

Tercer Parcial de Minería de Datos

Andrés Felipe Barrera Muñoz

Universidad de Investigación y Desarrollo (San Gil)

Ingeniería de Sistemas

Minería de Datos

2023

Contenido

Resumen	3
Objetivo	3
Características y Funcionalidades	3
Introducción	3
Desarrollo y Diseño.....	5
Base de Datos.....	8
Consultas y análisis mediante el aplicativo	9
Primer Punto	9
Segundo Punto.....	12
Tercer Punto.....	14
Cuarto Punto	17

Resumen

El aplicativo web desarrollado es una herramienta que permite realizar análisis de frecuencia en archivos CSV. CSV, que significa "Comma Separated Values", es un formato de archivo comúnmente utilizado para almacenar datos tabulares, donde cada línea representa una fila y los valores están separados por comas.

Este programa web mejora la eficiencia y la facilidad de análisis de frecuencia en conjuntos de datos CSV. Los usuarios pueden cargar archivos CSV desde su computadora. Una vez cargado el archivo, el aplicativo procesa los datos y muestra una visualización clara y concisa de la frecuencia de los valores de la columna seleccionada del archivo.

La interfaz del aplicativo es intuitiva y fácil de usar, lo que permite a los usuarios navegar sin problemas y obtener rápidamente los resultados deseados. También se brinda la opción de exportar los resultados del análisis a un archivo PDF, lo que facilita la posterior visualización de los datos.

Objetivo

Analizar datos generando su respectiva tabla de frecuencia y grafica obtenidos desde un .CSV además de guardar dicho análisis.

Características y Funcionalidades

Características	Funcionalidades
Importación de archivos CSV	Interfaz de usuario intuitiva
Análisis de frecuencia	Procesamiento de datos
Selección de columnas	Validación de datos
Visualización de resultados	Gráficos interactivos
Exportación de resultados	Funcionalidad de guardar y cargar

Introducción

Este informe proporciona una descripción detallada del desarrollo realizado para la materia de Minería de Datos, centrándose en el funcionamiento y la estructura del aplicativo

web desarrollado. El objetivo principal es entregar el aplicativo como parte del último parcial de la materia, junto con algunas consultas realizadas en la base de datos GSOASIS.GDB.

El aplicativo web ha sido desarrollado utilizando HTML, CSS, JavaScript y PHP como lenguajes de programación. Estos lenguajes se utilizaron para crear la interfaz del aplicativo, aplicar estilos y efectos visuales, y manejar la lógica y el procesamiento de datos. Además, se integró phpMyAdmin como herramienta de administración de la base de datos.

El funcionamiento del aplicativo web consiste en permitir al usuario cargar un archivo CSV desde su dispositivo. Una vez cargado el archivo, el sistema procesa los datos y genera una tabla de frecuencia que muestra la cantidad de veces que aparece cada valor en el conjunto de datos. Además, se genera una gráfica correspondiente a la distribución de los valores en el archivo.

La estructura del aplicativo se basa en diferentes archivos y componentes. El archivo HTML define la estructura básica de la página y proporciona contenedores para mostrar los resultados del análisis. El CSS se utiliza para aplicar estilos y personalizar la apariencia del aplicativo, mientras que el JavaScript se encarga de manipular los elementos de la página y realizar las operaciones de análisis de frecuencia.

El lenguaje de programación PHP se utiliza en el aplicativo para manejar la lógica del lado del servidor. Permite procesar la carga del archivo CSV, leer y analizar los datos contenidos en él, y generar la tabla de frecuencia y la gráfica correspondiente.

Además de los lenguajes de programación utilizados en el desarrollo del aplicativo web, también se ha incorporado Python como una herramienta adicional para el tratamiento de los datos generados en el IBConsole. Python, conocido por su amplia variedad de bibliotecas y su flexibilidad en el análisis de datos, ha desempeñado un papel fundamental en el procesamiento de los datos extraídos de la base de datos GSOASIS.GDB. Mediante el uso de bibliotecas en este caso csv, datetime y tkinter, se ha logrado realizar tareas como la

limpieza, manipulación y transformación de los datos. La integración de Python en el flujo de trabajo del aplicativo web ha proporcionado una capacidad adicional para trabajar mejor los datos obtenidos del IBConsole, permitiendo el análisis en el aplicativo.

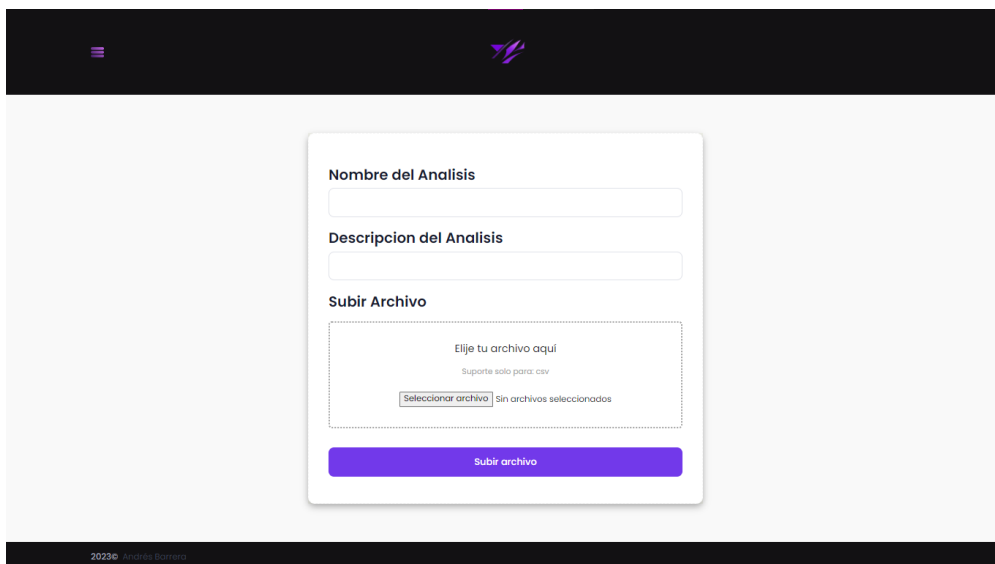
En cuanto a las consultas realizadas en la base de datos GSOASIS.GDB, se encuentran en proceso de finalización y se incluirán en el documento entregable como parte del parcial. Estas consultas permiten obtener información relevante de la base de datos y se integran en el aplicativo web para enriquecer el análisis de datos.

Desarrollo y Diseño

En cuanto al diseño del aplicativo web, se utilizó una plantilla como punto de partida, la cual fue modificada y adaptada según las necesidades específicas del proyecto. Se realizaron ajustes y personalizaciones para asegurar una interfaz coherente y atractiva.

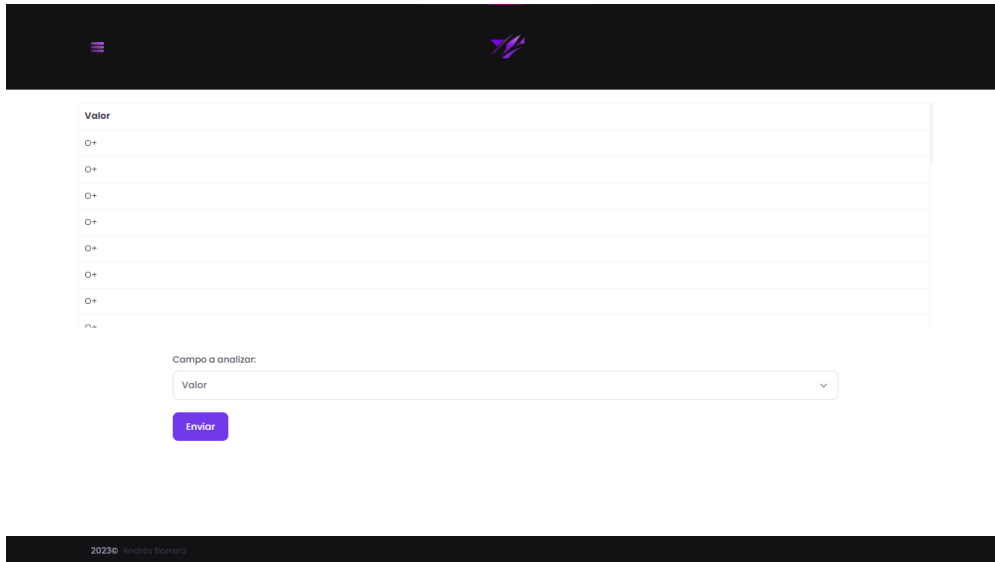
El diseño final consta de un total de 5 pantallas, cada una diseñada con el objetivo de brindar una experiencia intuitiva y fácil de usar para los usuarios. A continuación, se describen brevemente las pantallas desarrolladas:

1. Pantalla de carga de archivo: Esta pantalla permite al usuario seleccionar y cargar un archivo CSV desde su dispositivo.



The screenshot displays a web application interface for uploading a file. At the top, there is a dark header bar containing a hamburger menu icon on the left and a logo on the right. The main content area is light gray and features a white card with rounded corners. Inside the card, there are three sections: 'Nombre del Analisis' with a text input field, 'Descripcion del Analisis' with a text input field, and 'Subir Archivo'. The 'Subir Archivo' section includes a dashed border box with the text 'Elige tu archivo aqui' and 'Soporta solo para: csv'. Below this box is a button labeled 'Seleccionar archivo' and the text 'Sin archivos seleccionados'. At the bottom of the card is a large blue button labeled 'Subir archivo'. The footer of the application is a dark bar with the text '2023 © Andrius Borrero' on the left.

2. Pantalla de elección de columna a analizar: Una vez cargado el archivo CSV, esta pantalla muestra una lista de las columnas disponibles en el archivo. El usuario puede seleccionar la columna que desea analizar para generar la tabla de frecuencia y la gráfica correspondiente.



Valor

O+

O+

O+

O+

O+

O+

O+

O+

O+

O+

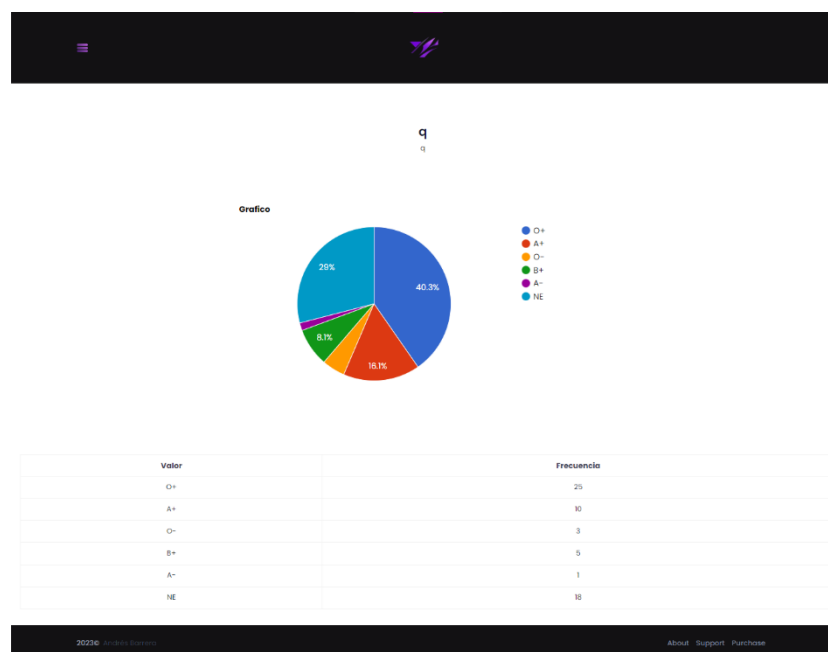
Campo a analizar:

Valor

Enviar

2023 © Andrés Barranto

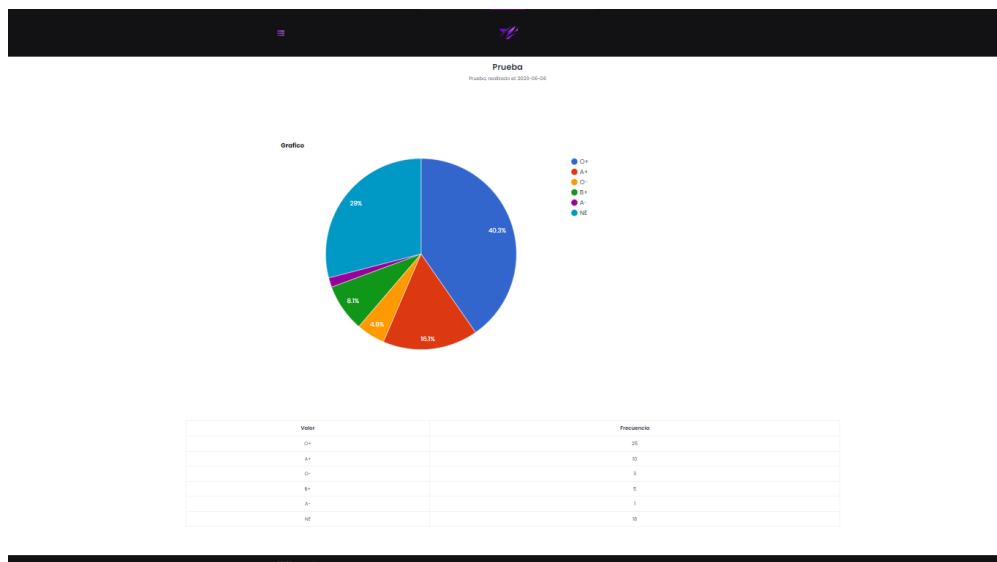
3. Pantalla de resultados de análisis: Después de seleccionar la columna a analizar, esta pantalla muestra los resultados del análisis en forma de tabla de frecuencia y gráfica. Proporciona una visualización clara y concisa de la distribución de los valores en la columna seleccionada.



4. Pantalla de análisis guardados: En esta pantalla, se muestran los análisis previamente guardados. Permite acceder a los análisis anteriores y consultar sus resultados en cualquier momento.

Resultados de Analisis				
Nombre	Descripción	Fecha	Cargar	Eliminar
Hombres y Mujeres	Cantidad de Hombres y Mujeres que generan Citas Médicas	2023-05-28	Cargar	Eliminar
Citas por Centro de Costo	Cantidad de citas por centro de costo	2023-05-28	Cargar	Eliminar
Ingresos por Centro de Costo	Cantidad de ingresos por centro de costo	2023-05-28	Cargar	Eliminar
Embarazadas y No Embarazadas	Cantidad de mujeres embarazadas y no embarazadas	2023-05-28	Cargar	Eliminar
Grupo Sanguíneo de Embarazadas	Cantidad por grupo sanguíneo de todas las embarazadas	2023-05-28	Cargar	Eliminar
Estado Civil de Embarazadas	Estado civil de todas la embarazadas registradas	2023-05-29	Cargar	Eliminar
Edades de Embarazadas	Edades agrupadas de las embarazadas registradas	2023-05-29	Cargar	Eliminar

5. Pantalla de carga de análisis antiguos: Esta pantalla permite al usuario cargar análisis previamente guardados desde el sistema. Proporciona una lista de los análisis guardados para que el usuario pueda seleccionar el que desea cargar y revisar nuevamente.

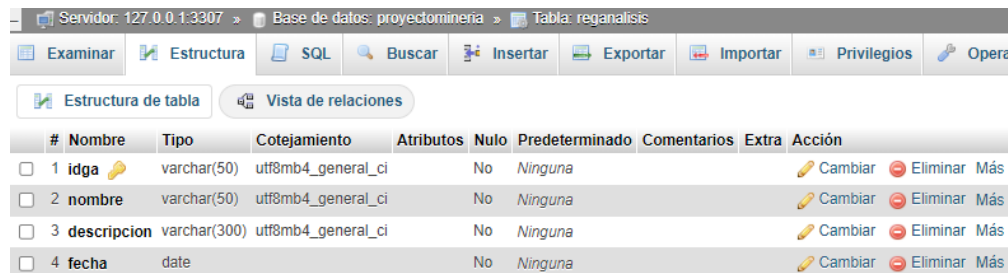


En cuanto al desarrollo del aplicativo web, todos los archivos relacionados se encuentran disponibles en mi repositorio personal de GitHub. Enlace del repositorio:

<https://github.com/AndresD4rk/mineryproyect.git>

Base de Datos

En el apartado de la base de datos, se utiliza una tabla específica con el propósito de guardar los datos del análisis realizado. Esta tabla consta de cuatro columnas principales: "idga" (identificador del archivo) como llave primaria, "nombre" (nombre del análisis), "descripcion" (descripción del análisis) y "fecha" (fecha de realización del análisis).



#	Nombre	Tipo	Cotejamiento	Atributos	Nulo	Predeterminado	Comentarios	Extra	Acción
<input type="checkbox"/> 1	idga	varchar(50)	utf8mb4_general_ci		No	Ninguna			Cambiar Eliminar Más
<input type="checkbox"/> 2	nombre	varchar(50)	utf8mb4_general_ci		No	Ninguna			Cambiar Eliminar Más
<input type="checkbox"/> 3	descripcion	varchar(300)	utf8mb4_general_ci		No	Ninguna			Cambiar Eliminar Más
<input type="checkbox"/> 4	fecha	date			No	Ninguna			Cambiar Eliminar Más

La columna "idga" se utiliza como identificador único para cada registro de análisis y se configura como llave primaria para garantizar la unicidad de los datos almacenados. Esto permite un fácil acceso y referencia a cada análisis individual.

La columna "nombre" almacena el nombre asignado al análisis, que sirve como una etiqueta descriptiva para identificarlo de manera más rápida y sencilla.

La columna "descripcion" proporciona una breve explicación o resumen del análisis realizado. Este campo puede ser utilizado para proporcionar detalles adicionales sobre los parámetros utilizados, los objetivos del análisis o cualquier otra información relevante.

Finalmente, la columna "fecha" registra la fecha en la que se realizó el análisis. Esto permite llevar un registro de cuándo se realizó cada análisis, lo que puede ser útil para la posterior revisión y seguimiento de los resultados a lo largo del tiempo.

A continuación, se presentan los datos almacenados como resultado del análisis realizado en relación a las consultas planteadas para el último parcial de la materia de Minería de Datos.

idga	nombre	descripcion	fecha
64738de7a41dc	Hombres y Mujeres	Cantidad de Hombres y Mujeres que generan Citas Mé...	2023-05-28
647394b53fb7c	Citas por Centro de Costo	Cantidad de citas por centro de costo	2023-05-28
647398ae7dca1	Ingresos por Centro de Costo	Cantidad de ingresos por centro de costo	2023-05-28
6473c6678ac55	Embarazadas y No Embarazadas	Cantidad de mujeres embarazadas y no embarazadas	2023-05-28
6473ca67236a8	Grupo Sanguíneo de Embarazadas	Cantidad por grupo sanguíneo de todas las embaraza...	2023-05-28
6473d0c14170b	Estado Civil de Embarazadas	Estado civil de todas la embarazadas registradas	2023-05-29
6473d62c5a78d	Edades de Embarazadas	Edades agrupadas de las embarazadas registradas	2023-05-29

Consultas y análisis mediante el aplicativo

En esta sección se presenta el desarrollo de las consultas realizadas para el parcial utilizando el aplicativo para generar el análisis de las mismas.

Primer Punto

Cantidad de Hombres y Mujeres que generan Citas Médicas. Para esto se planteó inicialmente la siguiente consulta

```
Select b.sexo from cuentaingreso a
inner join usuarios b on b.numdocumento=a.carnet
```

SEXO
F
M
F
F
F
F
F
F

En esta consulta modificada, además de seleccionar el campo "sexo" de la tabla "usuarios", se utiliza la función de agregación COUNT(*) para contar la cantidad de registros que cumplen con la condición de la consulta. Luego, se utiliza la cláusula GROUP BY para agrupar los resultados según el campo "sexo". Esto por en problema al intentar exportar todos los datos de la anterior forma.

```
Select count(u.sexo),u.sexo from cuentaingreso c
inner join usuarios u on u.numdocumento=c.carnet
group by u.sexo
```

COUNT	SEXO
197777	F
111908	M
7	f
3	m

Luego esta información se pasa a un Excel en este caso con el fin de rellenar los datos para su posterior análisis.

Valor
M
M
M
M
M
M
M
M
M
M
M
M
M
M
M

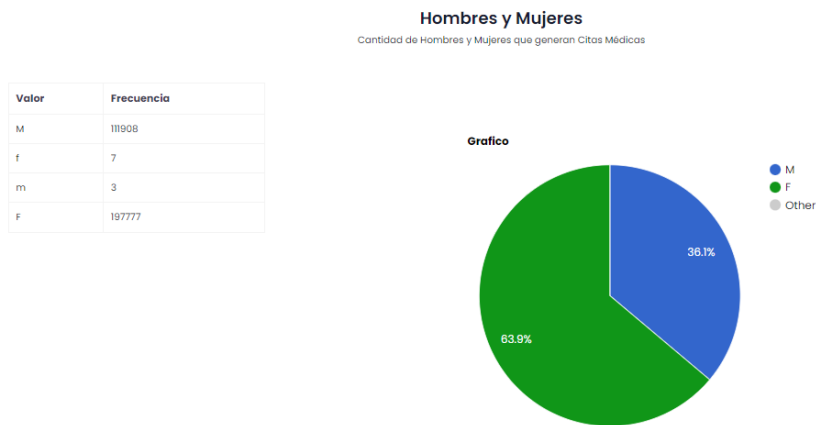
Ahora pasamos a la aplicación en esta se pone el nombre que se le desea dar al análisis, una descripción a este mismo y el archivo con los datos para el análisis.

The screenshot shows a web application interface with a dark header and footer. The main content area is light gray. A white card in the center contains the following elements:

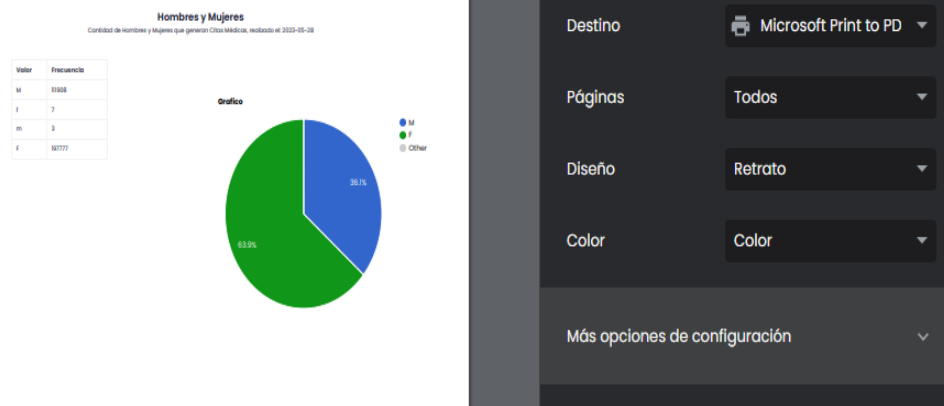
- Nombre del Analisis:** A text input field containing "Hombres y Mujeres".
- Descripcion del Analisis:** A text input field containing "Cantidad de Hombres y Mujeres que generan Citas Médicas".
- Subir Archivo:** A section with a dashed border containing:
 - The text "Elige tu archivo aqui".
 - The text "Soporte solo para: csv".
 - A button labeled "Seleccionar archivo" next to the filename "Hymporcitas.csv".
- A large blue button at the bottom labeled "Subir archivo".

The footer contains the text "2023 © Anedat" on the left and "About Support Purchase" on the right.

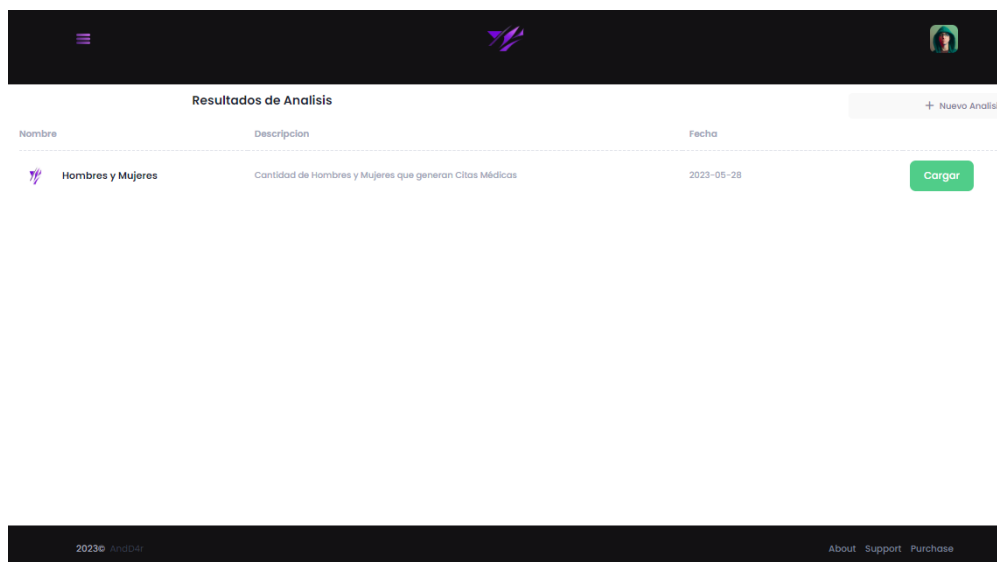
Dando como resultado la siguiente gráfica y tabla de frecuencia.



Esta información también se puede guardar como PDF al pulsar la letra “P”.



También se puede revisar nuevamente en el aplicativo.



Segundo Punto

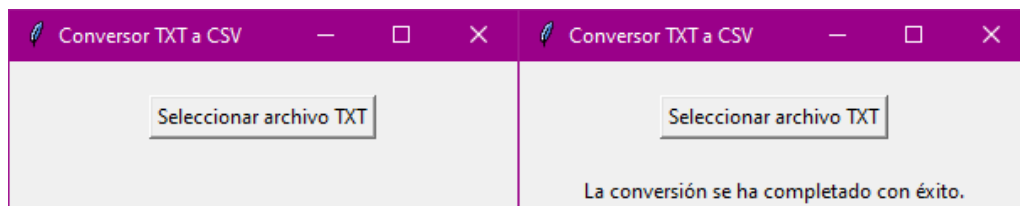
Cantidad de citas por centro de costo. En esta se necesita la cantidad de citas por centro costos dando como resultado la siguiente consulta.

```
select b.descripcion from citas a
inner join centrocostos b on
b.codigo=a.codigo costo
```

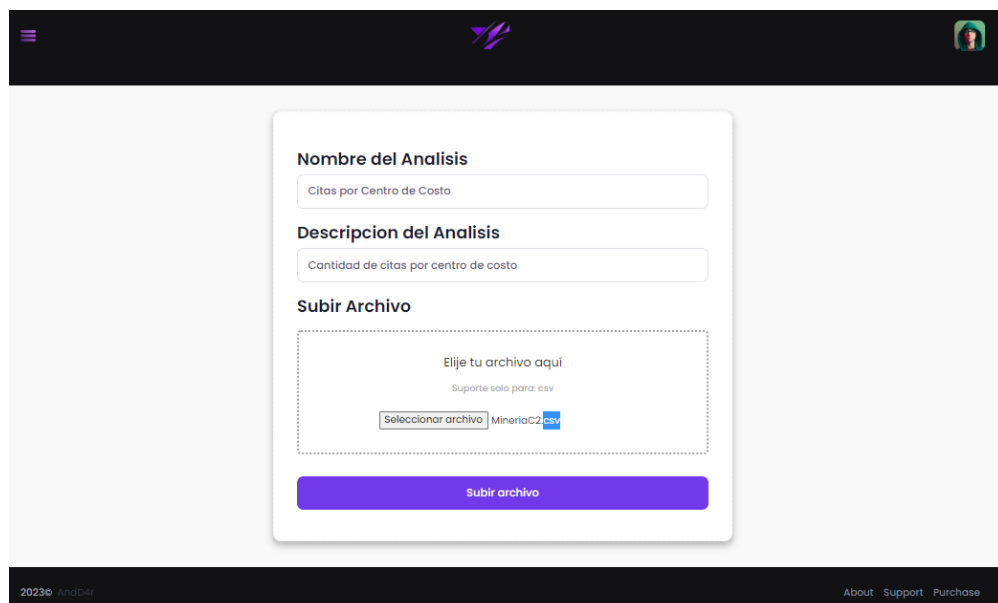
|

DESCRIPCION
PROMOCION Y PREVENCIÓN
PROMOCION Y PREVENCIÓN
PROMOCION Y PREVENCIÓN
PROMOCION Y PREVENCIÓN
PROMOCION Y PREVENCIÓN
PROMOCION Y PREVENCIÓN
PROMOCION Y PREVENCIÓN

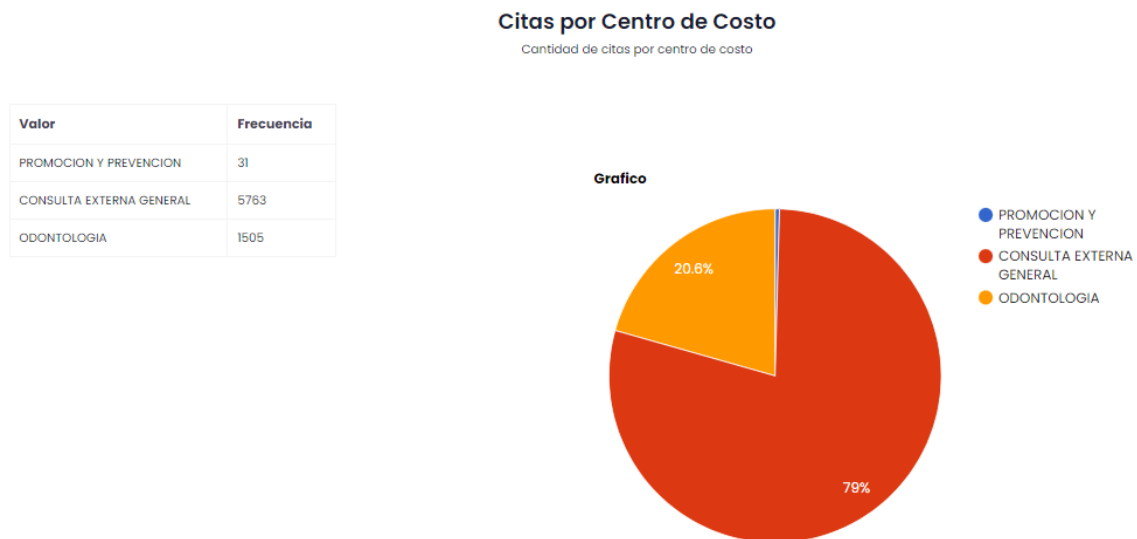
Se exportan los datos y se ingresan a un convertidor de TXT a CSV para poder ingresarlo en la aplicación.



Se revisa que no haya errores en la conversión para proseguir con el análisis. Y se ingresan todos los datos al aplicativo.



Y nos genera el análisis pertinente.



Además de permitir el mismo proceso de pulsar “P” para imprimir el informe

28/5/23, 12:52 Metronic - The World's #1 Selling Bootstrap Admin Template by KeenThemes

Citas por Centro de Costo
Cantidad de citas por centro de costo

Valor	Frecuencia
PROMOCION Y PREVENCIÓN	31
CONSULTA EXTERNA GENERAL	5763
ODONTOLOGIA	1505

Grafico

Legend:

- PROMOCION Y PREVENCIÓN
- CONSULTA EXTERNA GENERAL
- ODONTOLOGIA

Imprimir 1 hoja de papel

Destino:

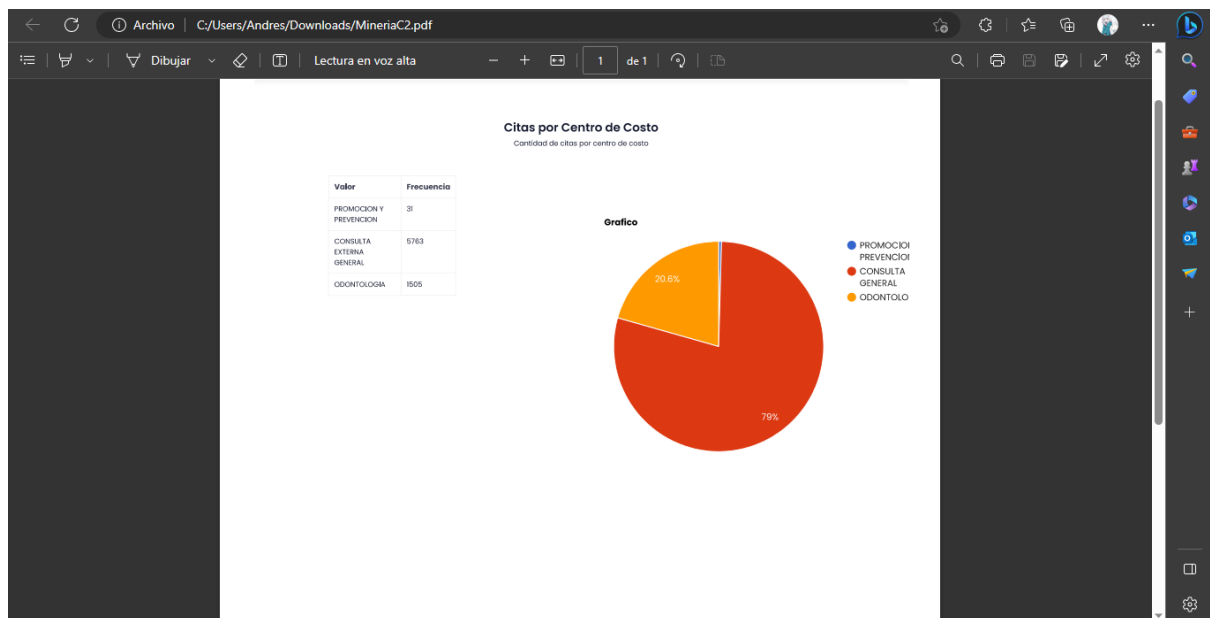
Páginas:

Diseño:

Color:

Más opciones de configuración

localhost:myproyecto.php 1/1



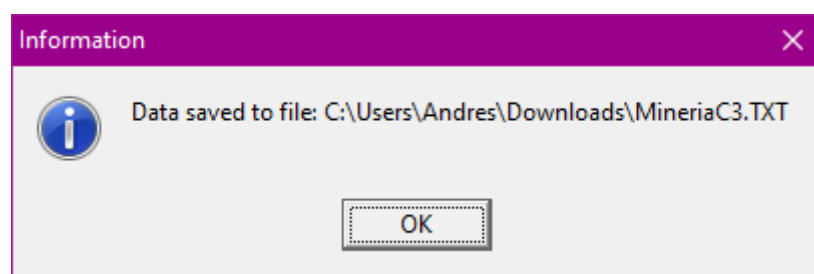
Tercer Punto

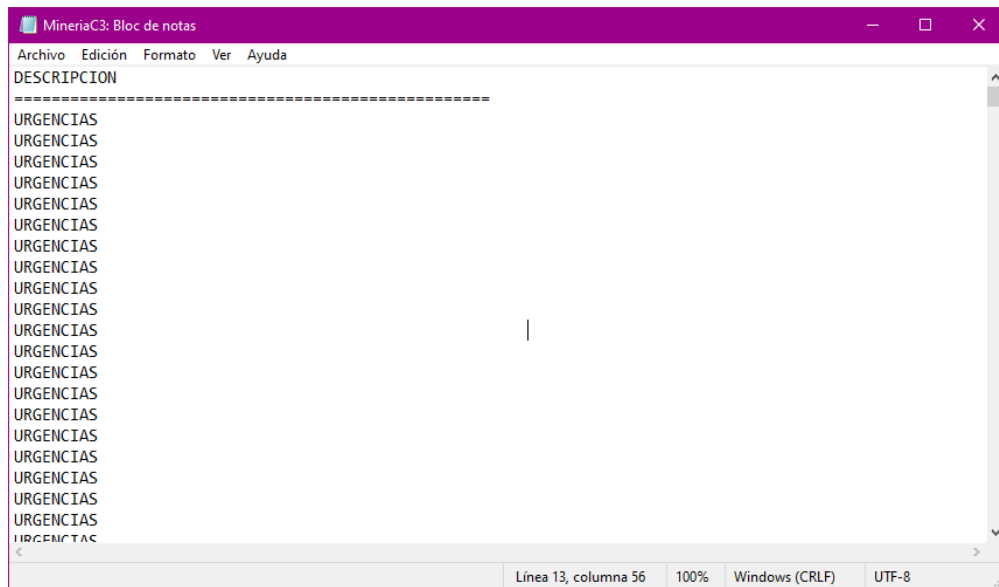
Ingresos por centro de costo. Para esta es un proceso igual al anterior, pero se genera la consulta para sacar los ingresos por centro de costos.

```
select c.descripcion from ingreso i
inner join centros costos c on
c.codigo=i.centrocosto
```

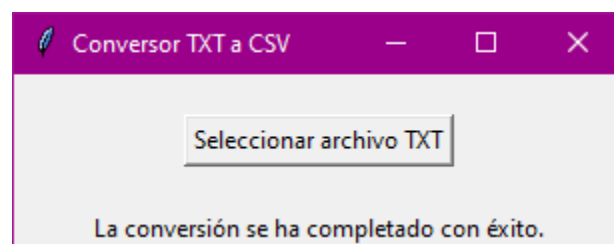
DESCRIPCION
URGENCIAS
URGENCIAS
URGENCIAS
URGENCIAS
URGENCIAS

Se guardan los datos y se quita el separador que se ve que la posterior imagen a la continua.





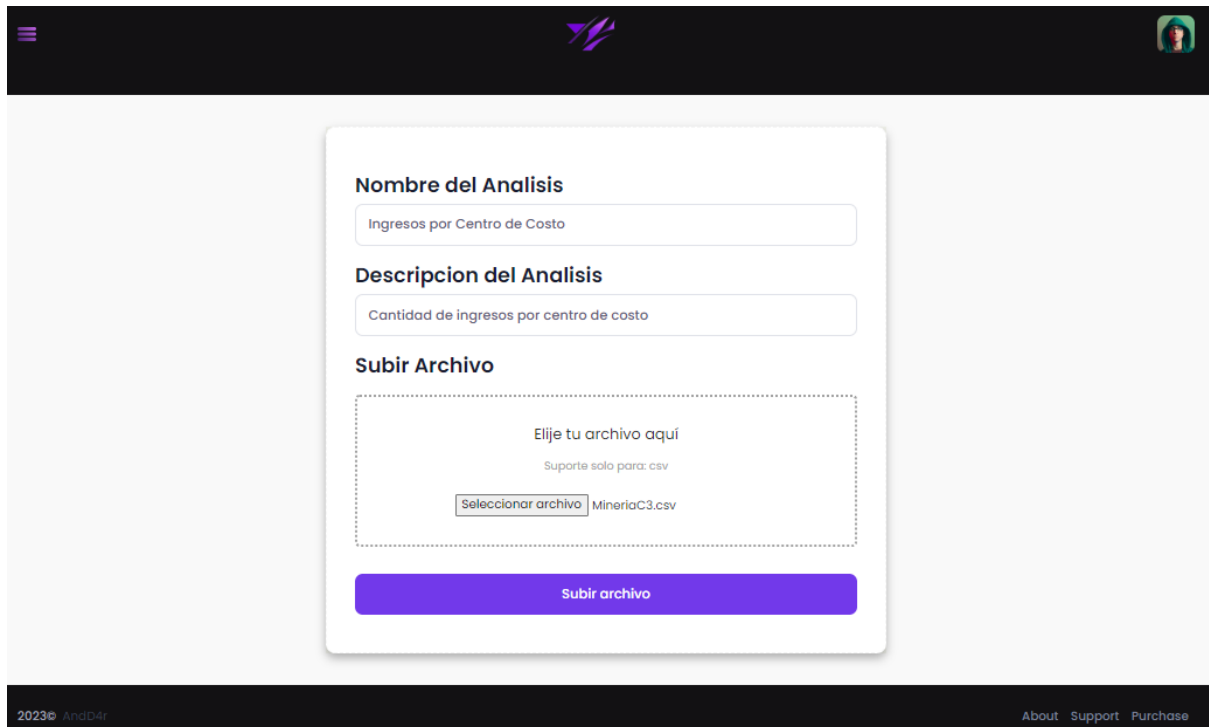
Se realiza la conversión.



Se revisan los datos.

[illegible]

Se usa el aplicativo para subir el análisis.



Nombre del Analisis

Ingresos por Centro de Costo

Descripcion del Analisis

Cantidad de ingresos por centro de costo

Subir Archivo

Elije tu archivo aquí

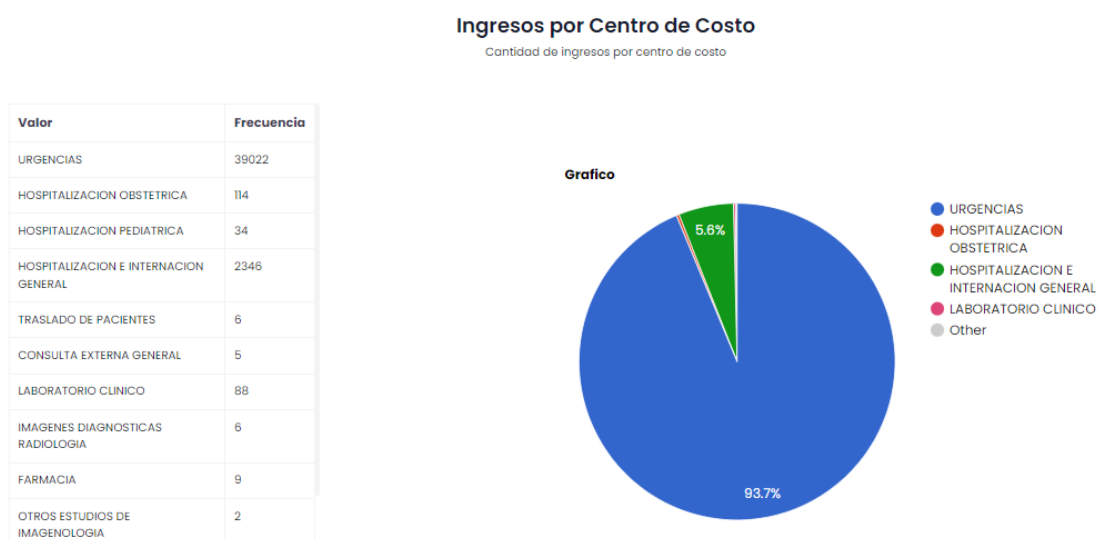
Soporte solo para: csv

Seleccionar archivo MineriaC3.csv

Subir archivo

2023 © AndD4r About Support Purchase

Y se aprecia el resultado del análisis.



Cuarto Punto

Generar un Informe de mujeres embarazadas a la fecha haciendo un análisis poblacional de mujeres en la BASE DE DATOS. Para el último punto se analizan diferentes factores en las mujeres embarazadas. El primer análisis es la cantidad de mujeres embarazadas frente a las no embarazadas en la tabla usuarios.

```
SELECT
  (SELECT COUNT(*) FROM usuarios u
   INNER JOIN embarazada e ON
   e.idhistoria = u.idhistoria
   WHERE u.sexo = 'F') AS mujeres_embarazadas,
  (SELECT COUNT(*) FROM usuarios u
   LEFT JOIN embarazada e ON
   e.idhistoria = u.idhistoria
   WHERE u.sexo = 'F' AND e.idhistoria IS NULL) AS mujeres_no_embarazadas
FROM
  rdb$database;
```

<	1: 1	Client dialect 3	Transaction is ACTIVE.	AutoDDL: ON
Data	Plan	Statistics		
	MUJERES_EMBARAZADAS	MUJERES_NO_EMBARAZADAS		
▶	62	15825		

Pasamos la consulta a un repetidor el cual exporta a CSV.

Exportador CSV

Valor:

Repeticiones:

Agregar

EMBARAZADAS: 62
NO EMBARAZADAS: 15825

Eliminar

Exportar a CSV

Se revisa el resultado.

Valor
EMBARAZADAS
EMBARAZADAS
EMBARAZADAS
EMBARAZADAS
EMBARAZADAS
EMBARAZADAS
EMBARAZADAS

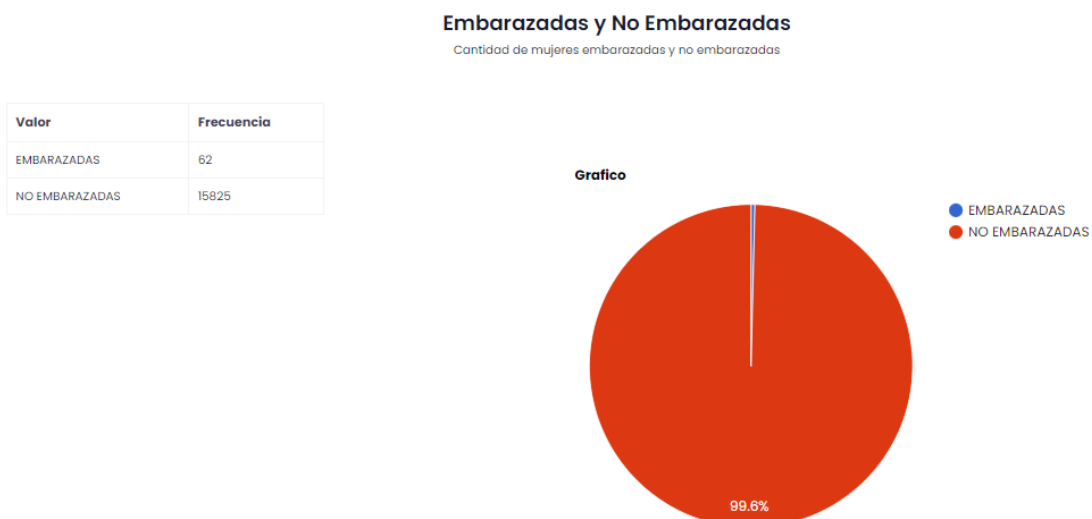
Y se procede al análisis en el aplicativo.

The screenshot shows a web application interface with a dark header. The main content area is light gray. A white card in the center contains the following sections:

- Nombre del Analisis:** A text input field containing "Embarazadas y No Embarazadas".
- Descripcion del Analisis:** A text input field containing "Cantidad de mujeres embarazadas y no embarazadas".
- Subir Archivo:** A section with a dashed border containing the text "Elige tu archivo aquí" and "Suporte solo para: csv". Below this is a button labeled "Seleccionar archivo" and a file name "MineriaC4v1.csv".
- A large blue button at the bottom labeled "Subir archivo".

The footer of the application shows "2023 © AndD4r" on the left and "About Support Purchase" on the right.

Se observa los resultados de este análisis.

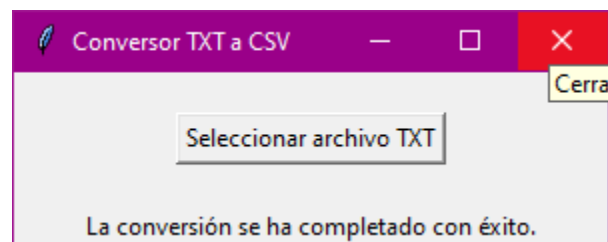


Ahora se pasa a analizar el grupo sanguíneo de las embarazadas mediante la siguiente consulta.

Select c11 from embarazada

<	1: 1	Client d
Data	Plan	Statistics

C11
O +
A +
A+
O +
O -



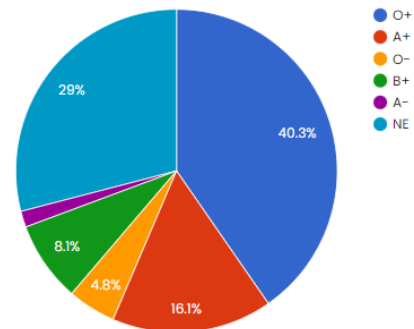
Se usa una herramienta que permite la repetición de valores y genera un CSV directamente incluyendo los datos de la consulta anterior pero agrupada.

Grupo Sanguíneo de Embarazadas

Cantidad por grupo sanguíneo de todas las embarazadas

Valor	Frecuencia
O+	25
A+	10
O-	3
B+	5
A-	1
NE	18

Grafico

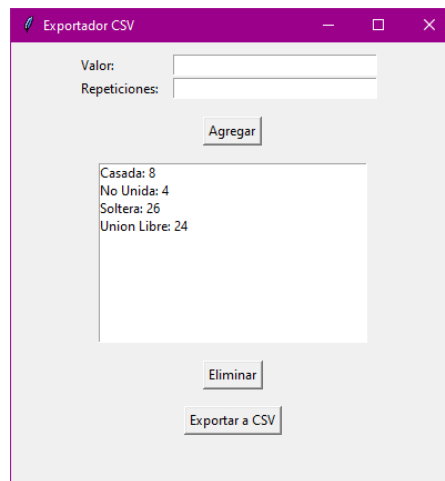


Ahora el estado civil de las embarazadas mediante la siguiente consulta.

```
SELECT count(a.estadocivil),a.estadocivil
FROM USUARIOS a
INNER JOIN EMBARAZADA b ON b.IDHISTORIA = a.IDHISTORIA
group by a.estadocivil
```

<		
5: 1	Modified	Client dialect 3 Transaction is ACTIVE. Aut
Data	Plan	Statistics
COUNT	ESTADOCIVIL	
8	C	
4	N	
26	S	
24	U	

Se usa la herramienta mostrada anteriormente para realizar el mismo proceso con los datos generados en la consulta y poder cambiar la etiqueta que lleva.



Exportador CSV

Valor:

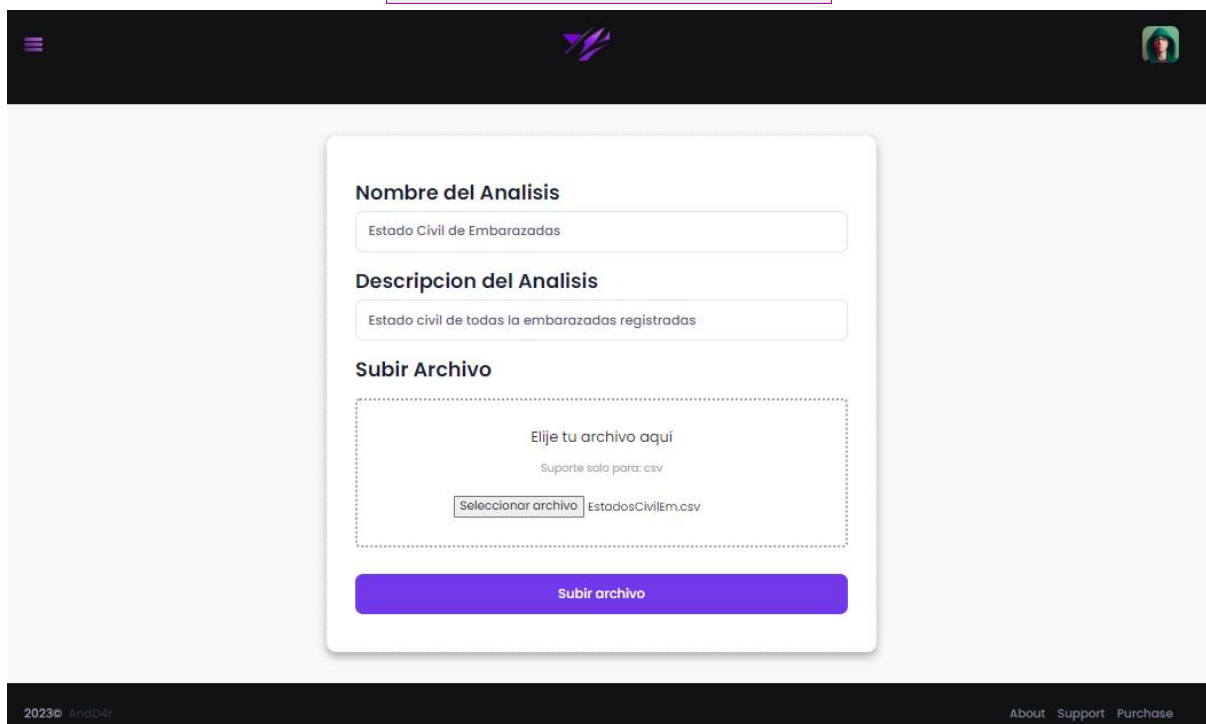
Repeticiones:

Agregar

Casada: 8
No Unida: 4
Soltera: 26
Union Libre: 24

Eliminar

Exportar a CSV



Nombre del Analisis

Estado Civil de Embarazadas

Descripcion del Analisis

Estado civil de todas la embarazadas registradas

Subir Archivo

Elige tu archivo aqui

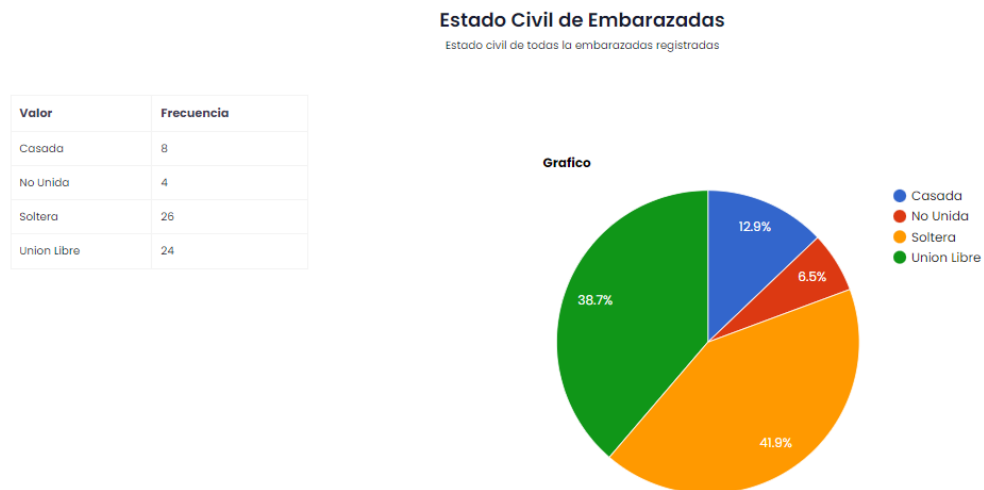
Soporte solo para: csv

Seleccionar archivo EstadosCivilem.csv

Subir archivo

2023 © And4r

About Support Purchase



Y por último análisis las edades de las embarazadas sacando las fechas de nacimiento de las embarazadas con la siguiente consulta.

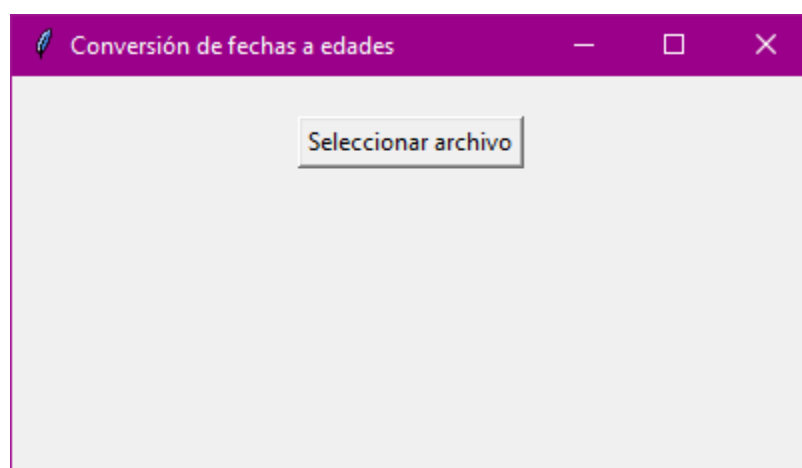
```
SELECT u.fechanacimiento
FROM usuarios u
INNER JOIN embarazada e ON e.idhistoria = u.idhistoria
```

1: 1 Client dialect 3 Transaction is ACTIVE. AutoDDL




Data Plan Statistics

FECHANACIMIENTO
12/12/1993
26/06/1996
2/04/1995
25/12/1986
27/01/1995

Se usa un conversor de fechas a edades el cual en este caso exporta todo a un CSV.



Se ingresan los datos en el aplicativo.



Nombre del Analisis
Edades de Embarazadas

Descripcion del Analisis
Edades agrupadas de las embarazadas registradas

Subir Archivo

Elije tu archivo aqui
Soporte solo para: csv




Seleccionar archivo

Edades embarazadas.csv

Subir archivo

2023© AndD4r

About Support Purchase



Edad	FECHANACIMIENTO
29	12/12/1993
26	26/06/1996
28	2/04/1995
36	25/12/1986
28	27/01/1995
27	29/03/1996
27	30/01/1996

Campo a analizar:

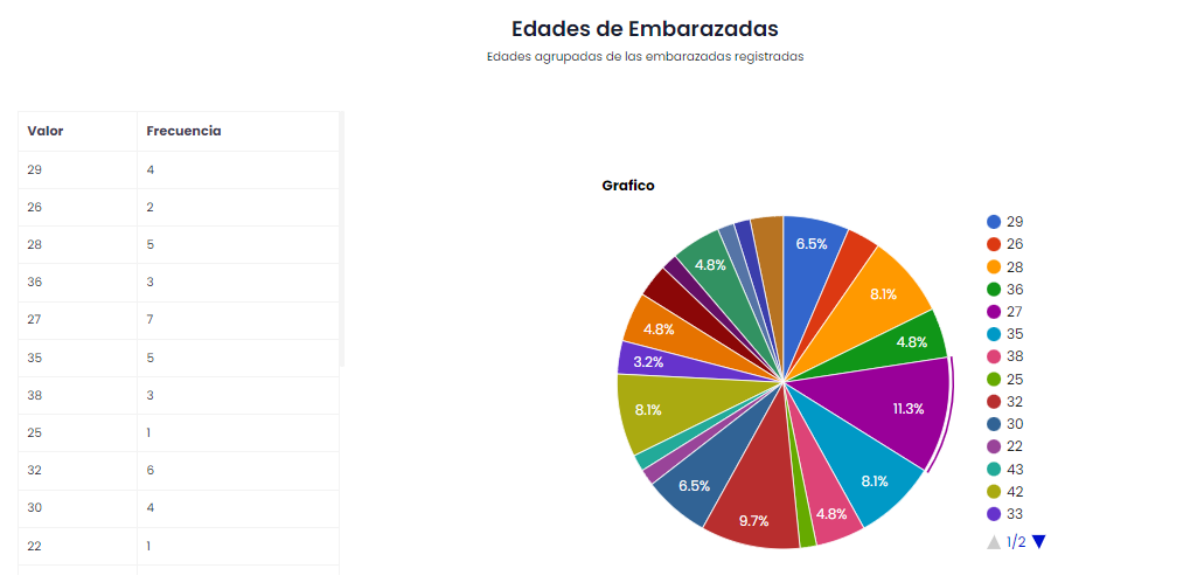
Edad

Enviar








2023© AndD4r

About Support Purchase

Se observan los resultados.



Además, se pueden observar todos los análisis que se realización

Resultados de Analisis			+ Nuevo Analisis
Nombre	Descripción	Fecha	
 Hombres y Mujeres	Cantidad de Hombres y Mujeres que generan Citas Médicas	2023-05-28	Cargar
 Citas por Centro de Costo	Cantidad de citas por centro de costo	2023-05-28	Cargar
 Ingresos por Centro de Costo	Cantidad de ingresos por centro de costo	2023-05-28	Cargar
 Embarazadas y No Embarazadas	Cantidad de mujeres embarazadas y no embarazadas	2023-05-28	Cargar
 Grupo Sanguíneo de Embarazadas	Cantidad por grupo sanguíneo de todas las embarazadas	2023-05-28	Cargar
 Estado Civil de Embarazadas	Estado civil de todas la embarazadas registradas	2023-05-29	Cargar
 Edades de Embarazadas	Edades agrupadas de las embarazadas registradas	2023-05-29	Cargar

2023© AndD4r

About Support Purchase