Inteligencia Artificial para una Salud Pública Mejorada: Clasificación y Prevención de la Obesidad en América Latina mediante Machine Learning.

Aplicaciones de Machine Learning en la Evaluación de la Obesidad en países de América Latina

Jiliar Antonio Silgado Cardona

ean universidad

Contenido

Resumen Ejecutivo

Resumen del Proyecto

En respuesta al creciente problema de obesidad en América Latina, especialmente en países como México, Perú y Colombia, este proyecto busca implementar un enfoque de machine learning que permita evaluar y clasificar de forma precisa los niveles de obesidad. Con un análisis exhaustivo de los patrones de alimentación y las condiciones físicas de los individuos, el proyecto desarrolla un modelo predictivo capaz de identificar factores de riesgo y clasificar a los individuos según su nivel de obesidad, desde Peso Insuficiente hasta Obesidad Tipo III.

La meta es brindar a las autoridades y profesionales de salud una herramienta basada en datos que optimice la asignación de recursos y apoye decisiones informadas en la prevención y el tratamiento de la obesidad.

Objetivos del Proyecto

1. Desarrollar un modelo de machine learning supervisado para clasificar con precisión los niveles de obesidad, abarcando categorías desde Peso Insuficiente hasta Obesidad Tipo III.
2. Identificar patrones clave en los hábitos alimenticios y condiciones físicas que influyen en la obesidad.
3. Optimizar la asignación de recursos en programas de prevención mediante evaluaciones objetivas basadas en datos precisos.
4. Apoyar la toma de decisiones en salud pública con un sistema de evaluación que combine datos generados sintéticamente (77%) y datos reales de usuarios (23%).

Hallazgos Clave

- La inclusión de los clústeres generados mediante K-means en el modelo supervisado mejora significativamente la precisión de la clasificación de

niveles de obesidad, proporcionando una perspectiva más segmentada y contextualizada.

- Los reentrenamientos periódicos del modelo con datos actualizados y la optimización continua de hiperparámetros mejoran su desempeño, haciendo que las clasificaciones sean más precisas y ajustadas a la evolución del problema.
- Las visualizaciones y el análisis interpretativo con SHAP o LIME ofrecen una comprensión profunda de las variables que influyen en los niveles de obesidad, permitiendo identificar factores de riesgo específicos.
- El enfoque integral de este modelo predictivo no solo apoya la toma de decisiones en salud pública, sino que también optimiza la asignación de recursos en programas de prevención, proporcionando una herramienta robusta para afrontar el problema de obesidad de manera efectiva y sostenible.

Introducción

La obesidad se ha convertido en un desafío de salud pública de gran magnitud en América Latina, afectando a países como México, Perú y Colombia, donde los índices de obesidad han aumentado significativamente en las últimas décadas. Este problema, impulsado por factores como los cambios en los hábitos alimenticios, el sedentarismo y las condiciones socioeconómicas, afecta la calidad de vida de millones de personas y representa un reto para los sistemas de salud pública, que enfrentan la necesidad de estrategias de prevención y tratamiento más efectivas.

Ante esta situación, las aplicaciones de machine learning ofrecen una solución prometedora, permitiendo analizar grandes volúmenes de datos para identificar patrones complejos en los hábitos y condiciones físicas asociados con la obesidad. A través del uso de técnicas avanzadas de aprendizaje automático y análisis de datos, este proyecto busca desarrollar un modelo predictivo que clasifique los niveles de obesidad con mayor precisión y objetividad, proporcionando una base sólida para la formulación de políticas de salud pública más focalizadas y eficientes. Con datos que incluyen atributos como hábitos alimenticios y características físicas, el modelo no solo apoya la identificación temprana de personas en riesgo, sino que también optimiza la asignación de recursos para intervenciones preventivas.

Este enfoque no solo representa una innovación en la evaluación de la obesidad, sino que también refuerza el papel de la inteligencia artificial en la mejora de la salud pública en América Latina, ofreciendo una herramienta poderosa para enfrentar uno de los problemas más apremiantes de la región.

Metodología

La metodología del proyecto para la evaluación de la obesidad en América Latina mediante machine learning sigue un enfoque estructurado en cinco fases principales, desde la obtención de los datos hasta la implementación y supervisión del modelo predictivo. A continuación, se detalla cada fase de la metodología utilizada:

1. **Obtención de los Datos**: Los datos empleados en este proyecto provienen de una combinación de registros generados sintéticamente (77%) y datos recopilados de usuarios mediante una plataforma web (23%), abarcando información demográfica, hábitos alimenticios y condiciones físicas de individuos en América Latina. Esta combinación de fuentes garantiza un conjunto de datos amplio y representativo, fundamental para la construcción de un modelo predictivo robusto y confiable.
2. **Preparación de los Datos**: La fase de preparación incluyó un análisis y limpieza exhaustivos para detectar y corregir valores faltantes o inconsistentes. Posteriormente, se aplicó una reducción de dimensionalidad a través del Análisis de Componentes Principales (ACP), lo cual permitió simplificar el conjunto de datos sin pérdida significativa de información. Esta reducción facilita la visualización y optimiza el rendimiento en fases posteriores, mejorando la precisión y eficiencia del modelo.
3. **Entrenamiento del Modelo:**

   La fase de entrenamiento incluyó la creación de dos modelos de machine learning, uno no supervisado y otro supervisado:

   - **Modelo No Supervisado:** Se utilizó un análisis de agrupamientos (clustering) mediante K-means. El número óptimo de clústeres se determinó utilizando el método del codo, lo cual permitió identificar grupos homogéneos de individuos basados en características comunes relacionadas con sus hábitos y condiciones físicas. Estos clústeres se

añadieron como características adicionales para mejorar la capacidad predictiva del modelo supervisado.

- **Modelo Supervisado:** Se probó el rendimiento de varios modelos de clasificación, incluyendo Random Forest, Logistic Regression, Support Vector Classifier (SVC) y XGBoost, eligiendo finalmente Random Forest como el modelo con mejor desempeño inicial en la clasificación de niveles de obesidad. Para maximizar la precisión y optimizar el modelo, se aplicaron técnicas avanzadas de optimización de hiperparámetros, utilizando GridSearchCV y RandomizedSearchCV. Esto permitió ajustar los parámetros de ambos modelos y establecer la configuración con el mejor rendimiento.

4. **Integración del Modelo**: Una vez completada la fase de entrenamiento, el modelo fue implementado en un punto de conexión, habilitando su uso en tiempo real para predecir niveles de obesidad en nuevos registros de datos. Esta implementación permite la aplicación del modelo en entornos de salud y políticas públicas, donde es esencial disponer de una herramienta que clasifique con rapidez y precisión los niveles de obesidad, facilitando decisiones informadas.
5. **Supervisión del Modelo**: Para garantizar la efectividad continua del modelo, se estableció un proceso de supervisión y reentrenamiento periódico con datos nuevos. Esta supervisión incluye la generación de visualizaciones de los resultados, lo cual permite interpretar cómo los diferentes atributos afectan la clasificación de obesidad. Además, se aplicaron técnicas de interpretabilidad, como SHAP (SHapley Additive exPlanations) y LIME (Local Interpretable Model-agnostic Explanations), para evaluar el impacto de cada variable en las predicciones, proporcionando una comprensión clara y transparente del modelo.
6. **Informe Final**: El proyecto concluye con un informe exhaustivo que documenta los hallazgos y resultados de cada fase de la metodología. Este informe incluye análisis detallados y recomendaciones específicas para la formulación de políticas de salud pública y estrategias de prevención en función de los resultados obtenidos, proporcionando una base sólida para enfrentar el problema de la obesidad en América Latina con un enfoque basado en datos.

Resultados

Análisis de Modelos Random Forest

Para la clasificación de niveles de obesidad, se compararon tres configuraciones del modelo Random Forest: sin ajuste de hiperparámetros, ajustado con GridSearchCV, y ajustado con RandomizedSearchCV. Los resultados mostraron que el modelo optimizado con GridSearchCV tuvo el mejor desempeño general. A continuación se presentan los resultados específicos:

- **Random Forest Sin Ajuste de Hiperparámetros:**
  - F1-Score: 0.6477
  - Recall: 0.6572
- **Random Forest Ajustado con GridSearchCV:**
  - F1-Score: 0.6749
  - Recall: 0.6809
- **Random Forest Ajustado con RandomizedSearchCV:**
  - F1-Score: 0.6256
  - Recall: 0.6430

El modelo ajustado con GridSearchCV logró un balance óptimo en precisión, recall y F1-Score, destacándose en comparación con los otros dos enfoques. Aunque las diferencias en las métricas son relativamente pequeñas, esta configuración permite obtener una ligera mejora, lo que indica que el ajuste de hiperparámetros mediante GridSearchCV optimiza de manera más efectiva el desempeño del modelo Random Forest.

Análisis de Modelos K-Means

El algoritmo de clustering K-Means también se probó en tres configuraciones diferentes: sin ajuste de hiperparámetros, ajustado con GridSearchCV, y ajustado con RandomizedSearchCV. Los resultados se evaluaron en términos de cohesión y separación, utilizando métricas como la inercia, el índice de silueta, y el coeficiente de Calinski-Harabasz:

- **K-Means Sin Ajuste:**
    - Desempeño inferior en cohesión y separación, con una inercia alta y un índice de silueta bajo. Esto sugiere mayor dispersión y menor calidad en la separación de los clústeres.
- **K-Means Ajustado con GridSearch:**
    - Mejoró la separación de los clústeres y redujo la inercia, aunque con una leve disminución en la cohesión, evidenciada por una caída en el coeficiente de Calinski-Harabasz.
- **K-Means Ajustado con RandomSearch:**
    - Mostró la mejor separación de clústeres, con el índice de silueta más alto y la inercia más baja. Sin embargo, esta mejora en la separación se acompañó de una ligera pérdida de cohesión, reflejada en el coeficiente de Calinski-Harabasz.

En conclusión, el modelo K-Means ajustado con RandomizedSearchCV se identificó como el más equilibrado en términos de separación, logrando una segmentación de los clústeres que permite una mejor representación de los grupos en el análisis.

Interpretabilidad de las Variables

Utilizando técnicas de interpretabilidad como SHAP y LIME, se analizaron las contribuciones específicas de diferentes variables en el modelo. Las principales variables interpretadas y su impacto en el modelo son:

- **FCVC (Frecuencia de Consumo de Vegetales Completos):** La distribución muestra un efecto positivo del consumo de vegetales en el modelo. La mayoría de los puntos de datos se encuentran en el área positiva, destacando que un mayor consumo de vegetales está asociado con resultados favorables.
- **NCP (Nivel de Consumo de Calorías):** La distribución de NCP sugiere que un mayor consumo de calorías tiene un impacto positivo en la predicción, especialmente en el rango de 0.2 a 0.3. Los valores más bajos de consumo calórico se asocian con un impacto negativo en las predicciones.
- **TUE (Tiempo de Ejercicio):** Un bajo tiempo de ejercicio se asocia con un impacto negativo en el modelo. Los puntos positivos en el área derecha del gráfico indican que un mayor tiempo de ejercicio tiene un impacto positivo.

- **CH2O (Consumo de Azúcares):** El consumo de azúcares presenta una influencia significativa en el modelo. La distribución homogénea muestra que tanto el aumento como la reducción de azúcares afectan la predicción de manera considerable.
- **FAF (Frecuencia de Actividad Física):** Un mayor nivel de actividad física está relacionado con un impacto positivo en el modelo, mientras que la falta de actividad física se asocia con un efecto negativo en la predicción.

Los resultados de esta actividad reflejan la efectividad de los modelos Random Forest y K-Means ajustados con técnicas de optimización de hiperparámetros para clasificar niveles de obesidad y segmentar grupos con características comunes. La integración de interpretabilidad permite una comprensión más profunda de los factores de riesgo y su impacto, proporcionando información valiosa para futuras políticas de salud pública en la región.

Análisis de Clusters

En el análisis de clusters, se identificaron patrones y características distintivas en cada grupo. La agrupación se basa en la variable de interés, **NObeyesdad**, permitiendo observar similitudes y diferencias en las concentraciones y distribuciones de los valores dentro de cada cluster.

Clusters 0 y 1

Estos clusters presentan una **alta concentración en valores bajos de NObeyesdad**. Sin embargo, existen diferencias en la homogeneidad de las observaciones:

- **Cluster 0** tiene una concentración más intensa y un pico pronunciado, lo que indica que la mayoría de las observaciones en este grupo tienen valores bajos de NObeyesdad y se encuentran muy cerca entre sí.
- **Cluster 1** también se agrupa en valores bajos, pero muestra una ligera dispersión en comparación con el Cluster 0, lo que sugiere menos homogeneidad en sus valores.

Este patrón implica que ambos clusters contienen observaciones similares en cuanto a niveles bajos de NObeyesdad, con una mayor homogeneidad en el Cluster 0 y una mayor variabilidad en el Cluster 1.

Clusters 2, 3, y 9

Estos clusters destacan por tener **distribuciones sesgadas hacia valores altos**, con picos múltiples que sugieren una variabilidad notable en los datos:

- **Cluster 2** y **Cluster 9** son multimodales, lo que significa que presentan concentraciones en varios puntos específicos dentro de su rango. Esto sugiere la existencia de subgrupos dentro de estos clusters, que podrían estar formados por observaciones con características similares.
- **Cluster 3** también muestra una inclinación hacia valores altos, aunque su distribución es menos compleja en comparación con los Clusters 2 y 9.

Este comportamiento multimodal en algunos clusters sugiere una mayor diversidad interna en estos grupos, posiblemente reflejando la presencia de características compartidas por subgrupos de observaciones.

Clusters 4, 5, y 6

Estos clusters presentan concentraciones unimodales en ciertos valores específicos, aunque existen diferencias en las tendencias de sesgo:

- **Cluster 4** muestra una concentración en valores altos de NObeyesdad, indicando que las observaciones en este grupo son más homogéneas en niveles elevados de la variable.
- **Clusters 5** y **6** están sesgados hacia valores bajos y presentan concentraciones en clases intermedias. Esto implica que estos clusters agrupan observaciones que se comportan de manera similar en rangos intermedios, con una tendencia hacia valores más bajos.

Los patrones unimodales observados en estos clusters sugieren una estructura interna más homogénea en comparación con los clusters multimodales, con agrupaciones en torno a valores específicos de NObeyesdad.

Clusters 7 y 8

Estos clusters presentan un **sesgo hacia valores altos** y concentran observaciones en clases medias, aunque con menor intensidad que los clusters con valores bajos:

- La estructura de los clusters 7 y 8 es similar, con una mayor dispersión en las observaciones hacia valores altos, pero sin una concentración dominante. Esto indica que estos clusters contienen observaciones con una mayor variabilidad y no presentan un pico de concentración específico.

En resumen, los **Clusters 0 y 1** representan observaciones en niveles bajos de NObeyesdad, con una estructura más homogénea en el Cluster 0. **Clusters 2, 3 y 9** son multimodales, mostrando variabilidad y subgrupos en valores altos, mientras que **Clusters 4, 5 y 6** presentan patrones unimodales en valores específicos con una clara tendencia de sesgo. Finalmente, **Clusters 7 y 8** tienen mayor dispersión y sesgo hacia valores altos, con una distribución más uniforme en clases medias.

Este análisis sugiere que cada cluster agrupa datos con características distintas en cuanto a niveles de NObeyesdad, permitiendo observar patrones de homogeneidad y variabilidad que son útiles para interpretaciones y aplicaciones específicas en estudios sobre obesidad.

Conclusiones

La segmentación de datos mediante el análisis de clusters, especialmente utilizando el modelo K-Means optimizado, ha demostrado ser una herramienta valiosa para mejorar la comprensión y estructura de los datos en relación con los niveles de obesidad (NObeyesdad). La agrupación permitió identificar patrones y subgrupos con características específicas, reflejando comportamientos de consumo y actividad física que influyen en los diferentes niveles de obesidad.

El uso de clusters como características adicionales en el modelo supervisado (Random Forest) aumentó la precisión en la predicción, mejorando métricas clave como el F1-Score y el Recall. Esta integración permitió que el modelo no solo clasifique los niveles de obesidad de manera más efectiva, sino también reconozca patrones subyacentes en los datos, reforzando la interpretación de los resultados.

El modelo Random Forest ajustado con **GridSearchCV** mostró ser el de mejor rendimiento entre las opciones evaluadas, al ofrecer un balance superior entre precisión y recall. Este ajuste demuestra la importancia de optimizar los hiperparámetros, ya que incluso pequeñas mejoras en las métricas pueden tener un impacto significativo en la calidad de las predicciones. El análisis de interpretabilidad confirmó la relevancia de variables como **frecuencia de consumo de vegetales, nivel de consumo de calorías, tiempo de ejercicio, consumo de azúcares y actividad física frecuente** como factores clave en la clasificación de obesidad.

Recomendaciones para Futuros Trabajos

1. **Exploración de otros métodos de clustering**: Considerar técnicas de agrupamiento adicionales como DBSCAN o clustering jerárquico podría aportar más profundidad y diversidad en los patrones de segmentación, especialmente en grupos con alta variabilidad como los Clusters 2 y 9.
2. **Mejora de la calidad de datos**: La precisión del modelo puede beneficiarse de la inclusión de más datos demográficos, información genética o factores socioeconómicos, que permitan refinar aún más los clústeres y mejorar la capacidad predictiva.
3. **Evaluación de otros modelos supervisados**: Además de Random Forest, podría explorarse el desempeño de modelos de redes neuronales o modelos de boosting adicionales, ya que su capacidad para manejar grandes volúmenes de datos podría mejorar el rendimiento en datasets más complejos.
4. **Implementación de análisis longitudinal**: Estudios que aborden la evolución temporal de los factores analizados permitirían observar cómo cambian los patrones de obesidad a lo largo del tiempo, ofreciendo un enfoque preventivo que podría ser útil para formular políticas de salud pública.
5. **Uso de técnicas de interpretabilidad avanzadas**: La incorporación de herramientas interpretativas adicionales como SHAP o LIME, junto con visualizaciones intuitivas, mejorará la comprensión del impacto de cada variable en el modelo, facilitando el uso de estos resultados en entornos clínicos o de políticas públicas.

En conjunto, la segmentación mediante clusters y el ajuste de modelos supervisados representan una estrategia sólida para mejorar las predicciones y comprensiones en estudios de obesidad. Estas recomendaciones buscan guiar futuros trabajos hacia enfoques integradores que optimicen el rendimiento predictivo y fortalezcan la interpretabilidad y aplicabilidad de los modelos.

Referencias

- **Reducción de Dimensionalidad y Análisis de Componentes Principales (ACP)** Esquivel, F., & Morales, C. (2019). *Análisis de Componentes Principales para la Reducción de Dimensionalidad en Datos Multivariados*. Revista Colombiana de Estadística, 42(1), 75–89. doi:10.15446/rce.v42n1.74301

- **Método de K-Means y Optimización de Clústeres** López, M. D., & Guerrero, R. (2015). *Algoritmo K-means para Agrupamiento de Datos en Ciencias Sociales: Un Estudio Comparativo*. Revista de Estadística y Matemáticas Aplicadas, 7(2), 25–32.

- **Optimización de Hiperparámetros con GridSearchCV y RandomizedSearchCV** Pérez, A. C., & Gómez, R. E. (2018). *Búsqueda de Hiperparámetros para Modelos de Clasificación: Comparación entre Grid Search y Randomized Search*. Revista Latinoamericana de Investigación en Inteligencia Artificial, 22(2), 135–148.

- **Visualización y Análisis de Resultados en Modelos Predictivos** Mendoza, J., & López, F. (2019). *Visualización de Datos para el Análisis de Modelos Predictivos en Python*. Revista Científica de Computación, 18(1), 63–71.