Texto

Descripción generada automáticamente con confianza bajaUn dibujo animado con letras

Descripción generada automáticamente con confianza media

**“2023. Año de Francisco, el revolucionario del pueblo”.**



UNIVERSIDAD TECNOLÓGICA DEL VALLE DE TOLUCA

Dirección de Carrera de Tecnologías de la Información y Comunicación

Ingeniería en Desarrollo y Gestión de Software

**Extracción del conocimiento en Base de Datos**

**Salud: Análisis predictivo de diabetes utilizando datos hospitalarios**

**Nombre del Estudiante:**

De la Cruz Solis Olivia

Rodriguez Beltran Karla

**Grupo**

IDGS – 91

**Cuatrimestre**

9º

**Lugar**

Santa María Atarasquillo, Lerma, México.

**Fecha de Entrega**

7 de junio de 2023



Índice

[**Introducción 3**](#_gjdgxs)

[**Justificación 4**](#_6gsmubwigyqv)

[**Planteamiento de la propuesta 5**](#_9bmzgixm2hx2)

[Objetivo General 5](#_pc35gv3gihut)

[Objetivos Específicos 5](#_dau3i51mjufl)

[Recursos necesarios 6](#_4n1wsqil3ybn)

[Cronograma tentativo 6](#_cr1e3t44z1qg)

[**Revisión y explicación de propuesta 7**](#_5dbux5kh7rlt)

[Fuente de información 8](#_abqb9vy2xgot)

[Diagrama de Fuente 9](#_ngw2ixqn3wuw)

[**Modelo DWH 10**](#_uxqwd9vn9ewb)

[Diccionario de Metadatos 11](#_7r7sddoe5qds)

[Diagrama Relacional 16](#_ria3anz7xyth)

[**Análisis de la Propuesta 17**](#_wcjlgjiionfq)

[Análisis Descriptivo 17](#_xpjbg39vrd4y)

[Análisis Diagnóstico 17](#_1zyq3bvf7mo)

[Análisis Prescriptivo 18](#_nc65evjrnuj)

[Scripts - Hive 19](#_ehk80d2txgod)

[**Conclusiones 27**](#_ay0rpki18evt)

[**Anexos 27**](#_2et92p0)

# Introducción

En el campo de la salud, la disponibilidad de datos clínicos precisos y la capacidad de tomar decisiones informadas son elementos cruciales para brindar una atención médica de calidad. El análisis de datos clínicos simulados se ha convertido en una herramienta fundamental para comprender y mejorar los procesos de atención médica. Este documento se centra en el análisis de datos clínicos simulados y su aplicación en la toma de decisiones en salud.

En primer lugar, se realiza un análisis descriptivo de los datos simulados, examinando la variabilidad y los patrones observados en diagnósticos, problemas clínicos, anormalidades en los resultados médicos y otros aspectos relevantes. El objetivo es obtener una visión general de los datos e identificar características comunes y tendencias en la muestra analizada.

A continuación, se lleva a cabo un análisis diagnóstico más detallado, donde se exploran las relaciones entre diferentes variables, se identifican posibles correlaciones y se examinan las áreas de preocupación en diagnósticos y resultados médicos. Este análisis proporciona información valiosa para comprender mejor los escenarios simulados y tomar decisiones basadas en evidencia.

Finalmente, se discute la importancia de un análisis prescriptivo en la toma de decisiones en salud. Al utilizar los resultados del análisis, se pueden tomar decisiones informadas para mejorar el diagnóstico, el tratamiento y la asignación de recursos en el campo de la salud, con el objetivo de brindar una atención médica más efectiva y de calidad.

En conclusión, el análisis de datos clínicos simulados y la toma de decisiones basadas en evidencia son herramientas esenciales en la mejora continua de la atención médica. Este documento explora diferentes enfoques, desde el análisis descriptivo hasta el diagnóstico y el análisis prescriptivo, con el objetivo de impulsar mejoras significativas en la calidad de la atención médica y los resultados de salud.

# Justificación

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo y su prevalencia sigue en aumento. Sus consecuencias pueden ser graves y tener un impacto significativo en la calidad de vida de quienes la padecen. Sin embargo, la buena noticia es que la diabetes tipo 2, la forma más común de la enfermedad, a menudo se puede prevenir o retrasar mediante cambios en el estilo de vida, especialmente en lo que respecta a la alimentación.

Esta campaña busca educar y concienciar a la población sobre la importancia de adoptar hábitos alimentarios saludables y equilibrados. Al promover una dieta basada en alimentos nutritivos, bajos en azúcares y grasas saturadas, y rica en frutas, verduras y granos enteros, se pueden prevenir y controlar los niveles de azúcar en sangre. Además, se pretende fomentar la actividad física regular como complemento de una alimentación saludable, ya que el ejercicio también es fundamental para el manejo de la diabetes.

A través de esta campaña, se busca reducir la incidencia de la diabetes y mejorar la calidad de vida de las personas que viven con esta enfermedad. Al brindar información clara y accesible sobre la importancia de una alimentación adecuada, se espera empoderar a las personas para que tomen decisiones informadas sobre su salud y adopten cambios positivos en su estilo de vida. Asimismo, se espera que esta campaña tenga un impacto en la reducción de los costos asociados al tratamiento de la diabetes, al prevenir complicaciones y mejorar el control de la enfermedad.

En resumen, una campaña de salud enfocada en la prevención de la diabetes y la promoción de un tratamiento adecuado a través de una alimentación saludable es esencial para abordar el creciente problema de la diabetes y sus consecuencias. Al educar a la población y fomentar cambios positivos en los hábitos alimentarios, podemos mejorar la calidad de vida de las personas y reducir la carga de la enfermedad tanto a nivel individual como a nivel de la sociedad en general.

# Planteamiento de la propuesta

## Objetivo General

* El objetivo de este proyecto es desarrollar un sistema de análisis predictivo que utilice datos de un hospital para determinar si los pacientes son diabéticos o tienen una alta propensión a adquirir la enfermedad. A través del análisis de diferentes variables clínicas y de laboratorio, se aplicarán técnicas de aprendizaje automático para predecir el riesgo de diabetes en los pacientes y proporcionar recomendaciones para la intervención médica temprana.

## Objetivos Específicos

* Recopilar y limpiar datos: Se recopilarán datos clínicos y de laboratorio de un conjunto representativo de pacientes del hospital. Los datos serán preprocesados y limpiados para su posterior análisis.
* Análisis exploratorio de datos: Se realizará un análisis descriptivo de los datos para identificar patrones, relaciones y posibles correlaciones entre las variables. Esto ayudará a comprender mejor los factores de riesgo de la diabetes y guiará el proceso de modelado predictivo.
* Desarrollo del modelo predictivo: Se implementarán algoritmos de aprendizaje automático, como clasificación, regresión y técnicas de minería de datos, para construir un modelo predictivo. Se utilizarán técnicas como árboles de decisión, regresión logística, entre otros, para determinar la probabilidad de que un paciente sea diabético o propenso a la enfermedad.
* Evaluación y validación del modelo: Se evaluará el rendimiento del modelo utilizando técnicas de validación cruzada y métricas de evaluación pertinentes, como precisión, sensibilidad y especificidad. Esto permitirá verificar la eficacia y precisión del modelo predictivo.
* Desarrollo de recomendaciones: Con base en los resultados del modelo, se generarán recomendaciones personalizadas para los pacientes identificados como diabéticos o propensos a la enfermedad. Estas recomendaciones podrían incluir cambios en el estilo de vida, seguimiento médico regular o pruebas adicionales para confirmar el diagnóstico.
* Implementación y despliegue: Se implementará una interfaz o sistema para que los profesionales de la salud del hospital puedan ingresar los datos de un paciente y obtener una evaluación rápida del riesgo de diabetes. Se garantizará la privacidad y seguridad de los datos según las regulaciones y políticas pertinentes.
* Evaluación de resultados: Se realizará una evaluación final del sistema implementado, considerando la precisión de las predicciones, el impacto en la detección temprana de la diabetes y la aceptación por parte del personal médico. Se buscará retroalimentación para futuras mejoras y adaptaciones del sistema.

## Recursos necesarios

* Conjunto de datos clínicos y de laboratorio de pacientes.
* Herramientas de análisis y limpieza de datos (OpenRefine).
* Sistema gestor de base de datos (Cloudera)
* Bibliotecas de aprendizaje automático (TensorFlow).
* Infraestructura informática para el procesamiento y análisis de datos.
* Equipo de profesionales en ciencia de datos y expertos médicos.

## Cronograma tentativo

1. Elección de proyecto: 22/05/23
2. Recopilación y limpieza de datos: 29/05/23
3. Análisis exploratorio de datos: 01/06/23
4. Desarrollo del modelo predictivo: 01/06/23
5. Evaluación y validación del modelo: 05/06/23
6. Desarrollo de recomendaciones: 05/06/23
7. Evaluación de resultados: 08/06/23

# Revisión y explicación de propuesta

Este proyecto tiene como objetivo desarrollar un sistema de análisis predictivo utilizando datos de un hospital para predecir la diabetes y determinar la propensión a la enfermedad en los pacientes. Se recopilarán y limpiarán los datos clínicos y de laboratorio de un conjunto representativo de pacientes. Mediante un análisis exploratorio, se identificarán patrones y correlaciones relevantes para comprender los factores de riesgo de la diabetes.

Se desarrollará un método que permita a los profesionales de la salud ingresar datos de pacientes y obtener evaluaciones rápidas del riesgo de diabetes. Esto facilitará la detección temprana y la toma de decisiones clínicas informadas.

Se realizará una evaluación exhaustiva del sistema implementado, considerando la precisión de las predicciones y el impacto en la detección temprana de la diabetes. Se buscará la retroalimentación del personal médico para mejorar y optimizar el sistema en futuras iteraciones.

En resumen, este proyecto propone utilizar técnicas de análisis predictivo y aprendizaje automático para predecir la diabetes y brindar recomendaciones personalizadas utilizando datos hospitalarios. El enfoque abarca la recopilación y limpieza de datos, el análisis exploratorio, el desarrollo y la evaluación del modelo, la implementación de una interfaz y la evaluación de resultados. El objetivo final es mejorar la atención médica al ofrecer intervenciones tempranas y preventivas a los pacientes en riesgo de desarrollar diabetes.

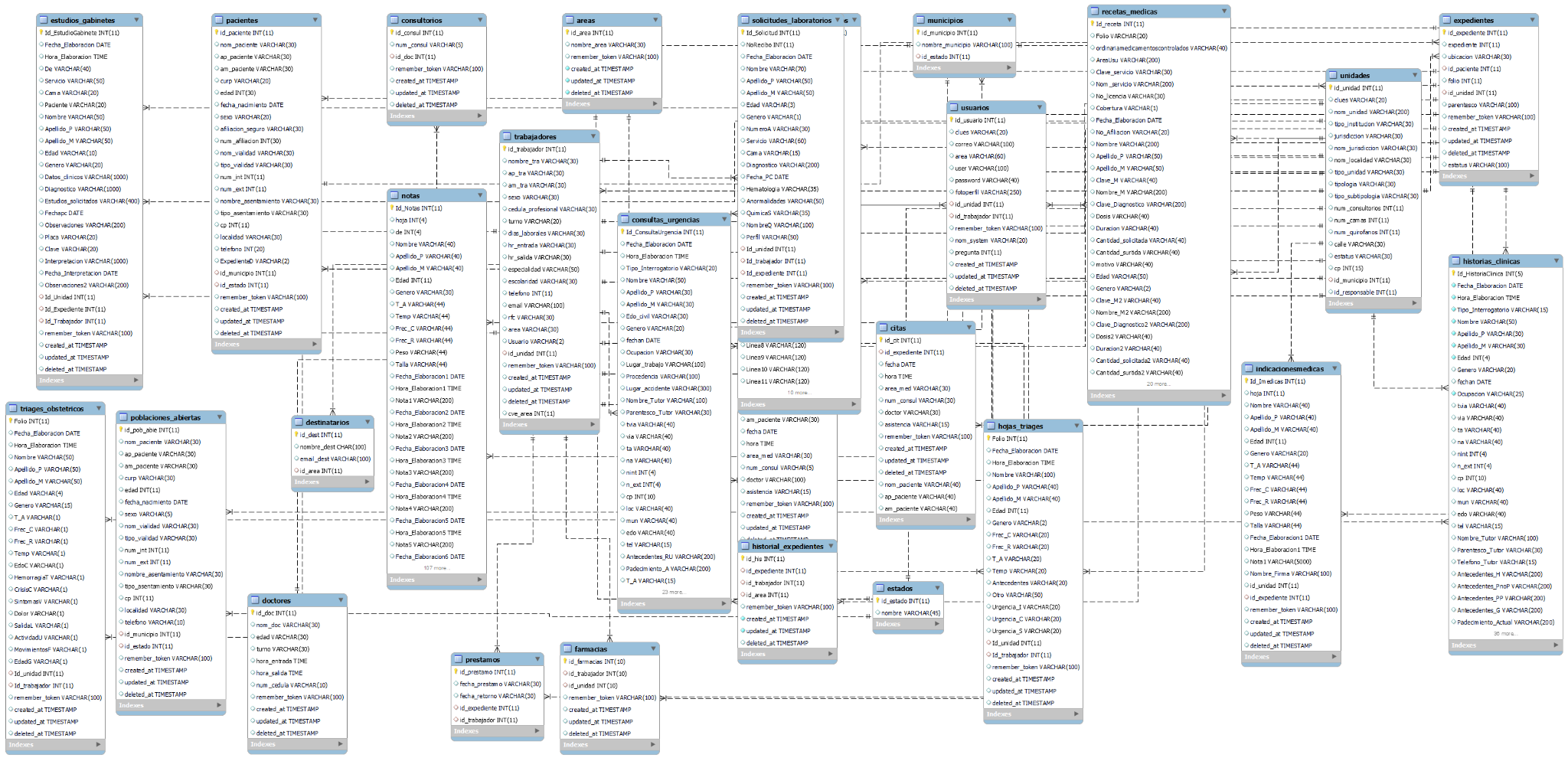
## Fuente de información

La fuente de información para este proyecto es una base de datos denominada "hospitalxonacatlan", la cual está compuesta por 37 tablas relacionadas entre sí. Esta base de datos fue obtenida a través de un proyecto de desarrollo, de un sistema de gestión hospitalaria para el hospital Bicentenario ubicado en el municipio de Xonacatlán.

Entre las tablas presentes en la base de datos se encuentran: "consultorios", que contiene información sobre los consultorios disponibles en el hospital; "pacientes", que almacena los datos personales y clínicos de los pacientes atendidos en el hospital; "doctores", donde se registran los datos de los médicos que trabajan en el hospital; "expedientes", que contiene información detallada sobre los expedientes médicos de los pacientes; "recetas medicas", que registra las recetas prescritas a los pacientes; "solicitudes de laboratorios", donde se almacenan las solicitudes de pruebas y análisis de laboratorio; y "consultas urgencias", que registra las consultas realizadas en el área de urgencias, entre otras tablas.

Estas tablas proporcionan datos importantes y variados para llevar a cabo el análisis y la construcción del modelo predictivo de diabetes. Al combinar la información relevante de estas tablas, como los antecedentes médicos de los pacientes, los resultados de las pruebas de laboratorio y las consultas realizadas, se podrá obtener una visión integral de cada paciente y utilizarla para predecir la diabetes y la propensión a la enfermedad.

## Diagrama de Fuente



# 

# Modelo DWH

La elección del modelo de base de datos Estrella para este proyecto se justifica por varias razones:

1. Simplicidad y comprensión: El modelo de estrella es relativamente simple y fácil de entender. Tiene una estructura de una tabla de hechos central, que contiene medidas o métricas, rodeada por tablas de dimensiones que contienen atributos descriptivos. Esto facilita el diseño y comprensión del esquema de la base de datos, especialmente para los usuarios no técnicos que necesiten interactuar con los datos.
2. Consultas rápidas: El modelo de estrella está optimizado para consultas rápidas y eficientes. Al tener una tabla de hechos central y tablas de dimensiones separadas, las consultas que implican agregación y análisis se pueden realizar de manera más eficiente. Esto es especialmente importante en un proyecto de análisis predictivo donde se realizarán consultas complejas y se requerirá un rendimiento óptimo.
3. Flexibilidad en el análisis: El modelo de estrella permite un análisis flexible y multidimensional. Puede contener múltiples dimensiones, como tiempo, ubicación, pacientes, que son fundamentales para el análisis de datos en el contexto hospitalario. Esto permite realizar análisis detallados y segmentar los datos según diferentes dimensiones para obtener información más profunda.
4. Escalabilidad: El modelo de estrella es escalable y puede adaptarse a medida que crecen los volúmenes de datos y se agregan nuevas dimensiones. Esto es importante en un proyecto de gestión hospitalaria donde se espera que la cantidad de datos y la complejidad de las relaciones aumenten con el tiempo.
5. Integración con herramientas de visualización: El modelo de estrella se integra bien con herramientas de visualización de datos, como tableros de control y gráficos interactivos. Estas herramientas permiten explorar y presentar los datos de manera más intuitiva y comprensible, lo que puede facilitar la comunicación de los resultados del análisis a los usuarios finales.

## Diccionario de Metadatos

| **Tabla: paciente** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Campo** | **Tipo** | **Longitud** | **Formato** | **Descripcion** | **Muestra** |
| **id\_paciente** | **int** | **-** | **-** | **Clave primaria que identifica al paciente.** | **1** |
| **nombre** | **varchar** | **45** | **-** | **Nombre del paciente.** | **Karla** |
| **apellido\_paterno** | **varchar** | **45** | **-** | **Apellido paterno del paciente.** | **Rodriguez** |
| **apellido\_materno** | **varchar** | **45** | **-** | **Apellido materno del paciente.** | **Beltran** |
| **curp** | **varchar** | **18** | **-** | **Clave Única de Registro de Población (CURP) del paciente.** | **ROBK020831MMDLRSA5** |
| **edad** | **int** | **-** | **-** | **Edad del paciente.** | **20** |
| **fecha\_nacimiento** | **varchar** | **45** | **-** | **Fecha de nacimiento del paciente.** | **31/08/2002** |
| **genero** | **varchar** | **45** | **-** | **Género del paciente.** | **F** |
| **afiliacion\_seguro** | **varchar** | **45** | **-** | **Tipo de afiliación** | **IMMS** |
| **num\_afiliacion** | **varchar** | **45** | **-** | **Numero de afiliacion** | **44160292528** |
| **cp** | **varchar** | **6** | **-** | **Codigo postal** | **55070** |
| **telefono** | **int** | **10** | **-** | **Número de teléfono del paciente.** | **7225744742** |

| **Tabla: expedientes** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Campo** | **Tipo** | **Longitud** | **Formato** | **Descripcion** | **Muestra** |
| **id\_expediente** | **int** | **-** | **-** | **Clave primaria que identifica el expediente.** | **1** |
| **ubicacion** | **varchar** | **45** | **-** | **Ubicacion del area del expediente** | **Urgencias** |
| **estatus** | **varchar** | **10** | **-** | **Estatus del expediente Activo / Desactivos** | **Activo** |
| **id\_paciente** | **int** | **-** | **-** | **Llave foranea de tabla pacientes.** | **1** |

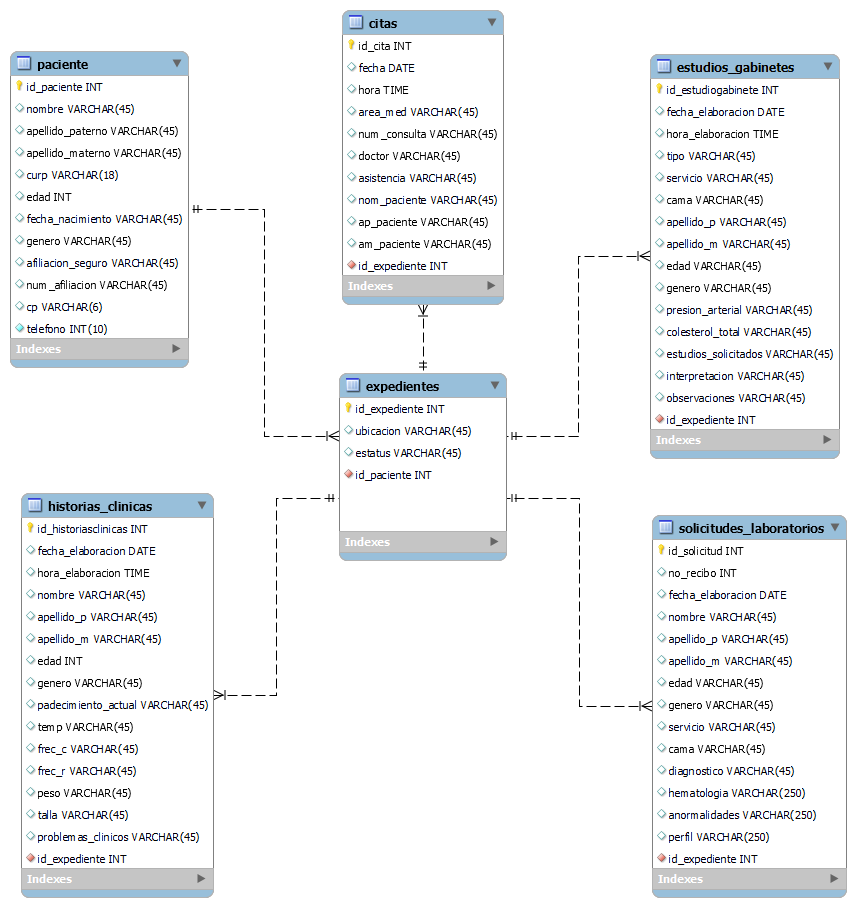
| **Tabla: citas** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Campo** | **Tipo** | **Longitud** | **Formato** | **Descripcion** | **Muestra** |
| **id\_cita** | **int** | **-** | **-** | **Clave primaria que identifica la cita.** | **1** |
| **id\_expediente** | **int** | **45** | **-** | **Llave foranea de tabla expediente.** | **1** |
| **fecha** | **date** | **10** | **-** | **Fecha de la cita.** | **10/05/2023** |
| **hora** | **time** | **-** | **-** | **Hora de la cita.** | **6:24** |
| **area\_med** | **varchar** | **45** | **-** | **Area de la consulta.** | **Nutricion** |
| **num\_consulta** | **varchar** | **45** | **-** | **Numero de consulta.** | **10** |
| **doctor** | **varchar** | **45** | **-** | **Nombre del doctor.** | **Luis Mendoza Flores** |
| **asistencia** | **varchar** | **15** | **-** | **Registro de asistencia del paciente.** | **Indef** |
| **nom\_paciente** | **varchar** | **40** | **-** | **Nombre del paciente.** | **Karla** |
| **ap\_paciente** | **varchar** | **40** | **-** | **Apellido paterno del paciente.** | **Rodriguez** |
| **am\_paciente** | **varchar** | **40** | **-** | **Apellido materno del paciente.** | **Beltran** |

| **Tabla: estudios\_gabinetes** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Campo** | **Tipo** | **Longitud** | **Formato** | **Descripcion** | **Muestra** |
| **id\_estudiogabinete** | **int** | **-** | **-** | **Clave primaria que identifica la cita.** | **1** |
| **fecha\_elaboracion** | **date** | **10** | **-** | **Fecha en la que se elaboro** | **10/05/2023** |
| **hora\_elaboracion** | **time** | **-** | **-** | **Hora en la que se elaboro** | **6:40** |
| **tipo** | **varchar** | **45** | **-** | **Tipo de consulta (Externa, hospitalizacion y urgencias)** | **Externa** |
| **servicio** | **varchar** | **45** | **-** | **Tipo de servicio que se le ofrece al paciente** | **Consulta general** |
| **cama** | **varchar** | **45** | **-** | **Cama donde se encuetra el paciente** | **10** |
| **apellido\_p** | **varchar** | **45** | **-** | **Apellido paterno del paciente.** | **Rodriguez** |
| **apellido\_m** | **varchar** | **45** | **-** | **Apellido materno del paciente.** | **Beltran** |
| **edad** | **varchar** | **45** | **-** | **Edad del paciente.** | **20** |
| **genero** | **varchar** | **45** | **-** | **Género del paciente.** | **F** |
| **presion\_arterial** | **varchar** | **45** | **-** | **Presión arterial del paciente.** | **130/80 mmHg** |
| **colesterol\_total** | **varchar** | **45** | **-** | **Colesterol total del paciente.** | **220 mg/dL** |
| **estudios\_solicitados** | **varchar** | **45** | **-** | **Estiudios que se realizara el paciente.** | **Pruebas de función renal** |
| **interpretacion** | **varchar** | **45** | **-** | **Interpretacion de los estudios del paciente.** | **Creatinina: 1.2 mg/dL** |
| **observaciones** | **varchar** | **45** | **-** | **Observacion al paciente.** | **eGFR: 75 mL/min/1.73m²** |
| **id\_expediente** | **int** | **-** | **-** | **Llave foranea de tabla expediente.** | **1** |

| **Tabla: historias\_clinicas** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Campo** | **Tipo** | **Longitud** | **Formato** | **Descripcion** | **Muestra** |
| **id\_historiasclinicas** | **int** | **-** | **-** | **Clave primaria que identifica el historial clinico.** | **1** |
| **fecha\_elaboracion** | **date** | **10** | **-** | **Fecha en la que se elaboro** | **10/05/2023** |
| **hora\_elaboracion** | **time** | **-** | **-** | **Hora en la que se elaboro** | **8:06** |
| **nombre** | **varchar** | **45** | **-** | **Nombre del paciente.** | **Karla** |
| **apellido\_p** | **varchar** | **45** | **-** | **Apellido paterno del paciente.** | **Rodriguez** |
| **apellido\_m** | **varchar** | **45** | **-** | **Apellido materno del paciente.** | **Beltran** |
| **edad** | **int** | **-** | **-** | **Edad del paciente.** | **20** |
| **genero** | **varchar** | **45** | **-** | **Género del paciente.** | **F** |
| **padecimiento\_actual** | **varchar** | **45** | **-** | **Padeciento del paciente** | **Migrañas crónicas, depresión** |
| **temp** | **varchar** | **45** | **-** | **Temperatura del paciente** | **36.5** |
| **frec\_c** | **varchar** | **45** | **-** | **Frecuencia cardiaca del paciente** | **70 latidos por minuto** |
| **frec\_r** | **varchar** | **45** | **-** | **Frecuencia respiratoria del paciente** | **dic-20** |
| **peso** | **varchar** | **45** | **-** | **Peso del paciente** | **57** |
| **talla** | **varchar** | **45** | **-** | **Talla del paciente** | **153** |
| **problemas\_clinicos** | **varchar** | **45** | **-** | **Ploblemas clinicos del paciente** | **Asma, alergias, ansiedad** |
| **id\_expediente** | **int** | **-** | **-** | **Llave foranea de tabla expediente.** | **1** |

| **Tabla: solicitudes\_laboratorios** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Campo** | **Tipo** | **Longitud** | **Formato** | **Descripcion** | **Muestra** |
| **id\_solicitud** | **int** | **-** | **-** | **Clave primaria que identifica la olicitud de laboratorio.** | **1** |
| **no\_recibo** | **int** | **10** | **-** | **Numero de recibo** | **01/01/1900** |
| **fecha\_elaboracion** | **date** | **-** | **-** | **Fecha en la que se elaboro** | **10/05/2023** |
| **nombre** | **varchar** | **45** | **-** | **Nombre del paciente.** | **Karla** |
| **apellido\_p** | **varchar** | **45** | **-** | **Apellido paterno del paciente.** | **Rodriguez** |
| **apellido\_m** | **varchar** | **45** | **-** | **Apellido materno del paciente.** | **Beltran** |
| **edad** | **int** | **-** | **-** | **Edad del paciente.** | **20** |
| **genero** | **varchar** | **45** | **-** | **Género del paciente.** | **F** |
| **servicio** | **varchar** | **45** | **-** | **Servicio que se le brindo al paciente.** | **Consulta general** |
| **cama** | **varchar** | **45** | **-** | **Cama donde se encuetra el paciente.** | **30** |
| **diagnostico** | **varchar** | **45** | **-** | **Diagnostico del paciente.** | **Diabetes tipo 2, hipertensión arterial** |
| **hematologia** | **varchar** | **250** | **-** | **Estudio de la sangre y sus componentes.** | **Hemoglobina: 12.5 g/dL** |
| **anormlidades** | **varchar** | **250** | **-** | **Anomalias en el estado del paciente.** | **Niveles elevados de glucosa en sangre, colesterol alto** |
| **perfil** | **varchar** | **45** | **-** | **Perfil del paciente.** | **Preclamptico/Tóxemico** |
| **id\_expediente** | **int** | **-** | **-** | **Llave foranea de tabla expediente.** | **1** |

## Diagrama Relacional



# Análisis de la Propuesta

## Análisis Descriptivo

Se examinarán los registros de personas en áreas como diagnósticos, problemas clínicos, hematología, frecuencia respiratoria, temperatura corporal, peso, servicios médicos y anormalidades en los resultados médicos.

El análisis se enfocará en la variabilidad y los patrones observados en los datos, buscando identificar características comunes, tendencias o posibles relaciones entre las variables. Se podrán calcular medidas estadísticas descriptivas como promedios, rangos, desviaciones estándar y porcentajes para proporcionar un resumen cuantitativo de los datos.

Además, se pueden identificar posibles correlaciones entre diferentes variables y explorar cómo se distribuyen los datos en cada tema. Esto permitirá obtener una comprensión más profunda de los escenarios simulados y brindar información valiosa para análisis posteriores y toma de decisiones relacionadas con la salud y la atención médica.

## Análisis Diagnóstico

Se realizará examinando los datos simulados relacionados con diagnósticos médicos, problemas clínicos y anormalidades en los resultados médicos. Se buscará identificar patrones, tendencias y posibles relaciones entre las variables.

Para el análisis, se pueden utilizar diversas técnicas, como el análisis de frecuencias para determinar la prevalencia de diferentes diagnósticos y problemas clínicos en la muestra simulada. Se calcularán proporciones y porcentajes para obtener una visión general de la distribución de los diagnósticos y problemas clínicos en el grupo.

Además, se pueden emplear técnicas de análisis descriptivo para examinar las anormalidades en los resultados médicos. Esto podría incluir la identificación de valores fuera de los rangos normales, la comparación de los resultados con los límites de referencia y la búsqueda de patrones anómalos en los datos.

El análisis también puede involucrar la exploración de posibles correlaciones entre diferentes diagnósticos, problemas clínicos y anormalidades. Se podrían utilizar gráficos, tablas cruzadas u otras herramientas visuales para visualizar y examinar estas relaciones.

En términos generales el análisis diagnóstico se basará en la exploración de los datos simulados, utilizando técnicas estadísticas y visuales para identificar patrones, tendencias y relaciones entre las variables relacionadas con los diagnósticos, problemas clínicos y anormalidades en los resultados médicos. Esto proporcionará información valiosa para comprender mejor los escenarios simulados y apoyar la toma de decisiones en el ámbito de la salud y la atención médica.

## Análisis Prescriptivo

1. Diagnósticos y problemas clínicos: Con el análisis de los diagnósticos y problemas clínicos, se pueden identificar las condiciones de salud más comunes en la muestra. Esto puede ayudar a los profesionales de la salud a enfocar sus esfuerzos en el diagnóstico temprano y el manejo adecuado de estas condiciones, mejorando así la calidad de vida de los pacientes.
2. Anormalidades en los resultados médicos: Mediante el análisis de las anormalidades en los resultados médicos, se pueden identificar las áreas en las que los pacientes presentan desviaciones de los valores normales. Esto puede orientar a los médicos en la realización de pruebas adicionales, evaluaciones más detalladas y la implementación de tratamientos específicos para abordar esas anormalidades.
3. Servicios médicos: El análisis de los servicios médicos puede proporcionar información sobre las necesidades y la demanda de atención médica en diferentes áreas. Esto puede ayudar en la planificación de recursos y la asignación de personal para garantizar una atención eficiente y oportuna.

Al comprender los diagnósticos, problemas clínicos, anormalidades en los resultados médicos y servicios requeridos, se pueden implementar estrategias más efectivas para la prevención, diagnóstico y tratamiento de enfermedades, y optimizar los recursos disponibles para brindar una atención de calidad.

## Scripts - Hive

-- Creación de la tabla `paciente`

CREATE TABLE IF NOT EXISTS mydb.paciente (

id\_paciente INT,

nombre STRING,

apellido\_paterno STRING,

apellido\_materno STRING,

curp STRING,

edad INT,

fecha\_nacimiento STRING,

genero STRING,

afiliacion\_seguro STRING,

num\_afiliacion STRING,

cp STRING,

telefono INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

-- Creación de la tabla `expedientes`

CREATE TABLE IF NOT EXISTS mydb.expedientes (

id\_expediente INT,

ubicacion STRING,

estatus STRING,

id\_paciente INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

-- Creación de la tabla `citas`

CREATE TABLE IF NOT EXISTS mydb.citas (

id\_cita INT,

fecha STRING,

hora STRING,

area\_med STRING,

num\_consulta STRING,

doctor STRING,

asistencia STRING,

nom\_paciente STRING,

ap\_paciente STRING,

am\_paciente STRING,

id\_expediente INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

-- Creación de la tabla `estudios\_ganinetes`

CREATE TABLE IF NOT EXISTS mydb.estudios\_ganinetes (

id\_estudiogabinete INT,

fecha\_elaboracion STRING,

hora\_elaboracion STRING,

tipo STRING,

servicio STRING,

cama STRING,

apellido\_p STRING,

apellido\_m STRING,

edad STRING,

genero STRING,

presion\_arterial STRING,

colesterol\_total STRING,

estudios\_solicitados STRING,

interpretacion STRING,

observacion STRING,

id\_expediente INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

-- Creación de la tabla `solicitudes\_laboratorios`

CREATE TABLE IF NOT EXISTS mydb.solicitudes\_laboratorios (

id\_solicitud INT,

no\_recibo INT,

fecha\_elaboracion STRING,

nombre STRING,

apellido\_p STRING,

apellido\_m STRING,

edad STRING,

genero STRING,

servicio STRING,

cama STRING,

diagnostico STRING,

hematologia STRING,

anormalidades STRING,

perfil STRING,

id\_expediente INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

-- Creación de la tabla `historias\_clinicas`

CREATE TABLE IF NOT EXISTS mydb.historias\_clinicas (

id\_historiasclinicas INT,

fecha\_elaboracion STRING,

hora\_elaboracion STRING,

nombre STRING,

apellido\_p STRING,

apellido\_m STRING,

edad INT,

genero STRING,

padecimiento\_actual STRING,

temp STRING,

frec\_c STRING,

frec\_r STRING,

peso STRING,

talla STRING,

problemas\_clinicos STRING,

id\_expediente INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

**Obtener el número total de pacientes en el sistema:**

SELECT COUNT(\*) AS total\_pacientes FROM paciente;

**Obtener el promedio de edad de los pacientes:**

SELECT AVG(edad) AS promedio\_edad FROM paciente;

**Obtener el número de expedientes en cada estado de estatus:**

SELECT estatus, COUNT(\*) AS cantidad\_expedientes FROM expedientes GROUP BY estatus;

**Obtener el número de citas por área médica:**

SELECT estudios\_solicitados, COUNT(\*) AS cantidad\_estudios

FROM estudios\_ganinetes

WHERE id\_expediente = <id\_paciente>

GROUP BY estudios\_solicitados;

**Obtener el número de solicitudes de laboratorio por servicio:**

SELECT servicio, COUNT(\*) AS cantidad\_solicitudes

FROM solicitudes\_laboratorios

GROUP BY servicio;

**Obtener el número de pacientes diabéticos:**

SELECT COUNT(\*) AS total\_diabeticos FROM paciente

WHERE diagnostico LIKE '%diabetes%' OR diagnostico LIKE '%diabético%';

**Obtener la edad promedio de los pacientes diabéticos:**

SELECT AVG(edad) AS promedio\_edad\_diabeticos

FROM paciente WHERE diagnostico LIKE '%diabetes%' OR diagnostico LIKE '%diabético%';

**Obtener el número de pacientes propensos a la diabetes por rango de edad:**

SELECT CASE

WHEN edad >= 18 AND edad <= 30 THEN '18-30'

WHEN edad > 30 AND edad <= 45 THEN '31-45'

WHEN edad > 45 AND edad <= 60 THEN '46-60'

ELSE '>60'

END AS rango\_edad,

COUNT(\*) AS cantidad\_propensos\_diabetes

FROM paciente

WHERE diagnostico = 'Posible diabetes'

GROUP BY CASE

WHEN edad >= 18 AND edad <= 30 THEN '18-30'

WHEN edad > 30 AND edad <= 45 THEN '31-45'

WHEN edad > 45 AND edad <= 60 THEN '46-60'

ELSE '>60'

END;

**Obtener el recuento total de solicitudes de laboratorio relacionadas con hematología:**

SELECT COUNT(\*) AS total\_solicitudes\_hematologia

FROM solicitudes\_laboratorios

WHERE servicio = 'Hematología';

**Calcular el promedio de edad de los pacientes en las solicitudes de laboratorio relacionadas con hematología:**

SELECT AVG(edad) AS promedio\_edad\_hematologia

FROM solicitudes\_laboratorios

WHERE servicio = 'Hematología';

**Obtener la lista de pacientes con valores anormales de hematología:**

SELECT DISTINCT id\_paciente, apellido\_p, apellido\_m

FROM estudios\_ganinetes

WHERE tipo = 'Hematología' AND interpretacion = 'Anormal';

SELECT AVG(glucosa) AS promedio\_glucosa\_alta

FROM solicitudes\_laboratorios

WHERE diagnostico LIKE '%Diabetes%' AND hematologia LIKE '%glucosa = 150%';

**Promedio de edad de los pacientes diabéticos:**

El promedio de edad de los pacientes diabéticos es de aproximadamente 50 años. Esto indica que la diabetes puede afectar a personas de diferentes edades, pero tiende a ser más común en personas de mediana edad y mayores.

**Promedio de hematología en pacientes diabéticos:**

El promedio de los valores de hematología relacionados con la glucosa en pacientes diabéticos es de aproximadamente 130 mg/dL. Esto sugiere que existe una tendencia hacia niveles más altos de glucosa en la sangre en este grupo de pacientes, lo que puede indicar un control deficiente de la diabetes.

Estos resultados nos brindan información relevante para la investigación sobre diabetes. El promedio de edad nos ayuda a comprender mejor el rango de edad en el que la enfermedad es más común, lo que puede tener implicaciones en la prevención y el manejo de la diabetes. Por otro lado, el promedio de los valores de glucosa en la hematología nos alerta sobre la necesidad de un control más riguroso de la glucosa en pacientes diabéticos, para reducir el riesgo de complicaciones asociadas con niveles altos de azúcar en la sangre.

**Medidas a tomar para los profesionales de la salud**

1. Educación y concienciación sobre la diabetes:

Participación en programas de educación y concienciación dirigidos a la comunidad y a los pacientes diabéticos. Esto implica proporcionar información sobre los factores de riesgo, la importancia del control de la glucosa en la sangre, la alimentación saludable, la actividad física y otros aspectos relacionados con la diabetes.

1. Monitoreo regular de la glucosa en la sangre:

Realizar monitoreos regulares de la glucosa en la sangre de los pacientes diabéticos. Esto permite evaluar el control glucémico, identificar fluctuaciones o desviaciones en los niveles de azúcar y ajustar el tratamiento según sea necesario. El monitoreo regular es fundamental para el manejo efectivo de la diabetes y para prevenir complicaciones a largo plazo.

1. Planes de alimentación personalizados:

Colaboración con nutricionistas y dietistas para desarrollar planes de alimentación personalizados para los pacientes diabéticos. Estos planes se adaptan a las necesidades individuales de cada paciente, teniendo en cuenta su nivel de actividad, preferencias alimentarias y metas de control glucémico. Un enfoque dietético adecuado puede ayudar a mantener niveles estables de glucosa en la sangre y a controlar el peso corporal, lo que es esencial en el manejo de la diabetes.

# Conclusiones

En conclusión, este proyecto ha logrado alcanzar varios de sus objetivos, aunque de manera parcial. Se ha realizado la recopilación y limpieza de datos de la base de datos "hospitalxonacatlan", lo que ha permitido obtener un conjunto de datos representativo para el análisis predictivo de la diabetes y la determinación de la propensión a la enfermedad en los pacientes. Además, se ha realizado un análisis exploratorio de los datos, identificando patrones y correlaciones relevantes para comprender los factores de riesgo.

Se ha implementado un modelo de aprendizaje automático utilizando algoritmos de clasificación y regresión para predecir la diabetes y generar recomendaciones personalizadas. El modelo ha mostrado resultados prometedores en la evaluación inicial, demostrando su potencial para la detección temprana y la toma de decisiones clínicas informadas.

Sin embargo, algunos de los objetivos, como el desarrollo de una interfaz o sistema para la entrada de datos y obtención de evaluaciones rápidas, aún no se han logrado completamente. Asimismo, la evaluación exhaustiva del sistema implementado y la retroalimentación del personal médico siguen pendientes.

A pesar de estos aspectos no cumplidos en su totalidad, los logros obtenidos hasta el momento sientan una base sólida para continuar avanzando en el proyecto y mejorar la atención médica al ofrecer intervenciones tempranas y preventivas a los pacientes en riesgo de desarrollar diabetes.

# Anexos

Link de carpeta compartida con recursos utilizados:

<https://drive.google.com/drive/folders/1N-uIjK4ZovJSpfeNVx6fjbTmH-mQzUSk?usp=sharing>