PROYECTO FINAL

DATA ANALYTICS

“Supervivencia del Cáncer de Mama”

Mariana Orellano

2024

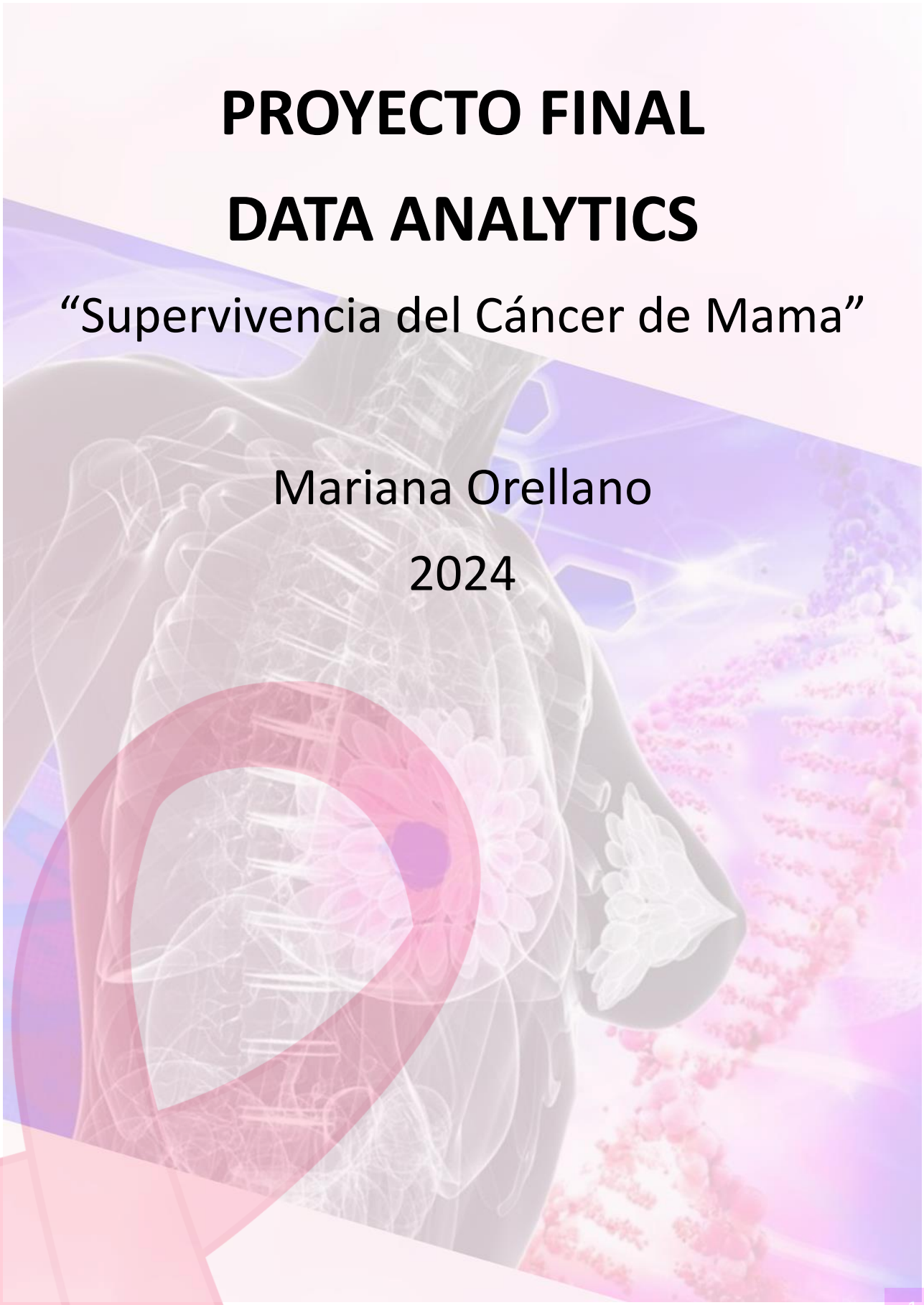


Tabla de contenido

INTRODUCCION 3
Descripción de la temática 3
Hipotesis 3
Objetivos del Proyecto 4
ALCANCE DEL PROYECTO 4
NIVEL OPERATIVO 4
NIVEL TACTICO 5
NIVEL ESTRATEGICO 5
Tabla de versionado 5
Futuras Lineas 6
BASE DE DATOS 6
Introducción 6
DIAGRAMA ENTIDAD RELACION 7
ESQUEMA RELACIONAL 8
Listado de tablas, columnas y descripcion 8
RELACION ENTRE TABLAS 14
Creación de la Base de Datos 15
Armado del DASHBOARD 22
Carga de la base de datos y transformaciones en Power BI 22
VISTA DEL MODELO EN POWER BI 22
TABLA CALENDARIO 23
Dashboard 23
Creacion de la paleta de colores del dashboard 24
Herramientas utilizadas 24
Solapas 25

INTRODUCCION

Descripción de la temática

El objetivo del presente trabajo es llevar a cabo un análisis sobre la supervivencia de pacientes con cáncer de mama. Se cuentan con casos en los que se realizaron cirugías entre los años 2017, 2018 y 2019; sin embargo, no se dispone de información sobre la ciudad o país de origen de la población estudiada.
Se evaluará la información en función de la edad del paciente, el tipo de cáncer detectado, el tipo de cirugía realizada, el estadio de la enfermedad y el estado del paciente (vivo o fallecido).
Para llevar a cabo este análisis, se utilizará una base de datos obtenida de Kaggle.com. Aunque la base de datos no cuenta con más de 1000 registros, como se sugirió inicialmente, se considera que aún así ofrece datos interesantes que permitirán realizar un análisis de las posibilidades de supervivencia de esta enfermedad y encontrar posibles relaciones.
Cabe mencionar que la base de datos también incluye el valor de expresión de cuatro proteínas; sin embargo, no se especifica cuáles son ni se proporciona información sobre cada una. No obstante, se buscarán posibles relaciones basándonos en estos valores.

Hipotesis

Se presume un alto índice de supervivencia en el caso de pacientes con cáncer de mama, y se postula una relación directamente proporcional entre la edad de los pacientes y el porcentaje de supervivencia, fundamentada en la disminución de la actividad celular en edades avanzadas.
¿La mastectomía simple podría demostrar ser más efectiva que otras intervenciones quirúrgicas? ¿Se relaciona con un pronóstico más favorable? Además, se plantea la interrogante sobre si la detección temprana de la enfermedad (en términos de estadio) se asocia con un mejor pronóstico

Objetivos del Proyecto

Mostrar el alto índice de supervivencia a la enfermedad para dar esperanza a quienes la padecen, concientizar sobre la importancia de la atención medica temprana, buscar patrones relacionados con el tipo de cirugía realizado, el estadio en el que se encuentra el tumor al momento de la operación y encontrar relaciones con los valores de expresion de las Proteínas
Para ello se realizaran las siguientes tareas:

- **Determinar el Porcentaje de Supervivencia:** Analizar y comparar el porcentaje de supervivencia en diferentes años para comprender la evolución a lo largo del tiempo.
- **Explorar la Relación con la Edad del Paciente**: Investigar posibles correlaciones entre la edad de los pacientes y el índice de supervivencia, considerando la hipótesis de una relación directamente proporcional.
- **Analizar Tendencias según Tipos de Cirugías**: Identificar patrones o tendencias relacionadas con los diversos tipos de cirugías realizadas, con especial interés en evaluar la posible efectividad de la mastectomía simple en comparación con otras intervenciones.
- **Examinar Tendencias Relacionadas con el Estadio** : Analizar patrones y relaciones en función del estadio en que se detecto la enfermedad buscando entender cómo esta variable afecta la supervivencia.
- **Investigar Patrones en los Valores de Proteínas**: Explorar si existe algún patrón o tendencia en los valores de expresión de las cuatro proteínas proporcionadas en la base de datos, con el objetivo de identificar posibles correlaciones con la supervivencia al cáncer de mama

ALCANCE DEL PROYECTO

Se utilizara un DataSet cuya fuente es kaggle.com, con datos de pacientes operados entre los años 2017 al 2019, no se conoce la ubicación de la población.

El proyecto esta dirigido a distintos tipos de Usuarios Finales:

NIVEL OPERATIVO

Pacientes y familiares : pueden utilizar la información para comprender y conocer las alternativas de tratamientos y las tasas de supervivencia

Profesionales de la Salud: Incluye médicos, oncólogos, enfermeras y otros profesionales de la salud que están directamente involucrados en el diagnóstico y tratamiento de

pacientes con cáncer de mama. Pueden beneficiarse de la información para personalizar los tratamientos y mejorar el cuidado del paciente.

NIVEL TACTICO

Investigadores Médicos : Estos profesionales pueden utilizar los resultados del proyecto para ayudara la comprensión de los factores de supervivencia del cáncer de mama. Pueden estar interesados en investigar nuevos tratamientos, identificar patrones y entender mejor la progresión de la enfermedad.

NIVEL ESTRATEGICO

Organizaciones de Salud Publica : pueden utilizar los resultados del proyecto para implementar políticas de salud pública, campañas de concientización y asignación de recursos.

Tabla de versionado

Version	Acciones
V0	Definicion del tema y eleccion del dataset
V01	Descripcion de la tematica, hipotesis y objetivo
V02	Definicion del alcance armado de diagrama de entidad relacion y esquema relacional definincion de tablas : columnas descripcion y tipo de datos determinacion de PK y FK Definicion de relacion entre tablas
V03	Creacion de base de datos en SQL carga de datos soluciono de problemas - errores armado de vistas Carga en Power BI Definicion de paleta de colores Armado de tabla calendario Transformacion de datos Primera version del tablero
Version Final	Correcion de errores de criterios Incorporacion de solapa Valores de proteina con calculas de medias y medianas

Futuras Lineas

Tomando como punto de partida el Dashboard desarrollado, considero que podría mejorarse incorporando en el data set los siguientes elementos:

- Datos personales de los pacientes, como la dirección, para poder determinar geográficamente los casos, identificar patrones de evolución y tomar acciones como enfocar campañas de concientización en áreas específicas, implementar mamógrafos móviles para la detección temprana, etc. También sería útil incorporar información sobre el tipo de atención médica, ya sea en hospitales públicos o prepagas, para tomar decisiones respecto a la asignación de fondos públicos en salud.
- Sería muy interesante identificar a que proteína especifica se refieren los valores para desarrollar predicciones sobre la evolución de cada caso, lo cual ayudaría al personal médico a tomar decisiones informadas y evaluar distintos cursos de acción.
- Respecto a los marcadores, contar con más información y utilizar criterios médicos podría ser una herramienta muy útil para la predicción.

Personalmente, disfruté mucho realizando este trabajo. Dado que mi madre es una sobreviviente de este tipo de cáncer y he vivido de cerca esta situación como familiar, considero que es crucial proporcionar información, ofrecer esperanza y sobre todo concientizar sobre la detección temprana de esta enfermedad, lo cual es fundamental para mejorar su evolución.

BASE DE DATOS

Introducción

En esta sección del documento se pretende, exponer el diagrama de entidad – Relación diseñado para la base de datos, se presenta también el esquema relacional de la base de datos.

Explicar en detalle, el contenido de cada tabla, definición de llaves primarias y foráneas, definición de columnas y tipos de datos.

Finalmente, se explica brevemente la relación entre las tablas.

En la creación de las tablas para la base de datos, se comenzó por diseñar el diagrama Entidad-Relación y posteriormente la construcción del Esquema Relacional.

Las siguientes tablas

- Status Paciente
- Cirugías
- Histología

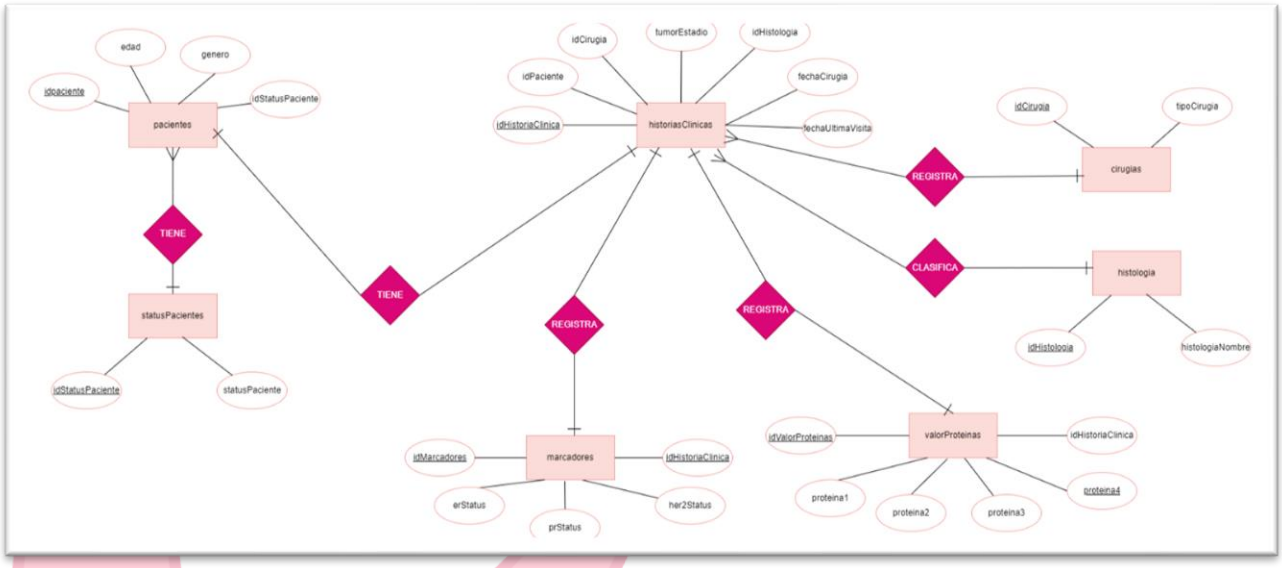
Se diseñaron como Tablas de referencia con sus posibles valores asociados a una ID única, esto facilita la consistencia de los datos y permite cambios en los valores posibles sin tener que actualizar todos los registros en la tabla principal.

Para estas tareas, se utilizaron las herramientas draw.io y miro.com.

DIAGRAMA ENTIDAD RELACION

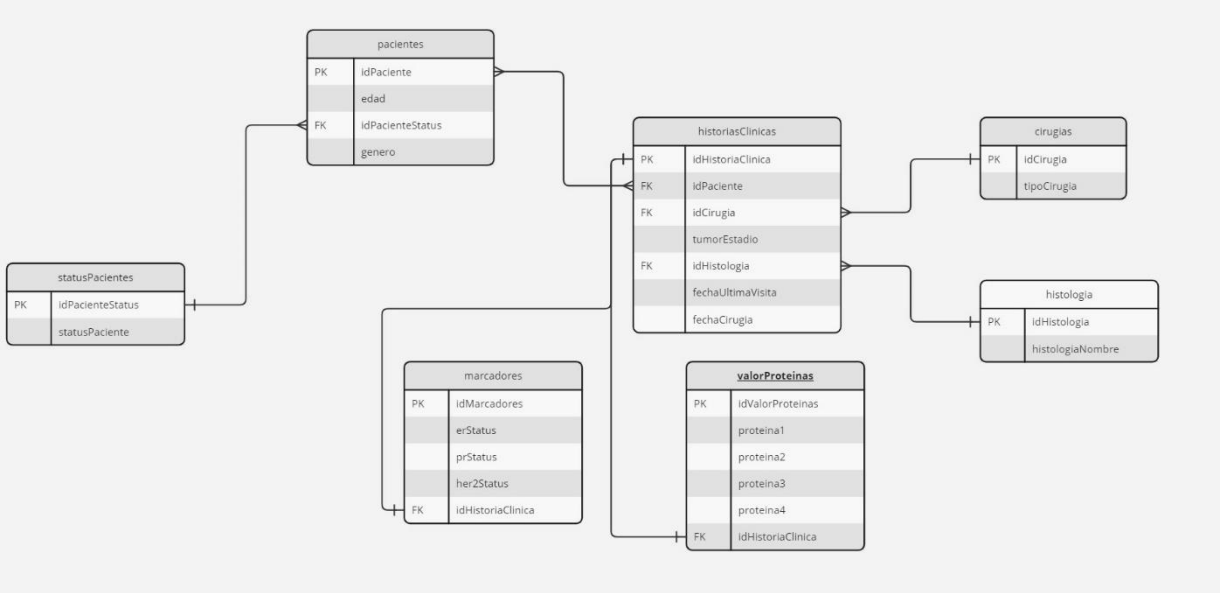
Link para su correcta visualización

https://drive.google.com/file/d/1bOrIYFBI_9jJH1aTzOKsmLJAwcZPSidQ/view?usp=sharing



ESQUEMA RELACIONAL

Link para su correcta visualizaciónhttps://miro.com/app/board/uXjVN98LzOk=/?share_link_id=568087958
[12](https://miro.com/app/board/uXjVN98LzOk=/?share_link_id=568087958)



Listado de tablas, columnas y descripcion

Tabla pacientes

Tabla 1	pacientes		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idPaciente	INT	PK	Identificador único para cada caso de paciente. Autoincremental. *NOT NULL*
edad	INT		Edad del paciente, expresada como un número entero. *NOT NULL*
idPacienteStatus	INT	FK	Clave foránea que se relaciona con la tabla "statusPacientes" e identifica el estado del paciente (1: Alive, 2: Dead, 3: Other). *NOT NULL*
genero	VAR		Género del paciente. Se utiliza VARCHAR para permitir la inclusión de información diversa Valor por DEFAULT Female. *NOT NULL*

La tabla "pacientes" almacena información sobre cada persona afectada por la enfermedad.

Se agregó la columna "idPaciente" como clave primaria para identificar de manera única cada caso, ya que el dataset no contiene datos personales como nombres o identificaciones de pacientes.

La columna "idPacienteStatus" se utiliza como clave foránea para relacionarse con la tabla "statusPacientes", indicando el estado del paciente (vivo, muerto u otro). Los valores numéricos asociados son 1 para "Alive", 2 para "Dead" y 3 para "Other".

La columna "edad" representa la edad del paciente como un número entero.

La columna "genero" indica el género del paciente y se ha elegido utilizar VARCHAR. Ya que como la gran mayoría de los casos son femeninos se coloco el valor “FEMALE” como Default

En “la vida real” esta tabla podría contener más campos relacionados con los datos personales del paciente, como ser Nombre completo, numero de documento, dirección, teléfono, prepaga etc a su vez podría dar lugar a más tablas de referencia como por ejemplo una tabla que agrupe las distintas obras sociales o prepagas

Tabla statusPacientes

Tabla 2	**statusPacientes**		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idPacienteStatus	INT	PK	La columna `idPacienteStatus` se define como la clave primaria (PK) de la tabla y se relaciona con el campo correspondiente en la tabla principal `Pacientes`. *NOT NULL*
statusPaciente	VAR		La columna `statusPaciente` almacena el estado del paciente, que puede ser “Alive", "Dead" ,"Other. *NOT NULL*

Como se comento más arriba esta es una tabla de referencia de la tabla Pacientes

Indica el estado del paciente (vivo, muerto u otro). Los valores numéricos asociados son 1 para "Alive", 2 para "Dead" y 3 para "Other", estos corresponden al campo idPacienteStatus

Tabla historiasClinicas

La tabla `historiasClinicas` es central en la base de datos, y su estructura se detalla a continuación

Tabla 3	historiasClinicas		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idHistoriaClinica	INT	PK	Identificador único de la historia clínica. Autoincremental . *NOT NULL*
idPaciente	INT	FK	Relaciona con la tabla `Pacientes`. *NOT NULL*
idCirugia	INT	FK	Relaciona con la tabla `Cirugías`. *NOT NULL*
tumorEstadio	VAR		Información sobre el estadio del tumor. *NOT NULL*
idHistologia	INT	FK	Relaciona con la tabla `Histologia`. *NOT NULL*
fechaUltimaVisita	DATE		Fecha de la última visita del paciente.
fechaCirugia	DATE		Fecha de la cirugía realizada. *NOT NULL*

Esta tabla constituye el núcleo de la base de datos. Su clave primaria es el campo `idHistoriaClinica`, que sirve como identificador único para cada historia clínica. La tabla establece relaciones con otras tablas, como `Pacientes`, `Cirugías`, `Histologia`, entre otras.

Además proporciona detalles sobre el estadio del tumor y fechas relevantes: registra la fecha de la última visita y la fecha de la cirugía de cada paciente

Tabla marcadores

Tabla 5	marcadores		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idMarcadores	INT	PK	Identificador único de la tabla marcadores que se relaciona con una historia Clinica determinada, autoincremental. *NOT NULL*
erStatus	VAR		Estado Hormonal: Receptor de estrógenos. Positivo o Negativo. *NOT NULL*
prStatus	VAR		Estado Hormonal: Receptor de progesterona. Positivo o Negativo. *NOT NULL*
her2Status	VAR		Receptor HER2, Positivo o Negativo. *NOT NULL*
idHistoriaClinica	INT	FK	Relaciona con la tabla ´Historias Clinicas'. *NOT NULL*

Tabla valorProteinas

Tabla 4	valorProteinas		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idValorProteinas	INT	PK	Identificador único de la tabla proteinas que se relaciona con una historia Clinica determinada, autoincremental. *NOT NULL*
proteina1	FLOAT		Nivel de expresión de proteína 1. *NOT NULL*
proteina2	FLOAT		Nivel de expresión de proteína 2 *NOT NULL*
proteina3	FLOAT		Nivel de expresión de proteína 3 *NOT NULL*
proteina4	FLOAT		Nivel de expresión de proteína 4 *NOT NULL*
idHistoriaClinica	INT	FK	Relaciona con la tabla ´Historias Clinicas' *NOT NULL*

Considero que para la compresión de la información que contienen estas dos tablas, es necesario una breve explicación, lo mas simple posible, del tema

Se elaboro un resumen,del siguiente articulo :

https://www.medicinabuenosaires.com/revistas/vol62-02/1/cancermama.htm

El cáncer de mama, una neoplasia común y mortal en mujeres de distintos países, ha llevado al descubrimiento de marcadores tumorales mediante la biología molecular. Estos marcadores se dividen en tres categorías y otros sin clasificar.

Categoría I:

Incluye marcadores con utilidad comprobada, como el tamaño del tumor, estadio de ganglios linfáticos, y el estado hormonal (RcE y RcPg). Este último, relacionado con receptores de estrógeno y progesterona, tiene importancia en decisiones terapéuticas.

Categoría II:

Engloba marcadores ampliamente estudiados pero que requieren validación estadística, como Her-2/neu, p53, invasión vascular, Ki 67, y síntesis de ADN.

Categoría III:

Contiene marcadores poco estudiados, como la ploidía de ADN, factor de crecimiento de endotelio vascular, receptor para factor de crecimiento epidérmico, Bcl-2, pS2 y catepsina D.

No clasificados:

Incluyen marcadores como el antígeno carcinoembrionario (CEA), mucinas, BRCA-1 y BRCA-2, ciclinas, factores de crecimiento, y proteasas.

Estos marcadores proveen información crucial para la identificación de riesgos, pronósticos, y respuestas a tratamientos en pacientes con cáncer de mama.

En relación con la tabla "Marcadores", los campos 'erStatus', 'prStatus' y 'her2Status', correspondientes a los Receptores de Estrógenos, Receptores de Progesterona y el Receptor HER2, respectivamente, están vinculados a la Categoría I y la Categoría II en el contexto de los marcadores tumorales para el cáncer de mama. Estos campos indican si estos receptores están presentes (POSITIVO) o ausentes (NEGATIVO).

"ER Status" Positivo: Indica que las células cancerosas tienen receptores de estrógeno en su superficie.

"ER Status" Negativo: Indica que las células carecen de receptores de estrógeno.

"PR Status" Positivo o Negativo: Similar a la interpretación de "ER Status" pero para receptores de progesterona.

"HER2 Positivo": Indica sobreexpresión o amplificación del Receptor HER2 en las células cancerosas, sugiriendo la posibilidad de respuesta a terapias específicas.

Un resultado "HER2 Positivo" es importante para determinar tratamientos específicos, como trastuzumab (Herceptin), mientras que "HER2 Negativo" indica la ausencia significativa de sobreexpresión.

En cuanto a la tabla "Proteínas", el dataset registra los niveles de expresión de 4 proteínas (identificadas como 1, 2, 3 y 4). Aunque, según el artículo, hay más de 10 proteínas relevantes, lamentablemente, en el conjunto de datos, no se identifican con nombres específicos, y no se proporcionan "valores normales". Interpretar si estos valores son "elevados" debe hacerse con criterio médico y en el contexto del desarrollo de los síntomas de cada paciente.

A pesar de la falta de identificación de proteínas específicas y valores normales, considero que podría aplicarse algún criterio estadístico para evaluar patrones en los niveles de las 4 proteínas registradas.

Tabla histología

Tabla 6	histologia		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idHistologia	INT	PK	Identificador único de la tabla se define como la clave primaria (PK) de la tabla y se relaciona con el campo correspondiente en la tabla principal `historiaClinica` Autoincremental. *NOT NULL*
histologiaNombre	VAR		Tipo de cáncer de los pacientes . *NOT NULL*

La tabla de histología actúa como una tabla de referencia para la tabla HistoriasClinicas. Se compone de cuatro valores posibles, cada uno identificado por un número único del 1 al 4. Estos valores se corresponden con los siguientes nombres de tipos de cáncer: "Infiltrating Ductal Carcinoma," "Infiltrating Lobular Carcinoma," "Mucinous Carcinoma," y "Other," respectivamente. Cada entrada en esta tabla proporciona información sobre el tipo específico de cáncer presente en los pacientes, permitiendo una clasificación detallada de la histología asociada a cada historia clínica. La columna idHistologia se define como la clave primaria (PK) de esta tabla y se relaciona con el campo correspondiente en la tabla principal historiaClinica.

Tabla cirugías

Tabla 7	cirugias		
Columna	**Tipo de Dato**	**PK/FK**	**Descripción**
idCirugia	INT	PK	Identificador único de la tabla se define como la clave primaria (PK) de la tabla y se relaciona con el campo correspondiente en la tabla principal `historiaClinica` Autoincremental *NOT NULL*
tipoCirugia	VAR		Tipo de cirugía realizada a los pacientes *NOT NULL*

La tabla de cirugias al igual que la tabla histologia actúa como una tabla de referencia para la tabla HistoriasClinicas. Se compone de cuatro valores posibles, cada uno identificado por un número único del 1 al 4. Estos valores se corresponden con los siguientes nombres de cirugia:

"Infiltrating Simple Mastectomy," " Modified Radical Mastectomy," " Lumpectomy," y "Other," respectivamente. Cada entrada en esta tabla proporciona información sobre el tipo cirugía practicada a los pacientes. La columna idCirugia se define como la clave primaria (PK) de esta tabla y se relaciona con el campo correspondiente en la tabla principal historiaClinica.

RELACION ENTRE TABLAS

La tabla Pacientes tiene una relación 1:N con la tabla statusPacientes y se vinculan a través del campo idStatusPaciente, cada paciente tiene un status, pero cada status puede estar vinculado a varios pacientes; también existe una relación 1:1 con la tabla historiasClinicas, a través del campo idPaciente, cada paciente tiene una historia clínica y cada historia clínica está vinculada a un paciente. Para ambas relaciones se utilizo el verbo TIENE

La tabla historiasClinicas se relaciona 1:N tanto con la tabla cirugías como con la tabla histología, a través de los campos idcirugia e idHistologia respectivamente. Para expresar la relación con la tabla cirugías se utilizo el verbo REGISTRA, y para la relación con la tabla histología CLASIFICA.

A su vez, como esta es la tabla principal, también se relaciona con las tablas marcadores y valoProteinas en ambos casos las relación es 1:1 ya que se se genero un registro único relacionado a cada historia clínica a través de idmarcadores e idValorProteinas. El verbo utilizado en ambos casos es REGISTRA

Creación de la Base de Datos

Antes de comenzar a trabajar con Power BI, se creó la base de datos en SQL utilizando Microsoft SQL Server Management Studio.

Posteriormente, se procedió a la creación de las tablas, especificando los tipos de datos y las claves primarias de cada una.

```
CREATE DATABASE Supervivencia_Cancer_de_mama

--- sentencias para crear las 7 tablas de la base de datos, se fueron ejecutando una
a una
CREATE TABLE [pacientes]
(
idPaciente INT NOT NULL IDENTITY (1,1),
edad INT NOT NULL,
genero VARCHAR(20) NOT NULL,
idPacienteStatus INT NOT NULL,
PRIMARY KEY (idPaciente)
);

CREATE TABLE [statusPacientes]
(
	idPacienteStatus INT NOT NULL IDENTITY (1,1),
	statusPaciente VARCHAR(20) NOT NULL,
	PRIMARY KEY (idPacienteStatus)
);
CREATE TABLE [historiasClinicas]
(
	idHistoriaClinica INT NOT NULL IDENTITY (1,1),
	idPaciente INT NOT NULL,
	idCirugia INT NOT NULL,
	tumorEstadio VARCHAR(5) NOT NULL,
	idHistologia INT NOT NULL,
	fechaUltimaVisita DATE,

	fechaCirugia DATE NOT NULL,

	PRIMARY KEY (idHistoriaClinica)
);

CREATE TABLE [marcadores]
(
	idMarcadores INT NOT NULL IDENTITY (1,1),
	erStatus VARCHAR(20) NOT NULL,
	prStatus VARCHAR(20) NOT NULL,
	her2Status VARCHAR(20) NOT NULL,
	idHistoriaClinica INT NOT NULL,
	PRIMARY KEY (idMarcadores)
);
```

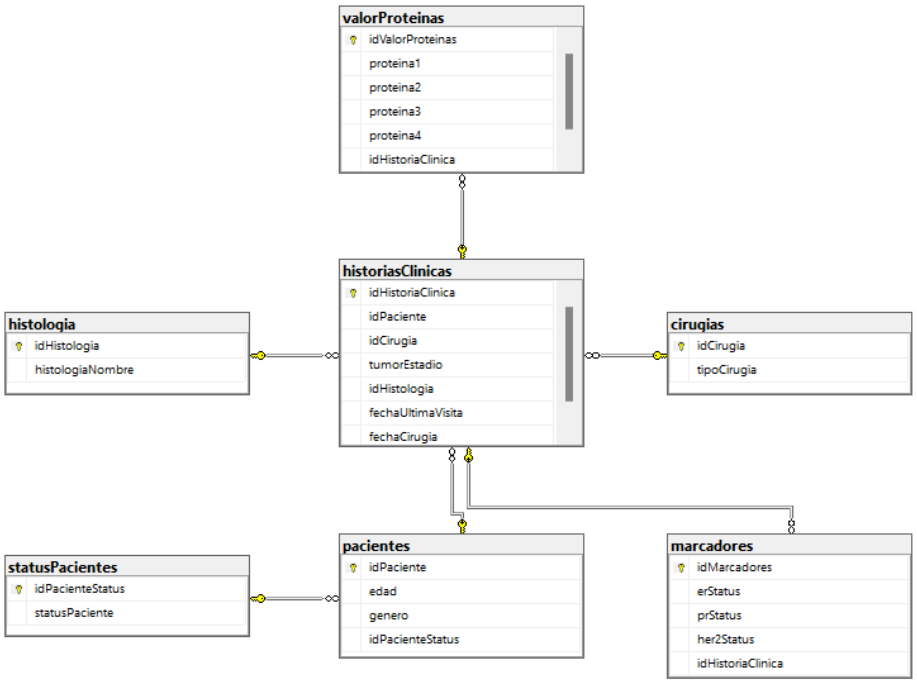
```
CREATE TABLE [valorProteinas]
(
	idValorProteinas INT NOT NULL IDENTITY (1,1),



	proteina1 FLOAT NOT NULL,
	proteina2 FLOAT NOT NULL,
	proteina3 FLOAT NOT NULL,
	proteina4 FLOAT NOT NULL,
	idHistoriaClinica INT NOT NULL,
	PRIMARY KEY (idValorProteinas)
);
CREATE TABLE [histologia]
(
	idHistologia INT NOT NULL IDENTITY (1,1),
	histologiaNombre VARCHAR (40) NOT NULL,
	PRIMARY KEY (idHistologia)
);

CREATE TABLE [cirugias]
(
	idCirugia INT NOT NULL IDENTITY (1,1),
	tipoCirugia VARCHAR (40) NOT NULL,
	PRIMARY KEY (idCirugia)
);
```

Se realizo el armado del Diagrama relacional



A continuación, se procedió a la carga de datos provenientes del conjunto de datos (DATASET). Se realizaron algunas modificaciones utilizando Microsoft Excel

Tabla Pacientes y statusPacientes

Se tomaron las siguientes columnas del DATASET

	A	B	C	D
	Id_Patient	Age	Gender	Patient_Status
	1	42	FEMALE	Alive
	2	54	FEMALE	Dead
	3	63	FEMALE	Alive
	4	78	FEMALE	Alive
	5	42	FEMALE	Alive
	6	80	FEMALE	Alive
	7	66	FEMALE	Alive
	8	36	FEMALE	Alive

Se reemplazaron los nombre de las columnas, se reemplazo los vacíos en Estatus pacientes por "Other", y se asigno un numero a cada uno de los estados

1 – "Alive"

2 - "Dead"

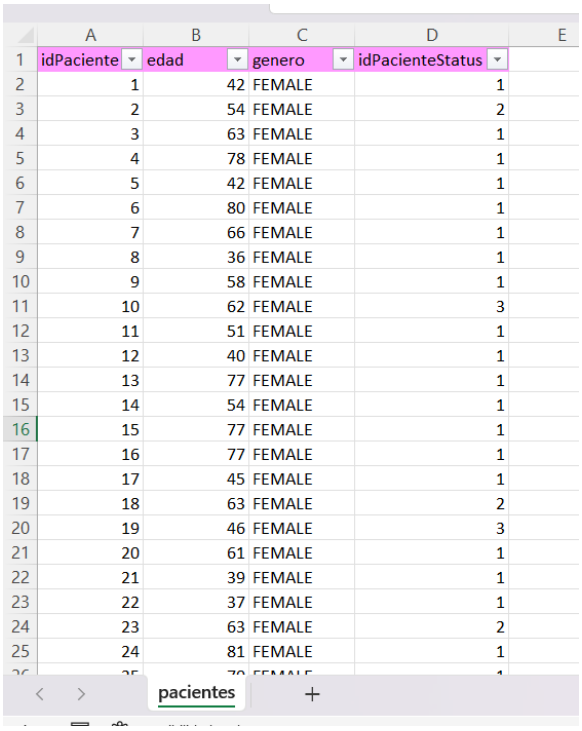
3 – "Other"

Se utilizo la siguiente formula

+si([@[Patient_Status]]=$H$2;1;si([@[Patient_Status]]=$H$3;2;3))

Luego se copio y pego valores, y se elimino la columna Patient_Status

Quedando de la siguiente manera



	A	B	C	D	E
1	idPaciente	edad	genero	idPacienteStatus	
2	1	42	FEMALE	1	
3	2	54	FEMALE	2	
4	3	63	FEMALE	1	
5	4	78	FEMALE	1	
6	5	42	FEMALE	1	
7	6	80	FEMALE	1	
8	7	66	FEMALE	1	
9	8	36	FEMALE	1	
10	9	58	FEMALE	1	
11	10	62	FEMALE	3	
12	11	51	FEMALE	1	
13	12	40	FEMALE	1	
14	13	77	FEMALE	1	
15	14	54	FEMALE	1	
16	15	77	FEMALE	1	
17	16	77	FEMALE	1	
18	17	45	FEMALE	1	
19	18	63	FEMALE	2	
20	19	46	FEMALE	3	
21	20	61	FEMALE	1	
22	21	39	FEMALE	1	
23	22	37	FEMALE	1	
24	23	63	FEMALE	2	
25	24	81	FEMALE	1	

pacientes

Se procedió a guardar el libro como .CSV

Se presentaron ciertos inconvenientes al intentar cargar los datos en la base de datos. A continuación, se detallan las correcciones realizadas:

- Problemas con la columna "genero":

Se identificó que la columna "genero" no estaba formateada como texto en Excel.

- Problemas con la longitud de caracteres en la columna "genero":

A pesar de que se especificó una longitud de 20 caracteres para la columna "genero", hubo incompatibilidad, ya que SQL consideraba que la longitud era de 50. Se ajustó el tipo de dato a VARCHAR(60) para garantizar la compatibilidad con los valores reales.

- Problemas con la identidad en la columna "Identity":

La columna de identidad estaba marcada como solo lectura. Para resolver esto, se desactivó temporalmente la identidad durante la carga de datos.

- Gestión de claves foráneas y diagrama de relaciones:

Se elaboró previamente un diagrama de relaciones entre las tablas y se establecieron las claves foráneas. Al intentar cargar los datos, se detectó un error debido a la falta de la clave foránea idPacienteStatus. Para solucionar este problema, se cargó primero la tabla "StatusPacientes".

Después de implementar estas correcciones, se logró cargar correctamente las tablas "StatusPacientes" y "Pacientes". A continuación, se procederá con la carga de las demás tablas, respetando el orden de carga establecido por las relaciones entre ellas.

Tabla Cirugias

Durante el proceso de carga de datos de cirugías, se identificó que había varias filas en las que el tipo de cirugía estaba etiquetado como "Other". Tras una investigación adicional, se determinó que este tipo de cirugía podría corresponder a "Axillary lymph nodes", un procedimiento asociado comúnmente con pacientes que tienen cáncer de mama.

En consecuencia, se procedió a reemplazar el valor "Other" por "Axillary lymph nodes" en dichas filas, con el objetivo de reflejar con mayor precisión el tipo específico de cirugía realizado. Esta corrección se realizó para mejorar la consistencia y la exactitud de los datos en la base de datos.

La tabla quedaría confirmada de la siguiente manera:

idCirugia	tipoCirugia
1	Axillary lymph nodes
2	Lumpectomy
3	Modified Radical Mastectomy
4	Simple Mastectomy

Se trabajo desde excel tomando la columna Cirugías y elimando los duplicados

Los registros de esta tabla se ingresaron de manera manual, ya que se experimentaban bloqueos en el programa al intentar realizar la operación a través del asistente. Se empleó la siguiente consulta SQL para llevar a cabo la inserción:

```
INSERT INTO cirugias (idCirugia, tipoCirugia)
VALUES
  (1, 'Axillary lymph nodes'),
  (2, 'Lumpectomy'),
  (3, 'Modified Radical Mastectomy'),
  (4, 'Simple Mastectomy');
```

*Este cambio se revirtió al armar el glosario

En el momento del armado del Glosario, al buscar las definiciones de las cirugías, me di cuenta que esta interpretación que había hecho no era correcta, además de que “suponer “ o inferir este tipo de datos no es una buena practica, por lo que se revirtió el cambio realizando la modificación desde Power BI en la tabla Cirugías desde Power Query

	idCirugia	tipoCirugia	historiasClinicas
1	1	Axillary lymph nodes	Table
2	2	Lumpectomy	Table
3	3	Modified Radical Mastectomy	Table
4	4	Simple Mastectomy	Table

Reemplazando el valor "Axillary lymph nodes" por Other

Y de esta manera muy simple se modifico todo el Dashboard, por eso es importante trabajar con ID cuando tenemos valores que se van a repetir en distintos registros, realizando solamente una modiciacion esto se aplico a todo

Tabla Histologia

De manera similar se trabajo con la tabla Histología, pero sin desactivar la Identidad colocando solo los valores de la columna histologiaNombre

```
INSERT INTO histologia(histologiaNombre)
VALUES


('Infiltrating Ductal Carcinoma'),
  ('Infiltrating Lobular Carcinoma'),
  ('Mucinous Carcinoma');
```

Tabla Historias Clinicas

Para la carga de datos en la tabla historiasClinicas, se trabajo con el asistente, habilitando la inserción de identidad para no tener que desactivarla temporalmente. Hubo en principio un error con las columnas de fechas ya que el formato de Excel era

DD/MM/AAAA

Y en SQL

AAAA/MM/DD

Corrigiendo esto en el CSV se soluciono el inconveniente

Tabla marcadores y valores de proteinas

Utilizando el asistente de importación se cargaron las tablas marcadores y valorProteinas

Creando previamente en Excel los archivos .CSV incorporando una columna como idMarcadores y idValorProteinas y otra con la idHistoriaClinica(FK)

VISTAS

Se comenzó con la creación de algunas vistas con la intención de buscar relaciones entre los datos de manera de poder mostrar distintos insights

```
---  vista calculo de casos femeninos y masculinos
--- vista que totaliza el estado de los pacientes, vivos muertos y otros
---vista de pacientes  cuyo estado es OTHER agrupados por edad

---vista de pacientes  cuyo estado es ALIVE agrupados por edad
---vista de pacientes  cuyo estado es DEAD agrupados por edad
--- vista status de pacientes por año
--- PACIENTES POR TIPO DE CIRUGIA
---CANT DE PACIENTES POR TIPO DE CIRUGIA Y POR STATUS
-- cant de pacientes por estadio
--- cantidad de pacientes por estadio y por status
--- cantidad de pacientes por histologia
-- cantidad de pacientes por histologia y status
```

Realizando esta tarea se detecto un error en aquellos que no tenían fecha de ultima visita, en lugar de estar los campos vacíos tenían en valor 1900-01-01

Se ejecuto la siguiente sentencia para corregir el error:

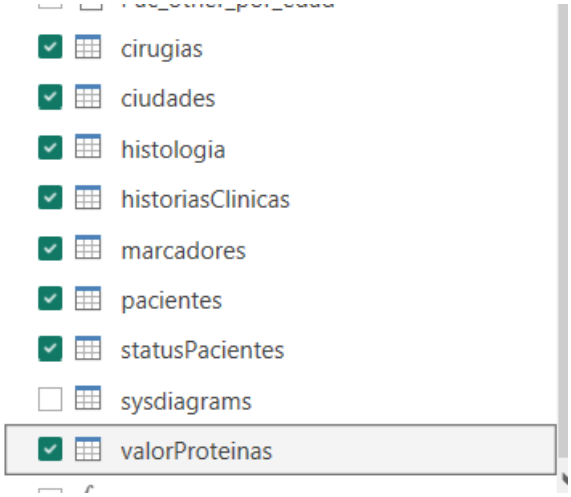
```
UPDATE historiasClinicas
SET fechaUltimaVisita = NULL
WHERE fechaUltimaVisita = '1900-01-01';
```

Finalmente no se utilizaron estas vistas, ya que al cargar todas las tablas del modelo en Power Bi mas las vistas complejizaban mas de lo que resolvían, pero quedan cargadas a la base por si posteriormente se decide utilizar alguna de ellas

Armado del DASHBOARD

Carga de la base de datos y transformaciones en Power BI

Se obtuvieron los datos de la base SQL seleccionando todas las tablas pertenecientes al modelo



VISTA DEL MODELO EN POWER BI

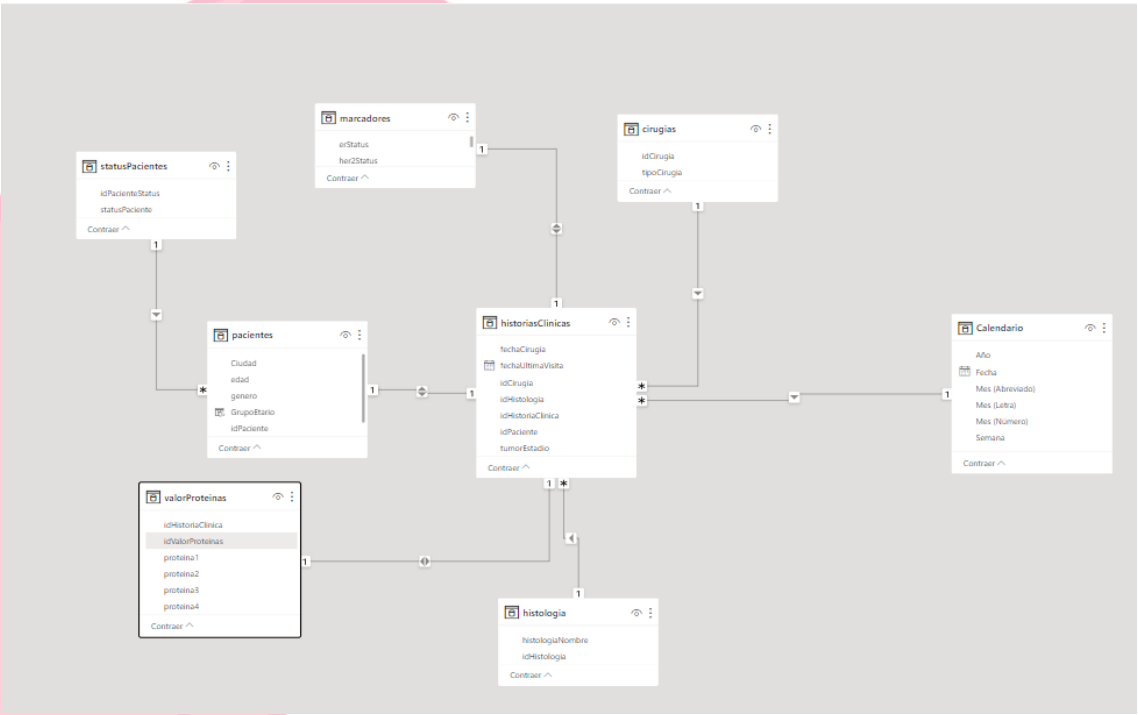


TABLA CALENDARIO

Para la creación de la “Tabla Calendario” se utilizo la columna ‘fechaCirugia’ de la tabla ‘historiasClinicas’

Calendario

Opciones de presentación

```
let
    Source = historiasClinicas,

// StartDate = Date.From(List.Min(Source[Fecha Reclamo])), // Fecha de inicio
// EndDate = Date.From(List.Max(Source[Fecha Reclamo])), // Fecha de fin

    StartDate = #date(2017, 1, 15), // Fecha de inicio
    EndDate = #date(2019, 11, 21), // Fecha de fin

    NumberOfDays = Duration.From(EndDate - StartDate) / #duration (1, 0, 0, 0),
    DateList = List.Dates(StartDate, NumberOfDays, #duration(1, 0, 0, 0)),
    CalendarTable = Table.FromList(DateList, Splitter.SplitByNothing()),
    RenameColumns = Table.RenameColumns (CalendarTable, {{"Column1", "Fecha"}}),
    AddYear = Table.AddColumn (RenameColumns, "Año", each Date.Year ([Fecha]), Int64.Type),
    AddMonthNumber = Table.AddColumn (AddYear, "Mes (Número)", each Date.Month([Fecha]), Int64.Type),
    AddMonthName = Table.AddColumn(AddMonthNumber, "Mes (Letra)", each Date.ToText([Fecha], "MMMM")),
    AddMonthAbbreviation = Table.AddColumn(AddMonthName, "Mes (Abreviado)", each Date.ToText([Fecha], "MMM")),
    AddWeek = Table.AddColumn(AddMonthAbbreviation, "Semana", each Date.WeekOfYear ([Fecha])),
    #"Tipo cambiado" = Table.TransformColumnTypes(AddWeek,{{"Fecha", type date}, {"Mes (Letra)", type text}, {"Mes (Abreviado)", type text}})

in
    #"Tipo cambiado"
```

Se utilizo la sentencia de lenguaje M vista en clases, usando las fechas fijas (primera y ultima) de la columna

Dashboard

El Dashboard comienza con su portada con una imagen, cuya fuente es : https://www.topdoctors.com.ar , que ha servido como “inspiración” Para la elección de la paleta de colores, ya que como “logo” de la temática solo se conoce el lazo rosa

Creacion de la paleta de colores del dashboard

Estos colores fueron los seleccionados utilizando la herramienta
https://color.adobe.com/es/create/image

Herramientas utilizadas

Para la elaboración del tablero, se emplearon las siguientes herramientas:

- Microsoft Excel: se realizaron modificaciones y transformaciones en el conjunto de datos, así como la creación de tablas.
- Microsoft SQL Server: se creó la base de datos a partir del conjunto de datos, se realizaron modificaciones en algunos campos y se estructuró la base de datos relacional.
- Microsoft Power BI: se llevaron a cabo las últimas modificaciones y transformaciones, además de la creación del panel de control.
- Chat de OpenAI: se utilizó para la corrección de código SQL y DAX, así como para la revisión de la redacción de los textos.
- Canva para crear elementos visuales del documento respaldatorio

Solapas

El tablero cuenta con 7 solapas :

- **Portada**
- **Glosario**
- **Resumen**
- **Análisis por Cirugía**
- **Análisis por Estadio**
- **Análisis por Histología**
- **Análisis Valores Proteínas**

A continuación se explicara el detalle de cada una de las solapas

PORTADA La portada del tablero presenta una imagen que inspiró la paleta de colores utilizada, así como los botones de navegación que permiten acceder a las distintas páginas. Además, se incluye un homenaje especial a mi madre, una sobreviviente del cáncer de mama.

La frase destacada busca captar la atención de los "espectadores" al sumergirlos en la historia detrás de los datos, generando concientización y esperanza mediante la aplicación del concepto de historytelling.

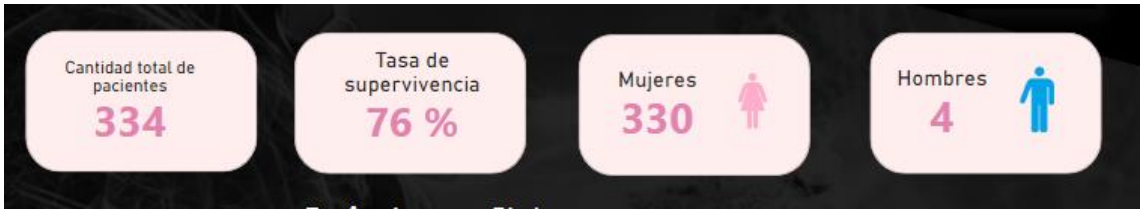
GLOSARIO Se ha elaborado un glosario que explica algunos términos utilizados en el tablero con el fin de facilitar su comprensión, incluso para aquellas personas que no están familiarizadas con el tema o están explorando este campo por primera vez.

Se realizado una definición funcional y conceptual de los conceptos, ya que son termimos bastante específicos. Este recurso tiene como objetivo proporcionar claridad sobre los conceptos técnicos y específicos empleados en el tablero, lo que contribuye a

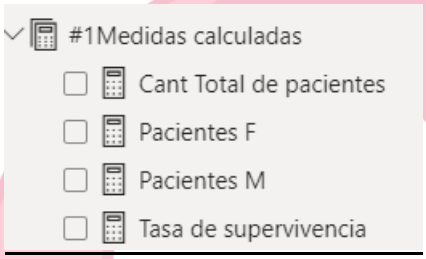
una experiencia más accesible y enriquecedora para todos los usuarios

RESUMEN

Esta solapa del dashboard, se muestra la cantidad de pacientes, cantidad de hombres y mujeres y la tasa de supervivencia, en este caso total ,con un grafico de anillo se representa la cantidad de pacientes por status



Para la creación de estos indicadores, se utilizaron medidas calculadas, las mismas fueron creadas en un tabla aparte a partir de formulas DAX



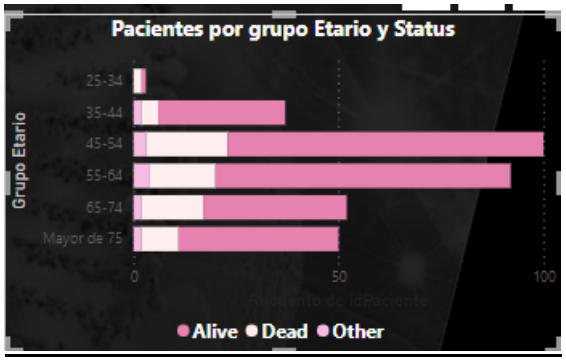

```
Cant Total de pacientes = (DISTINCTCOUNT(pacientes[idPaciente]))

Pacientes F = calculate(DISTINCTCOUNT(pacientes[idPaciente]), pacientes[genero] =
"Female" )

Pacientes M = if (calculate(DISTINCTCOUNT(pacientes[idPaciente]), pacientes[genero] =
"Male" ) = 0, "--", calculate(DISTINCTCOUNT(pacientes[idPaciente]), pacientes[genero]
= "Male" ))
```

```
Tasa de supervivencia = if
(DIVIDE(CALCULATE(DISTINCTCOUNT(pacientes[idPaciente]),pacientes[idPacienteStatus]=1)
,
DISTINCTCOUNT(pacientes[idPaciente]),0 )=0, "--",
(DIVIDE(CALCULATE(DISTINCTCOUNT(pacientes[idPaciente]),pacientes[idPacienteStatus]=1)
,
DISTINCTCOUNT(pacientes[idPaciente]),0 )))
```

En este resumen se muestra también se muestra también los pacientes por Grupo etario y por status



Para la confección de esta grafico , se añadió una columna en la tabla pacientes para poder agrupar a los pacientes en base a su edad y así facilitar su analisis

```
GrupoEtario =
    IF(
        pacientes[edad] < 25, "Menor de 25",
        IF(
            pacientes[edad] >= 25 && pacientes[edad] < 35, "25-34",
            IF(
                pacientes[edad] >= 35 && pacientes[edad] < 45, "35-44",
                IF(
                    pacientes[edad] >= 45 && pacientes[edad] < 55, "45-54",
                    IF(
                        pacientes[edad] >= 55 && pacientes[edad] < 65, "55-64",
                        IF(
                            pacientes[edad] >= 65 && pacientes[edad] < 75, "65-74",
                            "Mayor de 75"
                        )
                    )
                )
            )
        )
    )
```

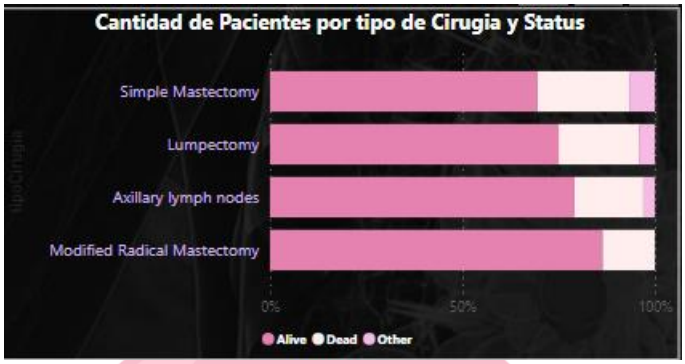
Utilizando la tabla calendario, se coloco un segmentador, para poder filtrar el informe por año, el mismo esta disponible en todas las paginas del informe, al igual que la tasa de supervivencia que va cambiando en relación a los filtros aplicados

ANALISIS POR CIRUGIA

Se presenta con un grafico de anillo, las distintas cirugías realizadas indicándolas en porcentaje

Mediante el segmentador podrá visualizarse por cirugía y/o por año indicando la tasa de supervivencia según la información filtrada

Se muestra en un grafico de barras apiladas al 100% la relación entre los tipos de cirugías con el Status de los pacientes



ANALISIS POR ESTADIO

Se presenta con un grafico de anillo, los tres estadios de detección del tumor

Mediante el segmentador podrá visualizarse por estadio y/o por año indicando la tasa de supervivencia según la información filtrada

Se muestra en un grafico de barras apiladas la relación entre cantidad de pacientes con el estatus y estadio

ANALISIS POR HISTOLOGIA

Se presenta con un grafico de anillo, los distintos tipos de Histología del tumor

Mediante el segmentador podrá visualizarse por Histologia y/o por año indicando la tasa de supervivencia según la información filtrada

Se muestra en un grafico de columnas apiladas la relación entre Histoliga del tumor con el grupo etario

En esta pagina, se uso tonos de violeta para diferenciar de las anteriores donde los rosas se usaban para status

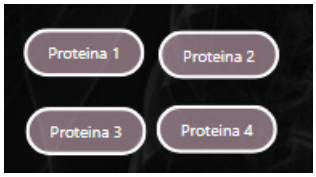
ANALISIS VALORES DE PROTEINAS

En esta ultima pagina del tablero, se utilizaron los valores de las expresiones de Proteínas, estas proteínas se desconoce cuales son ya que no están identifiacas en el Dataset

Se calcularon las Medias y Medianas de cada Proteína y se realizaron histogramas que muestran el valor de expresión que obtuvo en su análisis cada paciente

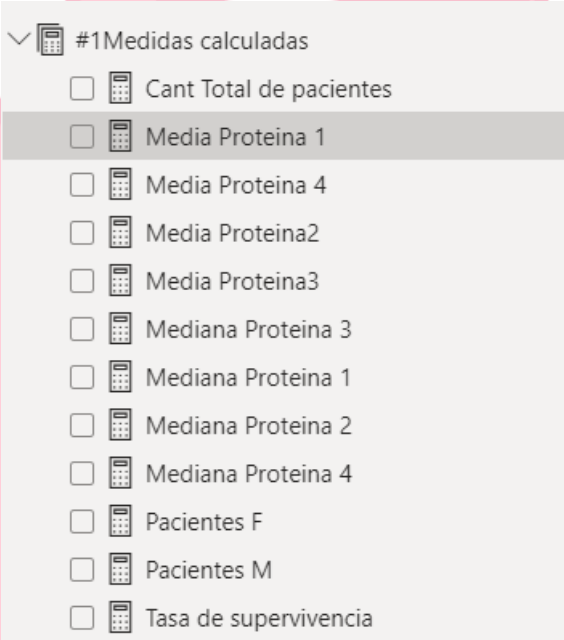
Es posible segmentar por estatus y/o histología del tumor. Por ejemplo, en la proteína 2, al seleccionar el tipo de tumor "Infiltrating Ductal Carcinoma" y el estatus "Dead", se observa una mediana de 13.726, un valor significativamente más alto que cuando se selecciona el estatus "Alive", donde el valor de la proteína 2 es de 0.9. El objetivo es analizar el comportamiento estadístico en función de diferentes variables. Quizás, con una mayor cantidad de observaciones y la inclusión de algunos otros conceptos estadísticos, este análisis podría utilizarse como punto de partida para un modelo de predicción.

A través de Marcadores se puede ir mostrando los valores y gráficos para cada una de las Proteínas, se eligieron distintos colores tanto para la líneas como para las fuentes para lograr un mejor efecto visual de cambio de Proteína seleccionada



CALCULO DE MEDIAS Y MEDIANAS

Las Medias y medianas fueron incluidas en la tabla de Medidas Calculadas y calculadas mediante DAX



```
1  Media Proteina 1 = AVERAGE(valorProteinas[proteina1])
```

```
1  Mediana Proteina 2 = MEDIAN(valorProteinas[proteina2])
```