CODER HOUSE

Identificación de Factores de Riesgo Predictivos para la Diabetes

- Estudiante: Florencia de la Rosa
- Profesor: German Rodriguez
- Tutor: Ignacio Fernández
- Comisión: #60895
- Preentrega: 03/07/2024

Tabla de Contenido

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.

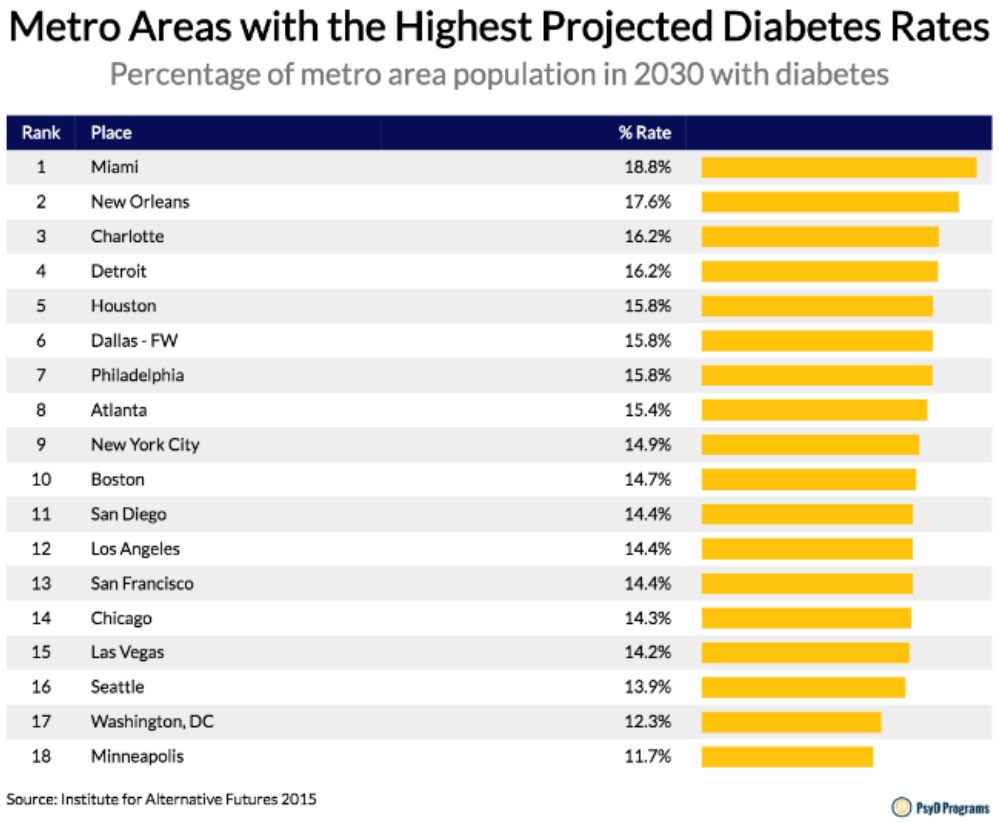
- Es una enfermedad crónica grave en la que los individuos pierden la capacidad de regular efectivamente los niveles de glucosa en la sangre, lo que puede llevar a una reducción de la calidad y expectativa de vida.

1. Contexto

El **Institute for Alternative Futures** pronosticó en el 2015 que para el 2030:

- Los estados con una prevalencia mayor de diabetes serán Miami, New Orleans, Charlotte, Detroit y Houston.

- 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.

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 Sistema de Vigilancia de Factores de riesgo del comportamiento (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 (Traducción): descripción

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

Característica (Traducción): descripción

- HighBP (High Blood Pressure o Presión arterial alta): 0 = no High BP, 1 = yes High BP
- HighChol (High cholesterol o Colesterol alto): 0 = no High Chol, 1 = yes High Chol
- CholCheck (Cholesterol check in the last 5 years o Chequeo de colesterol en los últimos 5 años): 0 = no high Chol check in 5 years, 1 = yes Chol check in 5 years
- BMI (Body Mass Index o Índice de Masa Corporal): es una medida que se calcula dividiendo el peso de una persona (en kilogramos) por el cuadrado de su altura (en metros).
- Smoker (Fumador): Have you smoked at least 100 cigarettes in your entire life? (Note: 5 packs = 100 cigarettes) 0 = no, 1 = yes
- Stroke (Derrame cerebral): Have you ever had a stroke? 0 = no, 1 = yes

Metadata

Glosario de datos

Característica (Traducción): descripción

- HeartDiseaserorAttack (Enfermedad o Infarto cardíaco): Coronary heart disease (CHD) or myocardial infarction (MI). 0 = no, 1 = yes
- PhysActivity (Actividad física): Physical activity in past 30 days – not including job. 0 = no, 1 = yes
- Fruits (Frutas): Consume fruits 1 or more times per day. 0 = no, 1 = yes
- Veggies (Vegetales): Consume vegetables 1 or more times per day. 0 = no, 1 = yes
- HyvAlcoholC (Heavy Alcohol Conssumption o Consumo de alcohol alto): 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 (Any Health Care o Alguna Asistencia Médica): Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes.
- NoDocbcCost (No Doctor because cost o No Doctor por el costo): 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.

Metadata

Glosario de datos

Característica (Traducción): descripción

- GenHlth (Genetic Health o Salud genética): Would you say that in general your health is: scale 1-5. 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.
- MenHlth (Mental Health o Salud mental): 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 (Physical Health o Salud física): 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 (Differential walk o Caminata diferente): Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
- Sex (Sexo): 0 = female 1 = male
- Age (Edad): 1 = 18-24 years, 2 = 25-29 years, 3 = 30-34 years, 4 = 35-39 years, 5 = 40-44 years, 6 = 45-49 years, 7 = 50-54 years, 8 = 55-59 years, 9 = 60-64 years, 10 = 65-69 years, 11 = 70-74 years, 12 = 75-79 years 13 = 80 or older
- Education (Educación): 1 = Never attended school or only kindergarten, 2 = Elementary school (Grades 1-8), 3 = Some high school (Grades 9-11), 4 = High school graduate (Grade 12 or GED), 5 = Some college or technical school (College 1-3 years), 6 = College graduate (College 4 years or more).
- Income (Ingresos económicos en dólares): 1 = Less than 10.000;2= 10.000- 15.000;3= 15.000- 20.000;4= 20.000- 25.000;5= 25.000- 35.000;6= 35.000- 50.000;7= 50.000- 75.000;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

Búsqueda y análisis de valores nulos y duplicados

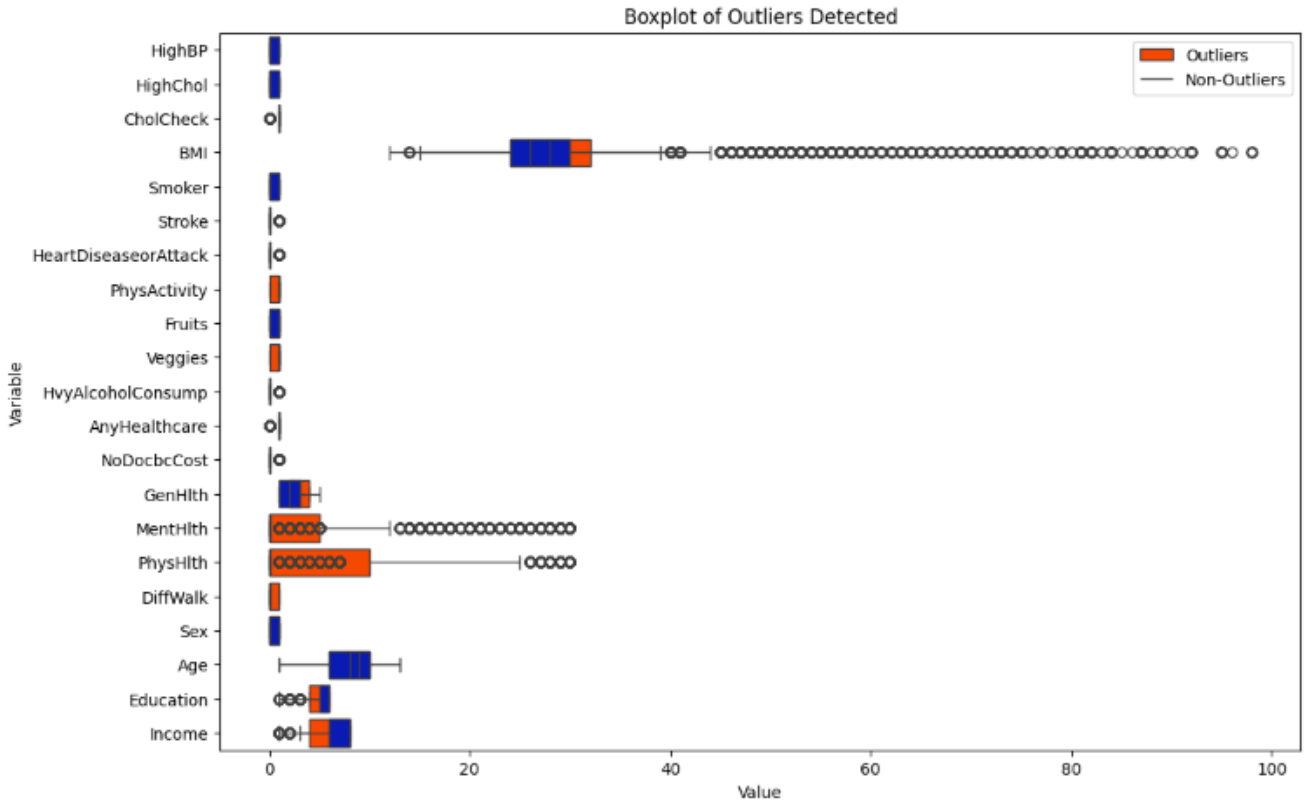
- Valores nulos

No hay valores nulos en el dataset.

- Valores duplicados

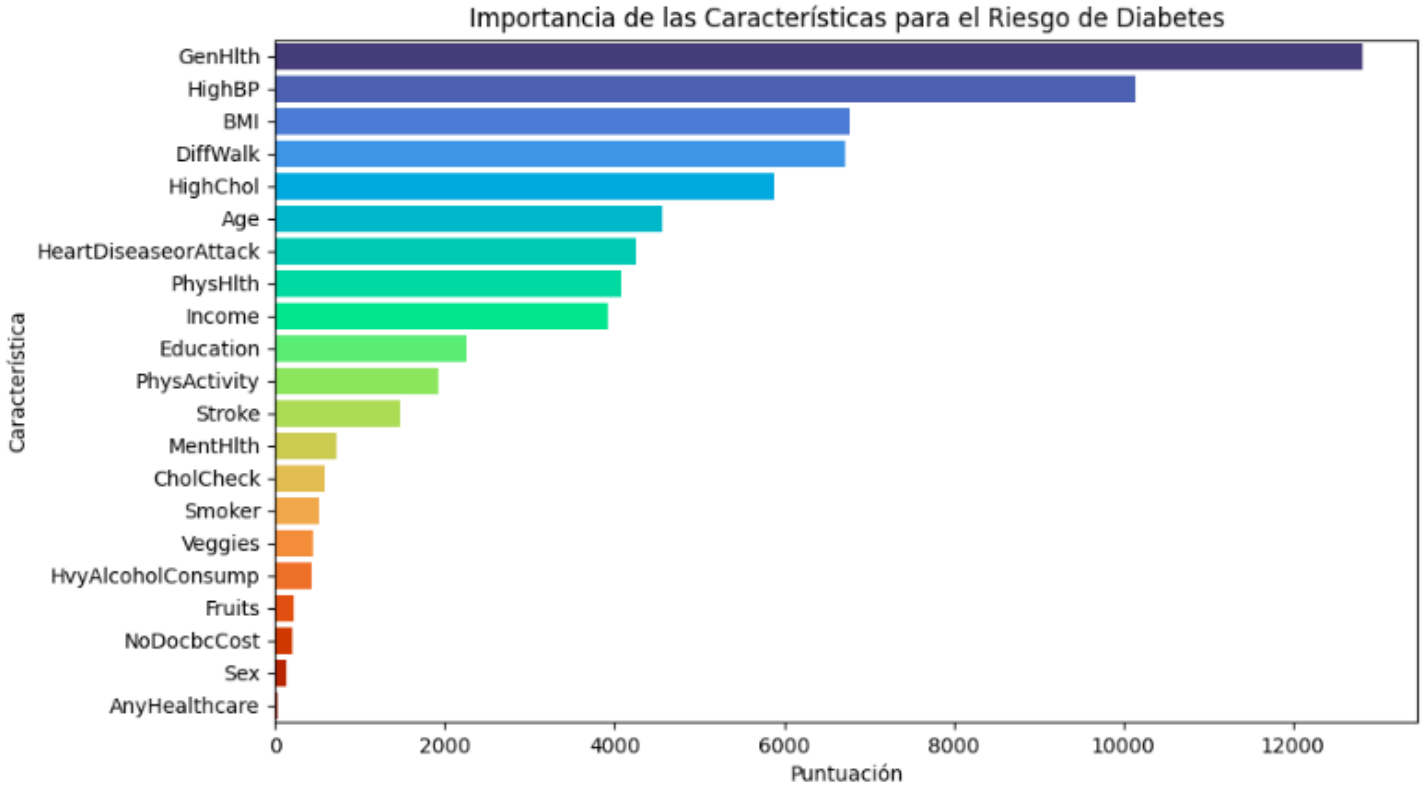
El modo de presentación de los datos da lugar a que diferentes personas completen con los mismos valores la encuesta y no necesariamente implicarían ser duplicados. Ya que la encuesta se presenta como "clean" en la fuente se procederá con la totalidad de los datos sin la eliminación de los identificados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

Búsqueda y análisis de valores outliers



Mas adelante tomaré decisiones para ver si es conveniente o no eliminar outliers. Aunque los outliers de BMI, MenHlth y PhysHlth son importantes para el estudio sobre la estimación de diabetes. Representan casos excepcionales que reflejan la variabilidad natural en la población y pueden proporcionar información relevante sobre condiciones médicas particulares o características únicas de los participantes. Mantener estos outliers permite realizar un análisis más completo y robusto, evitando la pérdida de información valiosa y sesgos en los resultados.

Identificación de variables importantes

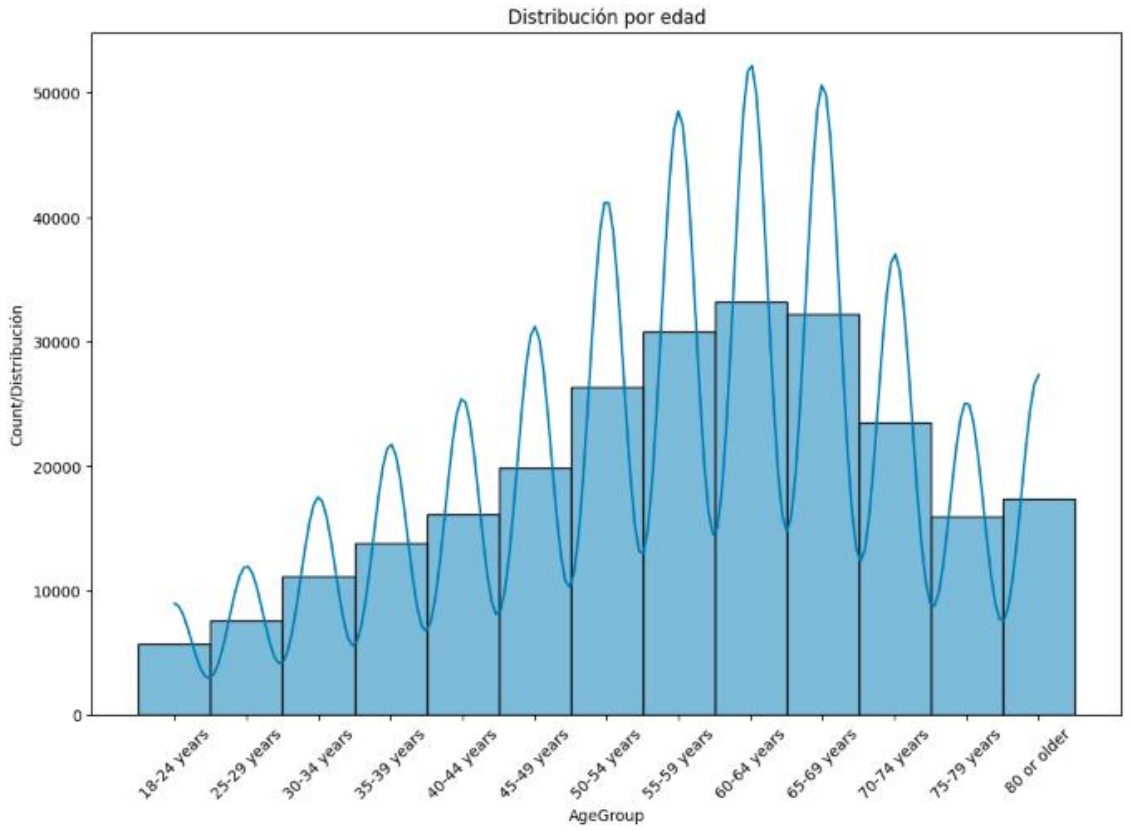


Las características con puntuaciones más altas (por ejemplo, GenHlth, HighBP, BMI, DiffWalk, HighChol, Age, HeartDiseaseorAttack etc.) son las más importantes para el modelo predictivo y probablemente tienen una fuerte relación con el estado de diabetes.

Se dividieron las características seleccionadas entre las que se encuentran relacionada con la salud y las demográficas o socioeconómicas.

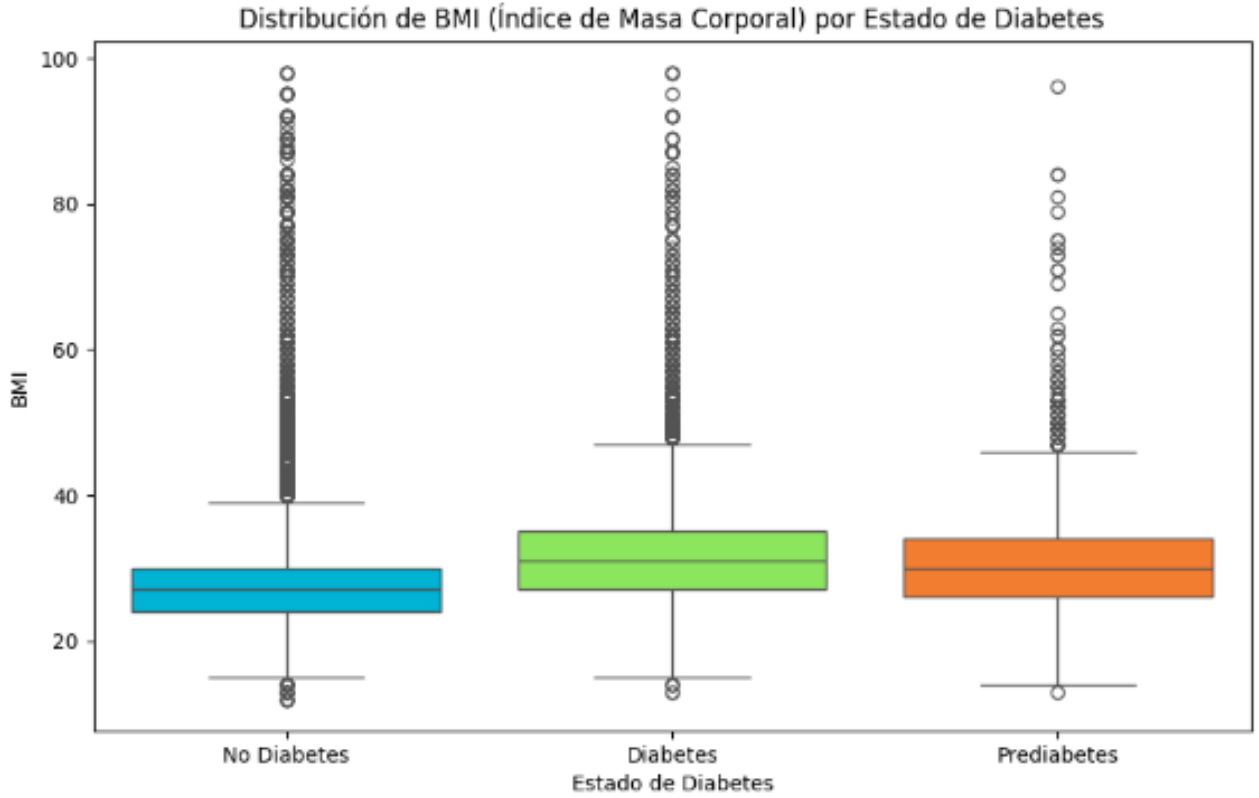
6. Storytelling

Análisis de la distribución por edad



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

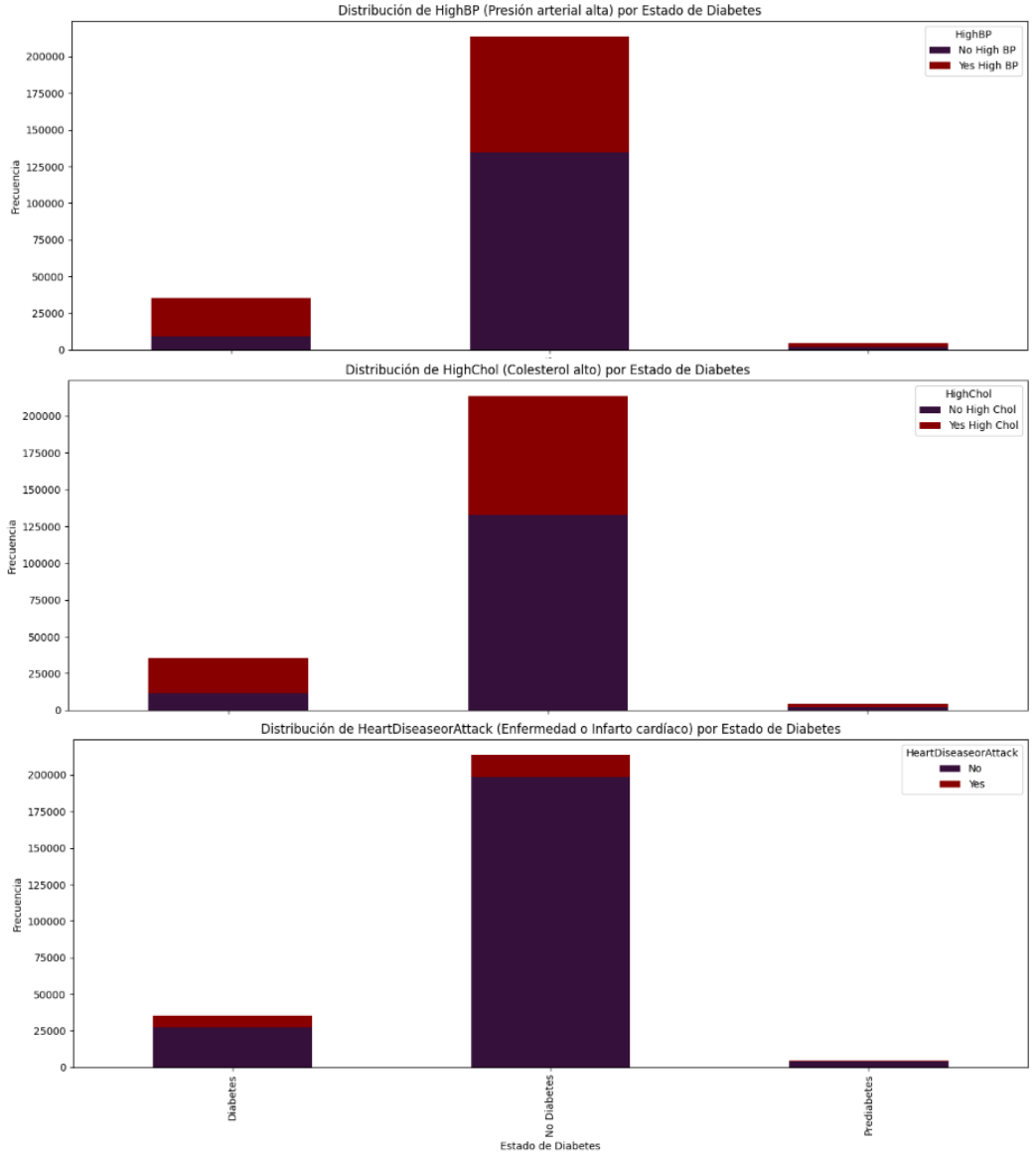
Distribución del IMC por estado de diabetes



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).

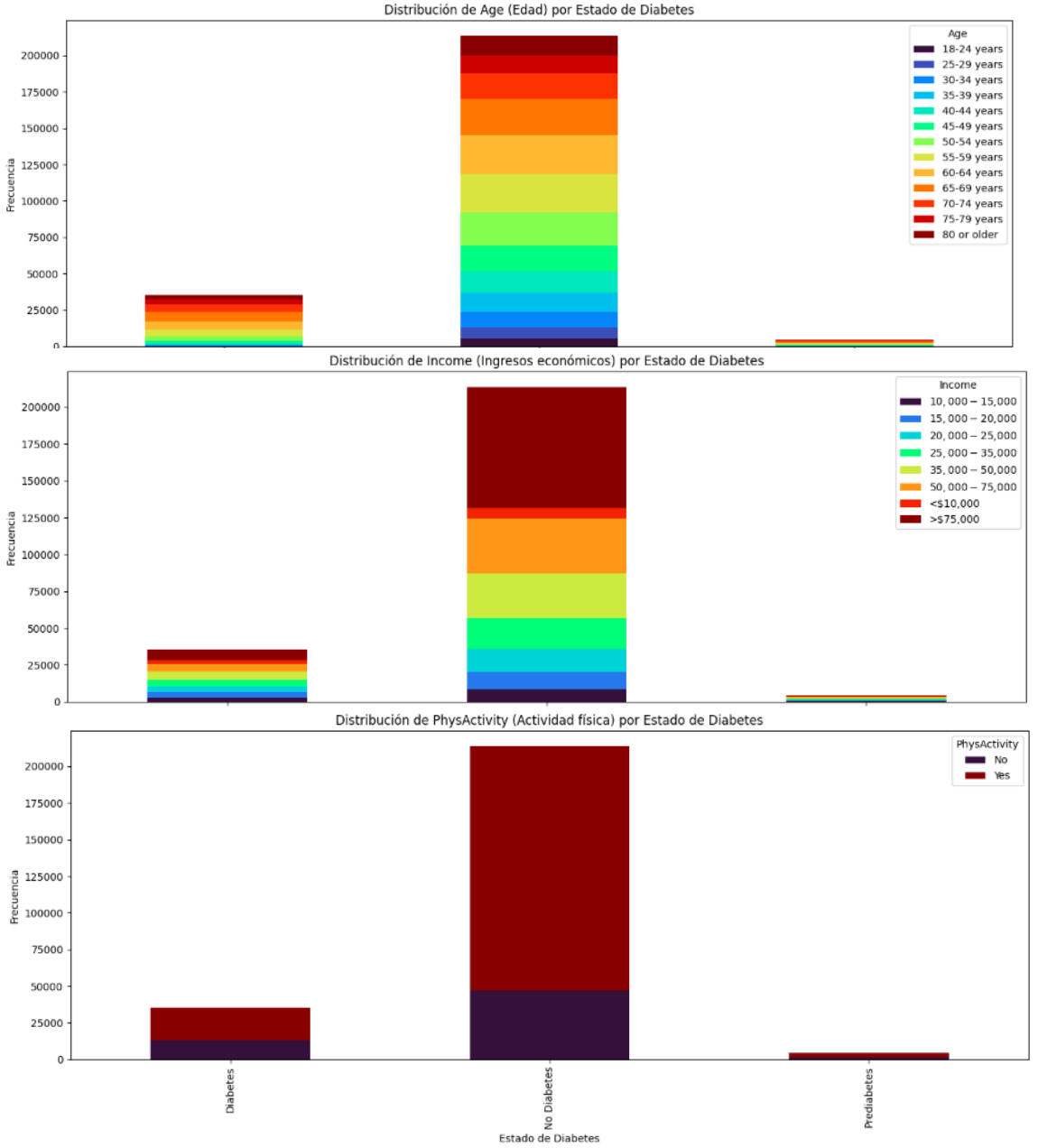
Distribución de HighBP, HighChol y HeartDiseaseorAttack por estado de diabetes

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



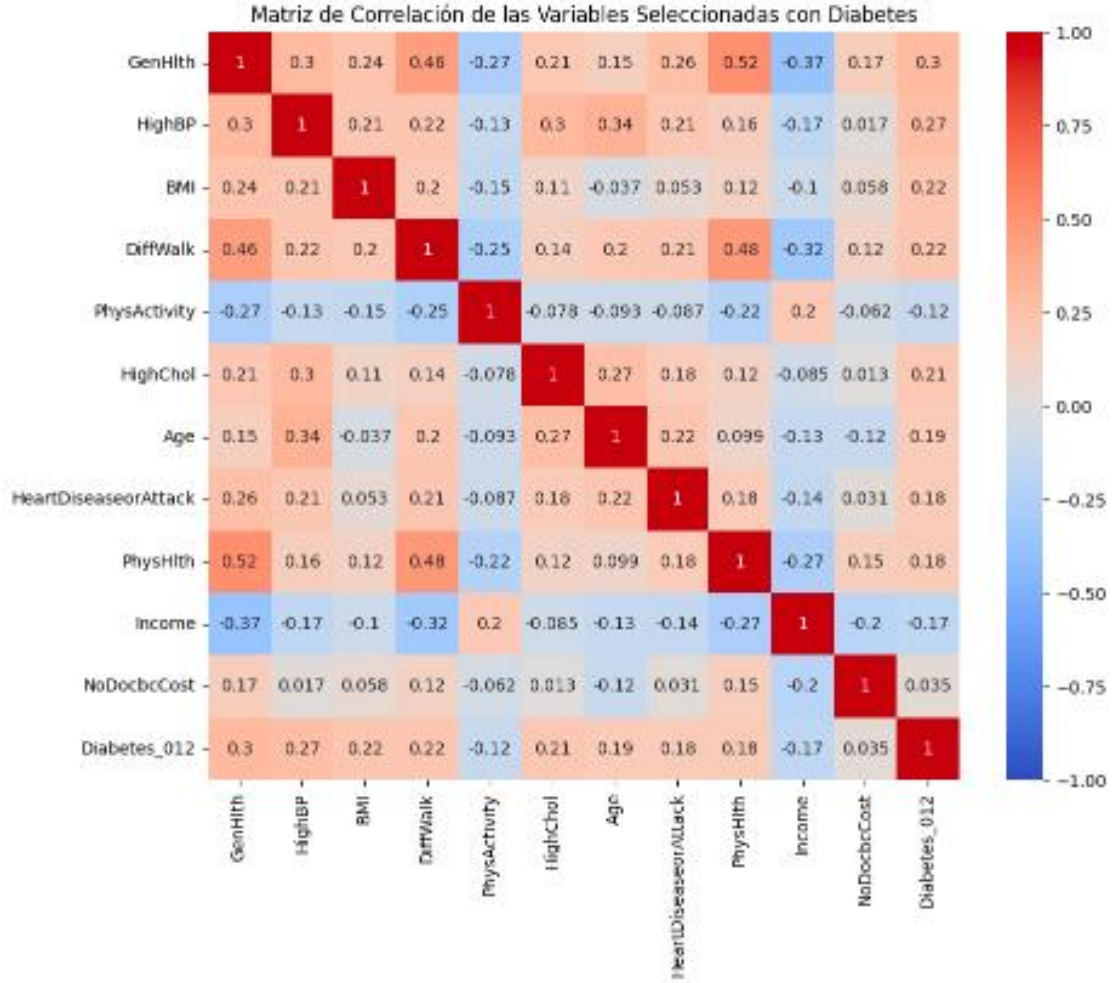
Distribución de Age, Income y PhysActivity por estado de diabetes

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.



Análisis de correlación

- 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.



7. Conclusiones preliminares

- 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**.
- 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.

CODER HOUSE

¡Muchas gracias!

Estudiante: Florencia de la Rosa
Profesor: German Rodriguez
Tutor: Ignacio Fernández
Comisión: #60895
Preentrega: 03/07/2024