CODER HOUSE

Identificación de Factores de Riesgo Predictivos para la Diabetes

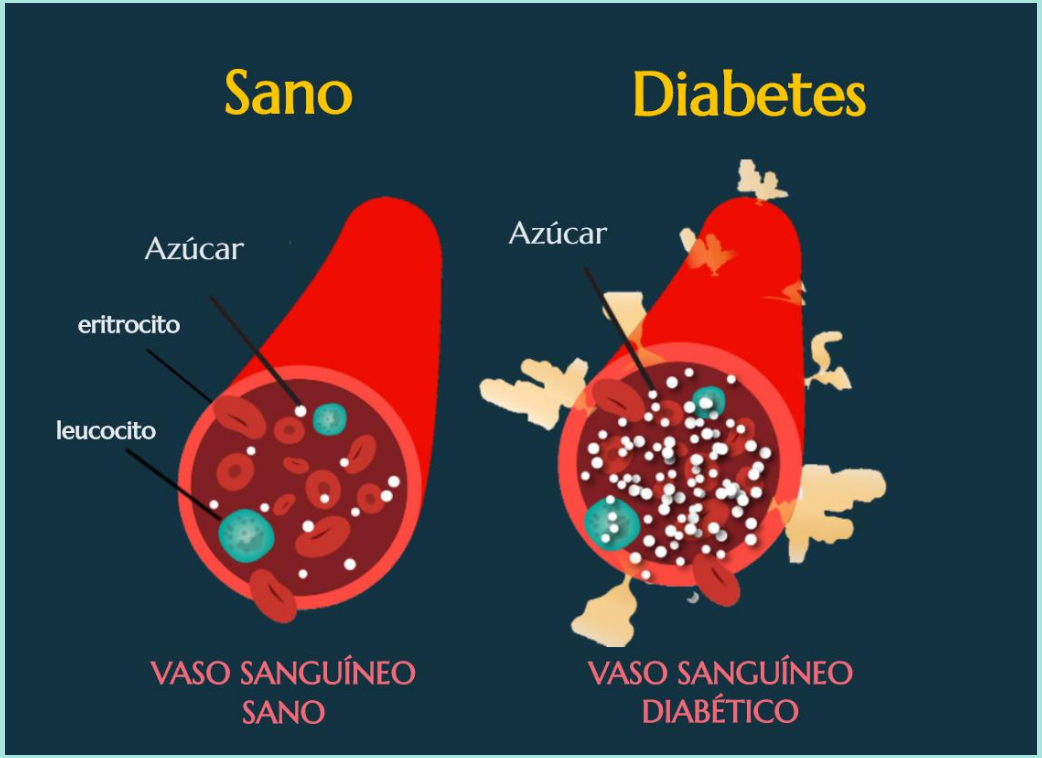
- Estudiante: Florencia de la Rosa
- Profesor: German Rodriguez
- Tutor: Ignacio Fernández
- Comisión: #60895
- Entrega Final: 21/08/2024

CODER HOUSE

Tabla de Contenido

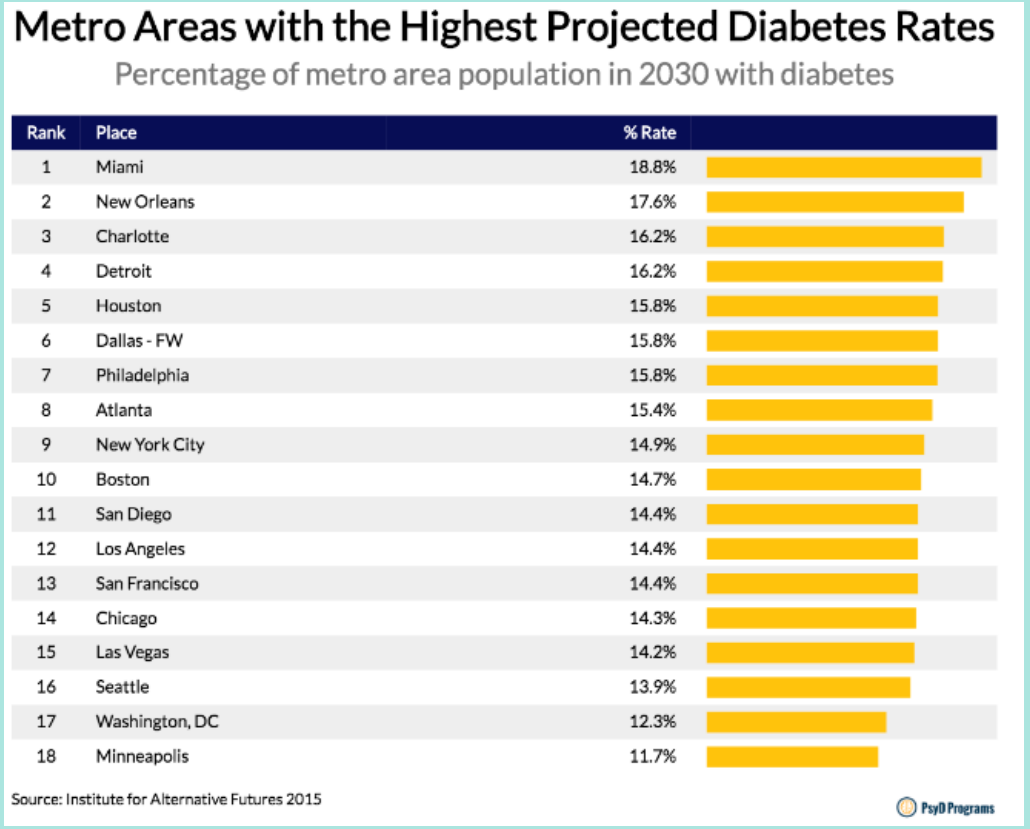
1. Contexto

La diabetes es una **enfermedad crónica grave** en la que los individuos **pierden la capacidad de regula**r efectivamente los **niveles de glucosa en la sangre**, lo que puede llevar a una **reducción de la calidad y expectativa de vida**.



1. Contexto

- La diabetes es una de las **enfermedades crónicas más prevalentes en los Estados Unidos**, afectando a millones de estadounidenses cada año y ejerciendo una carga financiera significativa en la economía.
- El **Institute for Alternative Futures** pronosticó en el 2015 que para el 2030 la tasa de diabéticos aumentará en un 38%; es decir, el 15,3% de la población la padecerá.



Metro Areas with the Highest Projected Diabetes Rates

Percentage of metro area population in 2030 with diabetes

Rank	Place	% Rate
1	Miami	18.8%
2	New Orleans	17.6%
3	Charlotte	16.2%
4	Detroit	16.2%
5	Houston	15.8%
6	Dallas - FW	15.8%
7	Philadelphia	15.8%
8	Atlanta	15.4%
9	New York City	14.9%
10	Boston	14.7%
11	San Diego	14.4%
12	Los Angeles	14.4%
13	San Francisco	14.4%
14	Chicago	14.3%
15	Las Vegas	14.2%
16	Seattle	13.9%
17	Washington, DC	12.3%
18	Minneapolis	11.7%

Source: Institute for Alternative Futures 2015

PsyD Programs

Fuente: https://laopinion.com/2019/05/30/los-ciudades-y-estados-de-eeuu-en-donde-la-diabetes-se-convertira-en-un-grave-problema-para-el-2030/

2. Hipótesis

Los factores como la edad, el índice de masa corporal (IMC), la presión arterial y los niveles de actividad física son altamente predictivos del riesgo de diabetes. Se espera que los pacientes de mayor edad, IMC y presión arterial, sumado a niveles menores de actividad física tienen una mayor probabilidad de padecer diabetes.

3. Objetivos y Alcance

- El **objetivo** es desarrollar un modelo de clasificación donde se pueda predecir el desarrollo de diabetes basado en información disponible en encuestas sobre variables relacionadas a la salud, hábitos, parámetros socioeconómicos y demográficos.

- Identificar los factores de riesgo más significativos que predicen el riesgo de diabetes en los individuos puede ser útil para profesionales de salud, estudiantes e investigadores de áreas relacionadas, ya que permitirá identificar patrones en los datos que puedan influir en el desarrollo de diabetes, ayudando en la toma de decisiones de prevención y tratamientos.

CODER HOUSE

4. Data Acquisition

Metadata

- **Descripción de los datos**

Respuestas sobre comportamientos de riesgo para la salud, condiciones de salud crónicas y el uso de servicios preventivos. Son encuestas telefónicas realizadas por el Behavioral Risk Factor Surveillance System (BRFSS) recopilada anualmente por los CDC en Estados Unidos.

- **Volumen estimado:**

250000 registros en el año 2015

- **Fuente:**

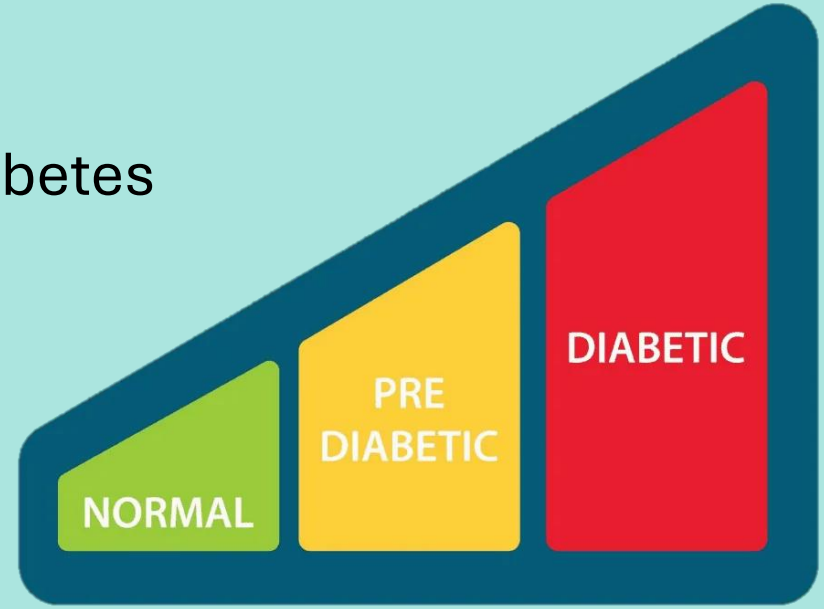
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Metadata

Glosario de datos

Variable objetivo: descripción

Diabetes_012: 0 = no diabetes, 1 = prediabetes, 2 = diabetes



CODER HOUSE

Metadata

Nombre de variable	Rol	Descripción
Diabetes_012	Objetivo	0 = no diabetes, 1 = prediabetes, 2 = diabetes
HighBP	Característica	0 = no high blood pressure (BP); 1 = high BP
HighChol	Característica	0 = no high cholesterol (Chol)
CholCheck	Característica	0 = no high Chol check in 5 years; 1 = yes Chol check in 5 years
BMI	Característica	Body Mass Index
Smoker	Característica	Have you smoked at least 100 cigarettes in your entire life? (Note: 5 packs = 100 cigarettes) 0 = no, 1 = yes
Stroke	Característica	(Ever told) you had a stroke. 0 = no, 1 = yes
HeartDiseaserorAttack	Característica	Coronary heart disease (CHD) or myocardial infarction (MI). 0 = no, 1 = yes
PhysActivity	Característica	Physical activity in past 30 days – not including job. 0 = no, 1 = yes
Fruits	Característica	Consume fruits 1 or more times per day. 0 = no, 1 = yes
Veggies	Característica	Consume vegetables 1 or more times per day. 0 = no, 1 = yes
HyvAlcoholC	Característica	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no, 1 = yes
AnyHealthcare	Característica	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes.
NoDocbcCost	Característica	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no, 1 = yes.

CODER HOUSE

Metadata

Nombre de variable	Rol	Descripción
GenHlth	Característica	Would you say that in general your health is: scale 1-5. 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.
MenHlth	Característica	Now thinking about your mental health, which includes stress, depression, and problems with emotions for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Característica	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Característica	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Sex	Característica	0 = female 1 = male
Age	Característica	13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Característica	Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary)
Income	Característica	Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more.

Importación del dataset

GitHub

'https://raw.githubusercontent.com/flordelarosa/Florencia_IFRPD-delaRosa/main/diabetes_012_health_indicators_BRFSS2015.csv'

CODER HOUSE

5. Data Wrangling

CODER HOUSE

Búsqueda y análisis de valores nulos y duplicados

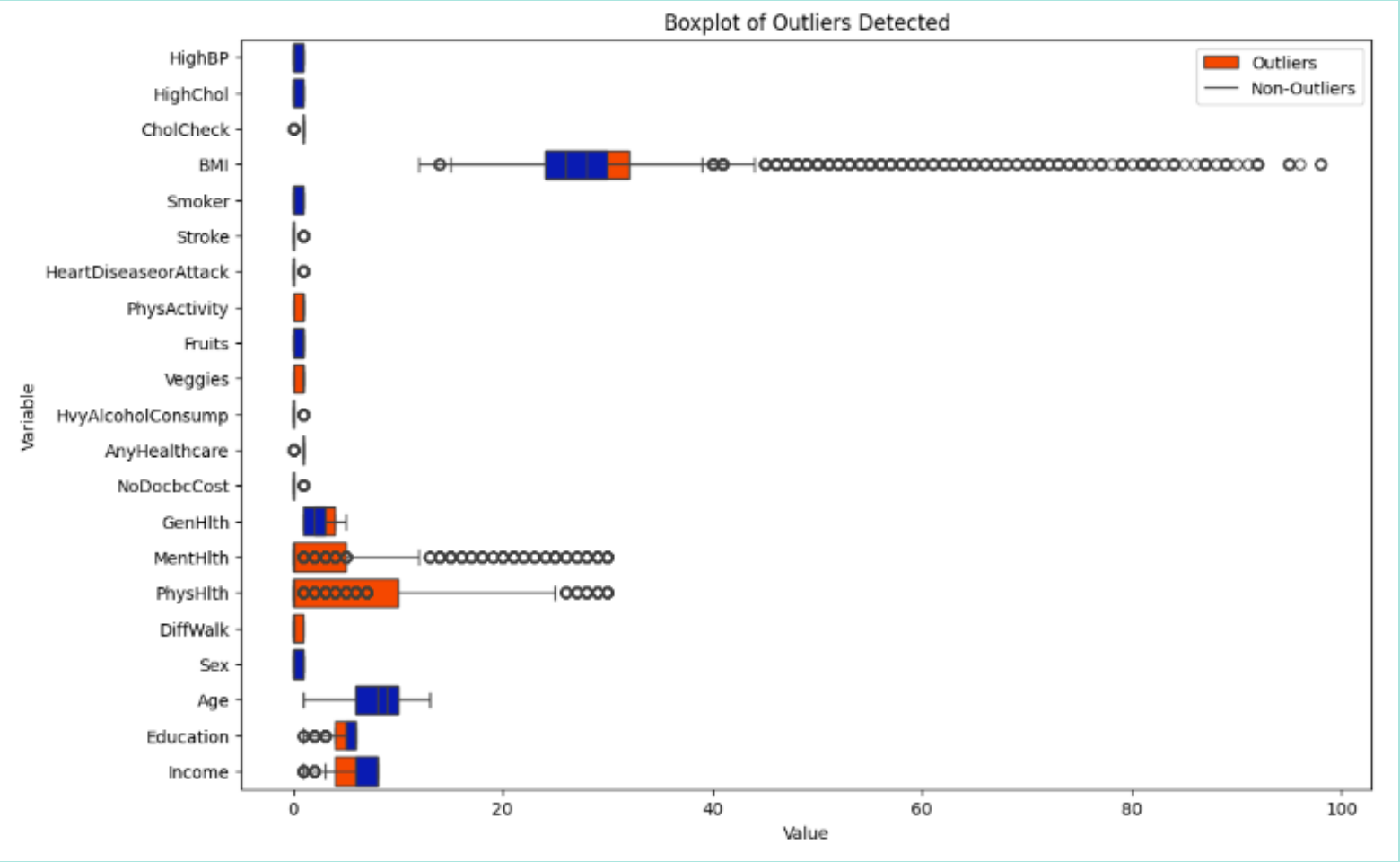
- **Valores nulos**

No hay valores nulos en el dataset.

- **Valores duplicados**

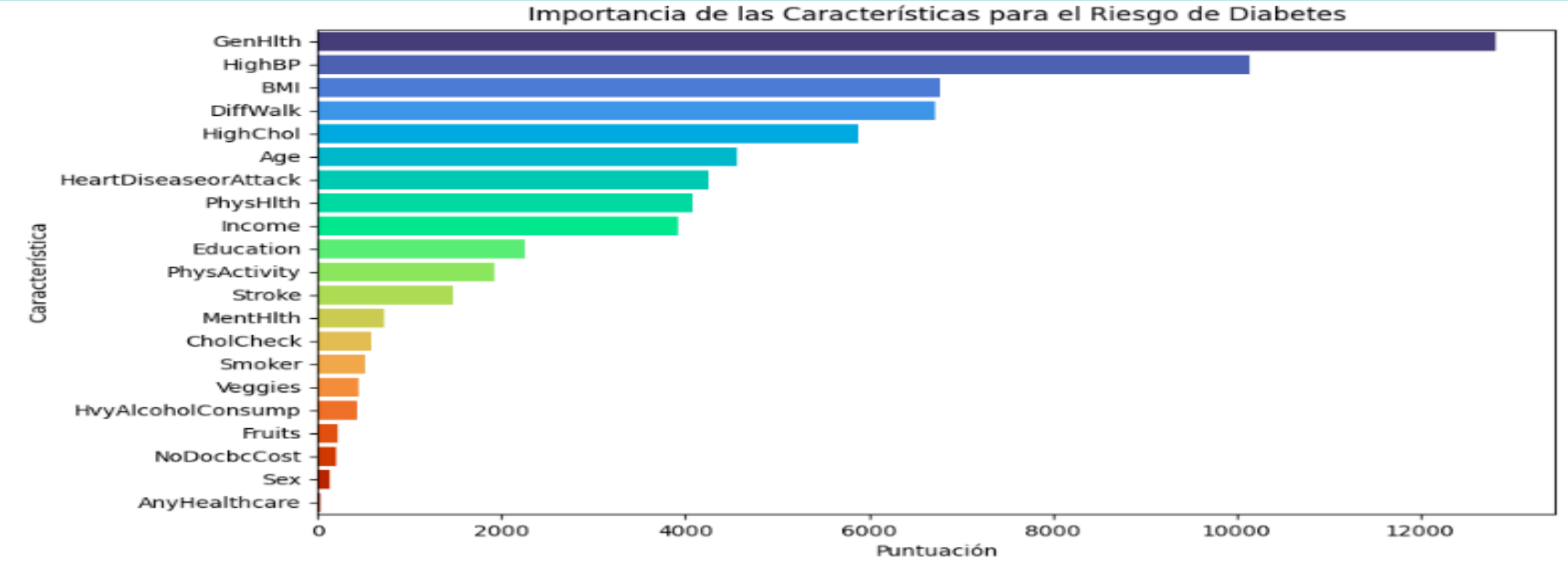
El modo de presentación de los datos da lugar a que diferentes personas completen con los mismos valores la encuesta y no necesariamente implicarían ser duplicados. Ya que la encuesta se presenta como "clean" en la fuente se procederá con la totalidad de los datos sin la eliminación de los identificados como "duplicados".

Búsqueda y análisis de valores outliers



Mantener los outliers en BMI, MenHlth y PhysHlth es clave para reflejar la variabilidad natural y evitar sesgos en el estudio sobre diabetes

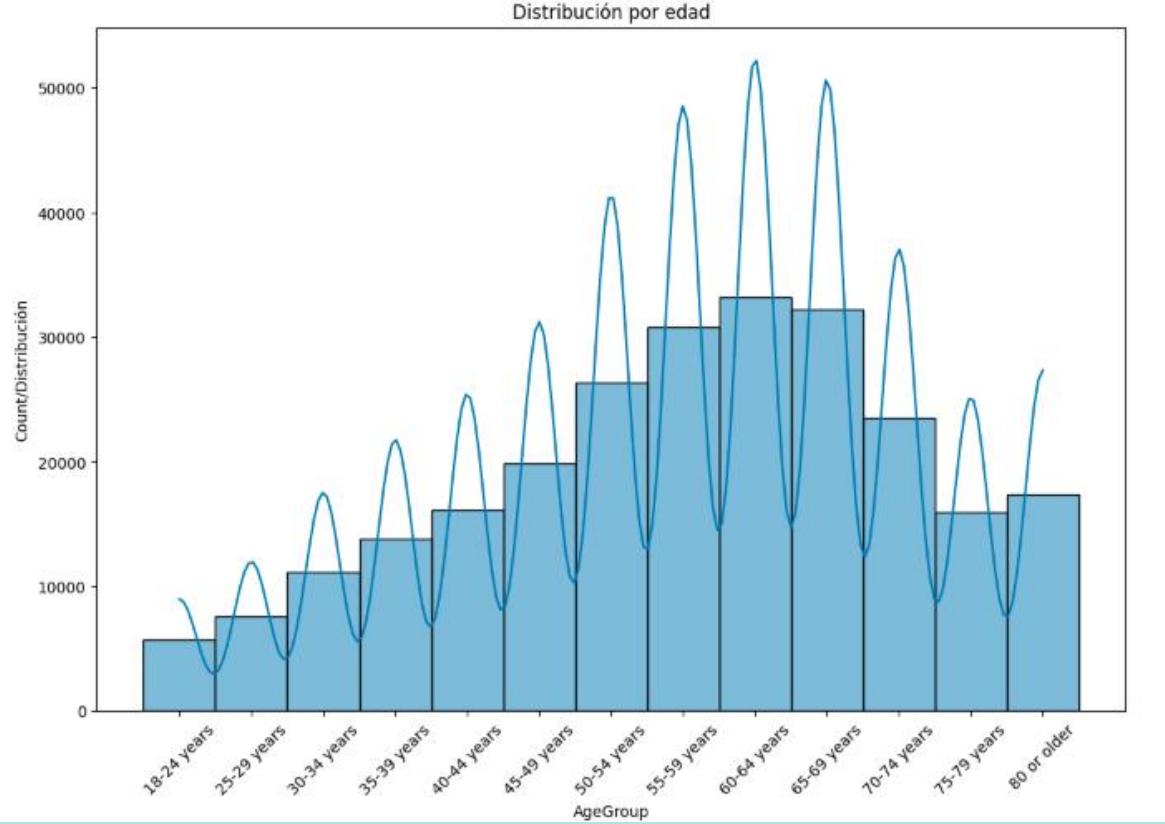
Identificación de variables importantes



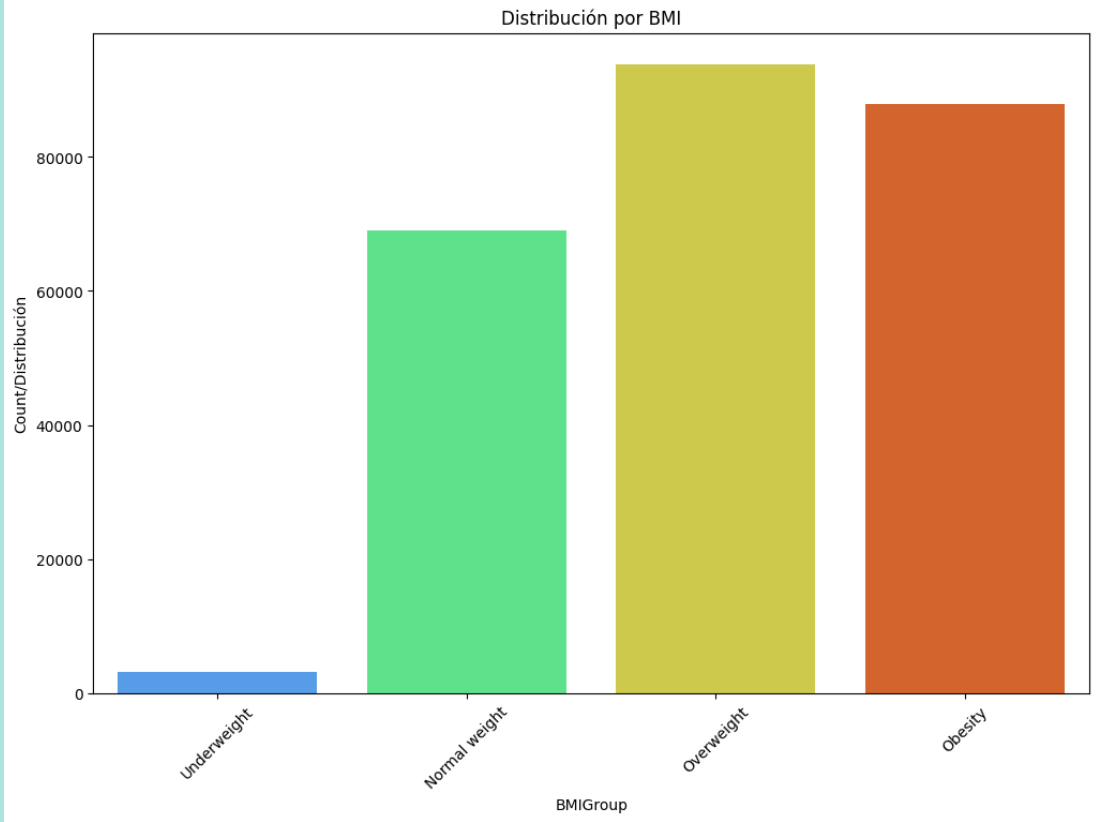
Las características más importantes para predecir la diabetes, como GenHlth, HighBP, BMI, DiffWalk, HighChol, Age, y HeartDiseaseorAttack; relacionadas a la salud y factores demográficos

6. Storytelling

Análisis de la distribución por edad y BMI



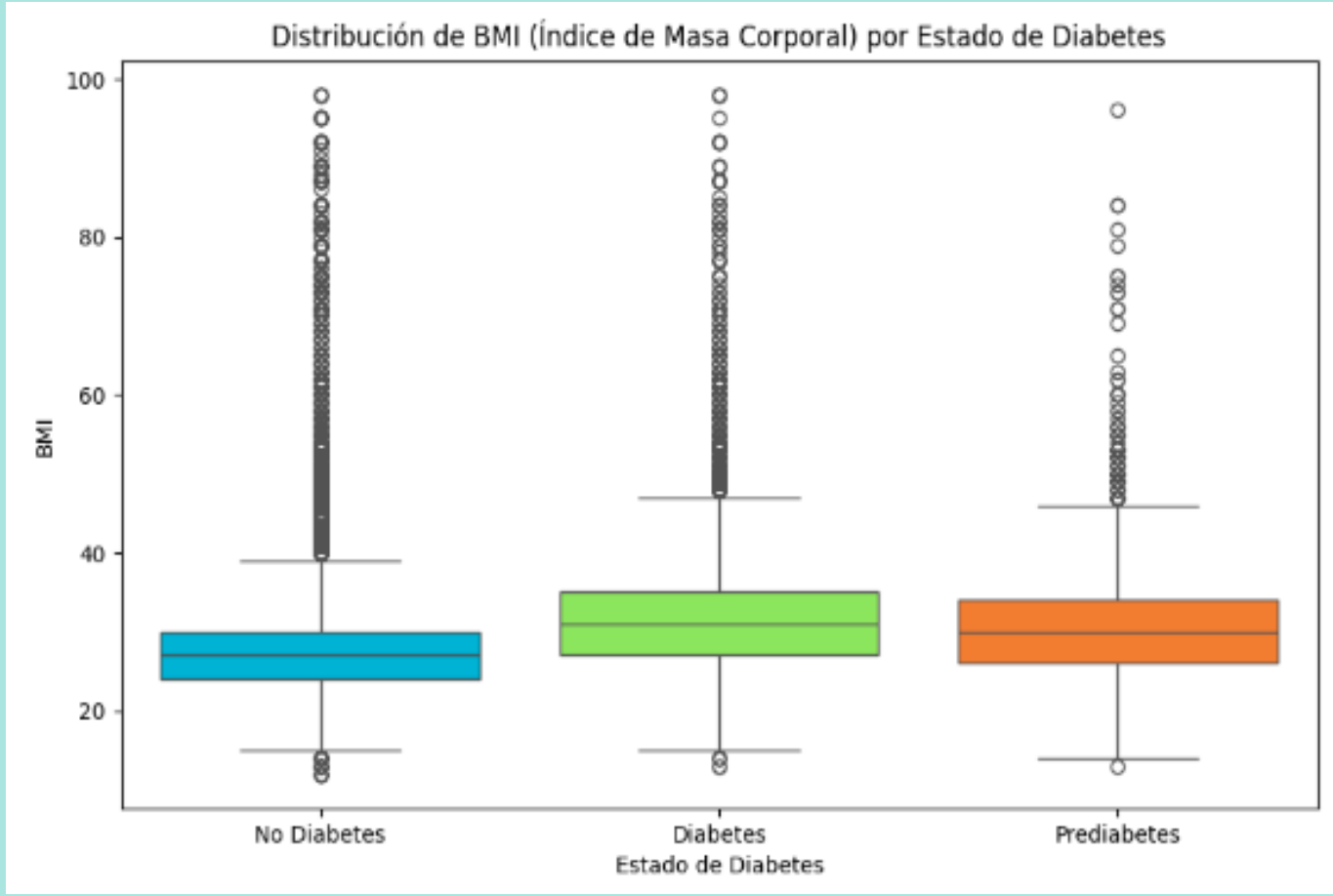
La media corresponde al grupo de **edad** de 55-59 años y existe una amplia distribución de edades



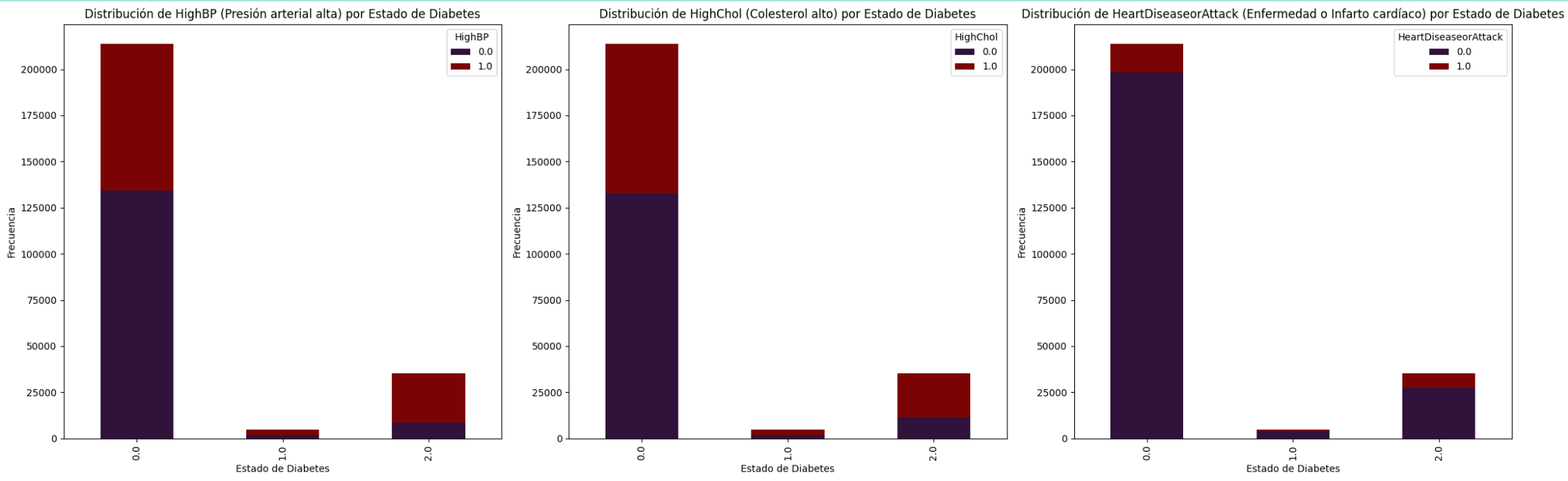
La mayoría de la población se encuentra en los grupos de sobrepeso y obesidad (**IMC** mayores)

Distribución del IMC por estado de diabetes

- El promedio de IMC indica sobrepeso
- Hay un amplio rango de IMC (mínimo = 12 y máximo = 98).
- **Aquellos con diabetes o prediabetes tienen un mayor IMC** (indicando mayor prevalencia de **obesidad y sobrepeso**)

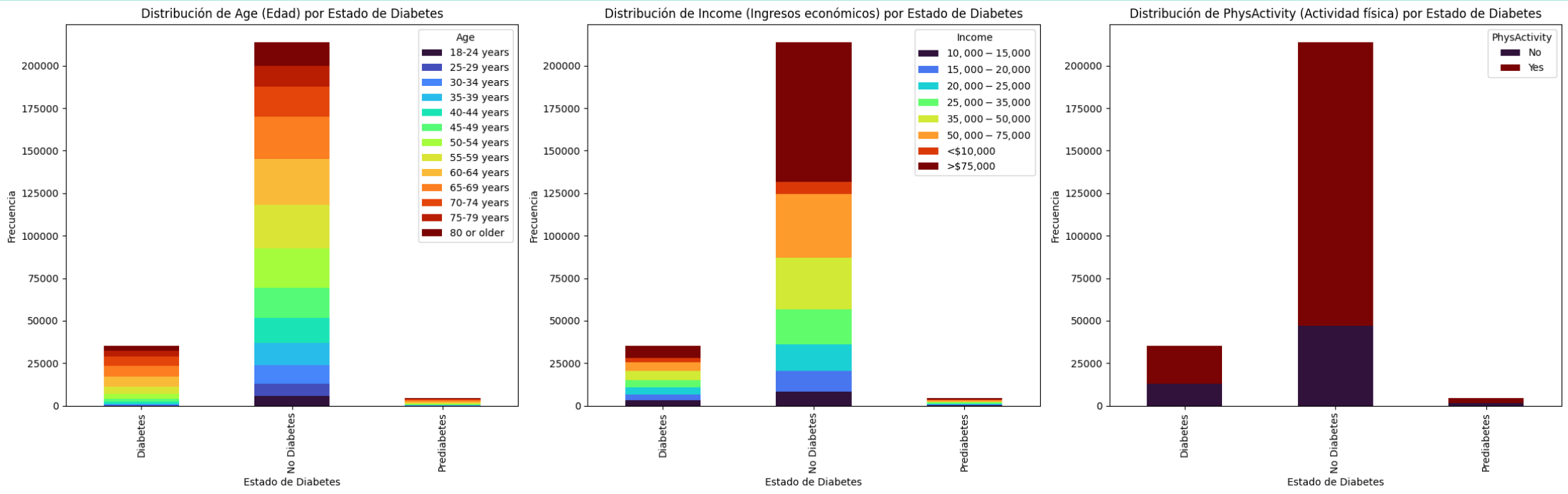


Distribución de HighBP, HighChol y HeartDiseaseorAttack por estado de diabetes



Aquellos con **diabetes** o **prediabetes** tienen una mayor incidencia de **presión arterial alta**, **colesterol alto** y **enfermedades o ataques cardíacos**

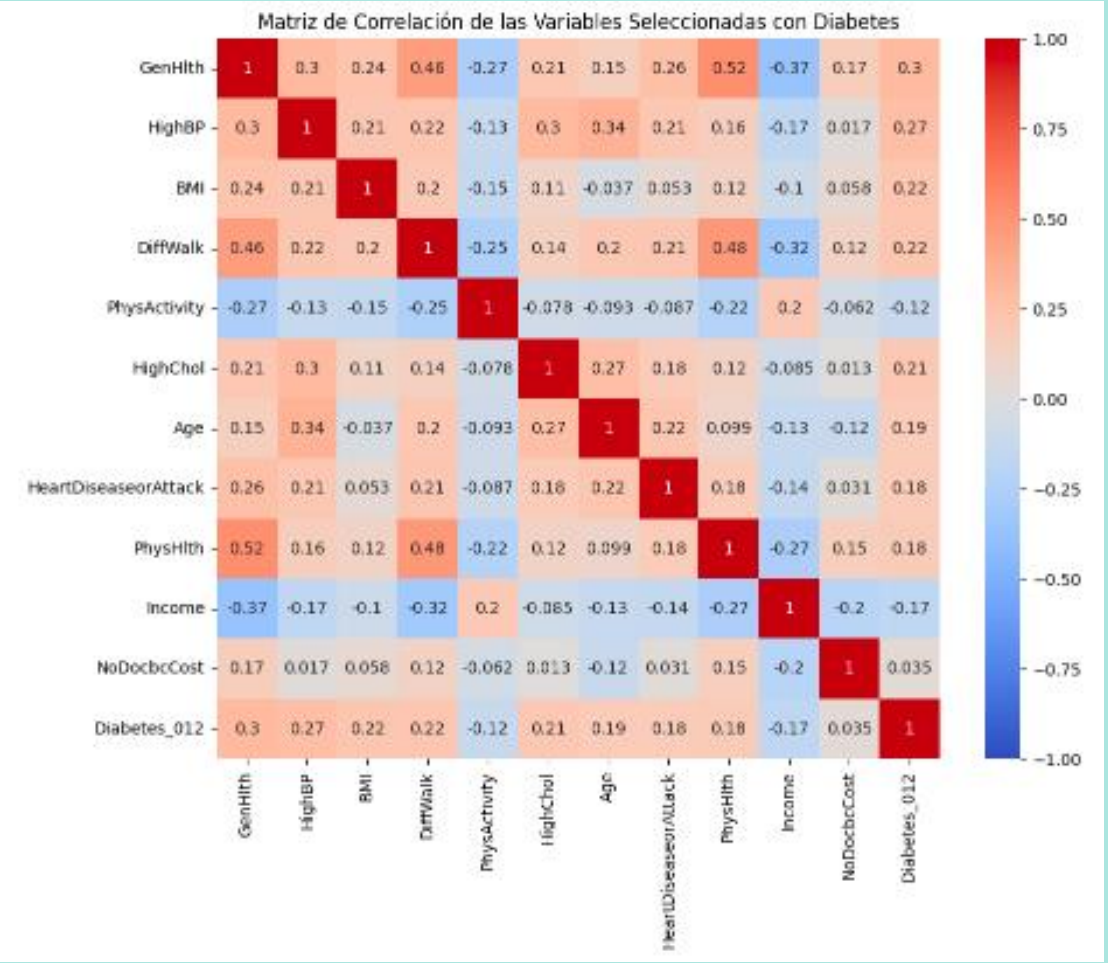
Distribución de Age, Income y PhysActivity por estado de diabetes



Las personas con **diabetes** o prediabetes realizan menos **actividad física**. La **edad** aumenta en las personas con diabetes, indicando que es más común en personas mayores. Mientras que los **ingresos promedios** son menores en aquellos que padecen la enfermedad.

Distribución del IMC por estado de diabetes

- Los factores correlacionados positivamente con la diabetes son una peor **salud general**, **hipertensión**, alto **IMC**, **dificultad para caminar**, **colesterol alto**, mayor **edad**, **enfermedades cardíacas** y mala **salud física**.
- Por otro lado, la **actividad física** e **ingresos económicos** mayores tienen una correlación negativa con la probabilidad de tener diabetes.



CODER HOUSE

7. Reporte EDA

Las personas con **diabetes** califican su **salud general** como peor y reportan más días de mala **salud física** que aquellas sin diabetes o con prediabetes.

Aquellos con diabetes o prediabetes tienen una mayor incidencia de **presión arterial alta**, **colesterol alto** y **enfermedades o ataques cardíacos**, así como un mayor **IMC** (indicando mayor prevalencia de obesidad y sobrepeso).

Las personas con diabetes o prediabetes presentan una mayor **dificultad para caminar** y menor **actividad física**.

7. Reporte EDA

La **edad** aumenta en las personas con diabetes, indicando que es más común en personas mayores. Mientras que los **ingresos promedios** son menores en aquellos que padecen la enfermedad y por ende en ocasiones no pueden **asistir al médico**.

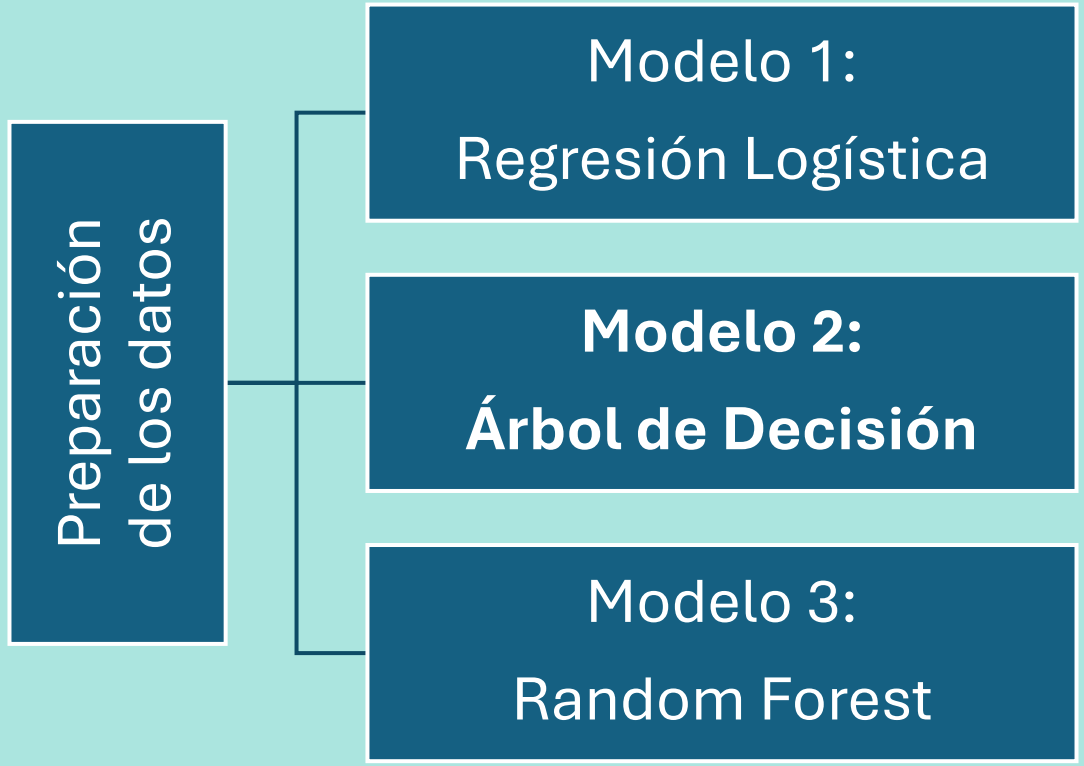
Los factores correlacionados positivamente con la diabetes son una peor **salud general**, **hipertensión**, alto **IMC**, **dificultad para caminar**, **colesterol alto**, mayor **edad**, **enfermedades cardíacas** y mala **salud física**.

Por otro lado, la **actividad física** e **ingresos económicos** mayores tienen una correlación negativa con la probabilidad de tener diabetes.

8. Selección del Modelo

8. Selección del Modelo

- El **modelo de regresión logística** clasifica bien la clase mayoritaria (**No Diabetes**) pero falla en Prediabetes y Diabetes, mostrando un sesgo hacia la clase mayoritaria.
- El **árbol de decisión** mejora un poco, pero **aún tiene dificultades con Prediabetes**.
- **Random Forest** es mejor para Diabetes pero **sigue fallando en Prediabetes**.
- **Conclusión:** El **árbol de decisión** es el mejor modelo **en términos de equilibrio**, aunque **todos tienen problemas con Prediabetes**.



9. Mejora del Modelo

9. Mejora del Modelo: Árbol de Decisión

1. División y Normalización de Datos
2. Balance de las clases con SMOTE
3. Reducción de la dimensionalidad con PCA
4. Búsqueda aleatoria de Hiperparámetros
5. Evaluación del modelo (validación cruzada y conjunto de prueba)

El modelo tiene un **buen rendimiento** en la clase mayoritaria (**No diabetes**) con un F1 Score de 0.83, pero muestra un **rendimiento bajo en las clases minoritarias** (Prediabetes y Diabetes).

Esto sugiere un problema de **desequilibrio de clases**

9. Mejora del Modelo: Cambio a XGBClassifier I

Se cambió al modelo XGBClassifier para mejorar la identificación de prediabetes y diabetes, ya que es más robusto y flexible que el DecisionTreeClassifier.

1. División y Normalización de Datos
2. Balance de las clases con SMOTE
3. Reducción de la dimensionalidad con PCA (a 5 componentes)
4. Búsqueda aleatoria de Hiperparámetros (RandomizedSearchCV)
5. Evaluación del modelo (validación cruzada y conjunto de prueba)

El XGBClassifier tuvo **buen rendimiento en la clase mayoritaria**, pero un desempeño **deficiente en las clases minoritarias**, con una precisión especialmente baja en la clase Prediabetes (3%)

El modelo necesita mejoras para **identificar mejor las clases minoritarias**

9. Mejora del Modelo: XGBClassifier II

Se realizaron ajustes al modelo, utilizando SMOTEENN, ampliando la búsqueda de hiperparámetros y aprovechando las capacidades de XGBoost, resultando en un modelo más robusto y preciso.

1. División y Normalización de Datos
2. Balance de las clases con **SMOTEENN**
3. Reducción de la dimensionalidad con PCA (a 5 componentes)
4. Búsqueda aleatoria de Hiperparámetros (RandomizedSearchCV)
5. Evaluación del modelo (validación cruzada y conjunto de prueba)

El modelo XGBClassifier mostró un **buen rendimiento** en la clase mayoritaria (**No diabetes**), con **dificultades en las clases minoritarias**, en la clase Prediabetes con una precisión del 3%.

Aunque **la clase Diabetes tuvo una mejora en el recall**, el desempeño general del modelo **necesita más ajustes** para mejorar la identificación de las **clases minoritarias**.

10. Conclusiones

CODER HOUSE

10. Conclusiones

Se **seleccionó el modelo XGBClassifier II** por ser el más adecuado para los objetivos.

Aunque **todos los modelos tienen un F1-score bajo para la clase de prediabetes**, el modelo **XGBClassifier II ofrece el mejor recall para** la clase 1 **(prediabetes) y un buen desempeño** en la clase 2 **(diabetes).**

Esto es crítico porque permite identificar tanto los casos de prediabetes como los de diabetes, lo que es fundamental para las intervenciones tempranas y el tratamiento efectivo.

10. Conclusiones

Los **factores de riesgo más significativos** identificados incluyen **peor salud general**, **hipertensión**, **alto IMC**, **dificultad para caminar**, **colesterol alto**, **mayor edad**, **enfermedades cardíacas** y **baja actividad física**.

Se necesitan más ajustes y podrían considerarse técnicas adicionales como el **ajuste de umbrales de decisión**, mejorar la **ponderación de las clases** o explorar **modelos más complejos** que manejen mejor las **clases desequilibradas**.

CODER HOUSE

¡Muchas gracias!

Estudiante: Florencia de la Rosa
Profesor: German Rodriguez
Tutor: Ignacio Fernández
Comisión: #60895
Entrega Final: 21/08/2024