



Universidad Nacional de Entre Ríos

Facultad de Ingeniería

Lic. en Bioinformática

Trabajo Integrador Final

Bases de Datos

“Análisis de Hipertensión Arterial”

Estudiante: Ríos Matías

Diciembre 2024

Indice

Indice.....	2
Introducción.....	3
Objetivo general.....	3
Requisitos mínimos.....	3
Metodología.....	4
Desarrollo.....	5
Propósito general.....	5
Objetivo.....	5
BioLAB.....	10
Identificación de entidades y relaciones.....	10
Entidades.....	10
Relaciones.....	12
Diagrama Entidad-Relación (DER).....	13
Componentes principales de un DER.....	13
Diagrama de tablas.....	16
Diccionario de datos.....	18
Tabla: Pacientes.....	18
Tabla: Examen_clinico.....	18
Tabla: Prueba_sangre.....	19
Tabla: Perfil_lipidico.....	19
Tabla: Marcadores_nutricionales.....	20
Tabla: Marcadores_inflamatorios.....	20
Tabla: Mediciones_fisicas.....	20
Consultas SQL varias.....	23
Árboles de consulta.....	29
Optimización heurística.....	29
Resumen de las reglas heurísticas para la optimización algebraica.....	29
Consulta SQL con optimización heurística.....	30
Implementación de Interfaz Grafica.....	35
Herramientas utilizadas:.....	35
Estructura de la Interfaz.....	36
Ventana Principal.....	36
Botón “Buscar Pacientes”.....	37
Módulo de Estadísticos.....	38
Resultados obtenidos.....	42
Conclusión.....	42
Bibliografía.....	43

Introducción

Este trabajo práctico final tiene como objetivo afianzar y evaluar los contenidos dictados durante la materia. Partiendo de una DB de interés de la carrera buscada en bibliografía por ustedes y acordada con la cátedra, se van a repasar todos los temas dictados en este curso.

Objetivo general

Buscar una base de datos aplicada a bioinformática en la literatura, o una nueva propuesta por el alumno acorde a la dificultad del curso, y a partir de ella realizar un trabajo de investigación y desarrollo en el que demuestre haber adquirido los conocimientos de la asignatura.

Requisitos mínimos

- ☐ Breve descripción de el uso de la DB
- ☐ Diagrama de entidad relación de la base de datos
- ☐ Diccionario de datos
- ☐ Diagrama de tablas
- ☐ Implementación de la base de datos en lenguaje SQL. Si no se tiene el DDL, crearlo (motor a elección, Postgres, MySQL, Oracle, etc)
- ☐ En caso de haber seleccionado una DB propia que no tenga datos, insertar datos suficientes (pueden ser ficticios) a todas las tablas de manera de poder realizar el resto del TP.
- ☐ Plantear al menos 3 consultas complejas en las que se utilicen operaciones avanzadas (Join, exist, in, etc.)
- ☐ Crear una consulta que utilice tres tablas, contenga una condición de igualdad y una condición de rango (>, >=, <, <=, between). Escribir el árbol de ejecución de la consulta en álgebra relacional y luego optimizarlo. Luego escribir la nueva consulta SQL en base al árbol optimizado.
- ☐ Utilizar una interfaz de acceso a datos para generar algún valor agregado. El valor agregado puede obtenerse generando gráficas, realizando test estadísticos o cualquier posibilidad que sea interesante y no pueda resolverse usando solamente SQL. Es probable que requieran conocimientos adquiridos en otras materias. La misma puede ser implementada en C, R, Python o en el lenguaje de preferencia.

Metodología

Se deberá presentar un informe escrito donde se detallen cómo se cumplen todos los requisitos mínimos. Se hará una defensa oral de este informe, donde deberán exponer su trabajo, mostrando las tareas y resultados obtenidos. Toda implementación deberá ser funcional de forma tal de poder mostrarlo y defenderlo en la instancia de evaluación

Desarrollo

Propósito general

El propósito de este trabajo práctico final es consolidar y evaluar los conocimientos adquiridos a lo largo de la asignatura, mediante la aplicación de técnicas y conceptos fundamentales en el análisis de datos biológicos. El proyecto implica la búsqueda y selección de una base de datos relacionada con el campo de la bioinformática, seguida de un análisis exhaustivo que permita demostrar la capacidad del estudiante para abordar un problema real. Para ello, elegí un dataset que se utiliza en estudios de salud pública, con un enfoque específico en la prevención de enfermedades mediante el monitoreo de marcadores bioquímicos y factores de estilo de vida.

Este trabajo se centra en el diseño, implementación y análisis de una base de datos que permita realizar un seguimiento integral de pacientes, combinando datos clínicos, bioquímicos y de estilo de vida para evaluar su riesgo de desarrollar enfermedades crónicas, como la hipertensión o la diabetes. A partir de estos datos, se pretende ilustrar cómo el uso adecuado de técnicas de bases de datos y la integración de información biomédica pueden facilitar la toma de decisiones en el ámbito de la salud.

Objetivo

El objetivo principal es seleccionar una base de datos aplicada a nuestra carrera, y a partir de ella desarrollar un trabajo de investigación que permita demostrar el dominio de los contenidos dictados durante la asignatura. Esto incluirá la creación de un modelo relacional que refleje las relaciones entre diferentes tipos de datos, así como el análisis de dichas relaciones para extraer conclusiones relevantes desde el punto de vista de la salud pública.

El trabajo busca no solo evaluar el conocimiento técnico adquirido en cuanto a la creación y manipulación de bases de datos, sino también la capacidad de interpretación y análisis de los datos obtenidos, vinculando los resultados con la realidad clínica de los pacientes.

En el desarrollo de este trabajo, me basé en un dataset público obtenido de la plataforma Kaggle, titulado Hipertensión Arterial México, disponible en [este enlace](#). Este conjunto de datos fue utilizado para estructurar el modelo entidad-relación descrito posteriormente.

El dataset en cuestión está compuesto por varias columnas que representan diferentes características o mediciones de los pacientes. Cada columna del dataset se detalla a continuación:

Nombre de Columna	Descripción
Folio_ID	Identificador único para cada entrada en el conjunto de datos
Sexo (Gender)	Indica el género del paciente (1 para masculino, 2 para femenino)
Edad (Age)	Representa la edad del paciente en años
Concentración de Hemoglobina (Hemoglobin Concentration)	Mide la concentración de hemoglobina en la sangre del paciente
Temperatura Ambiente (Ambient Temperature)	Registra la temperatura ambiente durante la medición
Valor de Ácido Úrico (Uric Acid Value)	Indica el nivel de ácido úrico en la sangre del paciente
Valor de Albúmina (Albumin Value)	Representa la concentración de albúmina en la sangre
Valor de Colesterol HDL (HDL Cholesterol Value)	Indica el nivel de colesterol de lipoproteínas de alta densidad (HDL)
Valor de Colesterol LDL (LDL Cholesterol Value)	Representa el nivel de colesterol de lipoproteínas de baja densidad (LDL)
Valor de Colesterol Total (Total Cholesterol Value)	Indica el nivel total de colesterol en la sangre del paciente

Valor de Creatinina (Creatinine Value)	Representa la concentración de creatinina en la sangre
Resultado de Glucosa (Glucose Result)	Indica el resultado de la medición del nivel de glucosa
Valor de Insulina (Insulin Value)	Representa la concentración de insulina en la sangre
Valor de Proteína C Reactiva (C-Reactive Protein Value)	Indica el nivel de proteína C reactiva, un marcador de inflamación
Valor de Triglicéridos (Triglycerides Value)	Representa el nivel de triglicéridos en la sangre
Resultado de Glucosa Promedio (Average Glucose Result)	Indica el nivel promedio de glucosa
Valor de Hemoglobina Glucosilada (Glycosylated Hemoglobin Value)	Representa la concentración de hemoglobina glucosilada
Valor de Ferritina (Ferritin Value)	Indica el nivel de ferritina, una proteína que almacena hierro
Valor de Folato (Folate Value)	Representa el nivel de folato, una vitamina B
Valor de Homocisteína (Homocysteine Value)	Indica el nivel de homocisteína en la sangre
Valor de Transferrina (Transferrin Value)	Representa la concentración de transferrina, una proteína que transporta hierro
Valor de Vitamina B12 (Vitamin B12 Value)	Indica el nivel de vitamina B12 en la sangre

Valor de Vitamina D (Vitamin D Value)	Representa el nivel de vitamina D en la sangre
Peso (Weight)	Representa el peso del paciente
Estatura (Height)	Indica la altura del paciente
Medida de Cintura (Waist Measurement)	Representa la circunferencia de la cintura del paciente
Segunda Medición de Peso (Second Weight Measurement)	Representa una segunda medición de peso, si está disponible
Segunda Medición de Estatura (Second Height Measurement)	Representa una segunda medición de altura, si está disponible
Distancia Rodilla-Talón (Knee-Heel Distance)	Representa la distancia desde la rodilla hasta el talón
Circunferencia de la Pantorrilla (Calf Circumference)	Indica la circunferencia de la pantorrilla
Segunda Medición de Cintura (Second Waist Measurement)	Representa una segunda medición de la circunferencia de la cintura, si está disponible
Tensión Arterial (Blood Pressure)	Indica la presión arterial del paciente
Sueño en Horas (Sleep in Hours)	Representa el número de horas de sueño
Masa Corporal (Body Mass)	Indica el índice de masa corporal (IMC) del paciente

Actividad Total (Total Activity)	Representa la actividad física total del paciente
Riesgo de Hipertensión (Hypertension Risk)	La variable objetivo, indica si el paciente está en riesgo de desarrollar hipertensión (1 para en riesgo, 0 para no en riesgo)

Sin embargo, cabe aclarar que solo extraje los encabezados de las diferentes tablas que componen el dataset original. A partir de estos encabezados, inventé los datos ficticios correspondientes para cada entidad, con el fin de ilustrar cómo podrían utilizarse estos valores en un contexto bioinformático. Esta modificación me permitió personalizar los datos de acuerdo con los objetivos del trabajo integrador, mientras mantenía la estructura general y las variables relevantes del conjunto de datos original.

La base de datos creada para este trabajo está orientada al análisis de factores de riesgo en pacientes que podrían estar en riesgo de desarrollar enfermedades crónicas, tales como la hipertensión arterial, la diabetes, y otras condiciones relacionadas con el estilo de vida y marcadores bioquímicos. Esta base de datos incluye un conjunto de registros relacionados con mediciones clínicas, resultados bioquímicos, y datos sobre el estilo de vida de los pacientes. Además, se tienen en cuenta datos demográficos que permiten realizar análisis estratificados por grupos de edad y sexo.

BioLAB

Esta base de datos está diseñada para almacenar y gestionar información relacionada con pacientes y sus exámenes clínicos. Facilita la recopilación, análisis y consulta de datos médicos, lo que permite a los profesionales de la salud tomar decisiones informadas basadas en el historial médico y los resultados de pruebas de laboratorio.

Se seleccionó [PostgreSQL](#) como sistema de gestión de bases de datos debido a su fiabilidad, capacidad para manejar grandes volúmenes de datos, soporte para integridad referencial y además porque fue el gestor de bases de datos que nos proporcionó la catedra durante el cursado.

[DBeaver](#) fue elegido como herramienta de administración por su interfaz amigable, compatibilidad multiplataforma y funciones avanzadas como diseño de diagramas y manejo de consultas complejas. Esto facilita el desarrollo, mantenimiento y gestión eficiente de la base de datos.

Ambas herramientas combinan robustez y facilidad de uso, asegurando un entorno confiable para almacenar y manejar información clínica.

Toda la documentación y scripts relacionados se encuentran en el [repositorio de GitHub](#)

Identificación de entidades y relaciones

Entidades

Luego de un arduo analisis del dataset, y buscando las entidades y relaciones que podrian incluirse en esta base de datos, expongo la siguiente tabla con las entidades que ocuparé con sus respectivos atributos:

Entidad	Atributos
Paciente	<ul style="list-style-type: none">• folio_id (PK)• sexo• edad• peso• estatura• medida_cintura• segundamedicion_peso• segundamedicion_estatura• segundamedicion_cintura• tension_arterial• sueno_horas

	<ul style="list-style-type: none">• masa_corporal• actividad_total
Examen_clinico	<ul style="list-style-type: none">• examen_id (PK)• folio_id (FK a Paciente)• fecha_examen• temperatura_ambiente• riesgo_hipertension
Prueba_sangre	<ul style="list-style-type: none">• prueba_id (PK)• examen_id (FK a Examen_clinico)• concentracion_hemoglobina• valor_acido_urico• valor_albumina• valor_creatina• valor_glucosa• valor_insulina• valor_trigliceridos• valor_hemoglobina_glucosilada
Perfil_lipidico	<ul style="list-style-type: none">• perfil_id (PK)• examen_id (FK a Examen_clinico)• colesterol_hdl• colesterol_ldl• colesterol_total
Marcadores_nutricionales	<ul style="list-style-type: none">• marcador_id (PK)• examen_id (FK a Examen_clinico)• valor_ferritina• valor_folato• valor_vitamina_b12• valor_vitamina_d
Marcadores_inflamatorios	<ul style="list-style-type: none">• marcador_inflam_id (PK)• examen_id (FK a Examen_clinico)• valor_proteinac_reactiva• valor_homocisteina• valor_transferrina
Mediciones_fisicas	<ul style="list-style-type: none">• medicion_id (PK)• folio_id (FK a Paciente)• distancia_rodilla_talon• circunferencia_pantorrilla

Relaciones

1) Paciente ↔ Examen_clinico:

- Relación: Un paciente puede tener varios exámenes clínicos (1:N)
 - Llave foránea: Examen_clinico.folio_id referencia a Paciente.folio_id
- 2) Examen_clinico ↔ Prueba_sangre:
- Relación: Un examen clínico puede estar asociado con una o más pruebas de sangre (1:N).
 - Llave foránea: Prueba_sangre.examen_id referencia Examen_clinico.examen_id.
- 3) Examen_clinico ↔ Perfil_lipidico
- Relación: Un examen clínico puede estar asociado con un perfil lipídico (1:1).
 - Llave foránea: Perfil_lipidico.examen_id referencia a Examen_clinico.examen_id.
- 4) Examen_clinico ↔ Marcadores_nutricionales
- Relación: Un examen clínico puede estar asociado con un conjunto de marcadores nutricionales (1:1).
 - Llave foránea: Marcadores_nutricionales.examen_id referencia a Examen_clinico.examen_id.
- 5) Examen_clinico ↔ Marcadores_inflamatorios
- Relación: Un examen clínico puede estar asociado con marcadores inflamatorios (1:1).
 - Llave foránea: Marcadores_inflamatorios.examen_id referencia a Examen_clinico.examen_id.
- 6) Paciente ↔ Mediciones_fisicas
- Relación: Un paciente puede tener mediciones físicas específicas (1:1).
 - Llave foránea: Mediciones_fisicas.folio_id referencia a Paciente.folio_id.

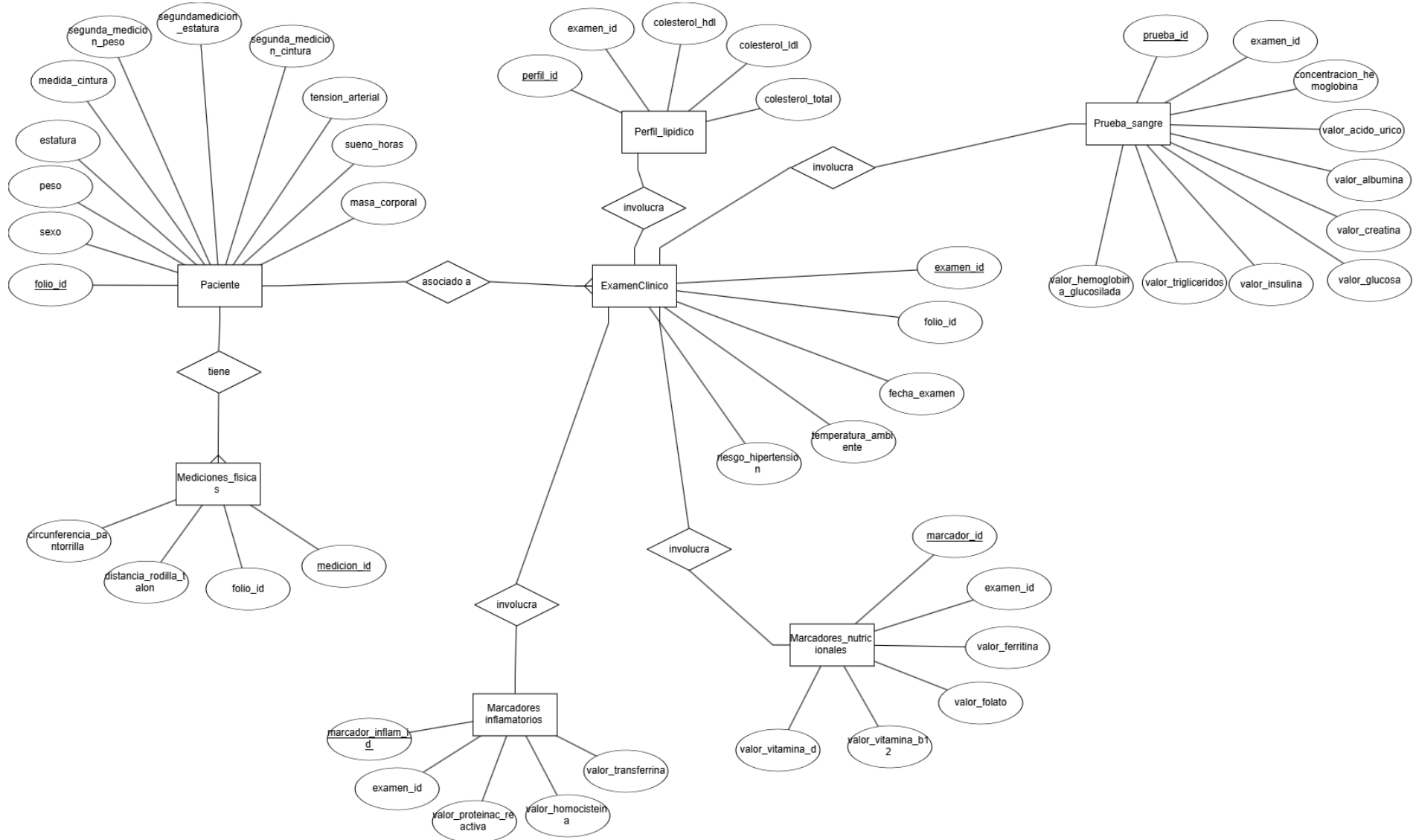
Este diseño permite que la base de datos sea flexible y escalable, permitiendo agregar diferentes exámenes y pruebas para cada paciente a lo largo del tiempo sin duplicar datos redundantes. Además, se mantiene la integridad referencial, ya que cualquier eliminación de un paciente o un examen clínico eliminaría automáticamente los registros correspondientes en las tablas relacionadas.

Diagrama Entidad-Relación (DER)

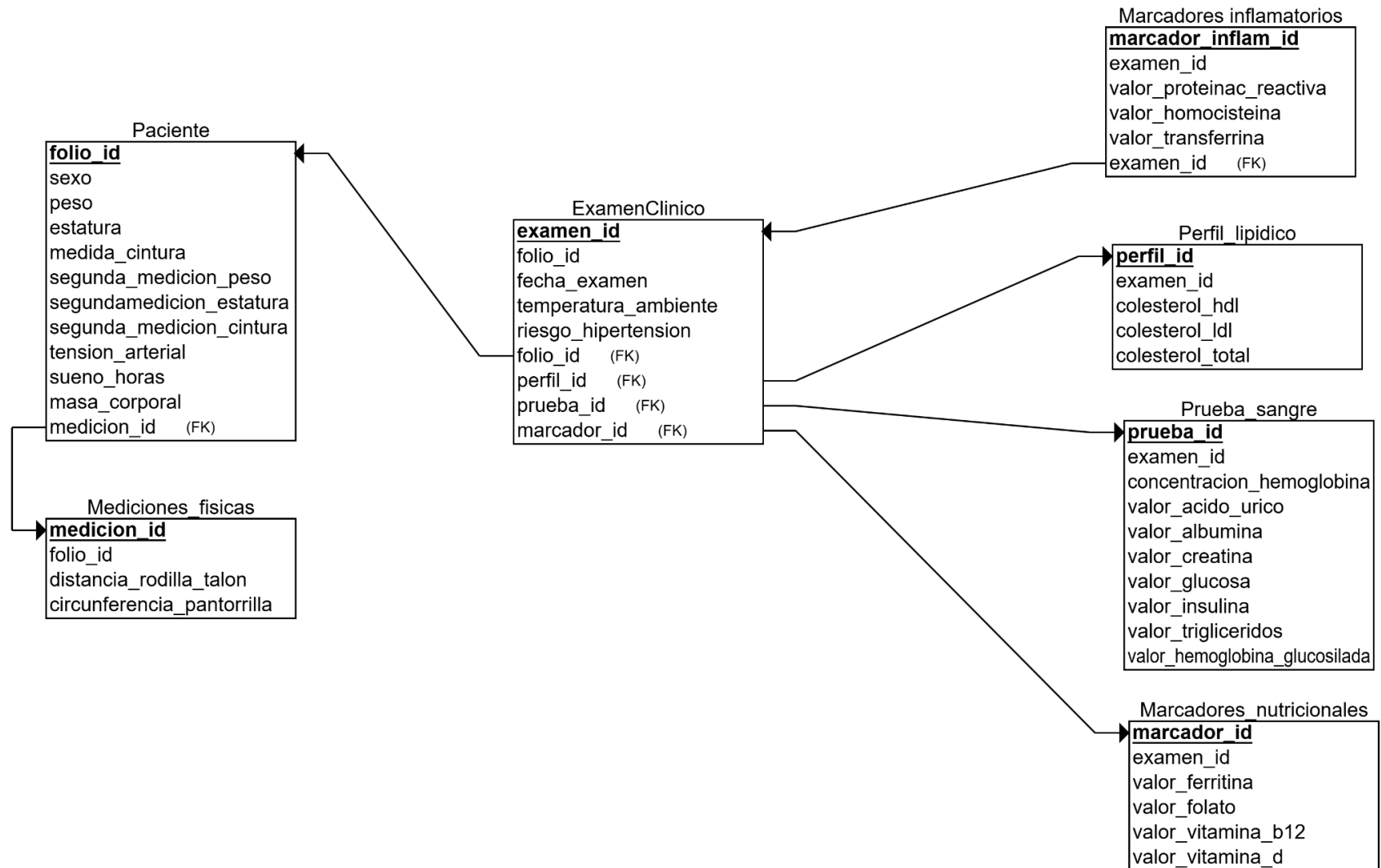
El DER (Diagrama Entidad-Relación) es una herramienta visual utilizada para modelar y representar la estructura de una base de datos a nivel conceptual. Describe las entidades involucradas, sus atributos y las relaciones entre ellas.

Componentes principales de un DER

- 1) Entidades → Representan "cosas" u "objetos" del mundo real que tienen relevancia para el sistema.
 - Pueden ser:
 - Entidades fuertes: Tienen existencia independiente y una clave primaria propia.
 - Entidades debiles: Dependen de otra entidad para existir, y su clave primaria incluye la clave de una entidad fuerte.
- 2) Atributos → Describen las propiedades o características de una entidad o relación.
 - Tipos de atributos:
 - Simples: No se pueden descomponer (por ejemplo, nombre).
 - Compuestos: Pueden descomponerse en subatributos (por ejemplo, dirección: calle, ciudad, código postal).
 - Derivados: Calculados a partir de otros atributos (por ejemplo, edad, a partir de la fecha de nacimiento).
 - Multievaluados: Pueden tener varios valores (por ejemplo, teléfonos).
- 3) Relaciones → Representan asociaciones entre entidades.
 - Tienen un nombre que indica el tipo de relación (por ejemplo, "tiene", "realiza").
 - Pueden incluir atributos que describan la relación.
- 4) Cardinalidad → Indica cuántas instancias de una entidad pueden estar asociadas con una instancia de otra.
 - Tipos comunes:
 - 1:1 (uno a uno): Una instancia de A se asocia con una instancia de B.
 - 1:N (uno a muchos): Una instancia de A se asocia con muchas instancias de B.
 - N:M (muchos a muchos): Muchas instancias de A se asocian con muchas de B.
- 5) Clave primaria (PK) → Atributo(s) único(s) que identifica(n) de manera exclusiva cada instancia de una entidad.
- 6) Clave foránea (FK) → Atributo(s) que referencia(n) la clave primaria de otra entidad, estableciendo una relación.



Esquema de tablas



Diccionario de datos

Tabla: Pacientes				
Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
folio_id	VARCHAR	15	Identificador único del paciente.	PK , No nulo
sexo	VARCHAR	10	Sexo del paciente (e.g., "Masculino", "Femenino").	
edad	INTEGER	N/A	Edad del paciente en años.	No negativo
peso	NUMERIC	(5,2)	Peso del paciente en kilogramos.	No negativo
estatura	NUMERIC	(5,2)	Estatura del paciente en metros.	No negativo
medida_cintura	NUMERIC	(5,2)	Circunferencia de cintura en centímetros.	No negativo
segundamedicion_peso	NUMERIC	(5,2)	Segunda medición de peso.	Opcional
segundamedicion_estatura	NUMERIC	(5,2)	Segunda medición de estatura.	Opcional
segundamedicion_cintura	NUMERIC	(5,2)	Segunda medición de circunferencia de cintura.	Opcional
tension_arterial	VARCHAR	15	Medición de la presión arterial (e.g., "120/80").	
sueno_horas	NUMERIC	(3,1)	Promedio de horas de sueño por día.	No negativo
masa_corporal	NUMERIC	(5,2)	Índice de masa corporal (calculado).	
actividad_total	NUMERIC	(5,2)	Nivel de actividad física (e.g., número de pasos diarios).	

Tabla: Examen_clinico				
Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
examen_id	SERIAL	N/A	Identificador único del examen clínico.	PK , No nulo
folio_id	VARCHAR	15	Referencia al paciente asociado.	FK → Paciente

fecha_examen	DATE	N/A	Fecha en la que se realizó el examen.	No nulo
temperatura_ambiente	NUMERIC	(4,1)	Temperatura ambiental al momento del examen (en °C).	Opcional
riesgo_hipertension	BOOLEAN	N/A	Indica si el paciente tiene riesgo de hipertensión (true/false).	Por defecto false

Tabla: Prueba_sangre				
Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
prueba_id	SERIAL	N/A	Identificador único de la prueba de sangre.	PK , No nulo
examen_id	INT	N/A	Referencia al examen clínico asociado.	FK → Examen_clinico
concentracion_hemoglobina	NUMERIC	(5,2)	Concentración de hemoglobina en g/dL.	No negativo
valor_acido_urico	NUMERIC	(5,2)	Nivel de ácido úrico en mg/dL.	No negativo
valor_albumina	NUMERIC	(5,2)	Nivel de albúmina en g/dL.	No negativo
valor_creatina	NUMERIC	(5,2)	Nivel de creatinina en mg/dL.	No negativo
valor_glucosa	NUMERIC	(5,2)	Nivel de glucosa en mg/dL.	No negativo
valor_insulina	NUMERIC	(5,2)	Nivel de insulina en µU/mL.	No negativo
valor_trigliceridos	NUMERIC	(5,2)	Nivel de triglicéridos en mg/dL.	No negativo
valor_hemoglobina_glucosilada	NUMERIC	(5,2)	Nivel de hemoglobina glucosilada (%).	No negativo

Tabla: Perfil_lipidico				
Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
perfil_id	SERIAL	N/A	Identificador único del perfil lipídico.	PK , No nulo
examen_id	INT	N/A	Referencia al examen clínico asociado.	FK → Examen_clinico
colesterol_hdl	NUMERIC	(5,2)	Nivel de colesterol HDL en mg/dL.	No negativo

colesterol_ldl	NUMERIC	(5,2)	Nivel de colesterol LDL en mg/dL.	No negativo
colesterol_total	NUMERIC	(5,2)	Nivel de colesterol total en mg/dL.	No negativo

Tabla: Marcadores_nutricionales

Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
marcador_id	SERIAL	N/A	Identificador único de los marcadores nutricionales.	PK , No nulo
examen_id	INT	N/A	Referencia al examen clínico asociado.	FK → Examen_clinico
valor_ferritina	NUMERIC	(5,2)	Nivel de ferritina en ng/mL.	No negativo
valor_folato	NUMERIC	(5,2)	Nivel de folato en ng/mL.	No negativo
valor_vitamina_b12	NUMERIC	(5,2)	Nivel de vitamina B12 en pg/mL.	No negativo
valor_vitamina_d	NUMERIC	(5,2)	Nivel de vitamina D en ng/mL.	No negativo

Tabla: Marcadores_inflamatorios

Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
marcador_inflam_id	SERIAL	N/A	Identificador único del marcador inflamatorio.	PK, No nulo
examen_id	INT	N/A	Identificador del examen clínico asociado.	FK → Examen_clinico
valor_proteinac_reactiva	NUMERIC	(5,2)	Concentración de proteína C reactiva en sangre.	Opcional
valor_homocisteina	NUMERIC	(5,2)	Nivel de homocisteína en sangre.	Opcional
valor_transferrina	NUMERIC	(5,2)	Concentración de transferrina en sangre.	Opcional

Tabla: Mediciones_fisicas				
Nombre del campo	Tipo de dato	Tamaño	Descripción	Restricciones
medicion_id	SERIAL	N/A	Identificador único de la medición física.	PK, No nulo
folio_id	VARCHAR	15	Identificador del paciente asociado.	FK → Paciente
distancia_rodilla_talon	NUMERIC	(5,2)	Medida de la distancia rodilla-talón (en cm).	Opcional
circunferencia_pantorrilla	NUMERIC	(5,2)	Circunferencia de la pantorrilla (en cm).	Opcional

Consultas SQL varias

1) Promedio de edad y peso por género

```
SELECT sexo, AVG(edad) AS promedio_edad, AVG(peso) AS  
promedio_peso  
FROM Paciente GROUP BY sexo;
```

	sexo character varying (10)	promedio_edad numeric	promedio_peso numeric
1	M	48.8328804347826087	74.6564198369565217
2	F	49.4941099476439791	74.6307329842931937

2) Pacientes con medidas de cintura incrementadas en la segunda medición (solo los primeros 10)

```
SELECT folio_id  
FROM Paciente  
WHERE segundamedicion_cintura > medida_cintura;  
LIMIT 10;
```

	folio_id [PK] character varying (15)
1	1001
2	1005
3	1008
4	1010
5	1012
6	1015
7	1017
8	1020
9	1022
10	1023

3) Promedio de masa corporal de pacientes con riesgo de hipertensión '

```
SELECT AVG(p.masa_corporal) AS promedio_masa_corporal
FROM Paciente p
JOIN Examen_clinico ec ON p.folio_id = ec.folio_id
WHERE ec.riesgo_hipertension = '1';
```

	promedio_masa_corporal numeric
1	26.3289731219848380

4) Los primeros 5 exámenes con temperaturas ambiente superiores al promedio

```
SELECT examen_id, temperatura_ambiente
FROM Examen_clinico
WHERE temperatura_ambiente > (SELECT AVG(temperatura_ambiente)
FROM Examen_clinico)
LIMIT 5;
```

	examen_id [PK] integer	temperatura_ambiente numeric (4,1)
1	1	28.7
2	2	27.1
3	3	27.9
4	6	28.5
5	9	26.9

5) Relación de colesterol total y LDL en pacientes con alto riesgo de hipertensión

```
SELECT p.folio_id, pl.colesterol_total, pl.colesterol_ldl
FROM Paciente p
JOIN Examen_clinico ec ON p.folio_id = ec.folio_id
JOIN Perfil_lipidico pl ON ec.examen_id = pl.examen_id
WHERE ec.riesgo_hipertension = '1'
LIMIT 10;
```

	folio_id character varying (15)	colesterol_total numeric (5,2)	colesterol_ldl numeric (5,2)
1	1000	171.26	72.46
2	1005	162.39	116.60
3	1006	162.12	103.19
4	1008	185.29	128.21
5	1011	165.38	96.33
6	1017	195.37	88.30
7	1018	197.46	96.96
8	1020	154.46	81.32
9	1027	150.61	70.04
10	1031	166.94	78.45

6) Usando vistas: valores promedio de ferritina y folato en pacientes mayores de 50 años

```

- Crear una vista para los pacientes mayores de 50 años
CREATE VIEW Pacientes_Mayores50 AS
SELECT folio_id, edad
FROM Paciente
WHERE edad > 50;

-- Crear una vista que combine los datos de marcadores
nutricionales y exámenes clínicos
CREATE VIEW Marcadores_Completos AS
SELECT mn.valor_ferritina, mn.valor_folato, ec.folio_id
FROM Marcadores_nutricionales mn
JOIN Examen_clinico ec ON mn.examen_id = ec.examen_id;

-- Calcular los promedios utilizando las vistas
SELECT AVG(mc.valor_ferritina) AS promedio_ferritina,
       AVG(mc.valor_folato) AS promedio_folato
FROM Marcadores_Completos mc
JOIN Pacientes_Mayores50 pm ON mc.folio_id = pm.folio_id;

```

	promedio_ferritina numeric	promedio_folato numeric
1	165.5929825783972125	10.2614181184668990

7) Promedio de colesterol y triglicéridos en pacientes con IMC superior a 30

```

SELECT AVG(pl.cholesterol_total) AS promedio_cholesterol,
AVG(ps.valor_trigliceridos) AS promedio_trigliceridos
FROM Paciente p
JOIN Examen_clinico ec ON p.folio_id = ec.folio_id
JOIN Perfil_lipidico pl ON ec.examen_id = pl.examen_id
JOIN Prueba_sangre ps ON ec.examen_id = ps.examen_id
WHERE p.masa_corporal > 30;

```

	promedio_cholesterol numeric	promedio_trigliceridos numeric
1	175.0865062500000000	100.9980000000000000

- 8) Obtener los pacientes cuyo colesterol total no supera los 190 mg/dL y cuyo colesterol HDL es mayor a 50 mg/dL, junto con la información de la fecha del examen y el riesgo de hipertensión.

```
SELECT P.folio_id, E.fecha_examen, PL.colesterol_total,
PL.colesterol_hdl, E.riesgo_hipertension
FROM Paciente AS P
JOIN Examen_clinico AS E ON P.folio_id = E.folio_id
JOIN Perfil_lipidico AS PL ON E.examen_id = PL.examen_id
WHERE PL.colesterol_total < 190 AND PL.colesterol_hdl > 50;
```

	folio_id character varying (15)	fecha_examen date	colesterol_total numeric (5,2)	colesterol_hdl numeric (5,2)	riesgo_hipertension boolean
1	1000	2021-07-20	159.29	55.38	true
2	1002	2023-09-19	170.60	57.27	false
3	1005	2021-06-05	172.36	54.54	false
4	1014	2023-11-27	176.98	59.86	false
5	1017	2022-06-10	158.51	56.87	false
6	1018	2022-03-14	152.55	59.86	false
7	1020	2022-07-15	161.95	59.34	true
8	1021	2021-02-06	182.57	53.56	true
9	1022	2020-07-21	163.35	53.34	false
10	1024	2023-07-17	156.26	54.93	false
11	1026	2020-12-04	180.11	51.09	true
12	1029	2022-12-01	171.47	52.94	false

- 9) Listar los pacientes con valores de proteína C reactiva mayores a 3 mg/L o niveles de homocisteína superiores a 15 µmol/L,

```
SELECT P.folio_id, E.fecha_examen, MI.valor_proteinac_reactiva,
MI.valor_homocisteina
FROM Paciente AS P
JOIN Examen_clinico AS E ON P.folio_id = E.folio_id
JOIN Marcadores_inflamatorios AS MI ON E.examen_id =
MI.examen_id
WHERE MI.valor_proteinac_reactiva > 3 OR MI.valor_homocisteina >
15;
```

	folio_id character varying (15)	fecha_examen date	valor_proteinac_reactiva numeric (5,2)	valor_homocisteina numeric (5,2)
1	1001	2023-09-07	6.54	5.04
2	1002	2023-09-19	8.00	8.59
3	1004	2021-08-20	8.48	13.58
4	1005	2021-06-05	9.42	4.81
5	1006	2020-01-25	8.79	10.77
6	1007	2022-01-05	4.07	6.43
7	1011	2020-07-28	4.59	14.50
8	1012	2021-04-14	3.68	13.98
9	1013	2020-06-20	8.83	6.44
10	1014	2023-11-27	8.89	12.88
11	1015	2021-03-11	6.89	10.11
12	1016	2023-02-12	8.37	10.04
13	1017	2022-06-10	8.43	14.91
14	1019	2023-10-25	4.33	12.95
15	1024	2023-07-17	9.73	12.34

- 10) Contar cuantos pacientes con actividad física total menor a 150 y cuyos niveles de ferritina o vitamina D están por debajo de los valores recomendados

```
SELECT count (*) as cantidad_valores_bajos
FROM Paciente AS P
JOIN Examen_clinico AS E ON P.folio_id = E.folio_id
JOIN Marcadores_inflamatorios AS MI ON E.examen_id =
MI.examen_id
WHERE MI.valor_proteinac_reactiva > 3 OR MI.valor_homocisteina >
15;
```

cantidad_valores_bajos bigint	2112
----------------------------------	------

- 11) Listar los pacientes que tienen niveles de glucosa en sangre superiores a 100 mg/dL y niveles de insulina mayores a 20 µU/mL. Mostrar folio, edad, el valor de glucosa, el valor de insulina y el riesgo de hipertensión

```
SELECT P.folio_id, P.edad, PS.valor_glucosa, PS.valor_insulina,
E.riesgo_hipertension
FROM Paciente AS P
JOIN Examen_clinico AS E ON P.folio_id = E.folio_id
JOIN Prueba_sangre AS PS ON E.examen_id = PS.examen_id
WHERE PS.valor_glucosa > 100 AND PS.valor_insulina > 20;
```


	folio_id character varying (15)	edad integer	valor_glucosa numeric (5,2)	valor_insulina numeric (5,2)	riesgo_hipertension boolean
1	1053	54	101.03	24.76	true
2	1075	75	103.28	24.62	false
3	1076	31	104.15	21.59	true
4	1086	26	103.32	20.23	false
5	1087	72	108.59	21.13	true
6	1099	44	100.76	24.83	true
7	1113	64	103.99	22.05	false
8	1134	40	105.81	23.26	true
9	1139	41	104.29	21.80	true
10	1172	51	109.49	20.77	false
11	1194	20	101.94	22.58	false
12	1264	49	105.60	23.08	true
13	1285	61	103.99	23.42	true
14	1319	49	101.23	21.61	false

- 12) Obtener los pacientes con riesgo de hipertensión y que duermen menos de 6 horas, además de mostrar su IMC (índice de masa corporal) y edad.

```
SELECT P.folio_id, P.edad, P.masa_corporal, P.sueno_horas,
E.riesgo_hipertension
FROM Paciente AS P
JOIN Examen_clinico AS E ON P.folio_id = E.folio_id
WHERE E.riesgo_hipertension = TRUE AND P.sueno_horas < 6;
```

	folio_id character varying (15)	edad integer	masa_corporal numeric (5,2)	sueno_horas numeric (3,1)	riesgo_hipertension boolean
1	1000	71	27.89	5.5	true
2	1041	77	16.87	5.2	true
3	1070	64	17.37	5.3	true
4	1082	39	28.27	5.2	true
5	1086	55	33.48	5.8	true
6	1090	71	31.03	5.2	true
7	1101	49	26.96	5.5	true
8	1107	34	22.89	5.9	true
9	1108	28	21.61	5.9	true
10	1110	76	21.67	5.2	true
11	1113	67	18.79	5.7	true
12	1114	20	24.83	5.7	true
13	1116	33	21.01	5.8	true
14	1125	52	26.10	5.3	true
15	1127	55	29.28	5.1	true
16	1136	33	36.63	5.7	true
17	1151	48	29.75	5.4	true
18	1175	32	23.45	5.7	true
19	1199	65	25.13	5.0	true

Árboles de consulta

Un árbol de consultas es una estructura de datos en forma de árbol que representa una consulta mediante una expresión de álgebra relacional. En esta representación, las relaciones de entrada se encuentran en los nodos hoja, mientras que las operaciones del álgebra relacional ocupan los nodos internos.

Para procesar un árbol de consultas, se ejecutan las operaciones de un nodo interno tan pronto como sus operandos están disponibles. Una vez que se realiza una operación, el nodo correspondiente se reemplaza por el resultado obtenido. Este proceso se repite sucesivamente con los nodos internos, avanzando hacia el nodo raíz. Al completar la ejecución en el nodo raíz, se obtiene la relación final que resulta de la consulta.

Optimización heurística

Al ejecutar una consulta, el analizador de consultas genera inicialmente un árbol de consultas que representa una expresión en álgebra relacional. Sin embargo, este árbol inicial suele ser ineficiente si se ejecuta tal cual. Para mejorar su rendimiento, se aplica la optimización heurística, transformándolo en un árbol de consultas equivalente pero más eficiente, lo que reduce los costos de ejecución.

Este proceso se logra mediante el uso de reglas de optimización heurística. Estas reglas consisten en transformaciones basadas en álgebra relacional que se aplican al árbol inicial para convertirlo en un árbol optimizado, listo para ser ejecutado de manera más eficiente.

Resumen de las reglas heurísticas para la optimización algebraica

La principal regla heurística es aplicar en primer lugar las operaciones que reducen el tamaño de los resultados intermedios. Eso significa ejecutar tan pronto como sea posible las operaciones SELECT para reducir el número de tuplas y las operaciones PROJECT para reducir el número de atributos. Esto se lleva a cabo desplazando las operaciones SELECT y PROJECT hacia abajo en el árbol lo más lejos posible. Además, las operaciones SELECT y JOIN más restrictivas (es decir, con las relaciones resultantes con el menor número de tuplas o con el tamaño absoluto menor) deberían ser ejecutadas antes que otras operaciones similares. Esto se hace reordenando los nodos hoja del árbol entre sí evitando los productos cartesianos y ajustando el resto del árbol adecuadamente.

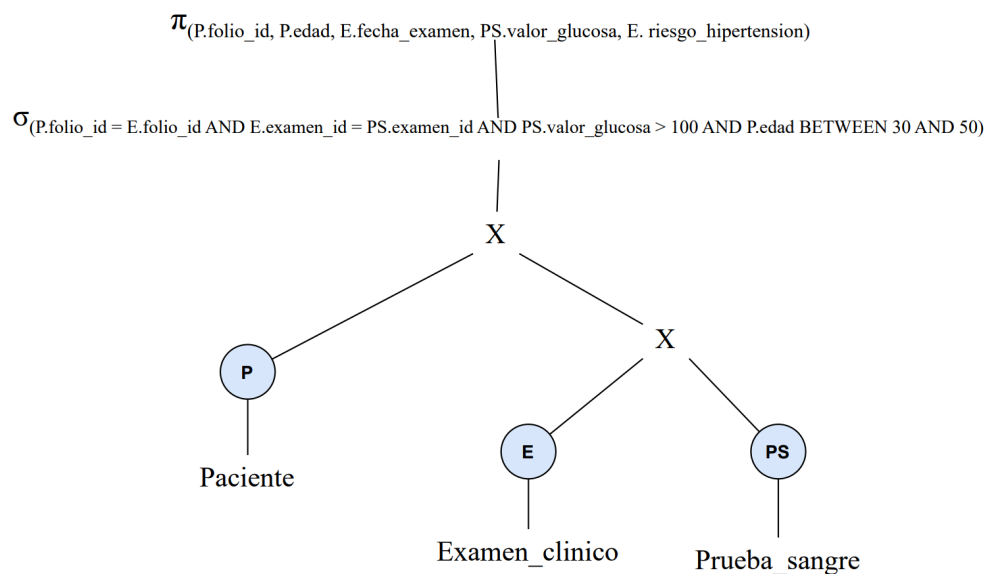
Consulta SQL con optimización heurística

Seleccionar los pacientes que tengan un nivel de glucosa mayor a 100 mg/dL, dentro de un rango específico de edad (por ejemplo entre 30 y 50 años). Recolectar el ID del paciente, su edad, la fecha del examen, el valor de glucosa y el riesgo de hipertensión

Consulta inicial

```
SELECT P.folio_id, P.edad, E.fecha_examen, PS.valor_glucosa,
E.riesgo_hipertension
FROM Paciente P,
     Examen_clinico E,
     Prueba_sangre PS
WHERE P.folio_id = E.folio_id
     AND E.examen_id = PS.examen_id
     AND PS.valor_glucosa > 100
     AND P.edad BETWEEN 30 AND 50;
```

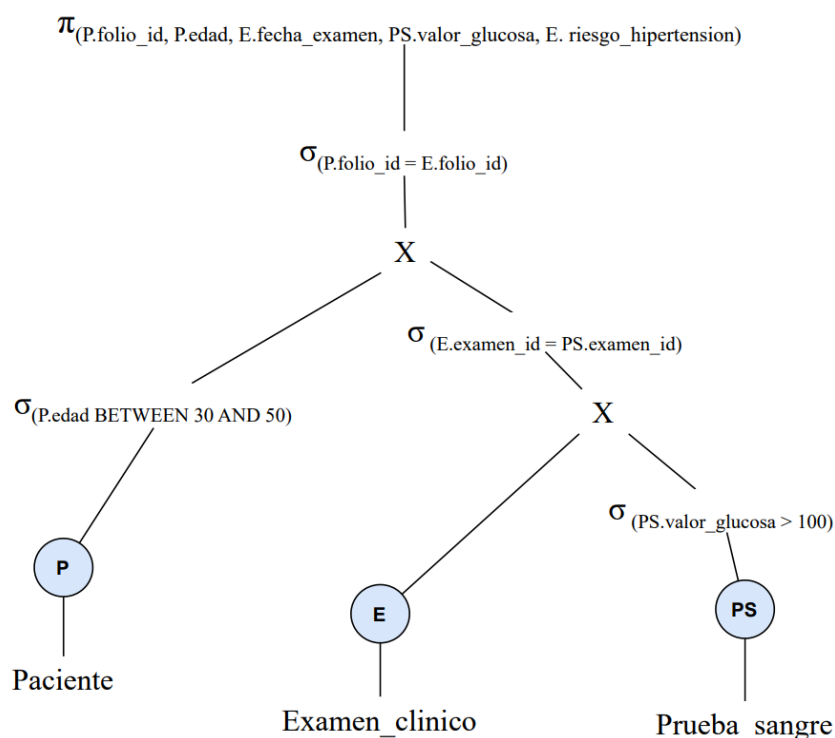
- 1) **Árbol inicial:** En este primer caso, el árbol inicial representa la consulta SQL, donde todas las condiciones del WHERE y las columnas especificadas en el SELECT están agrupadas en una misma operación de selección y proyección. Este enfoque es algo ineficiente porque todas las filas y columnas de las relaciones son cargadas desde el inicio, lo que aumenta el tamaño de los datos manejados, como así también no se reduce el volumen de datos hasta etapas posteriores de ejecución



Árbol de ejecución inicial: Procesamiento directo

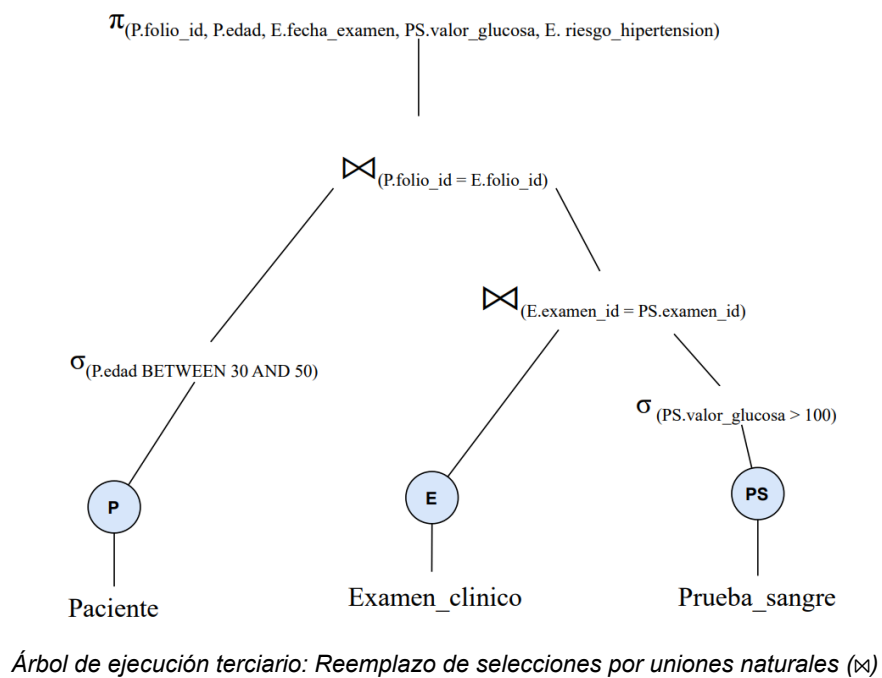
2) **Descomposición de selecciones:** En este caso se hace para reducir la cantidad de datos manejados. Se separan las condiciones de selección en operaciones independientes:

- σ (P.edad BETWEEN 30 AND 50): Reduce el número de filas de Paciente, eliminando aquellos fuera del rango de edad.
- σ (PS.valor_glucosa > 100): Filtra las filas de Prueba_sangre, dejando solo aquellas relevantes.
- σ (P.folio_id = E.folio_id) y σ (E.examen_id = PS.examen_id): Preparan las relaciones para realizar uniones eficientes.



Árbol de ejecución secundario: Descomposición de selecciones

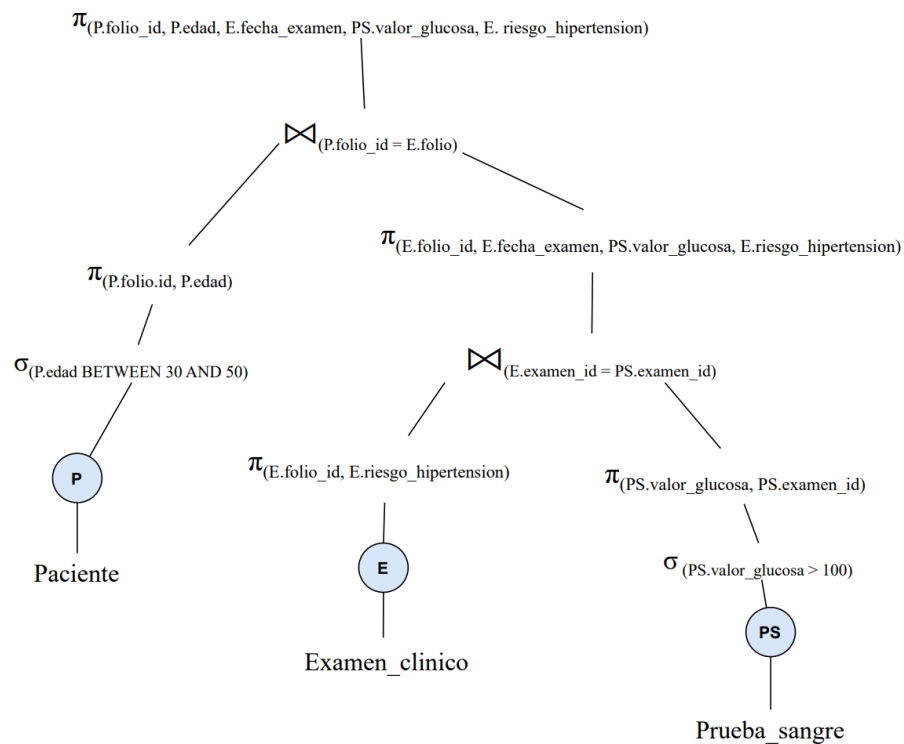
3) **Reemplazo de selecciones por uniones naturales (\bowtie):** En este caso se hace para simplificar y mejorar el rendimiento. Las condiciones de igualdad entre claves foráneas y primarias ($P.folio_id = E.folio_id$, $E.examen_id = PS.examen_id$) se transforman en uniones naturales. Es decir, reemplazo los productos cartesianos que luego tienen una selección por un JOIN o uniones naturales.



4) **Aplicación temprana de proyecciones (π):** En lugar de cargar todas las columnas de las relaciones, se seleccionan solo las necesarias, se hace bajando las selecciones:

- $\pi(P.folio_id, P.edad)$ en Paciente: Filtra solo las columnas relevantes para la consulta.
- $\pi(PS.valor_glucosa, PS.examen_id)$ en Prueba_sangre: Reduce las columnas a las utilizadas en las condiciones y resultados.
- $\pi(E.folio_id, E.fecha_examen, E.riesgo_hipertension)$ en Examen_clinico: Limita las columnas a aquellas involucradas en las uniones y el resultado final.

Este paso disminuye el volumen de datos que pasa por las uniones y reduce el uso de memoria y tiempo de ejecución.



Árbol de ejecución cuaternario y final: Aplicación temprana de proyecciones (π)

5) Orden de ejecución optimizado:

El árbol final anterior organiza las operaciones para minimizar costos:

- 1) Selecciones tempranas: Se eliminan las filas irrelevantes en etapas iniciales
- 2) Uniones naturales: Se combinan solo las filas necesarias, con atributos claves
- 3) Proyecciones: Se eliminan las columnas innecesarias para reducir mas el tamaño de los datos
- 4) Proyeccion final: Produce el resultado final con las columnas especificadas en el SELECT

Esto, como es de esperar, tiene un impacto en el rendimiento, en primer lugar usa menos recursos ya que al reducir datos desde etapas tempranas se disminuye el uso de memoria y disco, y en segundo lugar se da un procesamiento más rapido dado que operaciones como las uniones y las proyecciones son mas rapidas debido a la menor cantidad de datos procesados.

Por lo tanto la consulta final optimizada quedará de la siguiente manera:

Consulta optimizada

```
SELECT P.folio_id, P.edad, E.fecha_examen, PS.valor_glucosa,
E.riesgo_hipertension
FROM (
    SELECT folio_id, edad
    FROM Paciente
    WHERE edad BETWEEN 30 AND 50
) AS P
JOIN (
    SELECT folio_id, examen_id, fecha_examen, riesgo_hipertension
    FROM Examen_clinico
) AS E
ON P.folio_id = E.folio_id
JOIN (
    SELECT examen_id, valor_glucosa
    FROM Prueba_sangre
    WHERE valor_glucosa > 100
) AS PS
ON E.examen_id = PS.examen_id;
```

Implementación de Interfaz Grafica

La interfaz gráfica de la aplicación BioLAB fue diseñada para ofrecer una experiencia intuitiva y funcional que facilite el acceso a la información almacenada en la base de datos de pacientes. Se centra en mejorar la usabilidad para los usuarios finales, como personal médico o administrativo, quienes necesitan consultar datos de manera eficiente y visual. Este enfoque busca optimizar el tiempo que los usuarios invierten en tareas repetitivas, brindándoles herramientas que simplifiquen su flujo de trabajo.

Para su ejecución, debe correr el archivo app.py que se encuentra en el [repositorio de GitHub](#)

Herramientas utilizadas:

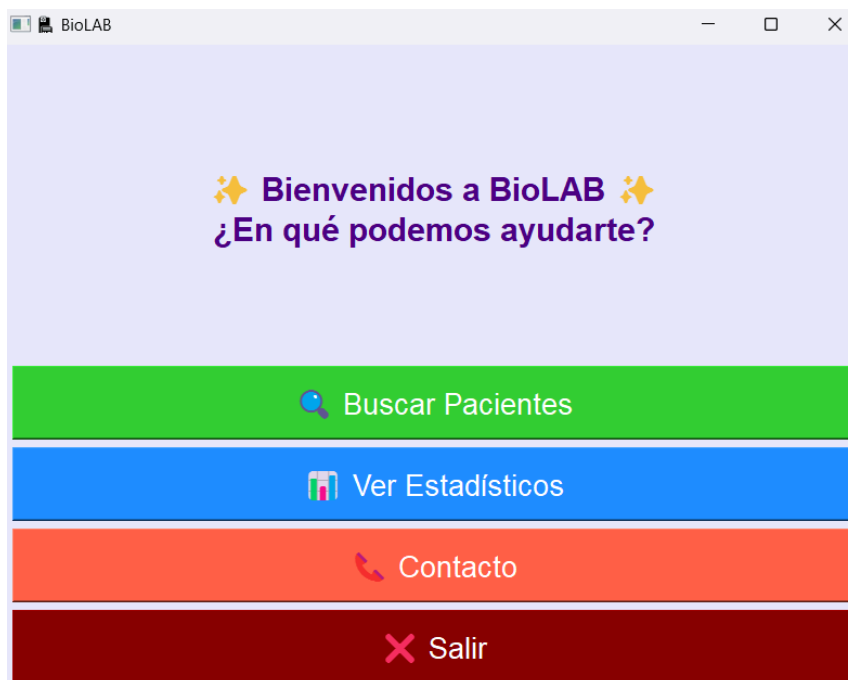
- 1) PyQt para la interfaz gráfica: PyQt fue elegido por su capacidad para crear interfaces gráficas modernas, funcionales y multiplataforma. Permite desarrollar aplicaciones con soporte para personalización y herramientas avanzadas.
- 2) PostgreSQL como sistema de Base de Datos: como mencioné anteriormente, se seleccionó por su robustez y seguridad, ideales para gestionar grandes volúmenes de datos. Su soporte para consultas complejas y cifrado asegura un manejo eficiente y protegido de la información.
- 3) Matplotlib para visualización de datos: matplotlib facilita la creación de gráficas claras y de alta calidad, fundamentales para el módulo de estadísticas. Su integración con PyQt mejora la experiencia del usuario al presentar datos de manera comprensible.
- 4) Python como lenguaje de programación: Python fue elegido por su sintaxis sencilla, amplia disponibilidad de bibliotecas como PyQt5 y psycopg2.
- 5) Git para control de versiones: se utilizó para gestionar cambios, colaborar eficientemente y mantener un historial claro del desarrollo.
- 6) Recursos adicionales:
 - a) Psycopg2: Para conectar Python con PostgreSQL de manera eficiente.
 - b) Qt Designer: Herramienta visual para diseñar interfaces PyQt.

Estructura de la Interfaz

Ventana Principal

La ventana principal actúa como el punto de partida y contiene cuatro opciones principales:

1. **Buscar Pacientes:** Permite a los usuarios localizar registros específicos mediante filtros como folio y sexo. Esto simplifica la navegación por los datos de pacientes.
2. **Ver Estadísticos:** Presenta opciones para visualizar gráficas que resumen información relevante, como distribuciones de IMC o análisis por género. Este módulo está orientado a decisiones basadas en datos.
3. **Contacto:** Proporciona información para soporte técnico.
4. **Salir:** Finaliza la aplicación de forma segura.

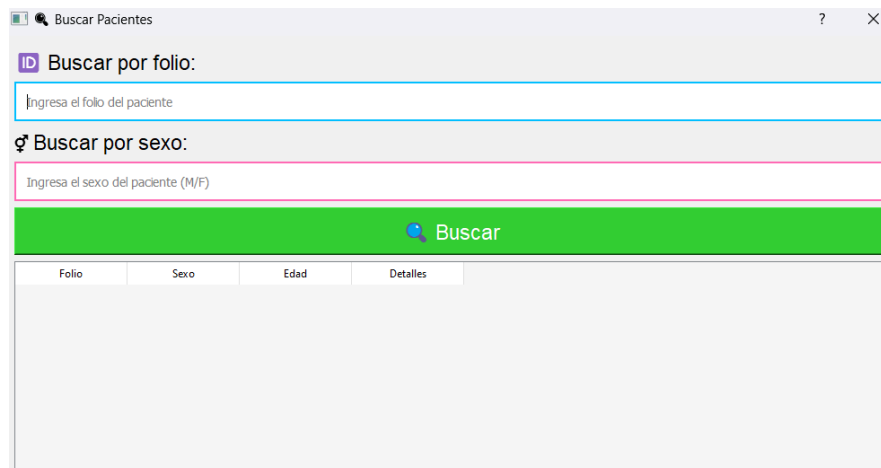


Esta organización permite una navegación clara y directa, evitando confusión y reduciendo la curva de aprendizaje. Además, su diseño modular facilita la incorporación de nuevas funcionalidades en el futuro, asegurando la escalabilidad de la aplicación.

Botón “Buscar Pacientes”


- **Diseño de Filtros:**

- **Búsqueda por Folio:** Proporciona un acceso directo a registros específicos mediante un identificador único.
- **Búsqueda por Sexo:** Permite segmentar rápidamente a los pacientes según su género.



The screenshot shows a window titled "Buscar Pacientes". It contains two search filters: "Buscar por folio:" with a text input field labeled "Ingresa el folio del paciente", and "Buscar por sexo:" with a text input field labeled "Ingresa el sexo del paciente (M/F)". Below these is a green button labeled "Buscar". At the bottom, there is a table header with columns: "Folio", "Sexo", "Edad", and "Detalles".

- **Resultados en tabla:** Los resultados de la consulta se muestran en una tabla con columnas clave: folio, sexo, edad y un botón para ver detalles. Esto asegura que la información esté bien organizada y sea fácilmente accesible.

	Folio	Sexo	Edad	Detalles
1	1000	F	80	 Ver Detalles

- **Botón de "Ver Detalles":** Este botón lleva al usuario a una ventana específica donde se presenta información detallada del paciente seleccionado. Esta funcionalidad es crucial para profundizar en los datos relevantes sin saturar la vista principal. Además, el diseño dinámico permite cargar la información en tiempo real desde la base de datos, garantizando que los datos siempre estén actualizados.



Detalles del Paciente: 1000

Detalles del Paciente con Folio: 1000

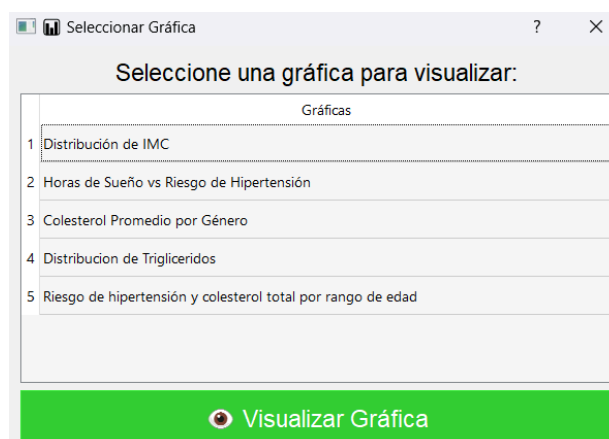
	Campo	Valor
1	folio_id	1000
2	sexo	F
3	edad	80
4	peso	62.03
5	estatura	1.75
6	medida_cintura	91.99
7	segundamedicion_peso	60.39
8	segundamedicion_estatura	1.75
9	segundamedicion_cintura	91.55
10	tension_arterial	132/75
11	sueno_horas	6.3
12	masa_corporal	20.25
13	actividad_total	159.92

Esta ventana permite explorar todos los campos disponibles del registro de un paciente específico. La información se presenta en un formato tabular que vincula el nombre del campo con su valor correspondiente, mejorando la claridad. Este diseño es especialmente útil para verificar información crítica de forma rápida.

Botón “Ver estadísticos”

El módulo de estadísticos fue diseñado para ofrecer visualizaciones claras y útiles sobre los datos almacenados:

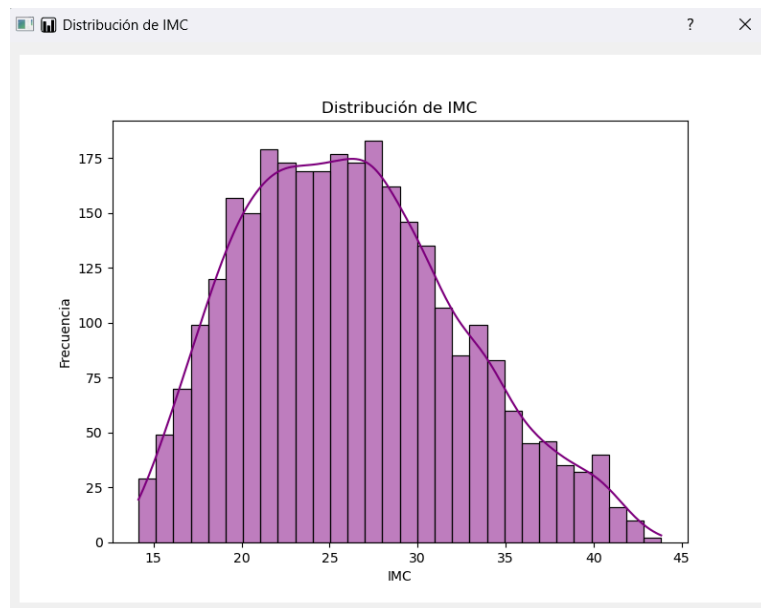
- **Selector de gráficas:** Una lista interactiva permite a los usuarios elegir la gráfica deseada. Cada opción tiene un nombre descriptivo que indica claramente el tipo de análisis que se presentará.



- **Visualización dinámica:** Una nueva ventana muestra la gráfica seleccionada utilizando matplotlib. Este enfoque asegura que los usuarios puedan explorar tendencias y patrones en los datos de forma interactiva.

Las gráficas disponibles incluyen:

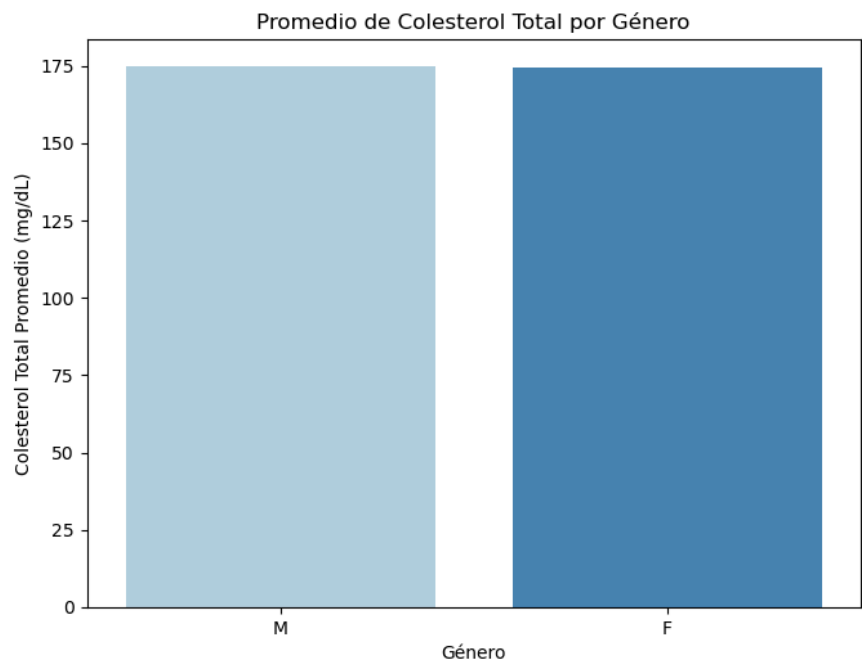
1. Distribución de IMC



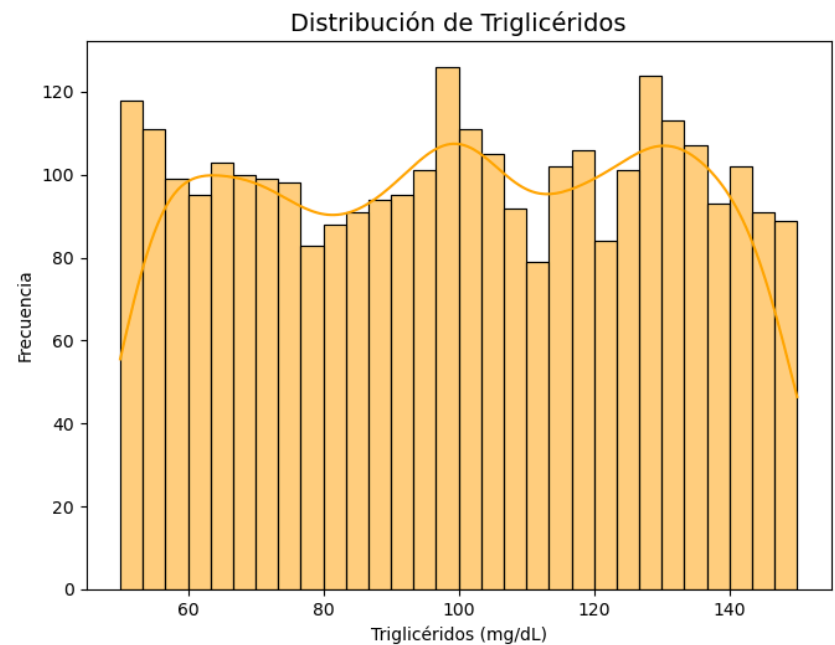
2. Horas de Sueño vs Riesgo de Hipertensión



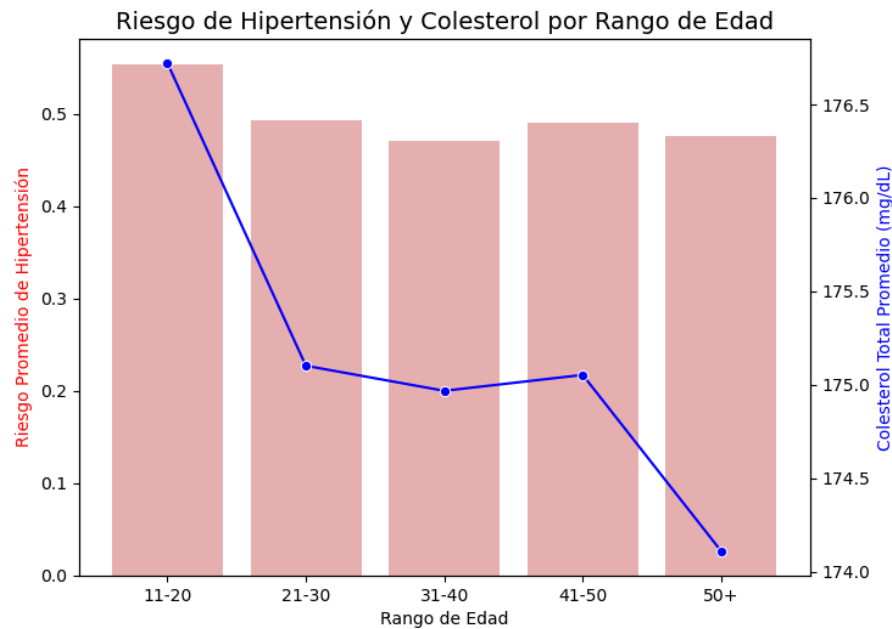
3. Colesterol Promedio por Género



4. Distribución de Triglicéridos.



5. Riesgo de Hipertensión y Colesterol Total por Rango de Edad.



Resultados obtenidos

Durante este trabajo, se diseñó e implementó una base de datos en PostgreSQL, con un modelo relacional que asegura integridad referencial y elimina redundancias. Se mostró como una consulta fue optimizada mediante reglas heurísticas, mejorando el rendimiento y reduciendo costos de ejecución. Se desarrolló una interfaz gráfica intuitiva con PyQt, que incluyó módulos para la búsqueda de pacientes, visualización de estadísticas y análisis interactivo con gráficos. El sistema resultante facilita el análisis clínico y bioquímico, siendo una herramienta útil para la toma de decisiones en salud pública.

Conclusión

A lo largo de este trabajo práctico, se aplicaron los conceptos adquiridos durante el cursado de la asignatura, combinando conocimientos teóricos con investigación autodidacta para llevar a cabo el desarrollo de la base de datos a partir del dataset seleccionado. Este trabajo permitió poner en práctica desde la estructuración y organización de la información hasta la elaboración de consultas avanzadas, asegurando el correcto funcionamiento del sistema. Se resalta la relevancia de gestionar eficientemente grandes volúmenes de datos para optimizar su accesibilidad y comprensión.

Bibliografia

- Fundamentos de bases de datos (Elmasri, Navathe): Elmasri, R., & Navathe, S. B. (2020). *Fundamentos de bases de datos* (7a ed.). Pearson.
- Web de Python Docs: Python Software Foundation. (2023). *Python Documentation*. <https://docs.python.org/3/>
- Web de Qt Designer: The Qt Company. (2023). *Qt Designer*. <https://www.qt.io/qt-designer>
- Web de Matplotlib: Matplotlib Developers. (2023). *Matplotlib Documentation*. <https://matplotlib.org/stable/contents.html>
- Web de Kaggle: Kaggle, Inc. (2023). *Kaggle Datasets*. <https://www.kaggle.com/datasets>.
- Rios, M. (2024). *BioLab-TPFinal-BD* [Repositorio en GitHub]. Recuperado de <https://github.com/MatiasRiosMR/BioLab-TPFinal-BD>