**Proyecto 2: Predicción del GRD**

# Introducción

En el ámbito hospitalario, la gestión eficiente de diagnósticos y procedimientos médicos es crucial para optimizar los recursos, mejorar la calidad de atención y reducir los costos operativos. Una de las metodologías más utilizadas para la agrupación de diagnósticos y la gestión de recursos es el sistema de Grupos Relacionados por Diagnóstico (GRD), que permite clasificar a los pacientes según sus condiciones clínicas y los procedimientos médicos realizados. Este sistema facilita la previsión de costos y el análisis de resultados clínicos, al mismo tiempo que estandariza el tratamiento de casos complejos.

El presente estudio se enfoca en el análisis de un dataset proveniente del Hospital El Pino, que contiene información detallada de diagnósticos (D01-D35) y procedimientos (P01-P30) asociados a pacientes, junto con datos demográficos como la edad, el sexo, el grupo relacionado de diagnóstico (GRD), y una clasificación diagnóstica adicional que agregamos seleccionando los dos primeros dígitos de la columna GRD. El objetivo principal es aplicar técnicas de análisis de datos y aprendizaje automático para identificar patrones relevantes dentro de los diagnósticos y procedimientos, con el fin de mejorar la precisión en la clasificación y predicción de estos.

A través de este análisis, se busca no solo explorar el comportamiento de los datos mediante estadísticas descriptivas, sino también desarrollar un modelo predictivo que permita apoyar la toma de decisiones médicas, optimizando así la asignación de recursos y mejorando la calidad del servicio hospitalario.

# Referencias bibliográficas

Para abordar un problema similar al análisis de diagnósticos médicos y procedimientos hospitalarios mediante técnicas de aprendizaje automático y análisis estadístico, se pueden utilizar diversas referencias de estudios previos.

**1. Problemas similares en el análisis de datos médicos:**

* **Aprendizaje automático en la medicina**: Se han utilizado algoritmos de aprendizaje automático para la clasificación y predicción de enfermedades basadas en grandes conjuntos de datos médicos, como en el caso de diagnósticos y procedimientos.
  + *Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A Survey of Recent Advances on Deep Learning Techniques for Electronic Health Record (EHR) Analysis.* IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604. https://doi.org/10.1109/JBHI.2017.2767063
  + \*Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259
* **Predicción de diagnóstico utilizando modelos basados en GRD**:
  + \*Saberzadeh-Ardestani, B., Rezapour, A., Kavosi, Z., & Abbasi-Moghadam, M. A. (2020). Development and Evaluation of a Case-Mix Prediction Model for Hospital Resource Consumption Based on Diagnosis-Related Groups (DRG). *Frontiers in Public Health*, 8, 401. https://doi.org/10.3389/fpubh.2020.00401
* **Estudios relacionados a la predicción de readmisiones hospitalarias y clasificación de diagnósticos**:
  + \*Zhou, S. M., Fernandez-Gutierrez, F., Kennedy, J., Cooksey, R., Atkinson, M., Denaxas, S., ... & Williams, R. (2016). Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Type 2 Diabetes Mellitus. *Journal of Biomedical Informatics*, 64, 37-43. https://doi.org/10.1016/j.jbi.2016.09.013

**2. Referencias para la generación de estadísticas descriptivas:**

* **Análisis estadístico en salud pública**:
  + \*Rosner, B. (2015). Fundamentals of Biostatistics. *Cengage Learning*. Este libro cubre los principios básicos del análisis estadístico en el campo de la biostatística y proporciona ejemplos detallados de cómo calcular estadísticas descriptivas, frecuencias, y análisis de datos biomédicos.
* **Guía para la generación de estadísticas descriptivas y análisis exploratorio**:
  + \*Altman, D. G. (1990). Practical Statistics for Medical Research. *Chapman & Hall/CRC*. Este texto es una referencia esencial para aplicar análisis descriptivos y exploratorios en datos médicos, incluyendo el uso de distribuciones de frecuencia y gráficos estadísticos.

# Objetivo del Estudio

El objetivo principal de este estudio es analizar y modelar los diagnósticos y procedimientos médicos del Hospital El Pino mediante el uso de técnicas de análisis de datos y aprendizaje automático. Se pretende identificar patrones en los datos que permitan mejorar la clasificación de diagnósticos y procedimientos, optimizar la gestión hospitalaria y apoyar la toma de decisiones clínicas. Específicamente, se busca desarrollar un modelo predictivo basado en Grupos Relacionados de Diagnóstico (GRD) y otras variables relevantes, que ayude a prever el comportamiento de los pacientes en términos de tratamientos, costos y resultados clínicos, contribuyendo así a la mejora de la calidad de la atención médica y la eficiencia operativa del hospital.

# Metodología

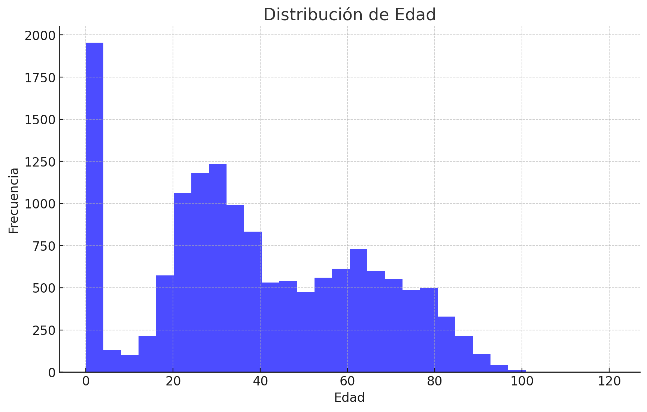
El estudio se llevará a cabo utilizando un enfoque basado en análisis de datos y técnicas avanzadas de aprendizaje automático. En primer lugar, se realizará una limpieza y exploración del dataset para asegurar la calidad de los datos, identificando posibles valores atípicos, datos faltantes y características relevantes. Posteriormente, se implementará un modelo predictivo utilizando Redes Neuronales Recurrentes (RNN), dado que estas son especialmente adecuadas para tratar datos secuenciales como los presentes en los registros hospitalarios. Las RNN permitirán capturar dependencias entre diagnósticos y procedimientos, mejorando la capacidad predictiva del modelo. Adicionalmente, se utilizarán técnicas de ingeniería de características para optimizar el rendimiento del modelo, y se evaluarán métricas como la precisión, el recall y la métrica F1 para seleccionar el modelo más adecuado. La implementación se realizará en un entorno de aprendizaje automático, asegurando la replicabilidad y escalabilidad del modelo para su uso en futuros estudios hospitalarios.

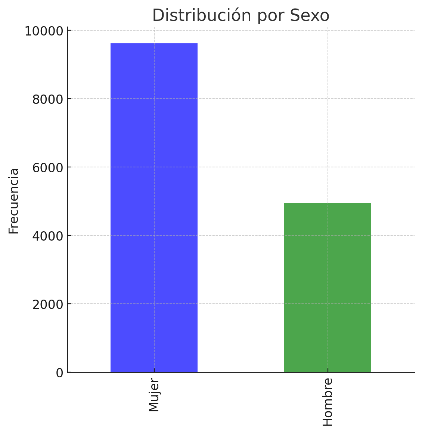
# Experimentos

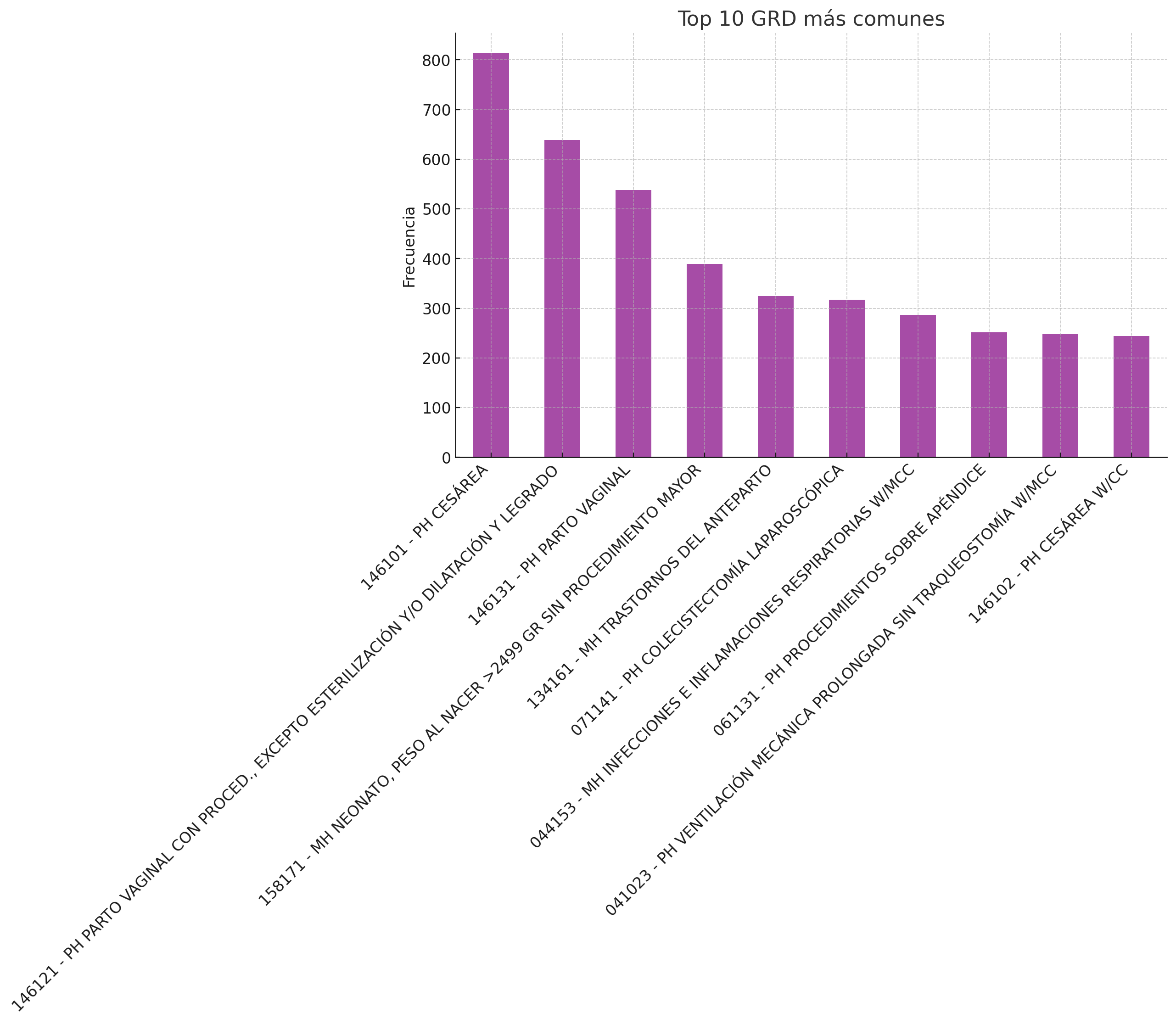
* 1. Análisis de datos
     1. Calidad de datos

Los datos nulos solo corresponden a diagnósticos o procedimientos de relleno por lo cual no se consideran nulos.

* + 1. Estadísticas descriptivas
    2. Gráficos







* 1. Resultados preliminares

Considerando el histograma por clasificación general:

A graph of a number of bars

Description automatically generated

Hay algunas con muy pocos datos, y serán candidatas a correr el modelo con todas las categorías comparado con la remoción de aquellas minoritarias y verificar cuál modelo da mejores resultados. Lo mismo se hará en atención a la distribución por edad, pues para menores (cerca del primer año de vida) hay muchos casos y potencialmente los diagnósticos de dicho segmento etáreo serán diferentes, por lo cual tal vez debamos también dividir la información.