica, la prevalencia en la población y la asociación con los resultados de interés (Steyerberg, 2009).

La limpieza de los datos implica la eliminación de variables irrelevantes o redundantes y la transformación de variables para que sean compatibles con los modelos de análisis. Diversas técnicas, como la imputación de valores faltantes y la normalización de variables, pueden aplicarse en esta etapa (García et al., 2016).

La construcción de modelos de regresión y de clasificación es esencial para predecir la probabilidad de una enfermedad en función de los datos de salud del paciente. Modelos como la regresión logística y los árboles de decisión son comunes en el análisis de datos médicos (Kuhn & Johnson, 2013). Estos modelos permiten identificar patrones y relaciones en los datos, lo que facilita la identificación de factores de riesgo y la implementación de intervenciones preventivas (Hastie et al., 2009).

La evaluación de la precisión y confiabilidad de los modelos es crucial para garantizar la calidad de las predicciones y las decisiones basadas en los resultados del análisis. Métricas como la sensibilidad, la especificidad y el área bajo la curva ROC pueden

utilizarse para evaluar el rendimiento de los modelos (Steyerberg, 2009). Además, la validación de los modelos mediante datos que no se utilizaron en la construcción de este ayuda a garantizar su generalización y aplicabilidad en contextos clínicos reales (Kuhn & Johnson, 2013).

La disertación se centra en la aplicación del modelo CRISP-DM al análisis de datos médicos pediátricos, con el objetivo de identificar factores que afectan la salud pediátrica y prevenir enfermedades en niños. La investigación aborda aspectos críticos como la selección y limpieza de variables, la construcción y evaluación de modelos predictivos y de clasificación, y la implementación de sistemas de alerta temprana. El marco teórico se basa en la literatura existente sobre el modelo CRISP-DM y su aplicación en el ámbito de la salud pediátrica, proporcionando una base sólida para el desarrollo y evaluación de estrategias de prevención y atención médica en el contexto de la salud infantil.

2.1. Pruebas de rendimiento

Las pruebas de rendimiento son una componente esencial en el campo de la ciencia de datos y el aprendizaje automático. Según Provost y Fawcett (2013), estas pruebas se utilizan para "evaluar la efectividad de los modelos predictivos y para proporcionar a los interesados información valiosa sobre la precisión y utilidad de estos modelos".

2.1.1 Aspectos

- Función: Se refiere a la capacidad del modelo para predecir correctamente la variable de salida basándose en las variables de entrada. Una función describe uno o más comportamientos o interacciones entre los datos y el modelo.
- Medida: La medida de la prueba puede ser cuantitativa, como la precisión, la sensibilidad o la especificidad del modelo, o cualitativa, como la interpretabilidad y utilidad de los resultados del modelo.

2.1.2 Objetivos

- Comparar diferentes modelos para determinar cuál ofrece un mejor rendimiento en función de las medidas de rendimiento seleccionadas.

- Identificar las áreas del modelo que necesitan mejoras y proporcionar recomendaciones basadas en los resultados de las pruebas de rendimiento.

2.2. Herramientas para el procesamiento de datos

R Studio es un framework ampliamente utilizado para el análisis de datos, y en el caso de este proyecto, puede facilitar el procesamiento y análisis de datos médicos pediátricos y factores de riesgo asociados (RStudio Team, 2020). Utilizando R Studio para el análisis de datos médicos pediátricos, es posible diseñar e implementar un layout adecuado para visualizar la información y abordar la problemática del manejo adecuado de la información de la base de datos para su procesamiento.

De acuerdo con Kaupp (2016), el diseño de un sistema centrado en el usuario requiere que su interfaz aporte valor y optimice las actividades diarias de los usuarios. Es fundamental establecer un flujo de trabajo claro que ayude a eliminar procesos innecesarios y transmitir las ideas de manera clara y sencilla, sin sobrecargar la interfaz ni causar confusión.

En el marco de nuestro estudio, es imprescindible recopilar datos médicos pediátricos y factores de riesgo de forma estructurada o no estructurada en una base de datos. Las bases de datos de código abierto más utilizadas y escalables incluyen MySQL, MSSQL y MongoDB. Gracias a R Studio, es posible establecer una conexión entre la base de datos y diversos paquetes para la manipulación y visualización de datos.

Uno de los mayores desafíos en el análisis de grandes volúmenes de datos es la visualización. Como señalan Mohd Ali et al. (2016), una herramienta de visualización efectiva debe permitir una interacción fluida con los datos, incluyendo la capacidad de manejar datos estructurados y no estructurados. Asimismo, debe ser capaz de identificar patrones y correlaciones, y es crucial seleccionar cuidadosamente la cantidad de datos a visualizar para evitar la pérdida de información relevante o la generación de visualizaciones densas y complicadas de interpretar.

La inclusión de todos los puntos de datos en la interfaz puede resultar en superposición de información y sobrecarga cognitiva para el usuario. Estos problemas se añaden a los desafíos habituales en la visualización de datos, como el ruido visual, la pérdida de información, la percepción excesiva de la imagen, el cambio rápido de imágenes y los altos requerimientos de rendimiento.

Existen diversas herramientas de visualización interactiva que facilitan la presentación efectiva de la información relevante. Entre las más populares se encuentran Tableau, Microsoft Power BI, Plotly y Gephi. Para nuestro proyecto, Plotly es una opción adecuada dado que es un software de código abierto que admite lenguajes de programación como Python o R, y facilita la generación de interfaces interactivas vinculadas a datos dinámicos de la base de datos.

Plotly es una plataforma basada en la web cuyas visualizaciones pueden ser compartidas y adaptadas a la perspectiva del usuario, proporcionando retroalimentación adecuada (Finch & Flenner, 2016). Sus principios subrayan la importancia de mantener la presentación simple pero relevante, utilizando diversas herramientas para crear una interfaz interactiva y útil.

Para abordar el análisis de datos médicos pediátricos y los factores de riesgo asociados en este proyecto, se propone utilizar métricas relevantes para la salud pediátrica y herramientas adecuadas para el procesamiento, análisis y visualización de datos, como R Studio y Plotly. Estas herramientas permitirán identificar patrones y correlaciones en los datos, lo que contribuirá a cumplir los objetivos específicos de la disertación, como la construcción de modelos de regresión y clasificación y la implementación de sistemas de alerta temprana para prevenir enfermedades pediátricas.

2.3. Reglas de asociación

Las reglas de asociación son una técnica de minería de datos que se utiliza para descubrir relaciones comunes dentro de un conjunto de datos. En el contexto de la salud pediátrica y el modelo CRISP-DM, las reglas de asociación pueden utilizarse para identificar factores que afectan la salud de los niños y posibles patrones en la incidencia de enfermedades.

Estas reglas brindan una perspectiva sobre la naturaleza de las asociaciones entre diversos aspectos relacionados con la salud pediátrica. Los hallazgos derivados de estas asociaciones pueden servir para filtrar información, analizarla y potencialmente construir un modelo predictivo basado en la observación de patrones (Giraldo, 2012).

2.4. Modelo CRISP-DM

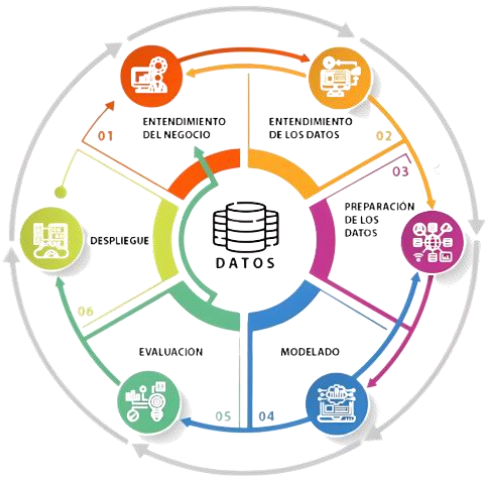
En base al objetivo de este proyecto, se realizará una búsqueda y análisis del estado del arte en salud pediátrica y minería de datos, así como la selección de herramientas y técnicas apropiadas para llevar a cabo el análisis de manera efectiva. Además, en este capítulo se detallará cada una de las fases del modelo CRISP-DM y cómo se aplicarán en el contexto del análisis de datos en salud pediátrica.

El modelo seleccionado para este proyecto es CRISP-DM que es un proceso estándar y largamente utilizado en la industria para la minería de datos (Chapman et al., 2000). Este modelo fue desarrollado por un consorcio de empresas y organizaciones, incluyendo NCR Corporation, DaimlerChrysler, SPSS Inc., Teradata y OHRA Verzekeringen en 1996.

La metodología CRISP-DM es un proceso estructurado y bien definido para la minería de datos que ha sido utilizado en diversas industrias para la resolución de problemas a través del análisis de datos. Este proceso consta de seis fases que ayudan a guiar el análisis de datos desde la definición del problema hasta la implementación de soluciones basadas en los resultados obtenidos.

A lo largo del proceso, se recomienda seguir un enfoque iterativo y flexible. Esto significa que los analistas de datos pueden volver a fases anteriores si es necesario, en función de los resultados obtenidos en cada etapa del proceso.

Figura 1. Fases del proceso de minería de datos con la metodología CRISP-DM



Fuente: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

Descripción de los pasos del modelo propuesto

2.4.1 Comprensión del negocio:

En esta etapa, se identifican los objetivos y requerimientos del proyecto desde una perspectiva pediátrica. Se deben entender los problemas de salud que afectan a los niños y cómo el análisis de datos puede contribuir a la prevención y tratamiento de enfermedades. Además, se debe establecer la comunicación con expertos en salud pediátrica para obtener una comprensión más profunda del dominio y sus desafíos.

2.4.2 Comprensión de los datos:

En esta fase, se recopilan, describen y exploran los datos disponibles. Para el caso de estudio en salud pediátrica, se deben identificar las variables relevantes para el análisis, entender la distribución de los datos y detectar valores faltantes. Además, se debe evaluar la calidad de los datos, considerando factores como la precisión, la integridad y la actualidad de la información (Wirth & Hipp, 2000).

2.4.3 Preparación de los datos:

Esta etapa involucra la limpieza de los datos, eliminación de variables irrelevantes o redundantes y la transformación de variables para que sean compatibles con los modelos de análisis. La preparación de los datos es esencial para garantizar la calidad y la confiabilidad de los resultados obtenidos en el análisis (Pyle, 1999).

2.4.4 Modelado:

En esta fase, se construyen modelos de regresión para predecir la probabilidad de una enfermedad en base a los datos de salud del paciente y se construyen modelos de clasificación para identificar patrones en los datos. Estos modelos pueden incluir algoritmos de aprendizaje supervisado y no supervisado, como árboles de decisión, regresión logística, máquinas de vectores de soporte y redes neuronales (Han, Kamber, & Pei, 2011).

2.4.5 Evaluación:

La evaluación de la precisión y confiabilidad de los modelos es fundamental en esta fase. Para ello, se deben aplicar métricas de evaluación, como la precisión,

el área bajo la curva ROC (AUC), la sensibilidad y la especificidad, entre otras. Además, se debe evaluar la relevancia y aplicabilidad de los modelos en el contexto pediátrico y realizar validaciones cruzadas para asegurar la generalización de los resultados (Kelleher, Mac Namee, & D'Arcy, 2015).

2.4.6 Implementación:

Finalmente, en la fase de implementación, los modelos desarrollados y validados se aplican para generar información útil y accionable en el ámbito de la salud pediátrica. Esto puede incluir la creación de sistemas de alerta temprana para identificar niños en riesgo de enfermedades, el desarrollo de estrategias de prevención y la implementación de intervenciones basadas en los resultados del análisis de datos. Es importante que los profesionales de la salud pediátrica colaboren estrechamente con los expertos en análisis de datos para asegurar que los resultados del proyecto sean efectivamente aplicados y se logren los objetivos establecidos en la fase de comprensión del negocio (Provost & Fawcett, 2013).

2.5 Estado del arte

Los diversos campos del conocimiento, como la ciencia, la medicina, la economía y la demografía, generan una gran cantidad de datos disponibles a través de numerosas fuentes. A menudo, estos datos se almacenan en bruto, sin filtrar ni preparar para su posterior utilización. Ante el desafío de entender, analizar y optimizar la información obtenida a partir de estos datos, han surgido herramientas de análisis y visualización de datos innovadoras y eficientes.

La visualización de información se refiere a la representación de datos en forma gráfica, con el objetivo de facilitar su interpretación y comprensión. Existe una amplia gama de gráficos, cuyo uso depende del tipo de contenido que se desea comunicar. Entre los más habituales se encuentran los diagramas corporativos, los histogramas y los gráficos circulares. Estos permiten a los encargados de tomar decisiones identificar relaciones y establecer patrones.

El análisis de información tiene como finalidad renovar la visión sobre las bases de datos, de modo que las interacciones del usuario con los datos puedan convertirse en

información útil para ajustar los procesos analíticos subyacentes. Este enfoque permite utilizar de manera sistemática los elementos o componentes de interés con el fin de optimizar el sistema y atraer la atención del usuario (Kelm, Fekete, Andreinko, & Kohihammer, 2008).

Card, Mackinlay, & Shneiderman (1999) ofrecieron una definición precisa del análisis y la visualización de información, describiéndolo como el uso de soporte informático interactivo y representaciones visuales de datos abstractos con el objetivo de ampliar la cognición.

La gran cantidad de información que se analizará debe ser categorizada según Lizasoain y Joaristi, quienes proponen dos tipos de variables: nominales y ordinales. De esta manera, las variables cualitativas se convierten en cuantitativas. Por otro lado, se mantiene una perspectiva clásica de dependencia, diferenciando las variables en explicativas y dependientes. Ambas perspectivas se definen en los manuales de análisis de datos estadísticos, lo cual es una herramienta valiosa para el desarrollo y la disminución del costo computacional.

En el campo de la salud pediátrica, un ejemplo relevante del análisis de datos en gran volumen es el estudio llevado a cabo por Amin, Prasad y Singh (2016). En este estudio, los investigadores aplicaron técnicas de minería de datos para analizar la prevalencia y los factores de riesgo de la desnutrición infantil en la India.

Los autores utilizaron datos demográficos y de salud de la Encuesta Nacional de Salud Familiar (NFHS) de la India, que incluía información sobre la salud, la nutrición y el estado socioeconómico de los niños menores de cinco años y sus familias. Mediante la utilización de algoritmos como árboles de decisión, regresión logística y redes neuronales, los investigadores identificaron factores de riesgo clave para la desnutrición infantil, como la falta de acceso a servicios de saneamiento, el bajo nivel educativo de la madre y la situación económica de la familia.

Además, el estudio presentó visualizaciones interactivas y dinámicas de los resultados para facilitar su interpretación y discusión. Estas visualizaciones incluían mapas geoespaciales que mostraban la prevalencia de la desnutrición en diferentes regiones de la India y gráficos de burbujas que ilustraban la relación entre la desnutrición y otros factores de riesgo.

Los tableros de control y los visualizadores facilitan el análisis del comportamiento de los datos, lo cual es crucial para la toma de decisiones informadas. Un ejemplo práctico se encuentra en la Carrera de Medicina de la Universidad Central del Ecuador, donde se implementó un tablero de control para generar indicadores de inserción laboral y competencias en los estudiantes graduados. En este proyecto, la información se obtuvo mediante encuestas, demostrando una aplicación efectiva de la inteligencia de negocios para evitar el trabajo manual y reducir el tiempo de recolección de datos (Lema Siguencia, 2016).

Otro caso del uso de visualizadores con el objetivo de generar indicadores se puede observar en la Universidad Politécnica Estatal del Carchi en 2019. Buscando agilizar las operaciones de las diversas áreas administrativas que utiliza la Dirección Académica de la institución, se desarrolló un tablero de control basado en herramientas de código abierto como Pentaho, facilitando la toma de decisiones institucionales (Castro Chauca, 2019).

En el contexto empresarial, Imptek Chova del Ecuador S.A. propuso mejorar la gestión estratégica a través del uso de un tablero de control y el análisis de datos para apoyar la toma de decisiones basada en indicadores de desempeño financieros, garantizando así la integridad de la información (Vélez de la Cruz, 2017).

En una línea similar, Mendoza (2015) sugiere la utilización del análisis de datos y herramientas de inteligencia de negocios para determinar la proyección de proformas presupuestarias, basándose en la metodología Kimball. En este caso, se destacó el uso de una base de datos histórica de la Autoridad Portuaria Puerto Bolívar. El producto final brindaba al usuario reportes, tableros de control y análisis que facilitaban la toma de decisiones empresariales, además de entender las necesidades operativas de cada departamento y gerencia (Mendoza Rodríguez, 2015).

La mayoría de los datos existentes están en constante evolución y contienen una gran cantidad de información que espera ser explotada, analizada y presentada. De este modo, se mejorará la toma de decisiones en cualquier ámbito relevante para el campo de aplicación de estas herramientas.

CAPÍTULO III – ANÁLISIS E INTERPRETACIÓN DE LA DATA

3 Preparación y análisis preliminar de los datos

Siguiendo las primeras etapas del modelo CRISP-DM, el objetivo es establecer un entendimiento sólido del negocio y de los datos disponibles, así como preparar los datos para su posterior análisis y modelado. Esta etapa es fundamental para garantizar la relevancia de los resultados obtenidos en el proceso de minería de datos.

Primero, se identificarán las necesidades del negocio y los usuarios finales, lo que permitirá orientar el análisis de datos y asegurar que los resultados sean útiles y aplicables. A continuación, se examinarán las fuentes de información y se describirán las variables disponibles en el conjunto de datos, proporcionando una visión general de la información con la que se trabajará.

Posteriormente, se llevará a cabo la preparación de los datos, que incluye la consolidación de la información en un repositorio, la limpieza de los datos y la selección de las características más relevantes. Estos pasos son cruciales para mejorar la calidad de los datos y facilitar el análisis posterior.

3.1 Entendimiento del negocio

3.1.1 Identificación de las necesidades del negocio

Es fundamental identificar y comprender las necesidades del negocio en relación con el análisis de datos de pacientes con condiciones neurológicas. Estas necesidades pueden incluir:

1. Comprender la distribución geográfica de las condiciones neurológicas para identificar áreas con mayor concentración de casos y posibles factores de riesgo asociados.
2. Analizar la relación entre las condiciones neurológicas y las características demográficas de los pacientes, como edad, género, grupo étnico y nivel socioeconómico.
3. Estudiar la evolución de las condiciones neurológicas a lo largo del tiempo y detectar tendencias emergentes en la incidencia y prevalencia de estas afecciones.

4. Identificar posibles correlaciones entre diferentes condiciones neurológicas y factores de riesgo asociados.
5. Comparar las características de los pacientes con diferentes condiciones neurológicas para identificar diferencias y similitudes entre estos grupos.

3.1.2 Identificación de los usuarios finales

- **Profesionales de la salud**: Profesionales de la atención médica que utilizan los resultados del análisis para mejorar la atención al paciente y desarrollar planes de tratamiento más efectivos.
- **Investigadores**: científicos y académicos que estudian las condiciones neurológicas y utilizan los resultados del análisis para generar nuevas hipótesis y guiar futuras investigaciones.
- **Pacientes y cuidadores**: personas directamente afectadas por las condiciones neurológicas que utilizan los resultados del análisis para tomar decisiones informadas sobre su atención médica y mejorar su calidad de vida.

3.2 Entendimiento de los datos

3.2.1. Ubicación de fuentes de información

Al inicio del proyecto, los datos fueron proporcionados en varios documentos de Excel por tres importantes clínicas del Distrito Metropolitano de Quito especializadas en atención pediátrica. Para facilitar el análisis y la manipulación, estos archivos fueron exportados a formato CSV y posteriormente combinados en una base de datos central que contiene la información relevante.

La base de datos resultante consta de 38,674 instancias y 147 variables, ocupando un espacio de memoria de 96,311,448 bytes. Se proporcionará información detallada de estas características, agrupándolas en bloques numerados según su relación, con el objetivo de apoyar la toma de decisiones y el preprocesamiento de los datos.

3.2.2. Análisis exploratorio de datos y variables

La descripción de los datos se presenta al lector en bloques donde cada característica posee relación con otra y permite agruparlas.

En el primer bloque de información se tiene una relación directa con la fecha de atención de los pacientes. Este bloque incluye las siguientes variables:

Tabla 1. Variables relacionadas a la fecha de atención de los pacientes

VARIABLE	DESCRIPCIÓN	TIPO DE DATO
ANO_ATENCION	Indica el año en que se realizó la atención, escrito en formato numérico de cuatro dígitos (día-mes-año, DD-MM-YYYY)	FECHA
MES_ATENCION	Indica el mes en que se realizó la atención	CADENA DE TEXTO
DIA_ATENCION	Describe el día en que se realizó la atención y puede ser desde lunes a domingo	CADENA DE TEXTO
HORA_ATENCION	Se presenta en formato de 24 horas (HH:MM)	HORA

Establecer la localización geográfica de los pacientes es un procedimiento esencial. Por lo cual, este factor se contempla como un segundo segmento de datos. Para obtener una visión más completa de la ubicación se utilizan los siguientes datos:

Tabla 2. Variables relacionadas al domicilio de los pacientes

VARIABLE	DESCRIPCIÓN	TIPO DE DATO
PROVINCIA	Indica la provincia donde se encuentra el domicilio del paciente	CADENA DE TEXTO
CANTON	Indica el cantón donde se encuentra el domicilio del paciente	CADENA DE TEXTO
DIRECCION	Describe la dirección del domicilio del paciente	CADENA DE TEXTO
CALLE_PRINCIPAL	Indica la calle principal del domicilio del paciente	CADENA DE TEXTO

CALLE_SECUNDARIA	Indica la calle secundaria del domicilio del paciente, si aplica	CADENA DE TEXTO
PUNTO_REFERENCIA	Descripción de un punto de referencia cercano al domicilio del paciente	CADENA DE TEXTO

Incorporando ahora un tercer bloque de información, el cual está compuesto por datos relacionados con el diagnóstico del paciente. Esto permite determinar su condición de salud y su tratamiento correspondiente. También se tienen en cuenta datos sobre la condición del paciente y el tipo de tratamiento. Dentro de este grupo, se categoriza a los pacientes según el tipo de enfermedad o síndrome que presenten, que se denomina "CATEGORIZACION".

Además, en este mismo bloque de información se puede identificar los medicamentos y su dosificación utilizados en el tratamiento. Esto permite analizar la efectividad y adecuación de los tratamientos empleados en función de la categorización de los pacientes.

Tabla 3. Variables relacionadas con el diagnóstico de los pacientes

VARIABLE	DESCRIPCIÓN	TIPO DE DATO
NOMBRE_DIAGNOSTICO	Nombre descriptivo del diagnóstico del paciente	CADENA DE TEXTO
CIE-10	Código de la Clasificación Internacional de Enfermedades, 10ª revisión, correspondiente al diagnóstico	CADENA DE TEXTO
FECHA_DIAGNOSTICO	Fecha en que se estableció el diagnóstico (formato DD-MM-YYYY)	FECHA
ANTECEDENTES	Antecedentes médicos del paciente	CADENA DE TEXTO

ALERGIAS	Información sobre cualquier alergia que el paciente pueda tener	CADENA DE TEXTO
EN_TRATAMIENTO	Indicador de si el paciente se encuentra actualmente en tratamiento	BOOLEANO
CATEGORIZACION	Categorización del paciente según el tipo de enfermedad o síndrome que presente	CADENA DE TEXTO
NOMBRE_MEDICAMENTO	Nombre del medicamento utilizado en el tratamiento	CADENA DE TEXTO
CANTIDAD_MEDICAMENTO	Cantidad de medicamento prescrita en el tratamiento	ENTERO
DOSIS_MEDICAMENTO	Dosis de medicamento utilizada en el tratamiento	CADENA DE TEXTO

La información personal de los pacientes se ubica en el cuarto bloque, donde se especifican datos relevantes individuales. Sin embargo, es importante señalar que, aunque esta información es personal y detallada, ha sido cuidadosamente anonimizada para cumplir con las normas de privacidad y protección de datos. Esto garantiza la confidencialidad y seguridad de los pacientes, al tiempo que permite la realización de un análisis adecuado.

Tabla 4. Variables con la información personal de los pacientes

VARIABLE	DESCRIPCIÓN	TIPO DE DATO
APELLIDO_PATERNO	Apellido paterno del paciente	CADENA DE TEXTO
APELLIDO_MATERNO	Apellido materno del paciente	CADENA DE TEXTO

PRIMER_NOMBRE	Primer nombre del paciente	CADENA DE TEXTO
SEGUNDO_NOMBRE	Segundo nombre del paciente	CADENA DE TEXTO
FECHA_DE_NACIMIENTO	Fecha de nacimiento del paciente (formato DD-MM-YYYY)	FECHA
EDAD	Edad del paciente	ENTERO
ESCALA_DE_EDAD	Escala utilizada para definir la edad	CADENA DE TEXTO
LUGAR_DE_NACIMIENTO	Lugar de nacimiento del paciente	CADENA DE TEXTO
CEDULA_PASAPORTE	Número de identificación del paciente (cédula o pasaporte)	CADENA DE TEXTO
TELEFONO	Número de teléfono del paciente	CADENA DE TEXTO
SEXO	Sexo del paciente	CADENA DE TEXTO
NACIONALIDAD	Nacionalidad del paciente	CADENA DE TEXTO
NOMBRE_PADRE	Nombre completo del padre del paciente	CADENA DE TEXTO
NOMBRE_MADRE	Nombre completo de la madre del paciente	CADENA DE TEXTO
CEDULA_PADRE	Número de identificación (cédula) del padre del paciente	CADENA DE TEXTO
CEDULA_MADRE	Número de identificación (cédula) de la madre del paciente	CADENA DE TEXTO

El siguiente bloque de información trata exclusivamente sobre el proceso de atención médica, comenzando desde el motivo de consulta del paciente. Las propiedades de este bloque son fácilmente entendibles, dado que sus denominaciones ilustran su vínculo con los datos.

Tabla 5. Variables relacionadas con la atención médica de los pacientes

VARIABLE	DESCRIPCIÓN	TIPO DE DATO
MOTIVO_CONSULTA	Describe la razón por la que el paciente solicitó la consulta.	CADENA DE TEXTO
ENFERMEDAD_ACTUAL	Describe la enfermedad o condición médica actual que padece el paciente.	CADENA DE TEXTO
EXAMEN_FISICO	Detalles del examen físico realizado al paciente durante la consulta.	CADENA DE TEXTO
ANALISIS	Análisis realizado por el médico basado en los hallazgos del examen físico.	CADENA DE TEXTO
PLAN	Plan de tratamiento o cuidado propuesto para el paciente.	CADENA DE TEXTO
NUMERO_EXPEDIENTE_MEDICO	Número asignado al expediente médico del paciente.	ENTERO
TIPO_CONSULTA	Tipo de consulta médica realizada (por ejemplo, inicial, de seguimiento, etc.)	CADENA DE TEXTO
MEDICO_RESPONSABLE	Nombre del médico a cargo de la consulta del paciente.	CADENA DE TEXTO

EQUIPO_MEDICO	Personal médico adicional involucrado en la atención del paciente.	CADENA DE TEXTO
RESUMEN_TRATAMIENTO	Breve descripción del tratamiento realizado o propuesto.	CADENA DE TEXTO

De manera similar, para el último bloque se obtiene información sobre los antecedentes prenatales, natales e información básica de la madre durante el embarazo.

Tabla 6. Variables con datos natales y prenatales de los pacientes

VARIABLE	DESCRIPCIÓN	TIPO DE DATO
GENERO	Sexo del paciente.	CADENA DE TEXTO
PROVINCIA_ID	Identificador de la provincia.	ENTERO
ANO_NACIMIENTO	Año de nacimiento del bebé.	ENTERO
ACIDO_FOLICO	Información sobre la ingesta de ácido fólico durante el embarazo.	BOOLEANO
TOXOPLASMOSIS	Presencia o ausencia de toxoplasmosis durante el embarazo.	BOOLEANO
HERPES_SIMPLE	Presencia o ausencia de herpes simple durante el embarazo.	BOOLEANO
VIH	Presencia o ausencia de VIH durante el embarazo.	BOOLEANO
SANGRADO	Información sobre la existencia de sangrado durante el embarazo.	BOOLEANO

AMENAZA_PARTO_PREMATURO	Información sobre si existió amenaza de parto prematuro.	BOOLEANO
VOMITO_EMBARAZO	Información sobre la presencia de vómitos durante el embarazo.	BOOLEANO
MEDICACION_EMBARAZO	Información sobre la medicación tomada durante el embarazo.	CADENA DE TEXTO
INFECCIONES	Presencia o ausencia de infecciones durante el embarazo.	BOOLEANO
HIPERTENSION_ARTERIAL	Presencia o ausencia de hipertensión arterial durante el embarazo.	BOOLEANO
DIABETES_GESTACIONAL	Presencia o ausencia de diabetes gestacional durante el embarazo.	BOOLEANO
PREECLAMPSIA	Presencia o ausencia de preeclampsia durante el embarazo.	BOOLEANO
REQUIERE_REANIMACION	Información sobre si el bebé requirió reanimación al nacer.	BOOLEANO
OXIGENOTERAPIA	Información sobre si el bebé requirió oxigenoterapia al nacer.	BOOLEANO
ICTERICIA	Presencia o ausencia de ictericia en el bebé.	BOOLEANO
HIPOGLICEMIA	Presencia o ausencia de hipoglucemia en el bebé.	BOOLEANO
EDAD	Edad de la madre al momento del nacimiento del bebé.	ENTERO

RUBEOLA_CITOMEGALOVIRUS	Presencia o ausencia de rubeola o citomegalovirus durante el embarazo.	BOOLEANO
SENO_MATERNO_EXCLUSIVO	Información sobre si se ha alimentado al bebé exclusivamente con leche materna.	BOOLEANO
APGAR1	Puntuación APGAR del bebé al minuto de nacer.	ENTERO
APGAR5	Puntuación APGAR del bebé a los cinco minutos de nacer.	ENTERO
PARTO_PREMATURO	Información sobre si el bebé nació prematuro.	BOOLEANO
ABORTOS_PREVIOS	Información sobre abortos previos de la madre.	ENTERO
BAJO_PESO	Información sobre si el bebé nació con bajo peso.	BOOLEANO
HOSPITALIZACION	Información sobre si el bebé requirió hospitalización después del nacimiento.	BOOLEANO

3.2.3. Identificación de patrones y relaciones iniciales

En esta sección, se exploran patrones y relaciones en los datos médicos pediátricos recopilados mediante la contextualización y análisis de distintas variables, como prevalencia de enfermedades específicas, factores de riesgo, efectividad de tratamientos y tendencias demográficas y geográficas. El análisis de estos aspectos puede revelar información valiosa sobre la correlación entre ciertas condiciones de salud y factores de riesgo, así como la eficacia de diferentes intervenciones médicas.

Para identificar dichos patrones, se utilizan diversas técnicas estadísticas y de aprendizaje automático. Estos modelos se aplican a los datos agrupados según marcos temporales (por ejemplo, año, semestre y mes), sexo del paciente y características demográficas o geográficas. Al evaluar el rendimiento de estos modelos en conjuntos de entrenamiento y prueba, se determina qué enfoques son más efectivos para descubrir relaciones en los datos médicos pediátricos.

3.3 Preparación de los datos

3.3.1. Consolidación de información en un repositorio

Con el objetivo de facilitar el acceso y la manipulación de los datos durante el proceso de análisis, se utilizó una base de datos no relacional MongoDB, que ofrece flexibilidad y escalabilidad en el almacenamiento y la gestión de la información. La elección de MongoDB se debe a su capacidad para manejar eficientemente datos complejos y heterogéneos, lo cual resulta especialmente útil en el contexto de datos médicos. Al consolidar la información en un único repositorio, se simplifica el proceso de análisis y se asegura una mayor coherencia y consistencia en el tratamiento de los datos a lo largo del proyecto.

3.3.2. Limpieza de la información

Esta etapa se enfoca en la limpieza y preparación de los datos obtenidos, donde se combinaron y examinaron los archivos proporcionados. Con base en esta información, se abordarán los siguientes aspectos clave para asegurar la calidad de los datos antes de avanzar a las siguientes fases del análisis:

- **Validación de factores**: Verificación de la consistencia y corrección de los factores en los datos.
- **Validación de caracteres**: Revisión y corrección de caracteres no válidos o incoherentes en los datos.
- **Validación de fechas**: Asegurar que las fechas tengan un formato adecuado para el análisis.
- **Búsqueda de valores no asignados**: Identificación y manejo de valores faltantes o no asignados.

- **Búsqueda de valores duplicados**: Localización y eliminación de registros duplicados en el conjunto de datos.

Es importante resaltar que, dado que los datos analizados corresponden a información médica, se garantizará la anonimización de los datos de los pacientes para proteger su privacidad. Cada uno de los aspectos mencionados anteriormente es crucial para garantizar la calidad y confiabilidad de los resultados del análisis, evitando así posibles errores o conclusiones erróneas.

3.3.3. Valores Perdidos

El manejo de valores perdidos o faltantes es un paso crucial al preparar la data para el análisis. Los valores perdidos pueden ocurrir por una variedad de razones, como errores en la recopilación de datos, no respuesta a ciertas preguntas en cuestionarios, o simplemente porque cierta información no estaba disponible en el momento de la recopilación de los datos.

En este caso se aplicaron dos métodos dependiendo de las variables afectadas. Se borró los registros cuando se notó que la cantidad de datos faltantes por paciente era demasiado extensa y se aplicó la imputación de datos faltantes. En este estudio, se utilizó el método de imputación media para los valores faltantes.

En este método, los valores faltantes para una variable dada son reemplazados por el valor medio de esa variable de los casos disponibles. Es importante tener en cuenta que cualquier método de imputación introduce cierta cantidad de error y puede llevar a estimaciones sesgadas si los datos no están faltando al azar. Por lo tanto, la imputación se utilizó con precaución y se realizó un análisis de sensibilidad para evaluar el impacto de la imputación en los resultados finales del estudio.

3.3.4. Manejo de Variables Categóricas

Las variables categóricas son aquellas que tienen un número fijo y generalmente pequeño de niveles o categorías posibles. En este conjunto de datos son el sexo del paciente, la provincia de origen y el diagnóstico médico CIE10. Aunque estas variables contienen información valiosa, los modelos de regresión y muchos otros tipos de análisis estadístico requieren datos numéricos. Por lo tanto, necesitamos

convertir nuestras variables categóricas en una forma que los modelos puedan manejar.

En este estudio, utilizamos la estrategia de codificación One-Hot para nuestras variables categóricas. Este método implica crear nuevas variables binarias para cada nivel de la variable categórica. Por ejemplo, para la variable "sexo", creamos dos nuevas variables: "sexo_masculino", donde si el paciente es masculino es 1, y "sexo_femenino", donde si el paciente es masculino es 0.

La codificación One-Hot es una estrategia de codificación simple y efectiva que puede manejar cualquier número de niveles categóricos y no hace suposiciones acerca de las relaciones entre los niveles. Sin embargo, puede llevar a un gran aumento en la cantidad de datos si la variable categórica tiene muchos niveles.

3.3.5. Selección de características

En esta lista de variables, se ha excluido intencionalmente información como nombres, apellidos y números de identificación. La razón detrás de esta decisión es asegurar que los datos del estudio sean anonimizados y se mantenga la privacidad de los pacientes. La anonimización de datos es una práctica importante en investigaciones médicas y de salud, ya que protege la identidad de los individuos y cumple con las regulaciones de privacidad y protección de datos.

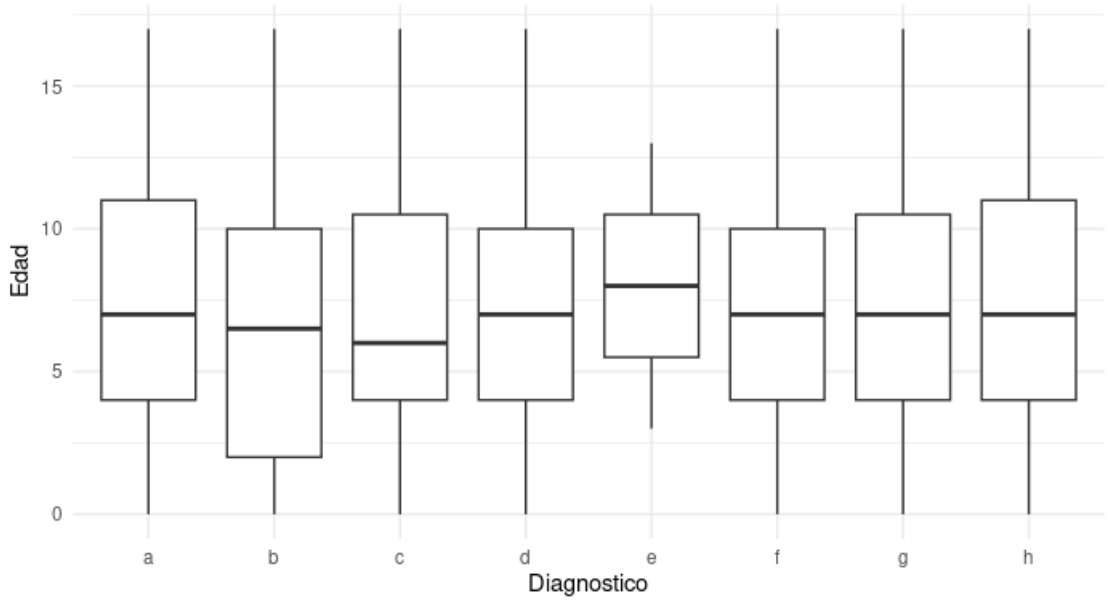
- Número de Caso: alfanumérico, valor incremental (hasta 10 caracteres)
- Fecha de atención: fecha (formato AAAA-MM-DD)
- Médico tratante: texto (hasta 100 caracteres)
- Edad: entero (hasta 3 dígitos)
- Fecha de nacimiento: fecha (formato AAAA-MM-DD)
- Sexo: categórico (2 niveles: masculino, femenino)
- Nacionalidad: texto (hasta 50 caracteres)
- Lugar de nacimiento: texto (hasta 100 caracteres)
- Estado Civil de los padres: categórico (hasta 5 niveles, por ejemplo, soltero, casado, divorciado, viudo, otro)
- Profesión de los padres: texto (hasta 100 caracteres)
- Diagnósticos médicos: texto (hasta 500 caracteres)

- Medicamentos recetados: texto (hasta 500 caracteres)
- Vacunas administradas: texto (hasta 500 caracteres), incluida la fecha de administración (formato AAAA-MM-DD) y el tipo de vacuna
- Signos vitales: numérico (hasta 6 dígitos para cada signo vital)
- Resultados de pruebas de laboratorio: texto (hasta 1000 caracteres)
- Historial médico: texto (hasta 1000 caracteres)
- Antecedentes familiares: texto (hasta 1000 caracteres)
- Observaciones médicas: texto (hasta 1000 caracteres)
- Provincia_ID: alfanumérico (hasta 5 caracteres)
- Año de Nacimiento: entero (4 dígitos)
- Ácido Fólico: categórico (2 niveles: sí, no)
- Toxoplasmosis: categórico (2 niveles: sí, no)
- Herpes Simple: categórico (2 niveles: sí, no)
- VIH: categórico (2 niveles: sí, no)
- Sangrado: categórico (2 niveles: sí, no)
- Amenaza de Parto Prematuro: categórico (2 niveles: sí, no)
- Vómito durante el Embarazo: categórico (2 niveles: sí, no)
- Medicación durante el Embarazo: categórico (2 niveles: sí, no)
- Infecciones: categórico (2 niveles: sí, no)
- Hipertensión Arterial: categórico (2 niveles: sí, no)
- Diabetes Gestacional: categórico (2 niveles: sí, no)
- Preeclampsia: categórico (2 niveles: sí, no)
- Requiere Reanimación: categórico (2 niveles: sí, no)
- Oxigenoterapia: categórico (2 niveles: sí, no)
- Ictericia: categórico (2 niveles: sí, no)
- Hipoglucemia: categórico (2 niveles: sí, no)
- Edad: entero (hasta 3 dígitos)
- Rubeola Citomegalovirus: categórico (2 niveles: sí, no)
- Seno Materno Exclusivo: categórico (2 niveles: sí, no)
- Apgar1: entero (hasta 2 dígitos)
- Apgar5: entero (hasta 2 dígitos)
- Parto Prematuro: categórico (2 niveles: sí, no)
- Abortos Previos: categórico (2 niveles: sí, no)

- Bajo Peso: categórico (2 niveles: sí, no)
- Hospitalización: categórico (2 niveles: sí, no)

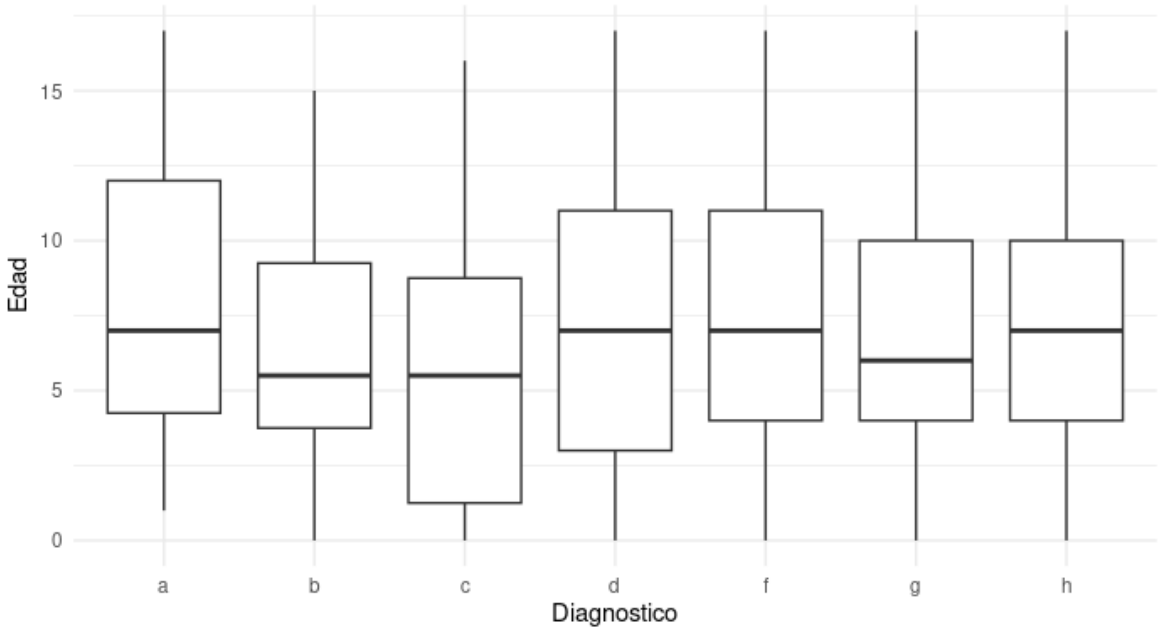
A continuación, se muestra la Figura 2, la cual es una representación visual de la distribución de las edades en relación con los diagnósticos en el conjunto de entrenamiento. Este gráfico proporciona una comprensión visual de cómo varían las edades entre diferentes categorías de diagnóstico.

Figura 2. Conjunto de entrenamiento Edad vs Diagnóstico



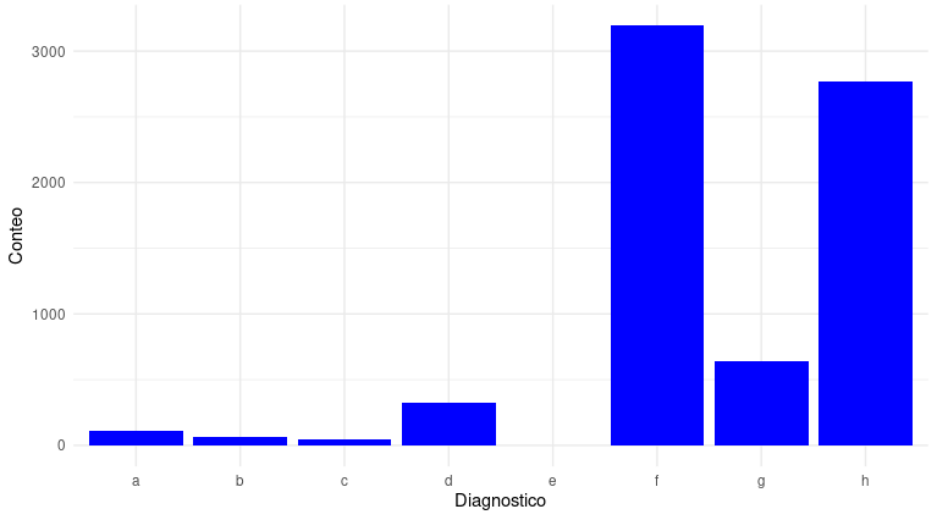
La Figura 3 ilustra la distribución de las edades en función de los diagnósticos en el conjunto de pruebas. Al igual que en la Figura 2, se emplea un gráfico de caja de bigotes para representar la mediana, los cuartiles y los posibles valores atípicos de la edad de los pacientes en distintos diagnósticos. Este gráfico es esencial para contrastar cómo las tendencias y la variación en la edad se mantienen o cambian en el conjunto de pruebas en comparación con el conjunto de entrenamiento.

Figura 3. Conjunto de pruebas Edad vs Diagnóstico



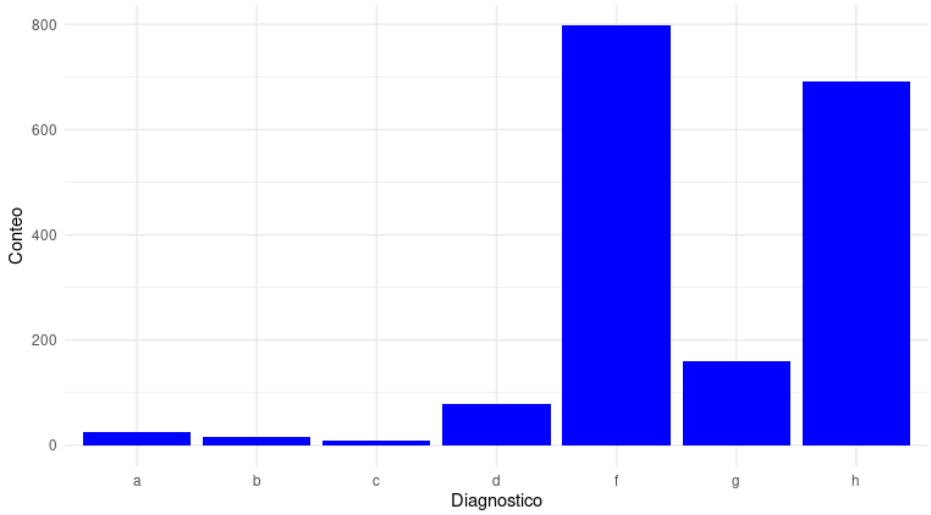
La Figura 4 muestra la distribución de frecuencia de los diagnósticos en el conjunto de entrenamiento. Se presenta un gráfico de barras en el que el eje vertical representa la cantidad de apariciones y el eje horizontal los distintos diagnósticos. Este gráfico permite entender la prevalencia de cada diagnóstico en el conjunto de entrenamiento, proporcionando una vista inicial sobre el balance de las categorías de diagnóstico en nuestro modelo.

Figura 4. Conjunto de entrenamiento frecuencia diagnósticos



La Figura 5 representa la distribución de frecuencias de los diagnósticos en el conjunto de pruebas mediante un gráfico de barras. Esta visualización es clave para entender la prevalencia de cada diagnóstico en el conjunto de pruebas, y permite verificar si la distribución en este conjunto es similar a la observada en el conjunto de entrenamiento, lo cual es esencial para validar la capacidad de generalización del modelo.

Figura 5. Conjunto de pruebas frecuencias diagnósticos



CAPÍTULO IV – IMPLEMENTACIÓN E INTERPRETACIÓN DE LOS RESULTADOS PROCESADOS

En este capítulo, se aborda la implementación del proyecto, centrándose en el modelado, la evaluación y el despliegue de los resultados. Se comienza con la selección de técnicas y algoritmos de minería de datos apropiados para abordar el problema en cuestión y luego se procede al entrenamiento y validación de los modelos.

A continuación, se evalúa el rendimiento de los modelos utilizando métricas apropiadas y se comparan los resultados con los objetivos del negocio, identificando posibles mejoras en el proceso. Por último, en la fase de despliegue, se diseñan visualizaciones efectivas para comunicar los resultados a los interesados, asegurándose de que la información obtenida a través del análisis y modelado de datos sea accesible y útil para la toma de decisiones.

4.1 Modelado

En la etapa de modelado, se aplicaron diferentes técnicas de aprendizaje automático para analizar la relación entre las variables disponibles y el resultado de interés. Para cada modelo, se seleccionaron y configuraron cuidadosamente los hiperparámetros apropiados. La configuración de estos hiperparámetros es un paso crítico, ya que su ajuste puede mejorar significativamente el desempeño del modelo. Dado que no existe una única configuración de hiperparámetros que funcione para todos los escenarios, se realizó una búsqueda en pro de conseguir el modelo que brinde mejor rendimiento, equilibrando el sesgo y la varianza para prevenir el sobreajuste o el subajuste. Así, el modelado permite identificar los patrones más significativos en los datos, y proporcionar insights valiosos sobre los factores que afectan la salud pediátrica.

4.1.1. Selección de Técnicas y Algoritmos de Minería de Datos

- **Modelo de regresión con vectores soporte**

  El funcionamiento de SVR se basa en el mismo principio de clasificación que las Máquinas de Vector Soporte (SVM), es decir, en la capacidad de ajustar un modelo a los datos de entrada. Sin embargo, SVR se adapta a la utilización de variables numéricas en lugar de categóricas. Además,

SVR es considerado una técnica no paramétrica debido a que su salida no depende de la distribución de las variables dependientes e independientes subyacentes, sino de su función de kernel. Esto permite la construcción de un modelo no lineal sin que varíen las variables explicativas, mediante la generación de una curva capaz de ajustar los datos y garantizar la separación entre esta y valores específicos. Esta característica ayuda a una fácil interpretación del modelo generado.

El algoritmo se implementa de la siguiente manera:

$$svm <- train(x = X\_train, y = y\_train, method = "svmRadial", trControl = trainControl(method = "cv", number = 5), preProcess = c("center","scale"))$$

A continuación, se presenta la precisión del modelo de regresión con vectores de soporte.

Tabla 7. Precisión modelo de regresión con vectores de soporte

Ajuste	Precisión.del.modelo
1	0.5510434
2	0.8314509
3	0.6190271
4	0.8844897

Autor: Leonardo Chamorro

- **KNN – Vecinos cercanos**

El algoritmo K-Vecinos Cercanos (KNN) se aplicó en este estudio para identificar patrones en los factores que afectan la salud pediátrica. Este modelo es útil gracias a su capacidad para manejar relaciones complejas y no lineales entre variables, ya que no asume una distribución específica de los datos. Sin embargo, el número óptimo de 'vecinos' (K) para considerar en las predicciones requerirá un ajuste cuidadoso para maximizar el rendimiento del modelo. Además, dado que KNN es sensible a la escala de las variables, se realizará un proceso de

normalización de datos previo a la implementación del modelo (Peterson, 2009).

Posteriormente, se implementa el algoritmo de la siguiente manera:

$$knn <- train(x = X\_train, y = y\_train, method = "knn", trControl = trainControl(method = "cv", number = 5), preProcess = c("center", "scale"), tuneGrid = expand.grid(k = k))$$

En la Tabla 8, se muestra la precisión del modelo KNN obtenida mediante el procedimiento anterior.

Tabla 8. Precisión modelo KNN Vecinos Cercanos

Ajuste	Precisión.del.modelo
1	0.7485853
2	0.7623715
3	0.6533197
4	0.6971167

Autor: Leonardo Chamorro

- **Modelo de regresión con árboles de decisión**

  Este fue desarrollado a partir del algoritmo de clasificación y regresión (CART) propuesto en 1984 por Breiman (Breiman, Friedman, Stone, & Olshen, 1984), este método se basa en árboles de decisión (DT, Decision Trees) y utiliza una segmentación binaria no paramétrica. En cuanto al uso de la regresión, este método es similar a los modelos mencionados en las subsecciones anteriores, ya que la variable dependiente es continua y se espera que los diferentes valores de los nodos terminales se reduzcan a la media de las observaciones en la región de análisis (Serna, 2009).

En este caso, el modelo se ajusta usando la función "train" del paquete "caret" en R, especificando "rpart" como método, y se realiza validación cruzada con cinco particiones de los datos.

$$dt <- train(x = X\_train, y = y\_train, method = "rpart", trControl = trainControl(method = "cv", number = 5), preProcess = c("center", "scale"))$$

En la Tabla 9, se presenta la precisión obtenida con el modelo de regresión con árboles de decisión.

Tabla 9. Precisión modelo de regresión con árboles de decisión

Ajuste	Precisión.del.modelo
1	0.4536739
2	0.7384877
3	0.7311939
4	0.7390925

Autor: Leonardo Chamorro

- **Modelo de regresión con bosques aleatorios**

En 2001, Breiman propuso el método de Bosques Aleatorios o Random Forests (Breiman, Random Forests, 2001). La regresión con Bosques Aleatorios funciona generando n cantidad de árboles de decisión aleatorios sobre el mismo conjunto de datos. La decisión final en la regresión se toma a partir del cálculo del promedio de las salidas o predicciones de todos los árboles. Sin embargo, presenta una limitante en la predicción de valores que se encuentran fuera del rango de los diversos valores del conjunto de entrenamiento.

La precisión del modelo de bosques aleatorios, similar al caso del algoritmo KNN, se mide utilizando la tasa de clasificación correcta. Esta métrica se calcula de la siguiente manera:

$$rf <- train(x = X\_train, y = y\_train, method = "rf", trControl = trainControl(method = "cv", number = 5), preProcess = c("center", "scale"))$$

$$accuracy <- confusionMatrix(y\_pred, y\_test)\$overall["Accuracy"]$$

En la Tabla 10, se presenta la precisión obtenida con el modelo de regresión con bosques aleatorios.

Tabla 10. Precisión modelo de regresión con bosques aleatorios

Ajuste	Precisión.del.modelo
1	0.7937062
2	0.8804330
3	0.8161501
4	0.8862298

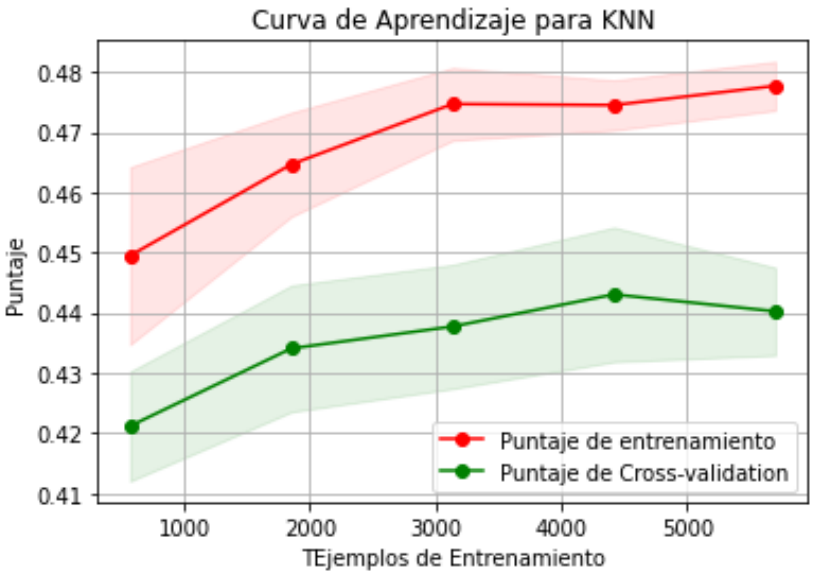
Autor: Leonardo Chamorro

4.2. Evaluación

4.2.1. Métricas de evaluación del rendimiento del modelo

Los modelos utilizados se evaluaron en función de su precisión y capacidad para generalizar los datos no vistos. Los hiperparámetros de cada modelo fueron cuidadosamente ajustados y validados mediante técnicas de validación cruzada para optimizar su rendimiento. El rendimiento de cada modelo se presentó en las tablas previas.

Figura 6. Curva de Aprendizaje para KNN

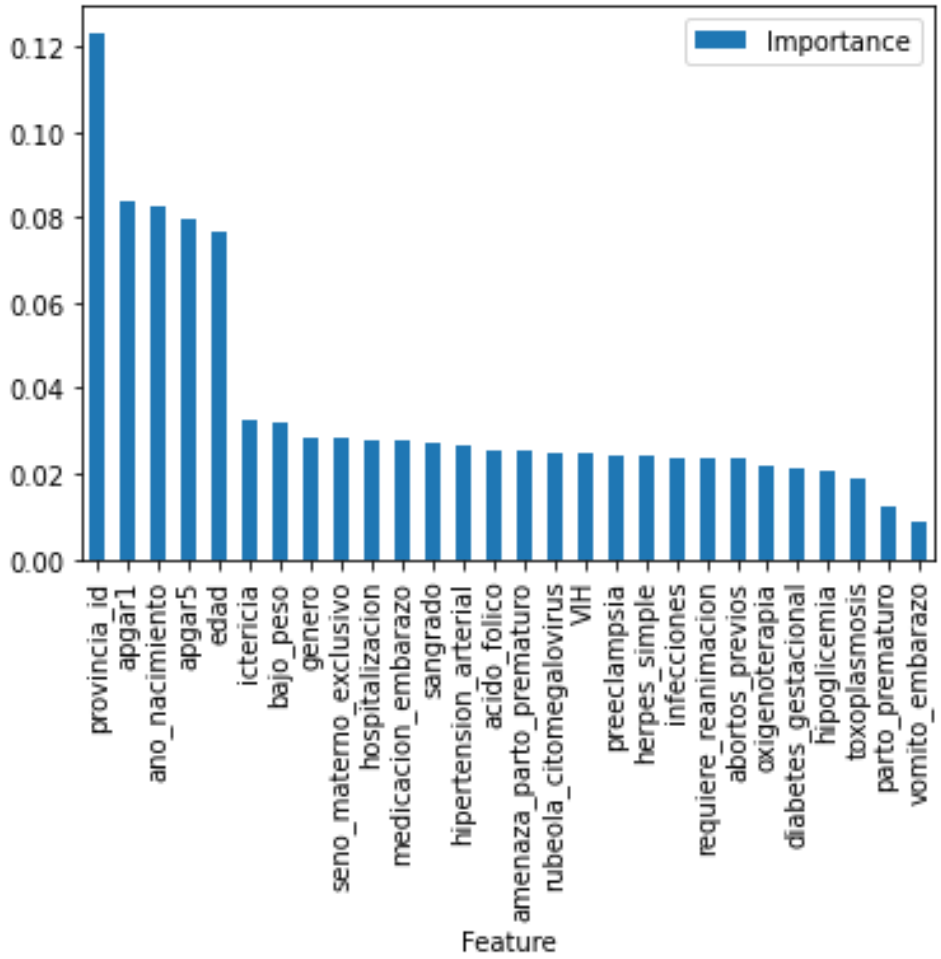


4.2.2. Comparación de resultados del modelo versus objetivos del negocio

El objetivo principal de este estudio es identificar los factores que más influyen en el diagnóstico de enfermedades neuropsicológicas en pacientes pediátricos en Ecuador. Los resultados obtenidos de los modelos permiten no sólo entender mejor estos factores, sino también hacer inferencias basadas en los datos y facilitar la toma de decisiones informada. El modelo ayuda a los médicos a prever las condiciones que pueden predisponer a los niños a desarrollar estas enfermedades, contribuyendo al objetivo de mejorar la salud pediátrica en Ecuador. Esto se evidencia en el siguiente gráfico.

La Figura 7 representa los factores evaluados en los modelos a través de un gráfico de barras. En el eje x se encuentran las principales características consideradas en el modelo. Este gráfico permite visualizar cuáles son los factores que más influyen en el diagnóstico de enfermedades neuropsicológicas en pacientes pediátricos. La longitud de cada barra indica el grado de importancia de cada característica en el modelo, facilitando la identificación de los factores más relevantes.

Figura 7. Factores evaluados en los modelos aplicados



4.3. Despliegue

El despliegue de la solución implica la implementación del modelo en el entorno de producción y su integración en el flujo de trabajo existente para que pueda ser utilizado de manera efectiva en la toma de decisiones relacionadas con la atención neuropsicológica de los pacientes.

4.3.1. Diseño de visualizaciones para la comunicación de resultados

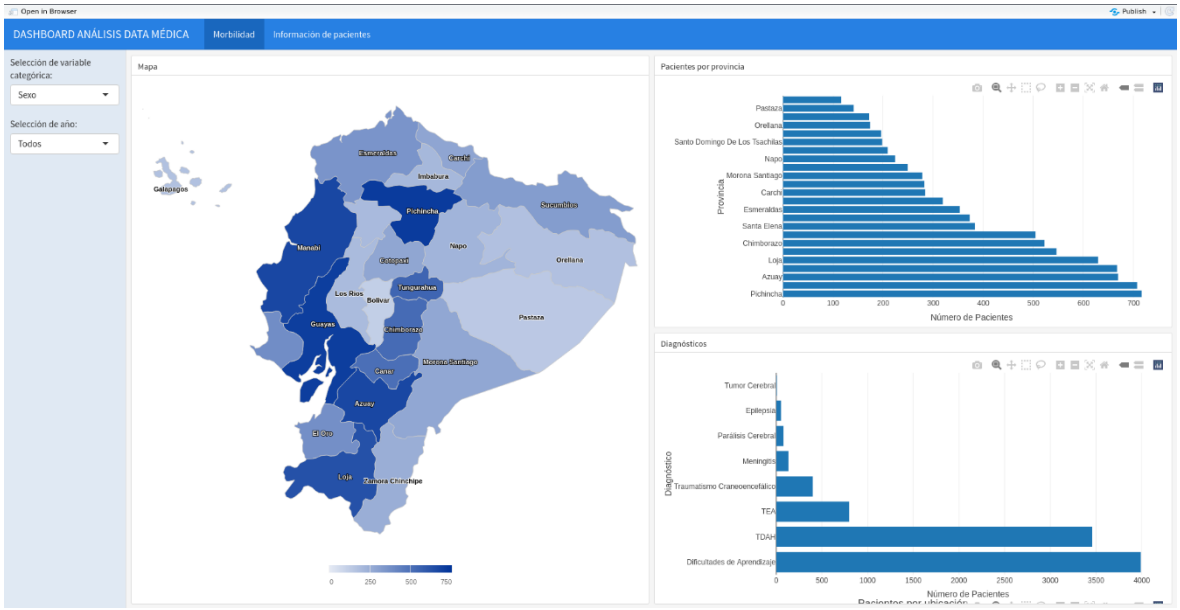
Para facilitar la interpretación de los resultados, se utilizó el lenguaje de programación R, que es ampliamente utilizado en el análisis de datos y permite la creación de visualizaciones dinámicas e interactivas a través de paquetes como Shiny y R Markdown (Santana & Farfán, 2014). Se desarrolló un panel de

control o 'dashboard' que ofrece una visión clara y fácilmente interpretable de los factores que influyen en los diagnósticos de enfermedades neuropsicológicas en niños en Ecuador.

La interfaz inicial del panel de control, mostrada en la Figura 8 presenta una serie de elementos clave para facilitar la navegación y la interpretación de los datos. La cabecera del dashboard incluye su título y las categorías de visualización que los usuarios pueden explorar. Un menú lateral a la izquierda proporciona una serie de parámetros ajustables, lo que permite a los usuarios personalizar las visualizaciones de acuerdo con sus necesidades específicas.

El primer conjunto de visualizaciones se centra en "Pacientes por ubicación". Aquí, se presenta un mapa interactivo que muestra la distribución geográfica de los pacientes pediátricos. Al pasar el cursor sobre las distintas provincias, los usuarios pueden ver el número de pacientes en cada una de ellas. Esta visualización se complementa con un diagrama de barras horizontal en la parte superior derecha, que representa los mismos datos. Además, otro diagrama de barras horizontal proporciona una visión detallada de los diagnósticos más frecuentes y relevantes en el contexto de este estudio.

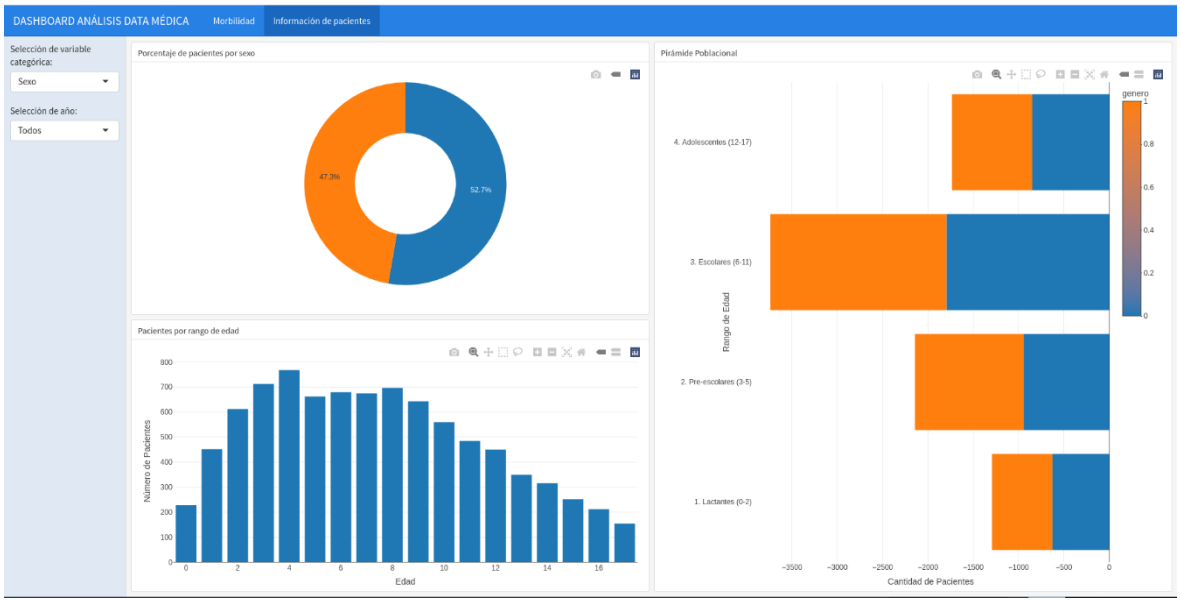
Figura 8. Interfaz de Ubicación Geográfica de Pacientes



En cuanto a la información demográfica de los pacientes, esta se presenta de manera generalizada pero informativa. El género de los pacientes se visualiza mediante un diagrama de rosquilla, proporcionando la distribución porcentual de hombres, mujeres y casos sin definir. Los rangos de edad, por otro lado, se muestran a través de un diagrama de barras verticales que refleja la distribución etaria de los pacientes a nivel nacional.

Por último, se incluye un gráfico que representa la pirámide poblacional, proporcionando una vista detallada de la distribución por edades y géneros de los pacientes. Este tipo de gráfico es especialmente útil para identificar tendencias y patrones demográficos que podrían tener relevancia para la planificación y la toma de decisiones en el contexto de la atención neuropsicológica pediátrica.

Figura 9. Interfaz del Área de Información de Pacientes



4.3.2. Diseminación de información a los interesados

Los resultados se han presentado en dashboards diseñados para proporcionar información relevante a diversas partes interesadas, como médicos, administradores de hospitales y formuladores de políticas de salud, de manera que puedan tomar decisiones informadas basadas en los resultados de los

modelos de regresión. El panel permite a los usuarios explorar los datos interactivamente y profundizar en áreas de interés particular. El código fuente para el panel de control se proporciona en el Anexo II.

4.4 Evaluación MOS

La Evaluación MOS (Mean Opinion Score) es una métrica crucial para medir la calidad global percibida por los usuarios de un sistema. En el contexto de este estudio, la evaluación MOS se ha empleado para determinar la usabilidad, interpretación y funcionalidad general de las visualizaciones de datos desarrolladas en el dashboard. Este proceso de evaluación se llevó a cabo en una muestra de médicos que laboran en las clínicas que brindaron acceso a los datos que alimentan nuestros modelos.

Durante esta fase de evaluación, se recogió tanto datos cualitativos como cuantitativos, principalmente a través de pruebas de usuario y encuestas de satisfacción, lo que permitió obtener un panorama más completo acerca de cómo estos profesionales de la salud interactuaron con nuestras herramientas y cómo percibieron su utilidad.

Los resultados de la evaluación MOS han demostrado que las visualizaciones y herramientas desarrolladas fueron, en su mayoría, bien recibidas, proporcionando una experiencia intuitiva y valiosa para los usuarios. Estos hallazgos no solo validan la eficacia y utilidad del trabajo que se ha realizado, sino que también establecen una base sólida para futuras mejoras y adaptaciones de nuestro sistema.

CAPÍTULO V – CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

1. A través de este estudio, se ha aportado un modelo práctico para el análisis de datos orientado a investigar la prevalencia de enfermedades neuropsicológicas en niños en Ecuador. La implementación de múltiples modelos de aprendizaje automático, con la correspondiente configuración de sus hiperparámetros, permitió identificar los factores más impactantes que contribuyen a la aparición de estas condiciones.

2. El modelo generado proporciona un recurso valioso para realizar predicciones sobre casos futuros en función de las variables de entrada. Estas variables se gestionan en el panel de control del tablero de la herramienta, el cual facilita la visualización de los resultados obtenidos.

3. El marco de trabajo CRISP-DM facilitó a realizar este estudio, por ser una metodología que permite un enfoque estructurado y eficaz para la aplicación de técnicas de aprendizaje automático en el ámbito de la salud. La metodología CRISP-DM proporcionó un camino claro y coherente, permitiendo una secuencia lógica de etapas desde la comprensión del negocio hasta el despliegue del modelo.

4. Se debe destacar que tal como sugiere este modelo, se requirió iterar varias veces entre algunas etapas, principalmente entre la modelización y la evaluación, para lograr un rendimiento óptimo y una mejor comprensión de los datos. Esto resalta la flexibilidad y adaptabilidad de la metodología CRISP-DM, que no solo permite, sino que también fomenta la revisión y la mejora continua de los modelos.

5. En la práctica realizada, la metodología CRISP-DM demostró ser una herramienta valiosa y poderosa por su modelo estructurado, además permite iterar entre sus etapas cuando fuere necesario, permitieron obtener resultados óptimos y proporcionaron un marco sólido y replicable para futuras investigaciones.

6. Una de las tareas de mayor complejidad en este proceso fue la recolección y depuración de los datos. La naturaleza de estos elementos como insumo esencial para el análisis subraya su importancia, sobre todo considerando la metodología utilizada. La eficacia del análisis dependía en gran medida de la calidad y exactitud de estos datos, de ahí que su recolección y limpieza fueran aspectos cruciales de este proceso.

7. El presente estudio contribuye a la investigación y comprensión de las enfermedades neuropsicológicas en la población pediátrica de Ecuador y proporciona una base para futuros estudios e intervenciones, y establece un precedente en la aplicación de técnicas de aprendizaje automático y visualización de datos en este campo.

5.2 Recomendaciones

1. Para un análisis de datos válido y confiable, es crucial contar con datos actualizados y depurados. La calidad de los datos influye directamente en la precisión de los resultados. Esto implica realizar una limpieza exhaustiva y garantizar que los datos estén actualizados. Al priorizar la calidad y actualización de los datos, se obtendrán conclusiones más significativas en el análisis de datos.

2. Antes de seleccionar el modelo a utilizar en el análisis de datos, es crucial comprender los datos y definir los resultados esperados. Es importante tener claridad sobre los objetivos y las necesidades específicas para garantizar una elección acertada del modelo de análisis de datos.

3. Es importante seleccionar las herramientas adecuadas para construir un dashboard que despliegue los resultados del modelo desarrollado de manera efectiva. La elección de herramientas funcionales permite un despliegue eficiente y comprensible de los resultados del análisis de datos. Esto proporciona una experiencia satisfactoria a los usuarios del panel de control.

4. Si bien los modelos desarrollados en este estudio han demostrado un rendimiento satisfactorio, aún existen áreas potenciales para mejorar y optimizar su eficacia. Por ejemplo, una mejora considerable podría derivarse de la inclusión de factores ambientales en los modelos, como la calidad del aire y la presencia de metales pesados en el entorno del paciente. Estos factores, aunque a veces se pasan por alto, han demostrado tener un impacto significativo en la salud neurológica y su inclusión podría mejorar aún más la precisión y la relevancia de las predicciones.

5. Luego de analizar los datos bajo el modelo CRISP-DM, se ve la necesidad de incluir datos sobre el estilo de vida de los padres, ya que factores como el estrés, la dieta y la exposición a sustancias tóxicas pueden influir en la salud neurológica de los niños.

6. Una vía adicional que se podría explorar es la aplicación de técnicas de aprendizaje profundo para analizar los registros médicos de los pacientes. Estos métodos podrían utilizarse para extraer información adicional y valiosa de los textos de los informes médicos, lo que podría ayudar a mejorar la precisión de las predicciones.

Bibliografía

- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, 487-499. https://doi.org/10.1016/B978-1-55860-277-4.50034-4

- Amin, R., Prasad, P. W., & Singh, R. K. (2016). Data mining approach to identify factors causing malnutrition among children in India. Procedia Computer Science, 79, 692-700.

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

- Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

- Card, S. (2008). Information visualization. Annual Review of Computer Science, 2, 151-178.

- Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann.

- Cardona Madariaga, A. M., González Rodríguez, G. A., Rivera Lozano, J. A., & Cárdenas Vallejo, J. A. (2013). Aplicación de regresión lineal para estimar la producción de energía eólica. Revista Ingenierías Universidad de Medellín, 12(22), 101-116.

- Castro Chauca, J. A. (2019). Implementación de un dashboard de inteligencia de negocios para la dirección académica de la Universidad Politécnica Estatal del Carchi. Repositorio Digital Universidad Politécnica Estatal del Carchi.

- Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. Artificial intelligence in medicine, 26(1-2), 1-24.

- Finch, D. & Flenner, A. (2016). Constructive alignment of interactive digital content for 21st century skills. In S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds.), Show Me The Learning. Proceedings ASCILITE 2016 Adelaide (pp. 178-182).

- Free, C., Phillips, G., Galli, L., Watson, L., Felix, L., Edwards, P., ... & Haines, A. (2013). The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. PLoS medicine, 10(1), e1001362.

- García, S., Luengo, J., & Herrera, F. (2016). Data preprocessing in data mining (pp. 9-37). Springer International Publishing.

- Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Morgan Kaufmann.

- Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 1-12. https://doi.org/10.1145/342009.335372

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

- Juan Camilo Giraldo Mejía. (2012). Minería de Datos: Conceptos, técnicas y aplicaciones. Ecoe Ediciones.

- Kaupp, J. (2016). Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications. IGI Global.

- Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

- Kelm, T., Fekete, J.-D., Andreinko, M., & Kohlhase, M. (2008). Cognition-Driven Visualization: Commented Example. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1679-1686.

- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer Science & Business Media.

- Lema Siguencia, M. A. (2016). Dashboard de inteligencia de negocios aplicado a la carrera de medicina de la Universidad Central del Ecuador. Repositorio Digital de la Universidad Central del Ecuador.

- Lizasoain, L., & Joaristi, M. (2012). Manual de análisis de datos en ciencias sociales. Pamploa: Servicio de Publicaciones de la Universidad de Navarra.

- Mendoza Rodríguez, J. R. (2015). Análisis de datos y diseño de un data warehouse para la Autoridad Portuaria Puerto Bolívar. Repositorio Digital de la Universidad de Guayaquil.

- Mohd Ali, N., Muda, Z., Yusof, R., & Rahman, N. A. (2016). Data visualization: a review of visual representation of data. Journal of Physics: Conference Series, 705(1), 012006. https://doi.org/10.1088/1742-6596/705/1/012006

- Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883. DOI: 10.4249/scholarpedia.1883

- Provost, F., & Fawcett, T. (2013). Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

- Pyle, D. (1999). Data preparation for data mining. Morgan Kaufmann.

- Smith, M. A., Barnett, N., Rosenberger, J. G., & Thomson, R. (2018). Applying CRISP-DM to investigate the determinants of infant and young child feeding practices among refugee and migrant mothers in Australia. Nutrients, 10(11), 1668.

- Serna, H. F. (2009). Regresión no paramétrica y árboles de decisión. Universidad Nacional de Colombia.

- Sistemas de soporte a la toma de decisiones. (2010). Técnicas de Minería de Datos. Universidad Tecnológica de Pereira.

- Steyerberg, E. W. (2009). Clinical prediction models: a practical approach to development, validation, and updating (Vol. 238). Springer Science & Business Media.

- Vélez de la Cruz, M. P. (2017). Uso del análisis de datos y un tablero de control para la mejora de la gestión estratégica en Imptek Chova del Ecuador S.A. Repositorio Digital de la Universidad de las Américas.

- Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers.

- Wirth, R., & Hipp, J. (2000). CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39.

- Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3), 372-390. https://doi.org/10.1109/69.846291

Glosario de Términos

1. **Algoritmo**: Conjunto definido de instrucciones o reglas para resolver un problema a través de un número finito de pasos.
2. **Aprendizaje Automático (Machine Learning)**: Subcampo de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para permitir a las máquinas mejorar su rendimiento en una tarea específica mediante la experiencia, sin ser explícitamente programadas.
3. **Base de datos**: Conjunto de datos almacenados y gestionados a lo largo del tiempo, que se utiliza para análisis de tendencias y predicciones.
4. **Dashboard**: Herramienta visual que presenta información de forma clara y concisa, generalmente en forma de gráficos y tablas, que permite a los usuarios monitorear y analizar el rendimiento o los indicadores clave.
5. **Data Warehouse (Almacén de datos)**: Sistema de almacenamiento de datos que contiene datos de múltiples fuentes, los cuales se limpian, transforman y catalogan para su análisis y consulta.
6. **Feature Engineering (Ingeniería de características)**: Proceso de utilizar el conocimiento del dominio para crear características que hacen que los modelos de aprendizaje automático funcionen mejor.
7. **Hiperparámetros**: Configuraciones ajustables que se establecen antes de entrenar un modelo de aprendizaje automático y que influyen en la eficacia del entrenamiento.
8. **Inteligencia de Negocios (Business Intelligence)**: Proceso de transformación de los datos brutos en información útil para fines estratégicos y tácticos de toma de decisiones empresariales.
9. **Lenguaje R**: Lenguaje de programación y entorno de software para análisis estadístico y gráficos.
10. **Minería de Datos**: Proceso que implica la exploración y análisis de grandes cantidades de datos para encontrar patrones significativos o reglas.
11. **Reglas de Asociación**: Técnica de minería de datos utilizada para encontrar relaciones frecuentes o patrones entre conjuntos de elementos en grandes bases de datos.
12. **Variables**: Características o atributos que pueden asumir diferentes valores o categorías y que se utilizan para medir fenómenos de interés en el análisis de datos.

Anexos

Anexo I

```
library(dplyr)
library(caret)
library(randomForest)
library(e1071)
library(ggplot2)

# Cargar el dataset
input_file <- "/home/leonardo/Documents/Tesis/pacientes_clean2.csv"
df <- read.csv(input_file)

# Preprocesamiento de datos
df$genero <- ifelse(df$genero == "Masculino", 1, 0)
df$fecha <- as.(df$fecha)
df$ano_nacimiento <- format(df$fecha, "%Y")

# Seleccionar las columnas deseadas para X
X <- df %>%
  select(genero, provincia_id, ano_nacimiento, acido_folico, toxoplasmosis, herpes_simple,
         VIH, sangrado, amenaza_parto_prematuro, vomito_embarazo, medicacion_embarazo,
         infecciones, hipertension_arterial, diabetes_gestacional, preeclampsia,
         requiere_reanimacion, oxigenoterapia, ictericia, hipoglicemia, edad,
         rubeola_citomegalovirus, seno_materno_exclusivo, apgar1, apgar5,
         parto_prematuro, abortos_previos, bajo_peso, hospitalizacion)

# Variable de salida y
y <- df$diagnostico
```

```
# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)  # Fijar la semilla para reproducibilidad
train_index <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X[train_index, ]
y_train <- y[train_index]
X_test <- X[-train_index, ]
y_test <- y[-train_index]

# Imputar valores faltantes en el conjunto de entrenamiento y prueba
X_train <- na.omit(X_train)
X_test <- na.omit(X_test)

# KNN
k <- 50  # Número de vecinos a considerar
knn <- train(x = X_train, y = y_train, method = "knn", trControl = trainControl(method = "cv", number = 5), preProcess = c("center", "scale"), tuneGrid = expand.grid(k = k))

# Obtener predicciones en el conjunto de prueba
y_pred <- predict(knn, newdata = X_test)

# Calcular la precisión del modelo
accuracy <- confusionMatrix(y_pred, y_test)$overall["Accuracy"]
print(paste("Precisión del modelo KNN:", accuracy * 100))

# Árboles de Decisión
dt <- train(x = X_train, y = y_train, method = "rpart", trControl = trainControl(method = "cv", number = 5),  preProcess = c("center", "scale"))

# Obtener predicciones en el conjunto de prueba
y_pred <- predict(dt, newdata = X_test)
```

```
# Calcular la precisión del modelo

accuracy <- confusionMatrix(y_pred, y_test)$overall["Accuracy"]

print(paste("Precisión del modelo Decision Tree:", accuracy * 100))


# Random Forest

rf <- train(x = X_train, y = y_train, method = "rf", trControl = trainControl(method = "cv", number = 5), preProcess = c("center", "scale"))


# Obtener predicciones en el conjunto de prueba

y_pred <- predict(rf, newdata = X_test)


# Calcular la precisión del modelo

accuracy <- confusionMatrix(y_pred, y_test)$overall["Accuracy"]

print(paste("Precisión del modelo Random Forest:", accuracy * 100))


# SVM

svm <- train(x = X_train, y = y_train, method = "svmRadial", trControl = trainControl(method = "cv", number = 5),  preProcess = c("center", "scale"))


# Obtener predicciones en el conjunto de prueba

y_pred <- predict(svm, newdata = X_test)


# Calcular la precisión del modelo

accuracy <- confusionMatrix(y_pred, y_test)$overall["Accuracy"]

print(paste("Precisión del modelo SVM:", accuracy * 100))


# Gráficos

importance <- varImp(dt)

importance_df <- as.data.frame(importance$importance)


# Ordenar el DataFrame por importancia
```

```
importance_df <- importance_df[order(importance_df$Overall), ]

# Graficar la importancia de las características
ggplot(importance_df, aes(x = Feature, y = Overall)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Feature", y = "Importance") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Anexo 2

```
---
title: "DASHBOARD ANÁLISIS DATA MÉDICA"
output:
  flexdashboard::flex_dashboard:
    orientation: columns
    vertical_layout: fill
runHORA: shiny
---

{r setup, include=FALSE}
library(flexdashboard) # creación de dashboards
library(shiny) # dashboard interactivo
library(dplyr) # manipulación de datos
library(plotly) # gráficos
library(rjson) # lectura de archivos json
library(highcharter) # gráfico mapas
library(lubriFECHA) # cálculo de edades


{r data}
Categorical.Variables <- c("Sexo", "Provincia", "CIE-10")
```

```
Numeric.Variables <- c("Todos", "2017", "2018", "2019", "2020", "2021", "2022", "2023")

data_pacientes <- read.csv("/home/leonardo/Documents/Tesis/pacientes_clean2.csv")

# Crear una nueva columna con los rangos de edad
data_pacientes <- data_pacientes %>%
  mutate(RangoEdad = case_when(
    edad >= 0 & edad <= 2 ~ "1. Lactantes (0-2)",
    edad >= 3 & edad <= 5 ~ "2. Pre-escolares (3-5)",
    edad >= 6 & edad <= 11 ~ "3. Escolares (6-11)",
    edad >= 12 & edad <= 17 ~ "4. Adolescentes (12-17)",
    TRUE ~ "Otro"
  ))

# Sidebar {data-width=200 .sidebar}

{r}
selectInput("categorical_variable", label = "Selección de variable categórica:", choices = Categorical.Variables)
selectInput("numeric_variable", label = "Selección de año:", choices = Numeric.Variables)


# Morbilidad
## Columna 1
### Mapa

{r}
ecuador <- fromJSON(file = "https://raw.githubusercontent.com/Rusersgroup/mapa_ecuador/master/ec-all.geo.json")

mapa <- data_pacientes %>%
```

```
  group_by(provincia) %>%
  summarise(count = n())

mapa <- data.frame(mapa)

highchart() %>%
  hc_tooltip(followPointer = FALSE) %>%
  hc_add_series_map(ecuador, mapa,
    name = "Pacientes",
    value = "count", joinBy = c("name", "provincia"),
    dataLabels = list(
      enabled = TRUE,
      format = "{point.properties.woe-name}"
    )
  )

## Columna 2
### Pacientes por provincia
{r}
bar_chart <- data_pacientes %>%
  group_by(provincia) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(provincia = factor(provincia, levels = unique(provincia)))

plot_ly(data = bar_chart, x = ~count, y = ~provincia, type = "bar", orientation = "h") %>%
  layout(
    xaxis = list(title = "Número de Pacientes"),
    yaxis = list(title = "Provincia")
  )
```

Diagnósticos

{r}

```
diag_chart <- data_pacientes %>%
  mutate(diagnostico = recode(diagnostico,
    "a" = "Meningitis",
    "b" = "Parálisis Cerebral",
    "c" = "Epilepsia",
    "d" = "Traumatismo Craneoencefálico",
    "e" = "Tumor Cerebral",
    "f" = "Dificultades de Aprendizaje",
    "g" = "TEA",
    "h" = "TDAH"
  )) %>%
  group_by(diagnostico) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(diagnostico = factor(diagnostico, levels = unique(diagnostico)))

plot_ly(data = diag_chart, x = ~count, y = ~diagnostico, type = "bar", orientation = "h") %>%
  layout(
    xaxis = list(title = "Número de Pacientes"),
    yaxis = list(title = "Diagnóstico")
  )
```

{r}

```
bar_chart <- data_pacientes %>%
  group_by(provincia) %>%
  summarise(count = n()) %>%
```

```
  arrange(desc(count)) %>%
  mutate(provincia = factor(provincia, levels = unique(provincia)))

plot_ly(data = bar_chart, x = ~count, y = ~provincia, type = "bar", orientation = "h") %>%
  layout(
    title = "Pacientes por ubicación",
    xaxis = list(title = "Número de Pacientes"),
    yaxis = list(title = "Provincia")
  )



# Información de pacientes
## Columna 1
### Porcentaje de pacientes por sexo
{r}
sexo_counts <- data_pacientes %>%
  count(genero)

plot_ly(data = sexo_counts, labels = ~genero, values = ~n, type = "pie", hole = 0.5) %>%
  layout(
    showlegend = FALSE
  )



### Pacientes por rango de edad
{r}
# Calcular el conteo de pacientes por cada rango de edad
datos_edad <- data_pacientes %>%
  group_by(edad) %>%
  count()
```

```
plot_ly(datos_edad,
  x = ~edad, y = ~n, type = "bar",
  hovertemplate = paste("Edad: %{x}", "<br>", "Número de Pacientes: %{y}")
) %>%
  layout(
    xaxis = list(title = "Edad"),
    yaxis = list(title = "Número de Pacientes")
  )



## Column 2
### Pirámide Poblacional
{r}
datos_piramide <- data_pacientes %>%
  group_by(RangoEdad, genero) %>%
  summarise(Pacientes = n())

datos_piramide$Cantidad <- ifelse(datos_piramide$genero == "Masculino",
datos_piramide$Pacientes, -datos_piramide$Pacientes)

datos_piramide <- datos_piramide %>%
  arrange(RangoEdad)

# Crear el gráfico de barras doble
plot_ly(datos_piramide,
  x = ~Cantidad, y = ~RangoEdad, type = "bar",
  orientation = "h", color = ~genero, colors = c("#1f77b4", "#ff7f0e"),
  hovertemplate = paste(
    "Rango de Edad: %{y}", "<br>",
    "Cantidad de Pacientes: %{x}"
```

```
  )
) %>%
  layout(
    xaxis = list(title = "Cantidad de Pacientes"),
    yaxis = list(title = "Rango de Edad"),
    barmode = "relative",
    bargap = 0.2
  )
```