Estudio estadístico de factores de riesgo asociados con la obesidad

1. Introducción

Este trabajo se propone examinar en profundidad los factores de riesgo asociados con la obesidad en una población específica, con el objetivo de contribuir al desarrollo de estrategias de prevención efectivas y personalizadas.

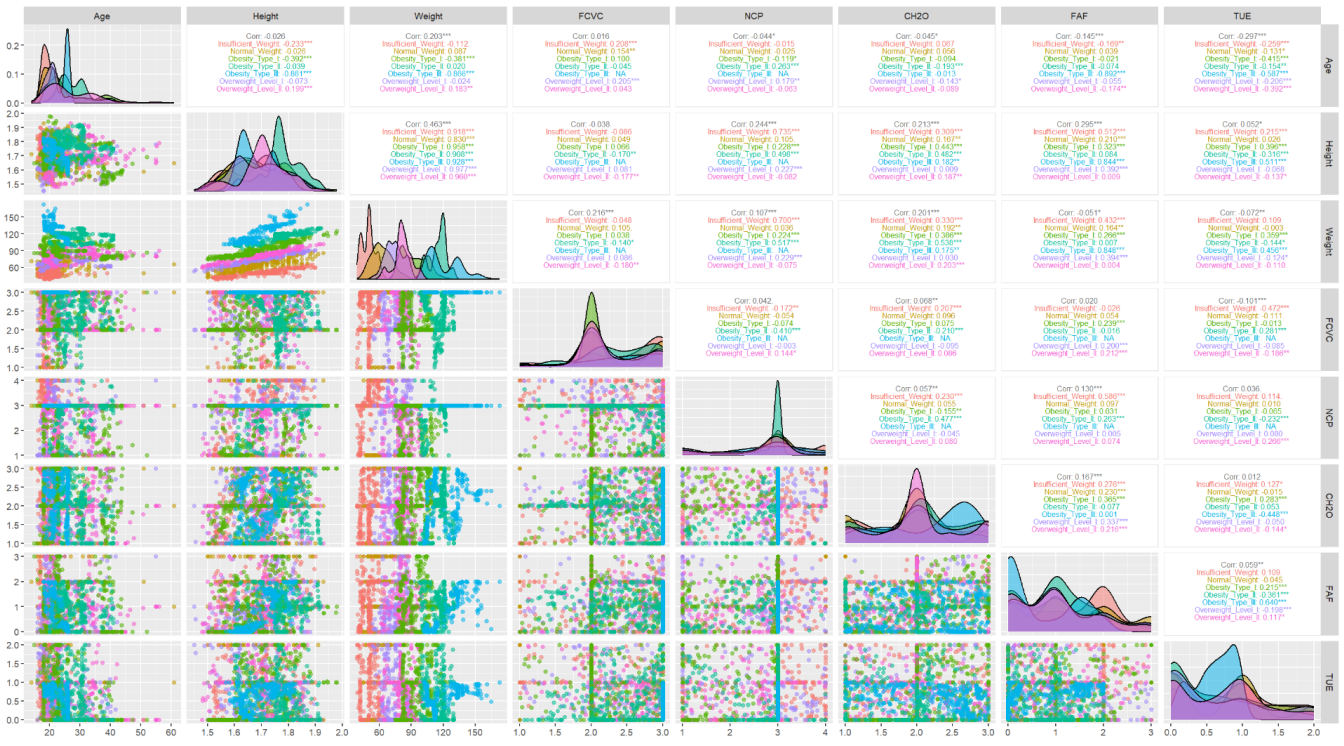
A lo largo de este análisis, exploraremos diversas variables que potencialmente influyen en el desarrollo de la obesidad, desde antecedentes familiares hasta hábitos de vida cotidianos con el fin de poder dar explicaciones detalladas acerca de cuáles son las causas de tal característica.

Mediante la aplicación de técnicas estadísticas avanzadas vistas en clase buscaremos identificar patrones, relaciones y grupos de riesgo que nos aporten la mayor información posible acerca de cuáles son los principales factores de riesgo asociados con la obesidad en la población estudiada.

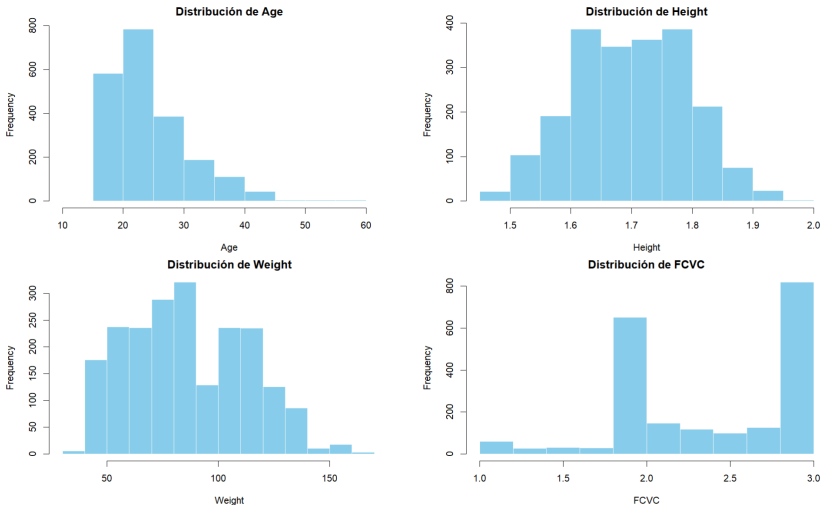
Descripción de las variables disponibles

Height	Continua	Metros
Weight	Continua	Kilogramos
family_history_with_overweight	Binaria	¿Algún miembro de la familia ha sufrido o sufre de sobrepeso?
FAVC	Binaria	¿Comes frecuentemente alimentos altos en calorías?
FCVC	Entera	¿Sueles comer vegetales en tus comidas?
NCP	Continua	¿Cuántas comidas principales tienes al día?
CAEC	Categórica	¿Comes algo entre comidas?
SMOKE	Binaria	¿Fumas?
CH2O	Continua	¿Cuánta agua bebes diariamente?
SCC	Binaria	¿Monitoreas las calorías que consumes diariamente?
FAF	Continua	¿Con qué frecuencia realizas actividad física?
TUE	Entera	¿Cuánto tiempo usas dispositivos tecnológicos como teléfono móvil, videojuegos, televisión, ordenador y otros?
CALC	Categórica	¿Con qué frecuencia bebes alcohol?
MTRANS	Categórica	¿Qué medio de transporte sueles utilizar?
Obesity_level	Categórica	Nivel de obesidad

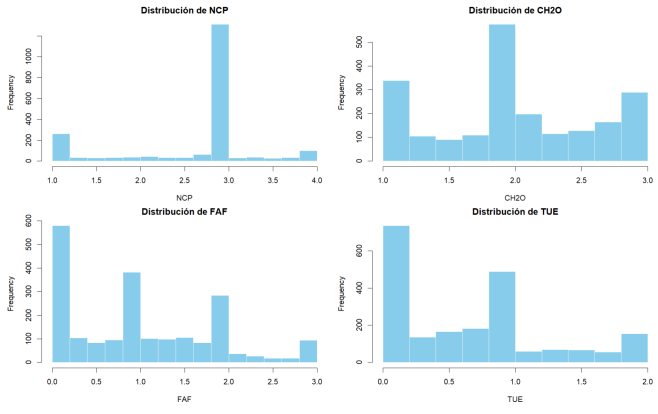
Breve estudio de la distribución y correlación de los datos



Visualización de la correlación de las variables numéricas con las distintas clases de la variable target.



Distribución de las 4 primeras variables numéricas.



Distribución de las últimas 4 variables numéricas.

Tras realizar un breve estudio de la naturaleza de los datos, podemos observar que hay ciertas variables que siguen una distribución ciertamente normal y otras las cuales se mueven en un rango de (1-3) lo que nos hace pensar que las variables han sido tratadas por medio de un `MinMaxScaler` **.** Esto nos facilita la creación de los modelos siguientes puesto que no es necesario realizar ningún tratamiento a los datos.

2. Metodología

Para analizar los factores de riesgo asociados con la obesidad, utilizaremos las técnicas vistas en clase de análisis multivariante adaptándonos a la naturaleza de nuestros datos, es por eso que las técnicas usadas serán:

- **Análisis de Componentes Principales (PCA):** Para reducir la dimensionalidad de los datos e identificar patrones principales.
- **Análisis Discriminante Lineal (LDA):** Para clasificar los datos proyectándolos en un espacio de menor dimensión que maximice la separación entre los distintos tipos de obesidad.
- **Partial Least Squares (PLS):** Para identificar combinaciones de factores de riesgo que explican la mayor varianza en los niveles de obesidad.
- **Análisis de Clúster:** Para identificar grupos de individuos con características similares en cuanto a factores de riesgo.
- **Clasificador Naive Bayes:** Para clasificar según la probabilidad de que cada muestra pertenezca a las distintas clases.

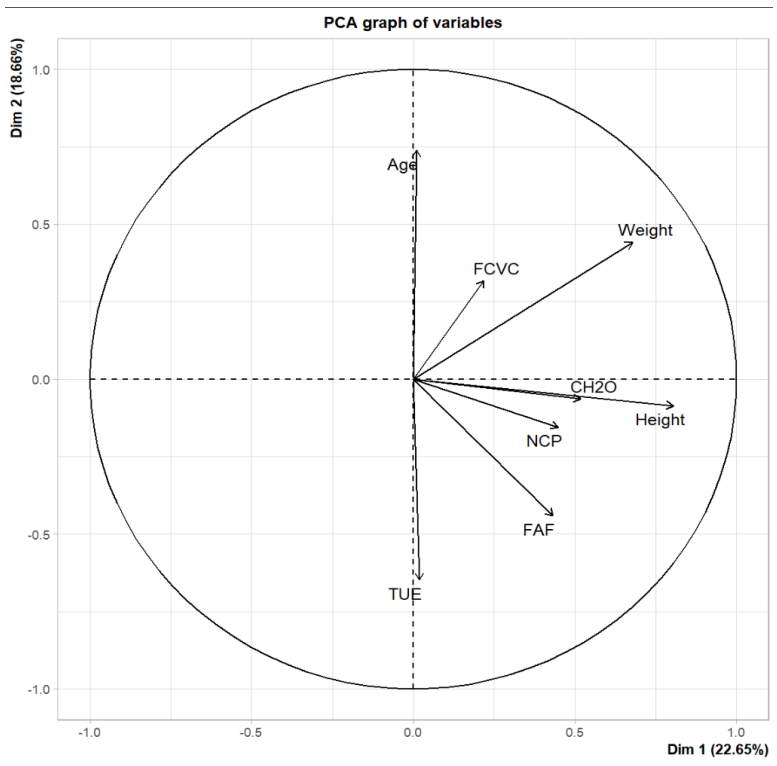
2.1 Análisis por medio de Componentes Principales (PCA)

Como hemos visto en el breve estudio de los datos realizado anteriormente, tenemos gran cantidad de datos no numéricos, además de que la variable target que nos interesa, `Obesity_level` , se trata de una variable categórica, por lo que, en este análisis, PCA no podrá encontrar componentes que maximicen la separación entre clases y no nos puede proporcionar información directa sobre la relación entre las variables predictoras y la variable categórica.

Es por eso que vamos a proceder a aplicar un PCA con variables cualitativas como suplementarias tal y como se ha comentado en clase. Para ello seguimos los siguientes pasos:

1. En primer lugar, realizamos una estandarización de las variables numéricas para que todas se encuentren en una misma escala, evitando así que aquellas con mayores magnitudes dominen o influyan desproporcionadamente en el análisis.
2. Procedemos a transformar todas las variables categóricas a tipo factor para que PCA pueda representarlas.
3. Tras esto, finalizamos aplicando PCA a las variables numéricas añadiendo las variables suplementarias.

La salida correspondiente es:



2.1.1. Interpretación de las Dimensiones

- Dimensión 1 (Dim 1): Representa aproximadamente el 22.6% de la varianza total.
- Dimensión 2 (Dim 2): Representa aproximadamente el 18.7% de la varianza total.

Estas dos dimensiones juntas explican el 41.3% de la variabilidad en los datos numéricos, lo que sugiere una representación no muy robusta de la estructura subyacente de los datos, aunque debemos tener en cuenta que las variables categóricas se tratan como suplementarias en este análisis, por lo cual nuestras conclusiones no van a ser del todo concluyentes.

2.1.2. Análisis de Variables

2.1.2.1. Longitud de las Flechas

- Variables con flechas largas: `Weight`, `Height` y `Age` tienen una fuerte influencia en la variabilidad de los datos.
- Variables con flechas cortas: `FCVC` (Consumo de vegetales) y `CH2O` (Consumo de agua) tienen menos impacto en la variabilidad general.

2.1.2.2. Variables Categóricas como Suplementarias

Las variables categóricas, incluida `Obesity_level`, se han tratado como suplementarias en el análisis PCA por lo que no podemos establecer relaciones causales directas, pero si nos permite mantener la integridad del análisis PCA que está diseñado principalmente para variables numéricas.

- Los niveles de obesidad han sido incluidos como variable suplementaria en el análisis, lo que significa que no han influido en la construcción de las componentes principales.

2.1.3. Implicaciones para el Análisis

Basándonos en este análisis PCA, podemos inferir que:

- Las variables numéricas explican una parte no muy significativa de la variabilidad en los datos (41.3%), por lo que no son suficientes por sí solas para explicar completamente los patrones de obesidad.
- La edad y las medidas antropométricas (peso y altura) son los factores numéricos más influyentes en la estructura de los datos.
- Los hábitos alimenticios y de hidratación muestran menor variabilidad en este análisis numérico.

2.1.4. Limitaciones y Consideraciones Futuras

Es importante considerar que:

- Al tratar las variables categóricas como suplementarias, perdemos parte de la información sobre las relaciones directas con la obesidad.
- Sería recomendable complementar este análisis con técnicas específicas para datos mixtos o métodos supervisados que puedan incorporar mejor la naturaleza categórica de algunas variables. Es por eso que vamos a proceder a realizar más técnicas de análisis multivariante con el objetivo de poder responder de la forma más precisa posible al problema planteado.

2.2. Análisis Discriminante Lineal (LDA)

Puesto que en el análisis anterior vimos que la gran mayoría de variables eran categóricas y LDA trabaja con medias, varianzas y covarianzas que no pueden calcularse directamente para variables categóricas, solo podremos trabajar con la variable target `Obesity_level` y las variables numéricas.

Esta decisión es crucial ya que, como bien se ha comentado en clase en numerosas ocasiones, tratar variables categóricas con medidas de variables numéricas mediante ciertas transformaciones (OneHot Encoding, Ordinal Encoding, etc..) es un error muy grave que no debe ocurrir en la fase de creación de modelo ya que repercutiría muy negativamente a la hora de generalizar con datos nunca antes vistos.

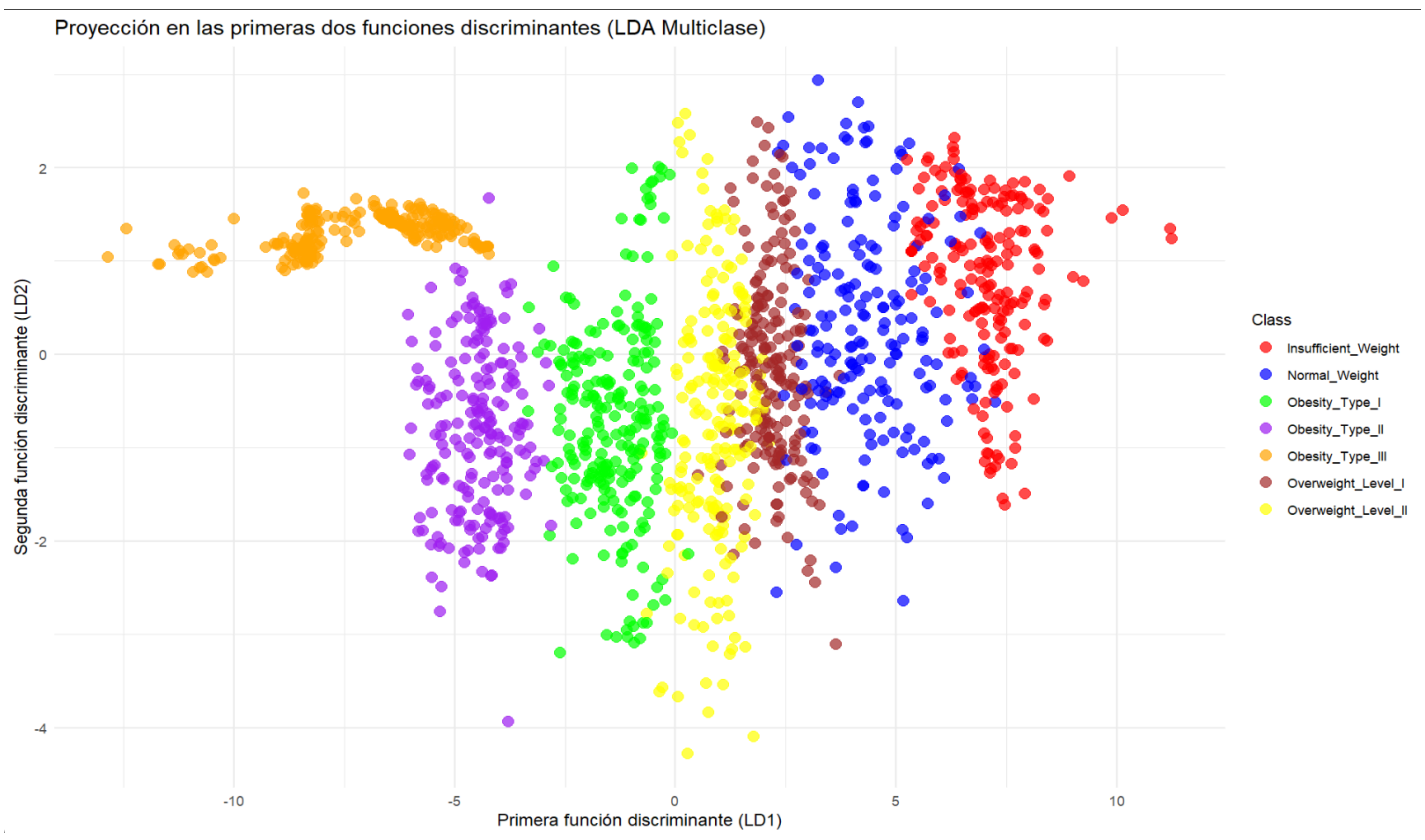
Es por eso que procedemos a realizar LDA siguiendo los siguientes pasos:

1. En primer lugar, eliminamos del conjunto de datos todas aquellas variables que no sean numéricas, además de mantener la variable objetivo `Obesity_level`.
2. Después, procedemos a dividir el conjunto de datos en 2 partes, *train* (70% del dataset) y *test* (30% del dataset), con el objetivo de poder poner a prueba la generalización del modelo con datos nunca antes vistos.
3. A continuación, procedemos a crear el modelo y proyectar los resultados.
4. Finalmente, ponemos a prueba el rendimiento del modelo calculando la tasa de *accuracy* en el conjunto de test, así como visualizar la matriz de confusión correspondiente al output.

2.2.1. Interpretación del Análisis LDA

El análisis LDA busca encontrar combinaciones lineales de las variables predictoras que mejor separen las diferentes clases de obesidad. La visualización proporcionada a continuación muestra la proyección de las observaciones en el espacio discriminante.

Resultados de la separación de clases del modelo



Resultados del *accuracy* del modelo para datos no vistos con anterioridad

```
> accuracy ← sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(paste("Precisión global:", round(accuracy, 2)))
 "Precisión global: 0.85"
```

Resultado de la matriz de confusión relativa al conjunto de datos no vistos con anterioridad

Actual / Predicted	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	83	7	0	0	0	0	0
Normal_Weight	8	62	0	0	0	21	1
Obesity_Type_I	0	0	91	3	0	0	2
Obesity_Type_II	0	0	0	89	1	0	0
Obesity_Type_III	0	0	0	10	89	0	0
Overweight_Level_I	0	0	0	0	0	66	15
Overweight_Level_II	0	0	1	0	0	27	58

2.2.2. Discusión de los resultados

El análisis LDA muestra resultados notables:

- **Precisión Global:** El modelo alcanza una precisión del 85% en el conjunto de test, lo que indica un buen rendimiento general en la clasificación de los niveles de obesidad.
- Destaca especialmente en la clasificación de casos extremos: peso insuficiente (83 casos correctos) y obesidad tipo III.
- Las principales confusiones se dan entre categorías adyacentes, especialmente en los niveles de sobrepeso.

La visualización muestra una clara separación entre grupos, aunque hay algunas zonas de solapamiento entre categorías cercanas.

2.2.3. Análisis de la Matriz de Confusión

La matriz de confusión revela varios patrones importantes:

- **Peso Insuficiente:** Excelente clasificación con 83 casos correctos y solo 7 confusiones con peso normal.
- **Peso Normal:** Buena clasificación, pero con algunas confusiones, principalmente con Sobrepeso Nivel I (21 casos).
- **Obesidad Tipo I, II y III:** Muy buena discriminación entre estos niveles, con pocas confusiones entre categorías adyacentes.
- **Niveles de Sobrepeso:** Mayor dificultad en la clasificación, con confusiones notables entre Sobrepeso Nivel I y II.

2.2.4. Fortalezas y Limitaciones

El análisis LDA presenta las siguientes características:

2.2.4.1. Fortalezas

- Alta precisión en la clasificación de casos extremos (peso insuficiente y obesidad tipo III).
- Buena capacidad para discriminar entre diferentes tipos de obesidad.
- Modelo interpretable y computacionalmente eficiente.

2.2.4.2. Limitaciones

- Solo puede utilizar variables numéricas, lo que excluye información potencialmente valiosa de variables categóricas.
- Mayor dificultad para discriminar entre categorías adyacentes, especialmente en los niveles de sobrepeso.
- Asume relaciones lineales entre las variables, lo que puede no ser siempre el caso en datos de obesidad.

2.2.5. Conclusiones del Análisis LDA

El análisis LDA ha demostrado ser una herramienta efectiva para la clasificación de niveles de obesidad, especialmente cuando se trabaja con variables numéricas. Sin embargo, la exclusión de variables categóricas y las confusiones entre categorías adyacentes sugieren que podría ser beneficioso complementar este análisis con otros métodos que puedan incorporar variables categóricas y capturar relaciones no lineales en los datos. Para ello procedemos con más técnicas de análisis multivariante.

2.3. Análisis por medio de mínimos cuadrados (PLS)

2.3.1. ¿Por qué no es conveniente realizar PLS?

La regresión por mínimos cuadrados parciales (PLS) no es una técnica apropiada para este conjunto de datos por varias razones fundamentales:

- **Naturaleza de la variable objetivo:** PLS está diseñado para predecir variables continuas, mientras que en nuestro caso, la variable `Obesity_level` es categórica con múltiples niveles discretos.
- **Funcionamiento de PLS:** Esta técnica busca maximizar la covarianza entre las variables predictoras X y la variable respuesta Y, asumiendo una relación lineal continua entre ellas. Este enfoque no es adecuado para problemas de clasificación multiclase.

- **Interpretación de resultados:** Los coeficientes y componentes de PLS se interpretan en términos de cambios continuos en la variable respuesta, lo cual no tiene sentido cuando tratamos con categorías discretas de obesidad.

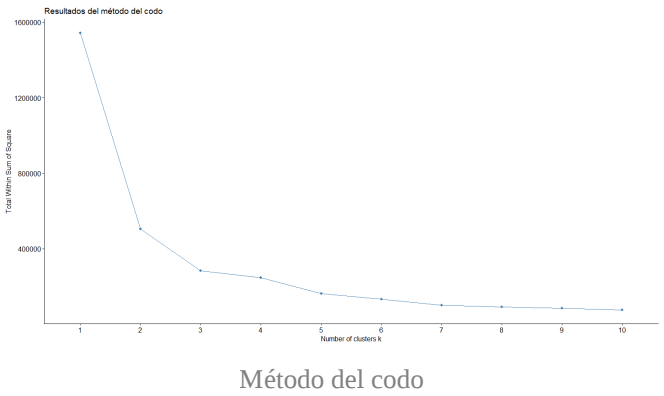
En su lugar, es más apropiado utilizar técnicas específicamente diseñadas para clasificación como las que hemos visto anteriormente (LDA, Naive Bayes, Clusters) o considerar otros métodos que pueden manejar adecuadamente problemas de clasificación multiclase.

2.4. Análisis de Cluster

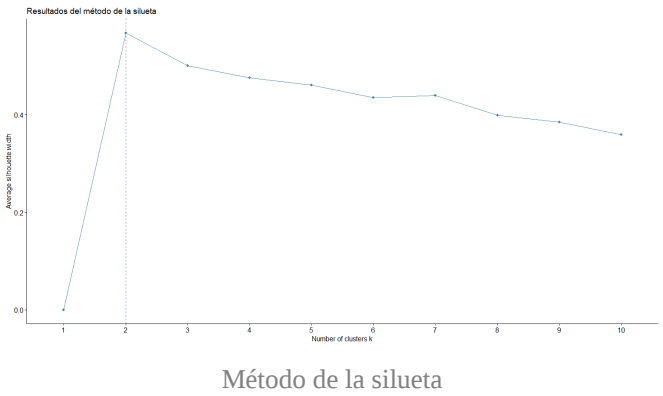
El análisis de clúster nos permite identificar grupos de individuos con características similares en función de las variables disponibles. Esto resulta particularmente útil para identificar patrones en los factores de riesgo asociados con la obesidad y explorar relaciones entre las variables.

En primer lugar, convertimos los datos categóricos (o binarios) a numéricos para poder medir las distancias.

Una vez hecho esto, buscamos el número de clusters óptimo para este dataset. Usamos dos métodos: el del codo y el de la silueta.



Método del codo



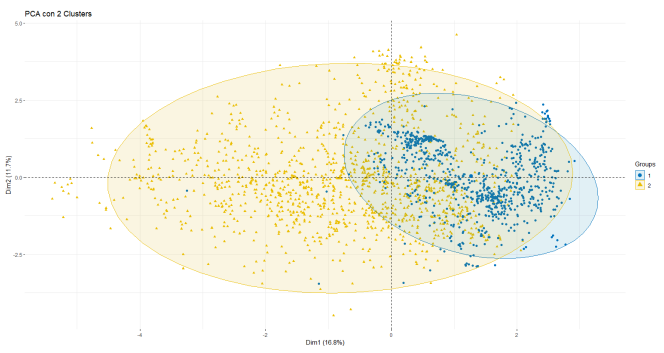
Método de la silueta

En el caso del método de la silueta, claramente indica que el número óptimo de clusters es 2, mientras que en el caso del método del codo es algo más ambiguo. Decidimos optar por 3 clusters, ya que la diferencia con 4 es mínima, y así se podrían comparar de forma efectiva ambos resultados.

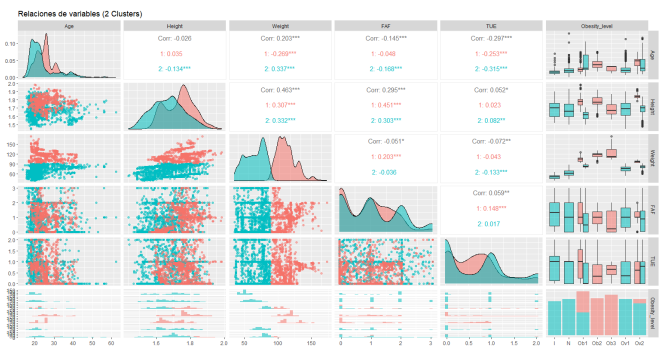
2.4.1. Visualización de Clusters

Para poder visualizar los clusters, las variables se han proyectado mediante PCA a un espacio bidimensional. Los clusters se han calculado sin usar la variable `Obesity_level`, ya que es nuestro target. Cabe destacar que en `Obesity_level` las categorías están desordenadas, apareciendo `Obesity` antes que `Overweight`.

Con 2 clusters, las agrupaciones quedan así:



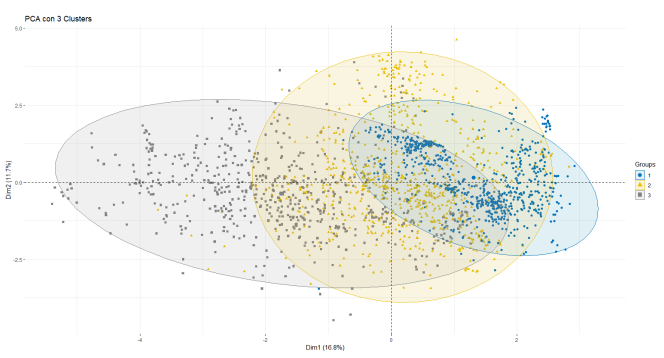
PCA con clusters



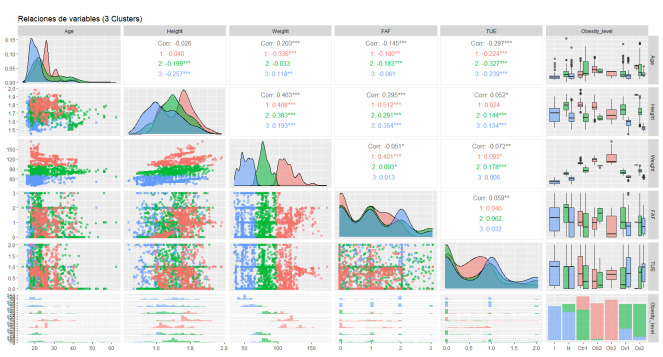
GGPairs con las variables que mejor se visualizan

Se puede ver que el grupo 1 (obeso) está bastante concentrado, mientras que el grupo 2 (desde insuficiente hasta sobrepeso de tipo 2) está mucho más disperso. En los histogramas se puede apreciar perfectamente como el peso está dividido en 2 grupos que apenas se solapan, cosa que coincide con lo observado en el gráfico de barras. El resto de variables, no mostradas aquí por cuestiones de visibilidad, no tienen tanta correlación o no se dividen tan bien.

Con 3 clusters obtenemos resultados muy similares:



PCA con clusters



GGPairs con las variables que mejor se visualizan

Los grupos actuales son:

1. Obesidad de tipo 2 y 3
2. Sobrepeso y obesidad tipo 1
3. Peso insuficiente y normal

2.4.2. Causas de riesgo

Comparando cómo los clusters dividen las distintas variables, podemos ver que las que mejor separadas están, y por tanto más relevancia han tenido a la hora de dividir, son `Age`, `Height` y `Weight`, coincidiendo con las conclusiones obtenidas en PCA y LDA.

2.5. Clasificador Naive Bayes

Se trata de un clasificador probabilísitco en el que asignaremos a una muestra la clase que maximice la probabilidad posterior, que es la probabilidad de que la clase *c* asignada sea la correcta para la muestra *x* observada.

Usando el Teorema de Bayes podemos decir que la probabilidad posterior es igual a la probabilidad de que se diera la muestra x sabiendo que la muestra observada es de la clase c (dicho de otra forma, mide la verosimilitud de que la muestra pertenezca a la clase *c*) multiplicado por la probabilidad de la clase y dividido entre la probabilidad de observar la muestra.
Como en el denominador tenemos un término que no está en función de la clase, podemos quitarlo y simplificar la expresión.

$$\hat{c} = \arg\max_{c \in C} P(c|x) = \arg\max_{c \in C} \frac{P(x|c) \cdot P(c)}{P(x)} = \arg\max_{c \in C}\ P(x|c) \cdot P(c)$$

A la hora de crear un modelo de este tipo, tenemos que suponer independencia entre cada una de las variables de nuestro dataset para poder descomponer *P(x|c)* en en un productorio de probabilidades de que los distintos valores de una muestra sean verosímiles con la clase.

$$P(x|c) = \prod_{i=1}^{m} P(x_i|c) \cdot P(c)$$

Una vez sabemos la teoría detrás del modelo, cargamos los datos y definimos nuestras variables como numéricas o categóricas. Esto es muy importante de cara al modelo, ya que para las categóricas se fijará en las frecuencias de los valores y para las numéricas calcularemos la probabilidad con la función de densidad de una distribución normal. En este caso, viene a ser más bien abuso del lenguaje porque esto no es realmente una “probabilidad”: en variables continuas no nos fijamos en las frecuencias de los valores sino en la verosimilitud que tienen unos valores en la clase.

Un matiz que nos ha causado bastantes problemas son las columnas definidas como continuas que no son ni edad, peso o altura. Estas variables aparentan haber sido transformadas anteriormente, por ejemplo, es raro tener valores continuos para la frecuencia con la que se hace deporte. Los valores enteros 0,1,2 y 3 están muy repetidos. Por todo ello, hemos decidido redondear los valores de estas columnas para tratarlas como categóricas y que el modelo se fije en sus frecuencias, ya que así conseguimos un mejor accuracy.

Después hacemos la partición de train y test, nos quedamos con el 70% de las muestras para train y las restantes para test.

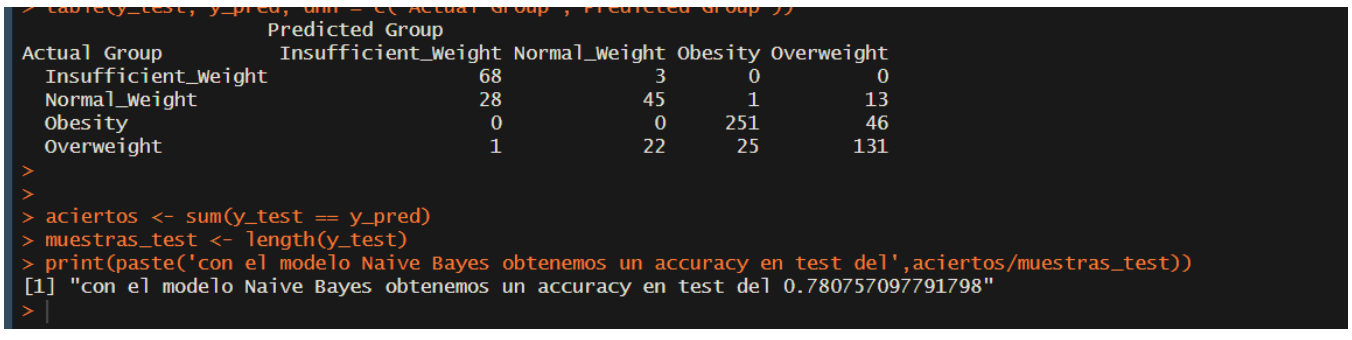
En Naive-Bayes suponemos que las variables continuas siguen una distribución normal, por lo tanto normalizamos a media 0 y error estándar 1 a aquellas variables que tratamos con as.numeric.

Hecho esto ya tendríamos los datos preparados para entrenar el modelo Naive bayes de la librería *e1071* con los datos de train, especificando que la variable objetivo es la de nivel de obesidad. Esta es la matriz de confusión probando el modelo con los datos de test.

```
> table(y_test, y_pred, dnn = c('Actual Group','Predicted Group'))
                    Predicted Group
Actual Group         Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweight_Level_I Overweight_Level_II
  Insufficient_Weight                  68            3             0               0                0                 0                  0
  Normal_Weight                        27           46             0               0                0                 7                  7
  Obesity_Type_I                        0            0            48              20                0                10                 28
  Obesity_Type_II                       0            0             8              90                0                 0                  0
  Obesity_Type_III                      0            0             1               0               92                 0                  0
  Overweight_Level_I                    1           10             8               0                0                54                 19
  Overweight_Level_II                   0           12            21               2                0                 7                 45
>
>
> aciertos <- sum(y_test == y_pred)
> muestras_test <- length(y_test)
> print(paste('con el modelo Naive Bayes obtenemos un accuracy en test del',aciertos/muestras_test))
[1] "con el modelo Naive Bayes obtenemos un accuracy en test del 0.69873817034703"
>
```

Las muestras cuyas predicciones han sido correctas se muestran en la diagonal. Si sumamos los valores de la diagonal y lo dividimos entre el número de muestras del conjunto test, obtenemos un accuracy del 0.7. Podemos observar que el modelo predice con mucha precisión las muestras de test pertenecientes a obesidad tipos 2 y 3 y poca precisión al predecir para muestras de las clases obesidad tipo 1, peso normal y sobrepeso nivel 2, en estos casos el accuracy es algo superior al 50%.

Aparte del propio análisis, hemos probado a tener un menor número de clases en la variable objetivo con el fin de obtener un mejor accuracy. Lo hemos hecho agrupando todas las muestras en peso infuficiente, peso normal, sobrepeso y obesidad, es decir, quitamos los tipos de obesidad y sobrepeso. Obtenemos un score notablemente más alto del 0.78. Si nuestra prioridad es la precisión y no nos importa centrarnos menos en los grados de obesidad y sobrepeso, optaríamos por tomar esta medida.



```
                    Predicted Group
Actual Group         Insufficient_Weight Normal_Weight Obesity Overweight
  Insufficient_Weight                  68             3       0          0
  Normal_Weight                        28            45       1         13
  Obesity                               0             0     251         46
  Overweight                            1            22      25        131
>
>
> aciertos <- sum(y_test == y_pred)
> muestras_test <- length(y_test)
> print(paste('con el modelo Naive Bayes obtenemos un accuracy en test del',aciertos/muestras_test))
[1] "con el modelo Naive Bayes obtenemos un accuracy en test del 0.780757097791798"
>
```

3. Conclusiones

Tras analizar los datos mediante diferentes metodologías estadísticas, podemos extraer las siguientes conclusiones sobre los factores de riesgo asociados a la obesidad:

3.1. Evaluación de las Metodologías

Análisis de Componentes Principales (PCA)

Este método nos permitió identificar las variables más relevantes en la variabilidad de los datos, destacando principalmente el peso, la altura y la edad como factores determinantes. Sin embargo, su naturaleza lineal limita la interpretación de relaciones más complejas.

Análisis Discriminante Lineal (LDA)

Confirmó los hallazgos del PCA, proporcionando una mejor separación entre las clases de obesidad. Su ventaja radica en la capacidad de discriminar entre grupos, aunque también está limitado por la suposición de linealidad.

Análisis de Clusters

Reveló patrones naturales en los datos, identificando principalmente dos o tres grupos distintivos de individuos. Este análisis fue particularmente útil para visualizar cómo se agrupan naturalmente los casos de obesidad, mostrando una clara separación entre individuos obesos y no obesos.

Naive Bayes

Proporcionó resultados satisfactorios en la clasificación (accuracy del 70%), especialmente en la identificación de casos de obesidad tipo 2 y 3. Su rendimiento mejoró significativamente (78%) al reducir las categorías de clasificación, aunque la suposición de independencia entre variables limita su precisión.

3.2. Factores de Riesgo Identificados

Basándonos en la convergencia de resultados de las diferentes metodologías, los principales factores de riesgo identificados son:

- **Variables Físicas:** El peso y la altura son los predictores más significativos, como se confirmó en todos los análisis realizados.
- **Factor Edad:** Demostró ser una variable importante en la clasificación y agrupación de casos de obesidad.
- **Patrones de Comportamiento:** Los análisis de clustering revelaron grupos distintivos basados en hábitos y características físicas.
- **Relación Peso-Altura:** Los análisis confirmaron que la relación entre el peso y la altura (como el índice de masa corporal, IMC) es un factor clave en la identificación de grupos de riesgo, especialmente en los análisis de clustering.
- **Interacción de Variables:** Aunque las metodologías como Naive Bayes asumen independencia, los resultados sugieren que existe interacción entre variables como peso, altura y edad, que influye significativamente en la clasificación de casos.
- **Diferenciación de Grupos:** Los análisis de clustering y LDA identificaron subgrupos específicos dentro de las categorías de obesidad, lo que destaca la importancia de considerar la heterogeneidad dentro de los casos al desarrollar estrategias de intervención.

3.3. Conclusión Final

La combinación de diferentes metodologías estadísticas nos ha permitido obtener una visión más completa y robusta del problema de la obesidad. Mientras que cada método tiene sus limitaciones individuales, su uso conjunto proporciona una comprensión más

profunda de los factores de riesgo. Los resultados sugieren que un enfoque multifactorial, considerando tanto variables físicas como comportamentales, es necesario para abordar efectivamente la prevención y tratamiento de la obesidad.

4. Bibliografía

Departamento de Matemática Aplicada UPM. (n.d.). *Clasificación Naive-Bayes*. https://dcain.etsin.upm.es/~carlos/bookAA/02.1_MetodosdeClasificacion-Naive-Bayes.html

¿Qué son los clasificadores Naive Bayes? (n.d.). IBM. Retrieved November 27, 2024, from https://www.ibm.com/es-es/topics/naive-bayes

MathWorks. (n.d.).
Introduction to K-Means Clustering https://www.mathworks.com/help/stats/k-means-clustering.html

Analytics Lane. (n.d.). Introducción al Análisis de Componentes Principales (PCA). Analytics Lane. 27 de noviembre de 2024, de https://www.analyticslane.com

r-bloggers**.** Cluster Analysis in R**.** April 20, 2021 by finnstats https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/

Datacamp.com .Principal Component Analysis in R Tutorial.Feb 13, 2023 https://www.datacamp.com/tutorial/pca-analysis-r

Rpubs.Análisis de Conglomerados (clusters).Gustavo Martínez-Valdes .2023-04-25 https://rpubs.com/gustavomtzv/903852

Rpubs.ANÁLISIS DE COMPONENTES PRINCIPALES (PCA) .Cristina Gil Martínez.Junio, 2018. https://rpubs.com/cristina_gil/pca