Tecnológico de Monterrey

“Matemáticas y ciencia de datos para la toma de decisiones”

“Proyecto de Ciencia de Datos MA1042”

Alumno: Matías Piedra Pichardo

Matrícula: A01772503

Índice:

Introducción:

La obesidad es una enfermedad compleja que implica un exceso de grasa corporal y no es solo un problema estético, sino un problema médico que aumenta el riesgo de diversas enfermedades y problemas de salud como enfermedades cardíacas, diabetes, presión arterial alta, colesterol alto, enfermedad hepática, apnea del sueño y ciertos tipos de cáncer. Las causas de la obesidad son multifactoriales, involucrando factores hereditarios, fisiológicos y ambientales, junto con la alimentación, la actividad física y las opciones de ejercicio. Afortunadamente, incluso una modesta pérdida de peso puede mejorar o prevenir muchos problemas de salud asociados con la obesidad. Para abordar la obesidad, se pueden implementar cambios en la alimentación, aumentar la actividad física y realizar cambios de conducta, así como considerar medicamentos recetados y procedimientos para bajar de peso.

En un esfuerzo por mejorar mi salud y bienestar general, llevo un registro detallado de todos los alimentos que consumo a lo largo de cada semana. Este registro incluye información sobre los alimentos consumidos, su contenido nutricional y si se consideran saludables o no. Con esta información, busco identificar patrones en mi dieta y hacer ajustes que me permitan alcanzar mis objetivos de salud. La ciencia de datos juega un papel fundamental en el análisis de registros alimentarios para mejorar la salud y el bienestar. Utilizando técnicas de ciencia de datos, es posible transformar grandes volúmenes de datos alimentarios en información valiosa y procesable. La recopilación y almacenamiento sistemático de datos detallados sobre cada alimento consumido permite organizar y acceder fácilmente a estos registros. La limpieza y preprocesamiento de datos garantiza la coherencia y precisión de la información, corrigiendo valores faltantes o erróneos y estandarizando los datos.

(Mayo Clinic, 2023)

Objetivo fundamental:

El objetivo principal de este proyecto es analizar mi dieta diaria para identificar patrones y tendencias en el consumo de alimentos. A partir de este análisis, se busca hacer recomendaciones para mejorar mis hábitos alimenticios y, por ende, mi salud general.
Como cliente y sujeto de estudio en este proyecto, estoy centrado en resolver problemas relacionados con desequilibrios en mi consumo nutricional, tales como excesos o deficiencias en calorías, carbohidratos, lípidos, proteínas y sodio. La meta es identificar estos desequilibrios y proporcionar soluciones que me ayuden a alcanzar un balance nutricional óptimo.
La Ciencia de Datos es la herramienta clave para lograr estos objetivos. Utilizando técnicas de análisis descriptivos y predictivos, puedo ajustar mi dieta diaria de manera informada. Para desarrollar estas soluciones, necesito aprender técnicas de

análisis de datos, visualización de datos y modelos predictivos específicos para la nutrición. El proceso involucra la recopilación y limpieza de datos, seguido de análisis descriptivos y predictivos, y la generación de visualizaciones claras que permitan interpretar los resultados de manera efectiva.
Las hipótesis planteadas en este proyecto son fundamentales para guiar el análisis. La hipótesis nula (H0) sugiere que no hay una relación significativa entre mi consumo de alimentos y la mejora de mis hábitos alimenticios. En contraste, la hipótesis alternativa (H1) plantea que existe una relación significativa entre mi consumo de alimentos y la mejora de mis hábitos alimenticios. Probar estas hipótesis es crucial para validar la efectividad del análisis.
La construcción del modelo está respaldada por elementos teóricos de nutrición básica, como el balance de macronutrientes y micronutrientes. Además, se utilizarán técnicas avanzadas de análisis de datos y modelos predictivos para identificar patrones y hacer recomendaciones dietéticas. Estos conceptos teóricos y técnicos proporcionan una base sólida para el desarrollo de un modelo que pueda ofrecer recomendaciones precisas y efectivas.

Definir métricas de éxito es esencial para evaluar el impacto de las recomendaciones. Las métricas incluirán la reducción en el consumo de alimentos no saludables, alcanzar un balance recomendado de macronutrientes (carbohidratos, proteínas, lípidos y sodio) y la disminución en la ingesta diaria de sodio, alineándose con las recomendaciones nutricionales. Estas métricas no solo medirán el progreso, sino que también proporcionarán un marco para realizar ajustes continuos y mejorar mis hábitos alimenticios a lo largo del tiempo.

Metodología

La metodología implementada en este proyecto sigue un flujo de trabajo sistemático y estructurado para asegurar que los objetivos sean alcanzados de manera efectiva. A continuación, se detallan cada una de las etapas del flujo de trabajo:

Recolección de datos

Descripción: El primer paso es registrar diariamente todos los alimentos consumidos, incluyendo detalles como porciones, ingredientes y horarios de consumo. Cada alimento debe estar acompañado por sus valores nutricionales específicos, como calorías, carbohidratos, proteínas, grasas, y sodio.
Herramientas: Para la recolección de datos, se utilizarán aplicaciones móviles o diarios digitales que permitan una entrada de datos fácil y precisa. También se puede considerar el uso de bases de datos nutricionales en línea para obtener la información nutricional de los alimentos consumidos.

Limpieza de datos

Descripción: Una vez recolectados los datos, es crucial asegurarse de su calidad. Esto implica verificar la precisión de los datos ingresados, corregir posibles errores y manejar datos faltantes o inconsistentes.
Herramientas: Se utilizarán lenguajes de programación como Python o R, junto con bibliotecas de manipulación de datos como Pandas para limpiar y preparar los datos.

Análisis descriptivo

Descripción: En esta etapa, se aplicarán técnicas de estadística descriptiva para entender los patrones en el consumo de alimentos. Esto incluye calcular medidas de tendencia central (media, mediana), dispersión (desviación estándar, rango) y distribución (histogramas, frecuencias).
Herramientas: Python o R, utilizando bibliotecas como Matplotlib, Seaborn, y Statsmodels para la visualización y análisis de datos.

Modelado predictivo

Descripción: Se desarrollarán modelos predictivos para identificar posibles mejoras en la dieta. Esto puede incluir técnicas de regresión, clasificación y clustering para prever el impacto de cambios en la dieta y sugerir optimizaciones.
Herramientas: Herramientas de Machine Learning como Scikit-learn en Python, así como técnicas de modelado como regresión lineal, árboles de decisión y clustering.

Visualización de datos

Descripción: Crear gráficos y visualizaciones que ayuden a interpretar los resultados del análisis descriptivo y predictivo. Las visualizaciones deben ser claras, informativas y accesibles para facilitar la toma de decisiones.
Herramientas: Bibliotecas de visualización de datos como Matplotlib, Seaborn, Plotly, y herramientas como Tableau para crear dashboards interactivos.

Descripción y preparación de los datos

Descripción de los datos:

Tipos de datos: Datos nutricionales de los alimentos consumidos (calorías, carbohidratos, lípidos, proteínas, sodio) y su clasificación como saludable o no.
Origen de los datos: Datos registrados manualmente en un archivo Excel basado en el consumo diario de alimentos.
Adecuación de los datos para el análisis: Los datos son adecuados ya que proporcionan la información nutricional necesaria para el análisis de patrones de consumo.

Ajustes a los datos:

Agregar datos: No se han agregado datos adicionales.
Integrar datos: Los datos están integrados en un solo archivo Excel.
Modificar datos: Algunos valores pueden necesitar normalización para el análisis.
Remover o eliminar información: Datos incompletos o erróneos serán removidos para asegurar la calidad del análisis.

Descripción de los datos mediante el uso de estadística descriptiva:

Python:
El archivo contiene un análisis de regresión lineal múltiple para predecir el contenido calórico de alimentos basado en sus componentes nutricionales (carbohidratos, lípidos y proteínas; sodio se elimina porque p es mayor 0.05). Los datos se cargan desde un archivo Excel (A01772503.xlsx), se limpian eliminando valores atípicos y se dividen en conjuntos de entrenamiento (80%) y prueba (20%). Se realiza un análisis exploratorio de datos, incluyendo histogramas, diagramas de caja y bigote, y análisis de correlación. El modelo se ajusta y se valida mediante la verificación de la suma de los residuos, pruebas de normalidad, homocedasticidad e independencia de los residuos. Las predicciones se hacen sobre el conjunto de prueba con valores faltantes imputados y se evalúan usando el error cuadrático medio (MSE) y el coeficiente de determinación (R^2 = 98.5%).

```
modelo = smf.ols("Calorias ~ Carbohidratos + Lípidos + Proteína", data=df_train).fit()
print(modelo.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Calorias   R-squared:                       0.985
Model:                           OLS   Adj. R-squared:                  0.985
Method:                Least Squares   F-statistic:                     6650.
Date:               Sat, 18 May 2024   Prob (F-statistic):          3.56e-278
Time:                       05:04:20   Log-Likelihood:                -1266.4
No. Observations:                310   AIC:                             2541.
Df Residuals:                    306   BIC:                             2556.
Df Model:                          3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -23.0106      1.774    -12.969      0.000     -26.502     -19.519
Carbohidratos   4.4926      0.049     91.575      0.000       4.396       4.589
Lípidos        10.6759      0.206     51.892      0.000      10.271      11.081
Proteína        4.5210      0.101     44.604      0.000       4.322       4.720
==============================================================================
Omnibus:                       35.608   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.209
Skew:                           0.331   Prob(JB):                     1.46e-33
Kurtosis:                       6.357   Cond. No.                         73.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

En base a esto podemos ajustar la siguiente ecuación:
y = -23.0106 (intercepto) + 4.4926x2(carbohidratos) + 10.679x3(lípidos) + 4.5210x4(proteína)

Minitab:

Ecuación de regresión

Calorias = -17.58 + 4.4231 Carbohidratos + 8.938 Lípidos + 4.4309 Proteína + 0.02046 Sodio

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-17.58	1.74	-10.08	0.000	
Carbohidratos	4.4231	0.0194	228.45	0.000	1.88
Lípidos	8.938	0.274	32.58	0.000	1.81
Proteína	4.4309	0.0894	49.58	0.000	1.79
Sodio	0.02046	0.00770	2.66	0.008	1.84

Resumen del modelo

S	R-cuadrado	R-cuadrado(ajustado)	R-cuadrado (pred)
20.3749	99.75%	99.75%	99.73%

Coeficientes:

- **Constante**: -17.58 (EE: 1.74, Valor T: -10.08, Valor p: 0.000)
- **Carbohidratos**: 4.4231 (EE: 0.0194, Valor T: 228.45, Valor p: 0.000, FIV: 1.88)
- **Lípidos**: 8.938 (EE: 0.274, Valor T: 32.58, Valor p: 0.000, FIV: 1.81)
- **Proteína**: 4.4309 (EE: 0.0894, Valor T: 49.58, Valor p: 0.000, FIV: 1.79)
- **Sodio**: 0.02046 (EE: 0.00770, Valor T: 2.66, Valor p: 0.008, FIV: 1.84)

Resumen del modelo:

- **S (Error estándar de la estimación)**: 20.3749
- **R-cuadrado**: 99.75%
- **R-cuadrado (ajustado)**: 99.75%
- **R-cuadrado (pred)**: 99.73%

Estos resultados indican que el modelo tiene un ajuste excelente, explicando el 99.75% de la variabilidad en las calorías basándose en los carbohidratos, lípidos, proteínas y sodio. Todos los coeficientes son estadísticamente significativos (valor p < 0.05), y los factores de inflación de la varianza (FIV) sugieren que no hay problemas graves de multicolinealidad.

Excel:

1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.99871058							
5	R Square	0.99742281							
6	Adjusted R S	0.9973924							
7	Standard Eri	19.1696113							
8	Observation:	344							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	48212503.1	12053125.8	32799.9419	0			
13	Residual	339	124573.685	367.473998					
14	Total	343	48337076.8						
15									
16		*Coefficients*	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-17.714307	1.63082468	-10.862178	8.7525E-24	-20.922117	-14.506497	-20.922117	-14.506497
18	Carbohidrato	4.43070162	0.01949429	227.282041	0	4.39235662	4.46904662	4.39235662	4.46904662
19	Lípidos	9.37643152	0.2668516	35.1372503	4.818E-115	8.85153804	9.90132501	8.85153804	9.90132501
20	Proteína	4.39152744	0.0909787	48.2698434	5.878E-154	4.21257357	4.5704813	4.21257357	4.5704813
21	Sodio	0.01390752	0.0074476	1.86738184	0.0627111	-0.0007418	0.02855685	-0.0007418	0.02855685
22									

Clarías=−17.71307+4.43070162 · Carbohidratos+9.37643152 · Lípidos+4.391527444 · Proteína+0.01390752 · Sodio

El valor de R^2 es 0.99742281, lo que indica que el modelo explica el 99.74% de la variabilidad en las calorías basándose en las variables predictoras.

Gracias al intercepto podemos saber que si no existen las otras variables, no hay algún alimento.
Construcción del Modelo
Para construir el modelo de predicción de calorías, utilizaremos una regresión lineal.

VENTAJAS Y DESVENTAJAS:

Ventajas:
Potente para análisis y manipulación de datos con bibliotecas como pandas, numpy, y scikit-learn.
Flexibilidad en la automatización de tareas y creación de pipelines de datos.
Amplia comunidad y abundante documentación.
Desventajas:
Curva de aprendizaje más pronunciada para quienes no están familiarizados con la programación.
Requiere instalación y configuración de un entorno de desarrollo.
Minitab:

Ventajas:

Intuitivo y fácil de usar para análisis estadísticos sin necesidad de programación.
Buena visualización y herramientas específicas para el análisis de datos industriales.
Desventajas:
Menos flexible que Python para la automatización y manipulación de grandes volúmenes de datos.
Es una herramienta de pago, lo que puede ser una barrera para algunos usuarios.
Excel:

Ventajas:
Muy accesible y conocido por una amplia audiencia.
Herramientas robustas para análisis y visualización de datos a pequeña escala.
Desventajas:
Limitado en capacidad de procesamiento para grandes conjuntos de datos.
Menos potente para análisis estadísticos avanzados y modelado predictivo comparado con Python y Minitab.

Interpretación de datos:

El modelo de regresión lineal múltiple revela que las calorías de los alimentos se pueden predecir con alta precisión (R^2 = 99.74%) utilizando carbohidratos, lípidos y proteínas como variables predictoras, mientras que el sodio no tiene un impacto significativo (p-valor = 0.062711).

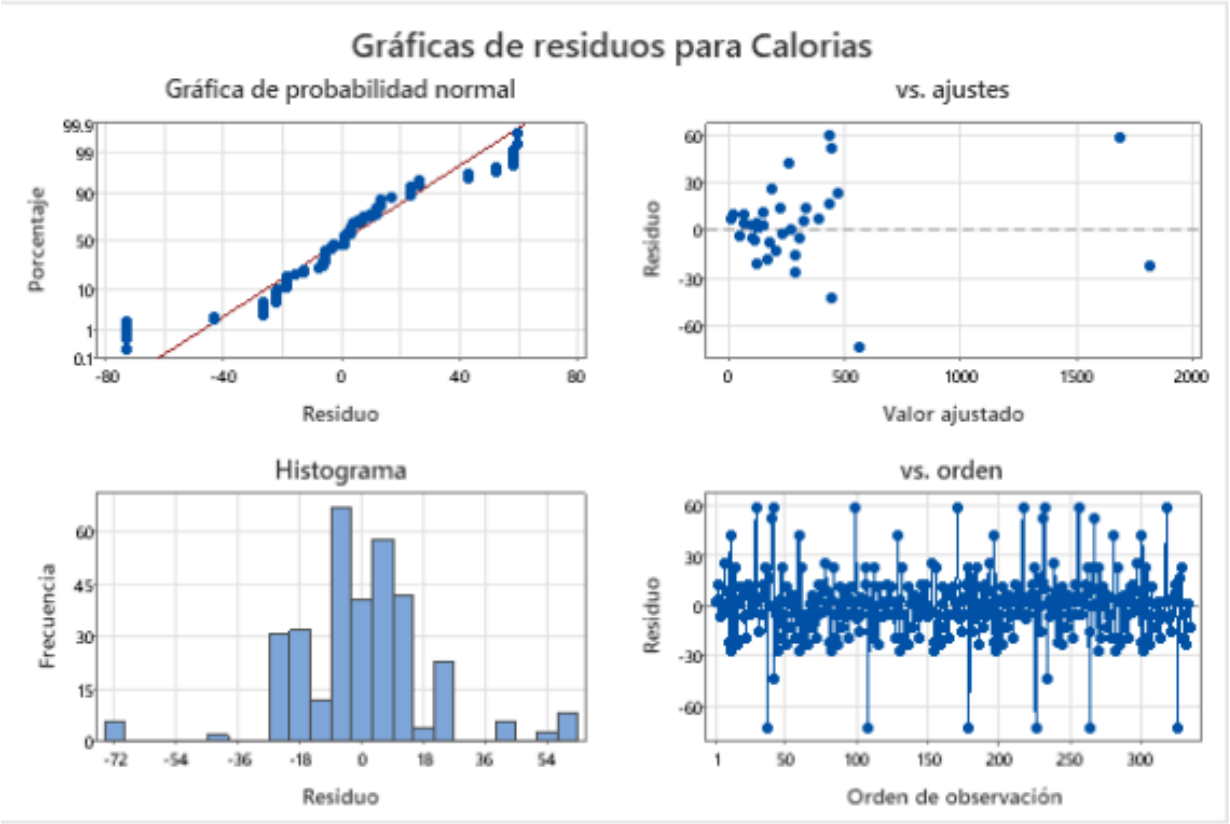
La ecuación resultante es:

$$\textbf{Clarías} = -17.71307 + 4.43070162 \cdot \textbf{Carbohidratos} + 9.37643152 \cdot \textbf{Lípidos} + 4.39152744 \cdot \textbf{Proteína} + 0.01390752 \cdot \textbf{Sodio}$$

Todos los coeficientes significativos sugieren que por cada unidad adicional de carbohidratos, lípidos y proteínas, las calorías aumentan en 4.43, 9.38 y 4.39 unidades respectivamente. El modelo es altamente significativo en su conjunto (F = 32799.9419, p-valor = 0), con un error estándar de 19.1663116, lo que confirma un ajuste preciso del modelo, permitiendo una predicción robusta del contenido calórico basado en los componentes nutricionales.

Comprobación:

Minitab:

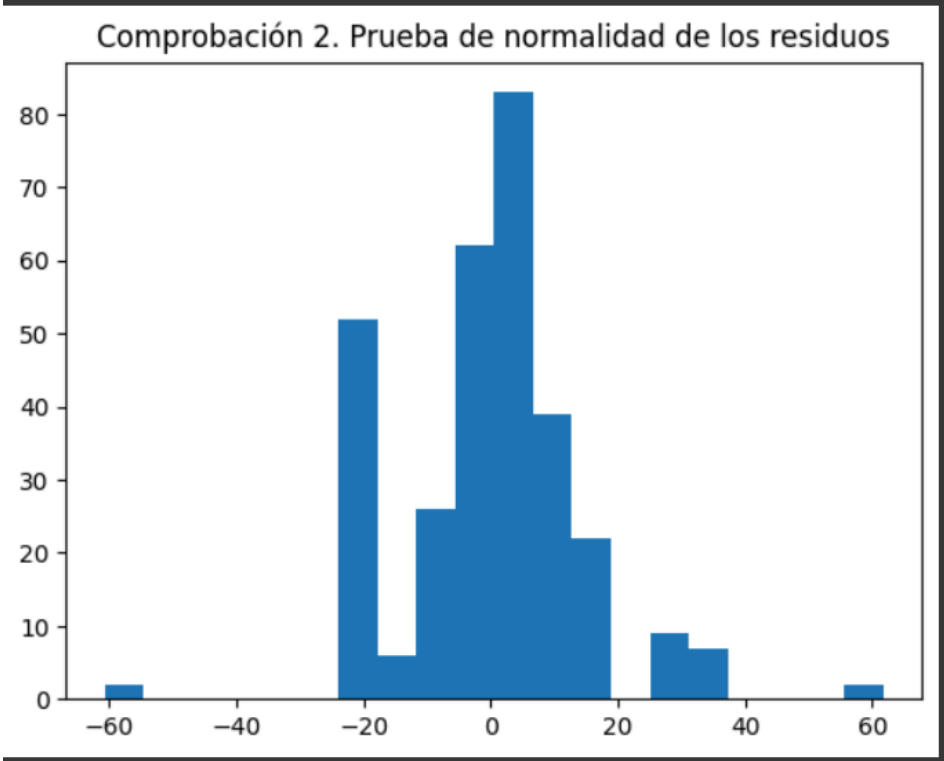


Python:

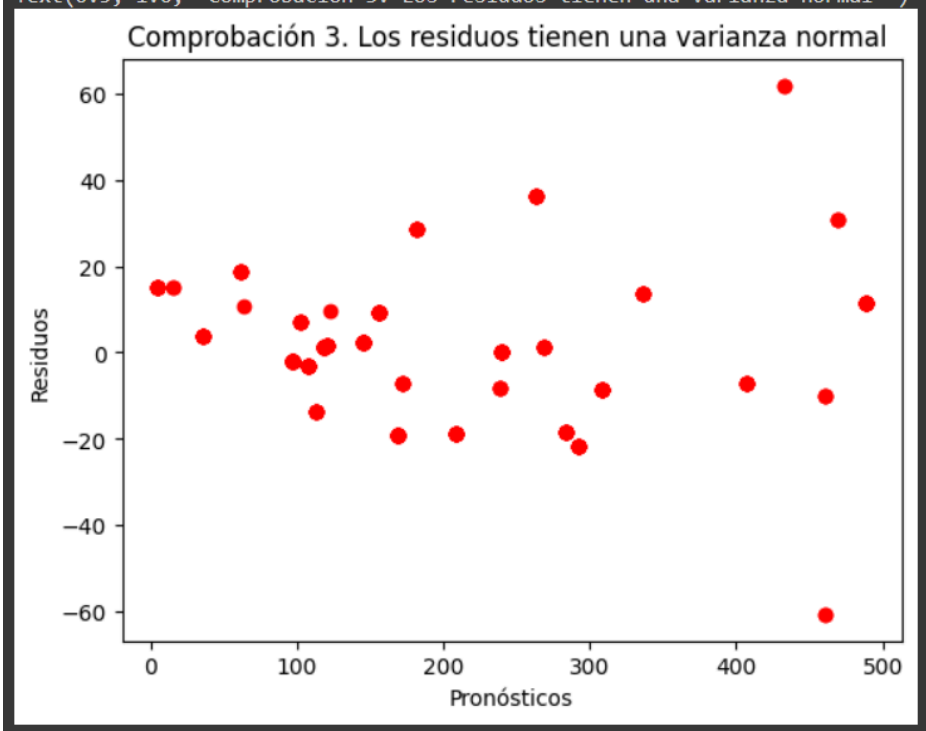
```
suma_res = sum(modelo.resid)
print("Comprobación 1. Ka suma de los residuos es: ", suma_res)
```

```
Comprobación 1. Ka suma de los residuos es:  -1.8616219676914625e-12
```

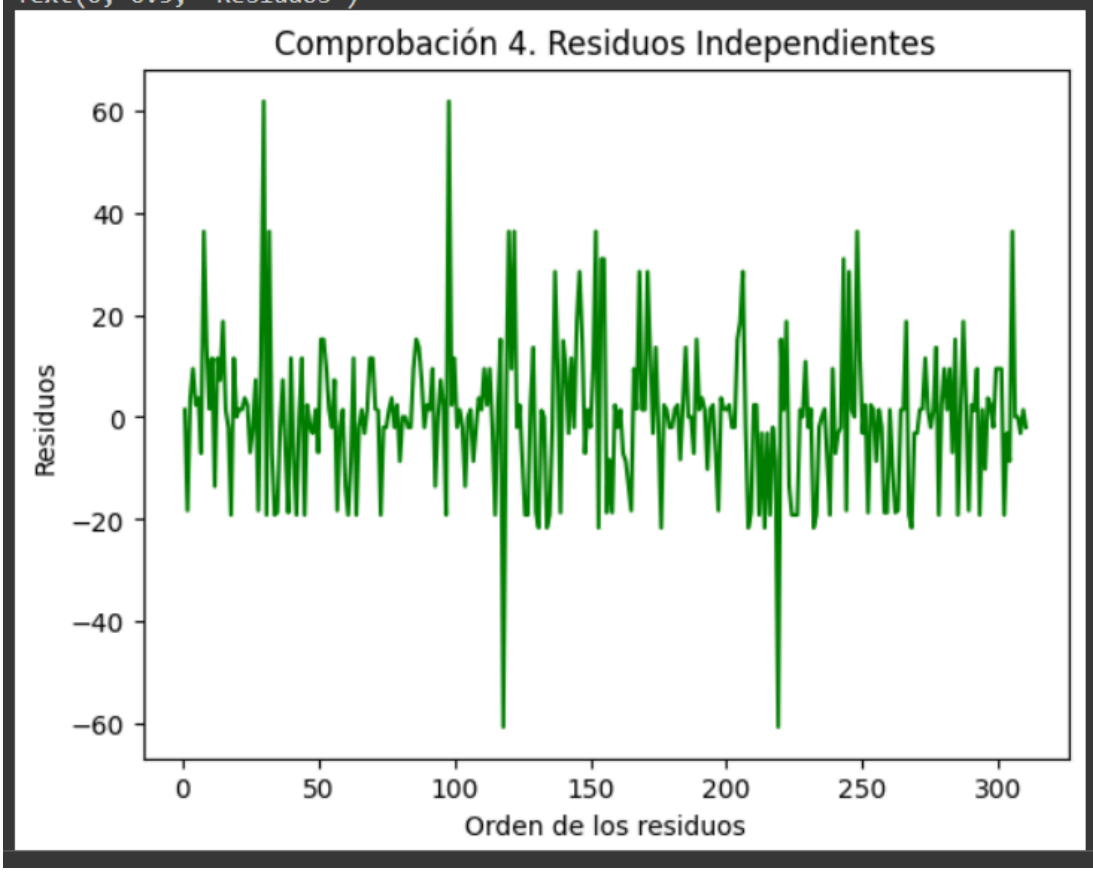
```
import matplotlib.pyplot as plt
plt.hist (modelo.resid,bins=20)
plt.title ("Comprobación 2. Prueba de normalidad de los residuos")
plt.show()
```



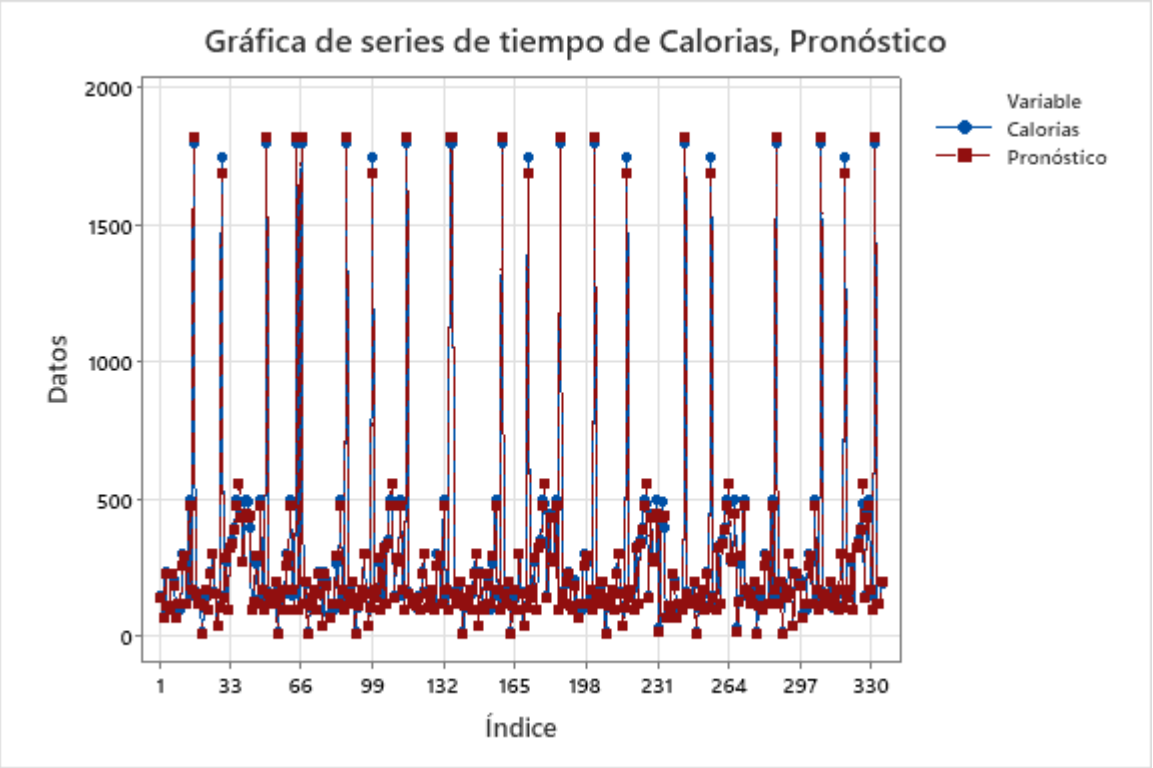
```
Text(0.5, 1.0, 'Comprobación 3. Los residuos tienen una varianza normal ')
```




```
Text(0, 0.5, 'Residuos')
```



Comparación de modelo con valores reales:



El gráfico de series de tiempo muestra que el modelo de pronóstico es efectivo para predecir las calorías reales, con ambos conjuntos de datos presentando picos periódicos significativos. Aunque hay algo de ruido y variabilidad, el modelo coincide estrechamente con los datos reales, indicando un buen desempeño.

Modelos:

Python:https://drive.google.com/file/d/1W5jvP2pgRha8WJEcu5vBgbT4YrL2Jpe5/view?usp=drive link

Excel:https://docs.google.com/spreadsheets/d/1Kpj9bulynWXd3H3b9Cn1h2f4TELwKbng/edit?usp=drive_link&ouid=101178778890081233610&rtpof=true&sd=true

Minitab:https://drive.google.com/file/d/1ptVQ1yCCEac6A6372YKEQ99cHKTZaiC5/view?usp=drive link

Conclusiones:

La obesidad es un problema médico complejo que requiere un enfoque multifactorial. Este proyecto ha utilizado técnicas avanzadas de ciencia de datos para analizar mi dieta diaria y mejorar mis hábitos alimenticios. A través de un registro detallado de los alimentos consumidos y su contenido nutricional, se han identificado patrones de consumo y se han hecho recomendaciones para alcanzar un balance nutricional óptimo.

Los modelos predictivos desarrollados en Python, Minitab y Excel muestran una alta precisión ($R^2 \approx 99.75\%$) en la predicción del contenido calórico basado en carbohidratos, lípidos y proteínas. Estos modelos permiten prever el impacto de diferentes combinaciones de alimentos en el consumo calórico total, facilitando la identificación de desequilibrios nutricionales y la generación de recomendaciones específicas para mejorar la salud.

Cada herramienta utilizada tiene sus ventajas: Python es potente y flexible para grandes volúmenes de datos, Minitab es intuitivo para análisis sin programación, y Excel es accesible para análisis a pequeña escala. La aplicación continua de estas técnicas permitirá ajustes y mejoras sostenibles en mis hábitos alimenticios, ayudando a manejar la obesidad y mejorar la salud a largo plazo.

References

Mayo Clinic. (2023, July 22). *Obesidad - Síntomas y causas*. Mayo Clinic. Retrieved June 14, 2024, from https://www.mayoclinic.org/es/diseases-conditions/obesity/symptoms-causes/syc-20375742