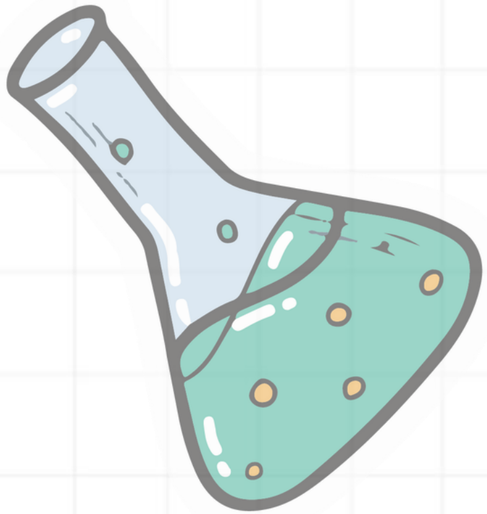
Proyecto de Datos Masivos y Ciencia de Datos en el Sector Salud

Integrantes:
Paola Michelle Figueroa Benítez
Leonard Jose Cuenca Roa
Lowenski Paredes Rosario
Carlos Damián Rodriguez Uitzil

Grupo: 1001
Equipo 01C

Tema del Proyecto

"Estimación de los niveles de obesidad en función de hábitos alimenticios y condición física mediante el análisis de datos masivos de historiales médicos electrónicos (HME)"



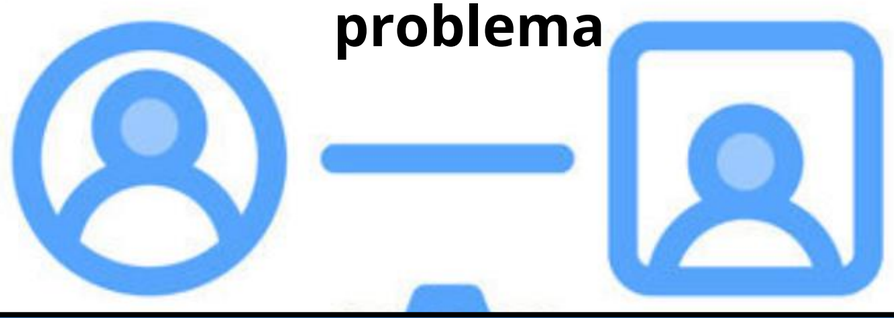
Sección 1

Definición del problema

Descripción del Proyecto

El proyecto busca analizar cómo los hábitos alimenticios y el nivel de actividad física contribuyen a los diferentes niveles de obesidad, utilizando grandes volúmenes de datos (Big Data) extraídos de historiales médicos electrónicos (HME) y dispositivos de monitoreo personal.

Definición del problema



La obesidad, clasificada como una enfermedad crónica, ha alcanzado proporciones epidémicas a nivel mundial. Afecta a millones de personas de diversas edades, géneros y contextos socioeconómicos. Se ha identificado como un factor de riesgo para una variedad de enfermedades graves, como la diabetes tipo 2, enfermedades cardiovasculares, apnea del sueño, hipertensión, y ciertos tipos de cáncer.

Importancia del Análisis de Datos en Salud

01

Mejora en la Atención Médica

El análisis de datos permite identificar tendencias en la salud de la población, facilitando la personalización de tratamientos y la anticipación de complicaciones, mejorando así la atención médica.

02

Optimización de Recursos Sanitarios

A través del análisis de datos, las instituciones de salud pueden evaluar el uso de recursos, lo que permite una gestión más eficiente y la reducción de costos sin sacrificar la calidad del servicio.

Contexto Global de la Obesidad

01

Prevalencia alarmante

La obesidad afecta a más de 650 millones de adultos en el mundo, con proyecciones que indican un aumento significativo en los próximos años.

02

Impacto en la salud

La obesidad está vinculada a enfermedades crónicas como diabetes y enfermedades cardiovasculares, aumentando la morbilidad y mortalidad en diversas poblaciones.

03

Desigualdades socioeconómicas

Las personas de bajos ingresos enfrentan mayores riesgos de obesidad debido a la falta de acceso a alimentos saludables y oportunidades de actividad física.

Proporciones de la Obesidad a Nivel Mundial

Aumento global de la obesidad

Desde 1975, la obesidad ha aumentado casi tres veces, afectando a millones de personas y generando preocupaciones significativas sobre la salud pública y los sistemas de atención médica.

Obesidad infantil en aumento

En 2020, aproximadamente 39 millones de niños menores de cinco años tenían sobrepeso u obesidad, lo que resalta la necesidad urgente de intervenciones preventivas en la infancia.

Desigualdades regionales marcadas

Las tasas de obesidad varían considerablemente entre regiones, siendo más altas en América del Norte, mientras que en Asia y África, aunque más bajas, están en aumento rápidamente.

Sección 2

Objetivos del Proyecto

Objetivos

Identificar patrones y tendencias

Utilizar técnicas de análisis de datos avanzadas, como el análisis de series temporales, minería de datos y análisis de patrones de consumo,

01

02

Evaluar la efectividad de intervenciones

Analizar la efectividad de distintas intervenciones en la reducción de la obesidad, utilizando métodos estadísticos avanzados para medir los resultados de programas de salud pública, dietas personalizadas y planes de ejercicios

Desarrollar modelos predictivos

Aplicar algoritmos de machine learning (como regresión logística, árboles de decisión, redes neuronales y algoritmos de clasificación) para predecir el riesgo de obesidad en individuos, teniendo en cuenta una variedad de factores, como el historial médico, los hábitos alimenticios y la actividad física

03

04

Proporcionar recomendaciones personalizadas

Desarrollar un sistema basado en datos que pueda generar recomendaciones personalizadas de cambios en los hábitos alimenticios y la actividad física de los usuarios, con el fin de prevenir y manejar la obesidad

Sección 3

Diseño del proyecto

Arquitecturas típicas de proyectos de datos masivos

Fuentes heterogéneas

Historiales médicos electrónicos (HME)

Datos estructurados como diagnósticos, tratamientos e historial de visitas, junto con antecedentes familiares (por ejemplo, historial de obesidad), resultados de laboratorio (colesterol, triglicéridos, hormonas) y notas clínicas detalladas, proporcionan una visión integral de la salud del paciente

Encuestas de salud y bienestar

Las encuestas administradas a los pacientes han proporcionado una valiosa cantidad de datos semiestructurados sobre su bienestar, estilo de vida y hábitos alimenticios. Esta información detallada incluye aspectos como la frecuencia y el tipo de alimentos consumidos, los niveles de actividad física, la presencia de factores psicosociales y los patrones de sueño.

01

02

03

04

Sensores de dispositivos de monitoreo (wearables)

Los relojes inteligentes, generan una gran cantidad de datos tanto estructurados como no estructurados. Estos dispositivos monitorizan diversos parámetros de salud, como la presión arterial, el ritmo cardíaco y los niveles de glucosa, proporcionando información valiosa sobre el estado físico de los usuarios.

Fuentes Adicionales

Las imágenes médicas, como resonancias magnéticas, tomografías computarizadas y radiografías, son herramientas fundamentales para evaluar la composición corporal y detectar comorbilidades. Estos estudios proporcionan una visión detallada del interior del cuerpo, permitiendo a los profesionales de la salud identificar problemas de salud subyacentes.

Arquitecturas típicas de proyectos de datos masivos

Extracción, transformación y carga (ETL)

01

Extracción

Las APIs facilitan el acceso a información estructurada en sistemas de historiales médicos electrónicos y plataformas de wearables. Sin embargo, para una extracción más profunda, la conexión directa a bases de datos, siempre y cuando se respeten los permisos establecidos, puede ser necesaria. Es crucial definir qué datos son relevantes, como medidas antropométricas y resultados de análisis. Paralelamente, las redes sociales ofrecen un tesoro de datos cualitativos.

02

Transformación

La limpieza de datos es un paso fundamental en cualquier análisis. Implica identificar y corregir errores como valores atípicos, duplicados o inconsistencias. Además, la normalización es crucial para convertir los datos a un formato común, facilitando su análisis. Por ejemplo, transformar las lecturas de dispositivos en formatos estandarizados o unificar las unidades de medida.

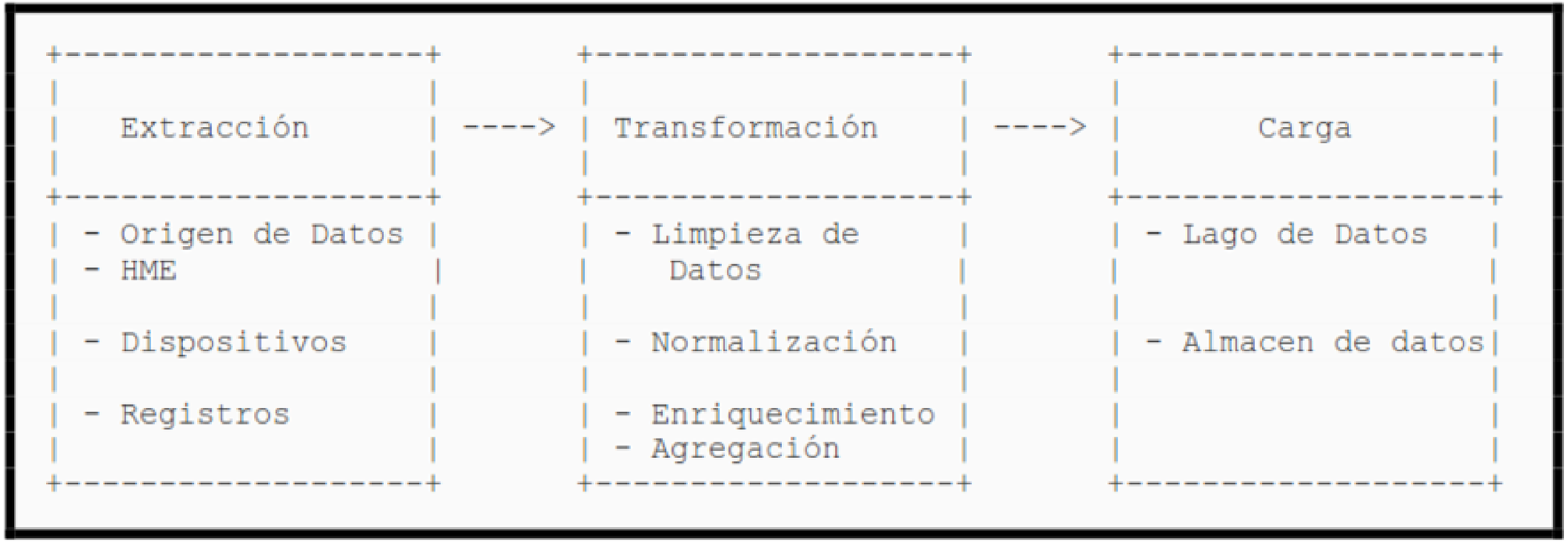
03

Carga

Almacenar los datos transformados en un Data Warehouse o Data Lake para su posterior análisis

Arquitecturas típicas de proyectos de datos masivos

Diagrama ETL



Arquitecturas típicas de proyectos de datos masivos

Almacenamiento

Data Lake

01

Los Data Lakes son ideales para almacenar una amplia variedad de datos, especialmente aquellos que no siguen una estructura rígida, como los provenientes de wearables o encuestas. Esta flexibilidad permite capturar y conservar datos crudos en su formato original, sin la necesidad de preprocesarlos inmediatamente.

Data Warehouse

02

Los datos estructurados derivados de los historiales médicos electrónicos, como los relacionados con los hábitos alimenticios y la condición física, serán almacenados en un Data Warehouse. Este sistema de gestión de datos está optimizado para ofrecer un acceso rápido y eficiente a grandes volúmenes de información, facilitando así la realización de análisis históricos y la extracción de conocimiento.

"El Data Warehouse se construirá sobre una base sólida de tablas de hechos y dimensiones. Las tablas de hechos contendrán los datos cuantitativos, como las medidas de los hábitos alimenticios y los indicadores de la condición física.

Arquitecturas típicas de proyectos de datos masivos

Tratamiento de los datos

01

Limpieza

La limpieza de datos es un proceso esencial para garantizar la calidad de la información. En primer lugar, se identifican y eliminan aquellos registros que presentan datos incompletos o erróneos. A continuación, se procede a la detección de outliers, es decir, valores atípicos que pueden sesgar los resultados del análisis. Estos valores extremos, como pesos o alturas inusuales, deben ser revisados y corregidos si es necesario

02

Integración

En primer lugar, es fundamental unificar los datos provenientes de diversas fuentes, como historias clínicas, sensores de dispositivos y encuestas. Este proceso implica establecer correspondencias claras entre las variables de cada fuente, por ejemplo, entre los códigos de diagnóstico utilizados en diferentes sistemas de clasificación.

03

Preparación para análisis

La normalización de variables es fundamental para garantizar la precisión en los análisis de datos. Al asegurar que todas las unidades de medida sean consistentes, se evitan sesgos y se facilita la comparación entre variables. Por otro lado, la transformación de variables permite adaptar los datos a los requerimientos específicos de cada análisis.

Arquitecturas típicas de proyectos de datos masivos

Visualización

La creación de dashboards en **Tableau** ofrece una forma innovadora de analizar la obesidad. Al convertir datos numéricos en representaciones visuales, como gráficos y mapas, podemos comprender rápidamente los factores que influyen en los niveles de obesidad.

Tableau nos permite identificar tendencias, comparar grupos y descubrir relaciones entre variables de manera sencilla. Por ejemplo, podemos visualizar cómo los patrones de consumo de alimentos varían según la ubicación geográfica o cómo la actividad física influye en el índice de masa corporal. Esta herramienta es invaluable para transformar datos en conocimientos accionables y para comunicar los resultados de manera clara y concisa

Arquitecturas típicas de proyectos de datos masivos

Nombre de la variable	Rol	Tipo	Demográfico	Descripción	Valores faltantes
Género	Característica	Categórico	Género		No
Edad	Característica	Continuo	Edad		No
Altura	Característica	Continuo			No
Peso	Característica	Continuo			No
family_history_with_overweight	Característica	Binario		¿Algún miembro de la familia ha sufrido o padece sobrepeso?	No
FAVC	Característica	Binario		¿Comes alimentos ricos en calorías con frecuencia?	No
FCVC	Característica	Entero		¿Sueles comer verduras en tus comidas?	No
NCP	Característica	Continuo		¿Cuántas comidas principales tienes al día?	No
CAEC	Característica	Categórico		¿Comes algún alimento entre comidas?	No
HUMO	Característica	Binario		¿Usted fuma?	No
CH2O	Característica	Continuo		¿Cuánta agua bebes al día?	No
SCC	Característica	Binario		¿Controlas las calorías que ingieres a diario?	No
FAF	Característica	Continuo		¿Con qué frecuencia realiza actividad física?	No
MARTES	Característica	Entero		¿Cuánto tiempo utilizas dispositivos tecnológicos como celular, videojuegos, televisión, computadora y otros?	No
CALC	Característica	Categórico		¿Con qué frecuencia bebes alcohol?	No
MTRANS	Característica	Categórico		¿Qué medio de transporte utilizas habitualmente?	No
NObeyesdad	Blanco	Categórico		Nivel de obesidad	No

DataSet

Para la elaboración del **Dashboard** se utilizo como fuente un **dataset** descargado desde el repositorio de Aprendizaje automático de la Universidad de California Irvine.

Este conjunto de datos incluye datos para la estimación de los niveles de obesidad en individuos de los países de México, Perú y Colombia, con base en sus hábitos alimenticios y condición física.

Los datos contienen 17 atributos y 2111 registros, los registros están etiquetados con la variable de clase NObesity (Nivel de Obesidad), que permite clasificar los datos utilizando los valores de Peso Insuficiente, Peso Normal, Sobrepeso Nivel I, Sobrepeso Nivel II, Obesidad Tipo I, Obesidad Tipo II y Obesidad Tipo III

Sección 4

Perfil del científico de datos

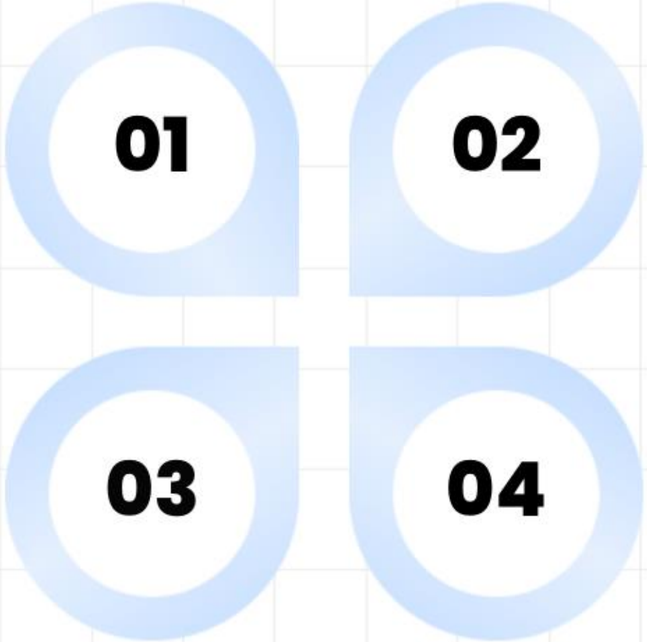
Perfil del científico de datos

Ciencias de la computación

Para esta posición, se requieren sólidos conocimientos en procesamiento de datos, incluyendo técnicas de integración y manejo de herramientas ETL como Apache NiFi o Talend. Será necesario dominar lenguajes de programación como Python o R, junto con bibliotecas especializadas como pandas y numpy, para llevar a cabo tareas de limpieza y análisis de datos.

Comunicación

El primer objetivo es elaborar un informe ejecutivo conciso y comprensible que sintetice los hallazgos clave del análisis. Es fundamental que el informe destaque cómo los datos obtenidos pueden influir de manera directa en la toma de decisiones médicas, proporcionando una base sólida para implementar estrategias y tratamientos más efectivos.



Matemáticas

se emplearán técnicas estadísticas como el análisis de regresión para identificar relaciones significativas entre el tratamiento administrado y la evolución de la salud de los pacientes. Asimismo, se utilizarán técnicas de clustering para agrupar a los pacientes en base a características comunes, lo que permitirá predecir de forma más precisa su respuesta a tratamientos específicos.

Negocios

Enfocar el proyecto hacia la mejora de la eficiencia en el tratamiento de enfermedades crónicas, lo cual tiene un impacto directo en la reducción de costos para las instituciones de salud y mejora en la calidad de vida de los pacientes.

Sección 5

Estrategias en almacenamiento masivo

Estrategias en almacenamiento masivo

Consideraciones para la Selección de Soluciones de Almacenamiento

Costo

Rendimiento

Escalabilidad

Seguridad

Integración

Estrategias en almacenamiento masivo

01

Data Mart

El diseño de un Data Mart para una enfermedad crónica como la diabetes tipo 2 es fundamental para la investigación y la gestión clínica. Al integrar tablas con datos demográficos, condición médica, tratamientos y evolución clínica, podemos obtener una visión holística de cada paciente.

02

Data Warehouse

Crear un Data Warehouse para almacenar los datos consolidados de todos los pacientes con enfermedades crónicas, estructurados para facilitar el análisis histórico y la toma de decisiones a nivel macro en la institución de salud.

03

Data Lake

Para facilitar el análisis de datos en su estado más crudo, como los provenientes de dispositivos wearables o encuestas, se utilizará un Data Lake. Este repositorio centralizado almacenará grandes volúmenes de datos no estructurados. Tanto AWS Lake Formation como Azure Data Lake Storage ofrecen soluciones robustas para la creación y gestión de estos lagos de datos en la nube.

Sección 6

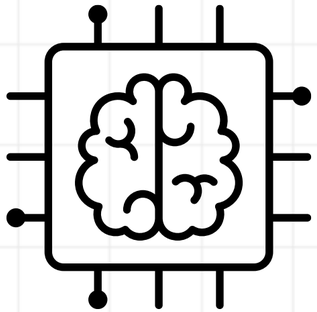
Estrategias de aplicación de la ciencia de datos y datos masivos

Estrategias de aplicación de la ciencia de datos y datos masivos

01

Inteligencia de negocio

Aplicar inteligencia de negocio para identificar qué tratamientos han sido más efectivos en el control de enfermedades crónicas y, a partir de allí, optimizar los recursos de salud



02

Analítica de negocio

Análisis de los datos clínicos para detectar patrones y prever complicaciones en pacientes con enfermedades crónicas, lo que permitirá tomar decisiones preventivas

Estrategias de aplicación de la ciencia de datos y datos masivos

Minería de datos

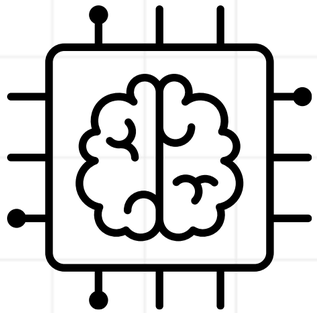
Utilizar minería de datos para descubrir patrones ocultos en los historiales médicos y en los datos de los dispositivos de monitoreo, como la relación entre el nivel de actividad física y el control de la diabetes

03

04

Aprendizaje automático

Desarrollar modelos predictivos utilizando aprendizaje automático para predecir el riesgo de complicaciones en pacientes con enfermedades crónicas. Por ejemplo, predecir la probabilidad de que un paciente con diabetes desarrolle insuficiencia renal.



Estrategias de aplicación de la ciencia de datos y datos masivos

Inteligencia artificial

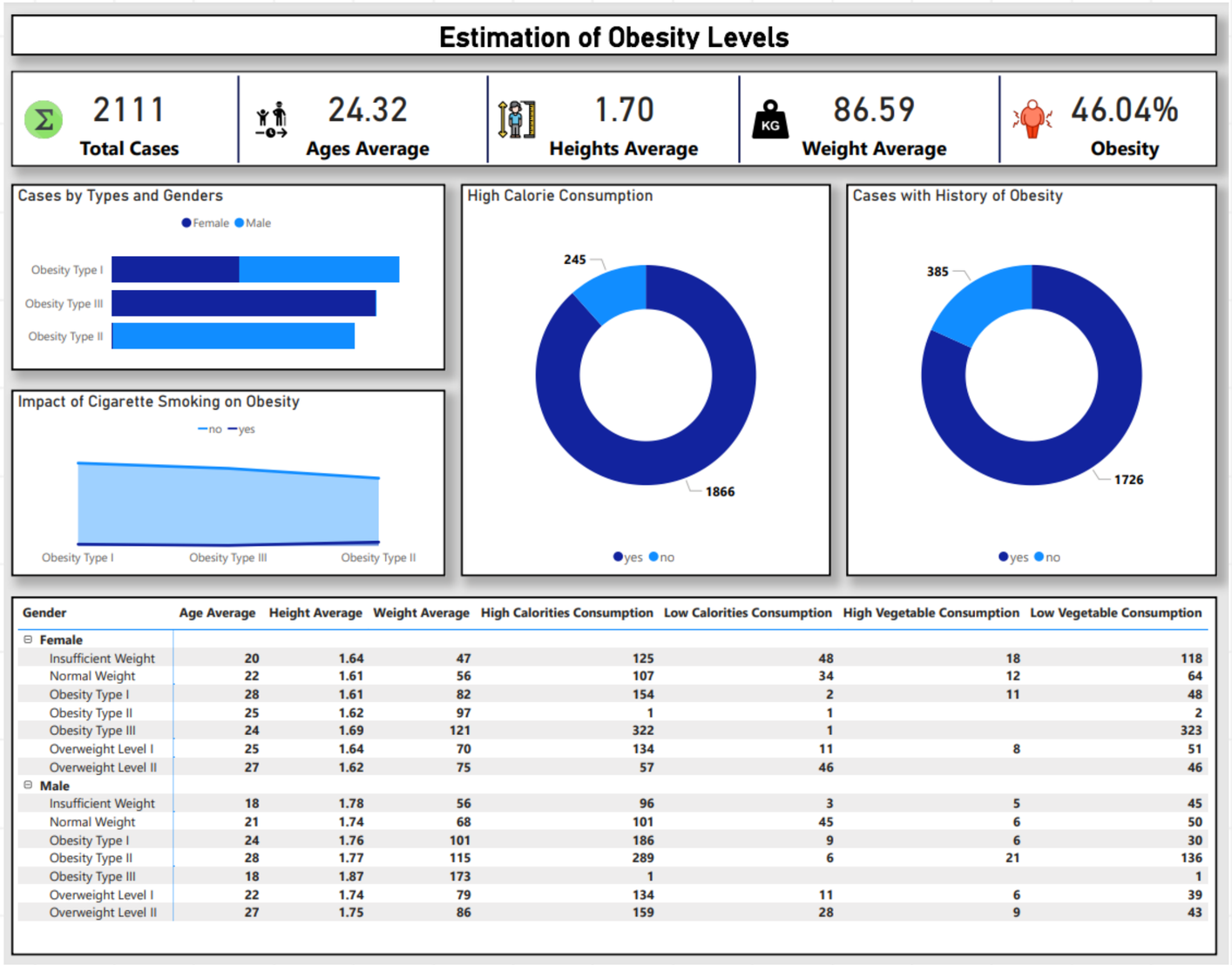
Explorar cómo la inteligencia artificial podría automatizar la personalización de tratamientos, utilizando los datos históricos de pacientes para recomendar tratamientos específicos según las características de cada individuo

05

Sección 7

Visualización de Datos

Dashboard



Gender	Age Average	Height Average	Weight Average	High Calorities Consumption	Low Calorities Consumption	High Vegetable Consumption	Low Vegetable Consumption
Female							
Insufficient Weight	20	1.64	47	125	48	18	118
Normal Weight	22	1.61	56	107	34	12	64
Obesity Type I	28	1.61	82	154	2	11	48
Obesity Type II	25	1.62	97	1	1		2
Obesity Type III	24	1.69	121	322	1		323
Overweight Level I	25	1.64	70	134	11	8	51
Overweight Level II	27	1.62	75	57	46		46
Male							
Insufficient Weight	18	1.78	56	96	3	5	45
Normal Weight	21	1.74	68	101	45	6	50
Obesity Type I	24	1.76	101	186	9	6	30
Obesity Type II	28	1.77	115	289	6	21	136
Obesity Type III	18	1.87	173	1			1
Overweight Level I	22	1.74	79	134	11	6	39
Overweight Level II	27	1.75	86	159	28	9	43

Cards:

1. Cantidad total de casos o muestras
2. Promedio de edades
3. Promedio de estaturas
4. Promedio de peso
5. Porcentaje de obesidad en la muestra

Visuales:

1. Casos de obesidad por tipo y que cantidad es hombre y mujer
2. Impacto del cigarrillo en el tipo de obesidad
3. Altos consumos de calorías versus bajo consumo
4. Impacto de los antecedentes en la familia

Tabla: por genero, que en que estado de peso se encuentra, en base a esos estados:

1. Promedio de edades
2. Promedio de estaturas
3. Promedio de peso alto consumo de calorías
4. Bajo consumo de calorías

Sección 8

Conclusiones

Conclusiones

Área de Salud

01

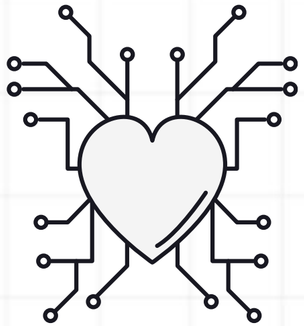
El análisis de datos se ha convertido en una herramienta indispensable en el sector de la salud, permitiendo una toma de decisiones más informada y personalizada. A través de la minería de grandes volúmenes de datos clínicos, genómicos y epidemiológicos, es posible identificar patrones, predecir riesgos y desarrollar tratamientos más efectivos. Esta disciplina ha revolucionado la investigación médica, la gestión de enfermedades crónicas y la optimización de recursos, contribuyendo significativamente a mejorar la calidad de vida de los pacientes.

Ciencias de Datos

02

La práctica y experiencia en el análisis de datos demuestran que esta disciplina va más allá de la simple manipulación de números. Requiere una combinación de habilidades técnicas, pensamiento crítico y conocimiento del dominio específico. La experiencia acumulada permite a los analistas de datos desarrollar intuiciones valiosas, identificar las preguntas correctas y seleccionar las herramientas adecuadas para cada problema. Asimismo, la colaboración con expertos de otras áreas es fundamental para garantizar la relevancia y el impacto de los resultados obtenidos

Gracias por su atención



Compromiso con la Salud

Agradecemos su atención y apoyo; juntos, podemos impulsar la ciencia de datos en salud, transformando información en acciones efectivas que mejoren la calidad de vida y promuevan un futuro más saludable para todos.