Documento técnico: Justificación del uso de variables NHANES para modelos de riesgo de diabetes.

1. Datos demográficos (NHANES – DEMO)

Variables:

Participant_ID: Identificador único del participante en el estudio NHANES.

Gender: Sexo del participante (1 = Hombre, 2 = Mujer).

Age_Years: Edad del participante en años.

Ethnicity: Categoría étnica del participante según clasificación NHANES.

Justificación técnica:

Las variables demográficas son la capa mínima necesaria para cualquier modelo de salud poblacional porque la prevalencia de diabetes tipo 2 no es homogénea entre grupos. La edad es uno de los predictores más fuertes: el riesgo aumenta a partir de la mediana edad por cambios en composición corporal, resistencia a la insulina y comorbilidades. El sexo puede capturar diferencias hormonales, patrones de grasa y conductas de salud. La etnia/grupo racial es importante porque NHANES está diseñado justamente para reflejar diferencias de riesgo entre subpoblaciones en EE.UU.; hay grupos con mayor prevalencia de obesidad, síndrome metabólico y diabetes, por lo que incluir esta variable mejora la calibración del modelo y evita sesgos sistemáticos.

El Participant_ID no se usa como predictor, pero es necesario para trazabilidad, auditoría de registros, unión con otros módulos (BMX, DIQ, MCQ, PAQ) y detección de duplicados.

Resumen DEMO (por qué se usa):

Este módulo entrega el contexto poblacional (edad, sexo, etnia) que condiciona el riesgo basal de diabetes. Sin DEMO el modelo sería menos generalizable porque no podría ajustar las diferencias de riesgo entre subgrupos.

2. Medidas corporales (NHANES – BMX)

Variables:

Weight_kg: Peso corporal del participante en kilogramos.

Height_cm: Altura del participante en centímetros.

BMI: Índice de masa corporal calculado como peso / (altura en metros)^2.

Justificación técnica:

El IMC es un indicador estándar y barato de exceso de peso. La evidencia epidemiológica muestra que el sobrepeso y, sobre todo, la obesidad son factores de riesgo principales para diabetes tipo 2 porque:

El exceso de tejido adiposo —en especial el adiposo visceral— genera resistencia a la insulina.

La resistencia crónica a la insulina lleva a hiperglucemia sostenida y a una mayor demanda pancreática, lo que acorta la “reserva” funcional beta-pancreática.

El IMC, aun siendo una medida gruesa, se correlaciona con otros marcadores cardiometabólicos (triglicéridos altos, HDL bajo, hipertensión), por lo que funciona bien como variable proxy cuando no hay laboratorio completo.

Weight_kg y Height_cm se mantienen por separado por dos razones:

Para recalcular IMC o generar indicadores alternativos si se detectan valores extremos.

Porque combinaciones “discordantes” (peso alto + estatura baja) a veces son más representativas de un fenotipo de riesgo que el IMC promedio.

Además, hay literatura que indica que la talla puede asociarse indirectamente al riesgo cardiometabólico (personas de menor estatura pueden tener distinta distribución de grasa o antecedentes de desarrollo) y en modelos ML esto puede aortar señal débil pero útil cuando se combina con edad y sexo.

Resumen BMX (por qué se usa):

Este módulo aporta la variable más directa de exceso de peso (IMC), que es uno de los determinantes más fuertes y modificables del riesgo de diabetes. También permite control de calidad (detectar outliers antropométricos).

3. Diabetes y percepción del riesgo (NHANES – DIQ)

Variables:

Diabetes Diagnosis: Diagnóstico médico informado de diabetes (1 = Sí, 2 = No, etc.).

Prediabetes_Diagnosis: Diagnóstico médico informado de prediabetes o niveles altos de glucosa (1 = Sí, 2 = No).

Perceived_Diabetes_Risk: Autoevaluación del riesgo percibido de desarrollar diabetes.

Justificación técnica:

Este módulo entrega información “cercana a la etiqueta”. Si el objetivo es construir un modelo de riesgo de diabetes no diagnosticada o detectar prediabetes, conocer el estado declarado permite:

Validar la consistencia del dataset (si la persona dice que tiene diabetes, sus medidas antropométricas y demográficas debieran ser coherentes con un fenotipo de riesgo).

Entrenar modelos de clasificación supervisada usando la presencia/ausencia de diagnóstico como target (o como una de las fuentes de la etiqueta).

Diferenciar entre diabetes ya instalada y estados intermedios (prediabetes), muy relevantes en tamizaje.

La variable de riesgo percibido agrega una dimensión conductual: personas que se perciben en riesgo suelen tener ya factores de riesgo presentes (obesidad, antecedentes familiares) o han recibido indicaciones médicas. En ML esto puede funcionar como variable de alta información cuando se cruza con DEMO y BMX.

Resumen DIQ (por qué se usa):

Este módulo entrega la información más directa sobre el estado de diabetes/prediabetes y permite definir o reforzar la etiqueta del modelo. También agrega una vista subjetiva (riesgo percibido) que puede mejorar la capacidad del modelo para identificar personas ya alertadas por el sistema de salud.

4. Condiciones médicas (NHANES – MCQ)

Variables:

Overweight_Diagnosis: Indica si el participante ha sido diagnosticado como con sobrepeso (1 = Sí, 2 = No).

Congestive_Heart_Failure: Presencia de insuficiencia cardíaca congestiva diagnosticada (1 = Sí, 2 = No).

Coronary_Artery_Disease: Presencia de enfermedad coronaria diagnosticada (1 = Sí, 2 = No).

Thyroid_Problem: Diagnóstico de enfermedad o trastorno de la tiroides (1 = Sí, 2 = No).

Jaundice_Diagnosis: Diagnóstico de ictericia (1 = Sí, 2 = No).

Family_History_Diabetes: Indica si algún familiar directo ha tenido diabetes (1 = Sí, 2 = No).

Justificación técnica:

Este bloque introduce comorbilidades y antecedentes, que son factores de contexto para la diabetes:

El antecedente familiar de diabetes es uno de los predictores no modificables más relevantes porque resume predisposición genética y también factores ambientales compartidos (dieta, estilo de vida).

El diagnóstico médico de sobrepeso confirma desde el punto de vista clínico lo que en BMX se ve como IMC alto, y puede servir para resolver inconsistencias (IMC en límite pero médico ya lo cataloga como sobrepeso).

Enfermedades cardiovasculares (insuficiencia cardíaca, enfermedad coronaria) suelen coexistir con síndrome metabólico. Si el paciente ya tiene afectación cardiovascular, es más probable que tenga o desarrolle desórdenes de glucosa.

Problemas de tiroides pueden alterar el metabolismo basal y el peso, lo que puede confundir al modelo si no se incluye.

Algunas variables como ictericia pueden servir más para control de calidad (descartar casos o señalar condiciones hepáticas que alteran laboratorio).

Incluir MCQ permite que el modelo no sólo vea “estado actual”, sino también carga de enfermedad y riesgo heredado, lo que suele mejorar AUC y calibración cuando el objetivo es detección temprana.

Resumen MCQ (por qué se usa):

Este módulo incorpora antecedentes familiares y comorbilidades que modulan el riesgo real de diabetes más allá del IMC. Es clave para distinguir a dos personas con el mismo peso pero con distinto riesgo biológico.

5. Actividad física (NHANES – PAQ)

Variable:

Total_MET_Score: Puntaje total de actividad física expresado en MET-minutos por semana.

Justificación técnica:

La actividad física es un factor protector frente a la diabetes tipo 2 porque mejora la sensibilidad a la insulina y ayuda al control del peso. Disponer de un puntaje cuantitativo (METs/semana) permite:

Incorporar al modelo una variable de estilo de vida.

Explicar casos de IMC moderadamente alto pero con bajo riesgo (personas activas).

Ajustar por sedentarismo, que suele agravar el efecto del exceso de grasa corporal.

En modelos ML, PAQ suele tener menor importancia que edad o IMC, pero aumenta la capacidad explicativa y puede ser decisiva en modelos de recomendación o estratificación de riesgo (a quién intervenir primero).

Resumen PAQ (por qué se usa):

Este módulo introduce el componente de estilo de vida. Permite que el modelo no sólo “castigue” por peso/edad, sino que también considere la actividad física como factor que puede reducir el riesgo.

Conclusión general

Tomados en conjunto, los módulos DEMO + BMX + DIQ + MCQ + PAQ entregan una vista multidimensional del riesgo de diabetes:

Quién es la persona (DEMO).

Cuál es su estado antropométrico (BMX).

Cuál es su estado/auto-reporte respecto a diabetes (DIQ).

Qué carga de enfermedad y antecedentes tiene (MCQ).

Qué tan activo es (PAQ).

Esta combinación es adecuada para un modelo de ML porque mezcla variables demográficas (lentas), clínicas (más directas), de antecedentes (moduladoras) y de comportamiento (modificables), lo que ayuda tanto a predecir como a interpretar el riesgo.