

PRA2 - Obsesidad en Latino América

Jinni Li

Junio 2024

Contents

1. Introducción	3
2. Metodología	3
3. Base de datos.	4
4. Análisis exploratorio del base de datos	5
4.1. Resumen de todos los variables en conjunto	5
4.2. Variables Numéricas	6
4.3. Variables Categóricas	9
4.3. Variables Binarias	10
5. Preprocesado de datos	13
5.1. Discretización	14
5.2. Factorización de variables categóricas	16
6. PCA	18
6.1. Componentes Principales	18
6.2. Relación entre las variables	24
7. Ramdon Forest	26
7.1. Discretización del resto de las variablas numéricas	26
7.2. Test de Phi y Cramer	28
7.3. Modelo	28
9. Conclusión	33

```
##  
## Attaching package: 'ggplot2'
```

```

## The following objects are masked from 'package:psych':
##
##      %+%, alpha

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##      filter, lag

## The following objects are masked from 'package:base':
##
##      intersect, setdiff, setequal, union

## corrplot 0.92 loaded

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##      combine

## The following object is masked from 'package:ggplot2':
##
##      margin

## The following object is masked from 'package:psych':
##
##      outlier

## Loading required package: lattice

##
## Attaching package: 'DescTools'

## The following objects are masked from 'package:caret':
##
##      MAE, RMSE

## The following objects are masked from 'package:psych':
##
##      AUC, ICC, SD

```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:randomForest':
##
##      combine

## The following object is masked from 'package:dplyr':
##
##      combine
```

1. Introducción

La obesidad es un desafío de salud pública en constante aumento en Colombia, Peru and Mexico. Según la Encuesta Demográfica y de Salud Familiar de Perú (ENDES) 2021, el 36,9 % de personas mayores de 15 años presentan sobrepeso, mientras que el 25,8% tiene obesidad. Estas cifras destacan la urgencia de abordar este problema de salud creciente en el país. Las tasas de obesidad son particularmente preocupantes en ciertos grupos demográficos, como los residentes del área urbana (26,9%) en comparación con el área rural (14,5%).

Los factores subyacentes que contribuyen a esta tendencia alarmante incluyen una dieta poco saludable, la falta de actividad física y factores genéticos. Estos elementos se combinan para crear un entorno propicio para el aumento de peso y la obesidad en la población peruana. La obesidad no solo afecta la salud física, sino que también está vinculada a una serie de enfermedades crónicas, como enfermedades cardíacas, accidentes cerebrovasculares, diabetes tipo 2 y ciertos tipos de cáncer. Además, puede tener un impacto negativo en la salud mental y la calidad de vida de los individuos afectados.

Sin embargo, esta situación presenta una oportunidad significativa para las empresas que operan en el sector de bienestar y fitness en Perú. A pesar de que muchas personas en Perú tienen recursos financieros limitados, existe una demanda creciente de productos y servicios de bienestar y salud asequibles en el país. Las empresas pueden enfocarse en ofrecer soluciones económicas que sean accesibles para una amplia gama de personas, como programas de ejercicio en el hogar, alimentos saludables a precios asequibles y servicios de asesoramiento a bajo costo. Para tener éxito en este sector en crecimiento, es crucial comprender el mercado local, ofrecer productos y servicios de calidad, desarrollar una estrategia de marketing efectiva, establecer asociaciones con empresas locales y brindar un servicio al cliente excepcional.

Además de comprender el mercado local y ofrecer productos y servicios de alta calidad, se necesita una comprensión más detallada de los posibles consumidores en el sector de bienestar y fitness en Perú. Para ello, es fundamental analizar en profundidad las características de los consumidores peruanos en relación con la salud y el bienestar. Aquí es donde entran en juego técnicas avanzadas de análisis de datos como el Análisis de Componentes Principales (PCA) y el árbol de decisión (En este caso Random Forest). Estas técnicas nos permiten reducir la dimensionalidad de nuestros datos, identificar patrones significativos y entender las relaciones subyacentes entre las diferentes variables. Y así, segmentar de manera más efectiva a nuestra población objetivo en grupos homogéneos con características similares. Esto nos brinda una visión más clara de los diferentes segmentos de mercado y nos ayuda a adaptar nuestras estrategias de marketing y productos para satisfacer las necesidades específicas de cada grupo.

Recursos: <https://www.gob.pe/institucion/minsa/noticias/634511-minsa-15-millones-de-personas-tienen-sobrepeso-y-obesidad> <https://m.inei.gob.pe/prensa/noticias/el-399-de-peruanos-de-15-y-mas-anos-de-edad-tiene-al-menos-una-comorbilidad-12903/>

2. Metodología

La metodología que vamos a aplicar en este proyecto siguen los siguientes pasos:

- **Análisis Exploratorio de Datos (EDA):** En esta primera fase, se realiza un análisis descriptivo para comprender mejor la distribución y las estadísticas resumidas de las variables. Esto incluye la identificación de tendencias, patrones y posibles relaciones entre las diferentes variables utilizando técnicas de visualización como diagramas de dispersión, histogramas.
- **Preprocesamiento de Datos:** Se limpian y organizan los datos para asegurar la integridad y la calidad del conjunto de datos.
- **Reducción de Dimensionalidad con PCA:** Se aplica el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de las variables, conservando la mayor cantidad posible de información en un espacio dimensional reducido. La selección del número óptimo de componentes principales o vectores singulares se realiza utilizando el método de “codos” (elbow)
- **La inclusión del modelo Random Forest** permite no solo predecir con alta precisión los niveles de obesidad, sino también identificar las variables más importantes que influyen en estas predicciones, como el peso, el género y los hábitos alimenticios.

Al seguir esta metodología detallada, se puede comprender mejor los factores que contribuyen al sobrepeso y la obesidad, desarrollar modelos predictivos precisos y generar recomendaciones efectivas para promover hábitos más saludables en la población.

3. Base de datos.

Dato extraído de: <https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

Se Incluye las siguientes variables:

Variables Categóricas:

- **Consumo de alimentos entre comidas (CAEC):** Indica la frecuencia con la que el individuo come entre las comidas principales, con opciones de No, A veces (Sometimes), Frecuentemente (Frequently) y Siempre (Always).
- **Consumo de alcohol (CALC):** Indica la frecuencia con la que el individuo consume alcohol, con opciones de No lo hago (I do not drink), A veces (Sometimes), Frecuentemente (Frequently) y Siempre (Always).
- **Medio de transporte habitual (MTRANS):** Representa el medio de transporte que el individuo utiliza regularmente, con opciones de Automóvil (Automobile), Motocicleta (Motorbike), Bicicleta (Bike), Transporte público (Public Transportation) y Caminata (Walking).
- **Nivel de obesidad (NObesidad):** Indica el nivel de obesidad basado en el índice de masa corporal, con categorías de Bajo peso (Underweight Less than 18.5), Normal (Normal 18.5 to 24.9), Sobrepeso (Overweight 25.0 to 29.9), Obesidad I (Obesity I 30.0 to 34.9), Obesidad II (Obesity II 35.0 to 39.9) y Obesidad III (Obesity III Higher than 40).

Variables Binarias:

- **Género (Gender):** Esta variable nos indica el género del individuo, diferenciando entre hombre (Male) y mujer (Female).
- **Historial familiar de sobrepeso (family_history_with_overweight):** Indica si hay antecedentes familiares de sobrepeso, con respuestas de Sí (Yes) o No (No).
- **Consumo frecuente de alimentos altos en calorías (FAVC):** Nos muestra si el individuo consume regularmente alimentos con alto contenido calórico, con opciones de Sí (Yes) o No (No).

- Fumador (SMOKE): Indica si el individuo fuma, con respuestas de Sí (Yes) o No (No).
- Monitoreo de calorías consumidas (SCC): Indica si el individuo monitorea las calorías que consume diariamente, con respuestas de Sí (Yes) o No (No).

Variables Numéricas:

- Edad (Age): Representa la edad del individuo, mostrando un rango de valores numéricos.
- Altura (Height): Indica la estatura del individuo, expresada en metros.
- Peso (Weight): Representa el peso del individuo, expresado en kilogramos.
- Número de comidas principales al día (NCP): Representa la cantidad de comidas principales que el individuo tiene diariamente, con valores numéricos.
- Consumo de verduras (FCVC): Representa la frecuencia de consumir verduras en el día a día, con valores numéricos.
- Consumo de agua diario (CH2O): Representa la cantidad de agua que el individuo consume diariamente, con valores numéricos.
- Frecuencia de actividad física (FAF): Representa la frecuencia con la que el individuo realiza actividad física, con valores numéricos.
- Tiempo de uso de dispositivos tecnológicos (TUE): Indica el tiempo que el individuo pasa utilizando dispositivos tecnológicos como teléfono celular, videojuegos, televisión, computadora, entre otros, con valores numéricos.

4. Análisis exploratorio del base de datos

4.1. Resumen de todos los variables en conjunto

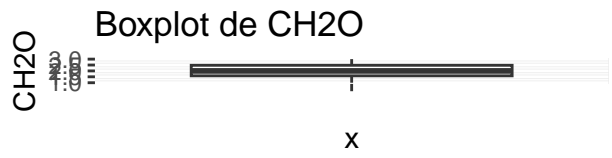
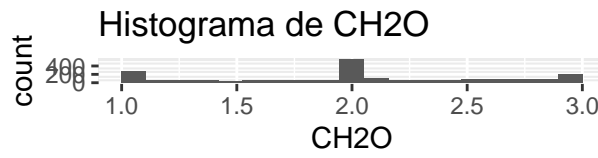
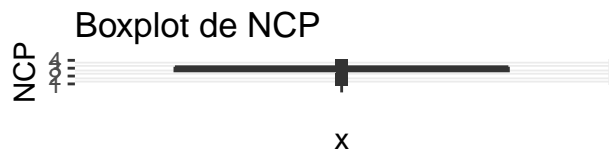
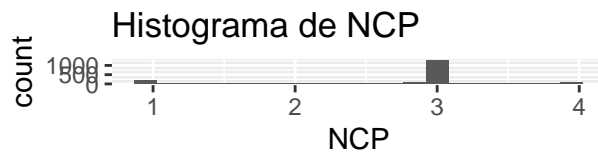
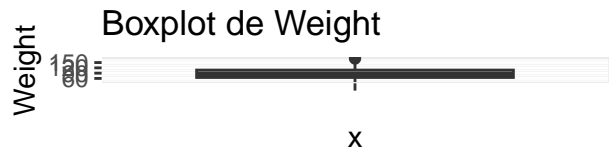
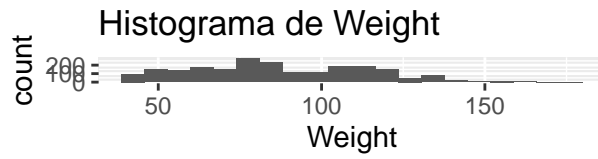
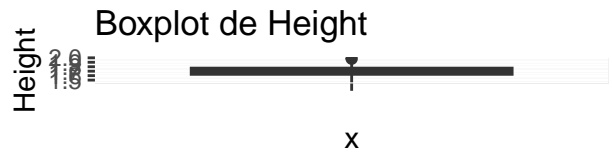
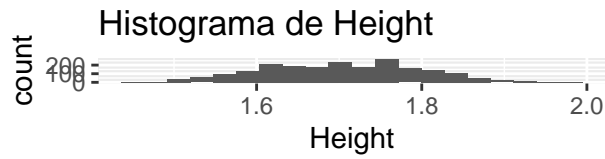
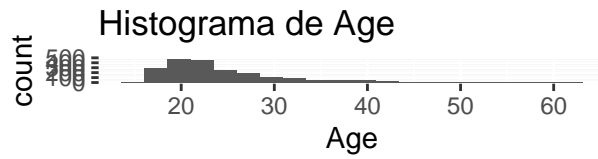
Para poder resumir las variables, se utiliza el comando “describe()” del paquete “psych”. Estos permiten obtener diferentes estadísticos que pueden ser útiles conjuntamente con un análisis exploratorio de las variables un poco más exhaustivo.

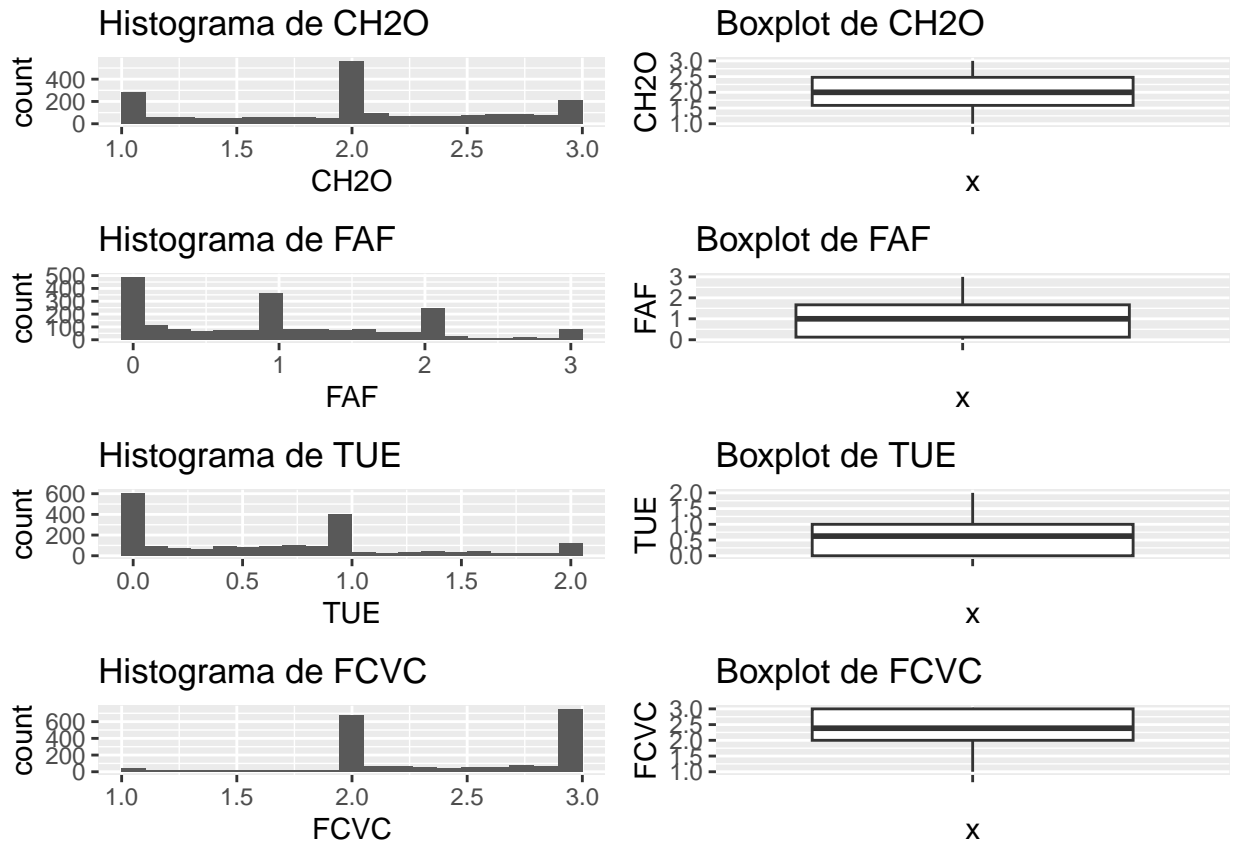
##	vars	n	mean	sd	median	trimmed	mad
## Gender*	1	2111	1.51	0.50	2.00	1.51	0.00
## Age	2	2111	24.31	6.35	22.78	23.34	4.78
## Height	3	2111	1.70	0.09	1.70	1.70	0.10
## Weight	4	2111	86.59	26.19	83.00	85.82	32.22
## family_history_with_overweight*	5	2111	1.82	0.39	2.00	1.90	0.00
## FAVC*	6	2111	1.88	0.32	2.00	1.98	0.00
## FCVC	7	2111	2.42	0.53	2.39	2.46	0.57
## NCP	8	2111	2.69	0.78	3.00	2.77	0.00
## CAEC*	9	2111	3.67	0.78	4.00	3.87	0.00
## SMOKE*	10	2111	1.02	0.14	1.00	1.00	0.00
## CH2O	11	2111	2.01	0.61	2.00	2.01	0.67
## SCC*	12	2111	1.05	0.21	1.00	1.00	0.00
## FAF	13	2111	1.01	0.85	1.00	0.94	1.19
## TUE	14	2111	0.66	0.61	0.63	0.59	0.72
## CALC*	15	2111	3.63	0.55	4.00	3.70	0.00
## MTRANS*	16	2111	3.37	1.26	4.00	3.55	0.00

## NObeyesdad*	17	2111	4.02	1.95	4.00	4.02	2.97
##	min	max	range	skew	kurtosis	se	
## Gender*	1.00	2.00	1.00	-0.02	-2.00	0.01	
## Age	14.00	61.00	47.00	1.53	2.81	0.14	
## Height	1.45	1.98	0.53	-0.01	-0.57	0.00	
## Weight	39.00	173.00	134.00	0.26	-0.70	0.57	
## family_history_with_overweight*	1.00	2.00	1.00	-1.64	0.70	0.01	
## FAVC*	1.00	2.00	1.00	-2.40	3.74	0.01	
## FCVC	1.00	3.00	2.00	-0.43	-0.64	0.01	
## NCP	1.00	4.00	3.00	-1.11	0.38	0.02	
## CAEC*	1.00	4.00	3.00	-2.13	3.06	0.02	
## SMOKE*	1.00	2.00	1.00	6.70	42.95	0.00	
## CH2O	1.00	3.00	2.00	-0.10	-0.88	0.01	
## SCC*	1.00	2.00	1.00	4.36	17.02	0.00	
## FAF	0.00	3.00	3.00	0.50	-0.62	0.02	
## TUE	0.00	2.00	2.00	0.62	-0.55	0.01	
## CALC*	1.00	4.00	3.00	-1.17	0.46	0.01	
## MTRANS*	1.00	5.00	4.00	-1.28	-0.20	0.03	
## NObeyesdad*	1.00	7.00	6.00	0.01	-1.19	0.04	

4.2. Variables Numéricas

Primero vamos a comenzar analizando las variables numéricas, con las cuales se pueden realizar análisis exploratorios más elaborados. Un primer análisis gráfico puede ayudar a la interpretación de la naturaleza de las variables y de la muestra en concreto. Para este análisis gráfico, se utilizarán histogramas y diagramas de caja.





Tal y como se pueden ver, la muestra para la variable “Age” parece contener sobre todo personas jóvenes de entre 18 y 30 años, mientras que otras edades mayores son más escasas, por lo que está sesgada hacia la izquierda y puede afectar a la generalización de los resultados de los diferentes análisis (o como mínimo se debería tener en cuenta). El diagrama de caja indica que la variabilidad (expresada a través del rango intercuartílico) se podría considerar relativamente pequeña, por lo que las edades más alejadas del rango anteriormente mencionado se podrían considerar valores atípicos (por su lejanía y siguiendo el criterio del IQR).

Cuando se mira la variable “Height”, se puede ver como la distribución es más parecida a una normal (por la campana simétrica que se forma), aún habiendo algunos picos entre medio. El diagrama de caja, sin embargo, apoya esta observación debido a la simetría que se muestra: la variabilidad por debajo y por encima de la mediana es casi la misma, y los bigotes son muy parecidos.

La distribución de la variable “Weight” vuelve a estar sesgada hacia la izquierda, lo cual es razonable al ser mucho más raro ver pesos muy altos para diferentes personas (ya sea por la distribución real o por el sesgo de supervivencia que habría, dado que las personas muy obesas tienden a tener más enfermedades y, por tanto, tienen más probabilidad de no ser recogidas en la muestra si estas suelen fallecer a una tasa más alta) que no pesos más cercanos a la media o menores (no muy menores, como se puede apreciar en la cola izquierda). La caja muestra como la mediana está cerca de 80kg y cómo hay más variabilidad para valores mayores a la mediana.

La variable “NCP” tiene una distribución más compleja, donde muchísimas observaciones se acumulan sobre el valor de 3 o en los extremos, indicando como 3 (o cercanos) suele ser un valor frecuente (la moda). Esto tiene sentido debido a que en la mayoría del mundo se suelen tener 3 comidas al día. El diagrama de caja muestra como hay una gran presencia de valores iguales o cercanos a 3 (porque la mediana y el cuartil 75% coinciden), pero como, debido a esta acumulación todos los valores fuera de un rango de entre 2 y 3,5 se considerarían valores atípicos o alejados de lo común.

Algo similar pasa con “CH20”, solo que ahora se concentra sobre todo en 2, lo cual se muestra en el

histograma. No obstante, el diagrama de caja permite ver una mayor simetría que antes y una distribución un tanto más “equitativa” entre los diferentes valores, dado que la caja parece simétrica alrededor de la media y no se detectan valores atípicos, debido seguramente a la gran presencia de observaciones en los valores extremos.

En el caso de la variable “FAF”, se puede ver como su distribución es de la misma naturaleza que las últimas dos, pero que está más sesgada hacia la izquierda. Como esta variable se refiere a la actividad física, se puede observar como la mayoría de personas realizan entre nada y poco ejercicio físico. Esto también es apoyado por el diagrama de caja, dado que la caja abarca valores bajos de 1 a 2. No obstante, se puede ver como habría más dispersión para valores mayores a la mediana debido al hecho de que es menos común ver observaciones con un nivel de actividad físico alto o medio.

A “TUE” ocurre lo mismo que con la última variable, apoyado por el histograma y por el diagrama de caja. En este caso, se muestra como hay una mayor cantidad de observaciones en valores que muestran un uso bajo de dispositivos tecnológicos.

Finalmente, la variable “FCVC” es parecido a “TUE”, solo que en espejo.

4.3. Variables Categóricas

Después de haber analizado las variables numéricas, podemos seguir con las categóricas. Para poder analizarlas gráficamente, nos apoyamos en el uso de un “pie plot” que permita ver la distribución porcentual de las observaciones en las diferentes categorías,

Gráfico de pastel de CAEC

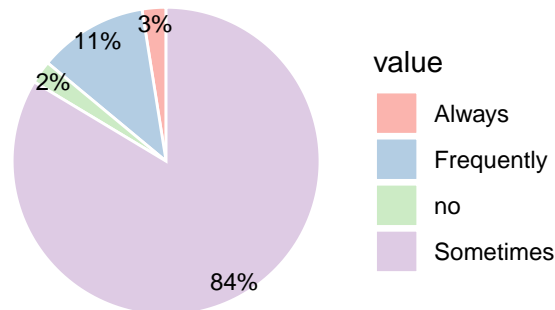


Gráfico de pastel de CALC

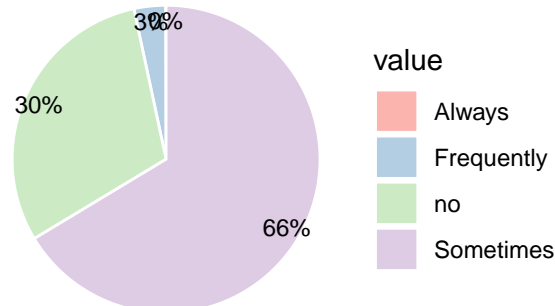


Gráfico de pastel de MTRANS

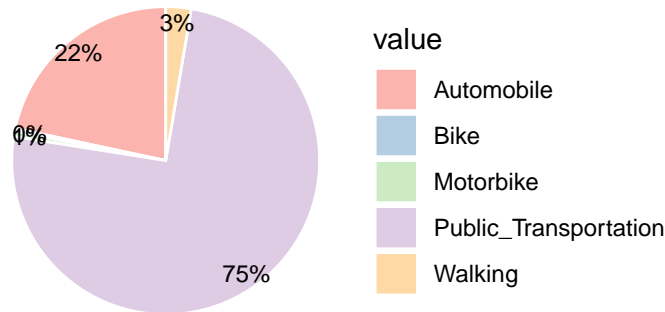
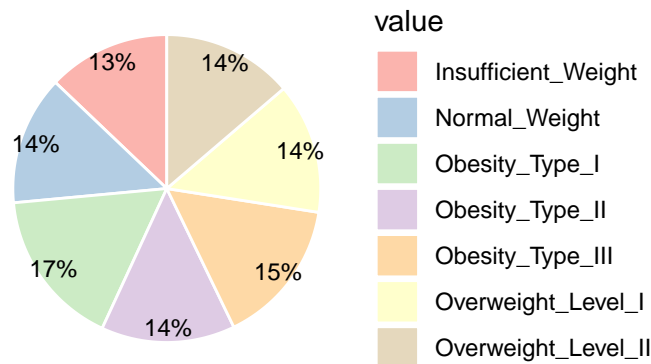


Gráfico de pastel de NObeyesdad



En el caso de la variable “CAEC”, se puede ver como la mayoría de observaciones en la muestra consumen comida entre horas algunas veces, representando un 84%. No obstante, se le otorga otro porcentaje más pequeño (del 11%) a aquellas personas que comen frecuentemente entre horas, siendo el no comer entre horas y el comer siempre entre horas algo poco frecuente. En el contexto del análisis de la obesidad, este es un resultado más positivo que los otros.

Mirando la variable “CALC”, se puede ver como las mayoría de las personas consumen alcohol de manera ocasional, pero esta vez hay un porcentaje importante (30%) que no consume alcohol. La categoría de “frecuente” representa solo un 3% (lo cual es esperable), mientras que no hay observaciones con un consumo de alcohol “demasiado” habitual.

Analizando la variable “MTRANS”, se puede ver como el 75% de la muestra utiliza el transporte público para desplazarse, mientras que un 22% utilizan automóviles privados. Pocas personas van andando, y casi ninguna va en moto o en bicicleta. Esto es común en países de latinoamérica dado que no suelen tener infraestructuras bien desarrolladas para otros vehículos como la bicicleta, por lo que se tendría que tener en cuenta.

Finalmente, la variable “NObeyesdad” muestra una distribución bastante equitativa sobre los diferentes niveles de obesidad, siendo esta distribución casi uniforme. Esto es preocupante desde un punto de vista médico, dado que quiere decir que la mayoría de personas sufren de diversos tipos de obesidad y sobrepeso, y el mismo número de personas sufre de peso insuficiente.

4.3. Variables Binarias

Por último, se puede analizar el último tipo de variables: las binarias o dicotómicas. Igual que antes, se hace uso de “pie plots” para poder interpretar.

Gráfico de pastel de Gender

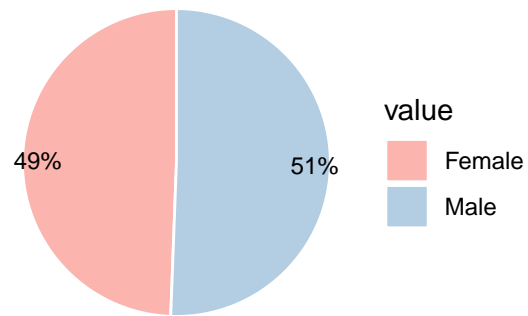


Gráfico de pastel de family_history_with_overweight

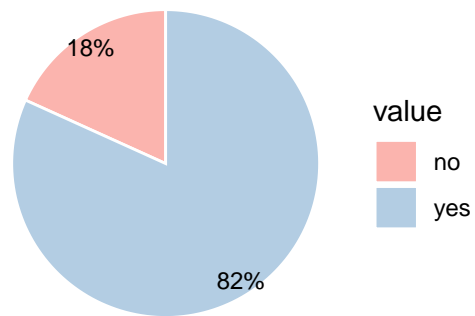


Gráfico de pastel de FAVC

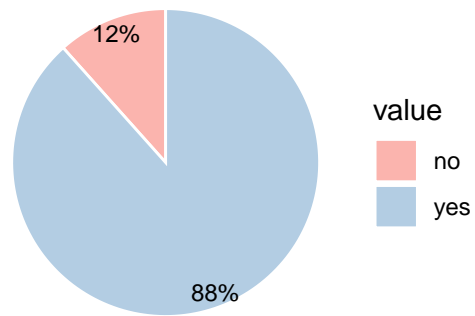


Gráfico de pastel de SMOKE

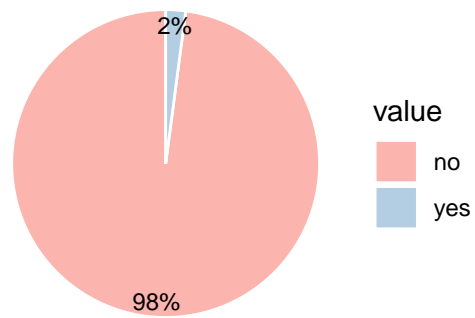
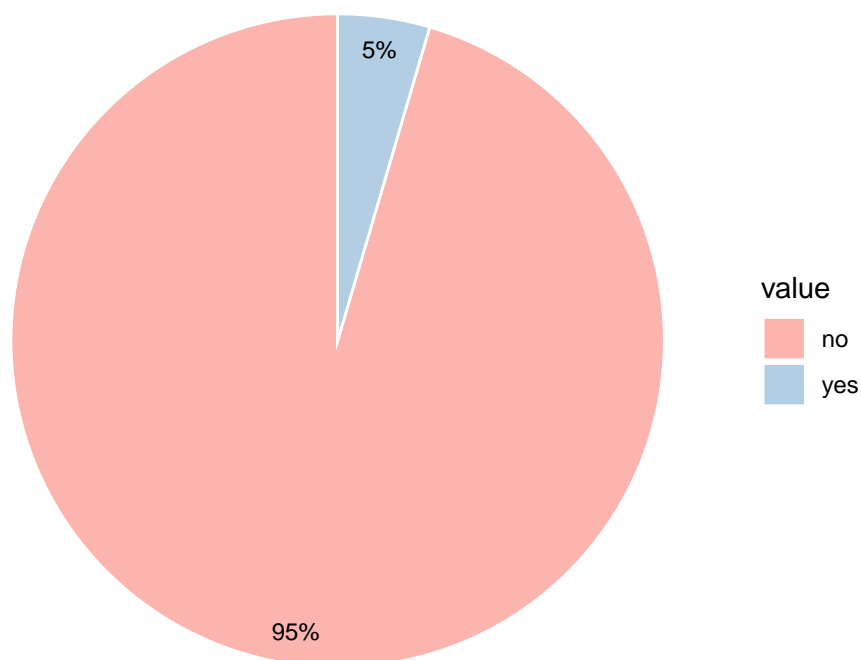


Gráfico de pastel de SCC



Tal y como se puede ver en la variable “gender”, la muestra contiene casi tantos hombres como mujeres, por lo que los resultados se podrían generalizar para ambos grupos.

El historial de sobrepeso de la familia, no obstante, tiene una distribución mucho menos equitativa, en donde la gran mayoría tiene algún familiar que ha sufrido o tiene sobrepeso de algún tipo, frente al 18% que no ha tenido ninguno.

La variable “FAVC” también muestra como la gran mayoría de las personas han consumido productos muy calóricos, frente al 12% que no.

El gráfico de la variable “SMOKE” muestra como casi la totalidad de las personas de la muestra no fuman, frente a un pequeño porcentaje del 2% que si fuman, concluyendo así que el análisis es sobre todo para no fumadores y que será difícil generalizar los resultados a fumadores si esto no se controla.

Finalmente, la variable “SCC” muestra como la gran mayoría de la muestra no monitorea sus calorías, frente a un 5% que si lo hace. Aunque esto es beneficioso desde el punto de vista de la salud, es una práctica poco común que normalmente realizan personas muy conscientes de su salud o deportistas.

5. Preprocesado de datos

Antes de entrar en el análisis de PCA, vamos a procesar los datos para que evitar que produzcan errores en el cálculo. Cabe destacar que no trataremos los outliers dado que consideramos que las observaciones extremas son relevantes para este análisis de obesidad. Como por ejemplo el rango de datos de mayores de 40 años, personas que comen más de 3 veces al día etc.

5.1. Discretización

Comenzamos con el proceso de discretización. Previamente, hemos identificado picos destacados en algunas variables numéricas, como por ejemplo TUE y FAF, que reflejan el uso de dispositivos electrónicos (como teléfonos móviles y televisores) y el nivel de actividad física semanal. Estas variables se originan en preguntas específicas del cuestionario original:

FAF: ¿How often do you have physical activity?

- I do not have
- 1 or 2 days
- 2 or 4 days
- 4 or 5 days

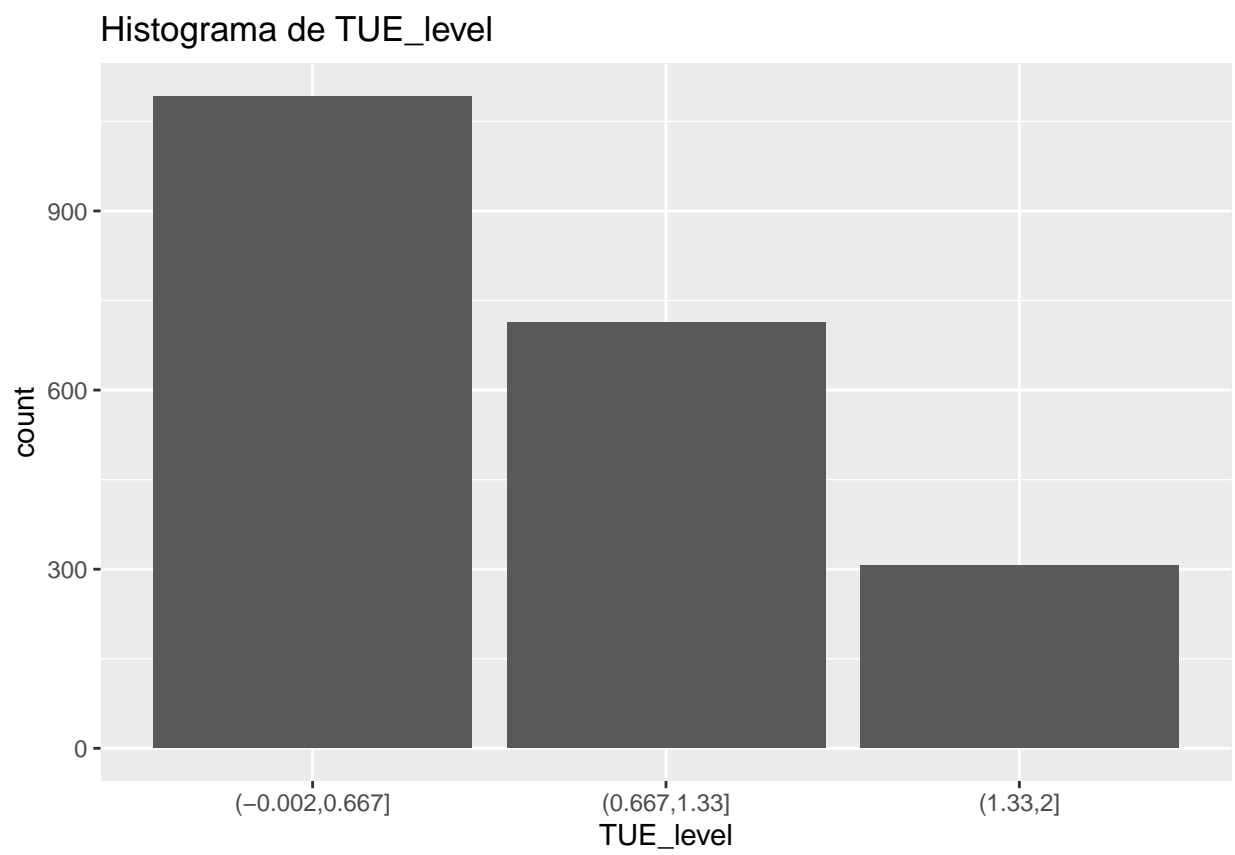
TUE: ¿How much time do you use technological devices such as cell phone, videogames, television, computer and others?

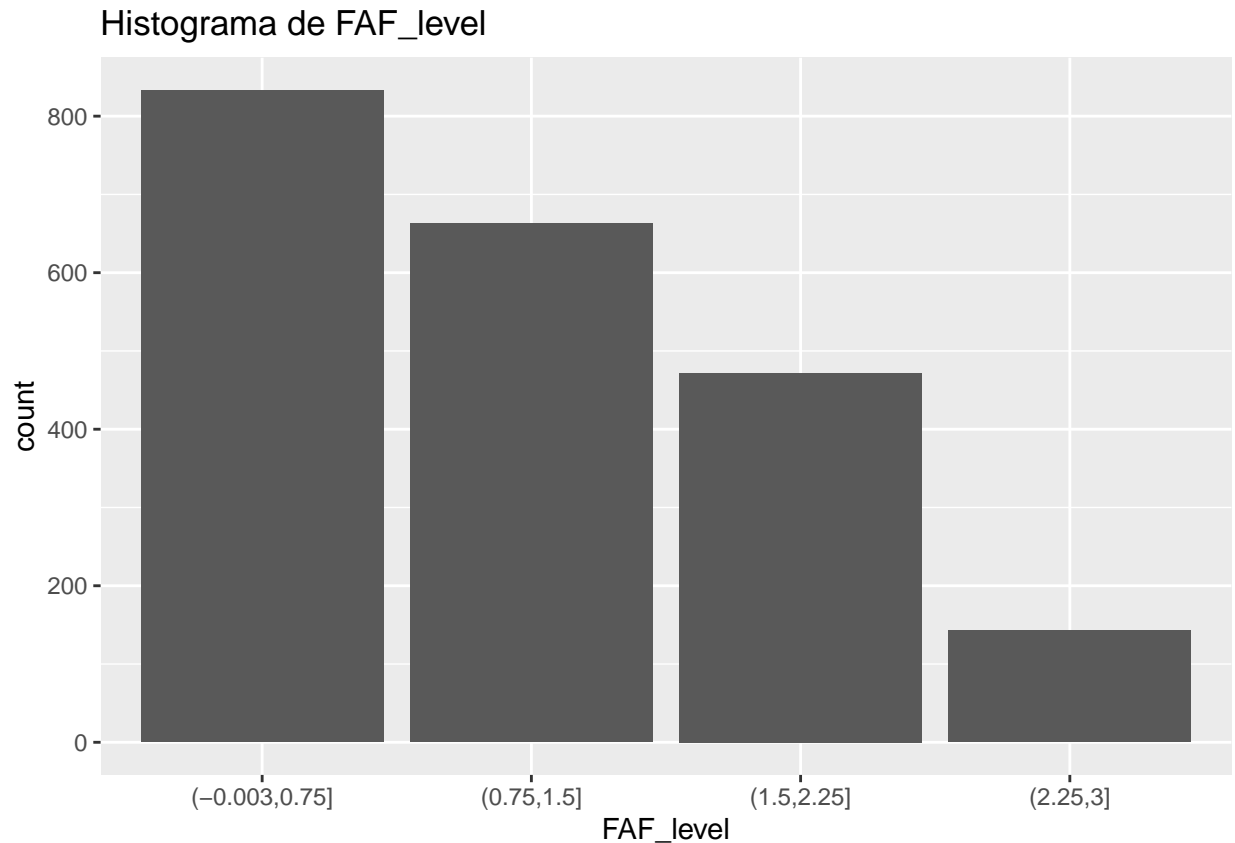
- 0-2 hours
- 3-5 hours
- More than 5 hours

En consecuencia, aplicamos la función `cut()` para discretizar TUE en 3 niveles y FAF en 4 en los intervalos especificados en el cuestionario.

Para TUE_level, el intervalo $(-0.002, 0.667]$ tiene 1092 observaciones, seguido por $(0.667, 1.33]$ con 713 observaciones y $(1.33, 2]$ con 306 observaciones. En cuanto a FAF_level, el intervalo $(-0.003, 0.75]$ contiene 833 observaciones, $(0.75, 1.5]$ tiene 663 observaciones, $(1.5, 2.25]$ cuenta con 472 observaciones, y $(2.25, 3]$ presenta 143 observaciones. Estos datos ofrecen una visión de cómo se distribuyen los valores discretizados en diferentes intervalos para estas dos variables.

Con el fin de visualizar la distribución de los valores discretizados, creamos un histograma para cada variable. Posteriormente, eliminamos las variables originales discretizadas, junto con MTRANS, debido a su ambiguo impacto en la obesidad.





Los bar plots creados siguen la misma distribución que las histogramas.

5.2. Factorización de variables categóricas

Las variables categóricas y binarias no se han tratado hasta ahora. En este paso, se convierten las variables categóricas y binarias en factores.

Observamos la distribución de las 2 variables nuevas creadas

```
## (-0.002,0.667] (0.667,1.33] (1.33,2]
##              1092         713         306

## (-0.003,0.75] (0.75,1.5] (1.5,2.25] (2.25,3]
##              833         663         472         143
```

Ahora transformamos las variables categóricas en variables numéricas. Esto implica asignar un valor numérico a cada categoría de las variables categóricas.

La asignación será de la siguiente manera:

- Gender:
 - Female: 1
 - Male: 0
- family_history_with_overweight:

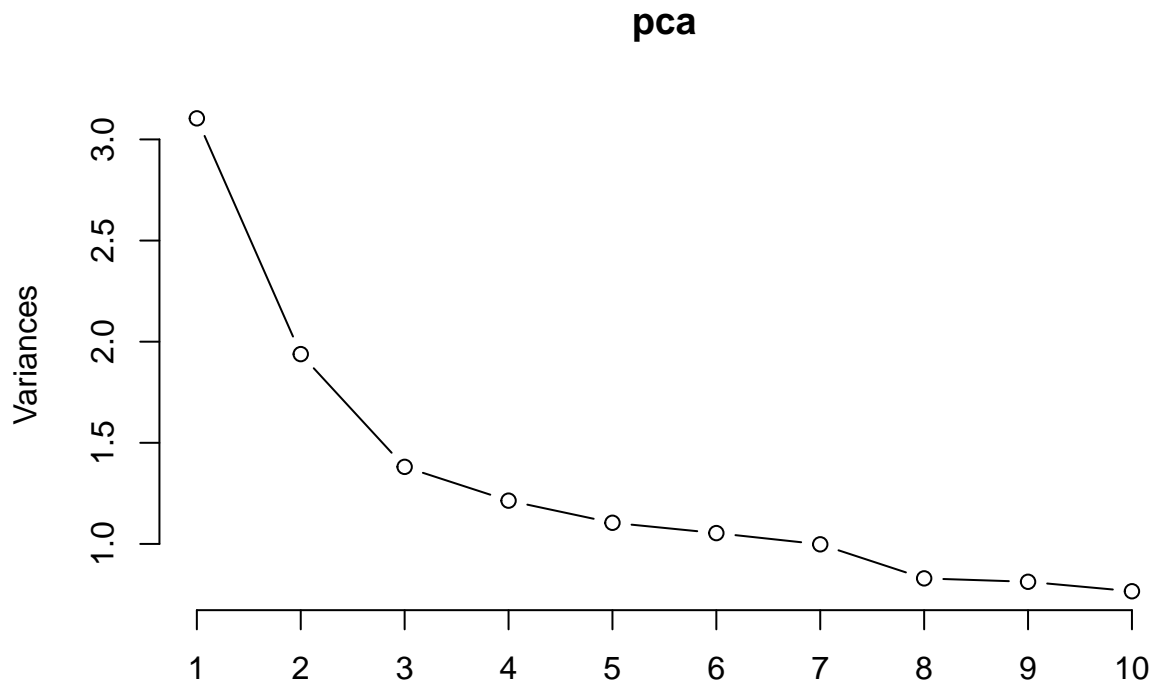
- yes: 1
 - no: 0
- FAVC:
 - yes: 1
 - no: 0
- SMOKE:
 - yes: 1
 - no: 0
- SCC:
 - yes: 1
 - no: 0
- NObeyesdad:
 - Insufficient_Weight: 0
 - Normal_Weight: 1
 - Overweight_Level_I: 2
 - Overweight_Level_II: 3
 - Obesity_Type_I: 4
 - Obesity_Type_II: 5
 - Obesity_Type_III: 6
- TUE_level:
 - (-0.002,0.667]: 0
 - (0.667,1.33]: 1
 - (1.33,2]: 2
- FAF_level:
 - (-0.003,0.75]: 0
 - (0.75,1.5]: 1
 - (1.5,2.25]: 2
 - (2.25,3]: 3
- CAEC:
 - no: 0
 - Sometimes: 1
 - Frequently: 2
 - Always: 3
- CALC:
 - no: 0
 - Sometimes: 1
 - Frequently: 2
 - Always: 3

6. PCA

6.1. Componentes Principales

Ahora normalizamos nuestro dato y realizamos la PCA.

```
## Importance of components:
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation  1.762 1.3923 1.17519 1.10184 1.05103 1.02645 0.99907
## Proportion of Variance 0.194 0.1212 0.08632 0.07588 0.06904 0.06585 0.06238
## Cumulative Proportion 0.194 0.3152 0.40148 0.47736 0.54640 0.61225 0.67464
##          PC8      PC9      PC10     PC11     PC12     PC13     PC14
## Standard deviation  0.91089 0.9016 0.87526 0.86085 0.81505 0.77083 0.70269
## Proportion of Variance 0.05186 0.0508 0.04788 0.04632 0.04152 0.03714 0.03086
## Cumulative Proportion 0.72650 0.7773 0.82518 0.87149 0.91301 0.95015 0.98101
##          PC15     PC16
## Standard deviation  0.53097 0.14803
## Proportion of Variance 0.01762 0.00137
## Cumulative Proportion 0.99863 1.00000
```



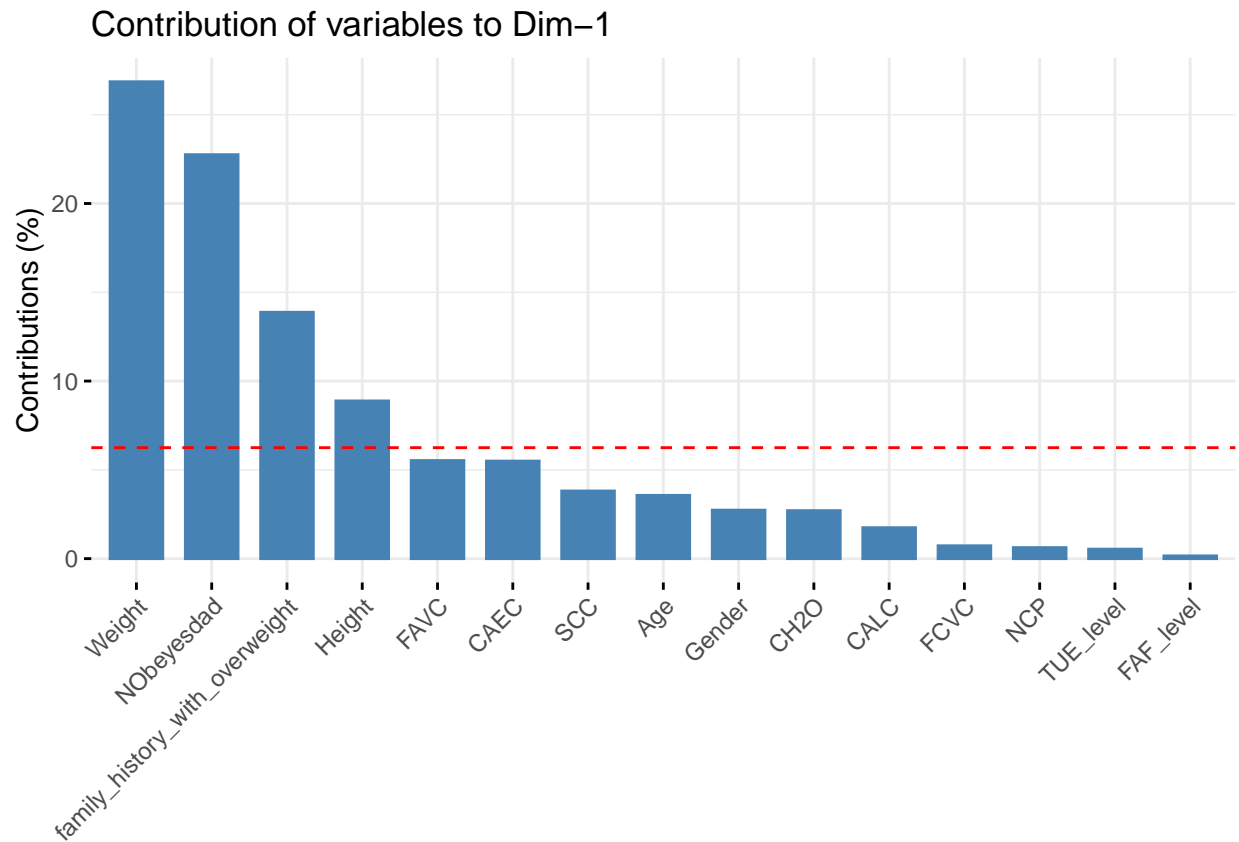
Con el resumen proporcionado y la gráfica de varianzas, vemos que el “codo” se encuentra en el punto 3. Esto sugiere que lo óptimo es coger los 3 primeros componentes principales para explicar nuestro caso (explica más o menos 40% de la varianza).

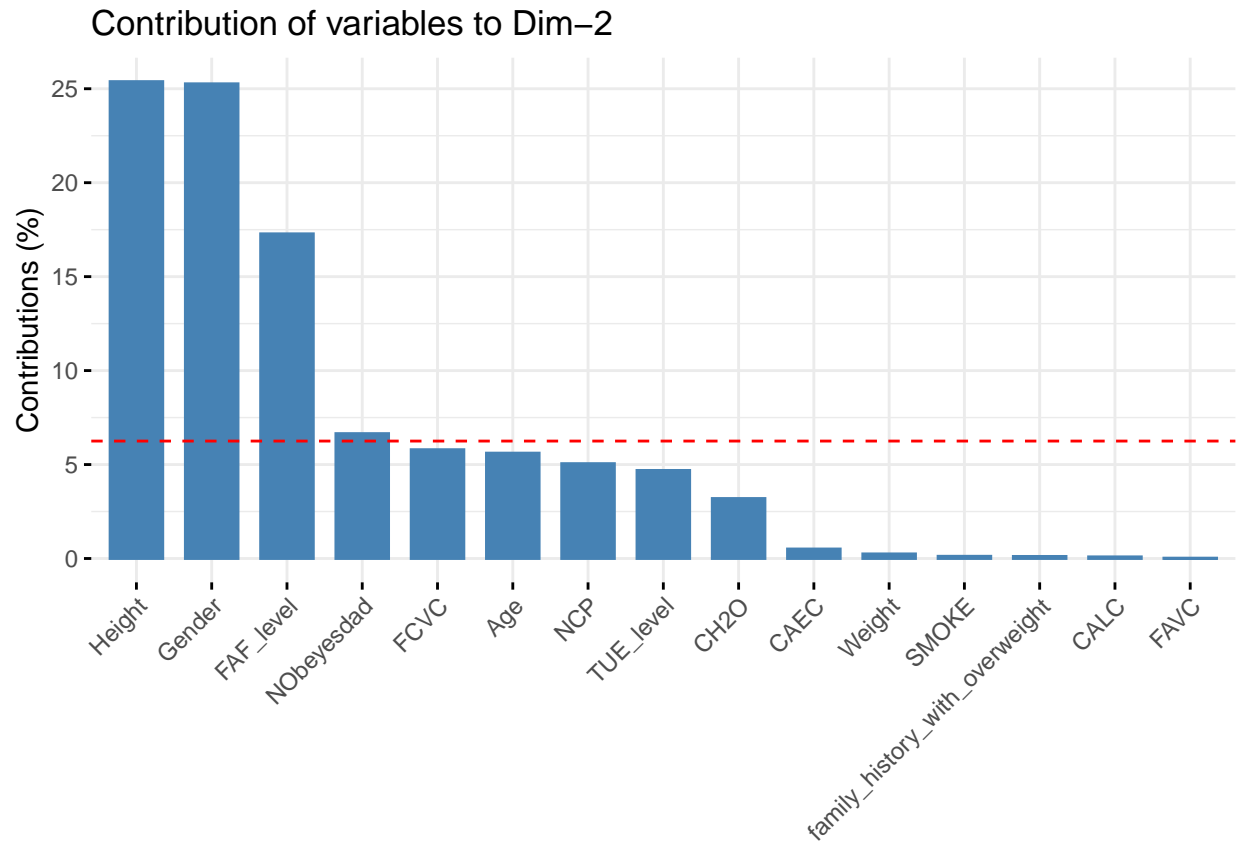
Veamos cuáles son los variables que contribuyen más en estas 3 PCs.

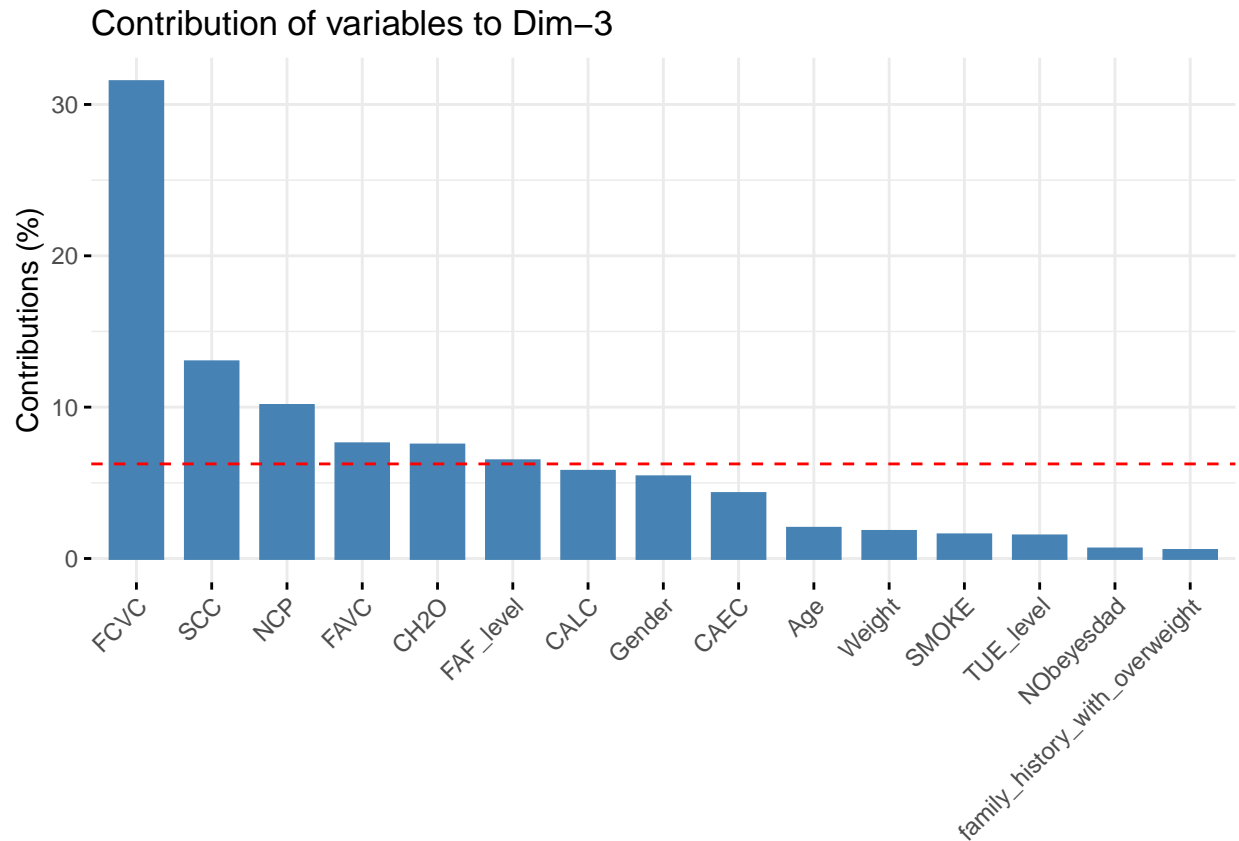
##	PC1	PC2	PC3
## Gender	-0.16528101	0.50263720	-0.23239677
## Age	0.18877865	0.23681820	0.14137394
## Height	0.29801140	-0.50379656	-0.06064230
## Weight	0.51824020	0.04957032	-0.13386806
## family_history_with_overweight	0.37247957	0.03336716	0.07292231
## FAVC	0.23504253	0.01415765	0.27530768
## FCVC	0.08517477	0.24066961	-0.56136056
## NCP	0.07896896	-0.22469491	-0.31803294
## CAEC	-0.23442116	-0.07120003	-0.20720296
## SMOKE	0.01596824	-0.03445260	-0.12506502
## CH2O	0.16443277	-0.17870392	-0.27385325
## SCC	-0.19519315	0.00793861	-0.36053515
## CALC	0.13206943	0.02966082	-0.24004416
## NObeyesdad	0.47696023	0.25782484	-0.07940804
## TUE_level	-0.07316571	-0.21650093	0.12225716
## FAF_level	-0.03938104	-0.41569306	-0.25415591

PC1 está asociada principalmente con características relacionadas al peso como peso, nivel de obesidad e historial familiar de sobrepeso por las altas cargas en Weight, NObeyesdad y family_history_with_overweight. PC2 parece estar relacionada con la actividad física, el género y la altura, por sus cargas en Height, Gender y FAF mientras que PC3 puede reflejar patrones de consumo de alimentos y la ingesta de agua por la contribución de FCVC (consumo de verduras) y SCC (Monitoreo de calorías consumidas).

Veámoslo mejor en un plot de contribución.

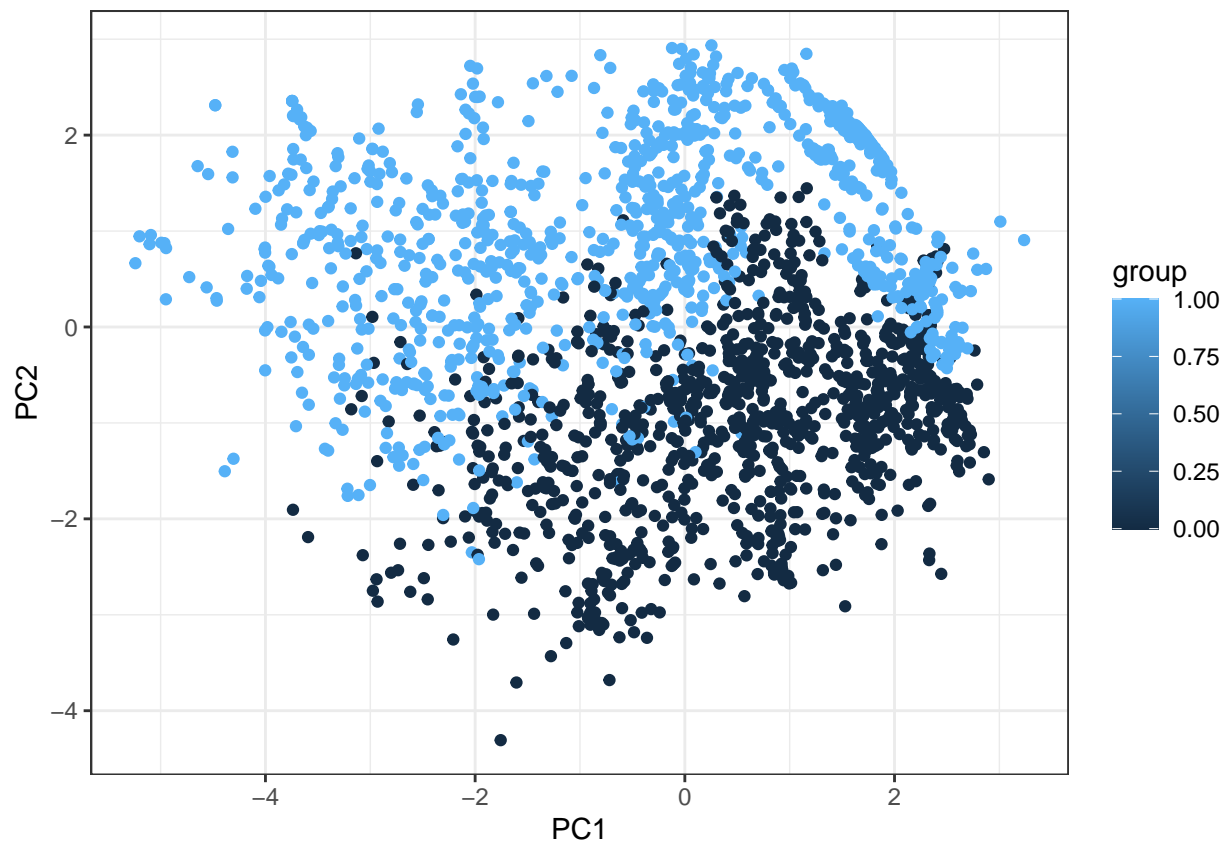






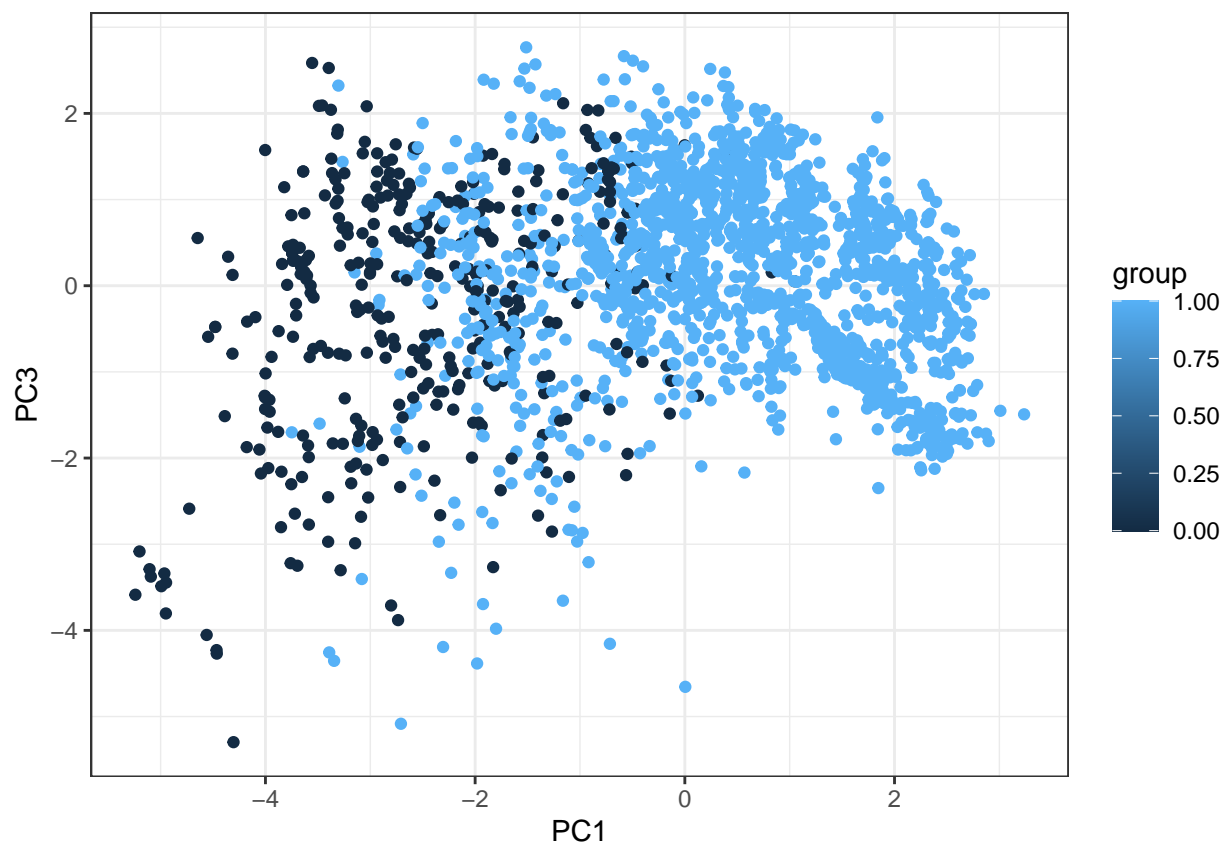
Los gráficos de dispersión abajo ilustran el impacto de cada variable en los tres componentes principales.

Comenzando con Gender en PC1 (Peso) vs PC2 (Condición física), no se evidencia una relación directa entre ambos componentes. Por otra parte, la disparidad de la variable Gender es más notable en PC2, donde las mujeres (Gender = 0) tienden a tener una puntuación más baja en este componente.



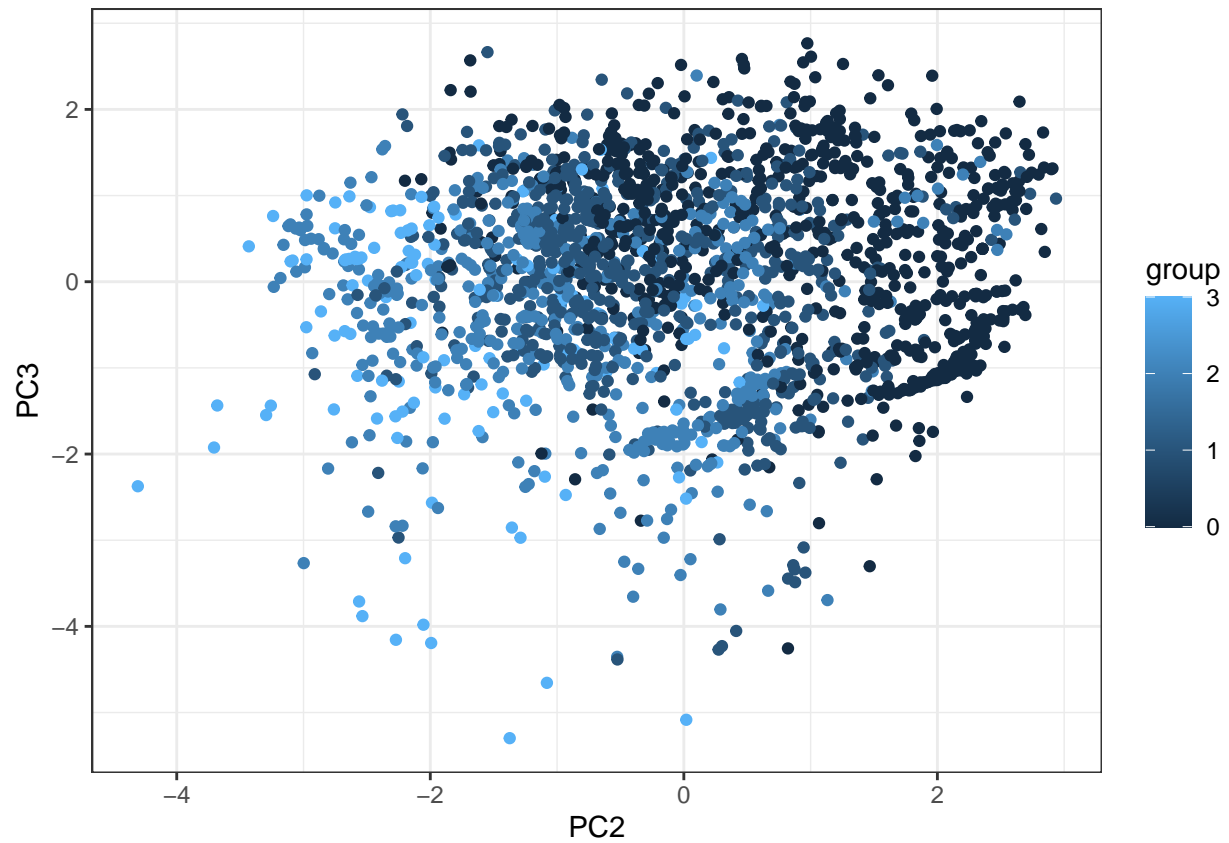
En el gráfico siguiente, comparamos PC1 con PC3, dividiendo los puntos según si las muestras tienen historial familiar de obesidad o no.

Se puede observar una relación más o menos positiva entre los dos componentes. Además, la variable de historial familiar de sobrepeso está mucho más relacionada con la PC1. En la PC3, aunque no es tan evidente como en la PC1, se observa que aquellos con historial familiar tienden a tener puntuaciones más altas en PC3. Esto podría relacionarse con una mejor organización de la dieta y un mayor consumo de verduras en comparación con aquellos sin historial.



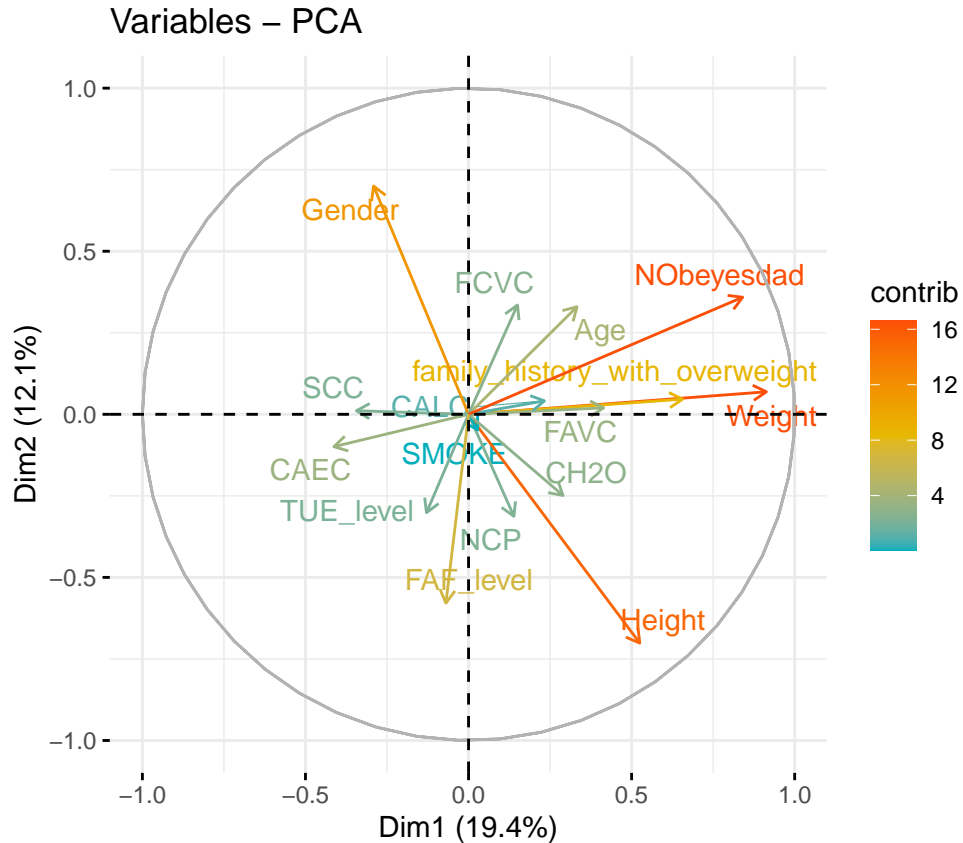
En la última comparación entre PC2 y PC3, utilizamos la variable FAF_level (nivel de actividad física). Se observa una relación positiva entre ambos componentes: una mejor condición física se asocia con una mejor dieta y un mayor consumo de verduras.

Sin embargo, en cuanto a la variable FAF, observamos una ligera contradicción con la afirmación anterior. Esto puede explicarse porque las personas que realizan más actividad física tienden a necesitar un mayor consumo de proteínas animales en comparación con el resto, lo que puede influir en sus puntuaciones en PC3.



6.2. Relación entre las variables

Creemos un biplot para observar mejor la relación que hay entre las variables.

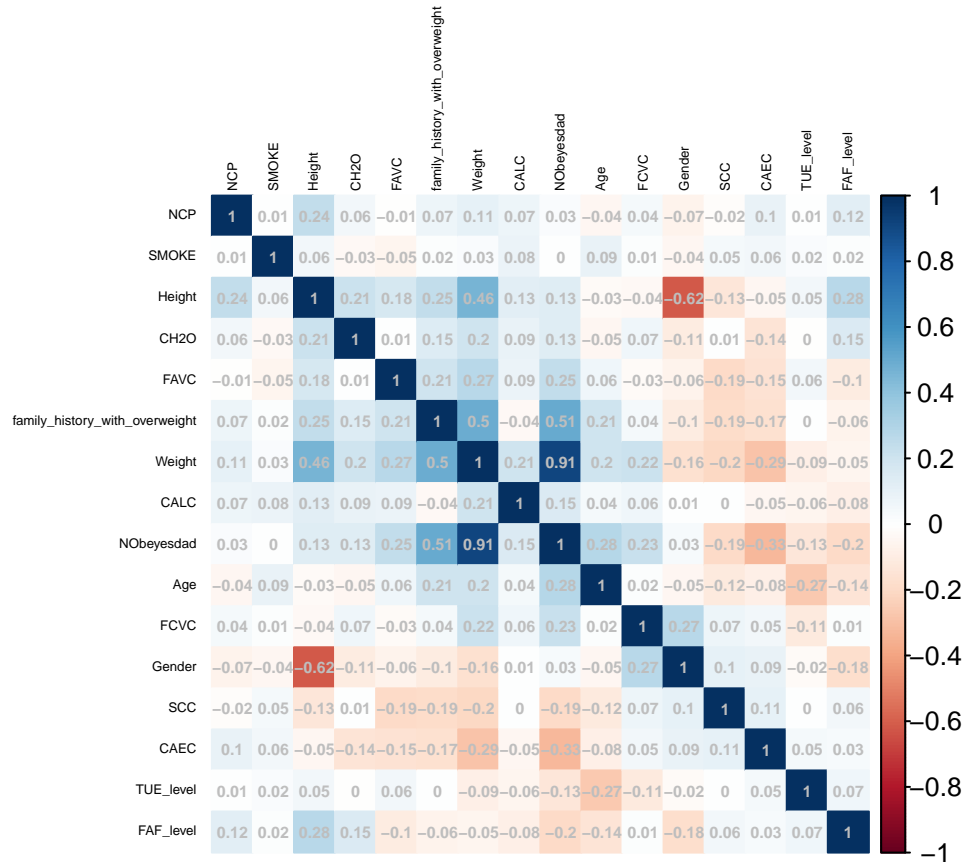


Este gráfico nos ilustra la correlación y la importancia de cada variable para las componentes principales. Notamos que las cuatro variables más influyentes son: Gender, Weight, NObeyesdad y Height.

Weight y NObeyesdad están altamente correlacionadas, lo que es comprensible ya que una representa el peso y la otra el nivel de obesidad.

Por otro lado, Gender se encuentra apartada del resto de las variables, mostrando una baja correlación con la mayoría y una fuerte correlación negativa con Height, lo que concuerda con las diferencias físicas entre hombres y mujeres.

Esta correlación también puede observarse en un corrplot.



7. Ramdon Forest

7.1. Discretización del resto de las variablas numéricas

Para la variable **age**, decidimos dividirla según las etapas de la vida, alineadas con la distribución observada anteriormente:

- 0 - 20 años: Este grupo incluye principalmente a estudiantes y a personas que recién ingresan al mercado laboral. Estas personas suelen estar en fases de educación o en los primeros años de sus carreras profesionales, con un poder de adquisición bajo y mucho tiempo para dedicarse al ejercicio físico.
- 20 - 35 años: En este intervalo se encuentra la población laboral activa. Las personas en este rango de edad generalmente están en una etapa de crecimiento profesional y desarrollo personal. Es un período de alta actividad económica y laboral. Tiempo de ejercicio físico limitada por el trabajo.
- 35 - 50 años: Este grupo representa una vida establecida. Las personas en este rango suelen haber alcanzado estabilidad en sus carreras y vida personal. Es una etapa caracterizada por ingresos estables y responsabilidades familiares y profesionales consolidadas. Tiempo de ejercicio físico limitada por el trabajo, y buscan actividades menos intensivas.
- 50+ años: Aquí se incluyen la parte de la población jubilada y los que si sitúan en los últimos años de la carrela profesional. Las personas de este grupo han dejado de trabajar o están cerca de hacerlo, centrando sus actividades en el retiro y la jubilación. Tienen bastante tiempo libre pero la capacidad para realizar ejercicios físicos son limitadas.

Esta clasificación nos permitirá analizar mejor cómo varían las diferentes características y comportamientos de los clientes a lo largo de las distintas etapas de la vida.

```
## Less than 20      20-35      35-50      50+
##           585        1358        158        10
```

La variable Height se divide en cinco grupos: “Under 1.60”, “1.60-1.70”, “1.70-1.80”, “1.80-1.90” y “1.90+”.

```
## Under 1.60  1.60-1.70  1.70-1.80  1.80-1.90  1.90 +
##           316        733        749        288        25
```

Los cuartiles de la variable Weight se utilizan para crear cuatro grupos: “Under 65.47”, “65.47-83.00”, “83.00-107.43” y “107.43+”.

```
## Under 65.47  65.47-83.00  83.00-107.43  107.43+
##           528        536        519        528
```

La variable FCVC se divide en dos grupos: “Under 2” y “2-3”.

```
## Under 2      2-3
##           802    1309
```

La variable NCP se divide en cuatro grupos: “Under 1”, “1-2”, “2-3” y “3-4”.

```
## Under 1      1-2      2-3      3-4
##           199      196     1488     228
```

```
##      Gender      family_history_with_overweight  FAVC      CAEC
##  Female:1043    no : 385                        no : 245    Always : 53
##  Male :1068    yes:1726                       yes:1866   Frequently: 242
##                                           no : 51
##                                           Sometimes :1765
##
##
##
```

```
## SMOKE      CH20      SCC      CALC
## no :2067    Min. :1.000    no :2015    Always : 1
## yes: 44    1st Qu.:1.585    yes: 96    Frequently: 70
##                      Median :2.000                      no : 639
##                      Mean :2.008                      Sometimes :1401
##                      3rd Qu.:2.477
##                      Max. :3.000
```

```
##      NObeyesdad      TUE_level      FAF_level
##  Insufficient_Weight:272 (-0.002,0.667]:1092 (-0.003,0.75]:833
##  Normal_Weight :287 (0.667,1.33] : 713 (0.75,1.5] :663
##  Obesity_Type_I :351 (1.33,2] : 306 (1.5,2.25] :472
##  Obesity_Type_II :297 (2.25,3] :143
##  Obesity_Type_III :324
##  Overweight_Level_I :290
##  Overweight_Level_II:290
```

```

##          age_group      Height_group      Weight_group      FCVC_group
## Less than 20: 585      Under 1.60:316      Under 65.47 :528      Under 2: 802
## 20-35          :1358      1.60-1.70 :733      65.47-83.00 :536      2-3      :1309
## 35-50          : 158      1.70-1.80 :749      83.00-107.43:519
## 50+           :  10      1.80-1.90 :288      107.43+      :528
##              1.90 +      : 25
##
##
##      NCP_group
## Under 1: 199
## 1-2      : 196
## 2-3      :1488
## 3-4      : 228
##
##
##

```

7.2. Test de Phi y Cramer

Calculamos las medidas de asociación Phi y V de Cramer para medir la relación entre nuestro variable target del árbol con el resto de las variables

```

##          Variable  Test_Phi Test_Cramer
## 1          Gender 0.5581939  0.5581939
## 2 family_history_with_overweight 0.5428051  0.5428051
## 3          FAVC 0.3324694  0.3324694
## 4          CAEC 0.6167477  0.3560794
## 5          SMOKE 0.1233855  0.1233855
## 6          CH2O 1.9412345  0.7925057
## 7          SCC 0.2414074  0.2414074
## 8          CALC 0.4004838  0.2312194
## 9      NObeyesdad 2.4494897  1.0000000
## 10      TUE_level 0.2786719  0.1970508
## 11      FAF_level 0.3293222  0.1901342
## 12      age_group 0.5133697  0.2963941
## 13      Height_group 0.4105095  0.2052547
## 14      Weight_group 1.1661906  0.6733005
## 15      FCVC_group 0.4390959  0.4390959
## 16      NCP_group 0.5531633  0.3193690

```

Las variables con mayor correlación con NObeyesdad son CH2O (consumo de agua) y Weight_group (grupos de peso), con valores de Phi y Cramer relativamente altos. Las variables como SMOKE, TUE_level, SCC, y FAF_level tienen una correlación baja con la variable objetivo. Las demás variables presentan correlaciones moderadas, lo que indica una relación significativa pero no dominante con la variable NObeyesdad.

Dado que no tenemos ninguna variable con coeficiente menor a 0.1, dejaremos todas las variables para crear nuestro modelo de Random Forest

7.3. Modelo

Dividimos el conjunto de datos en entrenamiento y prueba

```

## Tamaño del train_data: 1691

```

Tamaño del test_data: 420

Entrenamos el modelo Random Forest.

Y examinamos el resultado mediante una matriz de confusión.

Confusion Matrix and Statistics

```
##
##
##               Reference
## Prediction      Insufficient_Weight Normal_Weight Obesity_Type_I
##   Insufficient_Weight                47                2                0
##   Normal_Weight                    7                45                3
##   Obesity_Type_I                   0                1               62
##   Obesity_Type_II                  0                0                2
##   Obesity_Type_III                 0                0                0
##   Overweight_Level_I               0                5                1
##   Overweight_Level_II              0                4                2
##
##               Reference
## Prediction      Obesity_Type_II Obesity_Type_III Overweight_Level_I
##   Insufficient_Weight            0                0                0
##   Normal_Weight                  0                0                4
##   Obesity_Type_I                 0                0                5
##   Obesity_Type_II                59                0                0
##   Obesity_Type_III               0               64                0
##   Overweight_Level_I            0                0               46
##   Overweight_Level_II           0                0                3
##
##               Reference
## Prediction      Overweight_Level_II
##   Insufficient_Weight            0
##   Normal_Weight                  3
##   Obesity_Type_I                 7
##   Obesity_Type_II                1
##   Obesity_Type_III              0
##   Overweight_Level_I            5
##   Overweight_Level_II          42
##
## Overall Statistics
##
##               Accuracy : 0.869
##               95% CI : (0.833, 0.8998)
##   No Information Rate : 0.1667
##   P-Value [Acc > NIR] : < 2.2e-16
##
##               Kappa : 0.847
##
##   McNemar's Test P-Value : NA
##
## Statistics by Class:
##
##               Class: Insufficient_Weight Class: Normal_Weight
##   Sensitivity                0.8704                0.7895
##   Specificity                0.9945                0.9532
##   Pos Pred Value            0.9592                0.7258
##   Neg Pred Value            0.9811                0.9665
```

## Prevalence	0.1286	0.1357
## Detection Rate	0.1119	0.1071
## Detection Prevalence	0.1167	0.1476
## Balanced Accuracy	0.9325	0.8713
##	Class: Obesity_Type_I	Class: Obesity_Type_II
## Sensitivity	0.8857	1.0000
## Specificity	0.9629	0.9917
## Pos Pred Value	0.8267	0.9516
## Neg Pred Value	0.9768	1.0000
## Prevalence	0.1667	0.1405
## Detection Rate	0.1476	0.1405
## Detection Prevalence	0.1786	0.1476
## Balanced Accuracy	0.9243	0.9958
##	Class: Obesity_Type_III	Class: Overweight_Level_I
## Sensitivity	1.0000	0.7931
## Specificity	1.0000	0.9696
## Pos Pred Value	1.0000	0.8070
## Neg Pred Value	1.0000	0.9669
## Prevalence	0.1524	0.1381
## Detection Rate	0.1524	0.1095
## Detection Prevalence	0.1524	0.1357
## Balanced Accuracy	1.0000	0.8814
##	Class: Overweight_Level_II	
## Sensitivity	0.7241	
## Specificity	0.9751	
## Pos Pred Value	0.8235	
## Neg Pred Value	0.9566	
## Prevalence	0.1381	
## Detection Rate	0.1000	
## Detection Prevalence	0.1214	
## Balanced Accuracy	0.8496	

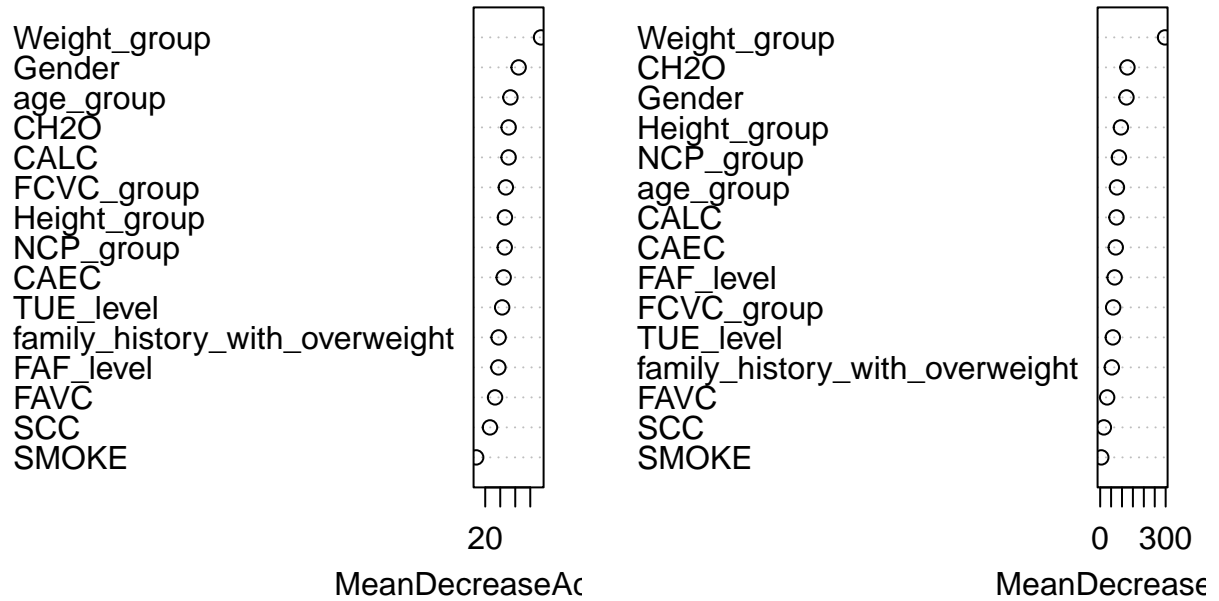
El modelo muestra un buen rendimiento general con una exactitud del 86.9% y un coeficiente Kappa de 0.847, indicando una alta concordancia entre las predicciones y las clases reales. Las clases Obesity_Type_III y Obesity_Type_II destacan con sensibilidad y especificidad perfectas (1.0000), señalando una predicción precisa en estas categorías.

Sin embargo, las clases Normal_Weight, Overweight_Level_I, y Overweight_Level_II presentan menores valores de sensibilidad y especificidad, sugiriendo dificultades en su clasificación. La exactitud balanceada es alta en la mayoría de las clases, lo que indica un buen equilibrio entre la capacidad del modelo para identificar correctamente los positivos y los negativos verdaderos.

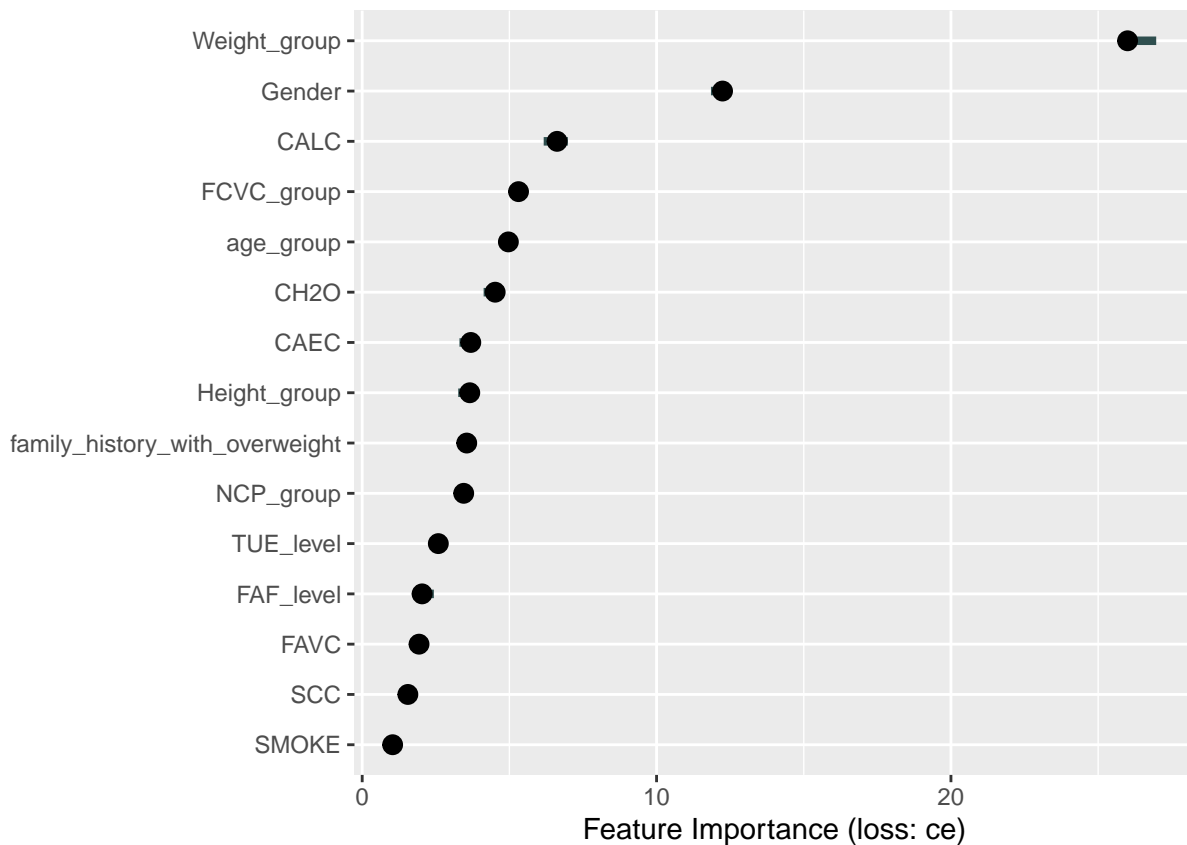
La matriz de confusión refleja una correcta clasificación en la mayoría de las categorías, con algunos errores menores en las clases de peso normal y sobrepeso. Estos resultados demuestran que, aunque el modelo es efectivo en general, existe margen de mejora en la clasificación de ciertas categorías específicas. En resumen, el modelo Random Forest es robusto y preciso, pero puede beneficiarse de ajustes adicionales para mejorar la clasificación en las categorías con menor rendimiento.

Ahora entramos en el análisis de la importancia para entender cuáles son las características más relevantes para la predicción de los niveles de obesidad (NObesidad).

rf_model



Weight_group emerge como la variable más importante, lo que sugiere que el peso tiene un impacto significativo en la clasificación de los niveles de obesidad, lo cual es obvio **Gender**, **age_group**, y **Height_group** también son altamente influyentes, indicando que el género, la edad, y la altura son factores críticos. Variables como **CH2O** (consumo de agua) también muestran relevancia considerable, subrayando la importancia de los hábitos de hidratación.



##	feature	importance.05	importance	importance.95
## 1	Weight_group	25.717241	26.000000	26.972414
## 2	Gender	11.848276	12.241379	12.413793
## 3	CALC	6.165517	6.620690	6.986207
## 4	FCVC_group	5.179310	5.310345	5.372414
## 5	age_group	4.882759	4.965517	5.186207
## 6	CH2O	4.124138	4.517241	4.579310
## 7	CAEC	3.303448	3.689655	3.917241
## 8	Height_group	3.262069	3.655172	3.862069
## 9	family_history_with_overweight	3.317241	3.551724	3.613793
## 10	NCP_group	3.255172	3.448276	3.558621
## 11	TUE_level	2.503448	2.586207	2.717241
## 12	FAF_level	1.862069	2.034483	2.434483
## 13	FAVC	1.765517	1.931034	1.986207
## 14	SCC	1.524138	1.551724	1.586207
## 15	SMOKE	1.034483	1.034483	1.062069
##	permutation.error			
## 1		0.44589001		
## 2		0.20993495		
## 3		0.11354228		
## 4		0.09107037		
## 5		0.08515671		
## 6		0.07746895		
## 7		0.06327617		
## 8		0.06268480		
## 9		0.06091070		

## 10	0.05913661
## 11	0.04435245
## 12	0.03489060
## 13	0.03311650
## 14	0.02661147
## 15	0.01774098

Las variables más influyentes en la predicción de los niveles de obesidad son **Weight_group**, **Gender**, y **age_group**, junto con hábitos como el consumo de alcohol y verduras. Estos resultados destacan la importancia del peso y el género, así como ciertos comportamientos alimenticios en la determinación de los niveles de obesidad, proporcionando información valiosa para futuras intervenciones y políticas de salud.

9. Conclusión

En este proyecto, hemos aplicado técnicas avanzadas de análisis de datos, específicamente el Análisis de Componentes Principales (PCA) y el modelo de Random Forest, para abordar el problema de la obesidad en Colombia, Perú y México. Los resultados obtenidos a partir de estas metodologías proporcionan insights valiosos sobre los factores que contribuyen al sobrepeso y la obesidad, y cómo pueden ser utilizados para desarrollar estrategias efectivas de intervención.

El PCA identificó factores clave que influyen en el riesgo de obesidad, destacando la importancia del peso, el historial familiar de sobrepeso, la actividad física y la dieta en la salud metabólica al revelar las relaciones lineales entre las observaciones y las variables originales. Estos componentes revelan que el peso, la actividad física y los hábitos alimenticios son factores críticos en la comprensión de la obesidad en la población estudiada.

El modelo de Random Forest demostró ser altamente preciso en la predicción de los niveles de obesidad (**NObesidad**), con una exactitud del 86.9% y un coeficiente Kappa de 0.847, lo que indica una alta concordancia entre las predicciones del modelo y las clases reales. Las variables más importantes identificadas por el modelo fueron: el peso, el género y la edad junto con los hábitos alimenticios reflejados en variables como FCVC (consumo de vegetales) y CH2O (consumo de agua).

Los resultados del PCA y el Random Forest proporcionan una comprensión detallada de los factores que contribuyen a la obesidad. Estos insights pueden ser utilizados para segmentar de manera más efectiva a la población objetivo en grupos con características similares, permitiendo desarrollar estrategias de marketing y productos adaptados a las necesidades específicas de cada grupo.

Además, las empresas del sector de bienestar y fitness pueden aprovechar estos hallazgos para ofrecer soluciones personalizadas y accesibles que promuevan hábitos más saludables en la población, mejorando así la calidad de vida y reduciendo la prevalencia de la obesidad en la región.

Para futuras investigaciones, se pueden explorar modelos predictivos más sofisticados y la incorporación de datos longitudinales para comprender mejor las tendencias temporales en los factores de riesgo de obesidad y su impacto en la salud a lo largo del tiempo. Estos enfoques avanzados ayudarán a impulsar la prevención y el manejo efectivo de la obesidad en América Latina, contribuyendo así a mejorar la calidad de vida y reducir la carga de enfermedades asociadas con esta condición.