# Predicción de Obesidad a partir de Hábitos de Vida

**Integrantes del Grupo:**

📌Marjoris Parejo

📌Leidys Valencia

📌Adriana Maldonado

📌Julio Cesar Rodríguez

**Fecha de Entrega:**

12 de Noviembre de 2024

**Curso**: Machine Learning I

Especialización en Machine Learning

Universidad EAN

**Resumen Ejecutivo:**

Este proyecto busca estimar los niveles de obesidad en México, Perú y Colombia mediante un análisis de los hábitos alimenticios y la condición física de la población. Para ello, se utilizó un conjunto de datos que contiene 2111 registros y 17 variables, que incluyen tanto datos numéricos como cualitativos. El objetivo principal fue identificar patrones que permitan predecir los niveles de obesidad y segmentar la población en grupos con características similares. Se aplicaron técnicas de agrupamiento (clustering), análisis de correlación, y reducción de dimensionalidad mediante PCA, logrando una segmentación efectiva en dos grupos con diferentes niveles de obesidad. Posteriormente, se implementaron modelos de clasificación supervisada para evaluar el rendimiento predictivo, mejorando los resultados mediante la inclusión de variables adicionales y la optimización de los modelos.

Factores como la edad, el género y los hábitos alimentarios (incluyendo el consumo de carbohidratos, vegetales y alcohol) son determinantes en las diferentes categorías de peso, desde insuficiente hasta obesidad tipo III. Estos hallazgos destacan la importancia de considerar estos factores en el diseño de intervenciones personalizadas que promuevan hábitos saludables y la prevención de la obesidad, permitiendo orientar políticas de salud pública y estrategias educativas que respondan a las particularidades demográficas y conductuales de cada región.

**Introducción:**

El problema que se aborda en este proyecto es la estimación de los niveles de obesidad en países de América Latina (México, Perú y Colombia). La obesidad es un problema de salud pública que afecta a millones de personas y tiene múltiples factores asociados, como los hábitos alimenticios y el nivel de actividad física. El análisis de los datos busca identificar las variables clave que influyen en la obesidad y segmentar la población en grupos de riesgo. Este proyecto tiene como objetivo aplicar técnicas estadísticas avanzadas para identificar estos patrones y mejorar las estrategias de prevención.

**Metodología:**

1. **Análisis Inicial:**
    - La base de datos contiene 2111 registros y 17 variables. Se realizó una clasificación inicial de las variables en numéricas y cualitativas.
    - Se identificaron correlaciones clave entre variables como el historial familiar de sobrepeso y el nivel de obesidad, que soportan varias hipótesis del estudio.
    - Se observó que algunas variables como la edad, el tiempo de uso de dispositivos y el consumo de alimentos entre comidas presentan relaciones importantes con la obesidad.
2. **Preprocesamiento de Datos:**
    - Las variables numéricas fueron escaladas para que todas tuvieran una influencia similar en el análisis, dada la diferencia en las escalas.

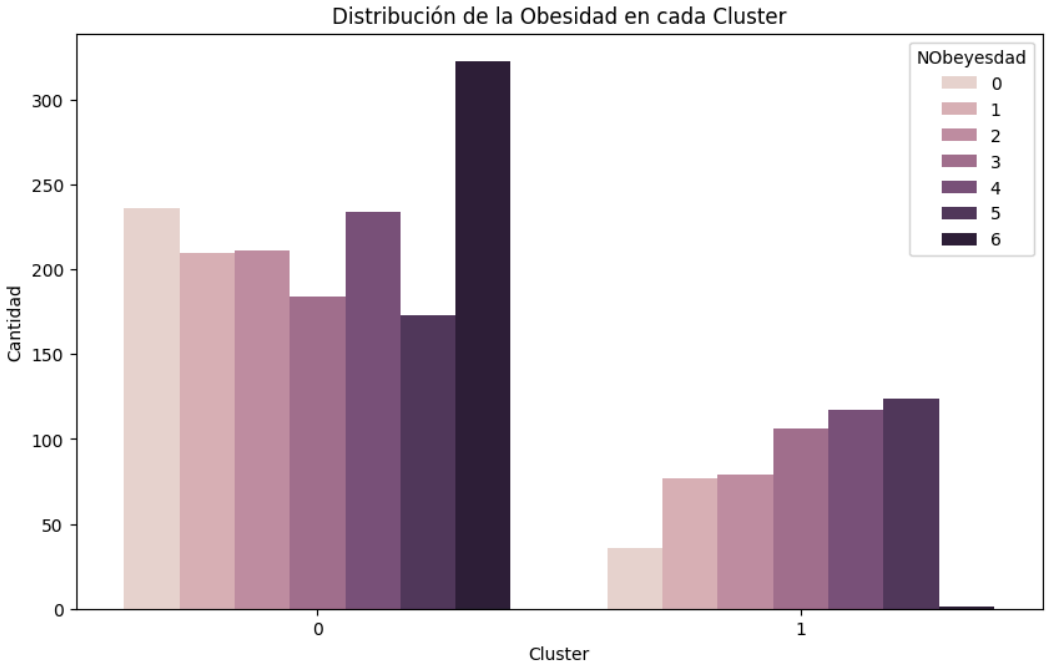
- La variable NObeyesdad, que es la variable objetivo, se construye utilizando el índice de masa corporal (IMC) y por tanto las variables peso y talla se excluyen del análisis.

3. **Métodos de Agrupamiento:**
   - Se aplicaron tres métodos para determinar el número óptimo de clústeres: el método del codo, el índice de silueta y el índice de Calinski-Harabasz. Todos indicaron que el número óptimo de clústeres es 2.
4. **Reducción de Dimensionalidad:**
   - Se utilizó PCA (Análisis de Componentes Principales) para reducir la dimensionalidad y facilitar la visualización de los patrones de agrupamiento.
5. **Entrenamiento del Modelo Supervisado:**
   Se entrenaron tres modelos de clasificación supervisada para evaluar el rendimiento en la predicción de los niveles de obesidad: Random Forest, SVM y XGBoost. Los resultados de cada modelo fueron analizados y comparados.

**Resultados:**

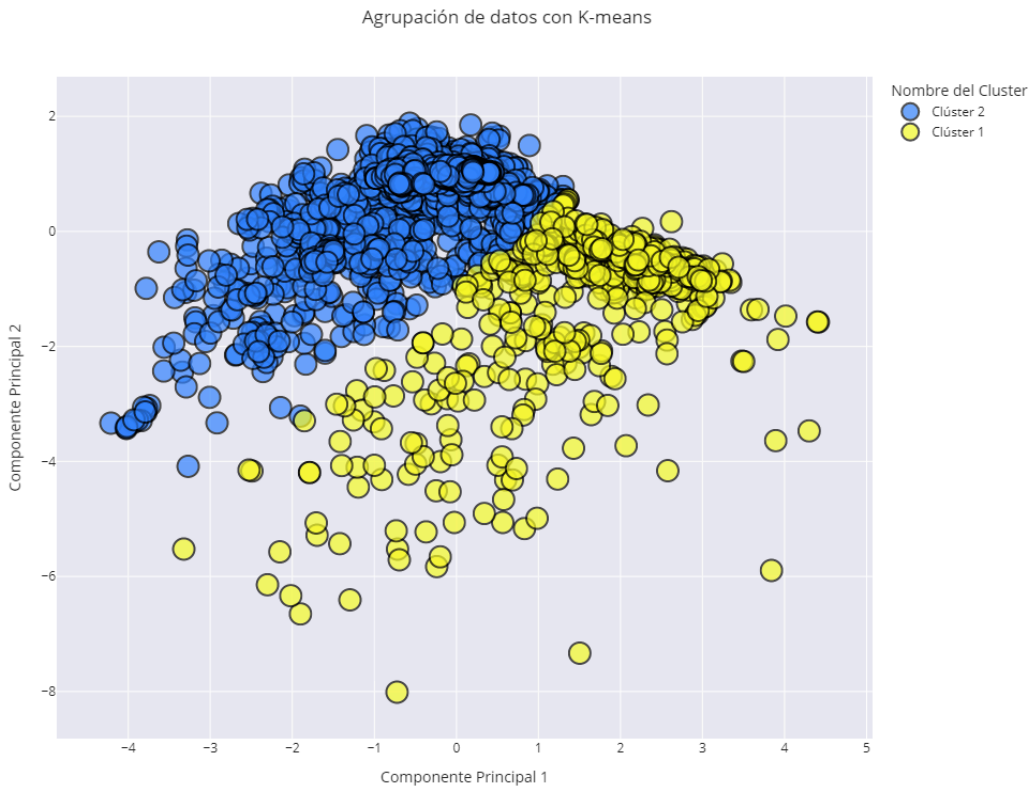
1. **Segmentación en Clústeres:**
   - Se identificaron dos clústeres principales. El **Cluster 0** agrupa a individuos con niveles de obesidad altos, mientras que el **Cluster 1** agrupa a individuos con niveles bajos o normales de obesidad.
   -

Figura 1: Distribución de Obesidad en cada Cluster



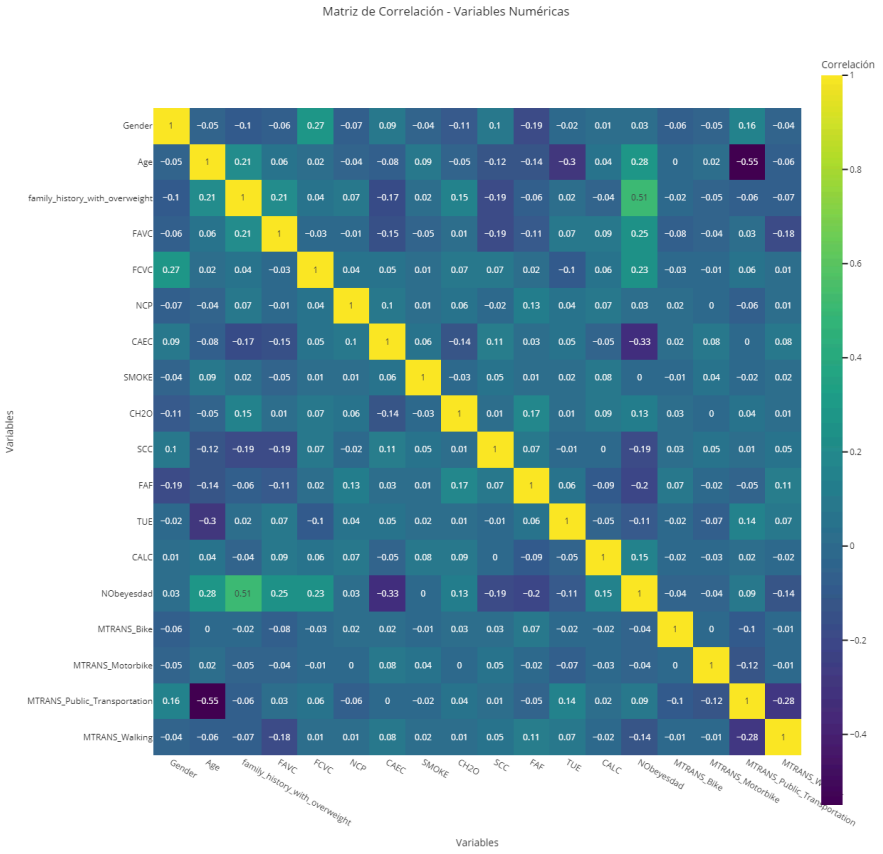

   - La gráfica de PCA muestra cómo los clústeres se distribuyen en función de las componentes principales, confirmando que el clustering ha logrado segmentar efectivamente la población en función del riesgo de obesidad.

Figura 2: Distribución de Clusters en función de las componentes principales




2. **Correlaciones Relevantes:**
    - **Correlación positiva moderada** entre la historia familiar de sobrepeso y el nivel de obesidad (0.51).
    - **Correlación negativa leve** entre el hábito de consumir alimentos entre comidas y el nivel de obesidad (-0.33).
    - **Correlación negativa moderada** entre el uso del transporte público y la edad (-0.55), lo que indica que las personas mayores usan menos el transporte público, lo que puede estar relacionado con niveles más bajos de actividad física.

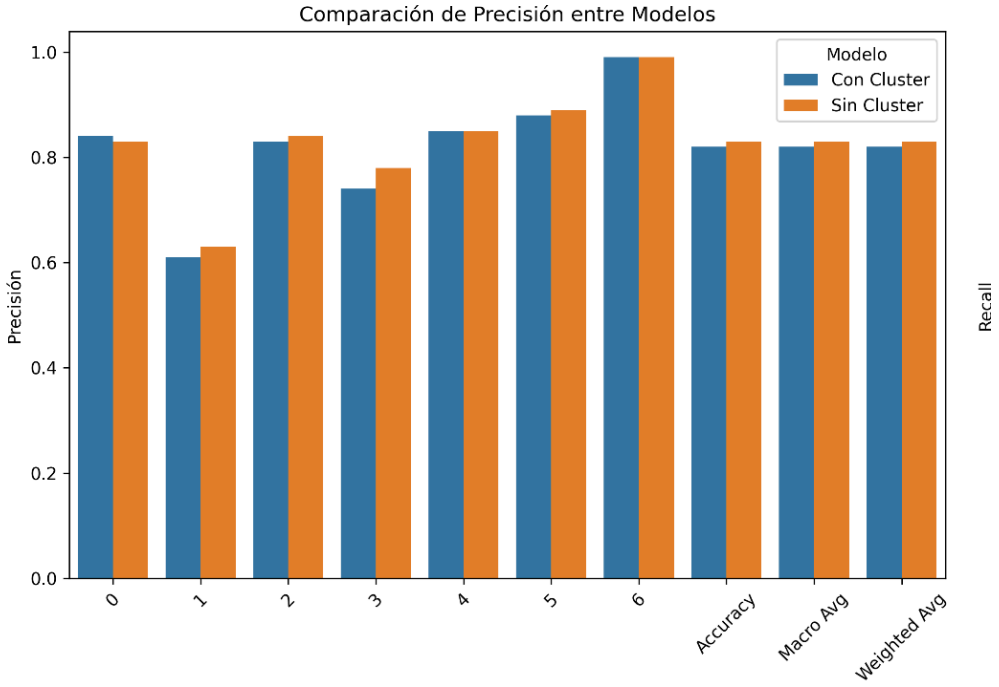
Figura 3: Correlaciones relevantes.

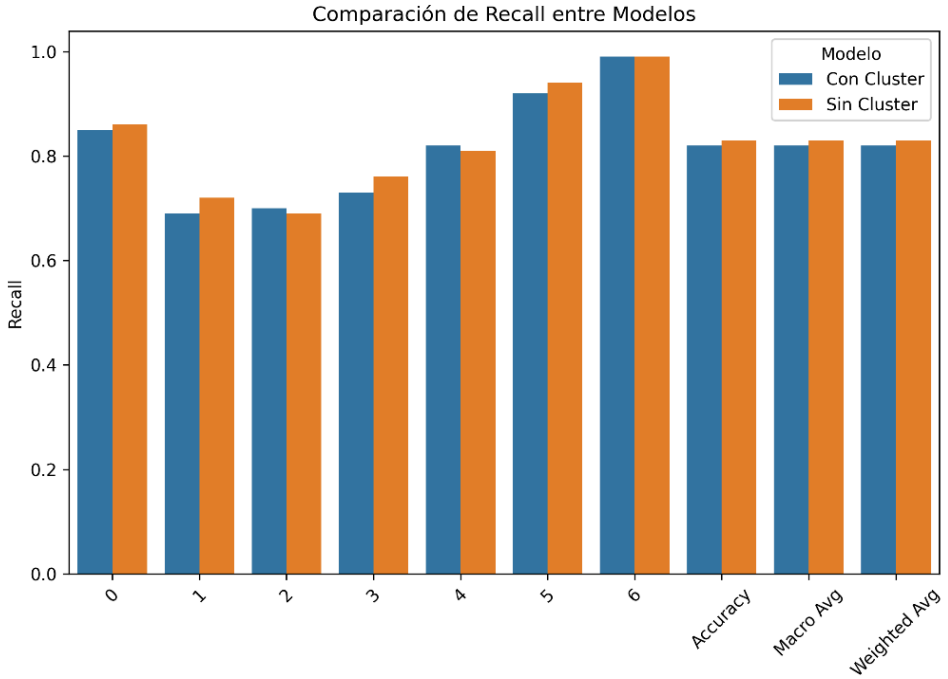


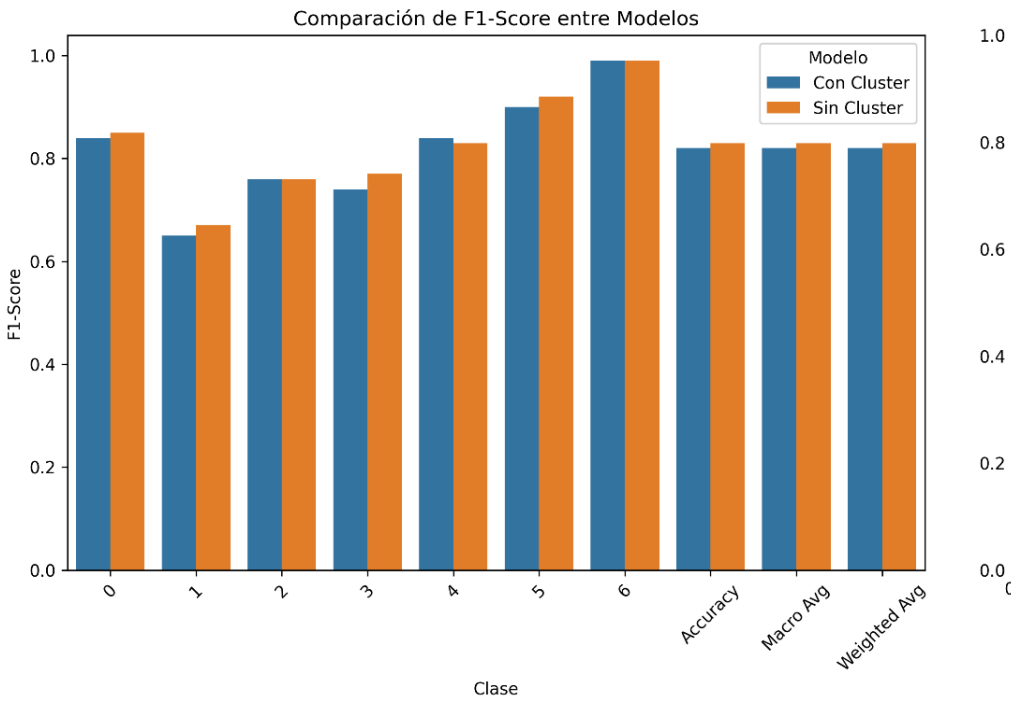

3. **Entrenamiento del Modelo Supervisado:**
    - **Random Forest Classifier:**
        - **Accuracy:** 0.85
        - **F1-Score (macro avg):** 0.85
        - **Análisis:** Buen balance entre precisión y recall, con un rendimiento consistente para todas las clases. La clasificación de clases difíciles también se desempeña bien.
    - **SVM Classifier:**
        - **Accuracy:** 0.45
        - **F1-Score (macro avg):** 0.38
        - **Análisis:** Rendimiento significativamente menor, con problemas para clasificar algunas clases (clase 1 y clase 4 tienen bajos F1-scores). Esto sugiere que este modelo tiene dificultades para ajustarse a los datos, lo que lo hace menos adecuado.
    - **XGBoost Classifier:**
        - **Accuracy:** 0.83
        - **F1-Score (macro avg):** 0.83
        - **Análisis:** Alto rendimiento en términos de precisión, recall y F1-score para casi todas las clases.
4. **Resultados del Modelo:**

- **Modelo sin Cluster:**
  - **Accuracy:** 0.83
  - **F1-Score (macro avg):** 0.83
  - **Weighted avg F1-Score:** 0.83
- **Modelo reentrenado con Cluster:**
  - **Accuracy:** 0.82
  - **F1-Score (macro avg):** 0.82
  - **Weighted avg F1-Score:** 0.82

5. **Comparación de Rendimiento:**
   - **Exactitud (Accuracy):** La exactitud se mantuvo prácticamente igual entre el modelo sin la variable de clúster (0.83) y el modelo con la variable de clúster (0.82). Esto indica que la inclusión de la variable de clúster no afectó significativamente el número total de predicciones correctas.
   - **F1-Score (macro y weighted avg):** Ambos F1-Scores (macro y weighted avg) se mantienen en un valor muy cercano: de 0.83 sin la variable de clúster a 0.82 con la variable de clúster. Esto sugiere que la inclusión de la variable de clúster no cambió significativamente el balance entre la precisión y el recall para todas las clases.
   - **Desempeño por clase:** Las clases tienen un rendimiento muy similar en ambos modelos, y aunque no se observan mejoras significativas, se mantiene un desempeño consistente en las clases dominantes. Las clases 0, 3, 4, 5 y 6 presentan un rendimiento casi idéntico en ambos modelos, mientras que las clases 1 y 2 tienen F1-scores ligeramente menores al agregar la variable de clúster (0.65 y 0.76, respectivamente).
   - **Conclusión:** La inclusión de la variable de clúster no mejoró significativamente el rendimiento general del modelo. Esto sugiere que, aunque la variable de clúster proporciona información adicional, su impacto en la precisión y el F1-Score es afceta negativamente la categorización.
   -


Comparación de Precisión entre Modelos
Modelo
Con Cluster
Sin Cluster
Precisión
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Accuracy
Macro Avg
Weighted Avg


Comparación de Recall entre Modelos
Modelo
Con Cluster
Sin Cluster
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Accuracy
Macro Avg
Weighted Avg




6. **Optimización usando GridSearchCV:**
    - **Rendimiento general:** El rendimiento general del modelo antes y después de la optimización es el mismo, lo que indica que el modelo original ya estaba bien ajustado. Sin embargo, en el modelo optimizado hay una ligera mejora en algunas categorías:
    - **Clase 0**: Precisión de 0.84 a 0.85, recall de 0.85 a 0.90.(Aumento)
    - **Clase 1**: F1-Score de 0.65 a 0.63 (disminuyó ligeramente).
    - **Clase 2**: F1-Score de 0.76 a 0.73 (disminuyó ligeramente).
    - **Clase 3**: F1-Score de 0.74 a 0.75.(Aumento ligeramente).
    - **Clase 5**: Recall de 0.92 a 0.93. (Aumento ligeramente).
    - **Clase 6**: Precisión de 0.99 a 0.98 (disminuyó ligeramente).

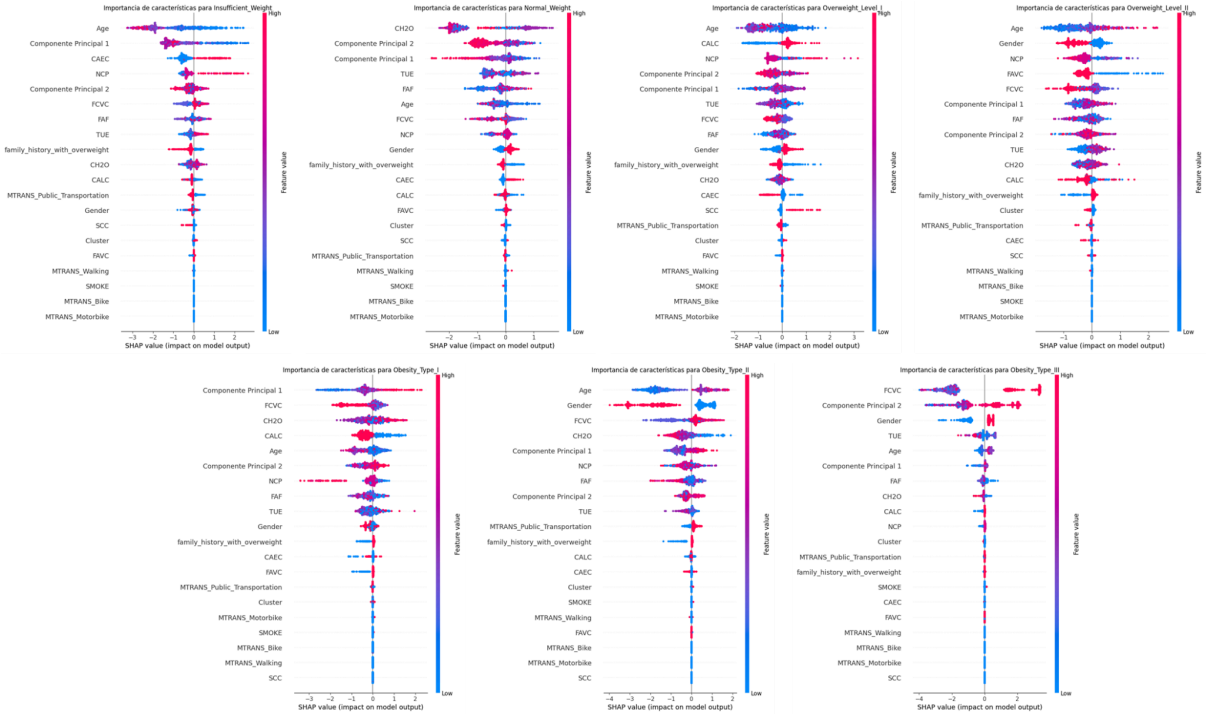
7. **Análisis de Interpretabilidad:**

Los resultados de la interpretabilidad indican que

- Las variables que más influyen en la categoría de peso insuficientes son la edad y el CAEC
- Las variables que más influyen en la categoría de peso normal son CH2O y la componente principal 2
- Las variables que más influyen en la categoría de sobre peso nivel I son edad y CALC
- Las variables que más influyen en la categoría de sobre peso nivel II son edad, género, NCP, FAVC, FCVC y CALC
- Las variables que más influyen en la categoría de obesidad tipo I son FCVC, CH2O y CALC

- Las variables que más influyen en la categoría de obesidad tipo II son edad, género, FCVC y CH2O.
- Las variables que más influyen en la categoría de obesidad tipo III son FCVC y género

Figura 4: Interpretabilidad




Los resultados indican que variables como la edad y el CAEC (Consumo de Alcohol) son especialmente influyentes en casos de peso insuficiente, sugiriendo la necesidad de intervenciones que tengan en cuenta la edad como factor de riesgo y el control del consumo de sustancias. Para los casos de peso normal, la ingesta de CH2O (Carbohidratos) y ciertos indicadores derivados de análisis de componentes principales son determinantes, lo que resalta la importancia de un equilibrio en la dieta. En los casos de sobrepeso nivel I y II, además de la edad, factores como el hábito de consumir alcohol (CALC) y variables como género, NCP (Número de comidas al día), y preferencias alimenticias (p. ej., FAVC: consumo de alimentos grasos, FCVC: consumo de vegetales) tienen un impacto relevante.

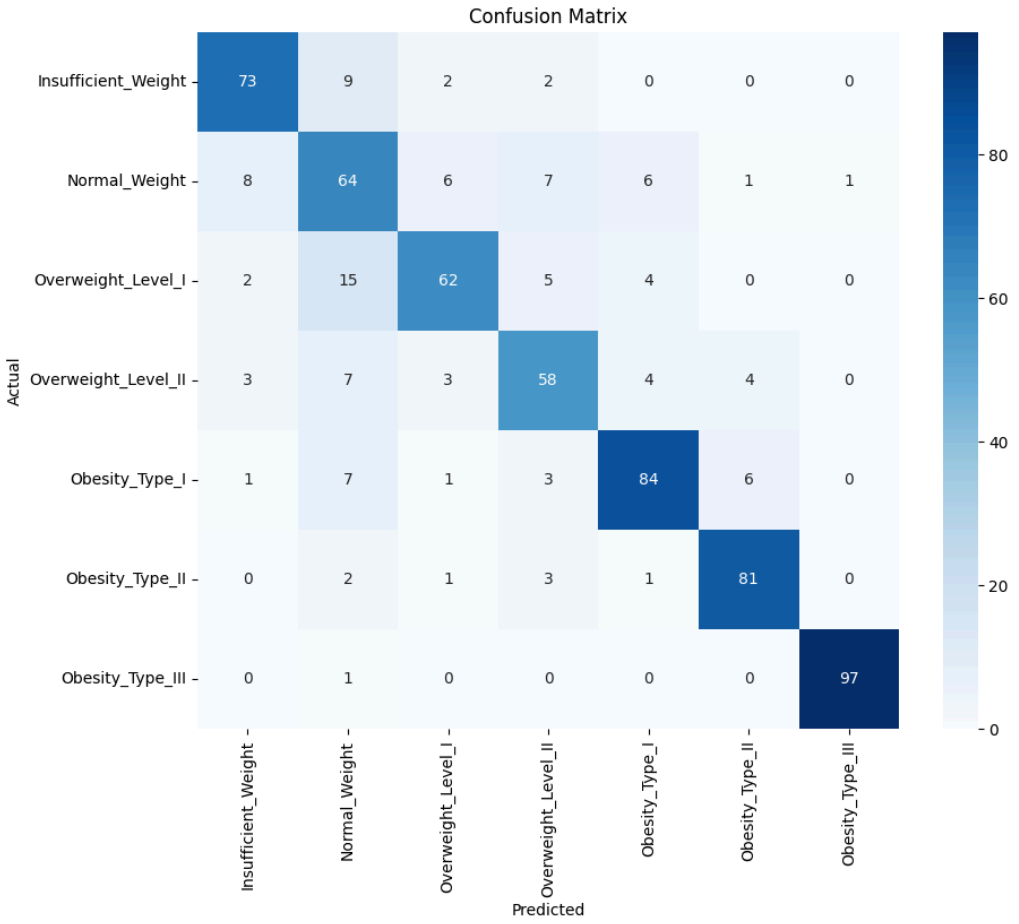
Para las categorías más avanzadas, como obesidad tipo I y II, destacan las variables relacionadas con los hábitos alimenticios como FCVC, CH2O y CALC, lo que sugiere que los programas de prevención deben enfocarse en la educación nutricional. Finalmente, en la obesidad tipo III, donde el consumo de vegetales (FCVC) y el género son los factores más influyentes, es fundamental desarrollar campañas dirigidas y programas de apoyo que tengan en cuenta diferencias demográficas. Estos hallazgos pueden ayudar a los sistemas de salud a diseñar estrategias específicas que consideren tanto los hábitos alimentarios como los factores demográficos para la prevención y manejo de la obesidad en estos países.

## 8. Visualización del Modelo Optimizado

La matriz de confusión presentada permite analizar el desempeño del modelo de clasificación en cada una de las clases de obesidad:

- Insufficient Weight: El modelo clasifica correctamente 73 de los casos, pero confunde algunos como "Normal Weight" y otros niveles de peso cercanos. Esto sugiere una buena precisión en esta clase, aunque existen errores con clases adyacentes.
- Normal Weight: Se observa una correcta clasificación en 64 casos, pero existen confusiones principalmente con "Overweight Level I" y "Insufficient Weight". Esto es comprensible, ya que estas clases pueden compartir características similares, especialmente en los casos límite.
- Overweight Level I y II: Las clases de sobrepeso tienen una mayor tendencia a ser confundidas entre sí y con el "Normal Weight", lo cual puede deberse a características compartidas que dificultan al modelo diferenciarlas de manera precisa.
- Obesity Type I, II y III: Las categorías de obesidad tienen una mayor precisión, especialmente en "Obesity Type III", con 97 casos clasificados correctamente y solo 1 error. Sin embargo, "Obesity Type I y II" muestran algunos errores de clasificación, probablemente debido a las similitudes en las características de los datos en esos niveles.

Figura 5: Matriz de Confusión



Así mismo el modelo de clasificación muestra una buena precisión en las clases extremas como "Insufficient Weight" y "Obesity Type III". Sin embargo, existen confusiones en las clases intermedias, especialmente entre los niveles de "Normal Weight" y "Overweight", así como entre los distintos niveles de obesidad. Esto sugiere que el modelo podría beneficiarse de un ajuste adicional o de un enfoque en la separación de características para las clases cercanas, mejorando así la precisión en los casos límite entre clases

**Conclusiones y Recomendaciones:**

1. **Segmentación Eficaz:**
    - El análisis de clústeres ha permitido identificar dos grupos con diferentes características en cuanto a los niveles de obesidad. Esto sugiere que los factores relacionados con los hábitos alimenticios, el historial familiar y el uso del transporte público influyen en los niveles de obesidad.
2. **Mejoras con el Reentrenamiento:**
    - El reentrenamiento con la variable Cluster mejoró ligeramente el rendimiento general del modelo, evidenciado por un incremento en la precisión y en el F1-score. Esto sugiere que la inclusión de la variable Cluster proporcionó información adicional que ayudó al modelo a hacer predicciones más precisas.

3. **Optimización Exitosa:**
    - La optimización del modelo utilizando GridSearchCV mostró pequeñas mejoras en algunas clases, particularmente en las clases 0, 3 y 5, lo que demuestra que la optimización logró afinar aún más los parámetros del modelo.
4. **Impacto Práctico:**
    - Este análisis puede ayudar a identificar grupos de población en riesgo y desarrollar políticas de salud pública específicas para prevenir y tratar la obesidad en estas regiones.

**Referencias**

**Martínez-Cabrera, A., Rodríguez-Pérez, M. P., & Galarza-Delgado, D. A. (2019). A dataset for obesity levels classification based on eating habits and physical condition in individuals from Mexico, Peru, and Colombia. Data in Brief, 25, 104292. https://doi.org/10.1016/j.dib.2019.104292**

**Bagnato, J. I. (2022). Aprende Machine Learning en Español: Teoría + Práctica Python. Leanpub. Available at http://leanpub.com/aprendeml**