

## Exploración de representaciones para identificar relaciones entre atributos

Alfredo Piero Mateos-Papis<sup>1</sup>, Christian Sánchez-Sánchez<sup>1</sup>,  
Héctor Jiménez-Salazar<sup>1</sup>, Natalí N Guerrero-Vargas<sup>2,3</sup>,  
Alberto Manuel Ángeles-Castellanos<sup>2,4</sup>, Carolina Escobar<sup>2,5</sup>

1 Universidad Autónoma Metropolitana,  
México

2 Universidad Nacional Autónoma de México,  
Facultad de Medicina,  
México

{csanchez, amateos, hjimenez}@cua.uam.mx  
natalinadi@facmed.unam.mx, mangleles\_castellanos@unam.mx  
cescobar@comunidad.unam.mx

**Resumen.** La motivación de este trabajo es el desarrollo alarmante del sobrepeso y obesidad en adultos jóvenes, lo cual no se explica tan solo por aspectos individuales como dieta, sedentarismo, estrés, u otros conocidos, sino que recientemente se ha encontrado explicación en aspectos sociales de la vida moderna, caracterizados por la actividad que se prolonga hacia la noche, que afectan al sistema circadiano. En este trabajo se analizan datos obtenidos de un cuestionario aplicado a jóvenes, alumnos de la carrera médico cirujano, sobre hábitos diarios que afectan al sistema circadiano, como sueño-vigilia, alimentación por la noche, exposición a luz nocturna, y uso de aditamento luminosos por la noche, además de datos como el índice de masa corporal (IMC), que determina la condición de sobrepeso u obesidad, para buscar la relación entre dichos hábitos y el IMC (atributos). Hasta donde los autores conocen, hay pocos trabajos que se apoyen en el Análisis de Componentes Principales (PCA), como sí lo hace el presente trabajo, donde la cantidad de atributos considerados entorpecen la obtención de dependencias de interés entre tales atributos. Así, a partir de buscar la relación entre el IMC con los otros atributos, aquí se propone un criterio para identificar los componentes que mejor muestran las relaciones entre un atributo, en este caso el IMC, y los otros. Es importante mencionar que con las técnicas tradicionales, desde las más simples, como la correlación entre las variables, hasta la representación directa en componentes principales, o la separación no lineal a través de Kernel PCA, como coseno o sigmoidal, no se pudieron identificar relaciones entre el IMC y otros atributos, pero esto sí se logró a través de una búsqueda que maximizara las proyecciones del atributo IMC en algunos de los componentes principales. En el trabajo se muestran, visualmente, los atributos que +más tienen relación con el IMC.

**Palabras claves:** Componentes principales, identificación de dependencias, ritmo circadiano.

## Exploring Representations for Identification of Relations between Attributes

**Abstract.** The motivation for this work is the alarming increase in overweight and obesity in young adults. This trend is not only explained by individual factors such as diet, sedentary lifestyle, stress, or other known factors. It has also recently been explained by social aspects of modern life, characterized by nighttime activity, which affects the circadian system. This work analyzes data obtained from a questionnaire administered to young students of medical and surgical studies regarding daily habits that affect the circadian system, such as sleep-wakefulness, nighttime eating, exposure to light at night, and use of luminous devices at night. This questionnaire also analyzes data such as the body mass index (BMI), which determines overweight or obesity, to identify relationships between these habits and BMI (attributes). To the authors' knowledge, few studies rely on Principal Component Analysis (PCA), as this study does. The number of attributes considered makes it difficult to determine relevant dependencies between these attributes. Thus, by searching for the relationship between BMI and other attributes, a criterion is proposed here to identify the components that best reflect the relationships between one attribute, in this case BMI, and the others. It is important to mention that traditional techniques, from the simplest, such as correlation between variables, to direct representation in principal components, or nonlinear separation through Kernel PCA, such as cosine or sigmoidal, failed to identify relationships between BMI and other attributes. However, this was achieved through a search that maximized the projections of the BMI attribute on some of the principal components. The attributes most closely related to BMI are visually shown in this paper.

**Keywords:** Principal components, identification of dependencies, circadian rhythm.

### 1. Introducción

El aumento progresivo y alarmante que ha tenido en los últimos 20 a 30 años el sobrepeso y la obesidad, en México y en algunos países del Mundo, no se logra explicar por las causas más exploradas en el ámbito individual, como la dieta, el sedentarismo, el estrés, alteraciones hormonales, fármacos, factores genéticos, etc. Este aumento puede deberse a la actividad y trabajo 24 h / 7 días del estilo de vida moderno [1-2], donde las personas: 1) prolongan su actividad hacia la noche reduciendo las horas dedicadas a dormir [3]; 2) tienen horarios irregulares de alimentación [4] y consumen más alimentos por la noche [5]; y 3) se exponen a contaminación lumínica por la noche [6-7]. Estas condiciones afectan al sistema circadiano de las personas, que tiene como función principal adaptar a los organismos a un ambiente cíclico dado por el día y la noche, y regular la homeostasis. Entre los grupos sociales mayormente expuestos a este nuevo estilo de vida están los adolescentes y adultos jóvenes, y los profesionales de la

salud [8, 9, 6], donde se identifica el “Jetlag social” [10]. Estudios clínicos y experimentales recientes señalan que la disrupción circadiana es un factor de riesgo para desarrollar enfermedades crónicas, incluyendo al sobrepeso y obesidad y enfermedades metabólicas [11].

Con el fin de conocer los factores que afectan al sistema circadiano que más inciden en el desarrollo de sobrepeso y obesidad, un grupo de investigadores de la Facultad de Medicina de la UNAM [12] aplicó un cuestionario a jóvenes alumnos de la carrera de médico cirujano, colectando datos de peso y altura; y hábitos diarios sobre sueño-vigilia, alimentación, exposición a luz nocturna y uso de aditamentos electrónicos por la noche, para generar datos con decenas de atributos, entre los cuales está el “Índice de Masa Corporal” (IMC), donde  $IMC = Kg\ de\ peso / (metros\ de\ altura)^2$ , que representa al sobrepeso y la obesidad. Este grupo pidió apoyo para el análisis de datos a un grupo de estudio de ciencia de datos y computación. Estos dos grupos son los autores de este trabajo.

La mayor parte de los estudios sobre los factores que disrumpen al sistema circadiano, que pueden causar alteraciones como sobrepeso y obesidad en las personas, obtienen sus datos de cuestionarios, o de pruebas con grupos de personas en condiciones controladas de laboratorio, y usan técnicas de análisis estadístico para obtener resultados; pero a conocimiento de los autores, hay pocos trabajos que apliquen el Análisis de Componentes Principales, PCA<sup>1</sup>, como técnica de análisis estadístico, que ayuda en situaciones como las del presente trabajo, donde tanto la cantidad de atributos considerados, aunque sean “solamente” decenas, como su alta correlación, entorpecen o impiden la obtención de resultados. El objetivo de este trabajo es buscar la relación o dependencia entre el IMC con los otros atributos de los datos, usando PCA.

Con relación al PCA, ésta es una técnica proviene del año 1900 [13], que se ha desarrollado con intensidad recientemente por el poder del cómputo. Con PCA se realiza una secuencia de transformaciones lineales de los datos “originales”, cada transformación con relación a lo que se llama un componente principal<sup>2</sup> (PC<sup>3</sup>), donde los datos “transformados” con relación al primer componente principal (PC (1)) son los más cercanos a los datos “originales”; los datos transformados con relación al segundo componente principal (PC (2)) son los siguientes más cercanos, y así sucesivamente. Se dice que PC (1) recaba la mayor información de los datos, y así sucesivamente [14]. Los datos en torno a algún PC, con relación a los datos en torno a cualquier otro PC, no están correlacionados. Estas características de PCA permiten seleccionar los primeros PC, desechando los restantes y reteniendo una mayor información sobre los datos originales, para reducir la cantidad de datos “transformados” y así simplificar el análisis. En la literatura sobre PCA se recomienda seleccionar los PC con más información [15-17].

En este trabajo no se tuvieron resultados claros con el uso de los PC con más información; en contraste, sí se obtuvieron resultados más claros al utilizar otros PC con menos información que sin embargo sí tenían una mayor información considerando

---

<sup>1</sup> Principal Component Analysis.

<sup>2</sup> Con relación al significado de PC ver el pie de página 6.

<sup>3</sup> Principal Component.

solo al IMC. Los autores consideran que este resultado es un hallazgo que ofrece una opción atractiva para el análisis de datos. El balance entre la información global y la información relativa a un solo atributo (el IMC) se realizó con el uso de un indicador propuesto en este trabajo. Los resultados de este trabajo se limitan a realizar operaciones que resuelven un problema particular, con observación visual en una gráfica que se llama Biplot, que presenta gráficamente los resultados del PCA.

En el Apéndice (Sección 8) se presenta la lista de los atributos correspondientes a los datos usados en este trabajo, que incluyen el identificador de dos letras con que se representan en el cuerpo del trabajo.

Las siguientes secciones de este trabajo son: Sección 2 - Trabajos Relacionados; Sección 3 - Descripción de los datos; Sección 4 - Método de análisis (e hipótesis); Sección 5 - Resultados; Sección 6 - Conclusiones y trabajo futuro.

## **2. Trabajos relacionados**

Esta sección presenta algunos trabajos relacionados que vinculen al IMC con factores del estilo de vida, indicando qué métodos de análisis utilizan.

Subu et al. [18], en su estudio buscan si existe relación entre la adicción a los videojuegos y el IMC, enfocado a 250 alumnos de secundaria en Indonesia. El análisis usa la herramienta SPSS, determinando dos clases de nivel de adicción: bajo y alto. El estudio mostró que de 177 estudiantes con nivel bajo de adicción a los videojuegos 126 tenían IMC normal, mientras que estudiantes con mayor nivel de adicción tendían a tener sobrepeso; sin embargo reportan que con sus datos la relación entre el nivel de adicción y el IMC fue débil e inversa, y plantean en el trabajo futuro tratar de entender esa relación y en su caso si existe causalidad alguna.

Johnson et al. [19] estudiaron la asociación del insomnio, o duración corta del sueño menor a 6 horas, con hipertensión, diabetes tipo 2 e IMC. Aunque en su estudio no encontraron evidencia de que el insomnio, o dormir menos de 6 horas, se asociara con el IMC, menciona que se apoyaron principalmente en otro estudio publicado por Crönlein et al. [20]. Es importante mencionar que ellos utilizaron 2,345 artículos, extraídos de 5 bases de datos, con 11 estudios individuales para una revisión cualitativa y 10 estudios para un metaanálisis, donde el tamaño de la muestra varió de 30 hasta 4,994 participantes; sin embargo, reportan que son muy pocos los estudios que busquen dicha relación.

Contrario a esto, los investigadores Bocicor et al. [21] también realizaron un estudio sobre la relación entre desorden de sueño y el incremento de IMC en pacientes adultos. En el estudio se sondearon 84 pacientes, entre 18 y 65 años, quienes respondieron un cuestionario (SAQ©) que ayudaba a determinar el nivel de desorden del sueño, donde un desorden medio de sueño lo obtienen quienes alcanzan entre 4 a 5 puntos y un severo desorden de sueño es obtenido con 6 a 7 puntos. La asociación entre el desorden de sueño y el IMC fue probada estadísticamente (Chi cuadrada y Prueba exacta de Fisher), y los investigadores reportan que sí encontraron una asociación entre el desorden de sueño y el IMC.

Por otro lado, se han presentado algunos estudios usando PCA y Biplot para estudiar problemas de salud, tal es el caso de Durankova et al. [22], que presentan un estudio de los efectos del estilo de vida en la salud física de alumnos de la Universidad en Eslovaquia. Los autores estudiaron el IMC en relación con nueve factores (lugar, comida, actividad, cantidad de actividad física, tipo de actividad física, consumo de alcohol, consumo de tabaco, estrés y dormir), en una muestra de 200 estudiantes (108 hombres y 92 mujeres). Los datos fueron obtenidos mediante un cuestionario. Los resultados mostraron diferencias entre ambos sexos, ya que mientras la actividad diaria y la actividad física fueron asociados son mayor obesidad en los hombres para las mujeres lo fue el consumo del tabaco.

También Imran et al. [23] presentan un estudio sobre el patrón de prevalencia de sobrepeso y obesidad y factores de riesgo asociado en la división de Malakand en Pakistán, con una muestra de 562 individuos, de los cuales obtuvieron datos sobre su estilo de vida, alimentación, presión arterial, consumo de tabaco, enfermedades y actividades de su vida. Para realizar revisar la relaciones los autores utilizaron análisis multivariado, análisis de clusters y técnicas de análisis de correspondencia canónica, reportan que la obesidad es mayor para mujeres casadas y en hombres la obesidad se relaciona por la comida y su estilo de vida.

Es importante mencionar que actualmente el uso de la computación puede ayudar a determinar relaciones entre diversos factores y el IMC, por ejemplo Zhang et al. [24] desarrollaron una revisión de artículos que buscan determinar el IMC basado en el análisis de los rostros humanos. Se presentan enfoques basados en predicciones matemáticas y aprendizaje automático (Machine Learning), principalmente clasificación, que toman en cuenta características geométricas de los rostros, extraídos de diferentes fuentes, reportando buenos resultados en la mayoría de los trabajos incluidos.

Otro ejemplo del uso de la tecnología para determinar el IMC es el presentado por Mark [25], donde a través de realizar una un dispositivo, basado en IOT, compuesto por un microcontrolador y un dispositivo para medir el ancho del cuerpo y otro para medir la altura, obtienen los datos que permiten determinar el IMC del usuario, donde se utilizó un algoritmo de regresión gaussiana que puede predecir si el peso de una persona es bajo, o peso normal, o sobrepeso, u obeso, con una exactitud reportada de 99.18%.

### **3. Descripción de los datos**

Entre los años 2018 y 2019 se aplicó el cuestionario de Cronobiología: “Hábitos circadianos”, a los estudiantes de primero y segundo año de la Licenciatura “Médico Cirujano”, de la Facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM) [12]. Este estudio fue aprobado por la comisión de Investigación y Ética de la División de Investigación Facultad de Medicina UNAM, FM/DI/097/2018.

El cuestionario colectó datos generales, entre ellos edad, peso y altura; y además colectó datos de hábitos diarios sobre patrones de: 1) sueño-vigilia; 2) horarios de alimentación; 3) exposición a luz nocturna; 4) uso de aditamentos electrónicos por la

noche. Las preguntas; se orientan tanto a patrones en días entre semana y en días de fines de semana. Se excluyeron los cuestionarios que aportaran información errónea en las unidades de peso y talla, o que tuvieran datos incompatibles en cuanto a horarios. Con los valores provenientes de las preguntas, se obtuvieron valores derivados comparativos, por ejemplo el *Jetlag* de la hora de despertar, que proviene calcular el valor absoluto de la resta entre la hora de despertar entre semana y la hora de despertar en fin de semana. Al final, se obtiene una matriz con 67 columnas, de los cuales se seleccionaron 34. Cada columna, para cada individuo, tiene un dato que le da al individuo una característica o rasgo, por eso se llama atributo.

El sobrepeso y obesidad se obtienen a través del IMC (Índice de Masa Corporal), que es proporcional al peso (en Kg), e inversamente proporcional al cuadrado de la estatura (en metros), tal que los investigadores clasifican el peso como: 1) Bajo peso ( $IMC < 18$ ), 2) peso normal ( $18 \leq IMC < 25$ ), 3) sobrepeso ( $25 \leq IMC < 30$ ), y 4) obesidad ( $IMC \geq 30$ ).

Con los cuestionarios y datos seleccionados, se forma una matriz de datos de 2485 renglones por 34 columnas, que equivale a 2485 cuestionarios (uno por individuo) de 34 atributos. En la matriz de datos, el individuo asociado a cada renglón se identifica con la matrícula de la UNAM, representada por las sigla: ID; y cada uno de los 34 atributos tiene una sigla. Las 34 siglas son: G1, G2, G3, G4, G5, S1, S2, S3, S4, S5, S6, C1, C2, C3, C4, C5, C6, C7, C8, C9, C10, C11, C12, C13, C14, L1, L2, L3, D1, D2, D3, D4, D5, D6, donde el atributo G4 es el IMC, que es el atributo de interés en este trabajo, por indicar la condición de sobrepeso y obesidad. Los atributos se refieren según la letra de inicio de la sigla: G  $\rightarrow$  información general; S  $\rightarrow$  sueño (horas de dormir); C  $\rightarrow$  consumo (alimentación); L  $\rightarrow$  luz (iluminación).

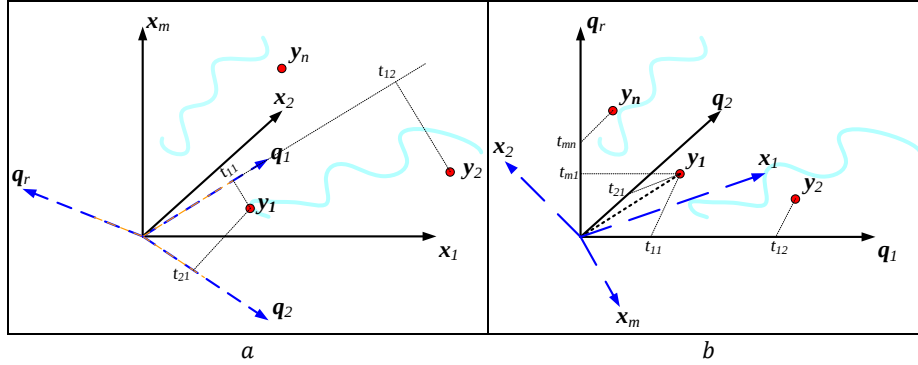
## 4. Método de análisis

La hipótesis de este trabajo es que mediante el análisis de componentes principales, y el B176iplot, se puede encontrar una relación que hay entre factores asociados (atributos) a la disrupción del sistema circadiano, y la presencia de sobrepeso y obesidad.

Con las variables  $m$  y  $n$  se representa el número de atributos, y el número de individuos, respectivamente (o sea  $m = 34$  y  $n = 2485$ ). Así, la matriz de datos tiene  $n$  renglones y  $m$  columnas, y se representa como  $Y_{n \times m}$ .

### 4.1 Análisis de componentes principales

Para estudiar la relación entre los atributos se puede representar gráficamente un espacio cartesiano de  $m$  dimensiones, con ejes coordenados ortonormales  $x_i$   $i = 1, \dots, m$ , donde cada eje coordenado corresponde a una dimensión y representa a un atributo. En este espacio se colocan  $n$  puntos,  $y_1, \dots, y_n$ , cada punto representando a un individuo, y cada punto teniendo las coordenadas de uno de los renglones,  $y_{1 \times m}, \dots, y_{n \times m}$ , de la matriz de datos  $Y_{n \times m}$ .



**Fig. 1.** Representación de: a) El plano original  $m$ -dimensional con ejes coordenados ortonormales,  $x_1, \dots, x_m$  (correspondientes a los  $m$  atributos), con  $n$  puntos,  $y_1, \dots, y_n$  (representando a  $n$  individuos), y  $r$  vectores,  $q_1, \dots, q_r$  (representando a las  $r$  columnas de la matriz de transformación  $Q_{m \times r}$  (ver **ecuación 1**)); b) El plano alternativo  $r$ -dimensional con ejes coordenados ortonormales,  $q_1, \dots, q_r$ , con proyecciones de los  $n$  puntos,  $y_1, \dots, y_n$ , y proyecciones de los  $m$  vectores,  $x_1, \dots, x_m$ .

Cuando  $m$  es “grande”, es difícil observar y obtener conclusiones de la representación en este espacio que llamamos “original”. Por lo anterior, es interesante poder encontrar un espacio cartesiano “alternativo”, con menos dimensiones, que represente “cercanamente” al espacio “original”, y por eso se plantea una transformación lineal matricial, que se puede lograr multiplicando la matriz  $Y_{n \times m}$  por una matriz de transformación  $Q_{m \times r} = [q_{1m \times 1} \quad q_{2m \times 1} \quad \dots \quad q_{rm \times 1}]$ , donde  $r \leq m$ , para obtener la matriz  $T_{n \times r}$ , tal como lo indica la **ecuación 1**. Al multiplicar  $Y_{n \times m}$  por cada columna,  $q_{im \times 1}$ , de  $Q_{m \times r}$  se produce la transformación  $t_{in \times 1}$ :

$$Y_{n \times m} Q_{m \times r} = \begin{bmatrix} y_{11 \times m} \\ y_{21 \times m} \\ \vdots \\ y_{n1 \times m} \end{bmatrix} [q_{1m \times 1} \quad q_{2m \times 1} \quad \dots \quad q_{rm \times 1}] =$$

$$T_{n \times r} = [t_{1n \times 1} \quad t_{2n \times 1} \quad \dots \quad t_{rn \times 1}] = \begin{bmatrix} t_{11} & t_{21} & \dots & t_{r1} \\ t_{12} & t_{22} & \dots & t_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1n} & t_{2n} & \dots & t_{rn} \end{bmatrix}. \quad (1)$$

La **Figura 1(a)** presenta al plano de  $m$  dimensiones con los puntos,  $y_1, \dots, y_n$ , y las  $r$  columnas,  $q_{1m \times 1}, \dots, q_{rm \times 1}$  representadas como los vectores  $q_1, \dots, q_r$ . Hay que notar que tanto vectores como puntos tienen  $m$  dimensiones.

La transformación pretende ver a los puntos  $y_1, \dots, y_n$  desde el punto de vista del espacio alternativo  $r$  dimensional, que tiene ejes coordenados que son los vectores  $q_1, \dots, q_r$  (que como se explica más adelante al calcularlos resultan ortonormales). En el espacio alternativo también se puede ver como vectores a los ejes coordenados originales  $x_1, \dots, x_m$ , correspondientes a los  $m$  atributos.

Antes de la transformación, es importante “normalizar” los valores de cada columna de  $Y_{n \times m}$  (para tener media cero y variancia uno en cada columna de la matriz), para no sesgar la transformación hacia uno o más atributos. Por facilidad, la matriz normalizada también se representa como  $Y_{n \times m}$ .

En la transformación, en el espacio “original” se pretende minimizar el promedio del error entre la ubicación de los puntos  $y_1, \dots, y_n$ , y la ubicación de sus proyecciones sobre cada uno de los vectores  $q_1, \dots, q_r$  (se dice que se minimizan los residuos [17]). De la minimización se obtiene la dirección y sentido de cada uno de los vectores  $q_1, \dots, q_r$ , representados en el espacio “original”. La primera minimización se hace con relación a  $q_1$ , la segunda con relación a  $q_2$  dado que se tiene la minimización con  $q_1$ , y así sucesivamente. Los últimos vectores  $q_i$  resultan ser los que tienen mayor residuo.

Se puede mostrar [17] que estas minimizaciones equivalen a maximizar la variancia de las proyecciones de los puntos  $y_1, \dots, y_n$  sobre los vectores  $q_1, \dots, q_r$ . Esto resulta en que la variancia de las proyecciones sobre  $q_1$  sea la mayor; la variancia de las proyecciones sobre  $q_2$  sea la siguiente mayor; y así sucesivamente.

Después de maximizar (ver [17]) resulta que los vectores  $q_1, \dots, q_r$ ,  $r \leq m$ , son los eigenvectores de la matriz  $Y_{m \times n}^T Y_{n \times m}$  (la matriz de covariancias de  $Y_{n \times m}$ ), tal que  $Y_{m \times n}^T Y_{n \times m} q_{i_{m \times 1}} = \varphi_i^2 q_{i_{m \times 1}}$ ,  $i = 1, \dots, r$ , donde el eigenvalor de  $q_{1_{m \times 1}}$  es  $\lambda_1 = \varphi_1^2$ , que es el máximo eigenvalor para esa matriz; el eigenvalor de  $q_{2_{m \times 1}}$  es  $\lambda_2 = \varphi_2^2$ , que es el segundo mayor eigenvalor, y así sucesivamente (estos valores aparecen en la **ecuación 2**).

Por ser  $Y_{m \times n}^T Y_{n \times m}$  una matriz simétrica, los eigenvectores son reales, y por ser matriz de covariancia todos los eigenvalores son positivos [17]. Además, por ser  $q_{i_{m \times 1}}$ ,  $i = 1, \dots, r$ ,  $r \leq m$ , eigenvectores, todos ellos son ortogonales cuando sus correspondientes eigenvalores son diferentes, y los eigenvectores que pudiesen corresponder a iguales eigenvalores se pueden seleccionar para ser ortogonales entre sí, de tal forma que se puede obtener un conjunto completo de eigenvectores:  $q_{i_{m \times 1}}$ ,  $i = 1, \dots, m$ , ortonormales<sup>4</sup>.

De la **ecuación 1** y la **Figura 1(a)**, es importante entender que cada columna  $t_{i_{n \times 1}}$  de  $T_{n \times r}$  tiene las coordenadas de proyección de los puntos  $y_1, \dots, y_n$  sobre  $q_i$ , y como resulta que la media y la variancia de  $t_{i_{n \times 1}}$  son iguales a cero y a  $\varphi_i^2/n$ , respectivamente, para  $i = 1, \dots, r$ ,  $r \leq m$ , entonces los valores  $\varphi_i^2$  se pueden asociar a las variancias de las proyecciones sobre  $q_i$ ,  $i = 1, \dots, r$ ,  $r \leq m$ , y las variancias relativas de las proyecciones se calculan como  $\varphi_i^2 / \sum_{j=1}^m \varphi_j^2$  ( $r = m$ ). Estas variancias relativas se suelen dar de forma porcentual. Entonces, cada vector  $q_i$  tiene un valor de variancia relativa<sup>5</sup>. A las variancias relativas se les asocia con cantidad de información [14].

La **Figura 1(b)** presenta una gráfica en un espacio “alternativo”  $r$ -dimensional, que tiene como ejes coordenados ortonormales a los vectores  $q_i$ ,  $i = 1, \dots, r$ , y con la que

<sup>4</sup> Es decir  $Q_{r \times m}^T Q_{m \times r} = I_r$ . Y en caso de que  $r = m$  se cumple que  $Q_{m \times m}^T Q_{m \times m} = I_m$ , donde  $I_r$  e  $I_m$  representan a matrices identidad de tamaño  $r$  y  $m$ , respectivamente.

<sup>5</sup> Además, la suma de las variancias de las columnas de  $T_{n \times r}$ , cuando  $r = m$ , es igual a  $m$ , o sea que la transformación no altera el valor total de la suma de variancias. También, la correlación entre cualesquiera dos diferentes columnas de  $T_{n \times r}$  es igual a cero.

se pretende facilitar la observación de los datos para encontrar la relación entre atributos, por ser un espacio de dimensión reducida, manteniendo la mayor cantidad de “información original”.

Esta técnica de transformación se llama *análisis de componentes principales* (PCA por sus siglas en inglés) [16,15], donde la palabra Componente Principal, se designa con las iniciales PC (por sus siglas en inglés).

En este trabajo, como en otros [17], cada componente principal  $PC(i)$ , corresponde al eje coordenado  $\mathbf{q}_i$  del plano “alternativo”, así que a cada  $PC(i)$  corresponde una variancia relativa (cantidad de información). En otros trabajos la nomenclatura cambia<sup>6</sup>.

#### 4.2 Descomposición en valores singulares

Esta subsección presenta brevemente la “descomposición en valores singulares”, para introducir el Biplot, que es la herramienta que se usa en este trabajo para descubrir las relaciones entre el IMC y otros atributos.

La “descomposición en valores singulares” se relaciona con PCA, y muestra que toda matriz de tamaño  $n \times m$ , como lo es  $Y_{n \times m}$ , se puede factorizar en un producto de tres matrices ([26] Pág. 284), ([14], Eq. 7.14), como lo indica<sup>7</sup> la **ecuación 2**, donde  $n > m$ :

$$Y_{n \times m} = P_{n \times n} \Phi_{n \times m} Q_{m \times m}^T = P_{n \times n} \begin{bmatrix} \varphi_1 & 0 & \cdots & 0 \\ 0 & \varphi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_m \\ \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} Q_{m \times m}^T. \quad (2)$$

La **ecuación 2** se puede reducir como en la **ecuación 3**:

$$Y_{n \times m} = P_{n \times m} \Phi_{m \times m} Q_{m \times m}^T = P_{n \times m} \begin{bmatrix} \varphi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi_m \end{bmatrix} Q_{m \times m}^T. \quad (3)$$

Los valores  $\varphi_i$ ,  $i = 1 \dots m$ , se llaman los valores singulares de  $Y_{n \times m}$ . Los elementos de las ecuaciones 2 y 3 se pueden ordenar, tal que  $\varphi_1 \geq \varphi_2 \geq \cdots \geq \varphi_m$ . De las ecuaciones 2 y 3 resulta la ecuación 4:

<sup>6</sup> Hay trabajos en donde la palabra “componente” se refiere a la proyección del punto  $\mathbf{y}_i$ ,  $i = 1, \dots, n$ , sobre el vector  $\mathbf{q}_j$ ,  $j = 1, \dots, r$ ,  $r \leq m$  (ver por ejemplo [15] Pág. 3), es decir que la proyección de  $\mathbf{y}_i$  sobre  $\mathbf{q}_j$  sería el componente  $j$  de  $\mathbf{y}_i$ . Cada elemento de esa proyección se llama *score*. Además, cada elemento de  $\mathbf{q}_j$  escalado se llama *loading* ([15] Pág. 3).

<sup>7</sup> Donde  $P_{n \times n}^T P_{n \times n} = I_n$  y  $Q_{m \times m}^T Q_{m \times m} = I_m$ .

$$Y_{m \times n}^T Y_{n \times m} Q_{m \times m} = Q_{m \times m} \Phi_{m \times m}^2$$

$$\Phi_{m \times m}^2 = \begin{bmatrix} \varphi_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi_m^2 \end{bmatrix}. \quad (4)$$

Poniendo la **ecuación 4** con más detalle resulta la **ecuación 5**:

$$Y_{m \times n}^T Y_{n \times m} [\mathbf{q}_{1 \times m} \quad \mathbf{q}_{2 \times m} \quad \cdots \quad \mathbf{q}_{r \times m}] =$$

$$= [\mathbf{q}_{1 \times m} \quad \mathbf{q}_{2 \times m} \quad \cdots \quad \mathbf{q}_{r \times m}] \begin{bmatrix} \varphi_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi_m^2 \end{bmatrix}. \quad (5)$$

En la **ecuación 5** se observa que los valores  $\varphi_i^2$  son los eigenvalores de  $Y_{m \times n}^T Y_{n \times m}$ , siendo las columnas  $\mathbf{q}_{i \times m}$ , de  $Q_{m \times m}$ , los correspondientes eigenvectores.

En la **ecuación 5** los valores  $\varphi_i$ ,  $i = r + 1, \dots, m$ , se suelen desechar para obtener la matriz  $\Phi'_{m \times m}$ , similar a  $\Phi_{m \times m}$  pero con  $\varphi_i = 0$ ,  $i = r + 1, \dots, m$ , obteniendo la aproximación  $Y_{(r) \times m}$  de  $Y_{n \times m}$ , expresada en la **ecuación 6**:

$$Y_{(r) \times m} = P_{n \times m} \Phi'_{m \times m} Q_{m \times m}^T. \quad (6)$$

La **ecuación 6** se puede simplificar aún más, como lo muestra la **ecuación 7**, donde  $\Phi'_{r \times r}$  es la parte de  $\Phi'_{m \times m}$  que tiene los elementos de la diagonal diferentes de cero:

$$Y_{(r) \times m} = P_{n \times r} \Phi'_{r \times r} Q_{r \times m}^T$$

$$Y_{(r) \times m} Q_{m \times r} = P_{n \times r} \Phi'_{r \times r}. \quad (7)$$

El segundo renglón de la **ecuación 7** resulta de post multiplicar la ecuación por  $Q_{m \times r}$ . Resulta que  $Y_{(r) \times m}$  tiene rango  $r$ . La “falta de ajuste” entre  $Y_{n \times m}$  y  $Y_{(r) \times m}$  es la suma de los cuadrados de todos los elementos de la matriz  $(Y_{n \times m} - Y_{(r) \times m})$ , y se representa como  $\|Y_{n \times m} - Y_{(r) \times m}\|^2$ , y es igual a  $\sum_{i=r+1}^m \varphi_i^2$ , así que mientras más pequeños sean los valores  $\varphi_i$ ,  $i = r + 1, \dots, m$ , la falta de ajuste es menor.

Se puede mostrar que  $Y_{n \times m} Q_{m \times r} = Y_{(r) \times m} Q_{m \times r}$ , y por eso, considerando la ecuaciones 1 y 7, podemos escribir la **ecuación 8**:

$$T_{n \times r} = Y_{n \times m} Q_{m \times r} = Y_{(r) \times m} Q_{m \times r} = P_{n \times r} \Phi'_{r \times r}. \quad (8)$$

### 4.3 Biplot

La matriz  $Y_{(r) \times m}$ , por tener rango  $r$ ,  $r \leq m$ , se puede representar como la multiplicación de dos matrices de rango  $r$ , que llamamos, matriz derecha, y matriz izquierda. Para esto, tomamos la **ecuación 7**, considerando también la **ecuación 8**, para formar la **ecuación 9**, donde  $L_{n \times r} = P_{n \times r} \Phi'_{r \times r} = T_{n \times r}$ , y  $R_{m \times r} = Q_{m \times r}$  (lo mismo

que  $R_{r \times m}^T = Q_{r \times m}^T$ , y donde las letras  $L$  proviene de *Left* (izquierda) y  $R$  proviene de *Right* (derecha) [27]. Hay más formas de conformar las matrices  $L$  y  $R$  (ver ecuaciones 2 a 5 en [28])<sup>8</sup>:

$$Y_{(r)n \times m} = P_{n \times r} \Phi'_{r \times r} Q_{r \times m}^T = [P_{n \times r} \Phi'_{r \times r}] Q_{r \times m}^T = T_{n \times r} Q_{r \times m}^T = L_{n \times r} R_{r \times m}^T. \quad (9)$$

La **ecuación 9** la podemos representar como en la **ecuación 10**, donde se indica el significado de las partes:

$$Y_{(r)n \times m} = T_{n \times r} Q_{r \times m}^T = \begin{bmatrix} t_{1n \times 1} & t_{2n \times 1} & \dots & t_{rn \times 1} \end{bmatrix} \begin{bmatrix} q_{11 \times m}^T \\ q_{21 \times m}^T \\ \vdots \\ q_{r1 \times m}^T \end{bmatrix} \quad (10)$$

Proyección punto 1, en  $r$  dimensiones  $\rightarrow \begin{bmatrix} t_{11} & t_{21} & \dots & t_{r1} \end{bmatrix}$

Proyección punto  $n$ , en  $r$  dimensiones  $\rightarrow \begin{bmatrix} t_{1n} & t_{2n} & \dots & t_{rn} \end{bmatrix}$

proyección vector (atributo 1) de  $r$  dimensiones  $\rightarrow \begin{bmatrix} q_{11} \\ q_{21} \\ \vdots \\ q_{r1} \end{bmatrix}$

proyección vector (atributo  $m$ ), de  $r$  dimensiones  $\rightarrow \begin{bmatrix} q_{1m} \\ q_{2m} \\ \vdots \\ q_{rm} \end{bmatrix}$

Ejes coordenados  $\rightarrow \begin{bmatrix} q_{11 \times m}^T \\ q_{21 \times m}^T \\ \vdots \\ q_{r1 \times m}^T \end{bmatrix}$

PC(1)  $\rightarrow \begin{bmatrix} q_{11 \times m}^T \\ q_{21 \times m}^T \\ \vdots \\ q_{r1 \times m}^T \end{bmatrix}$

PC( $n$ )  $\rightarrow \begin{bmatrix} q_{1n \times m}^T \\ q_{2n \times m}^T \\ \vdots \\ q_{rn \times m}^T \end{bmatrix}$

La **ecuación 10** contiene elementos que son proyecciones en  $r$  dimensiones de los puntos  $y_1, \dots, y_n$ ,  $m$ -dimensionales, y de los vectores  $q_1, \dots, q_r$ ,  $m$ -dimensionales. Estas proyecciones se ven en la gráfica de la **Figura 1-b**, que representa un espacio  $r$ -dimensional con  $r$  ejes coordenados ortonormales  $m$ -dimensionales  $q_j$ ,  $j = 1, \dots, r$ ,  $r \leq m$ , es decir, hay  $m$  ejes coordenados ortonormales  $q_j$ , pero solamente se usan  $r$  de ellos para la gráfica, la cual se llama Biplot [29,27,30]. El prefijo "bi" se refiere a que la gráfica tiene dos tipos de proyecciones: 1) "puntos" (individuos), y 2) "vectores" (atributos). El prefijo "bi" no quiere decir que un Biplot sea una gráfica bidimensional [30], aunque es común usar un Biplot bidimensional, es decir, con  $r = 2$ . Es frecuente que alguno de los dos grupos de elementos se grafique escalado para lograr tener en la gráfica elementos de un tamaño visual similar. Un Biplot que usa  $r = 2$  con  $q_1$  y  $q_2$ , se dice que es un Biplot con PC(1) y PC(2).

La forma de interpretar un Biplot es observar las proyecciones de los  $m$  vectores que representan a los  $m$  atributos. Si dos proyecciones son colineales en el mismo sentido, los atributos correspondientes se refuerzan; si las proyecciones son colineales en sentido contrario, los atributos correspondientes se contraponen; si las proyecciones son perpendiculares, los atributos correspondientes no están relacionados. La relación

<sup>8</sup> Con relación a la **ecuación 9**, considerando que  $r < m$ , no hay que caer en la tentación de volver a sustituir en círculo, tal que quedara  $Y_{(r)n \times m} = T_{n \times r} Q_{r \times m}^T = Y_{(r)n \times m} Q_{m \times r} Q_{r \times m}^T$  pues parecería que  $Q_{m \times r} Q_{r \times m}^T$  es igual a  $I_m$ , y no lo es, solamente se cumple que  $Q_{r \times m}^T Q_{m \times r} = I_r$ . Lo mismo pasa para la matriz  $P_{n \times r}$ , donde resulta que  $P_{n \times r} P_{r \times n}^T \neq I_n$ , y solamente se cumple que pues  $P_{r \times n}^T P_{n \times r} = I_r$ . Cuando  $r = m$  sí se cumplen todas estas expresiones.

de colinealidad y sentido se puede obtener mediante el producto punto de las proyecciones.

En el Biplot, también es de interés observar si las proyecciones de los puntos (individuos) clasificados según uno o más atributos de interés, se ubican en zonas dependiendo de los atributos elegidos. Estas zonas pueden tener una relación con la dirección y sentido de los diversos vectores (atributos), lo que es útil también para relacionar a los atributos.

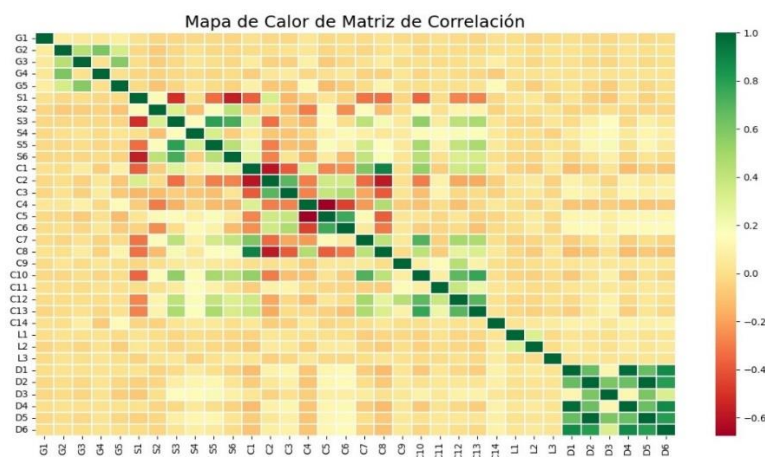
Es conveniente mencionar que para la realización de las operaciones matemáticas aquí presentadas, se usa el lenguaje Python con sus diversos componentes, como son los paquetes NumPy [31] y Pandas [32] y la biblioteca Scikit-learn (Sklearn) [33] de la cual se usan los paquetes: 1) preprocessing (del cual se usa la clase StandardScaler), y 2) decomposition (del cual se usa la clase PCA() que contiene algoritmos de aprendizaje de máquina no supervisado para reducción de dimensiones y la clase KernelPCA para trabajo con espacios no lineales). También se usa el módulo csv y otros paquetes para hacer representación gráfica, como Seaborn [34] y Matplotlib [35].

## 5. Resultados

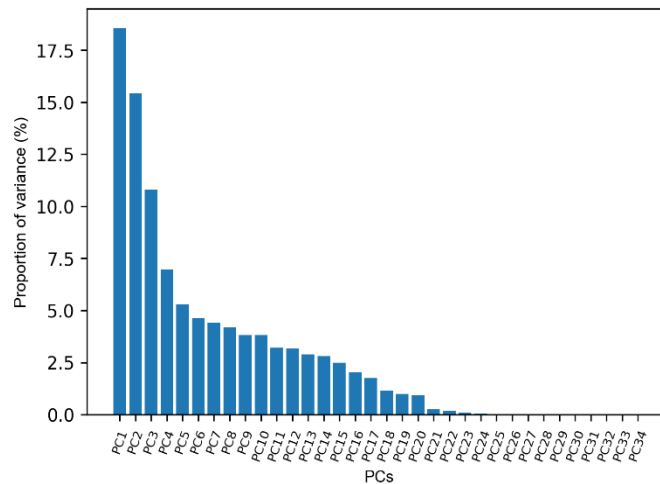
### 5.1 Exploración de las relaciones entre los atributos

Se tienen  $m = 34$  atributos, y  $n = 2485$  puntos (individuos), que forman la matriz de datos. En la **Figura 2** se observa el “mapa de calor” entre atributos, derivado de la matriz de datos, que indica las correlaciones entre los atributos. Con relación al atributo de interés G4 (IMC), no se observa relación con los demás atributos, excepto con G2 (peso).

A partir de la matriz de datos, se calculan los eigenvalores  $\varphi_i^2, i = 1, \dots, m$ . Por ejemplo  $\varphi_1^2 = 15,685.748$ ,  $\varphi_1^2/n = 6.312$ , que es la variancia de la primera columna de



**Fig. 2.** “Mapa de calor” entre atributos. Con relación al atributo G4 (IMC), no se observa claramente la relación que hay con los demás atributos, excepto con G2 (peso).



**Fig. 3.** Screeplot: Gráfica de barras de los valores relativos de los componentes principales.

$T_{n \times r}$  (de la ecuación 1), y  $\varphi_1^2 / \sum_{j=1}^m \varphi_j^2 = 0.186$ , que es el valor relativo porcentual de  $\varphi_1^2$  (o sea la información de PC(1)) es 18.6%. El valor relativo porcentual de  $\varphi_2^2$  (la información de PC(2)) es 15.447%, ..., y al final, el valor relativo porcentual de  $\varphi_{34}^2$  (la información de PC(34)) es  $1.046 \times 10^{-31}$  %. Estos valores se representan en la gráfica de la **Figura 3**, que se llama *Screeplot*<sup>9</sup>, que muestra los valores en forma de barras.

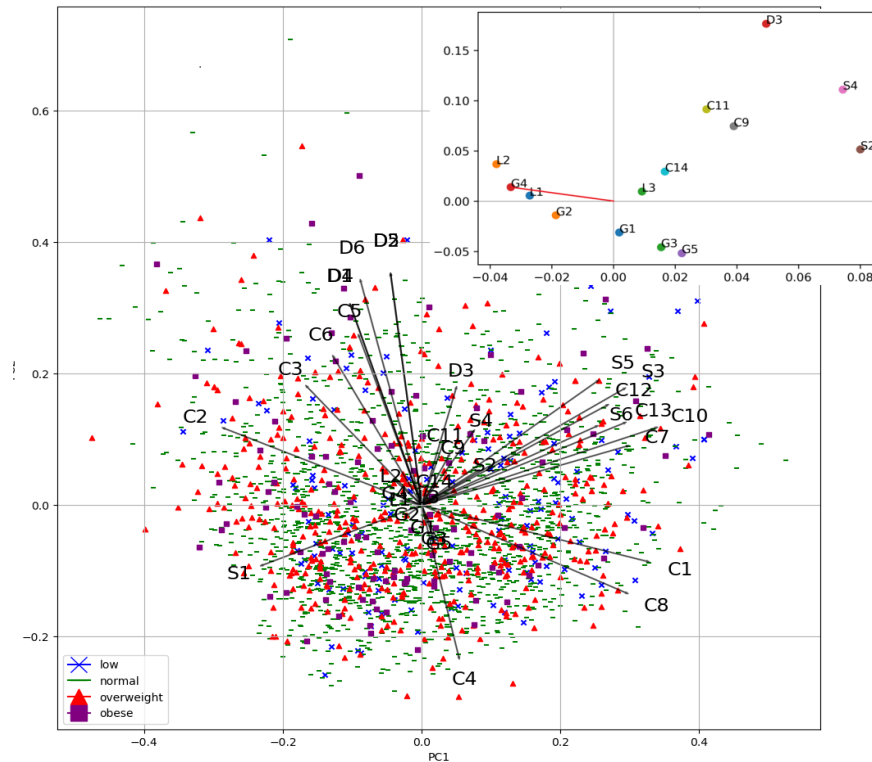
Los dos componentes principales, PC(1) y PC(2) reúnen un 34% de información (variancia relativa porcentual). Para reunir un 50% de información se requieren los primeros 4 componentes, y para reunir un 90% de información se requieren los primeros 14 componentes.

## 5.2 Resultados del análisis con los componentes PC (1) y PC (2)

El Biplot bidimensional ( $r = 2$ ) normalmente se realiza con los primeros 2 componentes PC(1)–PC(2), por ser los componentes de más información<sup>10</sup>. La **Figura 4** presenta el Biplot bidimensional con PC(1)–PC(2), con las proyecciones de  $m = 34$  vectores (que representan a 34 los atributos), y las proyecciones en dicho plano de  $n = 2485$  puntos (que representan a los individuos). Hay que notar que aun cuando los 34 vectores correspondientes a los atributos son ortonormales, sus proyecciones en el Biplot se ven oblicuas entre sí, y de tamaños diferentes. Los identificadores de algunas proyecciones de vector (atributo), que tienen similar dirección y sentido se

<sup>9</sup> La forma de la gráfica se parece a la forma de las piedras que se amontonan al lado de una montaña, que van bajando de altura conforme la distancia, que se llaman, en conjunto, *Scree* (del inglés).

<sup>10</sup> Dado que estos componentes no reúnen un porcentaje de información que llegue al 50%, se podría generar desconfianza sobre la validez representativa de los resultados.



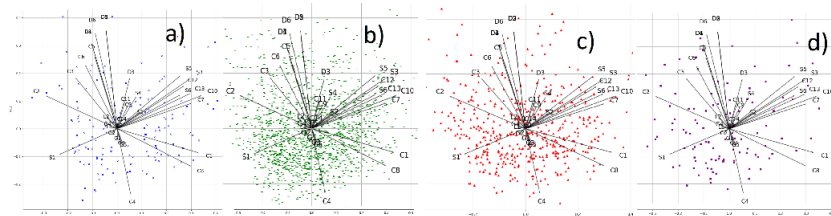
**Fig. 4.** Biplot bidimensional con PC(1)–PC(2)). Se observan las proyecciones de los 34 vectores que representan a los atributos, y de los 2485 puntos, que representan a los individuos. El símbolo de las proyecciones depende de su clasificación según el atributo G4 (IMC). La simbología es: Bajo peso × ; Peso normal – ; Sobre peso ▲ ; Obesidad ■ .

enciman: éste es el caso de los identificadores C7, C10, S3, S5, S6. Otros identificadores no se distinguen por estar muy aglomerados muy cercanos al centro del Biplot, como lo son L1, L2, L3, G1, G2, G3, G4, G5 y C14. Con relación a las proyecciones de los puntos, (individuos), se observa que hay mayor dispersión a lo largo del componente PC(1), que a lo largo del componente PC(2).

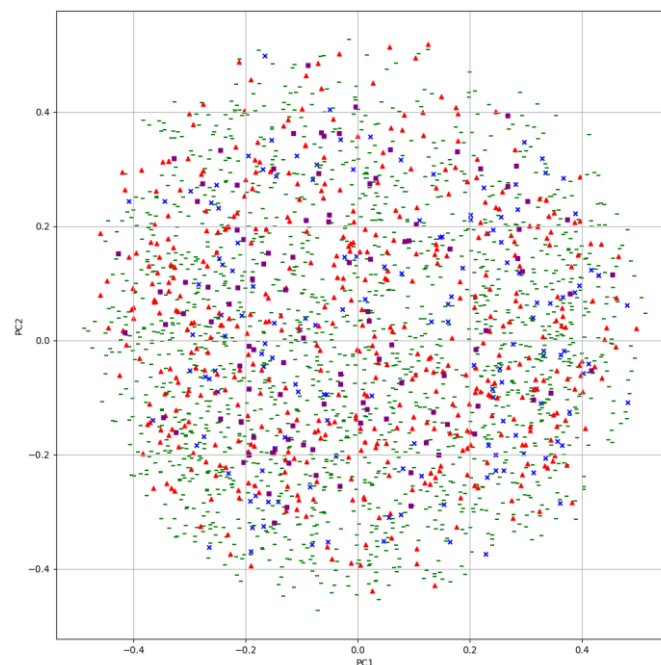
La **Figura 4** enfoca las proyecciones de los vectores más centrales (más pequeños), entre los que se encuentra G4 (IMC), cuya proyección es relativamente pequeña, lo que indica que este vector **es en gran medida perpendicular al plano PC(1)–PC(2)**. De aquí surgió la pregunta si este plano es significativo para observar lo que afecta a este atributo, aun cuando los componentes PC(1) y PC(2) son aquellos con más información.

Se dibujaron (marcaron) los puntos según los valores del atributo G4 (IMC), y no se observó que las proyecciones de los puntos en el Biplot se separaran según el valor de este atributo.

Las proyecciones, por separado, de los puntos (individuos) según G4 (IMC) en el Biplot bidimensional con PC(1)–PC(2)) se muestran la **Figura 5**.



**Fig. 5.** Cuatro imágenes de Biplot bidimensional con PC(1)–PC(2)), con las proyecciones de los 34 vectores (atributos) en cada imagen. Cada imagen contiene, además, las proyecciones de los puntos (individuos) clasificados según el atributo G4 (IMC). Las imágenes corresponden a: a) Bajo peso × ; b) Peso normal – ; c) Sobre peso ▲ ; d) Obesidad ■ .



**Fig. 6.** Biplot bidimensional con PC(1)–PC(2)), donde se observan las proyecciones de los 2485 puntos (que representan a los individuos), transformadas con el Kernel PCA Coseno. El símbolo de las proyecciones depende de su clasificación según el atributo G4 (IMC). La simbología es: Bajo peso × ; Peso normal – ; Sobre peso ▲ ; Obesidad ■ . No se observa separación de las proyecciones según el valor del atributo G4 (IMC).

Para buscar en el Biplot que las proyecciones de los puntos se separaran según el valor del atributo G4 (IMC), se usaron diversos Kernel PCA, que generalizan el análisis de PCA a dominios no-lineales. Los Kernels fueron: 1) Coseno, 2) Linear, 3) RBF, 4) Sigmoidal y 5) Polinomial, pero en ningún caso se logró obtener la separación buscada. En la Figura 6 se observa, como ejemplo, el Biplot referente al Kernel Coseno.

**Tabla 1.** Lista parcial de resultados del producto punto, de la proyección del vector del atributo G4 (IMC), con la proyección del vector de cada uno de los demás atributos, en el plano PC(1)–PC(2).

| Sigla | Prod. punto | Nombre de atributo                                                |
|-------|-------------|-------------------------------------------------------------------|
| C2    | 0.01108     | Tiempo entre cena y punto medio de sueño. (ideal más de 6 horas). |
| C3    | 0.00800     | Tiempo entre hora de cena y hora de dormir.                       |
| ...   | ...         | ...                                                               |
| L3    | -0.000169   | ¿Si te despiertas por la noche prendes la luz? "No"=0; "Otro"=1.  |
| C9    | -0.000244   | ABS( <i>Jet Lag</i> de comida principal).                         |
| ...   | ...         | ...                                                               |
| C10   | -0.009552   | ABS( <i>Jet Lag</i> de desayuno).                                 |
| C8    | -0.011672   | Promedio ventanas alimentación entre semana y fin de semana.      |
| C1    | -0.012005   | Ventana de alimentación.                                          |

Aún sin haber logrado observar alguna separación de las proyecciones de los puntos (individuos), con los componentes PC(1) y PC(2), según los valores del atributo G4 (IMC), en el Biplot, aquí se comparan las proyecciones de los vectores correspondientes a los atributos, para obtener un resultado de la relación de los atributos con relación al atributo G4 (IMC).

Para comparar la colinealidad y sentido de la proyección, en el plano PC(1)–PC(2), del vector correspondiente al atributo G4 (IMC), con la de los demás atributos, realizamos el producto punto de las proyecciones. La **Tabla 1** presenta una lista parcial de los resultados, tomados en orden descendente.

De acuerdo con la **Tabla 1**, podemos concluir que los atributos más influyentes en G4 son C2 y C3, los no relacionados son L3 y C9, y los más contrapuestos son C1, C8, C10.

### 5.3 Búsqueda de los componentes más significativos para G4 (IMC)

Ya que la proyección del vector correspondiente al atributo G4 (IMC) sobre el plano PC(1)–PC(2) fue más perpendicular a ese plano que la mayoría de los vectores de los otros atributos, buscamos otro plano en donde el vector correspondiente al atributo G4 (IMC) se proyecte mayormente.

Para esto, observamos la matriz  $Q_{r \times m}^T$ , para el caso de  $r = m$  (o sea  $Q_{m \times m}^T$ ), que indica proyecciones. Observamos la cuarta columna de la matriz  $Q_{r \times m}^T$ , que indica las proyecciones del vector relativo a G4 (IMC) sobre cada uno de componentes a PC(1), ..., P(34). Algunas proyecciones relevantes se ven en la **Tabla 2**, que contiene, para cada PC(*i*), su información (variancia relativa porcentual). La suma de los cuadrados de las 34 proyecciones es igual a 1.

Encontramos que donde se tiene la mayor proyección es en PC(26), de 0.5707, pero este eje coordenado tiene una variancia de  $\varphi_{26}^2/n = 0.00398$ , para una variancia relativa porcentual de 0.0117%, que es un valor de información “demasiado” pequeño, con relación a lo que tienen otros PC(*i*). Encontramos una mayor variancia relativa en PC(8) y PC(4), con valores de proyección, respectivamente, de  $-0.4823$  y de  $0.3668$ ,

**Tabla 2.** Proyección del vector relativo al atributo G4 (IMC), sobre algunos de los 34 PC(*i*), seleccionados. Se ordenan los renglones por valor del indicador que se propone en este trabajo.  $G4_{PC(i)}$  es la proyección.

| Consecutivo | Variancia Relativa (%)                              |                                | Indicador( <i>i</i> ) | PC( <i>i</i> ) | G4 <sub>PC(<i>i</i>)</sub> |
|-------------|-----------------------------------------------------|--------------------------------|-----------------------|----------------|----------------------------|
|             | $100 \times \varphi_i^2 / \sum_{j=1}^m \varphi_j^2$ | $= 100 \times \varphi_i^2 / m$ |                       |                |                            |
| 1           | 4.2091%                                             |                                | 5.44E-01              | PC(8)          | -0.482330806               |
| 2           | 6.9690%                                             |                                | 4.89E-01              | PC(4)          | 0.366788251                |
| 3           | 3.8271%                                             |                                | 3.68E-01              | PC(9)          | -0.337078093               |
|             |                                                     |                                |                       |                |                            |
| 11          | 0.0117%                                             |                                | 9.05E-02              | PC(26)         | 0.570684596                |
| 12          | 18.5577%                                            |                                | 6.15E-02              | PC(1)          | -0.033285789               |
| 15          |                                                     |                                |                       |                |                            |
| 16          | 15.4472%                                            |                                | 2.45E-02              | PC(2)          | 0.0140993                  |
|             |                                                     |                                |                       |                |                            |
| 32          | 0.0000%                                             |                                | 1.43E-26              | PC(32)         | -4.16439E-16               |
| 33          | 0.0000%                                             |                                | 9.49E-27              | PC(33)         | -2.78152E-16               |
| 34          | 0.0000%                                             |                                | 4.90E-27              | PC(34)         | -1.48889E-16               |

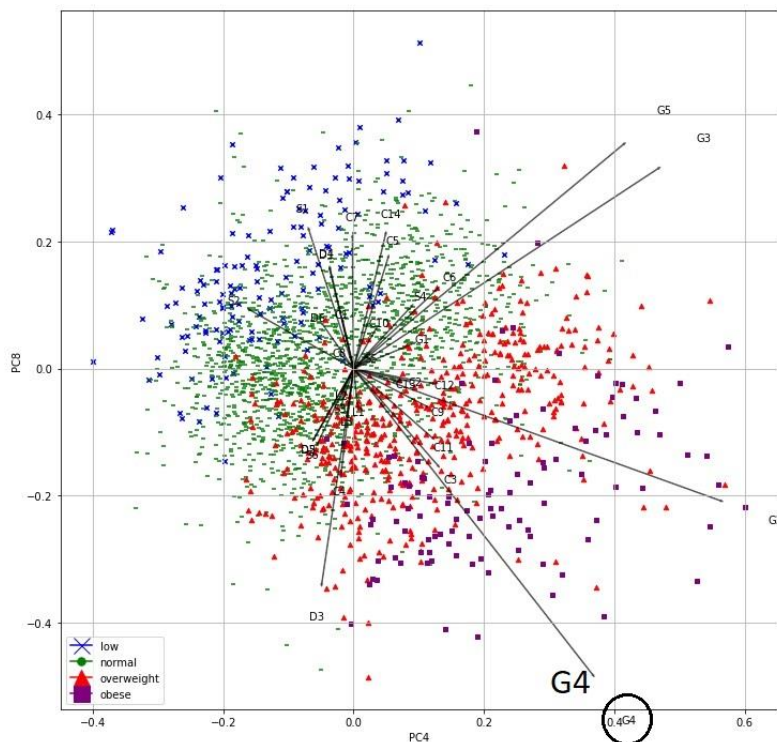
que tienen una variancia relativa, respectivamente, de 4.2091% y de 6.9690%. Por lo anterior, para establecer un balance entre proyección y variancia relativa, proponemos un indicador con la expresión 11, donde  $|G4_{PC(i)}|$  significa la magnitud de la proyección del vector correspondiente al atributo G4 (IMC) sobre PC(*i*). La razón de aplicar raíz cúbica al valor  $\varphi_i^2/n$  es disminuir su relevancia por ser muy pequeño. Según este indicador, los valores mayores fueron para PC(8) y PC(4), respectivamente. Los renglones de la **Tabla 2** se ordenan según el valor de este indicador:

$$\text{Indicador}(i) = \sqrt[3]{\varphi_i^2/n} \times |G4_{PC(i)}|. \quad (11)$$

Aun cuando los ejes coordenados PC(4) y PC(8) tienen una cantidad de información en conjunto de 11.18%, ésta es una información que se sospecha ser más significativa debido a la mayor proyección del vector correspondiente al atributo G4 (IMC) en el plano PC(4)-PC(8).

La **Figura 7** presenta el Biplot bidimensional con PC(4)–PC(8). **Es sorprendente** observar que en este Biplot sí se distinguen, “a simple vista”, zonas de localización para las proyecciones de los puntos por cada categoría de IMC (bajo peso, peso normal, sobrepeso y obesidad).

Las proyecciones de los puntos de bajo peso se concentran lejanas del centro y arriba a la izquierda, las de peso normal están centradas arriba a la izquierda, las de sobrepeso están centradas abajo a la derecha, y las de obesidad están lejanas del centro y abajo a la derecha. Se observa que la proyección del vector relativo al atributo G4 (IMC) apunta hacia las zonas donde están las proyecciones de los puntos de sobrepeso, y de obesidad.



**Fig. 7.** Biplot bidimensional con PC(4) y PC(8). Se observa la proyección de los 34 vectores que representan a los atributos, y de los 2485 puntos que representan a los individuos, clasificados según el atributo G4 (IMC). Para las proyecciones de los puntos (individuos) la simbología es: Bajo peso  $\times$ ; Peso normal  $\bullet$ ; Sobrepeso  $\blacktriangle$ ; Obesidad  $\blacksquare$ .

Realizamos la operación producto punto de la proyección, en el plano PC(4)–PC(8), del vector correspondiente al atributo G4 (IMC) por cada una de las proyecciones de los vectores correspondientes a los demás atributos. La **Tabla 3** presenta una lista parcial de resultados de esta operación, que no incluye el resultado relativo al atributo G4 (IMC), que tiene el mayor valor: 0.3671.

De acuerdo con la **Tabla 3**, podemos concluir que los atributos más influyentes en G4 (IMC) son G2, D3, C3, los no relacionados son L2, S5, C6, y los más contrapuestos son C7, S2 y S1.

Observamos que los resultados de proyectar en el plano PC(1)–PC(2) son diferentes a aquellos de proyectar en el plano PC(4)–PC(8).

## 6. Conclusiones y trabajo futuro

La descomposición por componentes principales podría no arrojar resultados útiles tan solo seleccionando a los componentes de más información, pero puede sí puede

**Tabla 3.** Lista parcial de resultados de la operación producto punto de la proyección del vector correspondiente al atributo G4 (IMC), en el plano PC(4)–PC(8), con la proyección de cada uno de los vectores correspondientes a los demás atributos.

| Sigla | Prod. punto  | Nombre de atributo                                                                                   |
|-------|--------------|------------------------------------------------------------------------------------------------------|
| G2    | 0.306754286  | Peso Kg.                                                                                             |
| D3    | 0.145614712  | Diferencia, en cama, antes dormir, uso celular o disp. Elec (entre semana a fin de semana), minutos. |
| C3    | 0.120905029  | Tiempo entre hora de cena y hora de dormir (entre semana).                                           |
| C11   | 0.096623535  | ABS( <i>Jet Lag</i> cena).                                                                           |
| ...   | ...          | ...                                                                                                  |
| L2    | 0.011971867  | Luz al dormir: "Obscuro"=0; "Tenue"=1; "Luz"=2; "Brillante"=3.                                       |
| S5    | 0.00201177   | ABS( <i>Jet Lag</i> punto medio sueño).                                                              |
| C6    | -0.013398287 | Tiempo entre hora de cena y hora de dormir (fin de semana).                                          |
| ...   | ...          | ...                                                                                                  |
| C7    | -0.10035604  | Diferencia ventanas alimentación (entre semana vs fin de semana).                                    |
| S2    | -0.102823015 | Horas dormido (fin de semana).                                                                       |
| S1    | -0.130563624 | Horas dormido (entre semana).                                                                        |

hacerlo al seleccionar componentes principales más enfocados al atributo que sea de interés. Esto pasó en este trabajo al seleccionar los componentes principales PC(4) y PC(8), que reunían apenas poco más del 11% de la información, en comparación de seleccionar los a los componentes principales con la mayor información, PC(1) y PC(2), que reunían el 34%, de la información. La utilidad se apreció en el Biplot bidimensional con PC(4)–PC(8), donde las proyecciones de los puntos (que representan a los individuos) ocuparon zonas según el atributo de interés, que indica el sobre peso y la obesidad, G4 (IMC).

Esta ubicación por zonas no se observó al trabajar con un Biplot bidimensional con PC(1) y PC(2). La clave para seleccionar a PC(4) y PC(8) fue el balance entre la proyección del vector asociado al atributo G4 (IMC), con la información de esa proyección. Para esto, en este trabajo se propuso un indicador que hiciera ese balance.

En próximos trabajos se puede realizar un estudio en un Biplot de tres dimensiones. También, se puede seguir analizando la utilidad del identificador que hemos presentado en la ecuación 11.

Según los autores del grupo de estudio de la ciencia de medicina, los resultados del análisis en las dos proyecciones coinciden con observaciones de estudios clínicos que indican que diversos hábitos como factores de riesgo para la salud promueven el desarrollo de sobrepeso y obesidad [36]. El atributo C3 (Tiempo entre hora de cena y hora de dormir entre semana) sale como reforzante en el análisis de las dos proyecciones, lo cual ubica como un factor más relevante entre los demás analizados.

Es importante considerar también que la disrupción circadiana no necesariamente incide en ganar peso; en algunos individuos puede incidir en perderlo. Esta situación puede confundir. Por esto, es importante perfeccionar las consideraciones del estudio.

## Apéndice. Lista de atributos de los datos

Cada atributo está numerado y entre paréntesis identificado por una letra que puede contener una ecuación si el atributo se deriva de otros. Si hay otro paréntesis es que el atributo se seleccionó para análisis: el paréntesis contiene el número de atributo seleccionado y una sigla de identificación de dos letras. La letra de inicio de la sigla indica el tipo de atributo: G → información general; S → sueño (horas de dormir); C → consumo (alimentación); L → luz (iluminación).

- **Atributos de identificación.** 1- Fecha ( $A$ ); 2- Carrera ( $B$ ); 3- Matricula ( $C$ ) (ID);
- **Atributos de información general.** 4- Edad años ( $D$ ) (1- G1); 5- Peso Kg ( $E$ ) (2- G2); 6- Estatura metros ( $F$ ) (3- G3); 7- IMC ( $G = E/F^2$ ) (4- G4); 8- Sexo: Fem.=1, Masc.=2 ( $H$ ) (5- G5);
- **Atributos marcadores de disrupción circadiana (sueño entre semana).** 9- Me despierto a las: ( $I$ ); 10- Me duermo a las: ( $J$ ); 11- Horas despierto ( $K = J - I$ ); 12- Horas dormido ( $L = 24 - K$ ) (6- S1); 13- Punto medio de sueño ( $M = J + L/2$ ); 14- Punto medio de sueño, horas de la mañana ( $N = M - 24$ );
- **Atributos marcadores de disrupción circadiana (sueño fin de semana).** 15- Me despierto a las: ( $O$ ); 16- Me duermo a las: ( $P$ ); 17- Horas despierto ( $Q = P - O$ ); 18- Horas dormido ( $R = 24 - Q$ ) (7- S2); 19- Punto medio de sueño ( $S = P + R/2$ ); 20- Punto medio de sueño horas de la mañana ( $T = S - 24$ );
- **Atributos marcadores de disrupción circadiana (sueño en general).** 21- Jet Lag hora de despertar ( $U = O - I$ ); 22- ABS(Jet Lag hora de despertar) ( $V = ABS(U)$ ) (8- S3); 23- Jet Lag hora de dormir ( $W = P - J$ ); 24- ABS(Jet Lag hora de dormir) ( $X = ABS(W)$ ) (9- S4); 25- Jet Lag punto medio de sueño ( $Y = S - M$ ); 26- ABS(Jet Lag punto medio de sueño) ( $Z = ABS(Y)$ ) (10- S5); 27- Diferencia de horas dormido ( $AA = R - L$ ) (11- S6);
- **Atributos marcadores de disrupción circadiana (consumo entre semana).** 28- Desayuno a las: ( $AB$ ); 29- Como a las: ( $AC$ ); 30- Ceno a las: ( $AD$ ); 31- Ventana de alimentación ( $AE = AD - AB$ ) (12- C1); 32- Punto medio de alimentación ( $AF = AB + AE/2$ ); 33- Tiempo entre cena y punto medio de sueño (ideal más de 6 horas) ( $AG = M - AD$ ) (13- C2); 34- Tiempo entre hora de cena y hora de dormir ( $AH = J - AD$ ) (14- C3);
- **Atributos marcadores de disrupción circadiana (consumo fin de semana).** 35- Desayuno a las: ( $AI$ ); 36- Como a las: ( $AJ$ ); 37- Ceno a las: ( $AK$ ); 38- Ventana de alimentación ( $AL = AK - AI$ ) (15- C4); 39- Punto medio de alimentación ( $AM = AI + AL/2$ ); 40- Tiempo entre cena y punto medio de sueño (ideal más de 6 horas) ( $AN = S - AK$ ) (16- C5); 41- Tiempo entre hora de cena y hora de dormir ( $AO = P - AK$ ) (17- C6);
- **Atributos marcadores de disrupción circadiana (consumo en general).** 42- Diferencia ventanas alimentación ( $AP = AE - AL$ ) (18- C7); 43- Promedio

- ventanas alimentación ( $AQ = (AE \times 5 + AL \times 2)/7$ ) (19- C8); 44- Jet Lag de comida principal ( $AR = AJ - AC$ ); 45- ABS(Jet Lag de comida principal) ( $AS = ABS(AR)$ ) (20- C9); 46- Jet Lag de desayuno ( $AT = AI - AB$ ); 47- ABS(Jet Lag de desayuno) ( $AU = ABS(AT)$ ) (21- C10); 48- Jet Lag cena ( $AV = AK - AD$ ); 49- ABS(Jet Lag cena) ( $AW = ABS(AV)$ ) (22- C11); 50- Promedio de Jet Lag metabólico ( $AX = (AS + AU + AW)/3$ ) (23- C12); 51- Jet Lag metabólico punto medio de alimentación ( $AY = AM - AF$ ); 52- ABS(Jet Lag metabólico punto medio de alimentación) ( $AZ = ABS(AY)$ ) (24- C13); 53- ¿Si te despiertas por la noche comes? ( $BA$ ); 54- ¿Si te despiertas por la noche comes? (numérico): "Nunca"=0, "A veces"=1 ( $BB$ ) (25- C14);
- **Atributos marcadores de disrupción circadiana (iluminación en general –solo por las noches).** 55- ¿La luz de calle ilumina tu cuarto? ( $BC$ ); 56- ¿La luz de calle ilumina tu cuarto? (numérico): "No"=0, "Un poco"=1, "Sí"=2 ( $BD$ ) (26- L1); 57- Luz al dormir ( $BE$ ); 58- Luz al dormir (numérico): "Obscuro"=0, "Tenue"=1, "Luz"=2, "Brillante"=3 ( $BF$ ) (27- L2); 59- ¿Si te despiertas por la noche prendes la luz? ( $BG$ ); 60- ¿Si te despiertas por la noche prendes la luz? (numérico): "No"=0, "Otro"=1 ( $BH$ ) (28- L3); 61- Puntuación total exposición luz ( $BI = BD + BF + BH$ );
  - **Atributos marcadores de disrupción circadiana (iluminación de dispositivos en general).** 62- En cama, antes dormir, uso celular o dispositivo electrónico, minutos (entre semana) ( $BJ$ ) (29- D1); 63- En cama, antes dormir, uso celular o dispositivo electrónico, minutos (fines semana) ( $BK$ ) (30- D2); 64- Diferencia, en cama, antes dormir, uso celular o dispositivo electrónico, minutos ( $BL = BK - BJ$ ) (31- D3); 65- En cama, antes dormir, total uso celular o dispositivo electrónico, minutos (entre semana) ( $BM = BJ \times 5$ ) (32- D4); 66- En cama, antes dormir, total uso celular o dispositivo electrónico, minutos (fines semana) ( $BN = BK \times 2$ ) (33- D5); 67- En cama, antes dormir, total uso celular o dispositivo electrónico, minutos ( $BO = BM + BN$ ) (34- D6).

## Referencias

1. Shrestha, P., Ghimire, L.: A Review about the Effect of Life Style Modification on Diabetes and Quality of Life. *Global Journal of Health Science*, vol. 4, no. 6, pp. 185–90 (2012). DOI: 10.5539/gjhs.v4n6p185.
2. Egger, G., Dixon, J.: Beyond Obesity and Lifestyle: A Review of 21st Century Chronic Disease Determinants. *BioMed Research International*, vol. 2014, pp. 731685 (2014). DOI: 10.1155/2014/731685.
3. Stoner, L., Castro, N., Signal, L., Skidmore, P., Faulkner, J., Lark, S., Williams, M.A., Muller, D., Harrex, H.: Sleep and Adiposity in Preadolescent Children: The Importance of Social Jetlag. *Childhood obesity* (Print), vol. 14, no. 3, pp. 158–164 (2018). DOI: 10.1089/chi.2017.0272.
4. Gill, S., Panda, S.: A Smartphone App Reveals Erratic Diurnal Eating Patterns in Humans that Can Be Modulated for Health Benefits. *Cell metabolism*, vol. 22, no. 5, pp. 789–798, (2015). DOI: 10.1016/j.cmet.2015.09.005.
5. Escobar, C., Ángeles-Castellanos, M., Espitia Bautista, E.N., Buijs, R.M.: Food During the Night is a Factor Leading to Obesity. *Revista Mexicana de Trastornos Alimentarios*, vol. 7, no. 1, pp. 78–83 (2016). DOI: 10.1016/j.rmta.2016.01.001.

6. Touitou, Y., Reinberg, A., Touitou, D.: Association Between Light at Night, Melatonin Secretion, Sleep Deprivation, and the Internal Clock: Health Impacts and Mechanisms of Circadian Disruption. *Life Sciences*, vol. 173, pp. 94–106 (2017). DOI: 10.1016/j.lfs.2017.02.008.
7. Guerrero-Vargas, N.N., Ángeles-Castellanos, M., Escobar Briones, C.: Los efectos adversos de la luz artificial por la noche. *Revista Digital Universitaria*, vol. 19, no. 3, pp. 1–18 (2018). DOI: 10.22201/codeic.16076079e.2018.
8. Ayala, G.X., Baquero, B., Klinger, S.: A Systematic Review of the Relationship Between Acculturation and Diet Among Latinos in the United States: Implications for Future Research. *Journal of the American Dietetic Association*, vol. 108, no. 8, pp. 1330–1344 (2008). DOI: 10.1016/j.jada.2008.05.009.
9. Miller, M.A., Kruisbrink, M., Wallace, J., Ji, C., Cappuccio, F.P.: Sleep Duration and Incidence of Obesity in Infants, Children, and Adolescents: A Systematic Review and Meta-Analysis of Prospective Studies. *Sleep*, vol. 41, no. 4 (2018). DOI: 10.1093/sleep/zsy018.
10. Roenneberg, T., Allebrandt, K.V., Merrow, M., Vetter, C.: Social jetlag and obesity. *Current Biology: CB*, vol. 22, no. 10, pp. 939–943 (2012). DOI: 10.1016/j.cub.2012.03.038.
11. Escobar, C., Ángeles-Castellanos, M., Miñana-Solís, M.C., Salgado Delgado, R.: Disturbance of Circadian Rhythms as a Predisposing Factor for Obesity and Metabolic Disease, de *Advances in Obesity-diabetes. Research at UNAM, El Manual Moderno*, pp. 1–17 (2010)
12. UNAM: Oferta Académica UNAM. Médico Cirujano, UNAM, 2022. [En línea]. Available: <http://oferta.unam.mx/medico-cirujano.html>. [Último acceso: 21 10 2022].
13. Stewart, G.W.: On the Early History of the Singular Value Decomposition, *SIAM Review*, vol. 35, p. 551–566 (1993). DOI: 10.1137/1035134.
14. Borg, I., Groenen, P.J.F.: *Modern Multidimensional Scaling. Theory and Applications*, Ney York: Springer (2005)
15. Jolliffe, I.T., Cadima, J.: Principal Component Analysis: A Review and Recent Developments, *Philosophical Transactions of the Royal Society A: Mathematical. Physical and Engineering Sciences*, vol. 374, p. 20150202 (2016). DOI: 10.1098/rsta.2015.0202.
16. Shlens, J.: A Tutorial on Principal Component Analysis, *CoRR*, vol. abs/1404.1100 (2014). DOI: 10.48550/arXiv.1404.1100.
17. Shalizi, C.R.: *Advanced Data Analysis from an Elementary Point of View (Principal Component Analysis chapter)*, Carnegie Mellon University, [on line]. Available: <https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/> [Último acceso: 18 10 2022].
18. Subu, M.A., Al-Yateem, N., Waluyo, I., Dias, J.M., Rahman, S.A., Agustino, R., Ahamed, I.S., Al Marzooqi, A.: Relationship Between Internet Gaming Addiction and Body Mass Index Status Among Indonesian Junior High School Students. *IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)* pp. 1391–1393 (2021). DOI: 10.1109/COMPSAC51774.2021.00200.
19. Johnson, K.A., Gordon, C.J., Chapman, J.L., Hoyos, C.M., Marshall, N.S., Miller, C.B., Grunstein, R.R.: The Association of Insomnia Disorder Characterised by Objective Short Sleep Duration with Hypertension, Diabetes and Body Mass Index: A Systematic Review and Meta-analysis. *Sleep Medicine Reviews*, vol. 59, pp. 101456 (2021). DOI: 10.1016/j.smrv.2021.101456.
20. Croenlein, T., Langguth, B., Busch, V., Rupperecht, R., Wetter, T.: Severe Chronic Insomnia is not Associated with Higher Body Mass Index. *Journal of Sleep Research*, vol. 24, no. 5, pp. 514–517 (2015). DOI: 10.1111/jsr.12294.

21. Bocicor, A.E., Buicu, G., Sabau, D., Varga, A., Tilea, I., Gabos-Grecu, I.: Association Between Sleep Disorder and Increased Body Mass Index in Adult Patients. *Acta Marisiensis - Seria Medica*, vol. 62, p. 221–224 (2016). DOI: 10.1515/amma-2016-0015.
22. Duranková, S., Csanády, A., Ždiľová, A., Bernasovská, J., Buková, A.: Effects of lifestyle on Physical Health in Slovak University Students. *Anthropological Review*, vol. 83, pp. 129–142 (2020). DOI: 10.2478/anre-2020-0010.
23. Imran, M., Khan, N., Shah, A.A., Ahmad, I.: Overweight and Obesity Prevalence Pattern and Associated Risk Factors Among the People of Malakand Division, Khyber Pakhtunkhwa Pakistan. *Arabian Journal for Science and Engineering*, vol. 44 (2018)
24. Zhang, Y., Li, W.: A Survey of The Relationship Between Human Faces And Body Mass Index (BMI), *IEEE Globecom Workshops*, pp. 1–6 (2021). DOI: 10.1109/GCWkshps52748.2021.9682060.
25. Mark, G.M., Bakunzibake, P., Mikeka, C.: Design of an IoT-based Body Mass Index (BMI) Prediction Model. 4th International Seminar on Research of Information Technology and Intelligent Systems, pp. 629–634 (2021). DOI: 10.1109/ISRITI54043.2021.9702866.
26. Leon, S.J.: *Linear Algebra with Applications*, 1 éd., New York: Macmillan Publishing Co., p. 338 (1980)
27. Gabriel, K.R.: The Biplot Graphic Display of Matrices with Application to Principal Component Analysis, *Biometrics*, vol. 58, no. 3, pp. 453–467 (1971)
28. Greenacre, M.J.: Biplots: the Joy of Singular Value Decomposition, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 4, pp. 399–406 (2012). DOI: 10.1002/wics.1200
29. Bedre, R.: Principal Component Analysis (PCA) and Visualization Using Python (Detailed Guide with Example), Creative Commons (<https://creativecommons.org/>), 07 11 2021. [on line]. Available: <https://www.reneshbedre.com/blog/principal-component-analysis.html>. [Último acceso: 07 07 2022].
30. Greenacre, M.: *Biplots in Practice*, B. Foundation, Ed., Bilbao: Spain (2010)
31. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array Programming with NumPy, *Nature*, vol. 585, pp. 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.
32. McKinney, W. et al.: Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference* (2010). DOI: 10.25080/Majora-92bf1922-00a.
33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830 (2011)
34. Waskom, M.L.: Seaborn: Statistical Data visualization, *Journal of Open Source Software*, vol. 6, pp. 3021 (2021). DOI: 10.21105/joss.03021.
35. Hunter, J.D.: Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, vol. 9, p. 90–95 (2007). DOI: 10.1109/MCSE.2007.55.
36. McHill, A.W., Wright, K.P.J.: Role of Sleep and Circadian Disruption on Energy Expenditure and in Metabolic Predisposition to Human Obesity and Metabolic Disease, *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity*, vol. 18, 1, pp. 15–24 (2017). DOI: 10.1111/obr.12503.