## Prevención de la diabetes

Semana 3

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Grupo 22**

**Integrantes:**

Angie Camila Valdes Fuentes

Diana Camila Cordoba Arenas

Robert Kenzo Medina Monsalve

Julian David Hernandez Correcha

## Despliegue de soluciones analíticas

UNIVERSIDAD DE LOS ANDES

2024

Prevención de diabetes 

# Problema a abordar:

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo y representa una carga significativa tanto para los sistemas de salud como para las economías nacionales. La detección y el manejo temprano son fundamentales para reducir las complicaciones a largo plazo y mejorar la calidad de vida de los pacientes. En este contexto, el uso de modelos predictivos se presenta como una herramienta eficaz para identificar a individuos con alto riesgo de desarrollar diabetes antes de que presenten síntomas clínicos evidentes. Mediante el análisis de datos de salud, demográficos y de estilo de vida, es posible anticipar la incidencia de la diabetes y personalizar intervenciones de prevención.

Con esto, se busca en un contexto organizacional facilitar la implementación de estrategias de prevención temprana y personalizada. Al prever quiénes tienen un mayor riesgo de desarrollar la enfermedad, las organizaciones pueden dirigir recursos y programas de intervención de manera más efectiva, como promover cambios en el estilo de vida, fomentar la actividad física, ofrecer educación en nutrición y brindar seguimiento médico adecuado. Esto tiene como objetivo último reducir la incidencia de la diabetes y sus complicaciones asociadas, mejorando así la salud y el bienestar de la población y, a su vez, reduciendo los costos asociados al tratamiento de la enfermedad.

**Pregunta de negocio y alcance del proyecto.**

¿Cómo pueden los modelos predictivos basados en indicadores de salud, hábitos de vida, y factores socioeconómicos, contribuir a la prevención de la diabetes al predecir con precisión la probabilidad de que un individuo desarrolle la enfermedad, permitiendo así la implementación de intervenciones preventivas personalizadas y dirigidas?

Este proyecto busca desarrollar un modelo predictivo robusto capaz de analizar datos de salud, demografía y estilo de vida para evaluar el riesgo de diabetes en una población determinada. A través de un enfoque de ciencia de datos, se pretende no solo identificar individuos con alto riesgo de desarrollar la enfermedad, sino también proporcionar información valiosa para intervenciones personalizadas. El modelo final permitirá a organizaciones de salud anticipar la aparición de diabetes, dirigir recursos de manera eficiente y priorizar programas de prevención como asesoría en estilo de vida, monitoreo médico, educación nutricional y promoción de actividad física.

El proyecto se centra en tres objetivos clave:

1. **Desarrollo y validación del modelo**: Crear y evaluar un modelo de alta precisión que pueda aplicarse a poblaciones diversas.
2. **Integración de datos seguros y anonimización**: Garantizar que los datos personales sean tratados con total confidencialidad y que el procesamiento cumpla con las normativas de privacidad.
3. **Implementación práctica para la toma de decisiones**: Proveer un sistema de predicciones que las organizaciones de salud puedan usar para asignar recursos preventivos.

Este enfoque global permitirá reducir la incidencia de la diabetes, mejorar la calidad de vida de los pacientes y optimizar el uso de los recursos en salud pública.

**Descripción de conjuntos de datos a emplear**

¿Qué datos se recopilarán o crearán? El dataset utilizado proviene del Centro para el Control y la Prevención de Enfermedades (CDC) y contiene indicadores de salud relacionados con la diabetes, como hábitos de vida (actividad física, consumo de tabaco y dieta), datos demográficos (edad, género, nivel socioeconómico), y factores médicos (presión arterial, niveles de glucosa y antecedentes familiares de diabetes). El conjunto de datos incluye observaciones de individuos anónimos a lo largo de varios años, con la finalidad de identificar patrones y predecir el riesgo de diabetes en la población.

¿Cómo se recopilarán o crearán los datos? Los datos son parte de un estudio longitudinal del CDC y fueron recopilados mediante encuestas y exámenes médicos. El acceso a estos datos se obtiene a través del repositorio de UCI, utilizando la herramienta ucimlrepo en Python para descargar el dataset. Los datos serán tratados y preprocesados en entornos de desarrollo seguros y accesibles sólo por personal autorizado del equipo de investigación. Los valores nulos y datos anómalos serán tratados para garantizar la calidad y precisión de los análisis posteriores.

La comunicación de los resultados se llevará a cabo al finalizar el proyecto mediante una presentación formal en la que participará todo el equipo. Esta presentación incluirá gráficos, tablas y visualizaciones interactivas creadas con herramientas como Looker/Power Bi, para explicar los hallazgos de manera clara y comprensible, abordando cómo las predicciones del modelo pueden contribuir a intervenciones preventivas.

# ¿Se requieren restricciones para compartir datos?

Los resultados, como los modelos predictivos y las visualizaciones, se presentarán de manera agregada, evitando detalles que puedan reidentificar a los individuos. Las categorías y grupos (clusters) identificados en los datos no tendrán nombres que sugieren comportamientos o características específicas, maximizando así la anonimidad de los participantes.

Es fundamental capacitar a todos los miembros del equipo en estas políticas de privacidad y confidencialidad para asegurar el cumplimiento de las normativas y garantizar la integridad y confidencialidad de los datos a lo largo del proyecto.

**Responsables y recursos**

# ¿Quién será responsable de la gestión de los datos?

Los principales responsables del manejo de los datos serán los data scientists y data managers del equipo. Estos se encargarán de todo el procesamiento, pre-procesamiento, análisis y mantenimiento de los datos proporcionados por el CDC. Su responsabilidad incluye asegurar la calidad, integridad y seguridad de los datos a lo largo del proyecto.

Sin embargo, todos los integrantes del equipo pueden tener un rol activo en este proceso, ya que el manejo de datos está estrechamente vinculado con el análisis de los resultados y la interpretación de la información en el contexto del problema de negocio. El equipo multidisciplinario trabajará en conjunto para asegurar que se tomen decisiones informadas basadas en los datos, optimizando así el uso y manejo de estos.

# ¿Qué recursos necesitará para ejecutar su plan?

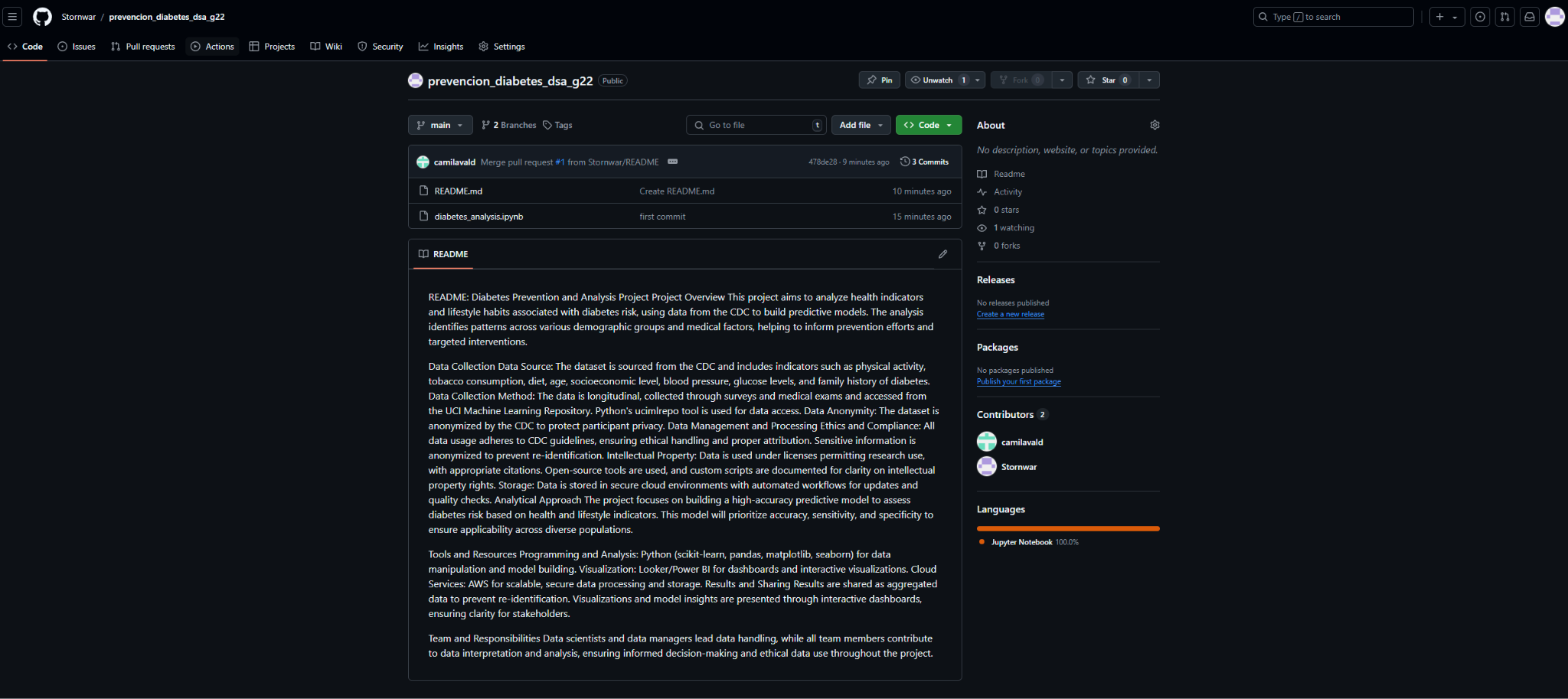
El equipo está capacitado para abordar la problemática con las herramientas necesarias para el desarrollo del proyecto. Se requerirá contar con un entorno de desarrollo adecuado (Python y sus bibliotecas especializadas como scikit-learn, pandas, matplotlib y seaborn) para realizar análisis de datos y modelado predictivo.

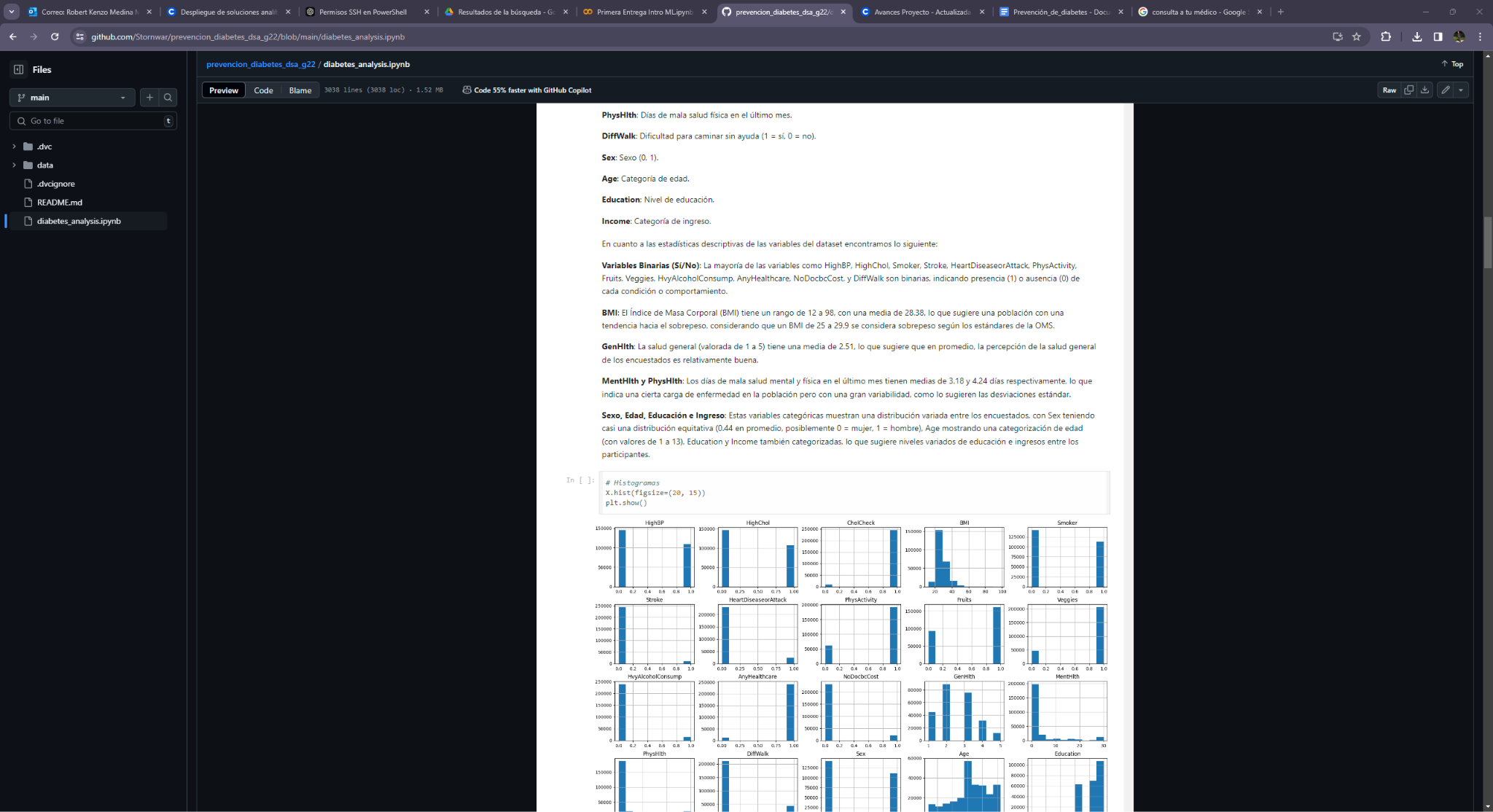
Además, se utilizarán herramientas de visualización de datos como Looker/Power Bi para crear dashboards interactivos que permitan analizar y comunicar los resultados de manera clara y efectiva.

En términos de almacenamiento y procesamiento de datos, se utilizarán servicios en la nube como AWS para garantizar el acceso seguro y escalable a los datos a lo largo del proyecto. Esto permitirá almacenar y procesar grandes volúmenes de datos de manera eficiente, así como integrar flujos de trabajo automatizados para la actualización y análisis de los datos.

## Repositorio Git en uso para el código

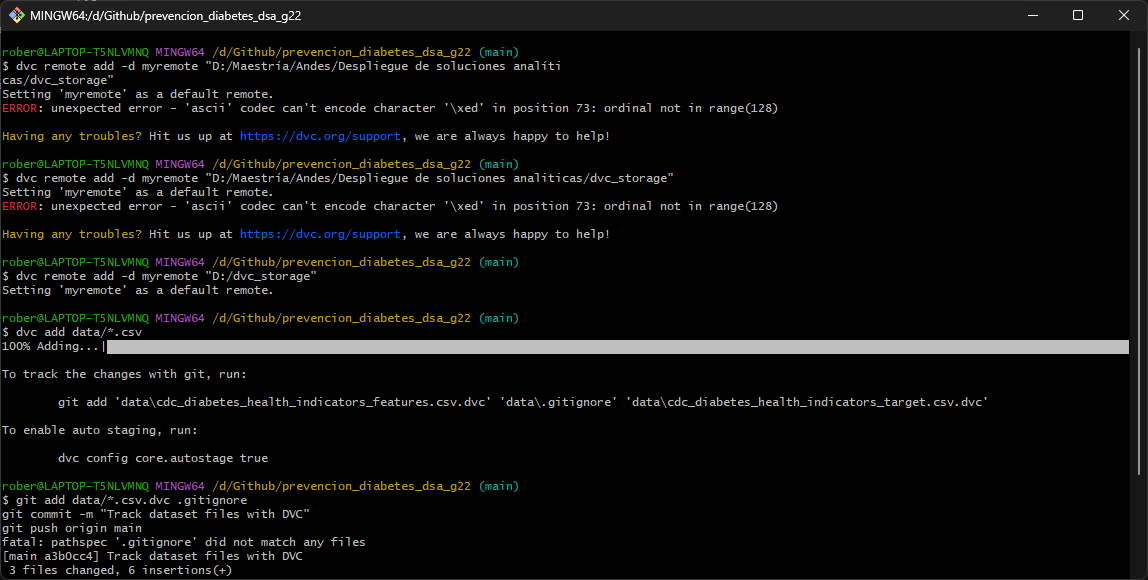
<https://github.com/Stornwar/prevencion_diabetes_dsa_g22>

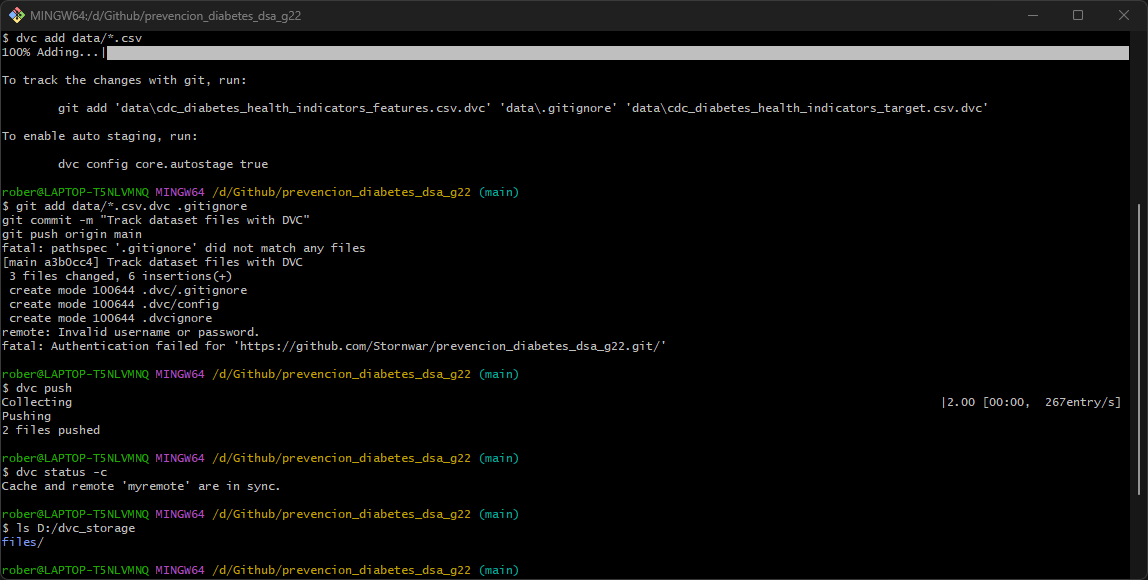


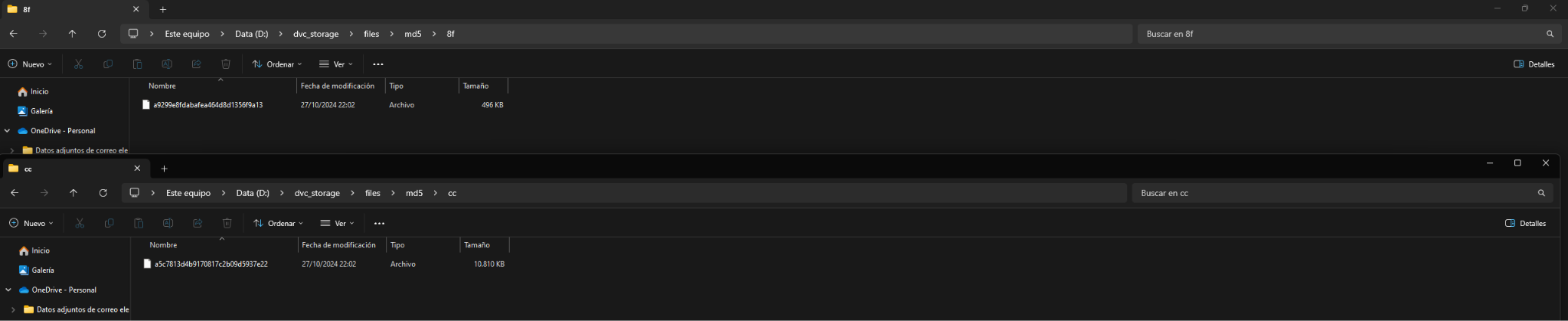


## Repositorio DVC en uso para los datos

Para esta entrega, el repositorio DVC se ha establecido localmente con los datos csv de los features y targets del dataset.









## Análisis descriptivo de datos

La base de datos está compuesta por:

Analizando cada variable, encontramos lo siguiente: **HighBP:** Presión arterial alta (1 = sí, 0 = no).

**HighChol:** Colesterol alto (1 = sí, 0 = no).

**CholCheck:** Chequeo de colesterol en los últimos 5 años (1 = sí, 0 = no).

**BMI:** Índice de masa corporal.

**Smoker:** Fumador (1 = sí, 0 = no).

**Stroke:** Ha tenido un derrame cerebral (1 = sí, 0 = no).

**HeartDiseaseorAttack:** Ha tenido enfermedad del corazón o un ataque al corazón (1 = sí, 0 = no).

**PhysActivity:** Actividad física (1 = sí, 0 = no).

**Fruits:** Consumo de frutas (1 = sí, 0 = no).

**Veggies:** Consumo de vegetales (1 = sí, 0 = no).

**HvyAlcoholConsump:** Consumo pesado de alcohol (1 = sí, 0 = no).

**AnyHealthcare:** Acceso a atención médica (1 = sí, 0 = no).

**NoDocbcCost:** No visitó al doctor debido al costo (1 = sí, 0 = no).

**GenHlth:** Salud general (valorado de 1 a 5).

**MentHlth:** Días de mala salud mental en el último mes.

**PhysHlth:** Días de mala salud física en el último mes.

**DiffWalk:** Dificultad para caminar sin ayuda (1 = sí, 0 = no).

**Sex:** Sexo (0, 1).

**Age:** Categoría de edad.

**Education:** Nivel de educación.

**Income:** Categoría de ingreso.

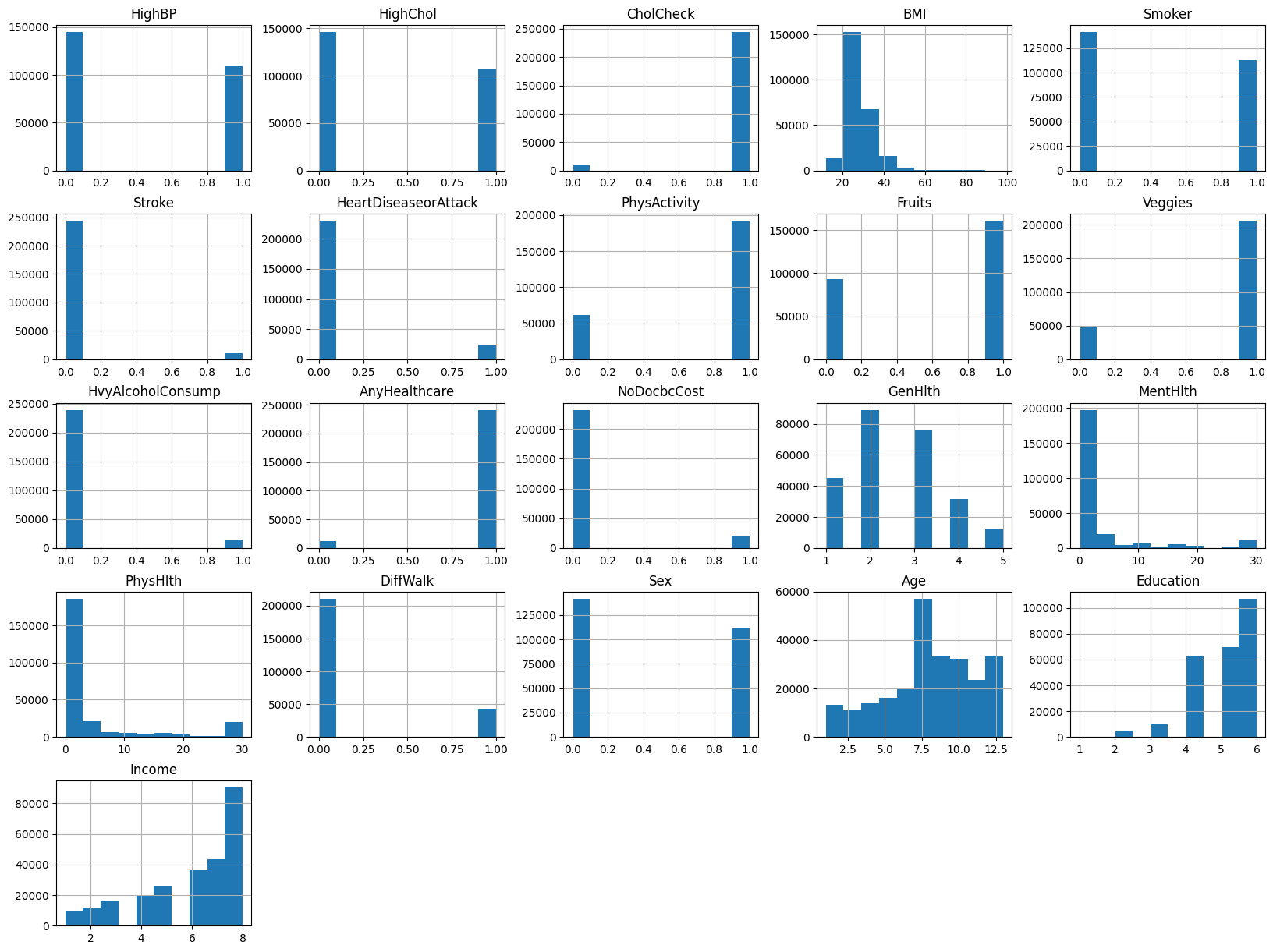
**Variables Binarias (Sí/No):** La mayoría de las variables como HighBP, HighChol, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, y DiffWalk son binarias, indicando presencia (1) o ausencia (0) de cada condición o comportamiento.

**BMI:** El Índice de Masa Corporal (BMI) tiene un rango de 12 a 98, con una media de 28.38, lo que sugiere una población con una tendencia hacia el sobrepeso, considerando que un BMI de 25 a 29.9 se considera sobrepeso según los estándares de la OMS.

**GenHlth:** La salud general (valorada de 1 a 5) tiene una media de 2.51, lo que sugiere que en promedio, la percepción de la salud general de los encuestados es relativamente buena.

**MentHlth y PhysHlth:** Los días de mala salud mental y física en el último mes tienen medias de 3.18 y 4.24 días respectivamente, lo que indica una cierta carga de enfermedad en la población pero con una gran variabilidad, como lo sugieren las desviaciones estándar.

**Sexo, Edad, Educación e Ingreso:** Estas variables categóricas muestran una distribución variada entre los encuestados, con Sex teniendo casi una distribución equitativa (0.44 en promedio, posiblemente 0 = mujer, 1 = hombre), Age mostrando una categorización de edad (con valores de 1 a 13), Education y Income también categorizadas, lo que sugiere niveles variados de educación e ingresos entre los participantes.



**Variables Binarias:** Las variables binarias como HighBP, HighChol, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, y DiffWalk muestran distribuciones que reflejan la proporción de respuestas afirmativas y negativas. Por ejemplo, se observa que una mayor proporción de participantes reporta tener acceso a atención médica (AnyHealthcare) y realizar actividad física (PhysActivity), mientras que una menor proporción reporta consumo pesado de alcohol (HvyAlcoholConsump) o haber tenido un derrame cerebral (Stroke).

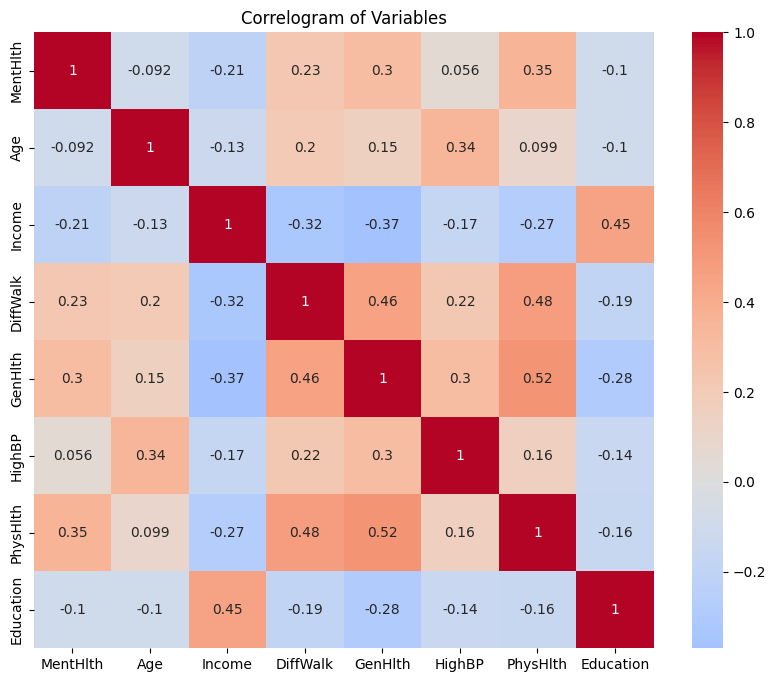
**BMI:** La distribución del Índice de Masa Corporal (BMI) parece sesgada hacia la derecha, indicando que una proporción significativa de la población tiene un BMI que se considera sobrepeso o incluso obesidad.

**Salud General (GenHlth):** La distribución de las percepciones de salud general muestra una tendencia hacia respuestas más positivas, con un pico en valores más bajos (mejor salud general).

**Días de Mala Salud Mental y Física (MentHlth y PhysHlth):** Ambas variables muestran una alta frecuencia de cero días reportados, lo que indica que muchos participantes no experimentaron días de mala salud mental o física en el último mes. Sin embargo, hay una cola larga hacia la derecha, lo que sugiere que una minoría reporta un número significativo de días afectados.

**Dificultad para Caminar (DiffWalk):** Una mayoría de participantes indica no tener dificultad para caminar, aunque hay una proporción no despreciable que sí reporta este problema.

**Variables Demográficas (Sex, Age, Education, Income):** Las distribuciones de estas variables sugieren una variada composición demográfica de los participantes. La edad muestra una distribución que sugiere una mayor proporción de participantes en categorías de edad media y avanzada. La educación y el ingreso muestran una variedad de niveles, con una distribución que permite inferir cierta relación entre nivel educativo y categoría de ingreso.



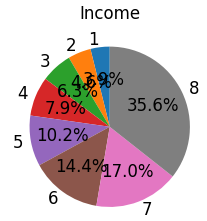
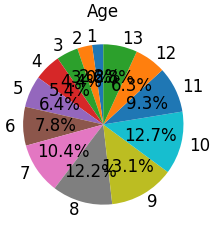
**DiffWalk** y **GenHlth** tienen una correlación positiva alta de 0.46, lo que sugiere que las personas con dificultades para caminar (DiffWalk) tienden a reportar un peor estado de salud general (GenHlth).

**PhysHlth** y **GenHlth** tienen una correlación positiva de 0.52, lo que indica que una peor salud física (PhysHlth) está asociada con una peor salud general (GenHlth).

**Income** y **Education** tienen una correlación positiva de 0.45, lo cual es común y sugiere que mayores niveles educativos suelen estar asociados con mayores ingresos.

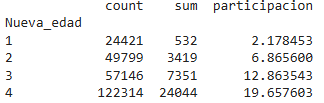
**Income** y **DiffWalk** tienen una correlación negativa de -0.32, lo que sugiere que los ingresos más altos están asociados con una menor probabilidad de tener dificultades para caminar.

**Income** y **GenHlth** tienen una correlación negativa de -0.37, lo cual indica que los ingresos más altos están asociados con una mejor salud general.

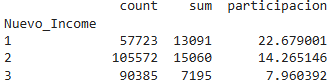


Se identifica que variables como age e income se pueden agrupar para poder tener una mejor interpretación de las variables a la hora de usar el modelo. La nueva agrupación se quedó de la siguiente manera:

**Nueva edad**



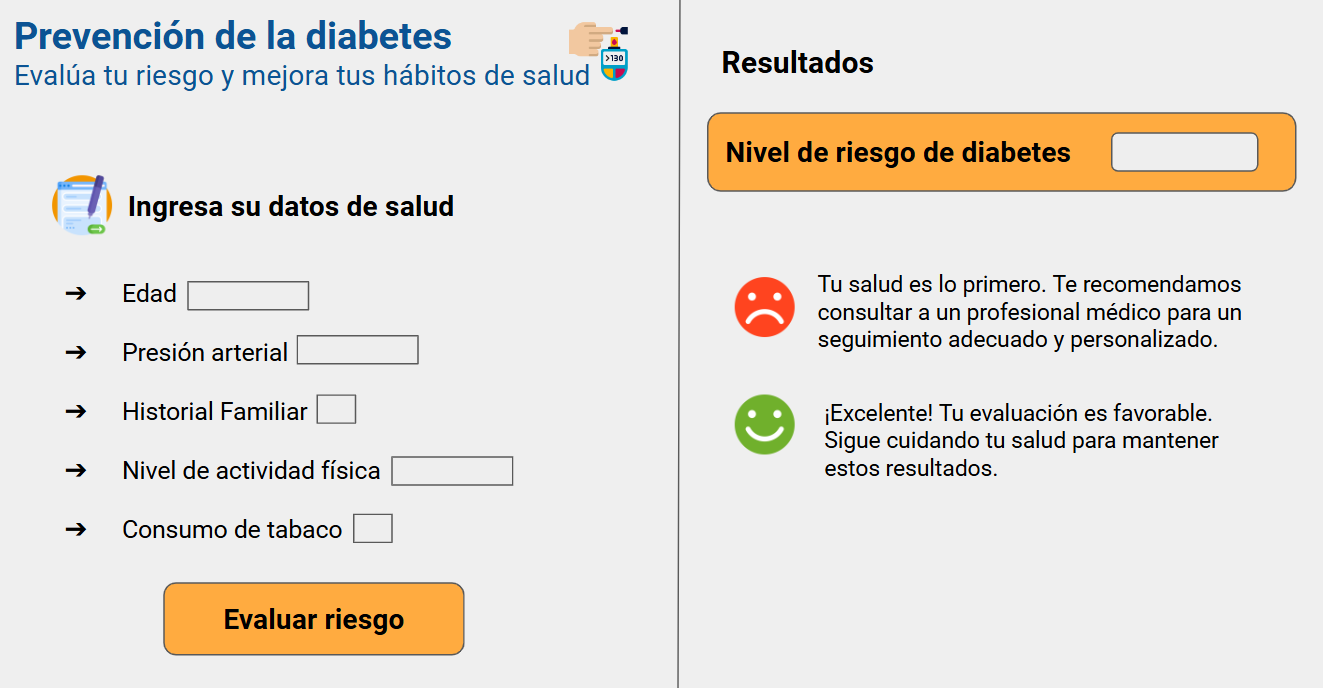
**Nuevo income**

****



**Maqueta del prototipo**

Esta herramienta busca ayudar a los usuarios a entender mejor su estado de salud y a motivarlos a adoptar hábitos saludables o a consultar a un profesional médico si es necesario. Para obtener un resultado, la persona deberá dejar sus datos en el formulario.



## Reporte de trabajo en equipo

Se realizó la configuración inicial del repositorio y de la implementación de DVC para el control de versiones de datos. Realizó el *first commit* y los commits "Track dataset files with DVC", asegurando una estructura de datos organizada y documentando los pasos para la integración de DVC, lo que facilita la colaboración y el manejo eficiente de datos en el proyecto. Además se hizo una exploración de los datos y la maqueta del prototipo.

