Diabetes Mellitus en EE.UU. 2015

Factores de Riesgo y Protectores

AUTORES:

Ángel Ciria

Agustín Nahuel Quiroga Baigorrí

Resumen

La diabetes mellitus es una enfermedad crónica grave en la que las personas pierden la capacidad de regular eficazmente los niveles de glucosa en la sangre y puede reducir tanto la calidad, como la esperanza de vida. La prevalencia mundial de la diabetes mellitus ha aumentado drásticamente en los últimos 20 años, de 30 millones de casos en 1985 se ha pasado a 177 millones en el año 2000. Basándonos en las tendencias actuales, más de 360 millones de personas padecerán diabetes en el año 2030 según *International Textbook of Diabetes Mellitus, 3rd ed. John Wiley & Sons, 2004.*

El motivo de este trabajo es encontrar factores de riesgo que desencadenan la enfermedad y hábitos saludables que ayuden a prevenirla. Para ello se trabajó con el Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS, por sus siglas en inglés) una encuesta telefónica relacionada con la salud que los centros para el control y la prevención de enfermedades en los Estados Unidos que se realiza anualmente. En nuestro caso seleccionamos el dataset que surge del año 2015 y consta de 253680 muestras y 22 atributos de los que podemos destacar: estadio de la enfermedad, si la persona posee presión sanguínea alta y/o colesterol alto, índice de masa corporal, apreciación de su salud mental como general, consumo de frutas, factores socioeconómicos, entre otros.

Introducción
- Objetivos
  - Principales
  - Secundarios
- Audiencia
- Contexto Clínico
- Contexto Histórico
- Contexto Analítico
- Problema
- Metadata
- Descripción de Variables
- Información Adicional

Análisis exploratorio de datos
- ¿Qué porcentaje de la población padece esta dolencia?
- ¿Cómo son los porcentajes de diabéticos en los distintos rangos etarios?
- ¿Qué información nos proporciona el índice de masa corporal?
- ¿Qué información nos proporciona el estado de salud general de la persona?
- ¿Cómo se relaciona la diabetes y el colesterol?
- ¿Cómo se relacionan la diabetes y la hipertensión?
- Enfermedad Coronaria / Ataque al miocardio
- Dificultad para subir escaleras
- Factores Socio-Económicos
  - ¿Cómo se distribuyen los ingresos según estatus de diabetes?
  - Nivel Educativo
- ¿Qué hábitos se pueden considerar preventores de la enfermedad?
  - Análisis para Frutas
  - Análisis para Vegetales
  - Análisis para Actividad Física

Selección de Algoritmo Apropiado
- Análisis de correlaciones
  - Matriz de correlaciones
  - Correlaciones con la variable objetivo
- Selección de características
  - Segmentación de la muestra por edades y balanceo
  - Algoritmos de selección
- Evaluación de Modelos
  - General

- Menores
- Intermedios
- Mayores
- Tuning de hiperparámetros

Conclusiones

Introducción

Objetivos

Las incógnitas que se buscan resolver son:

Principales

- ¿Se puede incidir el riesgo de padecer diabetes mellitus?
- ¿Se pueden detectar factores de riesgo que desencadenan la enfermedad y hábitos saludables que ayuden a prevenirla?

Secundarios

- ¿Qué porcentaje de la población padece esta dolencia?
- ¿Cómo es la distribución etaria de la muestra?
- ¿Cómo son los porcentajes de diabéticos en los distintos rangos etarios?
- ¿Qué información nos proporciona el índice de masa corporal?
- ¿Cómo se relaciona la diabetes y el colesterol?
- ¿Cómo se relaciona la diabetes y la hipertensión?
- ¿Cómo se distribuyen los ingresos según estatus de diabetes?
- ¿Qué hábitos se pueden considerar preventores de la enfermedad?

Audiencia

Autoridades de salud pública como privada que a partir de lo expuesto en este estudio determinarán políticas para prevenir y combatir esta dolencia.

Contexto Clínico

La diabetes es una enfermedad crónica grave en la que las personas pierden la capacidad de regular eficazmente los niveles de glucosa en la sangre y puede reducir tanto la calidad, como la esperanza de vida. El diagnóstico temprano conduce a cambios en el estilo de vida y a un tratamiento más eficaz, lo que convierte a los modelos predictivos del riesgo de diabetes en herramientas importantes para el público y los funcionarios de la salud.

Contexto Histórico

También es importante reconocer la escala de este problema, aproximadamente 62 millones de personas en las Américas (422 millones de personas en todo el mundo) tienen diabetes, la mayoría vive en países de ingresos bajos y medianos, y 244.084 muertes (1.5 millones en todo el mundo) se atribuyen directamente a la diabetes cada año. Tanto el número de casos como la prevalencia de diabetes han aumentado constantemente durante las últimas décadas. Según la Federación Internacional de Diabetes, en 2019 el gasto mundial fue del orden de los 760 mil millones de dólares y se prevé que crezca un 11% en 2045, hasta alcanzar los 845 mil millones de dólares.

Contexto Analítico

El Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS, por sus siglas en inglés) es una encuesta telefónica relacionada con la salud que los Centros para el Control y la Prevención de Enfermedades en los Estados Unidos realizan anualmente. Cada año, la encuesta recopila respuestas de más de 400.000 estadounidenses sobre comportamientos de riesgo relacionados con la salud, condiciones de salud crónicas y el uso de servicios preventivos. Se lleva a cabo

todos los años desde 1984. Para este proyecto, se utilizó un archivo .csv con un dataset disponible en Kaggle para el año 2015.

Problema

Con base en los datos recolectados mediante encuestas telefónicas y aplicando ciencia de datos y algoritmos de aprendizaje automatizado:

- ¿El carácter nominal de la mayoría de las variables perjudica el modelo?
- ¿Se puede proporcionar predicciones de alto grado de precisión sobre el riesgo de padecer diabetes?
- ¿Es posible determinar qué factores de riesgo son mejores predictivos?
- ¿Al utilizar un subconjunto de factores mejora la predicción?

Metadata

- Población de EE.UU. en el año 2.015: **320.738.994**.
- Muestras: **253.680**.
- Atributos o Variables: **22**.
    - Nominal: **15**.
    - Intervalo: **3**.
    - Ordinal: **1**.
    - Razón: **3**.
- No posee **ningún** valor nulo.
- Tiene un peso de memoria de **42,6 MB**.

- Posee **23.899** registros duplicados, pero dada la naturaleza de la base de datos no significa que cada registro duplicado sea de la misma persona si no que se tratan de distintas personas que han contestado de igual manera a las distintas preguntas.

Descripción de Variables

- **Diabetes_012**: Estadio de la enfermedad de Diabetes Mellitus.
    - 0 = Sin diabetes.
    - 1 = Prediabetes.
    - 2 = Diabetes.
- **HighBP**: Posee presión sanguínea alta.
    - 0 = No.
    - 1 = Si.
- **HighChol**: Colesterol alto.
    - 0 = No.
    - 1 = Si.
- **CholCheck**: Se realizó control de colesterol en los últimos 5 años.
    - 0 = No.
    - 1 = Si.
- **BMI**: Índice de masa corporal.
- **Smoker**: Fumó al menos 100 cigarrillos en su vida.
    - 0 = No.
    - 1 = Si.
- **Stroke**: Tuvo un derrame cerebral.
    - 0 = No.

    - 1 = Si.

- **HeartDiseaseorAttack**: Posee enfermedad coronaria o infarto de miocardio.

    - 0 = No.
    - 1 = Si.

- **PhysActivity**: Realizó actividad física en los últimos 30 días.

    - 0 = No.
    - 1 = Si.

- **Fruits**: Consume frutas 1 o más veces al día.

    - 0 = No.
    - 1 = Si.

- **Veggies**: Consume vegetales 1 o más veces al día.

    - 0 = No.
    - 1 = Si.

- **HvyAlcoholConsump**: Gran bebedor/a de alcohol.

    - Hombre: más de 14 tragos a la semana:

        - 0 = No.
        - 1 = Si.

    - Mujer: más de 7 tragos a la semana:

        - 0 = No.
        - 1 = Si.

- **AnyHealthcare**: Posee cobertura de salud.

    - 0 = No.
    - 1 = Si.

- **NoDocbcCost**: Hubo algún momento en los últimos 12 meses en que necesitó un doctor y no pudo acceder por el costo.
    - 0 = No.
    - 1 = Si.

- **GenHlth**: Diría usted que su salud general es:
    - 1 = Excelente.
    - 2 = Muy buena.
    - 3 = Buena.
    - 4 = Regular.
    - 5 = Mala.

- **MentHlth**: ¿Cuántos días de los últimos 30 padeció alguna enfermedad mental? Incluye estrés, depresión, problemas emocionales, etc.

- **PhysHlth**: ¿Cuántos días de los últimos 30 padeció alguna enfermedad física o lesión?

- **DiffWalk**: Dificultades serias para caminar o subir escaleras.
    - 0 = No.
    - 1 = Si.

- **Sex**: Género de la persona.
    - 0 = Femenino.
    - 1 = Masculino.

- **Age**: Intervalo de edad en el que se encuentra.
    - 1 = 18 a 24.
    - 2 = 25 a 29.
    - 3 = 30 a 34.
    - 4 = 35 a 39.
    - 5 = 40 a 44.
    - 6 = 45 a 49.
    - 7 = 50 a 54.
    - 8 = 55 a 59.
    - 9 = 60 a 64.

    - 10 = 65 a 69.
    - 11 = 70 a 74.
    - 12 = 75 a 79.
    - 13 = 80 o más.

- **Education**: Nivel educativo.

    - 1 = Nunca asistió a la escuela o solo al jardín de infantes.
    - 2 = Grados 1 a 8 (Elemental).
    - 3 = Grados 9 a 11 (High school).
    - 4 = Grado 12 o GED (Graduado High School).
    - 5 = College 1 a 3 años.
    - 6 = College 4 años o más (Graduado).

- **Income**: Ingresos anuales del hogar.

    - 1 = Menor a $10.000.
    - 2 = Mayor o igual a $10.000 y menor a $15.000.
    - 3 = Mayor o igual a $15.000 y menor a $20.000.
    - 4 = Mayor o igual a $20.000 y menor a $25.000.
    - 5 = Mayor o igual a $25.000 y menor a $35.000.
    - 6 = Mayor o igual a $35.000 y menor a $50.000.
    - 7 = Mayor o igual a $50.000 y menor a $75.000.
    - 8 = Mayor o igual a $75.000.

Información Adicional

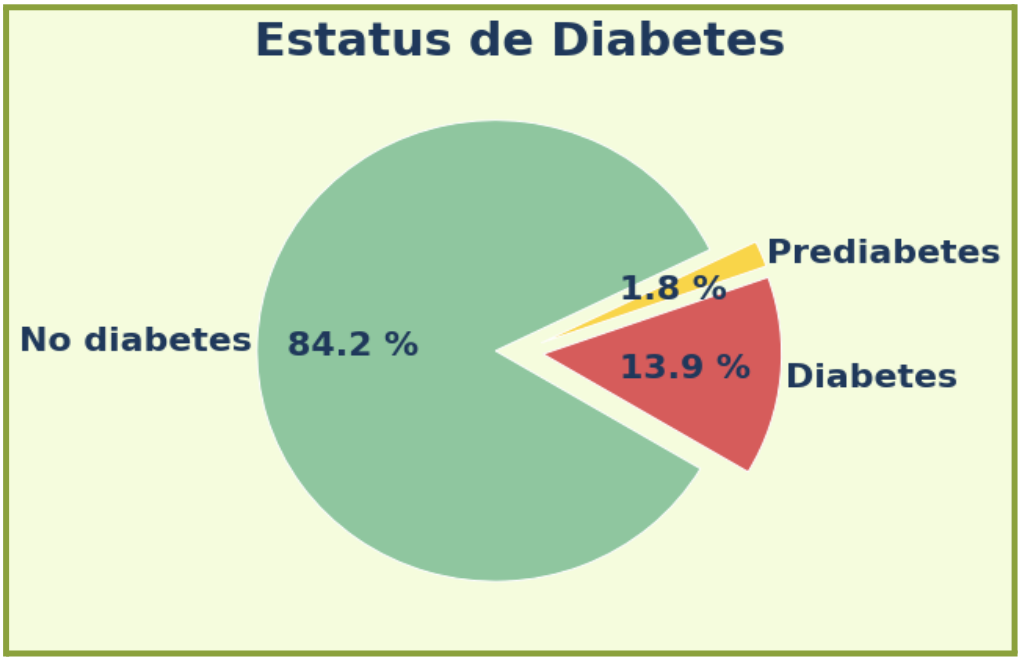
Se optó trabajar con la API de la Organización Mundial de la Salud, la cual se denomina Athena, con el fin de obtener información sobre la expectativa de vida de la población de la muestra. La documentación de dicha API se encuentra en el siguiente link.

La información extraída es la siguiente:

Género	Expectativa de Vida
Masculino	76,3
Femenino	80,8
No binario	78,6

Análisis exploratorio de datos

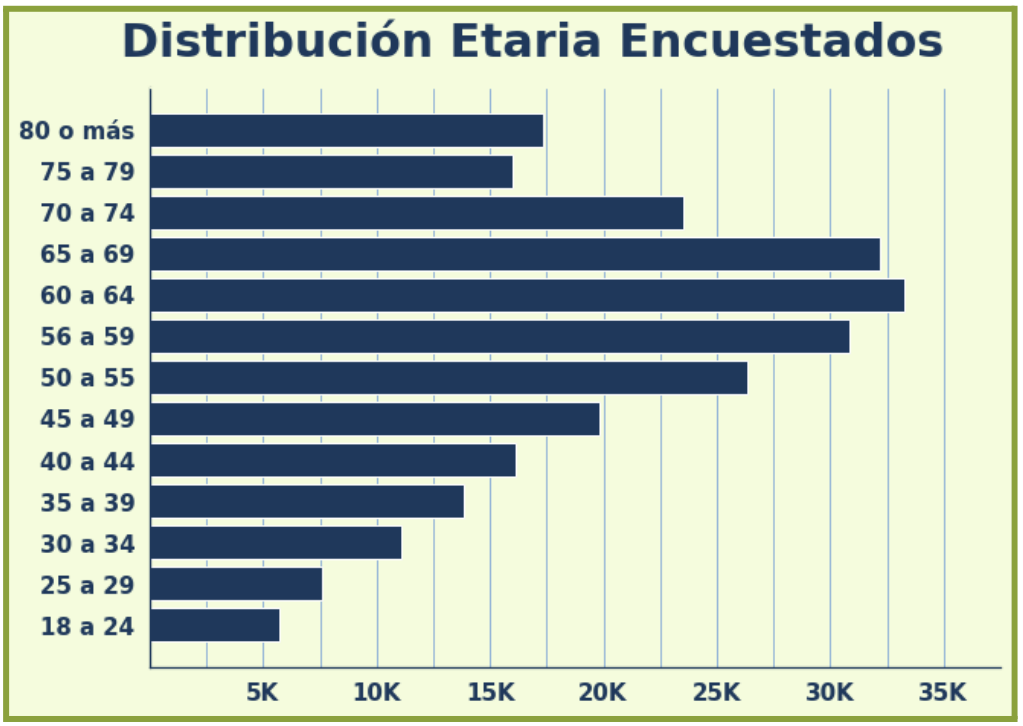
¿Qué porcentaje de la población padece esta dolencia?



Se observa en la visualización que el **84,2%** de la muestra **no padece diabetes** y que solamente un **1,8%** tiene **prediabetes**.

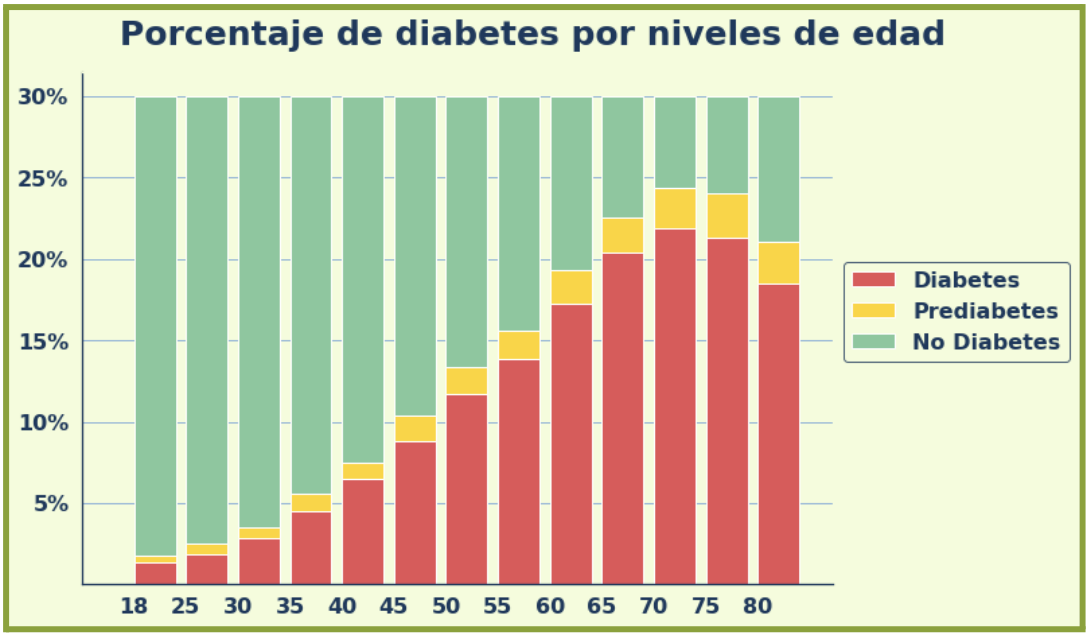
Esto también nos indica que nuestros datos se encuentran **desbalanceados** respecto de la variable a predecir y es objeto de consideración a la hora de entrenar a nuestro modelo.

¿Cómo son los porcentajes de diabéticos en los distintos rangos etarios?



Se distingue claramente que hay mayor cantidad de encuestados de los rangos etarios superiores. Ahora analizaremos en este caso la incidencia de la enfermedad en los distintos niveles de edad y lo visualizamos a través de un gráfico de barras apiladas.

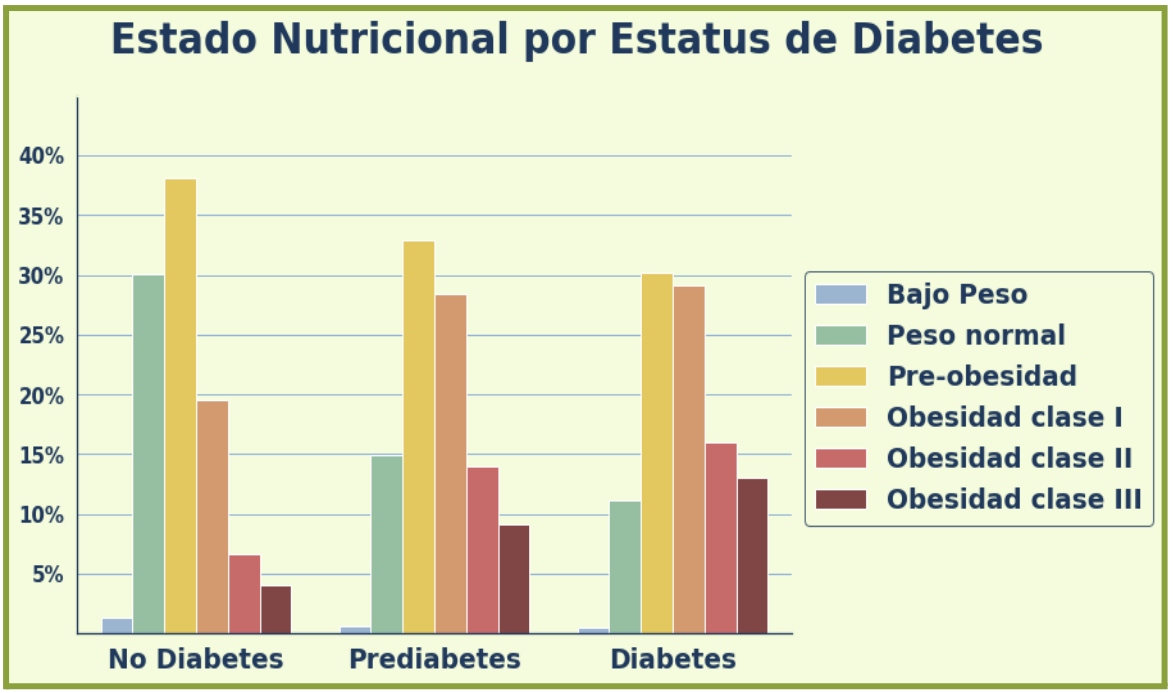
Se observa que los porcentajes de personas que padecen **diabetes** y **prediabetes crecen** al **aumentar la edad** de las personas. El rango etario con mayor porcentaje es el de **70 a los 74 años**, **22%** de personas. Luego de lo cual comienza a disminuir, esta disminución puede ser debido a que en el próximo rango etario se alcanza la expectativa de vida: 76 años para hombres y 81 para mujeres. Es decir, las personas con diabetes empiezan a fallecer por lo tanto **la expectativa de vida** de las **personas** con **diabetes** es **menor a la media**.



¿Qué información nos proporciona el índice de masa corporal?

Según la Organización Mundial de la Salud (O.M.S.) el estado nutricional basado en el Índice de Masa Corporal (I.M.C.) es el siguiente:

Debajo de 18,5	**Bajo peso**
18,5 a 24,9	**Peso normal**
25 a 29,9	**Pre-obesidad**
30 a 34,9	**Obesidad clase I**
35 a 39,9	**Obesidad clase II**
Arriba de 40	**Obesidad clase III**

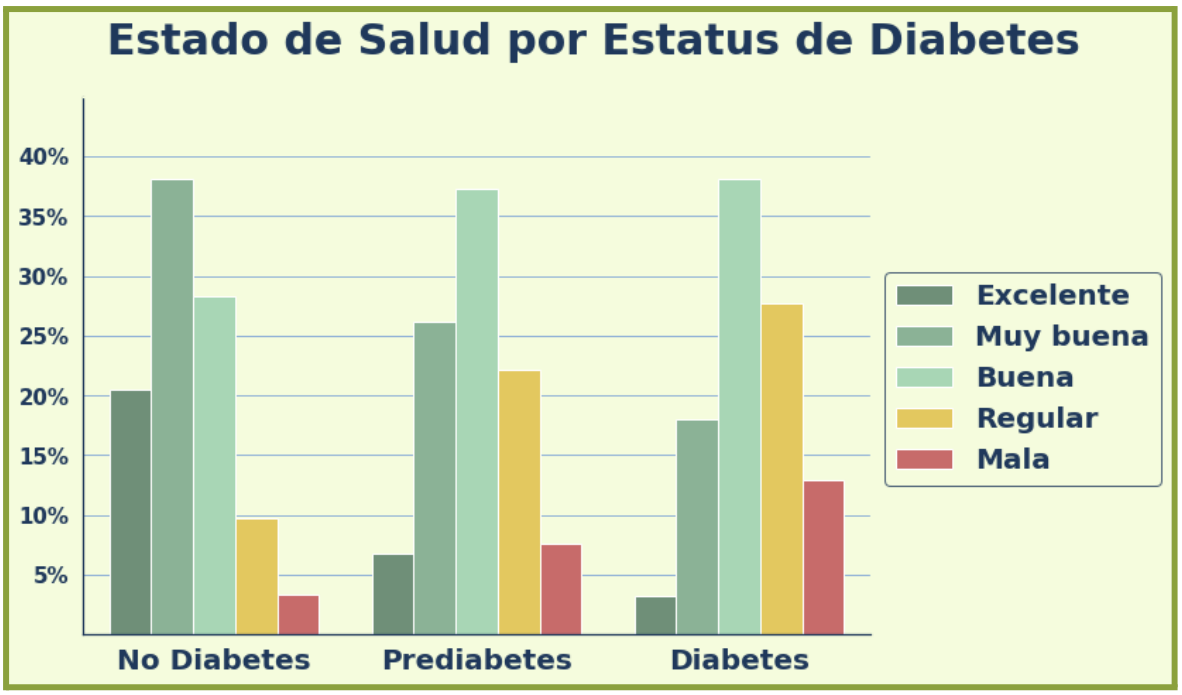


Se puede observar como el porcentaje con **bajo peso**, **peso normal** y **pre-obesidad** **disminuyen** considerablemente en el grupo con **prediabetes** y **diabetes**, **aumentando** notoriamente en estos grupos las **obesidades clase II y III**.

Claramente la **obesidad** puede considerarse un **factor de riesgo** para esta enfermedad.

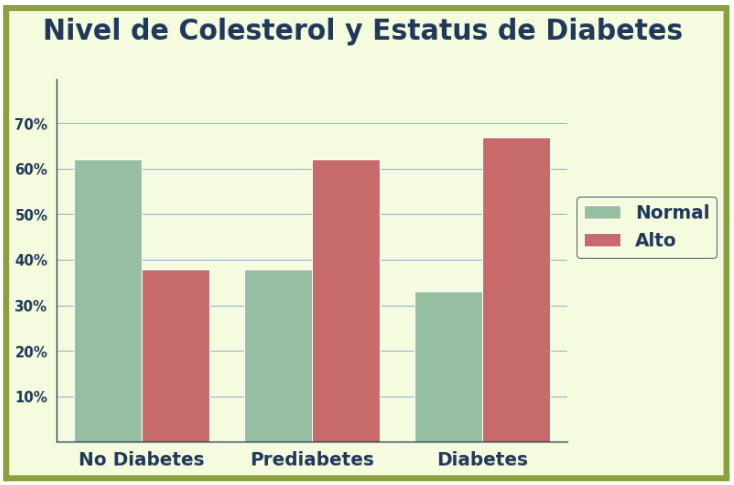
¿Qué información nos proporciona el estado de salud general de la persona?

Del análisis se puede concluir que el porcentaje de personas con un estado de salud **excelente** y **muy buena** **disminuye** considerablemente en el grupo con **prediabetes** y **diabetes**. En el grupo sin diabetes el 21% y 38% poseen una salud excelente y muy buena respectivamente, mientras que en las personas con diabetes es de 3% y 18%. Este decrecimiento se produce en función del **aumento** de las personas que poseen un estado de salud **regular** y **malo**.



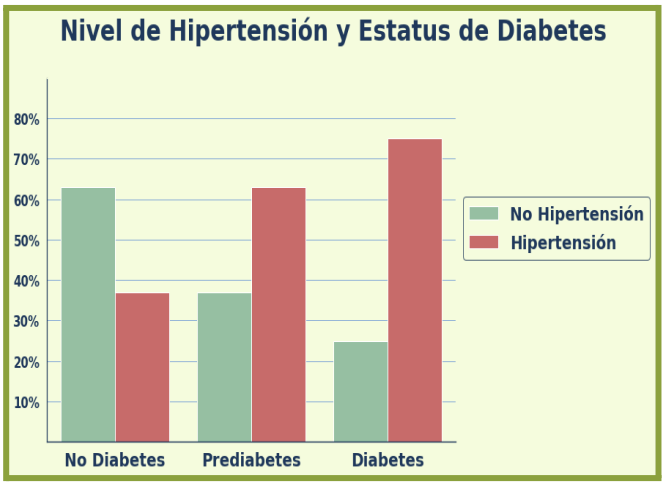
¿Cómo se relaciona la diabetes y el colesterol?

A partir de la visualización se puede observar que, entre las personas que no padecen diabetes, las que tienen niveles **normales de colesterol** representan el **62%** de la población mientras que las que poseen **colesterol alto** representan el **38%**. Mientras que para las poblaciones con **prediabetes** y **diabetes** estos porcentajes se intercambian, con prediabetes el **62%** tienen **colesterol alto** y con diabetes el **67%**. Estos datos avalan el hecho de considerar al **colesterol alto** como un **factor de riesgo** para esta enfermedad.

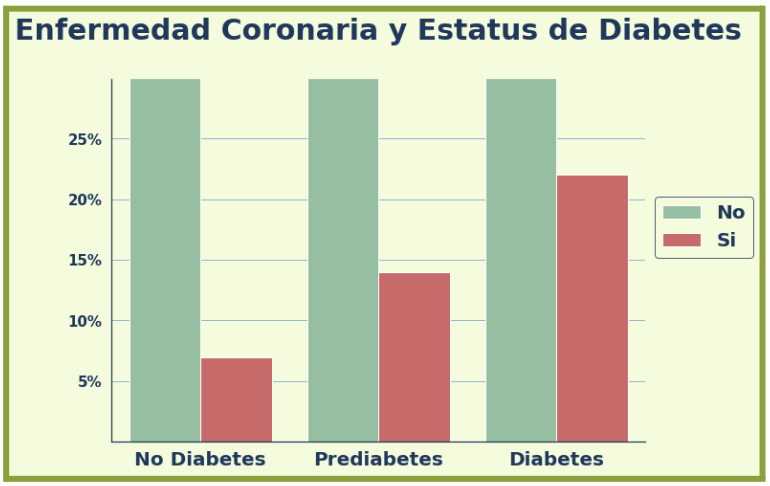


¿Cómo se relacionan la diabetes y la hipertensión?

A partir de la visualización se puede observar que, entre las personas que no padecen diabetes, las que tienen niveles **normales de tensión** representan el **63%** de la población mientras que las que poseen **hipertensión** representan el **37%**. Mientras que para las poblaciones con **prediabetes** y **diabetes** estos porcentajes se intercambian, con prediabetes el **63%** tienen **hipertensión** y con diabetes el **75%**. Estos datos avalan el hecho de considerar a la **hipertensión** como un **factor de riesgo** para esta enfermedad.

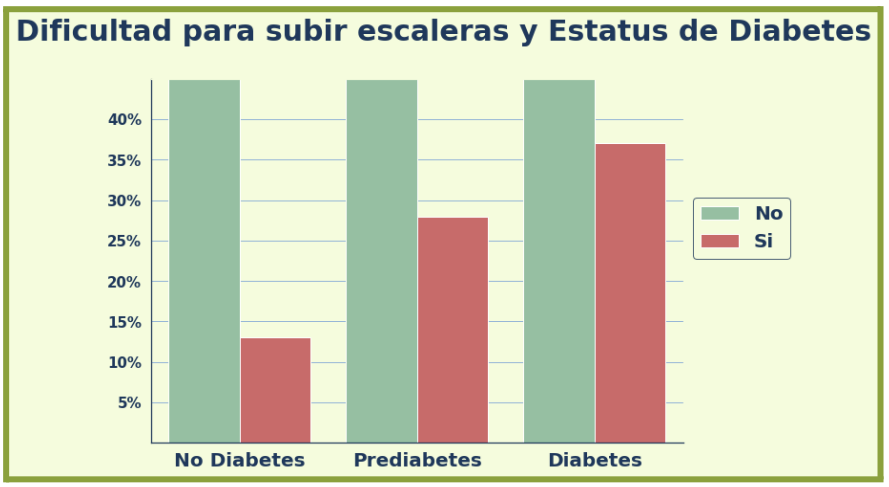


Enfermedad Coronaria / Ataque al miocardio



Se observa que la incidencia de personas con alguna enfermedad coronaria o que ha sufrido un ataque al miocardio se duplica en personas con prediabetes, de 7% al 14%, y se triplica en personas con diabetes alcanzando el 22%.

Dificultad para subir escaleras

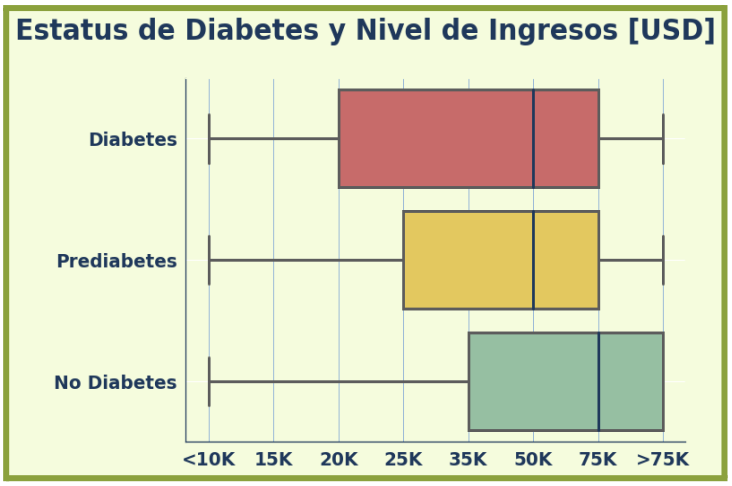


Se observa que la incidencia de personas con problemas serios para subir escaleras se duplica en personas con prediabetes del 13% al 28%, y casi se triplica en personas con diabetes alcanzando el 38%.

Factores Socio-Económicos

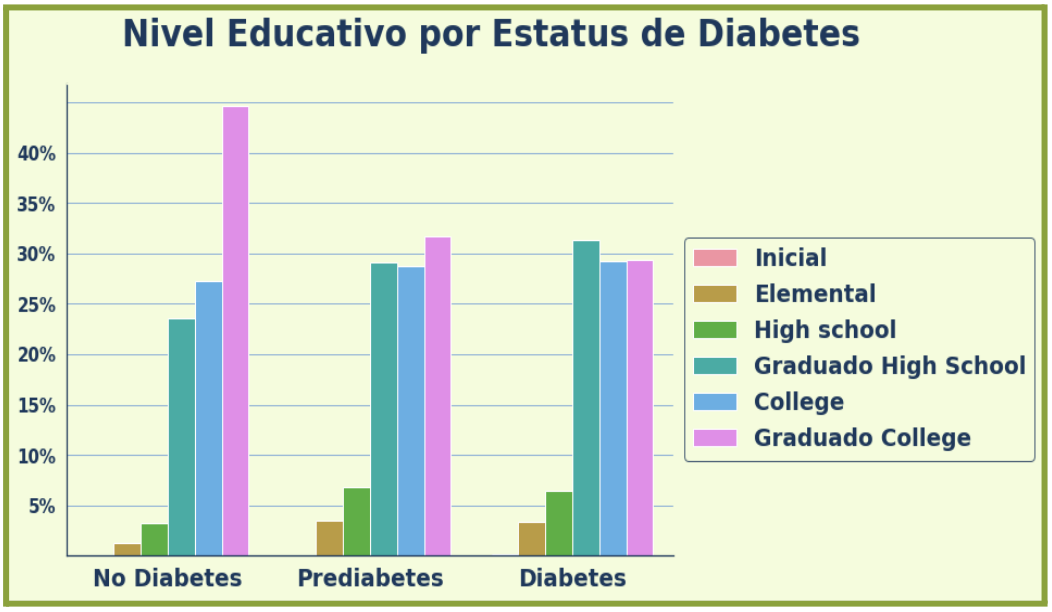
¿Cómo se distribuyen los ingresos según estatus de diabetes?

Se observa que del total de las muestras en promedio viven en hogares donde los ingresos anuales son iguales o menores a $50.000 dólares. De los tres estatus de diabetes, las personas con diabetes y prediabetes poseen menores



ingresos que las personas sin diabetes. Lo que indica que es una patología que tiene mayor incidencia en poblaciones vulnerables. También se observa un gran sesgo a la izquierda en la población que sufre diabetes, el 25% de las personas viven en hogares con ingresos menores a 20.000 USD.

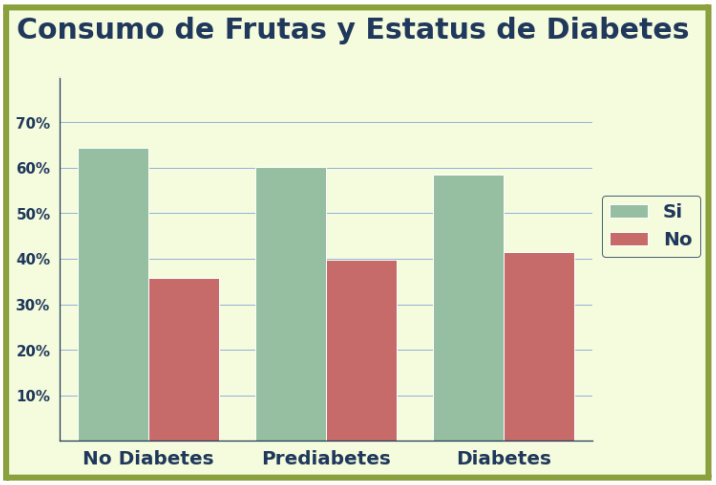
Nivel Educativo



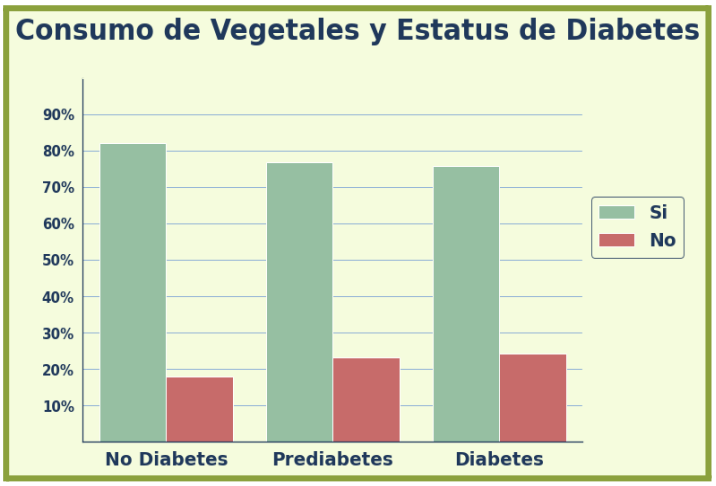
No se aprecia incidencia de Diabetes en los distintos Niveles Educativos más que la baja en el porcentaje de graduados del College, de 42% en las personas sin diabetes al 32% y 31% en personas con prediabetes y diabetes respectivamente.

¿Qué hábitos se pueden considerar preventores de la enfermedad?

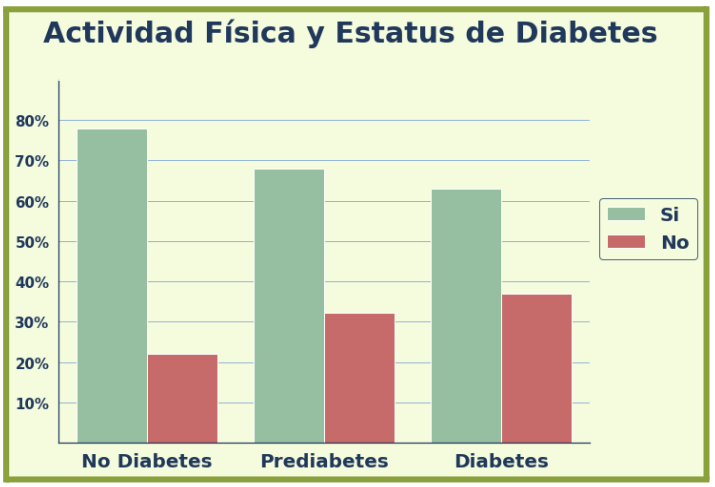
Análisis para Frutas



Análisis para Vegetales



Análisis para Actividad Física



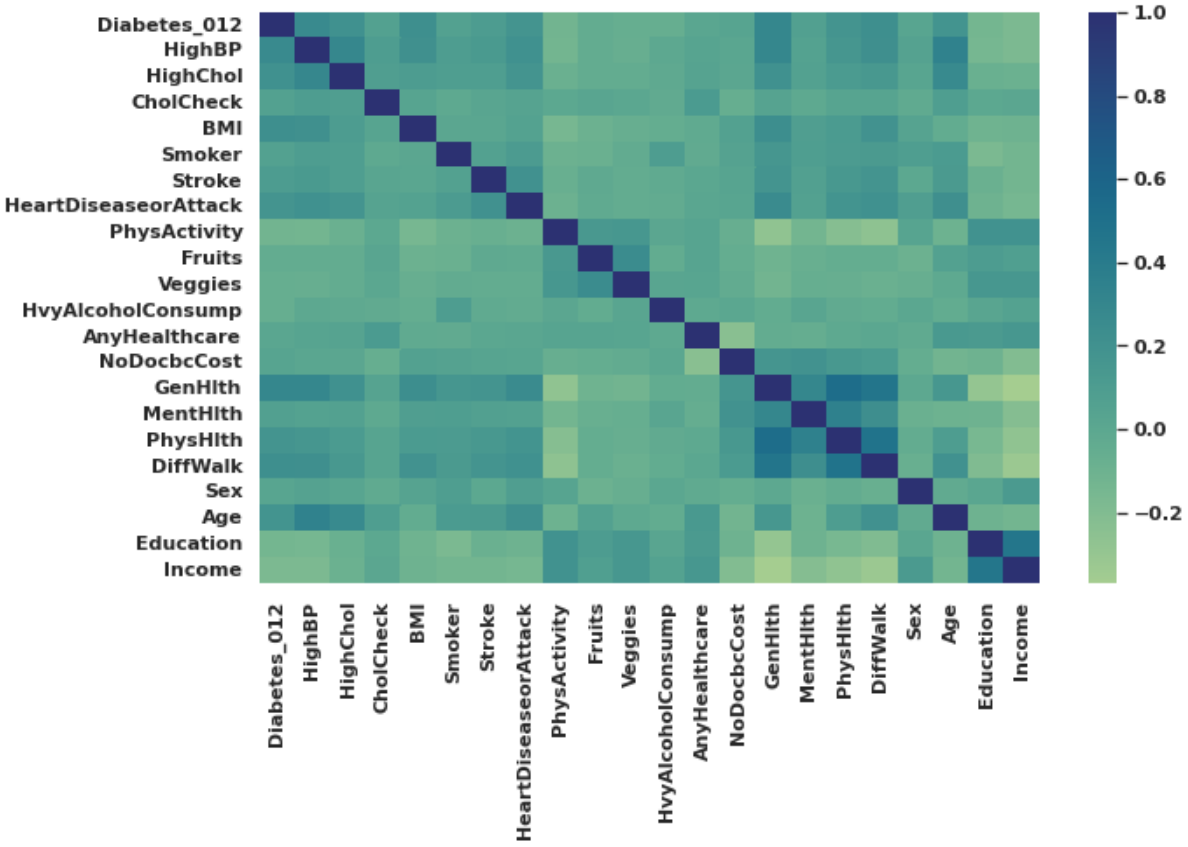
De los 3 factores saludables ninguno obtuvo una gran significancia en el estatus de diabetes. El factor en el que mayor diferencia se obtuvo entre las personas con diabetes y sin fue la actividad física, el 78% de las personas sin diabetes realiza alguna, contra un 63% de las personas con. Para los otros dos factores, consumo de frutas y vegetales, un 64% y 82% de los no diabéticos consumen respectivamente, mientras que sólo un 59% y 76% de los diabéticos consumen.

Selección de Algoritmo Apropiado

Análisis de correlaciones

Matriz de correlaciones

Analizamos las correlaciones entre las distintas variables del dataset y estos fueron los resultados:



De la matriz de correlaciones podemos observar que hay una fuerte correlación positiva entre las variables **GenHlth** y **PhysHlth**, y una correlación negativa fuerte entre **Income** y **GenHlth** (a mayores niveles de ingresos mejores condiciones de salud general).

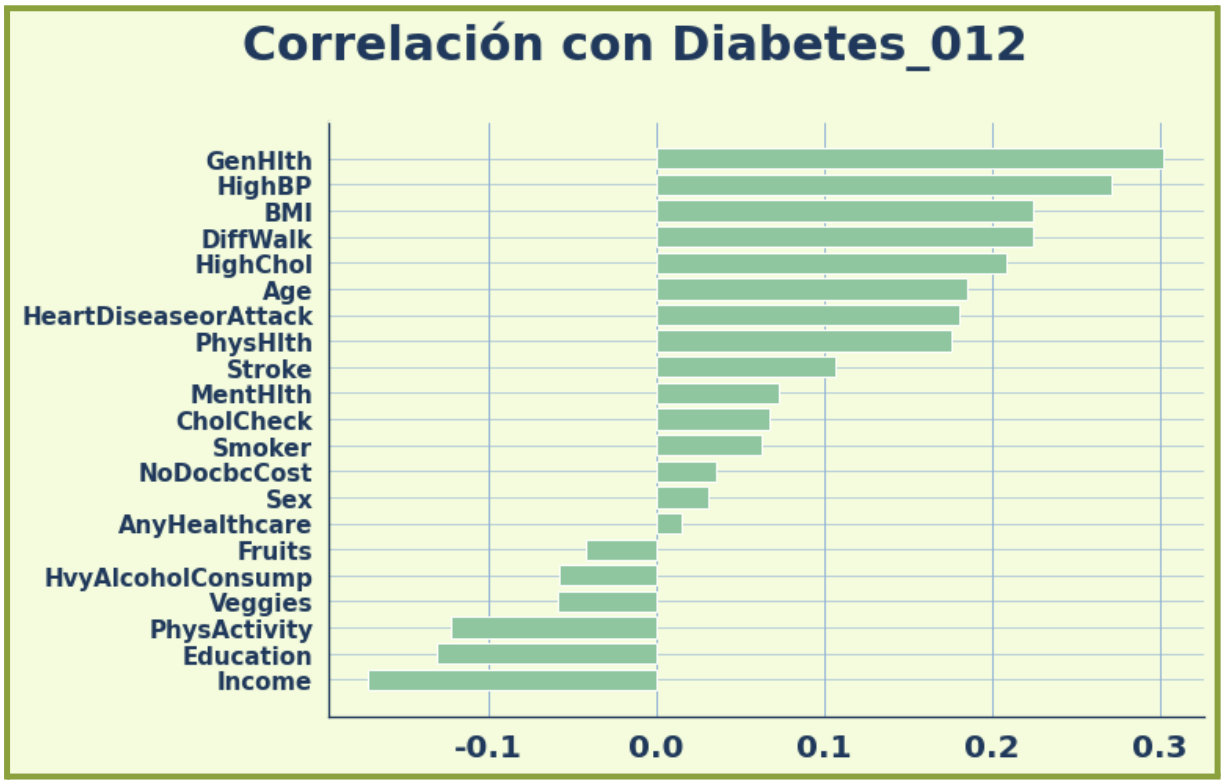
En la siguiente tabla se pueden ver el TOP 10 de las mayores correlaciones entre variables.

Variable 1	Variable 2	Correlación
GenHlth	PhysHlth	0.524364
PhysHlth	DiffWalk	0.478417
GenHlth	DiffWalk	0.456920
Education	Income	0.449106
GenHlth	Income	0.370014
MentHlth	PhysHlth	0.353619

HighBP	Age	0.344452
DiffWalk	Income	0.320124
Diabetes_012	GenHlth	0.302587
GenHlth	MentHlth	0.301674

Correlaciones con la variable objetivo

Considerando ahora las correlaciones de las distintas variables con la variable objetivo, estado de diabetes (Diabetes_012), los resultados son los siguientes.



A partir de la gráfica de correlación con la variable podemos observar que:

- GenHlth, HighBP, BMI, DiffWalk, HighChol, Age, HeartDiseaseorAttack, PhysHlth, Income, Education, PhysActivity, Stroke, Smoker, Veggies y HvyAlcoholconsump tienen una correlación significativa.

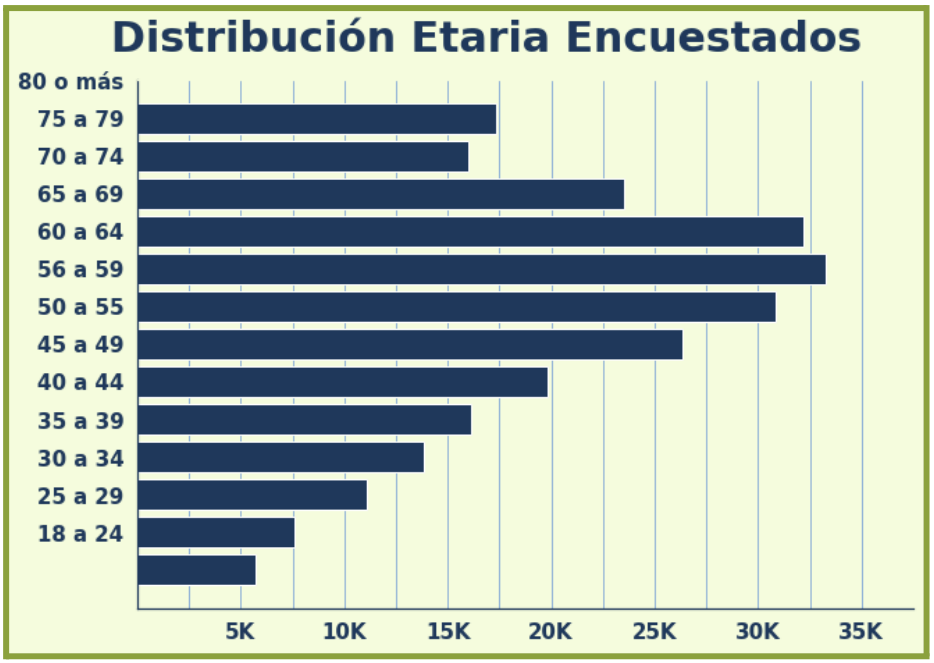
- Fruits, AnyHealthcare, NoDocbcCost y Sex tienen la menor correlación.

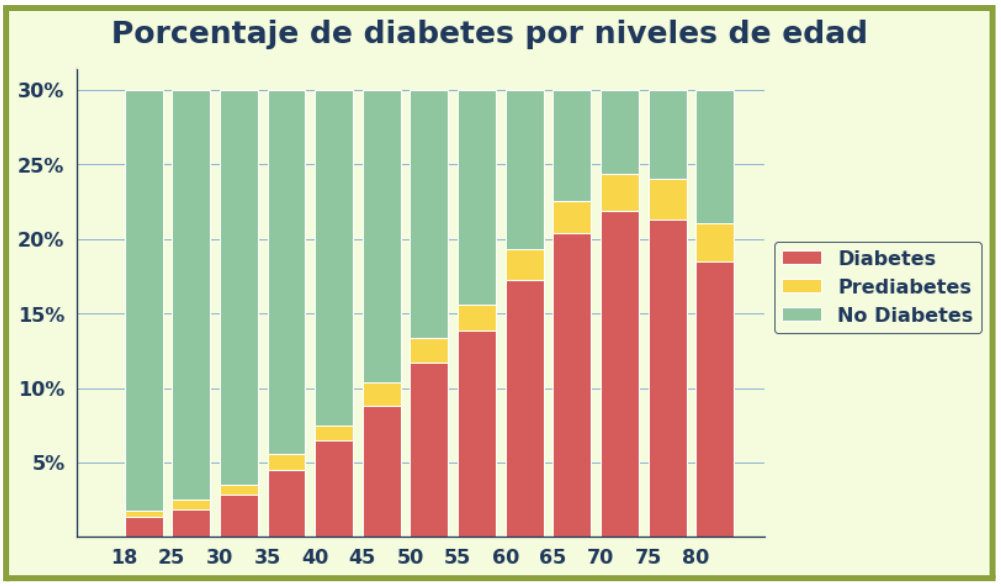
Selección de características

Dos técnicas que se implementaron con buenos resultados fueron las siguientes:

Segmentación de la muestra por edades y balanceo

Debido a la cantidad de muestras de los distintos rangos etarios, y a que el porcentaje de diabéticos aumenta en los rangos etarios mayores, se decidió realizar una división del DataSet.





Se separa la muestra en 3 grandes grupos etarios: mayores de 60, intermedios entre 60 y 40 y menores a 40 años.

Además se balancearon la cantidad de muestras de nuestra variable target mediante el algoritmo **SMOTE**, con el fin de que a la hora de entrenar se encuentre la misma cantidad de diabéticos, pre diabéticos y personas sin la patología.

Algoritmos de selección

Se probaron distintos métodos de selección de características cómo Sequential Feature Selection, devuelve todas las características como las mejores cuando evalúa exactitud (accuracy), o sea que ninguna es descartada por el algoritmo. Luego se implementaron algoritmos como chi cuadrado, PCA, t-SNE y MCA pero se obtuvieron resultados relevantes. Con el que sí se obtuvieron resultados favorables fue con el método **SelectKBest** los cuales se mostrarán a continuación.

Si elegimos las variables con **score > 0,08** trabajamos con **13 variables**

Variable	Score
BMI	0.226032
Age	0.222771

GenHlth	0.210261
Income	0.178846
HighBP	0.164013
Education	0.154055
HighChol	0.143348
Smoker	0.101654
Fruits	0.100500
PhysActivity	0.097725
PhysHlth	0.095442
Sex	0.092251
Veggies	0.089979

Evaluación de Modelos

Se probaron los modelos, DecisionTree, Random Forest, LightGBM, XGBoost, KNN y Bagging Classifier en el dataset en general y dividiendo las muestras en 3 rangos etarios: menores, intermedios y mayores, como se mencionó anteriormente y en todos los casos las muestras se encontraban balanceadas.

Los resultados fueron los siguientes:

General

Modelo/Métrica	Accuracy	Precisión	Recall	F1 Score	Tiempo de inferencia
Random Forest	0.929	0.931	0.929	0.929	9.821
Bagging Classifier	0.90	0.906	0.905	0.904	1.22

KNN	0.873	0.888	0.873	0.867	1357.61
Decision Tree	0.864	0.863	0.863	0.863	0.113
LightGBM	0.777	0.772	0.777	0.771	4.612
XGBoost	0.714	0.709	0.714	0.711	2.766

Las principales métricas para los algoritmos de clasificación resultan como mejor predictor, para este caso, a **Random Forest**. También se observa que el segundo algoritmo mejor posicionado es el **Bagging Classifier** y que incluso tiene un menor tiempo de inferencia, siendo **9 veces más rápido** que el Random Forest. Por otra parte, se observa que KNN es el tercero mejor posicionado en cuanto a métricas, fue muy rápido en tiempo de entrenamiento pero extremadamente lento en predicción.

Menores

Modelo/Métrica	Accuracy	Precisión	Recall	F1 Score	Tiempo de inferencia
Random Forest	0.987	0.987	0.987	0.987	0.877
Bagging Classifier	0.98	0.98	0.98	0.98	0.115
KNN	0.929	0.935	0.929	0.926	39.086
Decision Tree	0.965	0.965	0.965	0.965	0.018
LightGBM	0.962	0.963	0.962	0.962	0.705
XGBoost	0.813	0.809	0.812	0.81	0.381

Se obtienen mejores métricas en todos los algoritmos comparado al dataset sin agrupar por edades. Si analizamos detalladamente el Recall, Random Forest, Bagging Classifier y Decision Tree tuvieron los mejores resultados siendo el de RF 0.987.

Intermedios

Modelo/Métrica	Accuracy	Precisión	Recall	F1 Score	Tiempo de inferencia
Random Forest	0.953	0.954	0.952	0.953	1.222
Bagging Classifier	0.932	0.932	0.931	0.931	0.145
KNN	0.886	0.9	0.886	0.88	48.894
Decision Tree	0.896	0.896	0.896	0.896	0.015
LightGBM	0.875	0.874	0.874	0.872	0.761
XGBoost	0.771	0.765	0.77	0.765	0.427

Se obtienen mejores métricas en todos los algoritmos comparado al dataset sin agrupar por edades. Si analizamos detalladamente el Recall, Random Forest, Bagging Classifier y Decision Tree tuvieron los mejores resultados siendo el de RF 0.952.

Mayores

Modelo/Métrica	Accuracy	Precisión	Recall	F1 Score	Tiempo de inferencia
Random Forest	0.892	0.895	0.892	0.891	4.175
Bagging Classifier	0.857	0.858	0.857	0.855	0.428
KNN	0.835	0.85	0.835	0.824	249.71
Decision Tree	0.809	0.807	0.809	0.808	0.061
LightGBM	0.771	0.766	0.771	0.762	1.929
XGBoost	0.705	0.698	0.705	0.7	0.953

Se obtienen peores métricas en todos los algoritmos comparado al dataset sin agrupar por edades. Si analizamos detalladamente el Recall, Random Forest, Bagging Classifier y Decision Tree tuvieron lo mejores resultados siendo el de RF 0.891849

En las edades intermedias y menores las métricas resultaron ser mejores. Es recomendable dividir el dataset por edades.

Tuning de hiperparámetros

Se intentó, sin éxito, mejorar las métricas mediante el ajuste de los hiperparámetros utilizando **GridSearchCV** para los modelos de Random Forest y Decision Tree. No se obtienen mejoras con respecto a las métricas del modelo por defecto.

Conclusiones

- La cantidad de encuestados por **rango etario** son **muy dispares**.
- El rango etario con mayor porcentaje de personas que sufren diabetes es el de **70 a 74 años** con un **24,41%**.
- La **obesidad** puede considerarse un **factor de riesgo** para esta enfermedad.
- Las personas con algún tipo de **diabetes** poseen, mayoritariamente, **niveles altos** de **colesterol** e **hipertensión**.
- De los **3 hábitos** saludables analizados **ninguno** posee una **gran significancia** en el estatus de diabetes.
- Es una patología que tiene **mayor incidencia** en **poblaciones vulnerables**.
- En base a los resultados obtenidos, los factores **estado nutricional**, **colesterol alto** e **hipertensión** son **factores de riesgo** de la enfermedad.
- **Ningún hábito saludable** de la encuesta trajo aparejado una gran significancia en el estatus de diabetes. por lo tanto **no** es un **factor preventivo** de la enfermedad.

- **Segmentar la población** por **rangos etarios** a la hora de implementar los modelos predictivos **mejora las métricas**.

- De los modelos empleados, **Radom Forest** fue el que mejores resultados otorgó.

- Si se pondera el **tiempo de entrenamiento**, el mejor modelo es el **Bagging Classifier** siendo un **88% más rápido** que **Random Forest**.

- El modelo predice mejor a **pre diabéticos**, seguido de los no diabéticos, y por último a los diabéticos.