Modelo de predicción de obesidad optimizado y validado, aplicando técnicas de exploración no supervisada y supervisada

Elaborado por:

Carlos Gabriel Contreras Serrano

Universidad EAN

Facultad de ingeniería

Especialización en machine learning

Bogotá

14/11/2024

**Resumen Ejecutivo:**

El presente ejercicio busca entrenar un modelo para predecir obesidad con datos sintéticos. La reducción de variables mediante PCA logra ser una estrategia eficiente para reducir dimensionalidad y multicolinealidad. Pasamos de trabajar con 8 a modelar con 4 variables. Tras probar diferentes alternativas como regresión logística, k vecinos cercanos, support vector machine y decisión tree, el modelo ganador fue k vecinos cercanos con una precisión de 0,61. El uso de clusters como by, permitió mejorar este modelo en 1 punto dejando la precisión en 0,62. Aunque la mejora no parece alarmante es valiosa porque se da en control de segmentos más pequeños de obesidad. Lo anterior prueba las bondades del análisis de conglomerados como técnica de aprendizaje no supervisado que eleva la precisión en las estimaciones de las técnicas de aprendizaje supervisado.

**Introducción:**

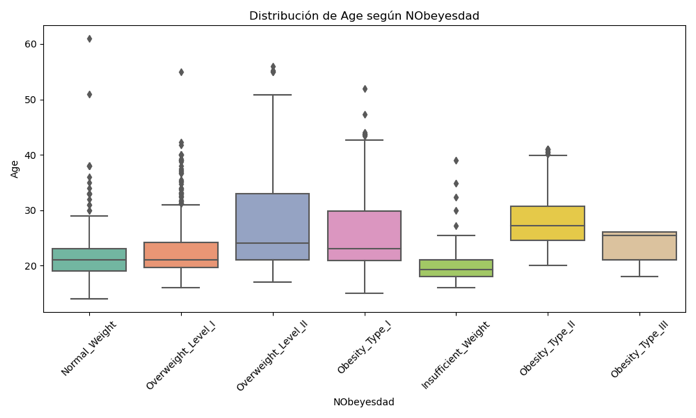
El objetivo del proyecto es Desarrollar un modelo de predicción de obesidad optimizado y validado, aplicando técnicas de exploración no supervisada y supervisada, y estructurarlo en una presentación completa y organizada. Se parte de un conjunto de datos que recopila información para estimar los niveles de obesidad en personas de México, Perú y Colombia, tomando en cuenta sus hábitos alimenticios y condición física. Incluye 17 atributos y 2111 registros, los cuales están clasificados mediante la variable de clase **NObesity** (Nivel de Obesidad). Esta variable permite categorizar los datos en las siguientes categorías: Peso Insuficiente, Peso Normal, Sobrepeso Nivel I, Sobrepeso Nivel II, Obesidad Tipo I, Obesidad Tipo II y Obesidad Tipo III. Del total de datos, el 77% fue generado de forma sintética utilizando la herramienta Weka y el filtro SMOTE, mientras que el 23% restante fue recolectado directamente de usuarios a través de una plataforma web1 el diccionario de datos se puede visualizar en la tabla 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nombre de variable** | **Rol** | **Tipo** | **Descripción** | **Valores perdidos** |
| Gender | Característica | Continuo |  | no |
| Age | Característica | Continuo |  | no |
| Height | Característica | Continuo |  | no |
| Weight | Característica | Continuo |  | no |
| family\_history\_with\_overweight | Característica | Binario | ¿Algún miembro de su familia ha sufrido o sufre de sobrepeso? | no |
| FAVC | Característica | Binario | ¿Consume alimentos altos en calorías con frecuencia? | no |
| FCVC | Característica | Entero | ¿Suele incluir vegetales en sus comidas? | no |
| NCP | Característica | Continuo | ¿Cuántas comidas principales realiza al día? | no |
| CAEC | Característica | Categórico | ¿Consume algún alimento entre comidas? | no |
| SMOKE | Característica | Binario | ¿Fuma? | no |
| CH2O | Característica | Continuo | ¿Cuánta agua bebe diariamente? | no |
| SCC | Característica | Binario | ¿Controla las calorías que consume diariamente? | no |
| FAF | Característica | Continuo | ¿Con qué frecuencia realiza actividad física? | no |
| TUE | Característica | Entero | ¿Cuánto tiempo dedica al uso de dispositivos tecnológicos como celular, videojuegos, televisión, computadora, entre otros? | no |
| CALC | Característica | Categórico | ¿Con qué frecuencia consume alcohol? | no |
| MTRANS | Característica | Categórico | ¿Qué medio de transporte utiliza habitualmente? | no |
| NObeyesdad | Target | Categórico | Nivel de obesidad | no |

Para iniciar el proceso, se ejecuta un análisis de varianza para determinar si almenos uno de los grupos es estadísticamente diferente en el promedio de todos los demás. Posterior a esto se genera una prueba de Dunn de comparaciones múltiples para estimar que grupo es mayor a los demás.

Esto permitió hacer un análisis descriptivo de la información y determinar univariadamente la capacidad de los predictores para describir el comportamiento de NObeyesdad Y.

Hay diferencias estadísticamente significativas en cada uno de los niveles de la variable NObeyesdad (F77,9, p: 0) y al comparar, todos los niveles reportados de obesidad tienen promedios de edad diferentes.



Desde la talla, solo la obesidad tipo II muestra tener tallas con diferencias estadísticamente significativas con respecto al resto del grupo (F:38.43 P 0)

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Cada uno de los niveles de la variable peso tiene comportamientos diferentes según el nivel de obesidad como es de esperarse. En este sentido esta variable se postula como un evidente predictor. Posiblemente sea conveniente no tenerla en consideración porque parece ser una variable que está indicando el nivel de obesidad de una forma muy marcada.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

No se evidencian diferencias estadísticamente significativas con respecto a la inclusión de vegetales en la alimentación con respecto al nivel de obesidad

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Las personas con peso insuficiente parecen reportar más comidas en el día, con contenido vegetariano

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Niveles bajos de obesidad, peso insuficiente y obesidad tipo III muestran mayor consumo de agua en comparación con los demás grupos. Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Las personas con peso insuficiente y obesidad tipo II muestran mayor tasa de actividad física en comparación con las demás personas.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Las personas con obesidad tipo I y tipo III muestra menor tiempo de exposición a pantalla con fines de entretenimiento

Gráfico, Gráfico de barras, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Metodología: Documentación de cada paso del proceso, desde la exploración no supervisada hasta el entrenamiento supervisado y la optimización del modelo.**

1. Teniendo en cuenta que la variable dependiente es cualitativa, categórica y posiblemente ordinal, se lleva a cabo un análisis de su distribución la cual se especificó al inicio en la introducción. Esta, tiene distribución uniforme.

Gráfico, Gráfico de barras

Descripción generada automáticamente

1. Se analiza la correlación entre las variables independientes, para conocer las consecuencias probables de multicolinealidad y en cualquier caso, reducir el número de variables dependientes para evitar dimensionalidad y sobre parametrización de cualquier estructura de modelo. La prueba de esfericidad de Barttlet y el índice de Kaiser Meyer y Olkim muestra que la matriz de correlación cuenta con adecuación muestral y esfericidad aunque débil. Posiblemente no exista correlación fuerte entre las variables. Con un KMO tan bajo, correr un PCA no parece indicado.

|  |  |
| --- | --- |
| **Gráfico, Gráfico de rectángulos  Descripción generada automáticamente** | Índice KMO: 0.4451  Chi-cuadrado Bartlett: 1679.3981, df: 28.0, p-value: 0.0000 |

1. Se cuenta con 8 variables y al intentar reducirlas mediante componentes principales, 4 componentes logran explicar el 77% de la varianza. Usando el método del codo se escoge 4 componentes.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

1. El circulo unitario muestra una baja correlación negativa entre edad y TUE (tiempo al celular con fines de entretenimiento) indicando que a menor edad mayor exposición y sedentarismo. En este análisis muestra como el peso esta correlacionado positivamente con la inclusión de vegetales en la dieta. El consumo de agua y la talla también muestran correlaciones positivas entre sí. Finalmente el número de comidas al día y la actividad física muestran correlación positiva igualmente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

1. Cada uno de los componentes goza de dos propiedades: distribución simétrica y hay ausencia de correlación entre ellos.

Gráfico, Histograma

Descripción generada automáticamente

1. Según la tabla 2, que muestra la carga de cada variable al factor, se interpreta que el componente 1 atrae peso, talla, consumo de agua u número de comidas al día indicando hábitos alimenticios, la componente 2 es edad, la 3 es ingesta de vegetales y sedentarismo tecnológico mientras que el PCA 4 hace referencia a actividad física y ejercicio.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PCA\_1** | **PCA\_2** | **PCA\_3** | **PCA\_4** |
| **Height** | 0.598077 |  |  |  |
| **Weight** | 0.503661 |  |  |  |
| **NCP** | 0.333123 |  |  |  |
| **CH2O** | 0.384655 |  |  |  |
| **Age** |  | 0.603802 |  |  |
| **FCVC** |  |  | 0.830807 |  |
| **TUE** |  |  | 0.045231 |  |
| **FAF** |  |  |  | 0.647829 |

1. Usando las variables componente, se prueban dos algoritmos de conglomeración, por un lado Jerárquico probando soluciones de 2 a 9 clusters y por otro lado K Medias generando soluciones entre 2 y 9 conglomerados. Se escoge la alternativa de K medias en 9 conglomerados al tener mejores indicativos del índice de silueta.

|  |  |
| --- | --- |
| **Gráfico, Gráfico de dispersión  Descripción generada automáticamente** | Gráfico, Gráfico circular  Descripción generada automáticamente |

1. Se evidencia una distribución simétrica de los sujetos entre los clústeres.
2. Se comparan 5 modelos en general, la precisión mas alta es de 6

|  |  |
| --- | --- |
| Modelo | Precisión |
| K-Nearest Neighbors | 0.61 |
| Random Forest | 0.59 |
| Support Vector Classifier | 0.58 |
| Decision Tree | 0.48 |
| Logistic Regression | 0.45 |

**Resultados:**

Partiendo del modelo K-Nearest Neighbors que en el proceso anterior tuvo una precisión del 0.61 se usa la variable cluster como group by y se obtiene una mejora de 1 punto de precisión.

**Imagen que contiene Diagrama

Descripción generada automáticamente**

**Conclusiones y Recomendaciones:**

La variabilidad de los datos puede generar procesos de confusión y alucinación en los modelos de machine learning. El uso de clusters permite entrenar el mismo algoritmo pero especializado en sujetos similares elevando la presión en la predicción.

Los cluster generan celdas de entrenamiento que permite a los modelos una menor concentración en la variabilidad para elevar la precisión y esto facilita resultados más precisos aunque no necesariamente más estables2

**Referencias**

1. Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. <https://doi.org/10.24432/C5H31Z>.
2. Xu, R., & Wunsch, D. (2009). *Clustering*. John Wiley & Sons.