BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA. CU2.
Facultad de ciencias de la computación

Predicción del riesgo de diabetes a partir de indicadores de salud y estilo de vida

ASIGNATURA:
Introducción a la ciencia de datos

DOCENTE:
JAIME ALEJANDRO ROMERO SIERRA

ALUMNO:
LUNA DAMIAN ANGEL GABRIEL

FECHA DE ENTREGA:
23/11/2025

Introducción

El objetivo principal de este proyecto es desarrollar un modelo de análisis y predicción del riesgo de diabetes usando indicadores de salud, hábitos y factores sociales, aplicando técnicas de ciencia de datos y aprendizaje automático. La idea es crear una herramienta confiable y fácil de usar en entornos clínicos y de salud pública que ayude a detectar temprano a personas con alto riesgo de desarrollar diabetes.

La diabetes, especialmente el tipo 2, ha crecido mucho en el mundo por cambios en los estilos de vida, aumento del sobrepeso y obesidad, además de factores genéticos y sociales. Pero la mayoría de los casos se pueden prevenir o retrasar si se detectan a tiempo los factores de riesgo. Un modelo predictivo con datos grandes tiene el potencial de convertirse en una herramienta importante para prevenir la enfermedad, tanto en la población general como en grupos específicos en riesgo.

Por eso, es importante buscar nuevas formas de detectar y predecir la diabetes, usando datos y técnicas de análisis avanzado. La disponibilidad de grandes bases de datos con información sobre la salud, hábitos y características de las personas crea una oportunidad única para encontrar patrones de riesgo y detectar la enfermedad antes de que sea irreversible.

Descripción general del conjunto de datos

El Diabetes Health Indicators Dataset es una base de datos desarrollado a partir de los registros de salud recopilados por el Centers for Disease Control and Prevention (CDC) en Estados Unidos.
Su principal objetivo es analizar y comprender los factores asociados con la aparición de la diabetes y la prediabetes en la población adulta.

Este dataset reúne información de miles de participantes, obtenida mediante encuestas nacionales que recogen tanto aspectos clínicos como de estilo de vida.
El conjunto contiene más de100 000 observaciones, donde cada fila representa a una persona y cada
columna a una característica relacionada con su salud o su entorno sociodemográfico.
Entre las variables más relevantes se incluyen indicadores como el índice de masa corporal (IMC), la presencia de presión arterial alta, colesterol elevado, nivel de actividad física, consumo de tabaco y alcohol, así como
condiciones médicas previas, como enfermedades cardíacas.

También incorpora variables demográficas importantes, como la edad, el sexo, el nivel educativo, los ingresos económicos y la raza o grupo étnico.

El valor central del dataset es una variable binaria o categórica que indica si la persona tiene diabetes, se encuentra en estado de prediabetes o no padece la enfermedad, lo que permite su uso en tareas de
clasificación y predicción médica.

El Diabetes Health Indicators Dataset ofrece una visión integral del estado de salud de una población amplia y diversa, constituyéndose como una herramienta útil para el análisis estadístico, la educación en ciencia de datos y el desarrollo de estrategias preventivas orientadas a mejorar la calidad de vida y
reducir la incidencia de la diabetes.

Metodología

Proceso de limpieza de datos

Primero se importó la biblioteca pandas y se cargó el archivo CSV original,
se examinaron las primeras filas, las clases de datos, los nombres de las columnas

y los valores nulos utilizando df.head(), df.info(), df.columns y df.isnull().sum(). respectivamente, esto facilitó la obtención de un diagnóstico global del estado inicial de la base de datos.

Se hallaron registros duplicados usando df.duplicated().sum() y ya que la cantidad de datos duplicados no representaba una cantidad notable, se eliminaron mediante df.drop_duplicates(). Esto aseguraba que cada fila del conjunto de datos representara un registro individual.

Se examinaron los valores únicos de cada columna utilizando un ciclo for, lo que facilitó identificar categorías, rangos y posibles errores de entrada.

Se encontraron valores erróneos como "Auto%#" en columnas numéricas. Se contabilizó cuántos valores incorrectos había por columna.

Se reemplazaron los valores erróneos por None para poder tratarlos como valores nulos.

Se utilizó el método backfill (bfill) para completar los valores ausentes. Si el último valor de la columna estaba vacío, continuaba sin completar.

Se cambió el nombre de las columnas del inglés al español utilizando un diccionario de traducción con df.rename().

Se encontraron columnas que debían ser números enteros o de punto flotante, y se cambiaron sus tipos usando astype(int) y astype(float).

Se revisó nuevamente el DataFrame utilizando df.info() para verificar la consistencia en los tipos de datos y la estructura final.

Por último, se exportó el DataFrame depurado como Base_limpia.csv sin incluir índice.

La limpieza del proceso permitió suprimir duplicados, rectificar valores erróneos, l lenar valores ausentes, estandarizar formatos de datos y traducir títulos de columnas. El conjunto de datos generado quedó preparado para su análisis y modelaje.

Análisis exploratorio de datos (EDA)

EL dataset contiene 100,401 registros y 31 columnas.

Tipos de datos que se manejan:

edad: numérico, edad del paciente en años
género: categórico, sexo del paciente
etnia: categórico, grupo étnico del paciente
nivel_educativo: categórico, máximo nivel de estudios alcanzado
nivel_ingresos: categórico, rango de ingresos económicos
estado_laboral: categórico, situación laboral del paciente
hábito_fumar: categórico, estado de consumo de tabaco
consumo_alcohol_semana: numérico, cantidad promedio semanal de alcohol
actividad_física_minutos_semana: numérico, minutos de actividad física por semana
puntaje_dieta: numérico, indicador de calidad de dieta
horas_sueño_día: numérico, horas que duerme por día
horas_pantalla_día: numérico, horas al día frente a pantallas
antecedente_familiar_diabetes: categórico, si tiene familiares con diabetes
antecedente_hipertensión: categórico, historial de hipertensión
antecedente_cardiovascular: categórico, historial de enfermedades cardiovasculares
imc: numérico, índice de masa corporal
relación_cintura_cadera: numérico, indicador de obesidad abdominal
presión_sistólica: numérico, presión arterial sistólica
presión_diastólica: numérico, presión arterial diastólica
frecuencia_cardiaca: numérico, latidos por minuto
colesterol_total: numérico, nivel total de colesterol
colesterol_hdl: numérico, nivel de colesterol HDL
colesterol_ldl: numérico, nivel de colesterol LDL
triglicéridos: numérico, nivel de triglicéridos
glucosa_ayuno: numérico, nivel de glucosa en ayunas
glucosa_postprandial: numérico, nivel de glucosa posterior a alimentos
nivel_insulina: numérico, nivel de insulina en sangre
hba1c: numérico, control de glucosa a largo plazo
puntaje_riesgo_diabetes: numérico, estimación del riesgo de desarrollar diabetes
etapa_diabetes: categórico, clasificación del estado del paciente
diabetes_diagnosticada: categórico, indica si tiene diagnóstico formal de diabetes

Resumen estadístico

Variables numéricas:

count 100401.000000
mean 50.118684
std 15.616850
min 18.000000
25% 39.000000
50% 50.000000
75% 61.000000
max 90.000000
Name: edad, dtype: float64
count 100401.000000

mean 2.004721
std 1.418460
min 0.000000
25% 1.000000
50% 2.000000
75% 3.000000
max 10.000000
Name: consumo_alcohol_semana, dtype: float64
count 100401.000000
mean 118.939473
std 84.456046
min 0.000000
25% 57.000000
50% 100.000000
75% 160.000000
max 833.000000
Name: actividad_física_minutos_semana, dtype: float64
count 100401.000000
mean 5.996742
std 1.781202
min 0.000000
25% 4.800000
50% 6.000000
75% 7.200000
max 10.000000
Name: puntaje_dieta, dtype: float64
count 100401.000000
mean 6.997764
std 1.094915
min 3.000000
25% 6.300000
50% 7.000000
75% 7.700000
max 10.000000
Name: horas_sueño_por_día, dtype: float64

count 100401.000000
mean 5.994320
std 2.464575
min 0.500000
25% 4.300000
50% 6.000000
75% 7.700000
max 16.800000
Name: horas_pantalla_día, dtype: float64
count 100401.000000
mean 25.611917
std 3.588629
min 15.000000
25% 23.200000
50% 25.600000
75% 28.000000
max 39.200000
Name: imc, dtype: float64
count 100401.000000
mean 0.856070
std 0.046851
min 0.670000
25% 0.820000
50% 0.860000
75% 0.890000
max 1.060000
Name: relación_cintura_cadera, dtype: float64
count 100401.000000
mean 115.789554
std 14.280768
min 90.000000
25% 106.000000
50% 116.000000
75% 125.000000
max 179.000000

Name: presión_sistólica, dtype: float64
count 100401.000000
mean 75.223763
std 8.209381
min 50.000000
25% 70.000000
50% 75.000000
75% 81.000000
max 110.000000
Name: presión_diastólica, dtype: float64
count 100401.000000
mean 69.633948
std 8.372390
min 40.000000
25% 64.000000
50% 70.000000
75% 75.000000
max 105.000000
Name: frecuencia_cardiaca, dtype: float64
count 100401.000000
mean 185.989054
std 31.988704
min 100.000000
25% 164.000000
50% 186.000000
75% 208.000000
max 318.000000
Name: colesterol_total, dtype: float64
count 100401.000000
mean 54.035129
std 10.264825
min 20.000000
25% 47.000000
50% 54.000000
75% 61.000000

max 98.000000
Name: colesterol_hdl, dtype: float64
count 100401.000000
mean 103.018416
std 33.446243
min 50.000000
25% 78.000000
50% 102.000000
75% 126.000000
max 263.000000
Name: colesterol_ldl, dtype: float64
count 100401.000000
mean 121.482266
std 43.370337
min 30.000000
25% 91.000000
50% 121.000000
75% 151.000000
max 344.000000
Name: triglicéridos, dtype: float64
count 100401.000000
mean 111.125995
std 13.601763
min 60.000000
25% 102.000000
50% 111.000000
75% 120.000000
max 172.000000
Name: glucosa_ayuno, dtype: float64
count 100401.000000
mean 160.024054
std 30.925267
min 70.000000
25% 139.000000
50% 160.000000

75% 181.000000
max 287.000000
Name: glucosa_postprandial, dtype: float64
count 100401.000000
mean 9.066071
std 4.951832
min 2.000000
25% 5.110000
50% 8.790000
75% 12.450000
max 32.220000
Name: nivel_insulina, dtype: float64
count 100401.000000
mean 6.521798
std 0.813378
min 4.000000
25% 5.970000
50% 6.520000
75% 7.070000
max 9.800000
Name: hba1c, dtype: float64
count 100401.000000
mean 30.211309
std 9.060686
min 2.700000
25% 23.800000
50% 29.000000
75% 35.600000
max 67.200000
Name: puntaje_riesgo_diabetes, dtype: float64

Variables categóricas:

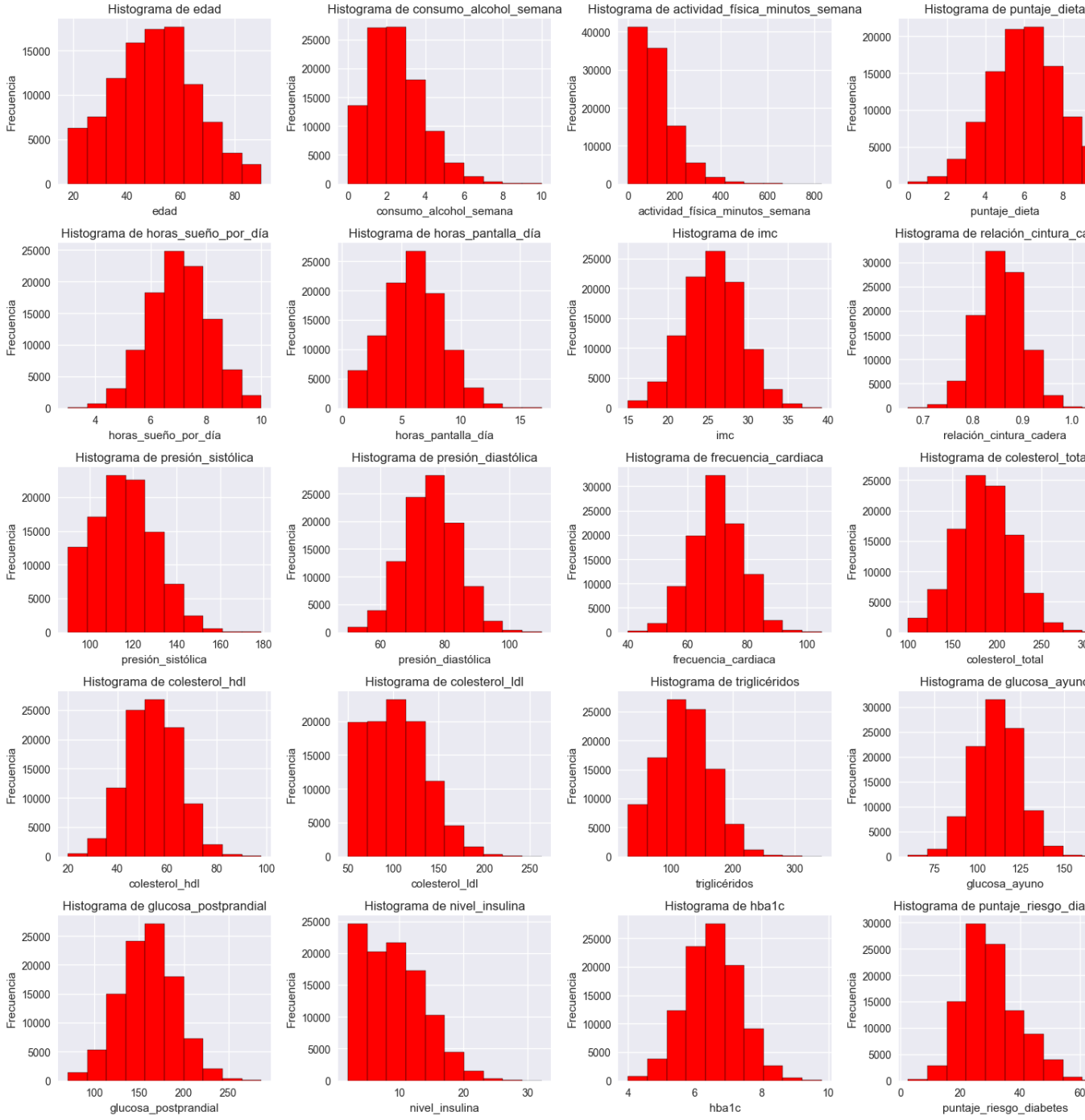
género
Mujer 50425

Hombre 47965
Otro 2011
Name: count, dtype: int64
etnia
Blanco 45176
Hispano 20168
Negro 18057
Asiático 11923
Otro 5077
Name: count, dtype: int64
nivel_educativo
Preparatoria 45113
Licenciatura 35123
Posgrado 15028
Sin estudios formales 5137
Name: count, dtype: int64
nivel_ingresos
Medio 35316
Medio-bajo 25326
Medio-alto 19888
Bajo 14857
Alto 5014
Name: count, dtype: int64
estado_laboral
Empleado 60354
Jubilado 21892
Desempleado 11943
Estudiante 6212
Name: count, dtype: int64
hábito_fumar
Nunca 60044
Fumador actual 20225
Exfumador 20132
Name: count, dtype: int64
antecedente_familiar_diabetes

0 78372
1 22029
Name: count, dtype: int64
antecedente_hipertensión
0 75265
1 25136
Name: count, dtype: int64
antecedente_cardiovascular
0 92434
1 7967
Name: count, dtype: int64
etapa_diabetes
Diabetes tipo 2 60027
Prediabetes 31988
Sin diabetes 7990
Diabetes gestacional 276
Diabetes tipo 1 120
Name: count, dtype: int64
diabetes_diagnosticada
1 60326
0 40075
Name: count, dtype: int64

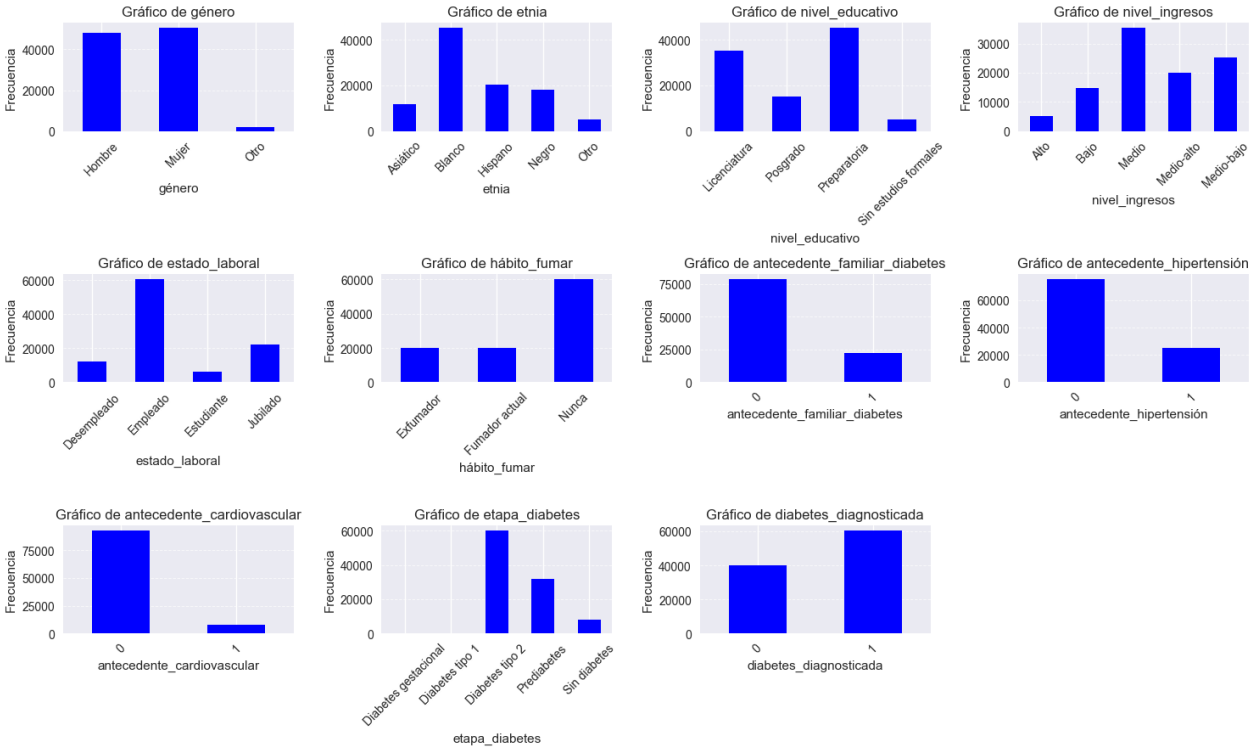
Visualización y distribución de variables individuales

Se usaron histogramas para variables numéricas:



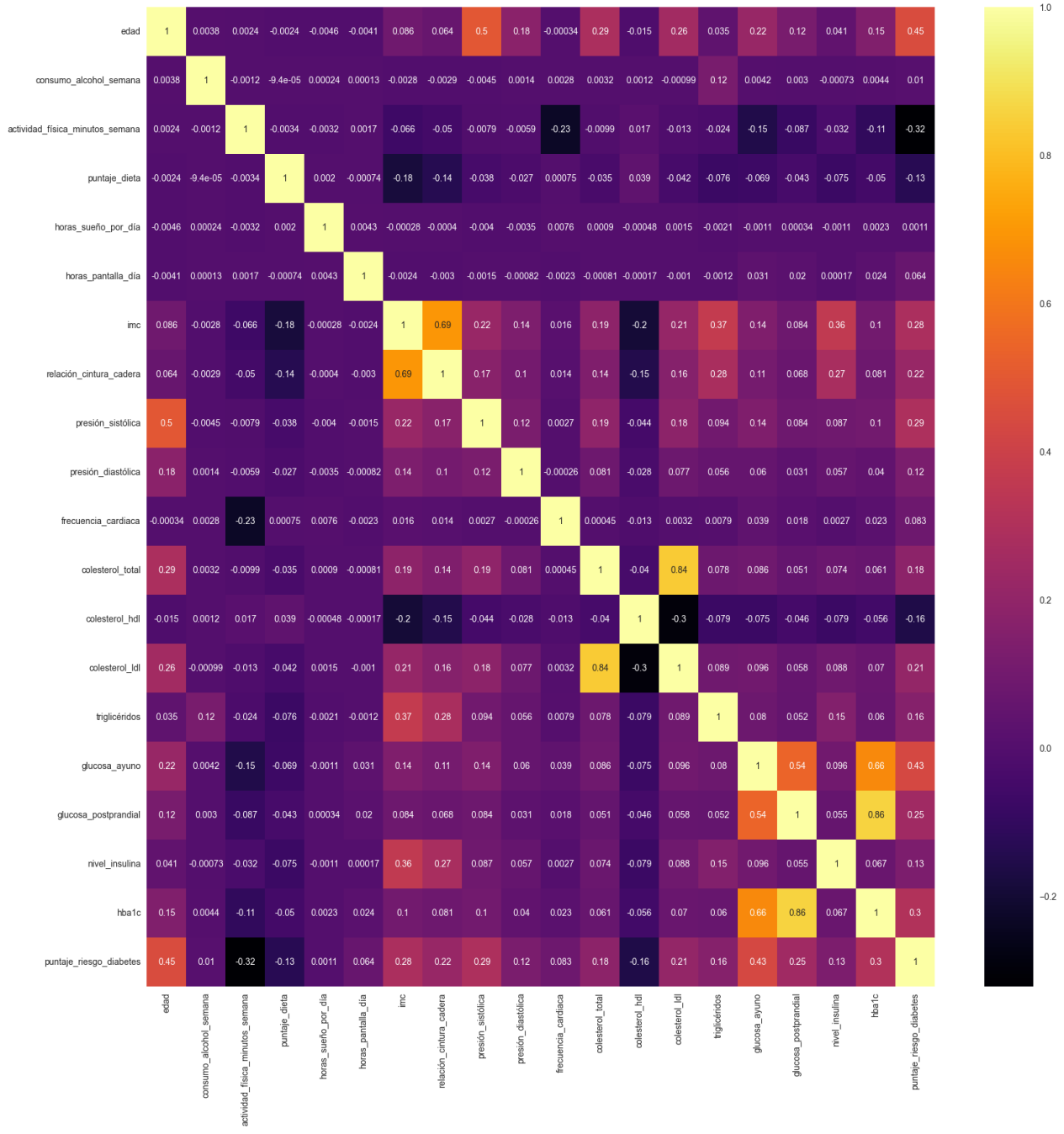
Los histogramas muestran que la mayoría de las variables del conjunto de datos tienen distribuciones aproximadamente normales, lo que significa un comportamiento poblacional estable, mientras que otras presentan una marcada asimetría hacia la derecha, especialmente aquellas relacionadas con estilo de vida y metabolismo, estas asimetrías sugieren la presencia de pocos valores muy altos que pueden influir en los análisis estadísticos por lo que mas adelante usaré una transformación logarítmica para disminuir el porcentaje de errores de las predicciones.

Después usamos gráficos de barras para visualizar la distribución de los datos categóricos:



Las gráficas de variables categóricas muestran que la población del estudio está equilibrada en género, con predominio de las etnias Blanco e Hispano, y presenta un nivel educativo mayoritariamente de licenciatura, la distribución del nivel de ingresos se concentra en los estratos medio, medio-bajo y medio-alto, mientras que la mayoría de los participantes se encuentra empleada, en hábitos, la mayor parte declara no haber fumado nunca, lo cual es favorable desde una perspectiva de salud pública, la mayoría no reporta antecedentes familiares de diabetes, hipertensión o enfermedades cardiovasculares, aunque existe un grupo importante que sí los presenta, en cuanto a la presencia de diabetes, la categoría con mayor frecuencia es diabetes tipo 2, seguida por prediabetes, lo que indica una carga significativa de esta enfermedad en la población

Correlación entre variables



Utilizamos un mapa de calor para determinar que variables tienen mayor relación entre sí, del mapa concluimos lo siguiente:

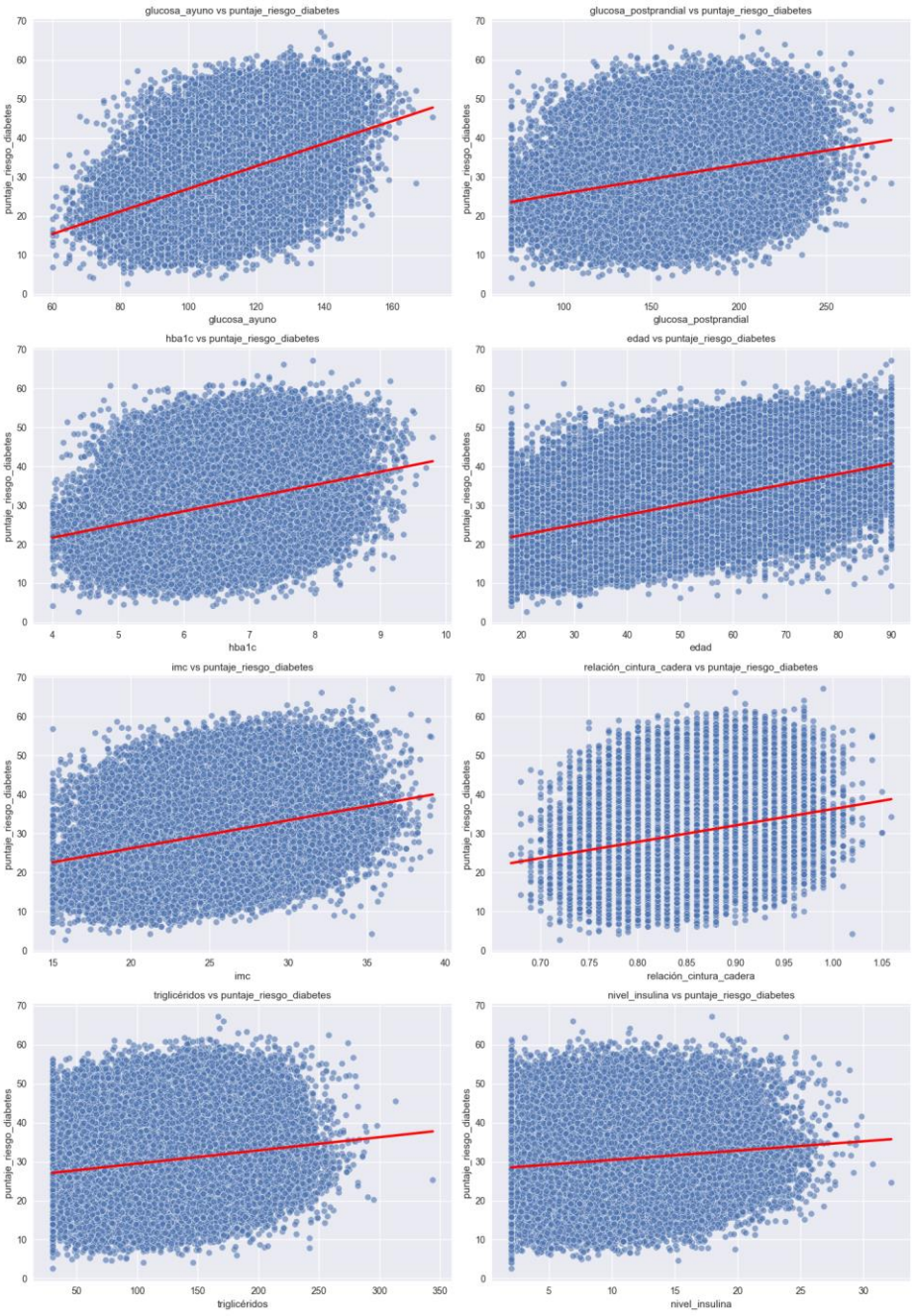
-Glucosa en ayuno y glucosa posprandial tienen la correlación más alta con el puntaje de riesgo, lo que indica que los niveles elevados de glucosa son el principal indicador de propensión a

diabetes

-HbA1c también muestra una correlación fuerte con ambas medidas de glucosa y con el riesgo general, confirmando que el control glucémico a largo plazo es un predictor clave

-Edad presenta una correlación moderada con el riesgo, lo que puede significar que la probabilidad de desarrollar diabetes aumenta conforme el paciente envejece

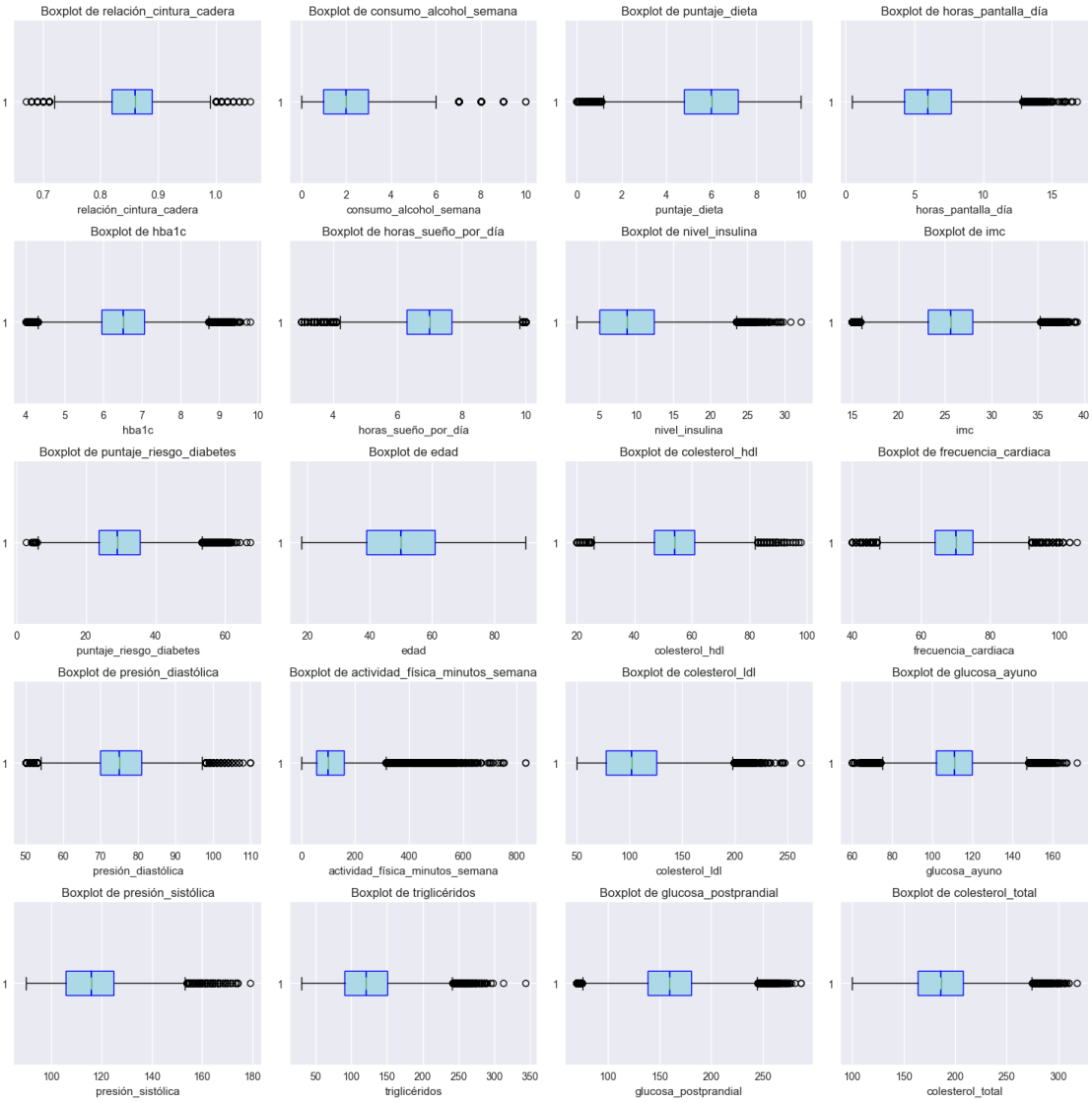
Basándonos en eso, visualizamos su relación mediante scatter plots:



Las gráficas muestran que las variables con mayor influencia en el puntaje de riesgo de diabetes son las relacionadas con metabolismo glucémico: glucosa en ayuno, glucosa posprandial y HbA1c, seguidas en menor medida por edad, IMC, relación cintura–cadera, y por último triglicéridos e insulina.

Análisis de valores atípicos

Para detectar valores atípicos dentro de nuestras variables utilizamos boxplots, ya que facilita tanto su detección como visualización.



Como se puede observar, gran parte de las variables numéricas contienen gran cantidad de valores atípicos, lo que mas adelante será perjudicial para el modelo de machine learning, por lo que decidí eliminarlos mediante una winsorisación al percentil 1 y 99, lo que cambia los valores atípicos por unos mas cercanos a la media, esto ya que nuestro objetivo es mantener la forma del dataframe conservando la mayor cantidad de registros posibles.

Esta es una lista de la cantidad de outliers corregidos en cada variable:

Variable: edad
  Límite inferior: 6.00
  Límite superior: 94.00
  Outliers corregidos: 0

Variable: consumo_alcohol_semana
  Límite inferior: -2.00
  Límite superior: 6.00
  Outliers corregidos: 460

Variable: actividad_física_minutos_semana
  Límite inferior: -97.50
  Límite superior: 314.50
  Outliers corregidos: 3227

Variable: puntaje_dieta
  Límite inferior: 1.20
  Límite superior: 10.80
  Outliers corregidos: 342

Variable: horas_sueño_por_día
  Límite inferior: 4.20
  Límite superior: 9.80
  Outliers corregidos: 901

Variable: horas_pantalla_día
  Límite inferior: -0.80
  Límite superior: 12.80
  Outliers corregidos: 299

Variable: imc
  Límite inferior: 16.00
  Límite superior: 35.20
  Outliers corregidos: 758

Variable: relación_cintura_cadera
  Límite inferior: 0.71
  Límite superior: 1.00
  Outliers corregidos: 280

Variable: presión_sistólica
  Límite inferior: 77.50
  Límite superior: 153.50
  Outliers corregidos: 537

Variable: presión_diastólica
  Límite inferior: 53.50
  Límite superior: 97.50
  Outliers corregidos: 744

Variable: frecuencia_cardiaca
  Límite inferior: 47.50
  Límite superior: 91.50
  Outliers corregidos: 861

Variable: colesterol_total
  Límite inferior: 98.00
  Límite superior: 274.00
  Outliers corregidos: 307

Variable: colesterol_hdl
  Límite inferior: 26.00
  Límite superior: 82.00
  Outliers corregidos: 566

Variable: colesterol_ldl
  Límite inferior: 6.00
  Límite superior: 198.00
  Outliers corregidos: 357

Variable: triglicéridos
  Límite inferior: 1.00
  Límite superior: 241.00
  Outliers corregidos: 298

Variable: glucosa_ayuno
  Límite inferior: 75.00
  Límite superior: 147.00
  Outliers corregidos: 737

Variable: glucosa_postprandial
  Límite inferior: 76.00
  Límite superior: 244.00
  Outliers corregidos: 638

Variable: nivel_insulina
  Límite inferior: -5.90
  Límite superior: 23.46
  Outliers corregidos: 329

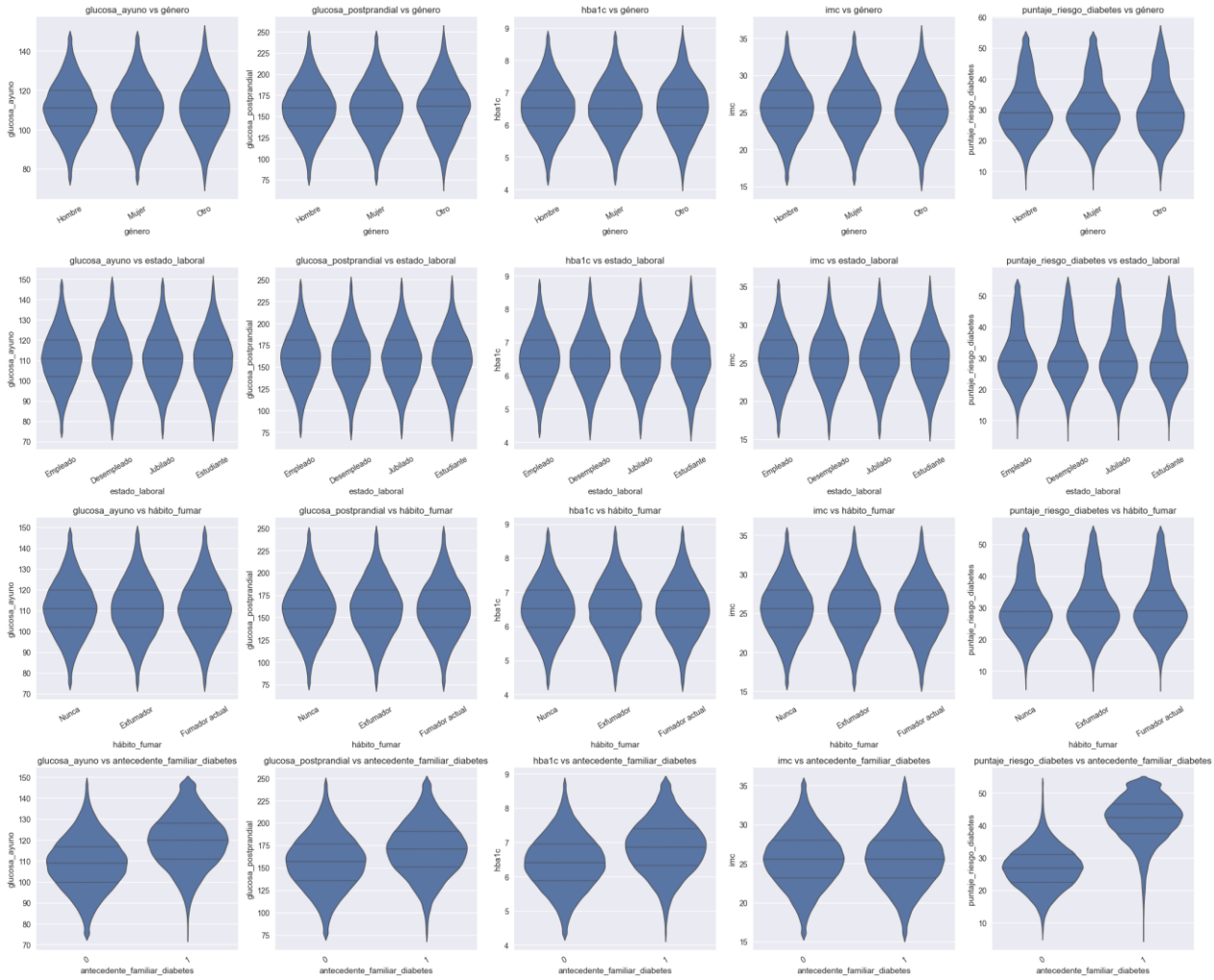
Variable: hba1c
  Límite inferior: 4.32
  Límite superior: 8.72
  Outliers corregidos: 616

Variable: puntaje_riesgo_diabetes
  Límite inferior: 6.10
  Límite superior: 53.30
  Outliers corregidos: 918

Después se hizo un análisis de los valores faltantes, sin embargo dicho análisis dio como resultado 0 en cada columna, por lo que la base no necesita otro tratamiento de limpieza.

Relación entre Variables Categóricas y Numéricas

Para esta parte utilicé violín plots para visualizar la relación entre las variables categóricas mas relacionadas con el puntaje de riesgo de diabetes:



Los violin plots permitieron evaluar cómo se distribuyen variables numéricas relacionadas con el riesgo de diabetes según distintas variables categóricas. En general, género y hábito de fumar mostraron distribuciones muy similares entre categorías, indicando que en este dataset no presentan un impacto notable sobre glucosa, HbA1c, IMC o puntaje de riesgo.

El estado laboral mostró diferencias leves, principalmente en la dispersión de los valores, pero sin cambios significativos en las medianas. La variable con mayor influencia observable fue el antecedente familiar de diabetes, ya que en este grupo se identificaron valores ligeramente más altos en glucosa posprandial, HbA1c y puntaje de riesgo.

Observaciones y hallazgos importantes

1. Perfil de la Población: Alto Riesgo Metabólico
El hallazgo más contundente es que esta base de datos no representa a una población general sana, sino a una población altamente patológica o de alto riesgo:
el 91.65% de los registros en la muestra ya tiene una alteración metabólica significativa (59.8% Diabetes Tipo 2 y 31.9% Pre-Diabetes).
Tipos Raros: La Diabetes Tipo 1 y Gestacional representan una fracción marginal (< 0.5%), lo que dificulta su predicción sin técnicas de balanceo de datos.

2. Factores de Riesgo Predominantes:
El Antecedente Familiar de Diabetes es el factor predictivo aislado más importannte (aprox. 25% de importancia), si tus padres la tienen, tu riesgo se dispara.
La Edad: Es el segundo factor más importante; a mayor edad, mayor acumulación de riesgo metabólico.
Estilo de Vida: La Actividad Física (min/semana) y el IMC son los factores modificables más influyentes.

3. Estado Clínico Promedio
HbA1c Promedio: 6.52%.
IMC Promedio: 25.6, sitúa a la media de la población en la categoría de Sobrepeso.
Glucosa en Ayuno: 111.1 mg/dL, un valor típico de Pre-Diabetes.

Descripción y justificación del modelo de machine learning

Se seleccionó el modelo Random Forest Classifier para predecir el riesgo de diabetes basándonos en los siguientes criterios:

Tipo de variable objetivo: Dado que la variable objetivo (etapa_diabetes) es categórica, buscamos resolver un problema de clasificación multiclase (Sin Diabetes, Pre-Diabetes, Diabetes Tipo 2). Random Forest maneja este tipo de clasificación de manera nativa y eficiente, asignando a cada paciente la categoría más probable por votación mayoritaria de los árboles.

Tamaño y naturaleza del dataset: Con un volumen de 100,401 registros y una mezcla de datos numéricos y categóricos, Random Forest es ideal porque no requiere una normalización estricta de los datos y es robusto frente al sobreajuste (*overfitting*) gracias a su técnica de ensamble (promediar múltiples árboles de decisión).

Interpretabilidad y Precisión: En el ámbito clínico, no basta con predecir; es necesario entender el *porqué*. Elegimos este modelo porque ofrece un excelente equilibrio: logra una alta precisión al capturar relaciones no lineales complejas (que modelos más simples como la Regresión Logística podrían perder) y, simultáneamente, proporciona la "Importancia de las Características", permitiéndonos explicar qué factores médicos (como HbA1c o Edad) son determinantes para el diagnóstico.

Resultados del modelo

El modelo tiene una precisión del 88.26%. Esto significa que de cada 100 pacientes que el modelo evalúa, clasifica correctamente la etapa de diabetes de 88 de ellos, lo cual representa un excelente rendimiento para un entorno médico.

Análisis por Categoría:

Diabetes Tipo 2:

Tiene una Precisión del 97%. Esto significa que cuando el modelo detecta un paciente con diabetes tipo 2, es muy probable que sea verdad.
sin embargo su Recall es del 85%, esto significa que se le escapan el 15% de los

diabéticos reales y los clasifica erróneamente en otra categoría.

Prediabetes:

Tiene un Recall del 95%., el modelo es una excelente herramienta, ya que detecta a casi todos los prediabéticos.

Sin Diabetes:

Detectando correctamente al 92% de las personas sanas.

Que datos usó el modelo para decidir:

El modelo le dió prioridad a los análisis de sangre:
HbA1c (26%) y Glucosa en ayuno (15%) fueron los factores principales.
El factor demográfico más importante fue la Edad (9%).
La Presión Sistólica y la Actividad Física jugaron un papel secundario pero relevante.

Interpretación de la Matriz de Confusión (Los errores)
La matriz nos dice dónde se equivocó el modelo exactamente:

Hubo 1,548 personas con Diabetes Tipo 2 que el modelo clasificó incorrectamente como Prediabetes. Esto confirma que hay una "zona gris" clínica donde los valores de estos pacientes están justo en el límite.

El acierto: Clasificó correctamente a 10,167 diabéticos y 6,084 prediabéticos, lo cual valida la utilidad del modelo para la gran mayoría de la población.

El modelo Random Forest alcanzó una precisión global del 88%, demostrando un error bajo y una alta fiabilidad para distinguir entre pacientes sanos y aquellos con Diabetes Tipo 2 o Pre-Diabetes. Sin embargo, mostró limitaciones para identificar las clases minoritarias ("Otros") debido al desbalance de los datos.

El análisis de importancia reveló que los indicadores clínicos ‘HbA1c’ y ‘Glucosa en Ayuno’ fueron determinantes para la predicción, seguidos por factores de riesgo no modificables como ‘Edad’ y ‘Antecedente Familiar de Diabetes’.

Para perfeccionar el modelo, se podrían implementar técnicas de balanceo de datos para mejorar la detección de Diabetes Tipo 1 y Gestacional.

El modelo Random Forest logró una precisión del 88%, validando su uso como herramienta de cribado clínico. Las variables más relevantes fueron ‘HbA1c’, ‘Glucosa’ y ‘Edad’.

Beneficios de usar un dashboard

¿Qué decisiones puede apoyar el usuario gracias al dashboard?

Permite al personal médico identificar rápidamente a los pacientes que requieren atención inmediata o pruebas confirmatorias, basándose en su perfil de riesgo (edad + antecedentes) antes incluso de tener resultados de laboratorio completos.

También ayuda a los directivos a decidir dónde invertir el presupuesto de prevención. Al ver que el riesgo se dispara a partir de los 45 años, pueden dirigir campañas de tamizaje específicamente a ese grupo demográfico en lugar de a la población general.

Asimismo, al visualizar que el 60% de la población ya tiene Diabetes Tipo 2, las aseguradoras o clínicas pueden decidir asignar más recursos a programas de control y prevención secundaria (evitar complicaciones) en lugar de solo prevención primaria, ajustando su estrategia operativa a la realidad de la población.

¿Qué insights se pueden obtener con solo mirar las gráficas?

El gráfico de pastel revela instantáneamente que estamos ante una población patológica, desmintiendo la idea de que la mayoría de los usuarios están "sanos".

El diagrama de dispersión (scatter plot) de glucosa vs. HbA1c muestra visualmente una correlación lineal perfecta, lo que valida la calidad de los datos: a mayor glucosa en ayuno, mayor daño crónico, sin excepciones visibles.

El gráfico de importancia de variables destaca que el "Antecedente Familiar" es un factor de riesgo crítico que el modelo pondera fuertemente, alertando que la historia clínica familiar es tan vital como un análisis de sangre.

¿Cómo se simplifica la interpretación del modelo o los resultados?

En lugar de mostrar fórmulas matemáticas complejas, el gráfico de barras horizontales explica en lenguaje sencillo qué está pensando la IA, mostrando qué variables tienen más peso en la decisión.

El gráfico de barras comparativo traduce la compleja matriz de confusión a un formato visual simple. Permite entender de un vistazo que el modelo es "conservador", lo que significa que la herramienta está configurada para pecar de precavida y no dejar escapar pacientes en riesgo.

Hallazgos principales y áreas de mejora

Aquí tienes los hallazgos principales y las posibles mejoras para tu proyecto de predicción de diabetes:
Hallazgos Principales

Alta Carga Patológica: El análisis reveló que el 91.6% de la población estudiada ya presenta alteraciones metabólicas (Diabetes Tipo 2 o Pre-Diabetes), desmintiendo la idea de una población mayoritariamente sana.

Jerarquía de Factores de Riesgo: Aunque los marcadores de laboratorio (HbA1c y Glucosa) son las predictoras dominantes, el modelo identificó que factores no modificables como la Edad y el Antecedente Familiar son determinantes críticos para el riesgo.

Comportamiento del Modelo: El algoritmo Random Forest alcanzó una precisión del 88%, demostrando un comportamiento "conservador" y seguro: tiende a clasificar casos limítrofes como "Pre-Diabetes" para minimizar los falsos negativos y no dejar sin atención a pacientes en riesgo.

Posibles Mejoras

Balanceo de Datos: Aplicar técnicas de sobremuestreo sintético para corregir el desbalance de clases, permitiendo que el modelo aprenda a detectar tipos de diabetes menos frecuentes (Tipo 1 y Gestacional) que actualmente ignora.

Exploración de Algoritmos: Probar modelos como XGBoost o LightGBM, que suelen tener un mejor rendimiento reduciendo errores en la "zona gris" entre Pre-Diabetes y Diabetes Tipo 2.

Referencias bibliográficas

Ismail, L., Materwala, H., & Kaabi, J. A. (2021). Association of risk factors with type 2 diabetes: A systematic review. *Computational And Structural Biotechnology Journal*, *19*, 1759-1785. https://doi.org/10.1016/j.csbj.2021.03.003

Diabetes risk factors. (2024, 15 mayo). Diabetes. https://www.cdc.gov/diabetes/risk-factors/index.html?utm_source=perplexity

P, A. (2023). Risk Factors of diabetes. *www.openaccessjournals.com*. https://doi.org/10.37532/jdmc.2023.6(3).61-66

Link de la base de datos:

https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset