PROYECTO DE DATOS MASIVOS Y CIENCIA DE DATOS

GESTIÓN Y ANÁLISIS DE GRANDES VOLUMENES DE DATOS EN EL SECTOR SALUD.

Descripción breve

El desarrollo del proyecto consiste en analizar información medica referente a los niveles de obesidad en función a los hábitos alimenticios y condición física.

Integrantes:

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

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)".

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. En las últimas décadas, el incremento de la obesidad ha sido impulsado por cambios sustanciales en los estilos de vida, entre los que se incluyen hábitos alimenticios poco saludables, el aumento del sedentarismo, y factores genéticos. La prevalencia de la obesidad en distintas poblaciones y el seguimiento de su evolución son esenciales para el desarrollo de intervenciones efectivas.

Este 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. La clave de este estudio está en la identificación de patrones ocultos que permitan desarrollar predicciones sobre el riesgo de obesidad, así como ofrecer intervenciones personalizadas.

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, para identificar tendencias comunes en los hábitos alimenticios y la condición física de individuos con obesidad.

- **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.

- **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.

- **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.

Diseño del proyecto

Arquitecturas típicas de proyectos de datos masivos:

Fuentes heterogéneas:

- **Historiales médicos electrónicos (HME):** Datos estructurados que incluyen información como diagnósticos, tratamientos, historial de visitas, etc.
    - Antecedentes familiares: Historia de obesidad en la familia.
    - Resultados de laboratorio: Niveles de colesterol, triglicéridos, hormonas, etc.
    - Notas clínicas: Observaciones detalladas de los médicos sobre el estado de salud del paciente y sus hábitos.
    - Datos estructurados que incluyen información como diagnósticos, tratamientos, historial de visitas, etc.

- **Sensores de dispositivos de monitoreo (wearables):** Datos semiestructurados o no estructurados provenientes de dispositivos como relojes inteligentes que monitorizan parámetros como la presión arterial, ritmo cardíaco y niveles de glucosa.
    - Datos de actividad física: Número de pasos, distancia recorrida, calorías quemadas.
    - Datos de sueño: Duración y calidad del sueño.

    - Datos de ubicación: Zonas frecuentadas (por ejemplo, restaurantes de comida rápida)
    - Datos semiestructurados o no estructurados provenientes de dispositivos como relojes inteligentes que monitorizan parámetros como la presión arterial, ritmo cardíaco y niveles de glucosa.

- **Encuestas de salud y bienestar:** Datos semiestructurados obtenidos a través de encuestas administradas a los pacientes sobre su bienestar, estilo de vida y hábitos alimenticios.
    - Hábitos alimenticios detallados: Frecuencia de consumo de diferentes grupos de alimentos, tamaño de las porciones, patrones de alimentación.
    - Nivel de actividad física: Tipo de ejercicio, frecuencia e intensidad.
    - Factores psicosociales: Estrés, depresión, ansiedad.
    - Hábitos de sueño: Horarios de sueño, dificultades para dormir.
    - Datos semiestructurados obtenidos a través de encuestas administradas a los pacientes sobre su bienestar, estilo de vida y hábitos alimenticios.

Fuentes Adicionales:

- Imágenes médicas: Resonancias magnéticas, tomografías computarizadas, radiografías, que pueden ayudar a evaluar la composición corporal y la presencia de comorbilidades.
- Datos de registros electrónicos de salud (RES): Información más allá de los HME, como reclamaciones de seguros, datos de farmacia, etc.
- Datos de sensores ambientales: Temperatura, humedad, niveles de contaminación, que podrían influir en los hábitos alimenticios y la actividad física.

Extracción, transformación y carga (ETL):

- **Extracción:**
    - Utilizar APIs para obtener datos de sistemas de historiales médicos electrónicos, datos de wearables a través de sus respectivas plataformas.
    - Historiales médicos electrónicos: Además de las APIs, podemos considerar la extracción directa de bases de datos, siempre y cuando existan los permisos necesarios. Es importante definir qué datos son

relevantes, como peso, talla, mediciones de cintura, resultados de análisis encuestas de salud de plataformas en línea.

- Redes sociales: Utilizar técnicas de web scraping para extraer información de plataformas como Twitter o Instagram, donde los usuarios suelen compartir sus hábitos alimenticios y rutinas de ejercicio.

- **Transformación:** Limpiar los datos, manejar valores nulos, convertir formatos y unificar las fuentes de datos. Por ejemplo, transformar las lecturas de dispositivos en formatos estandarizados.
  - Limpieza: Identificar y corregir errores en los datos, como valores atípicos, duplicados y datos inconsistentes.
  - Normalización: Convertir los datos a un formato común para facilitar el análisis. Por ejemplo, unificar las unidades de medida (kilogramos, libras), estandarizar los códigos de diagnóstico y codificar las variables categóricas (género, nivel de educación).
  - Enriquecimiento: Agregar datos contextuales relevantes, como información demográfica (edad, sexo, ubicación geográfica) o datos climáticos (temperatura, humedad), que podrían influir en los hábitos alimenticios y la actividad física.

- **Carga:** Almacenar los datos transformados en un Data Warehouse o Data Lake para su posterior análisis.

Almacenamiento:

- **Data Lake:** Los datos no estructurados o semi-estructurados (como los datos de los wearables o encuestas) se almacenarían en un Data Lake debido a la flexibilidad que estos proporcionan.

  Componentes:

  - Datos Crudos: Datos sin procesar directamente extraídos de las fuentes.
  - Datos Procesados: Datos que han sido transformados y limpiados para análisis específicos.

- **Data Warehouse:** Los datos estructurados provenientes de los historiales médicos electrónicos serían almacenados en un Data Warehouse, optimizado para consultas rápidas y análisis históricos.

Componentes:

- **Tablas de Hechos**: Contienen los datos cuantitativos sobre los hábitos alimenticios y la condición física.
- **Tablas de Dimensiones**: Contienen datos descriptivos como información demográfica, tipos de alimentos, y categorías de actividad física.

Tratamiento de los datos:

- **Limpieza:** Identificar y eliminar registros incompletos o erróneos.
  - Detección de outliers: Identificar valores atípicos que puedan distorsionar los resultados, como pesos o alturas extremadamente altos o bajos.
  - Corrección de errores: Corregir errores de digitación o inconsistencias en los datos, por ejemplo, fechas de nacimiento incorrectas o valores duplicados.
  - Imputación de datos faltantes: Rellenar los valores faltantes utilizando diferentes técnicas, como la media, la mediana o modelos de imputación más sofisticados.

- **Integración:** Unir los datos de las diferentes fuentes (historias clínicas, sensores de dispositivos y encuestas).
  - Mapeo de variables: Establecer correspondencias entre las variables de diferentes fuentes de datos, por ejemplo, entre los códigos de diagnóstico de diferentes sistemas de clasificación.
  - Creación de una base de datos unificada: Consolidar todos los datos en una única base de datos para facilitar el análisis.
  - Resolución de conflictos: Manejar las discrepancias entre los datos de diferentes fuentes, por ejemplo, si un paciente tiene diferentes alturas registradas en distintos sistemas. Unir los datos de las diferentes fuentes (historias clínicas, sensores de dispositivos y encuestas).

- **Preparación para análisis:** Normalizar las variables para asegurar que los análisis posteriores sean precisos, por ejemplo, asegurando que todas las unidades de medida sean consistentes.

- Normalización: Normalizar las variables para asegurar que los análisis posteriores sean precisos, por ejemplo, asegurando que todas las unidades de medida sean consistentes.
- Transformación de variables: Convertir las variables en un formato adecuado para el análisis, por ejemplo, categorizar variables continuas en grupos discretos.

Visualización:

- **Herramienta:** Utilizar **Tableau** para crear dashboards que permitan visualizar los indicadores que pueden estimar los niveles de obesidad de acuerdo con patrones de comportamiento en cuanto a hábitos alimenticios y condición física.

Un Dashboard, creado con la herramienta Tableau nos permitirá:

- Transformar datos en imágenes: En lugar de filas y columnas de números, podrás ver gráficas, mapas y otros elementos visuales que representan la información de manera clara y concisa.
- Identificar patrones y tendencias: Los gráficos nos ayudará a descubrir relaciones entre diferentes variables, como, por ejemplo, si existe una correlación entre el consumo de alimentos procesados y el aumento de peso.
- Comparar grupos: Podremos lograr comparar los hábitos alimenticios y la condición física de diferentes grupos de personas (por ejemplo, hombres vs. mujeres, diferentes rangos de edad) para identificar diferencias significativas.
- Contar historias con los datos: La visualización nos permitirá comunicar los hallazgos de manera efectiva a otros investigadores, profesionales de la salud y al público en general.

2. Perfil del científico de datos:

Ciencias de la computación:

- Necesitarás conocimientos en procesamiento de datos, técnicas de integración y herramientas ETL (como Apache Nifi o Talend).
- Uso de lenguajes de programación como Python o R para la limpieza y análisis de los datos, junto con bibliotecas como pandas y numpy.
- Conocimiento en procesamiento de datos, técnicas de NLP para analizar el texto en redes sociales, y herramientas de análisis de datos como Python, R, o plataformas como Hadoop.

- Experiencia en el uso de herramientas ETL como Apache Nifi o Talend para integrar datos provenientes de diferentes fuentes.

Matemáticas:

- Se aplicarán técnicas estadísticas como análisis de regresión para identificar correlaciones entre el tratamiento y la evolución de los pacientes.
- Uso de técnicas de clustering para segmentar a los pacientes según características similares y predecir su respuesta a ciertos tratamientos.
- Aplicación de modelos estadísticos y técnicas de aprendizaje automático para la predicción de brotes. Modelos como regresión logística o modelos de series temporales pueden ser utilizados para predecir la probabilidad de un brote en función de diversos factores.
- Técnicas de análisis espacial para entender cómo la movilidad afecta a la propagación de las enfermedades.

Comunicación:

- Crear un informe ejecutivo que explique de manera clara los hallazgos del análisis, destacando cómo los datos pueden influir en la toma de decisiones médicas.
- Diseño de presentaciones visuales interactivas para que médicos y responsables de la toma de decisiones en el área de salud comprendan fácilmente los resultados.

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.

3. Estrategias en almacenamiento masivo:

Data Mart:

- Diseñar un Data Mart para un área específica, como por ejemplo, para la diabetes tipo 2, que contenga tablas como:
    - **Datos Demográficos** (edad, género, ubicación).
    - **Condición médica** (diagnósticos de diabetes, complicaciones).
    - **Tratamientos aplicados** (medicación, cambios en el estilo de vida).
    - **Evolución clínica** (resultados de glucosa en sangre, presión arterial).

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.

Data Lake:

- Los datos no estructurados, como los obtenidos de wearables o encuestas, se almacenarán en un Data Lake para facilitar el análisis de estos datos en su forma bruta.
    - **AWS Lake Formation**: Facilita la configuración de un lago de datos seguro y escalable en AWS.
    - **Azure Data Lake Storage**: Ofrece almacenamiento de datos masivos y análisis de big data en la nube.

Nuevas tendencias en almacenamiento masivo:

- Explorar el uso de almacenamiento en la nube para reducir costos y facilitar la accesibilidad remota de los datos de los pacientes.

Almacenamiento en la Nube

- **Amazon Web Services (AWS)**
    - **Amazon S3:** Ofrece almacenamiento escalable y seguro para cualquier cantidad de datos. Ideal para almacenar grandes volúmenes de datos y acceder a ellos desde cualquier lugar.
    - **Amazon Redshift:** Almacén de datos en la nube que permite realizar análisis complejos y consultas rápidas sobre grandes conjuntos de datos.
- **Google Cloud Platform (GCP)**
    - **Google Cloud Storage:** Almacenamiento unificado de objetos que permite almacenar y acceder a datos de manera segura y escalable.
    - **BigQuery:** Almacén de datos totalmente gestionado que permite realizar análisis de datos a gran escala con SQL.
- **Microsoft Azure**
    - Azure Blob Storage: Servicio de almacenamiento de objetos optimizado para almacenar grandes cantidades de datos no estructurados.
    - Azure Synapse Analytics: Plataforma de análisis que combina almacenamiento de datos y análisis de big data.

Consideraciones para la Selección de Soluciones de Almacenamiento

1. **Escalabilidad:** Asegúrate de que la solución pueda crecer con tus necesidades de almacenamiento de datos.
2. **Seguridad:** Evalúa las medidas de seguridad ofrecidas, como el cifrado de datos y el control de acceso.
3. **Costo:** Considera el costo total de propiedad, incluyendo almacenamiento, transferencia de datos y costos de acceso.
4. **Integración:** Verifica la compatibilidad con las herramientas ETL y de análisis que planeas utilizar.
5. **Rendimiento:** Asegúrate de que la solución ofrezca el rendimiento necesario para tus cargas de trabajo de análisis de datos.

4. Estrategias de aplicación de la ciencia de datos y datos masivos:

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.

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.

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.

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.

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.