"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)"

Arquitecturas típicas de proyectos de datos masivos:

Fuentes heterogéneas:

- Historiales médicos electrónicos (HME):
    - Antecedentes familiares: Historia de obesidad en la familia.
    - Datos demográficos: Edad, sexo, raza, etnia, nivel socioeconómico.
    - Medicamentos prescritos: Fármacos relacionados con el control de peso o enfermedades metabólicas.
    - Resultados de laboratorio: Niveles de colesterol, triglicéridos, hormonas, etc.
    - Procedimientos médicos: Cirugías bariátricas, endoscopias, Método POSE Técnica restrictiva ya que modifica la capacidad gástrica, Balón intragástrico: Al igual que el método POSE el balón intragástrico se introduce en el estómago por vía endoscópica sin necesidad de cirugía.Banda gástrica: Técnica que consiste en la colocación de un anillo o banda con un calibre ajustable en la parte superior del estómago, Gastrectomía tubular “Manga gástrica” o sleeve gastrectomy que ha revolucionado las técnicas restrictivas ya que puede ser también considerada como técnica mixta o malabsortiva, la cual desempeña una gran eficacia en la reducción de peso de pacientes con obesidad severa a largo plazo.
    - 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 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:
  - 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.

- Datos de redes sociales: Información sobre las interacciones sociales del paciente y su exposición a contenido relacionado con la alimentación y la salud.
- Datos de ventas de alimentos: Información sobre los productos que compra el paciente y su frecuencia de compra.

- Consideraciones Importantes:
  - Calidad de los datos: Es fundamental evaluar la calidad, precisión y consistencia de los datos provenientes de diferentes fuentes, nos preocuparemos en mantener una privacidad y ética para asegurar el cumplimiento de las regulaciones de protección de datos y obtener el consentimiento informado de los pacientes, logrando una integración de datos, con el fin de desarrollar una estrategia para integrar y armonizar los datos provenientes de fuentes heterogéneas, lo que implica la transformación y limpieza de los datos, permitiendo garantizar un análisis estadístico apropiado que nos permita identificar patrones, correlaciones y relaciones causales entre las variables y para resultado final lograr componer una visualización de datos, que permita crear visualizaciones claras y concisas para comunicar los resultados de manera efectiva y apoyar en una mejor toma de desiciones para el campo y resolver la problemática de obesidad.

Extracción, transformación y carga (ETL):

- Extracción:
  - 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

de sangre (colesterol, triglicéridos), diagnósticos relacionados con la obesidad (diabetes, hipertensión) y tratamientos.

- Wearables: Explorar datos de actividad física (pasos, calorías quemadas), sueño, frecuencia cardíaca y datos biométricos como la composición corporal (si están disponibles).
- Encuestas de salud: Incluir preguntas sobre hábitos alimenticios detallados (frecuencia de consumo de diferentes alimentos, tamaño de las porciones), actividad física, consumo de alcohol y tabaco, historial familiar de obesidad, y factores socioeconómicos.
- 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:
  - 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:
  - Data Lake: Almacenar los datos en un formato raw para futuros análisis exploratorios y machine learning.
  - Data Warehouse: Crear un modelo de datos dimensional para análisis más estructurados y reportes.

- Bases de datos NoSQL: Considerar bases de datos como MongoDB para almacenar datos semiestructurados o no estructurados, como los obtenidos de las redes sociales.

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 ofrece para manejar diferentes tipos de datos.
- Data Warehouse: Los datos estructurados provenientes de los historiales médicos electrónicos serían almacenados en un Data Warehouse, optimizado para consultas rápidas y análisis históricos.
- Consideraciones Importantes: Estas dos estrategias de almacenamiento nos brindan una gran flexibilidad y consistencia en la gestión de nuestros datos, lo cual es fundamental al trabajar con una amplia variedad de información sobre hábitos alimenticios y estimación de niveles de obesidad. Al minimizar los datos duplicados e inconsistentes mediante procesos de limpieza y transformación, garantizamos la calidad y confiabilidad de nuestros análisis, permitiendo obtener resultados más precisos y relevantes.

  ¿Por qué ambas? Porque nos enfrentamos a una gran variedad de datos: estructurados, como los historiales médicos consolidados en bases de datos de hospitales y clínicas; y semi-estructurados o no estructurados, provenientes de redes sociales, encuestas y entrevistas (audio o video). Es necesario evaluar y consolidar todos estos datos para contrastar información y validar nuestro estudio sobre la estimación de niveles de obesidad en función de hábitos alimenticios y condición física. A través del análisis de grandes volúmenes de datos de historiales médicos electrónicos (HME), podremos obtener una visión más completa y precisa de la problemática.

Tratamiento de los datos:

- Limpieza:
    - 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:
    - 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 sistemasUnir los datos de las diferentes fuentes (historias clínicas, sensores de dispositivos y encuestas).
- Preparación para análisis:
    - 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.

- Creación de nuevas variables: Generar nuevas variables a partir de las existentes, como el índice de masa corporal (IMC) a partir del peso y la altura.
- Escalado de variables: Ajustar el rango de valores de las variables para que sean comparables, por ejemplo, utilizando la estandarización o la normalización

- Consideraciones Importantes:

  El tratamiento de datos es fundamental para obtener resultados precisos y confiables en el estudio de la obesidad a partir de historiales médicos electrónicos. Este proceso involucra varias etapas clave, la limpieza, se busca garantizar la calidad de los datos eliminando errores, valores atípicos e inconsistencias. Esto implica identificar y corregir información errónea, así como completar los datos faltantes utilizando métodos estadísticos. También la integración, es unir datos provenientes de diversas fuentes (historias clínicas, sensores, encuestas, redes sociales) en una única base de datos. Para ello, es necesario establecer correspondencias entre las variables y resolver cualquier discrepancia que pueda surgir, resaltando por ultimo pero no menos importante la preparación de los datos, estos se transforman y normalizan para que sean adecuados para el análisis estadístico. Esto implica convertir los datos en un formato uniforme, crear nuevas variables relevantes (como el IMC) y ajustar el rango de valores de las variables para facilitar las comparaciones de los mismos.

Visualización:

Debemos considerar que tendremos una enorme cantidad de datos sobre hábitos alimenticios, actividad física y condiciones de salud de un gran grupo de personas. Estos datos, por sí solos, pueden ser difíciles de entender y extraer conclusiones significativas. Aquí es donde entra en juego la visualización.

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.

Podemos generar visualizaciones útiles en este contexto:

- Gráficos de barras: Para comparar la frecuencia de diferentes hábitos alimenticios entre grupos de personas.
- Gráficos de línea: Para mostrar la evolución del peso a lo largo del tiempo en diferentes grupos de pacientes.
- Mapas de calor: Para visualizar la relación entre variables continuas, como el IMC y el consumo de calorías.
- Redes: Para representar las relaciones entre diferentes variables, como los hábitos alimenticios y el desarrollo de enfermedades crónicas.

- Consideraciones Importantes

La visualización hace que los datos sean más accesibles y fáciles de entender, incluso para personas sin conocimientos estadísticos avanzados. Podremos identificar patrones ocultos, que nos permite descubrir relaciones y tendencias que podrían pasar desapercibidas en un análisis puramente estadístico. Resaltando que comunicar los resultados es casi un arte y esto hay que hacerlo de manera efectiva, la visualización es una herramienta poderosa para comunicar los hallazgos de una investigación a un público más amplio, logrando el apoyo para la toma de decisiones, generando dashboards se pueden utilizar para monitorear el progreso de las intervenciones y tomar decisiones basadas en datos.

En resumen, utilizar Tableau para crear dashboards que permitan visualizar los indicadores que pueden estimar los niveles de obesidad de acuerdo a patrones de comportamiento en cuanto a habitos alimenticios y condición fisica la visualización de datos es una herramienta esencial para explorar y comprender la relación entre los hábitos alimenticios, la condición física y la obesidad. Al transformar los datos en imágenes, podemos identificar patrones, comunicar resultados y tomar decisiones más informadas para abordar este importante problema de salud pública.

2. Perfil del científico de datos:

Es complicado encontrar o incluso desarrollar un científico de datos especializado en salud y análisis sobre obesidad exclusivamente, pero enunciaremos las habilidades fundamentales que se desea para este estudio.

Se necesitará un profesional multidisciplinario que combina conocimientos sólidos en ciencias de la computación, matemáticas y estadística con una profunda comprensión del dominio de la salud. el objetivo principal es extraer conocimientos valiosos a partir de grandes volúmenes de datos de historiales médicos electrónicos (HME) para comprender mejor los factores que contribuyen a la obesidad y desarrollar estrategias de prevención y tratamiento más efectivas.

Habilidades y Conocimientos Clave

- Ciencias de la Computación:
  - Ingeniería de datos: Experiencia en la construcción de pipelines de datos robustos y escalables para recolectar, limpiar, transformar y cargar grandes conjuntos de datos heterogéneos.
  - Bases de datos: Conocimientos de bases de datos relacionales y no relacionales para almacenar y gestionar eficientemente los datos de salud.
  - Cloud computing: Habilidad para utilizar plataformas en la nube (AWS, GCP, Azure) para procesar y almacenar grandes volúmenes de datos.
  - Aprendizaje automático: Dominio de algoritmos de aprendizaje supervisado (regresión, clasificación) y no supervisado (clustering, reducción de dimensionalidad) para construir modelos predictivos y descubrir patrones en los datos.
  - Deep learning: Conocimientos básicos en redes neuronales profundas para abordar problemas complejos como el reconocimiento de imágenes médicas.
- Matemáticas y Estadística:
  - Estadística inferencial: Capacidad para realizar pruebas de hipótesis, construir intervalos de confianza y estimar parámetros poblacionales a partir de muestras
  - Análisis de series temporales: Experiencia en el análisis de datos que varían en el tiempo, como la evolución del peso a lo largo de varios años

    - Modelado estadístico: Habilidad para desarrollar modelos estadísticos complejos para explicar las relaciones entre diferentes variables.

- Dominio de la Salud:
    - Fisiología y nutrición: Conocimientos básicos sobre los procesos fisiológicos relacionados con la obesidad y los factores nutricionales que influyen en el peso.
    - Epidemiología: Comprensión de los conceptos epidemiológicos y la capacidad de interpretar estudios epidemiológicos sobre obesidad.
    - Terminología médica: Familiaridad con la terminología médica utilizada en los historiales clínicos electrónicos.

- Herramientas y Tecnologías:
    - Lenguajes de programación: Python (Pandas, NumPy, Scikit-learn), SQL.
    - Visualización de datos: Tableau para crear dashboards y gráficos informativos.
    - Herramientas de aprendizaje automático: TensorFlow, PyTorch

- 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.

- Habilidades Blandas:
    - Pensamiento crítico: Capacidad para analizar datos de manera crítica y extraer conclusiones significativas.
    - Comunicación: Habilidad para comunicar resultados complejos de manera clara y concisa a audiencias técnicas y no técnicas.

- Trabajo en equipo: Capacidad para colaborar con equipos multidisciplinarios de médicos, nutricionistas, ingenieros y otros profesionales de la salud.

- Funciones y Responsabilidades:
    - Recopilación y limpieza de datos: Obtener datos de diversas fuentes (HME, sensores, encuestas) y prepararlos para el análisis.
    - Exploración de datos: Identificar patrones, tendencias y relaciones en los datos utilizando técnicas de visualización y estadística descriptiva.
    - Modelado predictivo: Desarrollar modelos para predecir la probabilidad de desarrollar obesidad, la respuesta a diferentes tratamientos o la evolución del peso a largo plazo.
    - Evaluación de modelos: Evaluar la precisión y el rendimiento de los modelos utilizando métricas apropiadas.
    - Comunicación de resultados: Presentar los hallazgos de manera clara y concisa a través de informes, presentaciones y visualizaciones interactivas.
    - Colaboración con equipos multidisciplinarios: Trabajar en estrecha colaboración con médicos, nutricionistas y otros profesionales de la salud para traducir los hallazgos en acciones concretas

Consideraciones Importantes

El trabajo de un científico de datos especializado, sabemos que es complejo y mas en el área de la salud, consideramos que tiene un impacto significativo en la

mejora de la salud pública. Al identificar los factores de riesgo para la obesidad y desarrollar herramientas de predicción, se pueden diseñar intervenciones más personalizadas y efectivas para prevenir y tratar esta enfermedad. Además, los resultados de estas investigaciones pueden informar políticas públicas y guiar la asignación de recursos en el sector de la salud.

3. Estrategias en almacenamiento masivo:

Para este análisis de datos consideramos los siguientes puntos relacionados con el almacenamiento masivo:

Data Mart: Foco en la Obesidad

Un Data Mart dedicado a la obesidad permitiría un análisis más granular y específico de los datos relacionados con esta condición. Podría incluir tablas como:

- Datos antropométricos: Peso, altura, índice de masa corporal (IMC), perímetro de cintura.
- Hábitos alimenticios: Consumo de calorías, nutrientes, frecuencia de comidas, tipos de alimentos consumidos.
- Actividad física: Nivel de actividad física, tipo de ejercicio, tiempo dedicado al ejercicio.
- Factores socioeconómicos: Nivel educativo, ingresos, ocupación.
- Comorbilidades: Otras enfermedades asociadas a la obesidad (diabetes, hipertensión, enfermedades cardiovasculares).
- Intervenciones: Tratamientos farmacológicos, intervenciones quirúrgicas, programas de educación para la salud.

Beneficios del Data Mart:

- Agilidad en la consulta: Permite realizar consultas específicas sobre la obesidad de manera rápida y eficiente.
- Análisis detallado: Facilita la identificación de patrones y tendencias en los datos relacionados con la obesidad.
- Personalización: Se puede adaptar a las necesidades específicas de cada investigación o proyecto.

Data Warehouse: Visión Integral de la Salud

El Data Warehouse serviría como un repositorio central para todos los datos de salud, incluyendo aquellos relacionados con la obesidad. Además de los datos mencionados en el Data Mart, podría incluir:

- Datos genómicos: Información genética relacionada con la predisposición a la obesidad.
- Datos de imágenes médicas: Resonancias magnéticas, tomografías computarizadas, etc.
- Datos de dispositivos médicos: Información de sensores y dispositivos wearables.

Beneficios del Data Warehouse:

- Visión holística: Permite comprender la obesidad en el contexto de otras enfermedades y condiciones de salud.
- Análisis a largo plazo: Facilita el seguimiento de la evolución de la obesidad a lo largo del tiempo y la evaluación de la efectividad de las intervenciones.
- Inteligencia artificial: Permite aplicar técnicas de aprendizaje automático para descubrir patrones complejos y hacer predicciones.

Data Lake: Almacenamiento de Datos Brutos y No Estructurados

El Data Lake sería el repositorio ideal para almacenar datos no estructurados como:

- Datos de redes sociales: Información sobre hábitos alimenticios y actividad física compartida en redes sociales.
- Datos de sensores: Datos de acelerómetros, podómetros y otros dispositivos wearables.
- Datos de imágenes: Fotos de alimentos consumidos, registros de actividad física.
- Datos de audio: Grabaciones de entrevistas con pacientes.

Beneficios del Data Lake:

- Flexibilidad: Permite almacenar datos en su formato original sin necesidad de una estructura predefinida.
- Escalabilidad: Puede almacenar grandes volúmenes de datos de manera eficiente.
- Análisis exploratorio: Facilita la exploración de los datos y el descubrimiento de nuevos insights.

Nuevas Tendencias en Almacenamiento Masivo: La Nube

El almacenamiento en la nube ofrece varias ventajas para el almacenamiento de datos de salud:

- Escalabilidad: Permite ajustar la capacidad de almacenamiento según las necesidades.
- Accesibilidad: Facilita el acceso a los datos desde cualquier lugar y dispositivo con conexión a internet.
- Seguridad: Ofrece robustas medidas de seguridad para proteger los datos de los pacientes.

- Costo-efectividad: Permite pagar solo por el almacenamiento utilizado.

Beneficios específicos para la investigación de la obesidad:

- Colaboración: Facilita la colaboración entre investigadores de diferentes instituciones.
- Análisis en tiempo real: Permite realizar análisis en tiempo real de los datos generados por dispositivos wearables.
- Inteligencia artificial: Facilita la aplicación de técnicas de aprendizaje automático a gran escala

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.