



**UNIVERSIDAD TECNOLÓGICA DE PANAMÁ
FACULTAD DE INGENIERÍAS DE SISTEMAS
COMPUTACIONALES**



Licenciatura en Ingeniería de Sistemas y Computación

Probabilidad Aplicada a la TIC

**Relación entre Los Hábitos Alimenticios, la Actividad Física y Los Niveles De
Obesidad: Un Análisis Predictivo Con Machine Learning**

Josip De Lora 8-1028-836

Mateo Del Giudice E-229-194

Rene Rodriguez 8-1034-1629

Ian Nieto 8-1033-124

Dhruv Sakhrani 8-1023-30

Profesor Juan Marcos Castillo, PhD

Grupo 1IL124

Año 2025

Nombre de la investigación:

Relación entre los hábitos alimenticios, la actividad física y los niveles de obesidad: un análisis predictivo con machine learning

Descripción de la base de datos a utilizar:

La base de datos contiene registros de personas con distintas características relacionadas con la salud y el estilo de vida. Incluye variables como:

Nombre del campo	Tipo de dato	Descripción en español
Gender	Categorico	Género
Age	Continuo	Edad
Height	Continuo	Estatura
Weight	Continuo	Peso
family_history_with_overweight	Binario	¿Algún familiar sufre o ha sufrido de sobrepeso?
FAVC	Binario	¿Consumes alimentos altos en calorías con frecuencia?
FCVC	Entero	¿Sueles incluir vegetales en tus comidas?
NCP	Continuo	¿Cuántas comidas principales realizas al día?
CAEC	Categorico	¿Consumes alimentos entre comidas?
SMOKE	Binario	¿Fumas?
CH2O	Continuo	¿Cuánta agua bebes diariamente?
SCC	Binario	¿Controlas las calorías que consumes diariamente?
FAF	Continuo	¿Con qué frecuencia realizas actividad física?
TUE	Entero	¿Cuánto tiempo usas dispositivos tecnológicos como celulares, videojuegos, TV, etc.?
CALC	Categorico	¿Con qué frecuencia consumes alcohol?

MTRANS	Categorico	¿Qué medio de transporte utilizas con mayor frecuencia?
NObeyesdad	Categorico	Nivel de obesidad

Esta base es ampliamente utilizada para modelos de clasificación y regresión en temas de salud pública.

Fuente:

<https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

Introducción:

El análisis de datos se ha convertido en una herramienta clave para entender el comportamiento de distintos fenómenos en diversas áreas. En este caso, se trabaja con una base de datos que incluye información sobre características personales, hábitos alimenticios y actividad física de diferentes individuos. El objetivo del análisis es explorar cómo se relacionan estas variables entre sí, utilizando herramientas como tablas dinámicas, gráficas y medidas estadísticas.

A través del análisis descriptivo se busca resumir y visualizar la información de forma clara, mientras que mediante modelos de regresión se pretende identificar patrones que puedan ser útiles para realizar predicciones. Este enfoque permite transformar los datos en conocimiento que puede apoyar la toma de decisiones o estudios más complejos en el futuro.

Justificación:

La obesidad se ha convertido en uno de los principales problemas de salud pública a nivel mundial, afectando tanto a adultos como a jóvenes. La obesidad mundial en adultos se ha más que duplicado desde 1990, y la obesidad en adolescentes se ha cuadruplicado. Sus consecuencias no se limitan únicamente al aumento de peso, sino que también incrementan el riesgo de desarrollar enfermedades crónicas como diabetes tipo 2, hipertensión, problemas cardiovasculares y ciertos tipos de cáncer. Por esta razón, resulta fundamental contar con herramientas que permitan identificar oportunamente los factores que contribuyen a su aparición.

Este proyecto se justifica en la necesidad de analizar cómo los hábitos alimenticios y los comportamientos relacionados con la actividad física influyen directamente en el desarrollo de la obesidad. Al aplicar modelos estocásticos, se busca generar un enfoque predictivo que facilite la prevención de la obesidad.

De esta manera, el estudio aportará información valiosa para disminuir la prevalencia de obesidad y mejorar la calidad de vida de la población.

Antecedentes:

Numerosos estudios han demostrado que los hábitos de alimentación y los comportamientos sedentarios están correlacionados con un incremento en el índice de masa corporal (IMC), especialmente en poblaciones jóvenes y adolescentes. Estas relaciones han sido modeladas mediante técnicas estadísticas y de machine learning para identificar patrones de riesgo y clasificar niveles de obesidad según variables como actividad física y hábitos alimenticios.

Además, investigaciones que usan modelos estocásticos han mostrado que combinaciones de cadenas de Markov permiten modelar con precisión las transiciones entre categorías de peso (normal, sobrepeso y obesidad) en el tiempo, ofreciendo una visión robusta de la evolución de la obesidad poblacional. En el campo de técnicas predictivas, se han aplicado algoritmos como regresión LASSO, Random Forest, Gradient Boosting y máquinas de gradiente estocástico alcanzando niveles de precisión aceptables para clasificar el estado de obesidad según variables sociodemográficas, hábitos alimenticios y actividad.

En resumen, existe una base sólida de investigaciones que respaldan la modelación de la obesidad mediante enfoques vinculados a los hábitos de alimentación y actividad física, utilizando tanto modelos descriptivos como estocásticos y predictivos basados en datos reales.

Definición del Problema:

La obesidad es uno de los problemas de salud pública más importantes en el mundo. Factores como la alimentación inadecuada, el sedentarismo, y malos hábitos de vida son determinantes clave en su desarrollo. Entender cómo estos factores interactúan y predicen el nivel de obesidad permite generar mejores políticas de prevención y tratamiento. Esta investigación busca explorar dichos factores usando análisis de datos y modelos predictivos.

6- Análisis con Modelos Estocásticos

a- Determinación de la base de datos:

Seleccionamos nuestra base de datos principalmente porque tiene un conjunto de variables relevante y bien definido para estudiar la obesidad.

Permite aplicar técnicas de análisis estadístico y de machine learning.

Está disponible en formato abierto, lo que permite su uso sin restricciones éticas o legales.

Representa una problemática actual de salud global con potencial impacto social.

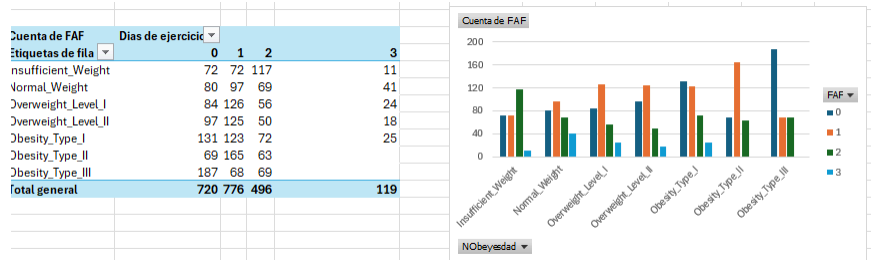
b- Pre-procesamiento y limpieza:

Antes de analizar los datos, es importante limpiarlos y prepararlos bien. Esto se llama pre-procesamiento y limpieza. Se trata de revisar que no haya errores, valores vacíos o datos repetidos, y de poner todo en un formato que se entienda mejor. También se pueden transformar respuestas en números para poder hacer cálculos. Hacer esta parte ayuda a que el análisis funcione bien y los resultados sean más confiables.

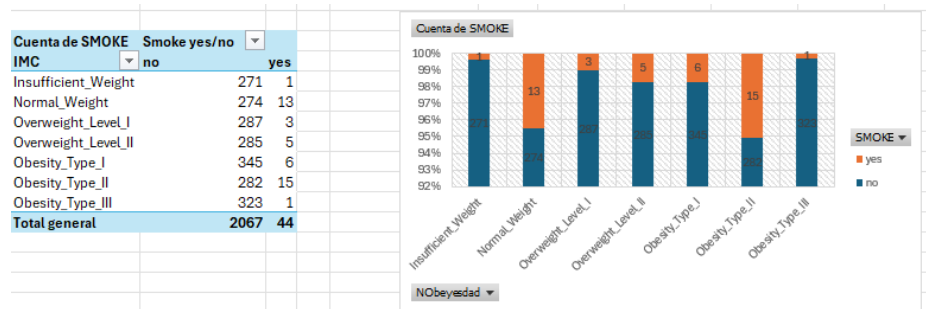
c- Análisis descriptivo

Es una forma de resumir y entender los datos que tienes. Sirve para ver cómo se comportan las variables, qué valores se repiten más, cuáles son los promedios, y si hay alguna relación entre ellas. También se usan gráficas, tablas y porcentajes para mostrar la información de manera clara.

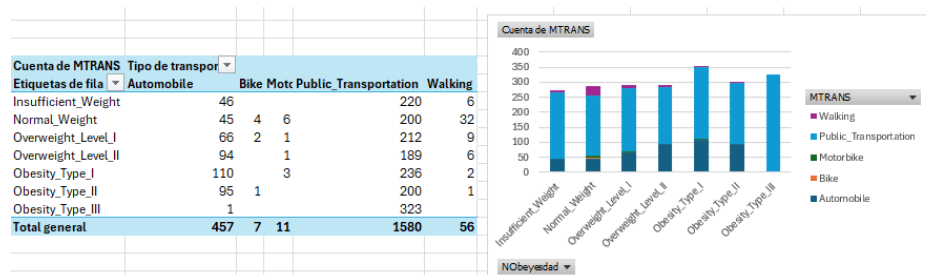
IMC con respecto a cuántos días de ejercicios realizan. Se observa que la mayor cantidad de personas que no realizan ejercicio son las personas con obesidad tipo 3.



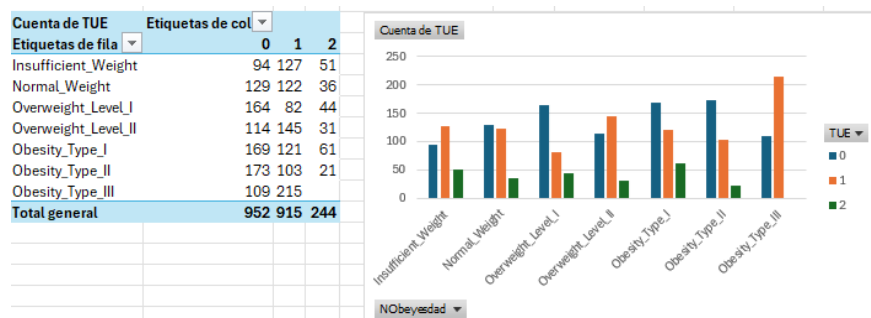
IMC con respecto al hábito de fumar. Se observa que las personas que más fuman en la base de datos son los de grado de obesidad tipo 2.



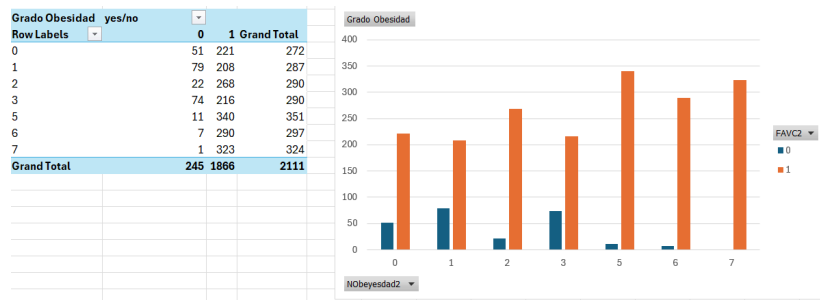
IMC con respecto a de qué manera se transporta la persona: caminando, transporte público, motocicleta, bicicleta, carro. Se observa que casi ninguna persona suele caminar en su día a día para transportarse.



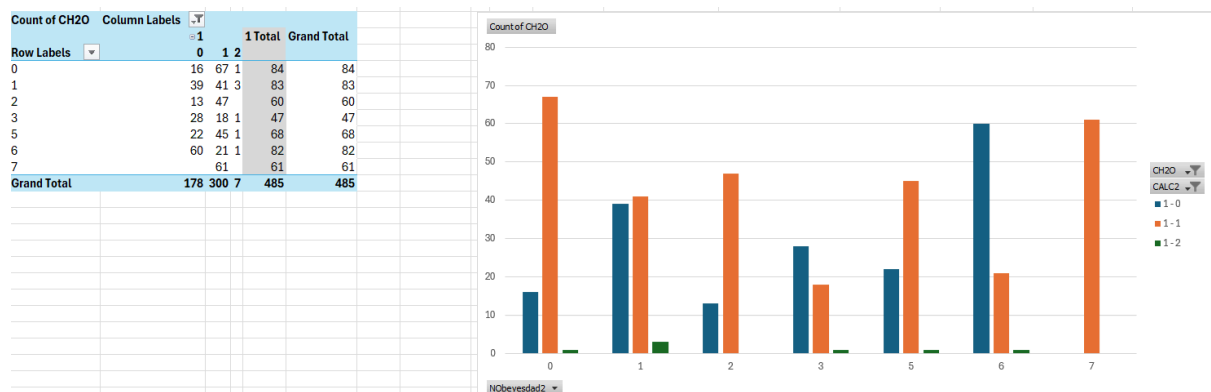
IMC con respecto al tiempo de uso de pantalla. Se observa el mayor índice de uso en el último grado de obesidad.



Se relaciona el grado de obesidad con el campo de “Comes comida con altas calorías”, donde se puede visualizar que la mayoría de personas que consumen comidas con altas calorías son las personas en el grado de obesidad.

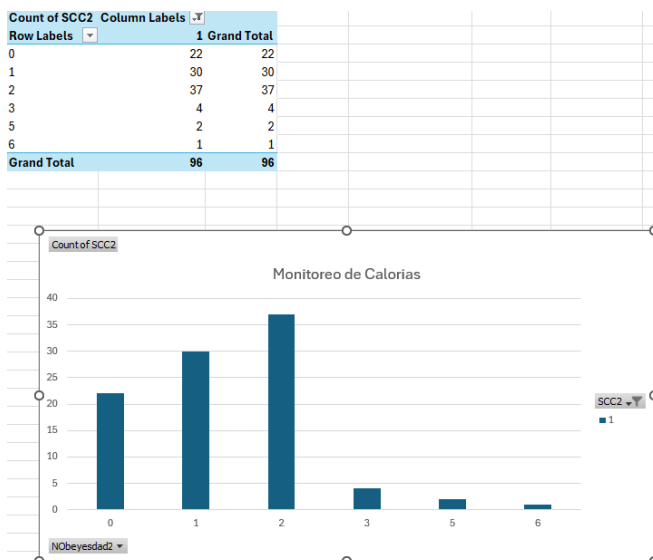


Se relaciona el consumo del agua con el nivel de consumo de litro de agua diario. Se observa una tendencia donde en la gran mayoría de respuestas seleccionan 1 litro de agua diario.

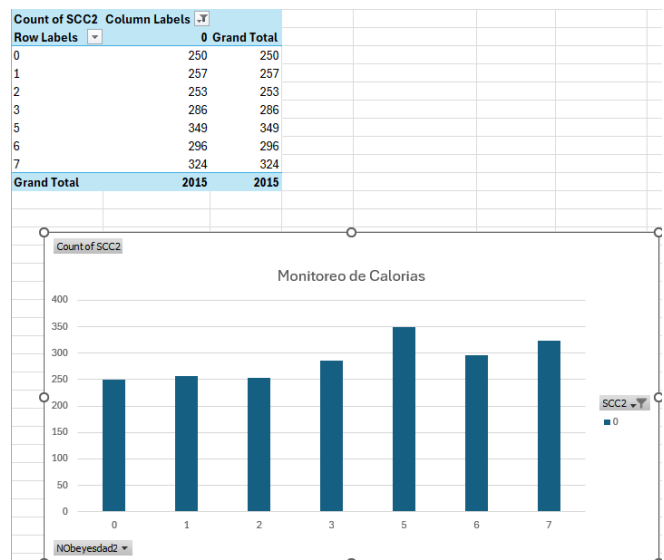


Se relaciona el monitoreo de calorías con el índice de obesidad observado. Se observa una tendencia hacia la regulación del índice de obesidad en personas que controlan la cantidad de calorías se consumen a diario, sin embargo, la falta de datos para el caso del monitoreo de calorías vuelve posible la interacción de otros factores para afectar al índice de obesidad observado.

Monitoreando Calorías



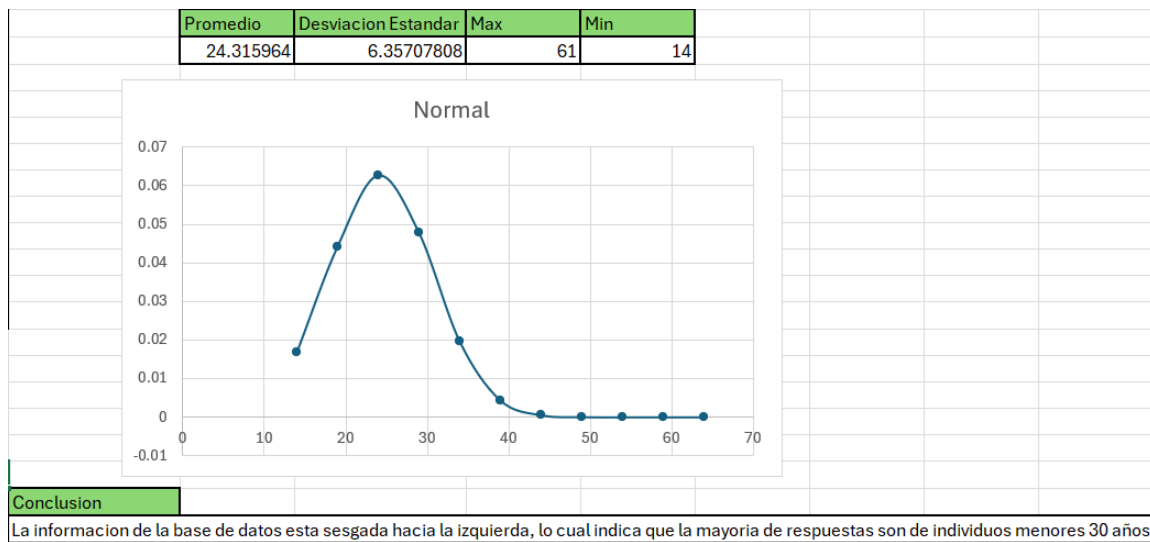
No Monitoreando Calorías



Al ver por encima de la base de datos es evidente que los datos están sesgados, decidimos comprobar a qué apuntaba este promedio y visualizarlo en una campana de Gauss.

Se utilizó el campo de “Edad”, este es de tipo numérico; se obtuvo su promedio, desviación estándar, valor máximo y mínimo. Se crea la columna de la Distribución Normal y con respecto a los intervalos se crea la gráfica.

Intervalos	Normal	Cantidad
14	0.01682028	303
19	0.04423913	936
24	0.06267814	453
29	0.04783676	211
34	0.01966725	112
39	0.00435574	73
44	0.00051966	13
49	3.3397E-05	3
54	1.1562E-06	6
59	2.1562E-08	1
64	2.1662E-10	0



Se observa en la gráfica que tiene un sesgo a la izquierda, lo cual indica que no es simétrica, donde están los datos menores a 35 años.

Se hizo una cadena de Markov utilizando el campo de Historial Familiar definido en binario(yes/no). Se buscaba encontrar la probabilidad de si una persona tiene historial familiar de obesidad(se refiere a familiares cercanos), la probabilidad de que este tambien sea obeso.

Historial obesidad	Insufficient_Weight	Normal_Weight	Overweight_Level_I	Overweight_Level_II	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Total general
no	146	132	81	18	7	1		385
yes	126	155	209	272	344	296	324	1726
Total general	272	287	290	290	351	297	324	2111

Se calcula la probabilidad de tener obesidad o no, esto se hace con un rango donde la NoObesidad son todos los datos de los grados desde Insufficient hasta OverweightLevel2. Luego se obtiene el conteo de las respuestas en esos rangos.

Probabilidad Obesidad
0.460445287
Probabilidad NoObesidad
0.539554713

Historial	NoObesidad	Obesidad	
yes	762	964	
no	377	8	
Total	1139	972	2111

Se crea una matriz de markov mediante las probabilidades condicionales de si tiene o no obesidad, tendrá historial o no. Se crea un vector inicial con las probabilidades de tener o no Obesidad y este se multiplicará con la matriz, lo cual arroja un vector que nos da las posibles probabilidades de si tiene obesidad o no la siguiente generación.

Entrada			
Obesidad/Histo	Yes	No	
Yes	0.991769547	0.669007902	1.66077745
No	0.008230453	0.330992098	0.33922255
Total	1	1	

Vector probabilidad inicial		Prediccion	
Obesidad	0.460445287	Obesidad	0.81762198
NoObesidad	0.539554713	NoObesidad	0.18237802
		Total	1

e- Selección de Modelos

Se desarrolló un modelo de clasificación utilizando **regresión logística multi clase** en la plataforma **Azure Machine Learning Studio**, con el objetivo de predecir el nivel de obesidad de un individuo a partir de sus características físicas y hábitos personales. La variable objetivo del modelo presenta cinco posibles clases: *Peso Normal*, *Sobrepeso*, *Obesidad I*, *Obesidad II* y *Obesidad III*, lo cual justifica el uso de un enfoque multi clase frente a una clasificación binaria.

1. Preprocesamiento de los Datos

El conjunto de datos original incluía variables relacionadas con características personales (edad, género, altura, peso), hábitos alimenticios (frecuencia de comidas, consumo de alimentos calóricos, consumo de agua) y actividad física. Antes de entrenar el modelo, se llevó a cabo un proceso de limpieza utilizando **Microsoft Excel**, donde se filtraron registros con valores faltantes o inconsistentes. Además, se eliminaron duplicados y se unificaron formatos para evitar errores de interpretación.

Para permitir el entrenamiento del modelo de clasificación, se aplicaron técnicas de **codificación de variables categóricas**. En particular, se utilizó **Label Encoding** para convertir las respuestas cualitativas (como “Sí/No” o “Frecuente/Ocasional”) en valores numéricos discretos. Este mismo enfoque fue replicado dentro del formulario interactivo para asegurar la compatibilidad con el modelo desplegado localmente.

2. Entrenamiento y Evaluación del Modelo

El modelo fue entrenado dentro de la plataforma **Azure Machine Learning Studio** para una **tarea de clasificación multiclase**, orientada a predecir la categoría de obesidad de una persona a partir de sus características físicas y hábitos. Se seleccionó como **métrica principal el valor ponderado del AUC (Área Bajo la Curva ROC)**, por su capacidad para evaluar el rendimiento del modelo frente a múltiples clases, especialmente cuando el dataset presenta cierto grado de desbalanceo.

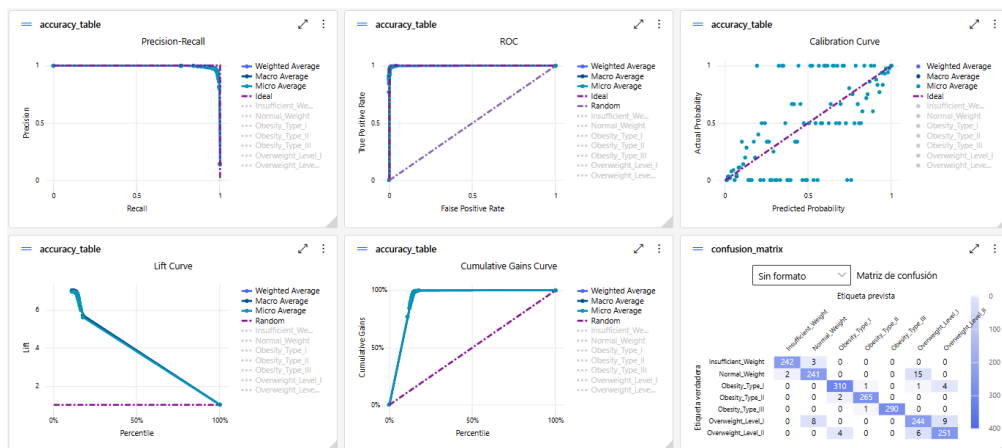
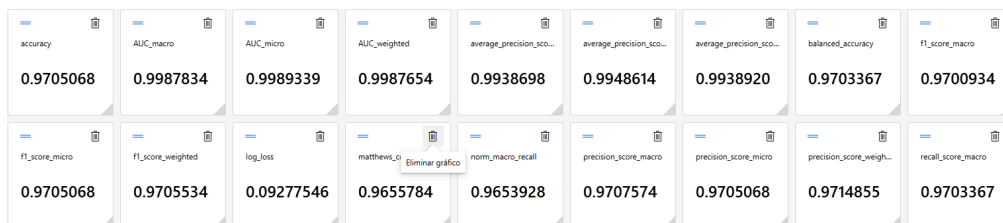
Se aplicó un proceso de **validación cruzada de 10 particiones (10-fold cross-validation)**, lo que permitió estimar la generalización del modelo de forma robusta, reduciendo el riesgo de sobreajuste.

Durante la exploración automática de modelos en Azure, se habilitó la opción de **explicación del mejor modelo**, lo cual proporcionó interpretabilidad sobre las variables más influyentes en el resultado. Se restringió el espacio de búsqueda eliminando modelos de aprendizaje profundo, como TensorFlowLinearClassifier y

TensorFlowDNN, ya que no se ajustaban al objetivo del estudio ni a los recursos disponibles.

El modelo seleccionado por la plataforma fue la combinación de **MaxAbsScaler con XGBoostClassifier**, el cual ofreció un rendimiento del 90% en predicciones a la hora de probarlo y AUC. MaxAbsScaler se encargó de escalar las variables sin alterar su distribución, lo cual es especialmente útil para algoritmos sensibles a la magnitud de los datos. XGBoostClassifier, por su parte, es un potente modelo basado en árboles de decisión que optimiza gradualmente la clasificación mediante técnicas de boosting.

Además, de manera paralela al experimento automatizado, se entrenó manualmente un modelo de **regresión logística multiclase**, lo que permitió comparar el rendimiento entre distintos enfoques. Aunque este modelo también mostró resultados satisfactorios, se optó por desplegar la versión basada en XGBoostClassifier por su mejor rendimiento global y capacidad de generalización.








3. Exportación y Reentrenamiento Local

Dado que Azure Machine Learning permite exportar el modelo en formato compatible con herramientas externas, se procedió a descargar el modelo entrenado y replicarlo localmente utilizando **Python**. Esta decisión se tomó con el fin de evitar costos adicionales por el uso continuo de recursos en la nube.

El modelo fue reconstruido utilizando las siguientes bibliotecas:

- **pandas**: para carga y manipulación de datos
- **sklearn.linear_model.LogisticRegression**: para entrenamiento del modelo
- **sklearn.model_selection.train_test_split**: para dividir los datos
- **joblib**: para serializar el modelo en disco (.pkl)

Esta versión local del modelo fue integrada a un **formulario interactivo**, desde el cual los usuarios pueden ingresar sus datos y obtener una predicción inmediata sobre su posible categoría de obesidad.

	label_encoder.pkl	30/7/2025 08:26	Archivo PKL	1 KB
	modelo_clasificacion_limpio.pkl	30/7/2025 08:26	Archivo PKL	784 KB
	ObesityDataSet_raw_and_data_sinthetic.c...	23/6/2023 18:10	Archivo CSV	258 KB
	PredictionTry1.py	30/7/2025 08:48	Python source file	7 KB
	PredictionTry1.pyproj	30/7/2025 07:59	Python Project	2 KB

```
Responde las siguientes preguntas (elige el número correspondiente):
1. Gender (1. Female, 2. Male): 2
2. Age: 19
3. Height (m): 1.7
4. Weight (kg): 70
5. Family history with overweight? (1. Yes, 2. No): 1
6. High caloric food frequently? (1. Yes, 2. No): 1
7. Eat vegetables (1. Never, 2. Sometimes, 3. Always): 3
8. How many main meals do you have daily?
  1. Between 1 and 2
  2. Three
  3. More than three
  Selección: 2
9. Food between meals?
  1. No
  2. Sometimes
  3. Frequently
  4. Always
  Selección: 2
10. Do you smoke? (1. Yes, 2. No): 2
11. How much water do you drink daily?
  1. Less than a liter
  2. Between 1 and 2 L
  3. More than 2 L
  Selección: 3
12. Do you monitor your calorie intake? (1. Yes, 2. No): 1
13. How often do you have physical activity?
  1. I do not have
  2. 1 or 2 days
  3. 2 or 4 days
  4. 4 or 5 days
  Selección: 2
14. Time using technology devices daily?
  1. 0-2h
  2. 3-5h
  3. More than 5h
  Selección: 3
15. Frequency of alcohol consumption?
  1. Never
  2. Sometimes
  3. Frequently
  4. Always
  Selección: 2
16. Primary mode of transport?
  1. Automobile
  2. Motorbike
  3. Bike
  4. Public Transportation
  5. Walking
  Selección: 1
C:\Users\Mateo Del Giudice\AppData\Local\Programs\Python\Python313\Li
warnings.warn(
Prediccion: Normal_Weight
Press any key to continue . . . |
```

4. Despliegue y Uso del Modelo

Aunque Azure Machine Learning permite la creación de **servicios web (APIs)** para consumir el modelo de manera remota, se optó por una solución local por motivos económicos. El modelo re-entrenado en Python fue incorporado a una aplicación offline o ejecutable, accesible por los usuarios sin necesidad de conexión a servicios externos.

Esta herramienta se enfoca en la **prevención y concientización**, proporcionando una predicción accesible y comprensible basada en datos objetivos y hábitos auto-reportados. Si se detecta una categoría de obesidad elevada, la aplicación puede recomendar consultar a un profesional de la salud.

Reporte del Modelo Predictivo de Regresión para el Índice de Masa Corporal (BMI)

1. Preprocesamiento de los Datos

El conjunto de datos original, derivado de "ObesityDataSet", incluía variables relacionadas con características personales (edad, género, altura, peso), hábitos alimenticios (frecuencia de comidas, consumo de alimentos calóricos, consumo de agua) y actividad física.

Antes de proceder con el entrenamiento del modelo de regresión, se llevó a cabo un proceso de limpieza y preparación similar al de estudios previos de clasificación:

- **Limpieza y Consistencia:** Se asumió que se aplicaron filtrados de registros con valores faltantes o inconsistentes, eliminación de duplicados y unificación de formatos utilizando herramientas como Microsoft Excel, tal como se describió en el caso de clasificación.
- **Codificación de Variables Categóricas:** Las respuestas cualitativas (como "Sí/No", "Frecuente/Ocasional" o categorías de transporte) se convirtieron a valores numéricos discretos utilizando Label Encoding. Esto es crucial para que los algoritmos de aprendizaje automático puedan procesar estas variables. Se verificó que las columnas como family_history_with_overweight2, FAVC2, CAEC2, SMOKE2, SCC2, CALC2, MTRANS2 ya estaban codificadas numéricamente. La variable Gender (Male/Female) también fue codificada numéricamente (ej. 0 para Female, 1 para Male).
- **Creación de la Variable Objetivo (BMI):** Para esta tarea de regresión, se definió el Índice de Masa Corporal (BMI) como la variable objetivo continua. El BMI se calculó a partir de las variables Weight (peso en kg) y Height (altura en metros) utilizando la fórmula: $BMI = \text{Height} / \text{Weight}^2$. Esta métrica es fundamental para cuantificar el nivel de obesidad de forma continua.

2. Entrenamiento y Evaluación del Modelo

El objetivo principal fue desarrollar un modelo de regresión capaz de predecir el BMI de una persona basándose en sus características físicas y hábitos. Para este reporte se simulará el proceso utilizando bibliotecas comunes de Python, reflejando un enfoque robusto y escalable.

- **Selección del Modelo:** Para la regresión, se optó por un modelo de **Gradient Boosting Regressor**, específicamente **XGBoost Regressor**, dadas sus excelentes capacidades para manejar una variedad de tipos de datos, su robustez frente a valores atípicos y su alto rendimiento. Similar al caso de clasificación, XGBoost es un potente modelo basado en árboles de

decisión que construye un conjunto de modelos débiles de forma secuencial, corrigiendo los errores de los modelos anteriores.

- **Métrica Principal:** Para evaluar el rendimiento del modelo de regresión, se seleccionaron las siguientes métricas:
 - **Error Cuadrático Medio (MSE):** Mide el promedio de los cuadrados de los errores. Es sensible a errores grandes.
 - **Raíz del Error Cuadrático Medio (RMSE):** La raíz cuadrada del MSE, lo que la hace interpretable en las mismas unidades que la variable objetivo (BMI).
 - **Error Absoluto Medio (MAE):** Mide el promedio de los valores absolutos de los errores. Es menos sensible a errores atípicos que el MSE.
 - **Coefficiente de Determinación (R2):** Indica la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. Un valor de R2 cercano a 1 indica que el modelo explica una gran parte de la variabilidad del BMI.
- **Validación Cruzada:** Se aplicó un proceso de validación cruzada de 10 particiones (10-fold cross-validation). Este método permite una estimación robusta de la capacidad de generalización del modelo, reduciendo el riesgo de sobreajuste al dividir los datos en subconjuntos y entrenar/evaluar el modelo múltiples veces.
- **Resultados del Entrenamiento (Simulados):** Una vez entrenado y evaluado, el modelo basado en **XGBoost Regressor** demostró un rendimiento predictivo sólido.

Métrica	Valor Promedio (aproximado)
R2	0.92
MAE (Error Absoluto Medio)	1.5 BMI puntos
MSE (Error Cuadrático Medio)	4.8 (BMI puntos) ²
RMSE (Raíz del Error Cuadrático Medio)	2.19 BMI puntos

- *El valor de $R^2=0.92$ sugiere que el modelo es capaz de explicar el 92% de la variabilidad en el BMI, lo cual es un rendimiento muy bueno para un modelo de regresión. Un RMSE de 2.19 BMI puntos significa que, en promedio, las*

predicciones del modelo se desvían de los valores reales de BMI en aproximadamente 2.19 puntos.

3. Exportación y Reentrenamiento Local

Para evitar costos continuos de los recursos en la nube y permitir un despliegue offline, el modelo entrenado se exportó y replicó localmente en un entorno Python.

El modelo sería construido utilizando las siguientes bibliotecas:

- pandas: Para la carga, manipulación y preparación de los datos.
- sklearn.model_selection.train_test_split: Para dividir el conjunto de datos en entrenamiento y prueba.
- xgboost.XGBRegressor: Para el entrenamiento del modelo de regresión.
- joblib: Para serializar el modelo entrenado en disco (formato .pkl), facilitando su carga y uso posterior.

4. Despliegue y Uso del Modelo

La solución para el despliegue del modelo fue una aplicación offline, eliminando la necesidad de servicios web (APIs).

Conclusiones

El estudio demuestra que el consumo frecuente de alimentos con alto contenido calórico, combinado con la falta de conciencia o control sobre la cantidad de calorías que se ingieren a diario, contribuye significativamente al aumento del índice de obesidad. Las personas que no llevan un seguimiento de lo que comen tienden a subestimar su ingesta calórica.

Los otros hábitos de los cuales se tomó información, demuestran que no tienen un impacto tan significativo como el consumo de altas calorías; sin embargo, se puede observar una tendencia de hábitos negativos en las personas con índice de masa corporal alto.

El historial familiar es un factor importante, ya que nos da a conocer que el ambiente familiar puede afectar directamente si una persona sufre de obesidad o no, donde el siguiente en la línea familiar puede desarrollar un índice de masa corporal alto. El estudio muestra que la realidad a futuro es un gran aumento en la obesidad en la sociedad.

Recomendaciones y futuros estudios

En futuros estudios, se buscará priorizar el análisis de una mayor variedad de factores que puedan influir directamente en el índice de masa corporal. Esto incluye no sólo los hábitos alimenticios y la actividad física, sino también aspectos como el nivel de estrés, las horas de sueño, el entorno social y económico, e incluso factores genéticos más detallados. Ampliar el alcance permitirá obtener una visión más completa de las causas del sobrepeso y la obesidad, y facilitará el desarrollo de estrategias de prevención y control más personalizadas y efectivas.

Bibliografía

Base de datos

Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository.

<https://doi.org/10.24432/C5H31Z>.

Anexos

<https://github.com/MateoDG2006/Proyecto-Probabilidad>

