Aplicando Análisis de Componentes Principales en la Composición de Alimentos para Rumiantes

Autor: Fabricio Javier Rivadeneira Zambrano
Universidad Laica Eloy Alfaro de Manabí, **ULEAM**
fabricio.rivadeneira@uleam.edu.ec
Manta, Ecuador

Rodolfo Andrés Rivadeneira Zambrano
Universidad Técnica de Manabí, **UTM**
rodolfo.rivadeneira@utm.edu.ec
Portoviejo, Ecuador

Silvia Mercedes Carvajal Rivadeneira
Unidad Educativa Julio Pierregrosse, **UEJP**
scarvajal@juliopierregrosse.edu.ec
Manta, Ecuador

Viviana Katiuska García Macías
Universidad Laica Eloy Alfaro de Manabí, **ULEAM**
viviana.garcia@uleam.edu.ec
Manta, Ecuador

DOI: https://doi.org/10.56124/encriptar.v7i14.008

Resumen

En el siguiente trabajo se presenta la aplicación del método de Análisis de Componentes Principales (PCA) empleado al análisis de variables cuantitativas para la reducción de dimensiones mediante la descomposición de la matriz de correlaciones en sus vectores propios y valores propios, aunque se pueden usar otros métodos de descomposición como SVD (Singular Value Decomposition). Esté método se aplica en los datos referentes a la composición nutricional de 150 alimentos o ingredientes para rumiantes, la composición de estos alimentos analizados en laboratorio conforma una tabla de 12 variables o columnas, de la cuales 8 son variables cuantitativas utilizadas en el análisis PCA, que representan los principales nutrientes necesarios para los rumiantes como son: porcentaje de Materia Seca, de Digestibilidad de Materia Seca, de Proteína Bruta, porcentaje de Proteína Degradable en el Rumen, de Fibra Detergente Neutro, porcentaje de Fibra, de Calcio, de Fósforo y, Energía Metabólica. Obteniendo como resultado la reducción de dimensión de la tabla de composición de alimentos e identificando cuatro ejes o componentes principales como factores importantes de nutrientes que inciden en la calidad de los alimentos para rumiantes.

Palabras clave: análisis en componentes principales; composición de alimentos; reducción de dimensión.

Applying Principal Component Analysis to the Composition of Ruminant Feeds

ABSTRACT

The following paper presents the application of the Principal Component Analysis (PCA) method used to analyze quantitative variables for dimension reduction by decomposing the correlation matrix into its eigenvectors and eigenvalues, although other decomposition methods such as SVD (Singular Value Decomposition) can be used.

This method is applied to data relating to the nutritional composition of 150 foods or ingredients for ruminants. The composition of these foods analyzed in the laboratory forms a table of 12 variables or columns, of which 8 are quantitative variables used in the PCA analysis, which represent the main nutrients needed by ruminants, such as: percentage of Dry Matter, Dry Matter Digestibility, Crude Protein, percentage of Rumen Degradable Protein, Neutral Detergent Fiber, percentage of Fiber, Calcium, Phosphorus and Metabolic Energy. The result is a reduction in the size of the feed composition table and four main axes or components are identified as important nutrient factors that affect the quality of feed for ruminants.

Keywords: principal component analysis; food composition; dimension reduction.

1. Introducción

Este trabajo busca identificar unos ejes principales a través del método de Análisis de Componentes Principales, que relacionan los diferentes nutrientes que se encuentran en los ingredientes utilizados en la alimentación de los rumiantes. Los nutrientes tienen un rol fundamental en la selección y costo de los alimentos y por ende influyen en la salud del animal, impactando en forma positiva su desarrollo y producción, por lo que es importante su análisis en la preparación de alimentos (Chibisa & Oba, 2020).

El método de análisis multivariable llamado Análisis de Componentes Principales o PCA por sus siglas en inglés de Principal Component Analysis, fue introducido por (Pearson, 1901) y formulada con rigor por (Hotelling, 1933) y, se aplica a tabla de datos descritas por ***p*** variables numéricas X_j.

$$X_{nxp} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Donde x_{ij} representa el valor de la variable *j* para el individuo *i*, $i = 1, \ldots, n$, $j=1, \ldots, p$.

El método PCA (Géron, 2022) transforma variables correlacionadas en nuevas variables no correlacionadas que serán combinaciones lineales de las variables originales previamente centradas o estandarizadas con varianza relevante o varianza irrelevante. Por lo que este procedimiento permitirá obtener una nueva tabla de datos *Y* reducida, y simplificará la interpretación de la información multivariada disponible.

$$X_{nxp} \rightarrow Y_{nxq}\,, \qquad con\ q \leq p$$

Por lo cual, se recabó datos de los nutrientes como: fibra, proteínas, grasas, calcio, fosforo y otros, de 150 alimentos que pertenecen a los grupos

de granos como: avena, maíz, cebada, sorgo, etc.; de grupos de fardo como: base alfalfa; de pastura; de grupos de silaje como: maíz picado grueso; de grupos de rollos como: el trébol rojo; de subproductos como: cascara de algodón, grasa y, de minerales como: el carbonato de calcio, conchilla entre otros. (Moraes & Fadel, 2020)

Se utiliza PCA para combinar las variables continuas que representa a cada uno de los nutrientes en los ingredientes, en un número más reducido de variables completamente no correlacionadas. Cabe recalcar que PCA es sólo una de tantas técnicas de análisis multivariantes, pero es una base para escalar hasta los llamados métodos para el análisis de múltiples tablas de datos (Rivadeneira, Figueiredo, Figueiredo, Carvajal, & Rivadeneira, 2016).

El aporte de este trabajo conforma los cuatro componentes principales o ejes que son resultado del PCA y que explica los nutrientes más importantes para tomar en cuenta en la nutrición animal, pudiendo estos nuevos ejes ser utilizados como variables independientes explicativas en el desarrollo de trabajos futuro para construir un modelo indicador de raciones alimenticias o algún modelo predictivo de cumplimiento nutricional de alguna ración alimenticia. El primer componente: nutrientes energéticos involucra a los alimentos que aportan altos valores de energía, el segundo componente: materia seca, calcio y fósforo involucra los alimentos que aportan altos valores de esos nutrientes, el tercer componente: proteína bruta involucra los alimentos con altos valores de proteínas y, el cuarto componente: proteína degradable involucra los alimentos con altos valores de dicha proteína.

El presente trabajo se inicia con una revisión de conceptos sobre PCA. A continuación, se describe los métodos y la metodología utilizada, los resultados obtenidos y algunas conclusiones pertinentes.

2. Materiales y métodos

2.1. Reducción de la Dimensión

El principal objetivo de PCA es caracterizar los individuos a través de la información principal contenida en la tabla, lo cual es transformar la matriz de datos originales X_{nxp} en una nueva tabla Y_{nxq} descrita por un número de variables, en general de menor número.

$$X_{nxp} \rightarrow Y_{nxq} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ \vdots & & \ddots & \vdots \\ y_{n1} & y_{n1} & \cdots & y_{nq} \end{bmatrix}, \qquad con\ q \leq p$$

Donde las nuevas variables Y_j es llamada *j-ési*ma *Componente Principal*:

$$Y_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ . \\ . \\ . \\ y_{nj} \end{pmatrix},$$

Donde y_{ij} es el valor de la *j*ma componente para el individuo *i*, *i* =1, … , *n* , *j* = 1, … , *q*.

Esta reducción de dimensión presupone alguna pérdida de información, debido que se pretende reconstruir el máximo de la variabilidad existente de los datos iniciales a través de un número reducido de nuevas variables.

Geométricamente, se pretende proyectar los *n* individuos que pertenecen a un espacio de dimensión *p*, en un subespacio *W* de dimensión *q* (*q* < *p*). La selección de este subespacio debe ser acuerdo al criterio: El promedio de la distancia al cuadrado entre los puntos proyectados (es medida de su dispersión) debe ser lo más grande posible. Por lo cual, se desea

distorsionar la configuración de puntos lo menos posible, es decir que sea mínima la deformación en la proyección y, por lo tanto, también las distancias entre ellos (en realidad solo pueden disminuir).

2.2. Espacio de Variables y de Individuos

En lo que se refiere a las variables, en este trabajo se considera asignar el mismo peso $p_i = 1/n$ a todos los individuos x_{ij} de la variable j, X_j

Como las variables no todas están expresadas en la misma unidad o tienen diferentes dispersiones, se deberá trabajar con datos estandarizados (datos centrados y reducidos), donde la distancia entre dos individuos no depende de las unidades de medidas debido a que las coordenadas de los estos nuevos vectores no tienen dimensión, así todas las variables tendrán la misma importancia, independientemente de su dispersión. Con estos datos se procede a diagonalizar la matriz de correlación *R*.

3. Metodología y Obtención de datos

Los datos seleccionados para aplicar PCA fueron extraídos de (Fernández, 2010) guardados en un repositorio digital de acceso abierto de un sitio de producción animal incluido entre las instituciones argentinas registradas en el Directorio de Instituciones de Investigación Agropecuaria en América Latina y el Caribe de la Organización de las Naciones Unidas para la Agricultura y la Alimentación (F.A.O.) en el 2004.

Los datos de composición de alimentos provienen fundamentalmente de análisis realizados en laboratorios, pero deben ser tomados sólo como orientativos. Es recomendable realizar los análisis de los alimentos según la variedad de ellos tomando en cuenta su respectiva localización geográfica.

El desarrollo de este trabajo fue realizado con la versión 4.2.3 del lenguaje R y, diversos paquetes de R: para lectura de datos (readxl), para la

verificación de supuestos de PCA (corrplot, psych, mvnormtest), para la presentación del PCA (FactoMineR y factoextra).

La metodología por seguir para el computo de PCA estará conformada por los siguientes siete pasos:

1. Analizar y estandarizar los datos.
2. Computar y analizar la matriz de correlación.
3. Calcular los valores propios de la matriz de correlación.
4. Cómputo y retención de los componentes principales.
5. Calcular los vectores propios de la matriz de correlación.
6. Ordenar los vectores propios e Interpretar el PCA.

3.1. Analizar y estandarizar los datos

Inicialmente los datos cargados conforman una tabla compuesta de 150 filas por 12 variables, de las cuales cuatro son variables categóricas y ocho son continuas, por lo que se procede a seleccionar a las continuas para aplicar el PCA (ver Tabla1).

Tabla 1. Descripción y análisis de variables originales seleccionadas.

Nombre de Variable	Descripción	Promedio	Varianza
% MS	Porcentaje de Materia Seca	65.53	945.18
% DIVMS	Porcentaje de Digestibilidad in Vitro de Materia Seca	62.52	273.67
EM (Mcal/kgMS)	Energía Metabólica	2.29	0.35
% PB	Porcentaje de Proteína Bruta	18.27	803.17
% PDR	Porcentaje de Proteína Degradable en el Rumen	56.91	577.54
% FDN	Porcentaje de Fibra Detergente Neutro	44.59	565.21
% Calcio	Porcentaje de Calcio	0.59	0.34
% Fósforo	Porcentaje de Fósforo	0.36	0.13

*N/A: no aplica

Fuente: Autor (2024).

Se observa en la Tabla 1, que es necesario la estandarización de las variables a fin de evitar sesgos en el análisis final debido a sus diferentes unidades, medías aritméticas y, sus diferentes varianzas.

3.2 Computar y analizar la matriz de correlación

A continuación, se computa la matriz de correlación entre las variables: *R* y, se prueba a algunos supuestos para aplicar de mejor manera el PCA, como el análisis de correlación. Se aplica la prueba de hipótesis de esfericidad Bartlett (Bartlett, 1951) -no confundir con la prueba de homogeneidad de varianzas de Bartlett- para la idoneidad del PCA, tomando como hipótesis nula y alternativa lo siguiente:

H_0: matriz de correlaciones es igual a una matriz de identidad (no existe correlaciones entre variables),

H_1: matriz de correlaciones no es igual a una matriz de identidad (existen correlaciones significativas entre variables. PCA idóneo),

dando como resultado un *p-value* semejante a $2.88x10^{-51}$ que es mucho menor a 1%, indicando la existencia de correlaciones entre variables, por lo que se puede proceder a la realización del ACP.

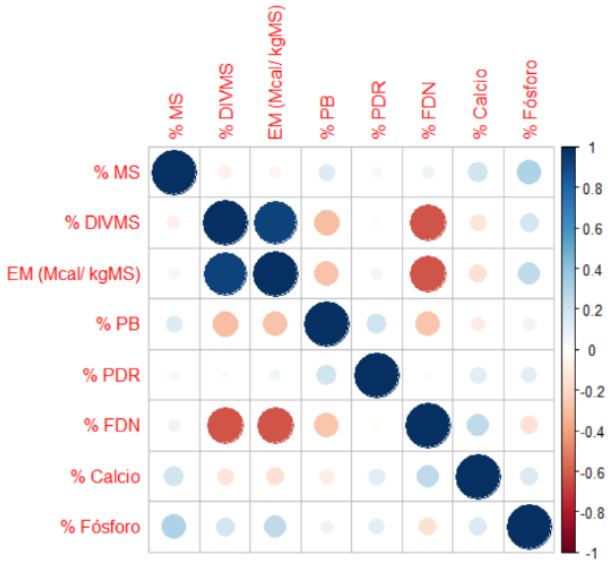
En la figura 1.a y 1.b, se indica las correlaciones entre las variables seleccionadas.

Figura 1.a. Matriz y Gráfico de correlaciones entre variables.

	% MS	% DIVMS	EM (Mcal/ kgMS)	% PB	% PDR	% FDN	% Calcio	% Fósforo
% MS	1.00	-0.08	-0.06	0.14	-0.05	0.07	0.19	0.30
% DIVMS	-0.08	1.00	0.92	-0.31	0.02	-0.63	-0.14	0.18
EM (Mcal/ kgMS)	-0.06	0.92	1.00	-0.29	0.06	-0.63	-0.16	0.25
% PB	0.14	-0.31	-0.29	1.00	0.19	-0.28	-0.10	0.07
% PDR	-0.05	0.02	0.06	0.19	1.00	-0.03	0.13	0.12
% FDN	0.07	-0.63	-0.63	-0.28	-0.03	1.00	0.25	-0.16
% Calcio	0.19	-0.14	-0.16	-0.10	0.13	0.25	1.00	0.16
% Fósforo	0.30	0.18	0.25	0.07	0.12	-0.16	0.16	1.00

Fuente: Autor (2024).

Figura 1 (b) Matriz y Gráfico de correlaciones entre variables.



Fuente: Autor (2024).

3.3 Generación del PCA

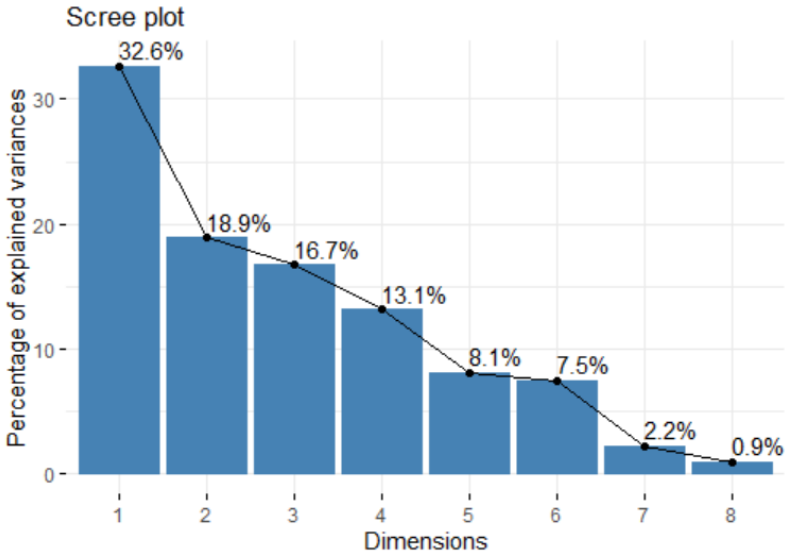
Los pasos que continúan del punto 2:

3. Calcular los valores propios de la matriz de correlación
4. Cómputo y retención de los componentes principales,
5. Calcular los vectores propios de la matriz de correlación
6. Ordenar los vectores propios e Interpretar el PCA.

Éstos se pueden agrupar en una sola función de lenguaje R que es: *prcomp()* o *PCA()* del paquete FactoMineR.

Se presenta la Fig. 2, de sedimentación (scree plot) mostrando el porcentaje o la cantidad de varianza de cada Componente Principal (CP), estos porcentajes se obtienen de los valores propios de la matriz de correlación de las variables originales, porque estos valores propios son iguales a las varianzas de los componentes principales y, cumple la propiedad que la varianza total de los CP es igual al número de variables o número de valores propios.

Figura 2. Porcentaje de varianza abarcada por cada Componente Principal.



Fuente: Autor (2024).

Ahora es el momento de la reducción del espacio dimensional, seleccionando o extrayendo los Componentes Principales de acuerdo con los siguientes criterios conocidos:

- Usando el criterio de *Cattel* (1966), se observa la gráfica representación de los valores propios (Fig. 2) y, se pueden retener los componentes principales cuya diferencia entre dos consecutivos valores propios es relativamente grande, en este caso se decide retener los cuatros primeros componentes principales.
- Usando el criterio de *Kaiser* (1958), se retiene los cuatro primeros componentes cuyas varianzas (valores propios) son mayores que el promedio de los valores propios o uno.
- Usando el criterio de *Pearson*, se selecciona el número más pequeño de componentes principales que juntos abarca del 80% a 90% del total de la varianza, en este caso con los cuatros primeros componentes se

tiene el 81.4% de la varianza o inercia total.

Por último, los vectores propios generados conforman una matriz de pesos de los Componentes Principales, ordenados de mayor a menor según la magnitud de sus valores propios, de tal forma que los pesos para el PC1 corresponden al primer vector propio de la matriz de correlación, del PC2 al segundo vector propio y así sucesivamente. Estos pesos (Fig. 3) son coeficientes de la combinación lineal de las variables estandarizadas del cual los scores de los Componentes Principales son calculados.

Figura 3. Vectores propios o pesos de los Componentes Principales

[,1]	[,2]	[,3]	[,4]
-0.07	0.54	0.16	-0.49
0.58	-0.04	0.15	0.04
0.59	0.01	0.14	0.04
-0.12	0.35	-0.71	-0.07
0.03	0.31	-0.18	0.80
-0.48	-0.10	0.38	0.09
-0.17	0.36	0.47	0.30
0.17	0.59	0.15	-0.10

Fuente: Autor (2024).

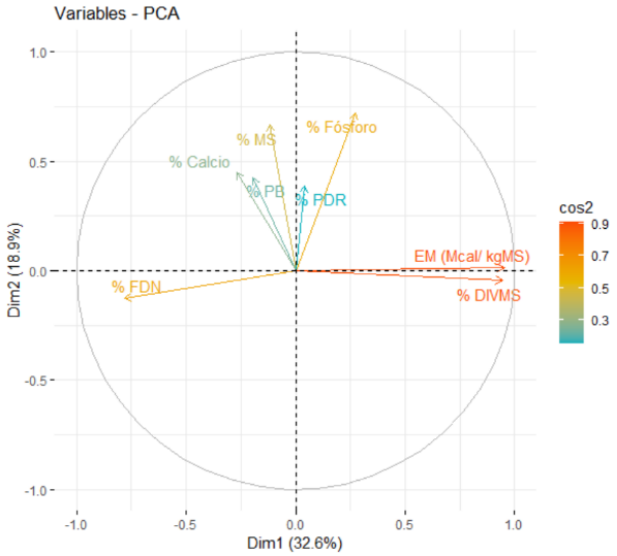
4. Resultados: Interpretar el PCA

En este último apartado se tiene la interpretación del PCA, basándose las correlaciones entre las Variables y los Componentes Principales y los gráficos de dichas correlaciones.

4.1. Primer Componente Principal

La figura 4 y la tabla 2 muestran las variables activas más correlacionadas con el primer Componente Principal:

Figura 4. Circulo de variables correlacionadas con el primer y segundo Componente Principal



Fuente: Autor (2024).

Tabla 2. Variables más correlacionadas con el primer Componente Principal.

Nombre de Variable	Correlación negativa	Correlación positiva
% DIVMS		0.94
EM (Mcal/kgMS)		0.95
% FDN	-0.78	

Fuente: Autor (2024).

El primer Componente Principal opone los alimentos con alto valores en *Porcentaje de Digestibilidad in Vitro de Materia Seca* y en *Energía Metabólica* a los alimentos con bajo valores en ellos.

4.2. Segundo Componente Principal

La figura 4 y la tabla 3 muestran las variables activas más

correlacionadas con el segundo Componente Principal:

Tabla 3. Variables más correlacionadas con el segundo Componente Principal.

Nombre de Variable	Correlación negativa	Correlación positiva
% MS		0.67
%PB		0,42
%Calcio		0,45
% Fósforo		0.72

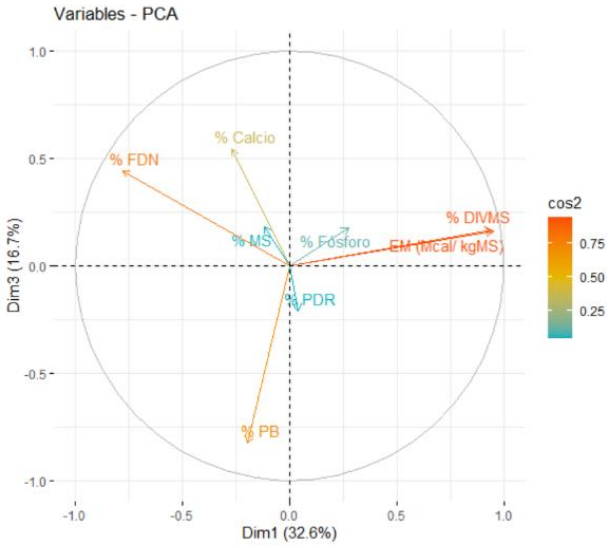
Fuente: Autor (2024).

El segundo Componente Principal opone alimentos con alto valor en *Porcentaje de Materia Seca, Porcentaje de Proteína Bruta, Porcentaje de Calcio* y en *Porcentaje de Fósforo* a los alimentos con bajos porcentajes en dichas variables.

4.3. Tercer Componente Principal

La figura 5 y la tabla 4 muestran las variables activas más correlacionadas con el tercer Componente Principal:

Figura 5. Circulo de variables correlacionadas con el tercer Componente Principal



Fuente: Autor (2024).

Tabla 4. Variables más correlacionadas con el tercer Componente Principal.

Nombre de Variable	Correlación negativa	Correlación positiva
% PB	-0.83	
%FDN		0,44
%Calcio		0,54

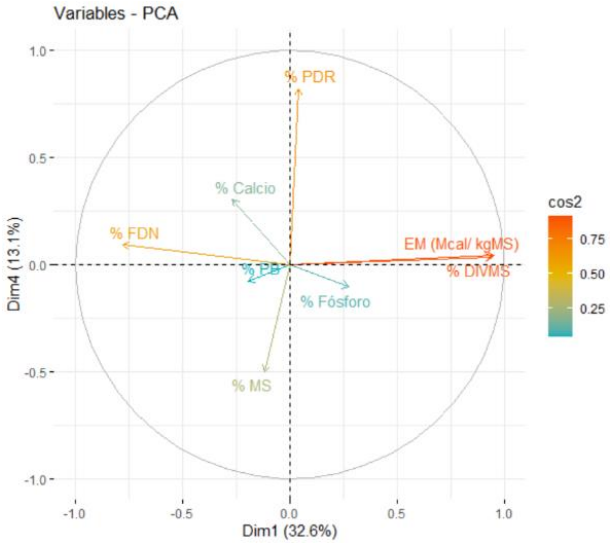
Fuente: Autor (2024).

El tercer Componente Principal opone alimentos con alto valor en *Porcentaje de Proteína Bruta* a los alimentos con bajos porcentajes en dicha variable.

4.4. Cuarto Componente Principal

La figura 6 y la tabla 5 muestran las variables activas más correlacionadas con el cuarto Componente Principal:

Figura 6. Circulo de variables correlacionadas con el cuarto Componente Principal



Fuente: Autor (2024).

Tabla 5. Variables más correlacionadas con el cuarto Componente Principal.

Nombre de Variable	Correlación negativa	Correlación positiva
%MS	-0.50	
% PDR		0.82

Fuente: Autor (2024).

El cuarto Componente Principal opone alimentos con un valor alto en *Porcentaje de Proteína Degradable en el Rumen* a los alimentos con valores bajos en dicha variable.

5. Conclusiones

El aporte de este trabajo al aplicar el Análisis en Componentes Principales que puede ser vista como perteneciente a la familia de los algoritmos de aprendizaje no supervisados, es proveer un enfoque a través de la reducción de dimensionalidad de una tabla conformadas por columnas de variables numéricas que representan los nutrientes de alimentos para rumiantes a solo cuatro Componentes Principales no correlacionados y, gracias a esta reducción de dimensión se pudieron identificar los nutrientes que son más importantes en los respectivos Componentes y por ende considerarlos en la selección alimentos. Estos componentes son:

- Primer Componente: relacionados a nutrientes energéticos e involucran a los alimentos que aportan altos valores de energía.
- Segundo Componente: relacionados a materias seca, calcio y fósforo e involucran a los alimentos que aportan altos valores de materias secas, calcio y fosforo.
- Tercer Componente: relacionados a proteína bruta e involucran a los alimentos que aportan altos valores de proteínas.
- Cuarto Componente: relacionados a proteína degradable e involucran a los alimentos que aportan altos valores de proteína degradable.

A futuro se podrían construir índices que permita clasificar ciertos alimentos debido a que los Componentes Principales son combinaciones lineales de las variables originales. También, para futuras investigaciones se podría aplicar un análisis de conglomerados de alimentos para corroborar su agrupación con respecto a los nutrientes como variables.

6. Referencias

Bartlett, M. S. (1951). The Effect of Standardization on a chi square Approximation in Factor Analysis. *Biometrika*(38), 337-344.

Chibisa, G.E., & Oba. (2020), M. Nutrition and Feeding of Ruminants.1ª edición. Wageningen Academic Publishers.

Fernández, H. H. (2010). Obtenido de Sitio Argentino de Producción Animal: https://www.produccion-animal.com.ar/tablas_composicion_alimentos/46-Tabla.pdf

Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3ª edición. Sebastopol, CA: O'Reilly Media.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417–441, and 498–520.

Moraes, L.E., & Fadel, J.G. (2020). Feed Efficiency in the Beef Industry. 1ª edición. Hoboken, NJ: Wiley-Blackwell.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine, 2*(11), 559–572. doi:10.1080/14786440109462720

Rivadeneira, F. J., Figueiredo, A. M., Figueiredo, F. O., Carvajal, S. M., & Rivadeneira, R. A. (2016). Analysis of Well-Being In OECD Countries through Statis Methodology. *HOLOS, 7*(32), 335-351. doi:10.15628/holos.2016.5003