Universidad Autónoma de Querétaro
Facultad de Ingeniería
Licenciatura en Matemáticas Aplicadas

USO DE MÉTODOS ESTADÍSTICOS EN LA SELECCIÓN DE VARIABLES CLASIFICATORIAS

Tesis

Que como parte de los requisitos para obtener el grado de
Licenciado en Matemáticas Aplicadas

Presenta:
Valeria Barón Villar

Dirigido por:
Dr. Eduardo Castaño Tostado

Dr. Eduardo Castaño Tostado Presidente	Firma
Dr. Mario Santana Cibrián Secretario	Firma
Mto. Juan Antonio Villeda Reséndiz Vocal	Firma
Mto. Gilberto Guillermo Sánchez de la Isla Sinodal	Firma

Índice

Resumen

Existen muchos métodos para la selección de variables, sin embargo, muchos tienen ciertas limitaciones cuando se manejan conjuntos de datos de grandes dimensiones. Es un problema estadístico interesante identi car las variables más relevantes en este tipo de escenarios, además del sinfín de aplicaciones que se le puede dar en otras áreas. Este trabajo de tesis se centrará en un método de selección de variables llamado ERGS, el cual efectúa la selección asignando cierta relevancia o peso a cada una de ellas por medio de la construcción de intervalos.

Summary

There are several variable selection methods, however many of them have certain limitations when it comes to high-dimentional data sets. To identify the most relevant variables in this type of scenarios is a very interesting statistical problem, furthermore there are endless aplications in other areas. This thesis will focus on a variable selection method called ERGS. This does makes the variable selection assigning certain relevance value or weight value to each of the variables by constructing e ective ranges.

1. Introducción

El poder de cómputo en los últimos años ha crecido, sin embargo, en muchas ramas de la ciencia se están recolectando tantas variables o mediciones que los algoritmos que los procesan pueden llegar a ser muy costosos computacionalmente; se debe contrastar la relevancia de incluir la mayor cantidad de variables contra el tiempo que lleva procesarlas para obtener resultados útiles.

La selección de variables se puede aplicar en problemas de clasi cación de unidades. En las ciencias ómicas (genómica, metabolómica, proteómica, etc.) es relevante porque hay demasiadas variables disponibles, lo cual di culta el análisis.

En este trabajo de tesis se profundizará sobre un método que se ha usado en este ámbito de la genómica, ERGS (e ective range based gene selection) [1], cuyo principio básico es asignar un mayor peso a las variables que discriminan mejor entre las diferentes clases.

Se comenzará con introducción al tema donde se observará el problema de selección de variables en diferentes áreas, junto con teoría estadística; se hará una descripción teórica del método ERGS y hará un análisis práctico en un ámbito controlado con una simulación y otro análisis en ámbito real con datos de una investigación.

El capítulo 2 contiene la teoría básica para el desarrollo de este trabajo, se abarcan temas como la importancia de la reducción de dimensión en una base de datos, el uso en otras áreas, el problema estadístico, algunos métodos y estrategias para la selección de variables y la evaluación de la selección.

En el capítulo 3 se analiza la construcción y funcionamiento del algoritmo ERGS. Como este método se retomó del artículo de Chandra & Gupta [1], en este capítulo también se reporta el trabajo sobre cáncer de colon donde los autorea aplicaron el algoritmo.

En el capítulo 4 se aplica una simulación a variables uniformes construida por Wang & Gevertz [5] para el análisis de los resultados que arroja el algoritmo, partiendo de un a base de datos y variable de referencia controladas.

En el capítulo 5 se aplica el algoritmo de una manera no controlada, usando los datos de una aplicación práctica sobre metabolitos de hombres y mujeres a ser utilizados en un diagnóstico de salud diferencial.

En el último capítulo es dónde se concluye el trabajo considerando los detalles que se abarcaron en los capítulos anteriores y una conclusión personal.

2. Antecedentes del problema

2.1. Selección de características

2.1.1. Reducción de dimensión

En al medida que hay más información y sensores en sistemas, las bases de datos crecen y tienen más variables. Se debe reducir de dimensión sin perder información, es posible que con la misma base de datos se puedan contestar diferentes preguntas ya que no todas son útiles según el objetivo de la investigación.

El que $n \quad p$ tiene un efecto negativo sobre el análisis exploratorio de los datos ya que casi ninguno de los procedimientos multivariados funcionan adecuadamente. Este escenario puede implicar que se tienen variables irrelevantes o redundantes, por lo que se deben remover para poder tener una mejora en la precisión de la clasi cación, que no sea caro computacionalmente [1], evitar deteriorar el rendimiento de algoritmos de aprendizaje [12] y lograr hacer experimentos más rápidos y menos costosos [5].

Análisis de datos masivos

Hoy en día estamos viviendo en la época de la revolución de los datos, esta expresión se re ere a la masividad de los datos en cuanto a volumen, velocidad, variedad, veracidad y valor; las 5 V's. Esto lo vemos re ejado ya que el 90 % de los datos que existen han sido creados en los últimos dos años como lo escribe Jacobson en el 2013 [11].

- Volumen: al día se generan 2.5 quintillones de bytes de datos, es decir, 2,684,354,270 GB. Para poder hacer un análisis de sentimiento sobre un producto, se deben analizar 12 terabytes de información para obtener buenos resultados [11].
- Velocidad: Según Visual Capitalist en el 2017 [15], en un minuto se genera mucha información en el mundo. Algunos ejemplos son: 16 millones de mensajes de texto, 3.5 millones de búsquedas en Google, 900,000 inicios de sesión en Facebook, 452,000 tweets mandados, 156 millones de correos electrónicos enviados, solo por mencionar algunos ejemplos.
- Variedad: a grandes rasgos existen dos tipos de datos, los datos estructurados y los no estructurados. Los datos estructurados son los tipos de datos que siguen un patrón el cual es fácil de

buscar; por otro lado los datos no estructurados se de nen como todo lo demás [16]. Un ejemplo de datos con estructura son los códigos postales o la Clave Única de Registro Poblacional (CURP), ya que es una cifra que contiene nombre, apellido, fecha y lugar de nacimiento siguiendo un orden especí co. Los datos no estructurados se encuentran en archivos de texto, audio, imagen o video; una publicación en cualquier red social está contribuyendo a la generación de datos no estructurados.

- Veracidad: se re ere a la incertidumbre de los datos. En muchos casos la calidad y precisión de la información es menos controlable. Por ejemplo los mensajes de la plataforma Twitter que tienen abreviaciones, errores tipográ cos, lenguaje coloquial, etc.[19]
- Valor: se relaciona con el potencial de transformar la información en ganancia. Se enfoca en la ciencia de datos tal como herramientas y métodos estadísticos y analíticos para la extracción de conocimientos y toma de decisiones. [19]

El análisis de datos masivos es más que una simple cuestión de tamaño; es la oportunidad de encontrar valor en diferentes tipos de bases de datos y contenidos, con el propósito de contestar preguntas que antes estaban fuera de alcance [11]. Por otro lado, se puede tener mucha información pero no necesariamente es de utilidad por lo que muchas ciencias comienzan a aplicar métodos de clasi cación y selección de variables; sin embargo es crucial que se apliquen de manera correcta y se conozcan sus limitaciones.

2.1.2. Aplicaciones

Estudios epidemiológicos

En países en vías de desarrollo ubicados en África, se implementó un programa para combatir la malaria. Vía mensaje de texto SMS los trabajadores de salud contestaban una encuesta para alertar a los funcionarios sobre brotes de esta enfermedad en diferentes regiones, esto para anticipar la cantidad de medicamento necesario y evitar la escasez. Así es como la recolección de datos contribuye con los servicios de salud en un país.

Se debe de tener muy claro las variables que in uyen sobre estos brotes de malaria, ya que cada encuesta tiene un costo y, aunque esto no representa una gran cantidad para un gobierno, se deben buscar maneras para e cientar costos en relación a la información obtenida [14].

Medir el estado de salud de una persona implica una enorme cantidad de variables, sin embargo con métodos de selección se pueden identi car aquellas variables que responden a la enfermedad de malaria y elaborar las preguntas en consecuencia dirigidas a la ciudadanía.

Metabolómica

De acuerdo con el Diccionario de Oxford [18], la metabolómica es el estudio cientí co de un conjunto de los metabolitos presentes en un organismo, célula o tejido. El objeto de estudio de esta ciencia son

los biomarcadores, es por eso que al hacer la selección de los mismos se puede llegar a diagnosticar, monitorear y/o predecir el riesgo a cierta enfermedad.
Al combinar dos o más biomarcadores es posible generar un diagnóstico más preciso e incluso distinguir entre diferentes enfermedades similares.

Genómica

La clasi cación de la expresión de los genes juega un papel importante en la predicción y diagnóstico de enfermedades, sin embargo es necesario hacer una selección de genes en primera instancia. Para bases de datos de este tipo es muy común el uso de métodos de selección de variables por la cantidad de genes que poseen los seres humanos, animales y otros objetos de estudio de esta ciencia.
No todos los genes contribuyen para hacer una clasi cación e ciente de las muestras, se necesita un algoritmo robusto de selección de estos para identi car los que son importantes, los cuales ayudarán a clasi car la muestra e cientemente [1].

2.1.3. Problema estadístico

Por la mayor parte del siglo veinte, los principales problemas prácticos consistían en tener un gran número de unidades o casos (n) y un número limitado o jo de variables (p), esto dadas las limitaciones de poder cómputo y de visualización grá ca.
Sin embargo, en los últimos veinte años comenzaron a surgir problemas prácticos que demandaban la evolución de nuevas tecnologías para la adquisición de datos y de instalaciones computacionales que ya estaban en evolución.
En términos de teoría asintótica, estas nuevas aplicaciones prácticas asumen que para una base de datos considerada como una matriz de dimensión $n \times p$, el número de variables p tiende mucho más rápido a in nito que el número de sujetos n, o bien ambas tienden a in nito con la misma razón [4], por lo que muchas herramientas estadísticas tradicionales y computacionales no son útiles.

Herramientas estadísticas tradicionales

Debido a los problemas que conlleva el escenario en que $n\ p$ es necesario reducir el número de variables a un subconjunto de variables para poder cumplir con las expectativas y objetivos de interés. Por ejemplo, en regresión lineal se tienen datos sobre n individuos en forma de parejas donde x_i es la variable predictora y y_i es una respuesta tal que $\{x_i, y_i;\ i = 1, \ldots, n\}$. La relación de estas parejas está dada por

$$y_i = \beta_1 + \beta_2 x_i + s_i,$$

donde los términos de error $s_i \sim N\left(0, \sigma^2\right)$ son independientes. Así, la relación entre la variable predictora y la esperanza de la respuesta sigue una línea recta con ordenada al origen β_1 y pendiente β_2; es decir,

$$E\left(y_i | x_i\right) = \beta_1 + \beta_2 x_i.$$

Este modelo recibe el nombre de regresión lineal simple, ya que solamente hay una variable predictora, los dos parámetros β_1 y β_2 son constantes desconocidas que se deben de estimar. Existe una notación matricial dada por

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{s}$$

donde el vector $\boldsymbol{y}$ de dimensión $n \times 1$ contiene las respuestas, el vector β contiene p parámetros en general excepto σ^2, el vector s contiene el ruido y X es las llamada matriz de diseño de dimensión $n \times p$. Para la regresión lineal simple, la primer columna de X tiene el valor de 1 para todas las entradas. La forma estándar de estimar β_1 y β_2, es usando mínimos cuadrados y obtener

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (y_i - \beta_1 - \beta_2 x_i)^2 ,$$

esto signi ca que $\hat{\boldsymbol{\beta}}$ minimiza la suma de cuadrados del lado derecho. La notación en general está dada por

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2.$$

Como se tomó como supuesto que s era Normal y sea I la matriz identidad $n \times n$, esto implica que $\boldsymbol{y} \sim N_n(X\beta, \sigma^2 I)$, donde N_n es una distribución normal multivariada. Por lo que la función de densidad conjunta de probabilidad para y es

$$p(\boldsymbol{y}|X, \beta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{||y - X\beta||_2}{\sigma}\right)^2\right).$$

Esto se puede ver como una función de los parámetros, es decir, y recibe el nombre de función de verosimilitud. Sea $\hat{\boldsymbol{\beta}}$ es el estimador máximo verosímil de $\boldsymbol{\beta}$ dados por

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} p(y|X, \beta).$$

Ahora, $\hat{\beta}$ satisface

$$X^T X\hat{\boldsymbol{\beta}} = X^T y$$

y

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

si y sólo si $X^T X$ es invertible. Más aún,

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

desde un punto de vista frecuentista, las estimaciones de intervalos para $\boldsymbol{\beta}$ pueden obtenerse con una modi cación sutil si σ^2 debe ser estimada. En general, para muchos escenarios de máxima verosimilitud

respecto a modelos paramétricos que involucran un número jo $p - 1$ variables y p parámetros en β, con un gran número de variables, casi seguramente

$$\hat{\boldsymbol{\beta}} \sim N_p \left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} \right),$$

para una cierta matriz $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$.

Pero, ¾por qué no se puede usar la regresión lineal tradicional en el escenario $p > n$?

El problema es que una característica necesario para el análisis de regresión lineal es que $X^T X$ sea invertible, esto sucede solamente si $p \leq n$. Si $p > n$ entonces $X^T X$ es una matriz singular por lo que los parámetros $\boldsymbol{\beta}$ no pueden ser estimados de manera única.

Una forma de solucionar esto y evadir el problema de la singularidad de la matriz $X^T X$ es mediante un método de regularización conocido como mínimos cuadrados penalizados o máxima verosimilitud penalizada [4].

Un ejemplo de esto es la regresión Ridge, cuya solución se puede escribir de la siguiente forma

$$\hat{\beta} = \left(X^T X + \lambda I \right)^{-1} X^T y.$$

Podemos ver que la constante λ es la que controla el valor de $\hat{\beta}$ ya que si $\lambda \to \infty$ entonces $\hat{\beta}$ disminuye

$$\hat{\beta} = \left(\frac{1}{\lambda} I \right) X^T y$$
$$= \frac{1}{\lambda} X^T y.$$

De esta forma, se busca λ tal que $X^T X + \lambda I$ sea invertible [4].

2.2. Estrategias de búsqueda para la selección de variables

Varios estudios demuestran la importancia de métodos de selección de variables para identi car las aquellas que son informativas. Según Chandra & Gupta (2011) [1], los métodos de selección de variables pretenden remover las variables redundantes e irrelevantes para mejorar la exactitud de una clasi cación, además de facilitar el trabajo de cómputo. A continuación se presentan algunos enfoques de estas estrategias y ejemplos particulares de cada uno.

2.2.1. Enfoque wrapper

Este método es bastante costoso computacionalmente, ya que se usa un subconjunto inicial de variables que se modi ca hasta obtener el mejor subconjunto.

Se usa un algoritmo de aprendizaje para que, con base a un criterio, se decida si se agregan o remueven variables del subconjunto. El proceso se muestra en la Figura 1.

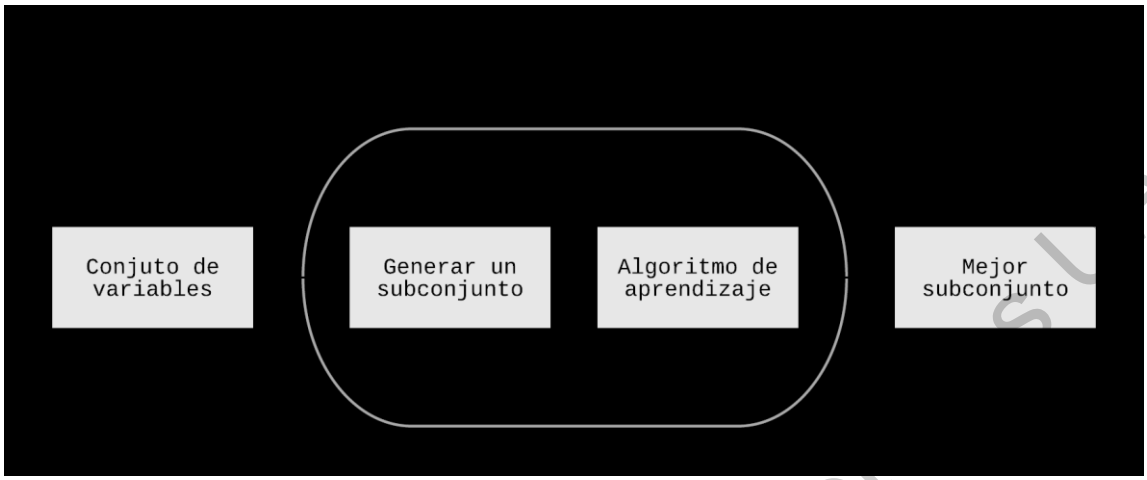

Figura 1: Método de envoltura.

Envoltura hacia adelante

Este tipo de selección de variables consiste en comenzar con un subconjunto vacío, en los pasos subsecuentes se agrega la variable que mejore el criterio. Esto se continúa hasta que la adición de una nueva variable no sea importante para el criterio. El proceso se muestra en la Figura 2.

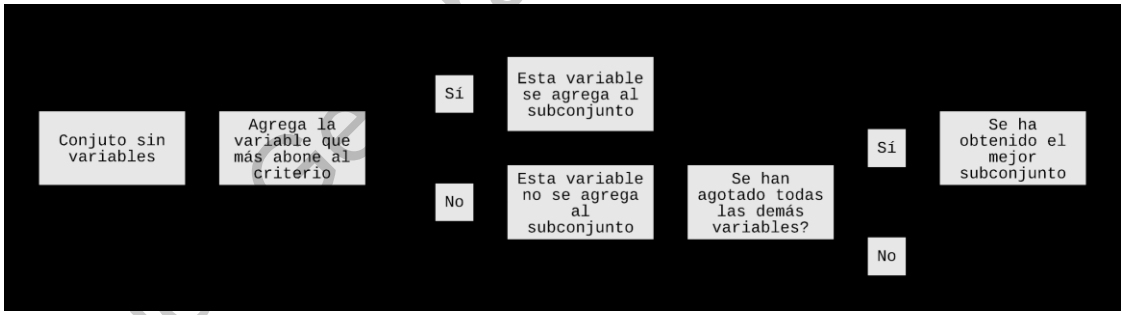

Figura 2: Método de envoltura hacia adelante.

Envoltura hacia atrás

El diagrama del proceso para este método es muy similar al método anterior pero, en este caso, se comienza al contrario. De inicio se tienen todas las variables, en cada iteración se remueve la variable que menos reste al criterio hasta que no se observe una mejora. El proceso se muestra en la Figura 3.

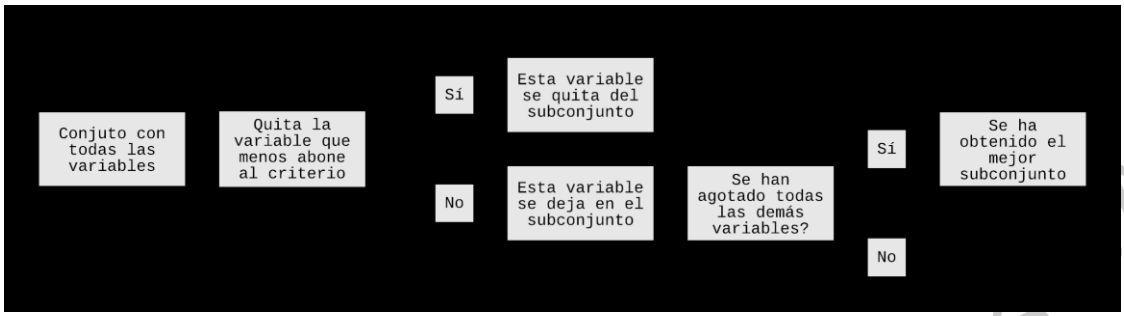

Figura 3: Método de envoltura hacia atrás.

Envoltura por pasos

Es la combinación de los dos métodos en envoltura antes mencionados. Se agregan las variables una a la vez, sin embargo, después de que se agrega una variable se evalúa a todas las variables que están en el modelo y se elimina a cualquiera que no sea importante para el criterio.

2.2.2. Enfoques incrustados

Este enfoque realiza la selección de variables durante el proceso de entrenamiento de un algoritmo de aprendizaje [8]. Se busca reducir el tiempo de cómputo que se usa para reclasi car diferentes subconjuntos, como en los métodos de envoltura. El enfoque principal es incorporar la selección de variables como parte del proceso de entrenamiento [6].

Validación cruzada anidada en modelación del algoritmo de aprendizaje

La validación cruzada anidada puede ser usada en el caso de que se tenga un pequeño tamaño muestral. El proceso es el siguiente:

1. Se divide el conjunto de datos aleatoriamente en dos subconjuntos, uno de entrenamiento (train) y uno de prueba (test).
2. Se usa un algoritmo de aprendizaje de forma iterativa para la selección de características, solamente para el conjunto de entrenamiento.
3. Se usa la validación cruzada para evaluar el modelo con los posibles subconjuntos de variables

Como resultado se obtiene un modelo óptimo, el cual es evaluado usando el conjunto de datos de prueba. Los pasos 1 y 2 se repiten *N* veces, tal que hagan *N* evaluaciones óptimas del modelo [3].

2.2.3. Enfoque de ltrado

Este es el enfoque comúnmente usado, el cual selecciona características sin involucrar algoritmos de minería de datos, los algoritmos de ltrado son evaluados basados con en base a cuatro criterios: distancia, información, dependencia y consistencia [1].

Esta estrategia ordena las variables sin usar algoritmos de aprendizaje, depende de las características generales de los datos [2].

En la mayoría de los casos se calcula la relevancia de las variables dándoles una puntuación (o peso), se ordenan y las q p variables con puntuación alta son consideradas como el subconjunto ltrado del conjunto de todas las características.

Después, se aplica un algoritmo de aprendizaje usando a este subconjunto obtenido como la entrada (Cakmak 2015) [2].

Una de las maneras más e cientes para evaluar el desempeño de un subconjunto de q variables y saber si se eligió correctamente el valor de q, es midiendo su capacidad de predicción esto se puede determinar por medio de las curvas de ROC.

Aún con las curvas de ROC no hay ninguna garantía teórica que las principales q variables será el subconjunto óptimo. Las curvas ROC se explicarán más adelante [3].

Estadístico t de Student

El t-estadístico es un método de ltrado usado en problemas de dos clases, por lo que en este caso $l = 1, 2$. Tenemos que μ_j y σ_j denotan la media y desviación estándar para la j-ésima característica y la l-ésima clase, respectivamente. Además n es el número de muestras en la l-ésima clase; entonces el estadístico t para cada variable se calcula como se sigue

$$T(X) = \frac{\cdot \mu^{i}_{1} - \mu^{i}_{2} \cdot}{\sqrt{(\sigma_1)^2/n_1 + (\sigma_2)^2/n_2}}.$$

Al igual que en los otros métodos de ltrado se buscan los mayores valores del estadístico t de las variables [2]. Lo que requiere el estadístico t para tener un valor relativamente mayor es que

$$|\mu_1 - \mu_2| > \sqrt{(\sigma_1)^2/n_1 + (\sigma_2)^2/n_2},$$

es decir, que la distancia entre las medias sea mayor que una varianza combinada. De esta manera se identi ca a las variables que estén alejadas entre sí.

El t-estadístico inicialmente se diseñó para examinar las diferencias entre dos muestras independientes y pequeñas que tengan distribución normal y homogeneidad en sus varianzas. La idea intuitiva sobre esta pruebademuestraque existe una curva que describe el comportamiento de la diferencia de medias

y permite calcular el área bajo la curva que representa la probabilidad de la diferencia entre ellas.[17]

Esta prueba es poderosa, que aunque una de las muestras no tenga distribución normal pero la otra sí y la razón de la varianza más grande a la más pequeña sea chica, esta prueba resulta adecuada al comparar dos medias.[17]

Fisher Score

Este tipo de puntuación se le conoce como el índice de Fisher y su propósito es encontrar un subconjunto de variables tales que las distancias de los datos dentro de cada clase sean las menores y las distancias de los datos entre las clases sean mayores, en la medida de lo posible.

Este criterio se calcula con la proporción entre la varianza entre clases y la varianza dentro de cada clase,

$$FS(X^i) = \frac{\sum_{j=1}^{c} n_j (\mu_j^i - \mu^i)^2}{\sum_{j=1}^{c} n_j (\sigma^i)^2}$$

donde el conjunto muestra de la j-ésima clase $X^j = (x^j_1, x^j_2, \ldots, x^j_{n_j})$ con n_j es el número de muestras en la j-ésima clase, μ^i es la media para la j-ésima clase con la correspondiente i-ésima variable y σ^i es la desviación estándar para la j-ésima clase con la correspondiente i-ésima variable.

Una vez que se calcularon los *FS* para cada variable se ordenan de forma descendiente y se consideran las primeras q variables con rangos mayores [2] . El valor de q varía dependiendo del objetivo de la investigación.

LDA

El análisis de discriminantes lineales (LDA) tiene como objetivo construir combinaciones lineales de variables que permita recuperar las categorías respuesta en un conjunto de datos.

A grandes rasgos, para hacer el análisis

1. Calcular los vectores de medias p-dimensionales para las diferentes categorías de una base de datos.
2. Calcular las matrices de dispersión, tanto la matriz entre categorías como la matriz dentro de la misma categoría.
3. Calcular los eigenvectores $(e_1, e_2, \ldots, e_p)$ correspondientes a sus eigenvalores $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ de ciertas matrices de dispersión.
4. Ordenar los eigenvectores de forma descendiente según el valor de sus eigenvalores y elegir los q eigenvectores con los mayores eigenvalores para formar una matriz W de $p \times k$, donde cada columna representa un eigenvector.

5. Usar W para transformar las muestras a un subespacio nuevo. Esto puede ser $Y = X \times W$ donde X es una matriz de $n \times p$ que representa las n muestras y Y el espacio de $n \times q$ muestras en el subespacio nuevo.

Aunque es un buen método tiene algunas limitaciones, una de éstas es que si las distribuciones con signi cativamente no Normales entonces las proyecciones no van a re ejar las estructuras complejas de la clasi cación de los datos.

Otro problema se da con un pequeño tamaño muestral, porque se necesita calcular la inversa de la matriz de dispersión interclase, la cual es singular cuando el número de muestras de entrenamiento es menor que el número de variables [8].

3. Método ERGS

3.1. Introducción

Efective Range Based Gene Selection es un algoritmo que se basa en amplitudes efectivas, es un método de ltrado ya que al crear estas amplitudes se obtienen puntuaciones o pesos que distinguen a las variables más relevantes; además ERGS también es un método supervisado ya que se trabaja con una variable de referencia.

A diferencia de la mayoría de los algoritmos de selección de variables más populares, el ERGS no requiere de una estrategia iterativa de búsqueda para la generación de subconjuntos. El principio básico del algoritmo es que se le asigna un peso mayor a las variables que discriminan las clases con mayor claridad, esto se hace con todas las variables de la base de datos.

El cálculo de las amplitudes efectivas, en parte, se basan en la desigualdad de Chebyshev, la cual es cierta para todas las distribuciones y se de ne como

$$P\left(|X - \mu_{ij}| \geq \gamma\sigma_{ij}\right) \leq \frac{1}{\gamma^2},$$

esta desigualdad es cierta para toda distribución y el valor de $\gamma = 1.732$ dado que las amplitudes contienen, al menos dos tercios de los casos.

El objetivo es poder seleccionar a las variables que mejor distingan a las diferentes clases; esto se va a determinar por los valores de w_i de cada variable. A continuación, en las secciones 3.2 y 3.3 se van a de nir las amplitudes y a explicar el algoritmo ERGS según Chandra & Gupta (2011) [1], respectivamente.

3.2. De nición de las amplitudes

Antes de desarrollar el algoritmo se deben de nir:

- $X = \{X_1, X_2, \ldots, X_d\}$ conjunto con n casos en d variables.
- $C = \{C_j\}$ el conjunto de clases donde $j = 1, 2, \ldots, L$
- p_j denota la probabilidad de la $j-ésima$ clase C_j, esta probabilidad está directamente relacionada

con el número de sujetos que contenga cada clase, i.e.

$$p_k = \frac{\text{Número de sujetos en la clase } k}{\text{Número de sujetos}} = \frac{\text{Número de sujetos en la clase } k}{n}$$

- μ_{ij} denota la media para la característica X_i en la clase C_j
- σ_{ij} denota la desviación estándar para la característica X_i y la clase C_j
- R_{ij} denota la amplitud efectiva para la i-ésima característica X_i y j-ésima la clase C_j, y se calcula como se sigue

$$R_{ij} = \left[r^{-}_{ij},\ r^{+}_{ij} \right] = \left[\mu_{ij} - (1 - p_j)\,\gamma\sigma_{ij}\ ,\ \mu_{ij} + (1 - p_j)\,\gamma\sigma_{ij} \right].$$

Se puede ver que $r^{-}_{ij} = \mu_{ij} - (1 - p_j)\,\gamma\sigma_{ij}$ es la cota inferior de la amplitud y $r^{+}_{ij} = \mu_{ij} + (1 - p_j)\,\gamma\sigma_{ij}$ es la cota superior de la amplitud.

3.3. Algoritmo

1. Calcule los rangos efectivos R_{ij} para todas las clases C_j y cada característica X_i.
2. Ordene de forma ascendiente los r^{-}_{ij} y de la misma manera los r^{+}_{ij}.
3. Calcule el área de traslape (OA_i) entre clases para la característica X_i usando la siguiente fórmula

$$OA_i = \sum_{j=1}^{k} \sum \phi_i(j,\ k)$$

donde

$$\phi_i(j,\ k) = \begin{cases} r^{+}_{ij} - r^{-}_{ij} & \text{si } r^{+}_{ij} > r^{-}_{ij} \\ 0 & \text{en otro caso} \end{cases}.$$

4. Calcule el coe ciente de área AC_i

$$AC_i = \frac{OA_i}{\text{máx} \left[\sum r^{+}_{ij} \right] - \text{mín} \left[\sum r^{-}_{ij} \right]}.$$

5. Calcule el coe ciente de área normalizado NAC_i

$$NAC_i = \frac{AC_i}{\text{máx}\,(AC_j)} \text{ para } j = 1, 2, \ldots, d.$$

6. Calcule los pesos, denotados por w_i, para cada característica X_i, como

$$w_i = 1 - NAC_i.$$

7. Seleccione las características X_i tales que $w_i > \theta$, donde θ es un valor umbral determinado dependiendo de las necesidades del estudio.

Se puede observar que si σ_{ij} incrementa entonces los pesos w_i para estas variables disminuyen, esto es por que

$$\begin{aligned} \sigma_{ij} \text{ incrementa} &\Longrightarrow R_{ij} \text{ incrementa} \\ &\Longrightarrow OA_i \text{ incrementa} \\ &\Longrightarrow AC_i \text{ incrementa} \\ &\Longrightarrow w_i \text{ decrementa.} \end{aligned}$$

Por esta razón R_{ij} considera la dispersión para el valor que asignará a los w_i, ya que σ_{ij} se reduce proporcionalmente al multiplicarse por $(1 - p_j)$, donde $0 < p_j < 1$, así se calculan pesos menores para estas variables. Esto no implica que los valores de R_{ij} estén exentos de ser afectados por valores atípicos, ya que también se calculan a partir de los valores de μ_{ij} que son una medida de tendecia central.

Los autores no dejan en claro el hecho de que se quieran incluir $\frac{2}{3}$ de los datos ni se demuestra el valor de γ se cumple para cada clases de todas las variables.

3.4. Colon

En el artículo de Chandra (2011) [1] se evalúa el método ERGS en diferentes bases de datos, en dos de estas reporta los resultados encontrados. Las bases de datos tiene las siguientes características:

- Colon: 62 muestras de células epiteliales de colon de pacientes con cáncer de colon. Las muestras consisten en biopsias tumorales recolectada de tumores y biopsias normales recolectadas de partes sanas del colon del mismo paciente. Cada gen representa una variables y el número de genes en la base de datos es de 2000.
- MLL: este conjunto llamado, Leucemia de Linaje Mixto contiene 72 muestras de 12,582 genes. Las muestras consisten en tres tipos de leucemia, 24 muestras de ALL, 20 muestras de MLL y 28 muestras de AML.

Después de implementar el método ERGS se presentan los 10 genes seleccionados para la base de datos de colon se puede ver en la Tabla 1 y para MLL en la Tabla 2.

Colon		
Número de acceso del gen	Nombre del gen	Descripción del gen
H08393	Hsa.6814	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
X63629	Hsa.2928	H.sapiens mRNA for p cadherin.
M22382	Hsa.831	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN)
R36977	Hsa.549	P03001 TRANSCRIPTION FACTOR IIIA
T56604	Hsa.6472	TUBULIN BETA CHAIN (Haliotis discus)
H40095	Hsa.773	MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN)
U30825	Hsa.5971	Human splicing factor SRp30c mRNA, complete cds
T47377	Hsa.3016	S-100P PROTEIN (HUMAN)
J05032	Hsa.601	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
M63391	Hsa.8147	Human desmin gene, complete cds

Tabla 1: Las 10 variables más importantes para el proceso de clasi cación identi cadas por el algoritmo ERGS para el conjunto de datos de cáncer de colon.

Leucemia de linaje mixto (MLL)		
Número de acceso del gen	Nombre del gen	Descripción del gen
328447_at	Hs.211582	gnl|UG|Hs#S417769 Homo sapiens myosin light chain kinase (MLCK) mRNA, complete cds
1389_at	Hs.1298	gnlUGHs#S1945 Human common ALL antigen (CALLA) mRNA, complete cds
36239_at	Hs.2407	gnlUGHs#S226199 H.sapiens mRNA for oct-binding factor
37539_at	Hs.79219	gnlUGHs#S1569334 Homo sapiens mRNA for KIAA0959 protein, partial cds
39931_at	Hs.38018	gnlUGHs#S952957 Homo sapiens mRNA for protein kinase, Dyrk3
266_s_at	Hs.278667	gnlUGHs#S885 Homo sapiens CD24 signal transducer mRNA, complete cds and 3 region
32579_at	Hs.7802	gnlUGHs#S6032 Human transcriptional activator (BRG1) mRNA, complete cds
963_at	Hs.166091	gnlUGHs#S5776 H.sapiens mRNA for DNA ligase IV
35260_at	Hs.52081	gnlUGHs#S1367627 Homo sapiens mRNA for KIAA0867 protein, complete cds
32872_at	Hs.202685	gnlUGHs#S1368108 Homo sapiens mRNA; cDNA DKFZp564I083 (from clone DKFZp564I083)

Tabla 2: Las 10 variables más importantes para el proceso de clasi cación identi cadas por el algoritmo ERGS para el conjunto de datos de leucemia.

Después de mostrar estos resultados, en el artículo [1] se asegura que estos resultados fueron correspondientes a los resultados probados clínicamente.

Se hicieron los programas (Anexos) para ambas bases datos, para la de cáncer de colon , se obtuvieron los mismos 10 genes más relevantes en el mismo orden; sin embargo, usando los datos de MLL se recuperaron 4 de los 10 genes más relevantes reportados en el artículo, dos en el mismo orden y dos en diferente orden. Los resultados se muestran en la Tabla 3 y Tabla 4.

Colon	
Resultados Chandra	Resultados programa
Hsa.6814	Hsa.6814
Hsa.2928	Hsa.2928
Hsa.831	Hsa.831
Hsa.549	Hsa.549
Hsa.6472	Hsa.6472
Hsa.773	Hsa.773
Hsa.5971	Hsa.5971
Hsa.3016	Hsa.3016
Hsa.601	Hsa.601
Hsa.8147	Hsa.8147

Tabla 3: Comparación de resultados para datos de colon.

MLL	
Resultados Chandra	Resultados programa
328447_at	328447_at
1389_at	1389_at
36239_at	34168_at
37539_at	753_at
39931_at	37576_at
266_s_at	38299_at
32579_at	32551_at
963_at	36239_at
35260_at	37187_at
32872_at	37539_at

Tabla 4: Comparación de resultados para datos de MLL.

No fue posible reproducir los resultaods del artículo [1] con el algoritmo ERGS y no se obtuvo respuesta por parte del autor cuando se le preguntó sobre los detalles de la implementación.

4. Simulación

Se desea probar el algoritmo ERGS con un ejemplo de simulación con el propósito de evaluarlo con condiciones controladas. Se va a construir una matriz M de 62×2000 para simular la dimensión de la base de datos de cáncer de colon pero compuesta de variables uniformes, se van a extraer aleatoriamente X_1, X_2 y X_3, estas variables uniformes se van a transformar en Y_1, Y_2, y Y_3 y se van a regresar a M ; por otro lado Y_1,Y_2 y Y_3 se usarán para generar la variable de referencia determinada por las funciones f_1,f_2,f_3, g_1, g_2 y g_3. Una vez contando con la construcción de M y de la variable de referencia, es aplicará el algoritmo ERGS; lo que debería de esperarse es que se recuperarán las variables Y_1, Y_2 y Y_3 como las variables con mayor in uencia sobre la variable de referencia, ya que desde un principio así fue como se construyó.

Para estas simulaciones se construyó la variable respuesta por medio de las funciones propuestas por Wang y Gevertz (2016) [5], tres funciones aditivas y tres funciones de interacción. La razón por la cual en este artículo se usaron estos escenarios fue porque creían que los datos simulados les ayudarían a encontrar el mejor método para elegir los genes correctos en su estudio sobre el cáncer; en este caso se van a usar los datos simulados para probar un solo método, esto ayudará a observar su comportamiento y robustez.

4.1. Variables

Inicialmente las variables que se toman de la matriz de variables uniformes son $X_1 \sim U(0, 10)$, $X_2 \sim U(0, 10)$ y $X_3 \sim U(0, 10)$, además las transformaciones que se les aplicarán están dadas por $Y_1 = X_1$, $Y_2 = X_1 + 0.35Z_2$, $Y_3 = Y_2 + 0.35Z_3$ y $Y_4 = {}^2Y_5 + {}^2Y_3 + {}^1e$ donde $Z_2 = 2X_2$, $Z_3 = 3X_3$ y $e \sim N(0, 1)$.
A continuación se hará el desarrollo de Y_2, Y_3y Y_4 según la dependencia entre las variables para que las variables queden en términos de distribuciones uniformes iniciales.

$$\begin{aligned} Y_2 &= X_1 + 0.35Z_2 \\ &= X_1 + 0.35\,(2X_2) \\ &= X_1 + 0.7X_2 \end{aligned}$$

así, $Y_2 = X_1 + 0.7X_2$.

$$\begin{aligned} Y_3 &= X_2 + 0.35Z_3 \\ &= X_1 + 0.7X_2 + 0.35\,(3X_3) \\ &= 1.05X_3 + X_1 + 0.7X_2 \end{aligned}$$

así, $Y_3 = X_1 + 0.7X_2 + 1.05X_3$.

$$\begin{aligned} Y_4 &= \frac{2}{3}Y_1 + \frac{2}{3}Y_2 + \frac{1}{3}e \\ &= \frac{2}{3}X_1 + \frac{2}{3}(X_1 + 0.7X_2) + \frac{1}{3}e \\ &= \frac{4}{3}X_1 + \frac{7}{15}X_2 + \frac{1}{3}e \end{aligned}$$

y así, $Y_4 = \frac{4}{3}X_1 + \frac{7}{15}X_{2p} + \frac{1}{3}e$.

4.2. Funciones

Ahora para la construcción de la variable referencia que a partir de Y_1, Y_2 y Y_3 se establecerán las funciones aditivas f_1, f_2 y f_3 y las funciones de interacción g_1, g_2 y g_3, todas éstas generan una respuesta con dos clases, es decir, $j = 1, 2$. En seguida se muestra cada una y su desarrollo.

4.2.1. Funciones aditivas

Para f_1

$$f_1 = \begin{cases} 1, & \text{si } 2Y_1 + 3Y_2 + 4Y_3 + s > c_1 \\ 0, & \text{en otro caso} \end{cases}$$

$$\begin{aligned} f_1 : 2Y_1 + 3Y_2 + 4Y_3 + s \\ &= 2X_1 + 3\,(X_1 + 0.7X_2) + 4\,(X_1 + 0.7X_2 + 1.05X_3) + s \\ &= 2X_1 + 3X_1 + 2.1X_2 + 4X_1 + 2.8X_2 + 4.2X_3 + s \\ &= 9X_1 + 4.9X_2 + 4.2X_3 + s, \end{aligned}$$

donde $c_1 = 88.4$ y $s \sim N(0, 1)$ es un término de ruido.
Para f_2

$$f_2 = \begin{cases} 1, & \text{si } Y_1^2 + Y_2^2 + Y_3 > c_2 \\ 0, & \text{en otro caso} \end{cases}$$

$$
\begin{aligned}
f_2 : & \; Y_1^2 + Y_2^2 + Y_3 \\
& = X_1^2 + (X_1 + 0.7X_2)^2 + X_1 + 0.7X_2 + 1.05X_3 \\
& = X_1^2 + X_1^2 + 1.4X_1X_2 + 0.49X_2 + X_1 + 0.7X_2 + 1.05X_3 \\
& = 2X_1^2 + X_1 + 1.4X_1X_2 + 1.19X_2 + 1.05X_3,
\end{aligned}
$$

donde $c_2 = 95$.
Para f_3

$$
f_3 = \begin{cases} 1, & \text{si } (Y_1 - \mu_1)^2 + (Y_2 - \mu_2) + (Y_3 - \mu_3) > c_3 \\ 0, & \text{en otro caso} \end{cases}
$$

$$
\begin{aligned}
f_3 : & \; (Y_1 - \mu_1)^2 + (Y_2 - \mu_2) + (Y_3 - \mu_3) \\
& = (X_1 - \mu_1)^2 + (X_1 + 0.7X_2 - \mu_2) + (X_1 + 0.7X_2 + 1.05X_3 - \mu_3) \\
& = (X_1 - \mu_1)^2 + 2X_1 + 1.4X_2 + 1.05X_3 - \mu_2 - \mu_3 \\
& = (X_1 - 5)^2 + 2X_1 + 1.4X_2 + 1.05X_3 - 8.3 - 13.6 \\
& = (X_1 - 5)^2 + 2X_1 + 1.4X_2 + 1.05X_3 - 21.9
\end{aligned}
$$

donde $c_3 = 4.9$, $\mu_1 = 5$, $\mu_2 = 8.3$ y $\mu_3 = 13.6$.

4.2.2. Funciones con interacción

Para g_1

$$
g_1 = \begin{cases} 1, & \text{si } Y_1 + Y_2 + Y_3 + Y_1Y_2 + Y_1Y_3 + Y_2Y_3 + Y_1Y_2Y_3 > k_1 \\ 0, & \text{en otro caso} \end{cases}
$$

$$
\begin{aligned}
g_1 : & \; Y_1 + Y_2 + Y_3 + Y_1Y_2 + Y_1Y_3 + Y_2Y_3 + Y_1Y_2Y_3 \\
& = X_1 + (X_1 + 0.7X_2) + (X_1 + 0.7X_2 + 1.05X_3) + X_1 (X_1 + 0.7X_2) + X_1 (X_1 + 0.7X_2 + 1.05X_3) \cdots \\
& \cdots + (X_1 + 0.7X_2) (X_1 + 0.7X_2 + 1.05X_3) + X_1 (X_1 + 0.7X_2) (X_1 + 0.7X_2 + 1.05X_3) \\
& = X_1 + X_1 + 0.7X_2 + X_1 + 0.7X_2 + 1.05X_3 + X_1^2 + 0.7X_1X_2 + X_1^2 + 0.7X_1X_2 + 1 - 05X_1X_3 + \cdots \\
& \cdots \\
& = X_1^3 + 1.4X_1^2X_2 + 1.05X_1^2X_3 + 3X_1^2 + 0.49X_1X_1^2 + 0.735X_1X_2X_3 \cdots \\
& \cdots + 2.8X_1X_2 + 2.1X_1X_3 + 3X_1 + 0.49X_2 + 0.735X_2X_3 + 1.4X_2 + 1.05X_3,
\end{aligned}
$$

donde $k_1 = 698$.

Para g_2

$$g_2 = \begin{cases} 1, & \text{si } Y_1 \cdot Y_2 \cdot Y_3 \cdot Y_4 > k_2 \\ 0, & \text{en otro caso} \end{cases}$$

$$g_2 : Y_1 \cdot Y_2 \cdot Y_3 \cdot Y_4 = X_1 (X_1 + 0.7X_2)(X_1 + 0.7X_2 + 1.05X_3) - \frac{4}{3} X_1 + \frac{7}{15} X_2 + \frac{1}{3} e^{\Sigma}$$

.

$$= X^3_1(1.33X_1 + 1.4X_3) + X_2 (X_1 (2.33X_1 + 1.47X_3) + X_2 (0.22X_1X_2 + X_1 (1.3X_1 + 0.34X_3))) .$$

donde $k_2 = 6000$.

Para g_3

$$g_3 = \begin{cases} 1, & \text{si } Y_1Y_2 > k_3 \text{ y } Y_3 < k_4 \\ 0, & \text{en otro caso} \end{cases}$$

$$g_3 : Y_1Y_2 = X_1 (X_1 + 0.7X_2) = X^2_1 + 0.7X_1X_2$$

y

$$Y_3 = X_1 + 0.7X_2 + 1.05X_3,$$

es decir,

$$g_3 = \begin{cases} 1, & \text{si } X^2_1 + 0.7X_1X_2 > k_3 \text{ y } X_1 + 0.7X_2 + 1.05X_3 < k_4 \\ 0, & \text{en otro caso,} \end{cases}$$

donde $k_3 = 10.5$ y $k_4 = 17$. Los valores de c_i y k_j en cada caso se determinaron para asegurar que las muestras están relativamente balanceadas.

Por los desarrollos anteriores, es más simple detectar:

- Que la variable X_1, X_2 y X_3 son las que tienen mayor peso para las funciones f_1, f_2, g_1 y g_2 cuando éstas toman el valor de 1.
- Para la función f_3; X_1, X_2 y X_3 sí tienen in uencia cuando la función toma el valor de 1; sin embargo, cuando los valores de X_1 son cercanos a la media poblacional μ_1 esta variable no contribuye signi cativamente.
- Para la función g_3, lo que sucede es una sobre expresión de los genes que interactúan (X_1 y X_2)

emparejado con una expresión (X_3). No es claro que X_1 tenga la mayor in uencia sobre g_3 dado que, X_3 tiene más peso sobre la segunda condición ($< k_3$) y se requiere que se cumplan ambas condiciones para que $g_3 = 1$.

4.3. Programa

El programa para hacer esta simulación sigue el siguiente proceso:

- De ne una matriz M de 62×2000 de variables $U(0, 10)$ que, para continuar con la notación en la siguiente sección, estas variables se llamarán m_j para $j = \{1, 2, 3, \ldots, 2000\}$.
- Genera una variable llamada pos para determinar aleatoriamente qué variables m_j se van a extraer de la matriz M.
- Éstas variables se van a nombrar X_1,X_2 y X_3, donde X_1 es la variable número pos[1] de la matriz M, X_2 es la variable número pos[2] de la matriz M y X_3 es la variable número pos[3] de la matriz M.
- Se realizan las correspondientes transformaciones para obtener Y_1, Y_2, Y_3 y Y_4, cabe aclarar que Y_4 es una variable que se genera a partir de Y_1 y Y_2, no se extrae de M.
- Ahora se aplica el algoritmo ERGS

El código se encuentra en la sección de Anexos.

4.4. Resultados

4.4.1. Funciones f_1, f_2, g_1 y g_2

Recordando que estas funciones en términos de X_1, X_2 y X_3 son:

$$f_1 = \begin{cases} 1, & \text{si } 9X_1 + 4.9X_2 + 4.2X_3 + s > 88.4 \\ 0, & \text{en otro caso} \end{cases}$$

$$f_2 = \begin{cases} 1, & \text{si } 2X_1^2 + X_1 + 1.4X_1X_2 + 1.19X_2 + 1.05X_3 > 95 \\ 0, & \text{en otro caso} \end{cases}$$

$$g_1 = \begin{cases} 1, & = X_1^3 + 1.4X_1^2X_2 + 1.05X_1^2X_3 + 3X_1^2 + 0.49X_1X_2^2 + 0.735X_1X_2X_3 \cdots \\ & \cdots + 2.8X_1X_2 + 2.1X_1X_3 + 3X_1 + 0.49X_2^2 + 0.735X_2X_3 + 1.4X_2 + 1.05X_3 \text{ si} > 698 \\ 0, & \text{en otro caso} \end{cases}$$

$$g_2 = \begin{cases} 1, & \text{si } X_1^3(1.33X_1 + 1.4X_3) + X_2\,(X_1\,(2.33X_1 + 1.47X_3) + X_2\,(0.22X_1X_2 + X_1\,(1.3X_1 + 0.34X_3))) \,. > 6000 \\ 0, & \text{en otro caso} \end{cases}$$

El histograma de la Figura 5, muestra un ejemplo del comportamiento de los datos con la simulación f_1, este comportamiento es similar cuando se hace la simulación usando las funciones f_2, g_1 y g_2. Sean W_m el conjunto de los valores de los pesos calculados por el método en orden descendiente y W_{mi} el conjunto de las posiciones en M de las variables correspondientes a los pesos de W_m. Como el orden en estos cojuntos es relevante, se usará el super índice j para indenti car los valores dentro de los conjuntos W_m y W_{mi} donde $j = \{1, 2, 3, \ldots, 2000\}$.

El histograma de pesos W_m detectados por el método ERGS están concentrados la mayoría en valores bajos, esto es por la poca relevancia de la mayoría de las variables. La variable con mayor peso se detecta fácilmente alejada de los demás, hasta el extremo derecho del histograma.

Para este caso en particular usando f_1, tenemos que la variable que determina qué variables se van a extraer de la matriz M es la siguiente,

$$\text{pos} = \{1435, 530, 1658\}$$

entonces $X_1 = m_{pos_1} = m_{1435}$, es decir, X_1 es la variable número 1435 de la matriz M ; $X_2 = m_{pos_2} = m_{530}$, es decir, X_2 es la variable número 530 de la matriz M y $X_3 = m_{pos_3} = m_{1658}$, es decir, X_3 es la variable número 1658 de la matriz M . El algoritmo ERGS debería detectar a X_1, X_2 y X_3 como las variables con mayor peso, ya que son éstas con las que se construye la variable respuesta mediante la función f_1.

Después de hacer las transformaciones correspondientes con f_1 y aplicar el algoritmo ERGS, se obtiene W_m que es la variable de los valores de los pesos en orden descendiente

$$W_m = 1, 0.6450959, 0.6333690, 0.6282232, \ldots$$

se obtiene W_{mi} que es la variable que contiene la posición en la matriz M de las variables correspondiente a su respectivo peso en el conjunto W_m

$$W_{mi} = 1435, 689, 530, 1658, \ldots,$$

esto quiere decir que la variable 1435 tiene peso de 1, la variable 689 tiene peso de 0.6450959, la variable 530 tiene peso de 0.6333690, la variable 1658 tiene peso de 0.6282232 y así sucesivamente. En este caso, el método fue exitoso, dado que las variables 1435, 530 y 1658 fueron detectadas dentro de los mayores pesos y éstas son $m_{1435} = X_1$, $m_{530} = X_2$ y $m_{1658} = X_3$; las variables con las que se construyó la variable respuesta.

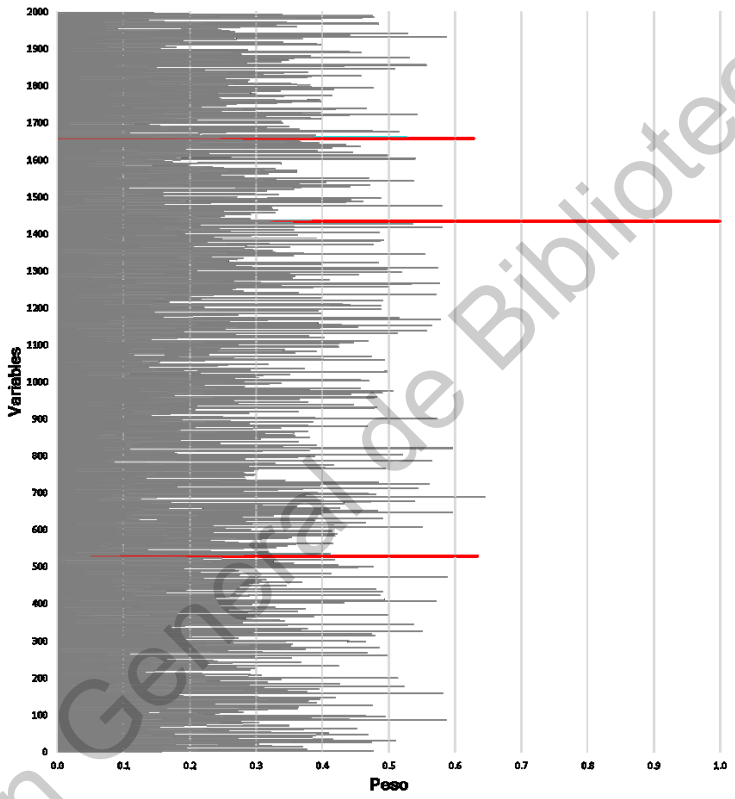

Figura 4: Variables por número de posición en la matriz contra su peso correspondiente.

Figura 5: Histograma de pesos usando f_1.

Al observar más a detalle el histograma con sólo los 20 mayores pesos, se distingue la variable m_{1435} con mayor peso con mayor facilidad (Figura 6), además en la Figura 4 se puede corroborar la frecuencia que muestran los histogramas y se resaltan las variables X_1, X_2 y X_3 que fueron detctadas.

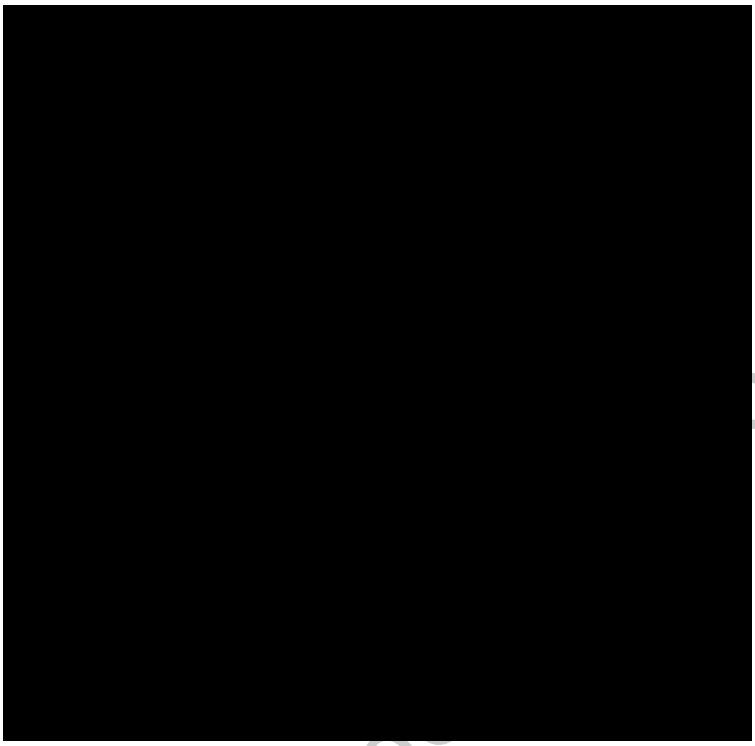

Figura 6: Histograma usando la función f_1 considerando los 20 pesos mayores w_i.

Así es como grá camente podemos observar que, en efecto, se detectan X_1, X_2 y X_3 que fueron construidas para ser las de mayor in uencia sobre la variable respuesta.

4.4.2. Función f_3

Recordando que esta función en términos de X_1, X_2 y X_3 es:

$$f_3 = \begin{cases} 1, & \text{si } (X_1 - 5)^2 + 2X_1 + 1.4X_2 + 1.05X_3 - 21.9 > 4.9 \\ 0, & \text{en otro caso} \end{cases}$$

Por otro lado, para la simulación usando f_3, los pesos de las variables no tienden a valores tan bajos como en la simulación antes mencionada. Diversas variables tienden a ser relevantes, esto lo podemos observar en la Figura 7. Para este caso en particular usando f_3, tenemos que la variable que determina qué variables se van a extraer de la matriz M es la siguiente,

$$\text{pos} = \{297, 220, 132\}$$

entonces $X_1 = m_{pos_1} = m_{297}$, es decir, X_1 es la variable número 297 de la matriz M; $X_2 = m_{pos_2} = m_{220}$, es decir, X_2 es la variable número 220 de la matriz M y $X_3 = m_{pos_3} = m_{132}$, es decir, X_3 es la variable número 132 de la matriz M. El algoritmo ERGS debería detectar a X_1, X_2 y X_3 como las variables con mayor peso, ya que son éstas con las que se construye la variable respuesta mediante la función f_3.

Después de hacer las transformaciones correspondientes con f_3 y aplicar el algoritmo ERGS, se obtiene W_m que es la variable de los valores de los pesos en orden descendiente

$$W_m = 0.7629462,\ 0.74923130,\ 0.7094006,\ 0.6986233,\ \ldots$$

se obtiene W_{mi} que es la variable que contiene el número (o nombre) de las variables correspondiente a su respectivo peso en W_m

$$W_{mi} = 1983,\ 1708,\ 611,\ 282,\ldots,$$

los valores de los pesos de las variables X_1, X_2 y X_3 resultaron ser

$$W_m^{22} = 0.5799801,$$

$$W_m^{89} = 0.4540688 \text{ y}$$

$$W_m^{25} = 0.5669151$$

respectivamente, donde

$$W_{mi}^{22} = 297,$$

$$W_{mi}^{89} = 220 \text{ y}$$

$$W_{mi}^{25} = 132.$$

Dicho de otra forma, la variable m_{297} fue la 22- variable más relevante, la variable m_{220} fue la 89- variable más relevante y la variable m_{13} fue la 25- variable más relevante con sus respectivos pesos

$$W_{m_{297}} = 0.5799801$$
$$W_{m_{220}} = 0.4540688$$
$$W_{m_{132}} = 0.5669151$$

El histograma de la Figura 7 nos muestra que los pesos de las variables X_1, X_2 y X_3 son menores que 0.6, que es donde comienza a haber mayor frecuencia, es decir, la mayoría de las variables tiene un peso menor que 0.6.. Esta simulación no fue tan exitosa como las anteriores, sin embargo, de 2000 variables detectó a X_1y a X_3dentro de las 25 más relevantes.

Figura 7: Histograma de pesos usando f_3.

4.4.3. Función g_3

Recordando que esta función en términos de X_1, X_2 y X_3 es:

$$g_3 = \begin{cases} 1, & \text{si } X_1^2 + 0.7X_1X_2 > 10.5 \text{ y } X_1 + 0.7X_2 + 1.05X_3 < 17 \\ 0, & \text{en otro caso,} \end{cases}$$

Para este caso en particular usando g_3, tenemos que la variable que determina qué variables se van a extraer de la matriz M es la siguiente,

$$\text{pos} = \{378, 1742, 1013\},$$

entonces $X_1 = m_{pos_1} = m_{378}$, es decir, X_1 es la variable número 378 de la matriz M ; $X_2 = m_{pos_2} = m_{1742}$, es decir, X_2 es la variable número 1742 de la matriz M y $X_3 = m_{pos_3} = m_{1013}$, es decir, X_3 es la variable número 1013 de la matriz M. El algoritmo ERGS debería detectar a X_1, X_2 y X_3 como las

variables con mayor peso, ya que son éstas con las que se construye la variable respuesta mediante la función g_3.

Después de hacer las transformaciones correspondientes con g_3 y aplicar el algoritmo ERGS, se obtiene W_m que es la variable de los valores de los pesos en orden descendiente

$$W_m = 0.7894975,\ 0.7037191,\ 0.6727974,\ 0.6394442,\ \ldots$$

se obtiene W_{mi} que es la variable que contiene el número (o nombre) de las variables correspondiente a su respectivo peso en W_m

$$W_{mi} = 1013,\ 884,\ 1142,\ 1187,\ \ldots$$

Los resultados son:

- La variable X_3, tiene un peso de 0.7894975 y fue la variable más relevante de acuerdo con el algoritmo ERGS.
- La variable X_1, tiene un peso de 0.5973526, sin embargo, X_1 fue la 12- variable más relevante.
- La variable X_2, tiene un peso de 0.4298197, sin embargo, X_2 fue la 355- variable más relevante.

Grá camente, el histograma de pesos está centrado en 0.4 esto indica que la concentración de pesos se centra en valores poco relevantes, Figura 8.

Figura 8: Histograma de pesos usando g_3.

En la Figura 9 se muestran las 20 variables con mayor peso, aquí se encuentran X_3 y X_1

Figura 9: Histograma usando g_3 con los 20 mayores w_i.

De las tres variables que se debieron de haber detectado con este método, se detectaron 2 dentro de las 12 variables con mayor relevancia. Esta función no fue tan exitosa como f_1, f_2, g_1 y g_2, ya que X_3 se identi có como la variable número 355 más relevante de 2000 variables.

4.5. Conclusión simulación

Estas simulaciones se verían re ejadas más precisamente para un método de envoltura dadas las variables que se construyen desde el inicio, unas dependen de otras. Sin embargo, sí es posible ver el éxito del método ERGS, que es un método de ltro, dado que detecta la variable que in uye sobre la variable de referencia que se genera y sobre las variables construidas dependientes de ésta.

5. Aplicación

5.1. Descripción

El estudio en el que esta tesis se apoyará para poder aplicar el método ERGS, fue realizado por la Facultad de Química de la Universidad Autónoma de Querétaro; el propósito del estudio fue con nes ajenos a este trabajo, sin embargo resulta interesante dadas las dimensiones de las bases de datos además de que las variables implicadas son metabolitos tomados de una población de 44 mujeres y 51 hombres.

La forma de ambas bases de datos es la siguiente: la primera columna llamada *numero* corresponde a un identi cador de cada sujeto, la segunda columna llamada FACTOR GPOS corresponde a la variable respuesta o variable referencia y, tanto para la población de mujeres como la de hombres, tienen 5 clases; las demás columnas son las mediciones de los metabolitos de cada sujeto con sus respectivos nombres.

La base de datos de mujeres tiene dimensión 44×94 y la base de datos de hombres de 51×95, siendo cada variable es un metabolito; ambas variables respuesta tienen las clases: BF %, BP, HDL-c, Healthy y TG. Los datos que son iguales a cero representan una ausencia de registro de resultados.

Se aplicó el método ERGS para las dos bases de datos, los resultados se muestran en las siguiente secciones.

5.2. Metabolitos de mujeres

En la Tabla 5, se muestran las diez variables con mayores pesos y los valores de éstos

Relevancia	Variables	w_i
1	Stearic.acid	0.8904811
2	Corticosterone	0.8511226
3	Gamma.Glutamyl.phenylalanine	0.8358236
4	Nervonic.acid	0.8339311
5	Lignoceric.acid	0.8336201
6	Erucic.acid	0.8043068
7	PC..18.0.18.2.	0.8023021
8	Arachidic.acid	0.8000730
9	TAG..16.0.16.0.16.0.	0.7917410
10	TAG..20.0.20.0.20.0.	0.7903822

Tabla 5: Las 10 variables con mayor peso.

Para visualizar el comportamiento de estas variables se construyeron diagramas de caja de cada una, véase la Figura 10.

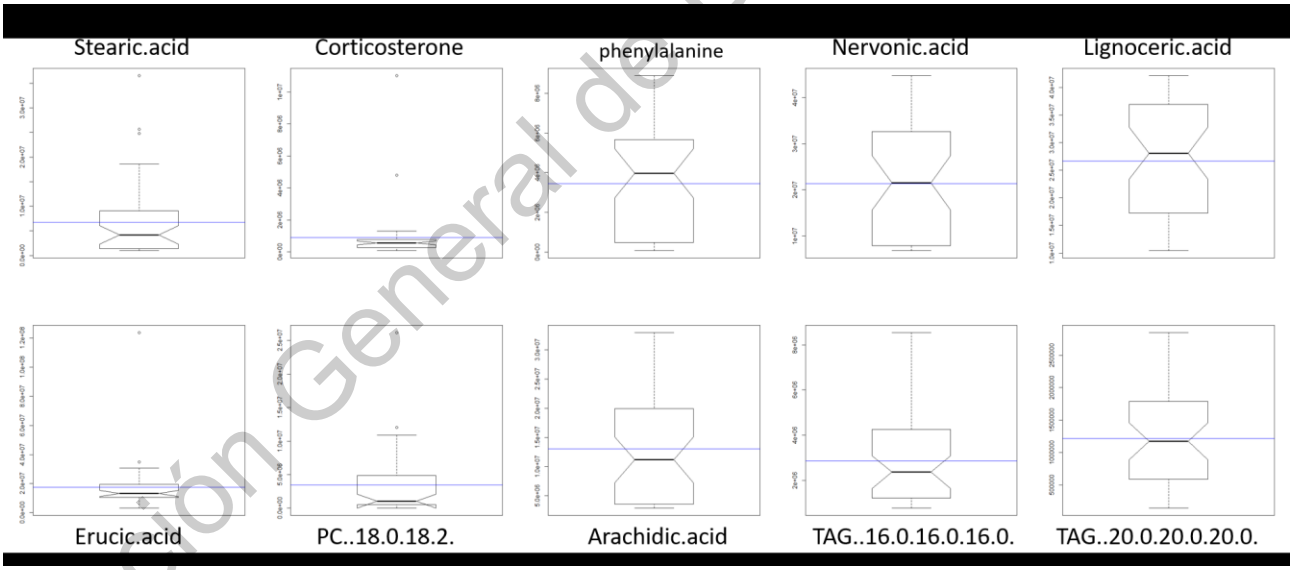

Figura 10: Diagramas de caja de las variables de los metabolitos de mujeres con mayor peso.

Cabe aclarar que la línea azul en cada grá ca representa la media para cada caso, esto es para observar cómo los datos atípicos tienden a crear una mayor separación entre la media y la mediana.

Cada variable está en su propia escala dada la forma de medición de cada metabolito, por esta razón, la Tabla 6 muestra un resumen de los valores más relevantes de cada variable.

Nombre de la variable	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo
Stearic.acid	**96,988**	**1,360,775**	**4,184,144**	**6,774,083**	**9,070,378**	**36,564,634**
Corticosterone	**75,154**	**286,078**	**564,831**	**896,023**	**761,310**	**11,033,524**
Gamma.Glutamyl. p	67,411	519,788	3,971,595	3,449,223	5,591,451	8,901,095
Nervonic.acid	6,824,724	7,965,791	21,550,722	21,368,807	32,625,726	44,754,968
Lignoceric.acid	10,434,753	17,317,190	28,087,818	26,670,985	36,911,094	42,186,576
Erucic.acid	**3,134,526**	**10,747,241**	**13,267,110**	**17,402,933**	**19,126,554**	**123,903,672**
PC..18.0.18.2.	**0**	**557,323**	**1,056,274**	**3,439,297**	**4,592,937**	**26,194,036**
Arachidic.acid	2,860,038	3,615,588	11,157,982	13,009,914	19,529,529	33,028,934
TAG..16.0.16.0.16.0.	772,976	1,226,627	2,366,146	2,863,041	4,255,563	8,542,759
TAG..20.0.20.0.20.0.	144,281	616,446	1,174,317	1,215,017	1,780,764	2,849,390

Tabla 6: Resumen de medidas por variable.

Las variables resaltadas en la tabla anterior, *Stearic.acid*, *Corticosterone*, *Erucic.acid* y *PC..*18.0.18.2. contienen datos atípicos, esto se puede detectar por las grá cas de la Figura 10 y los valores en la Tabla 6, dado que las cajas muestran dónde se encuentran la mayoría de los datos y los puntos son los atípicos.

En el caso de *PC..*18.0.18.2. resulta tener datos atípicos y su diagrama de caja pierde su forma cuadrada porque, de sus 45 observaciones, el valor de 10 de estas es cero mientras que su máximo es 26,194,036.

Se quitaron los datos atípicos de cada variable y se repitió el algoritmo, para saber si estos datos atípicos afectan la relevancia de los metabolitos de los hombres según el método ERGS.

Stearic acid

Su diagrama de caja original individual se muestra en la Figura 11.

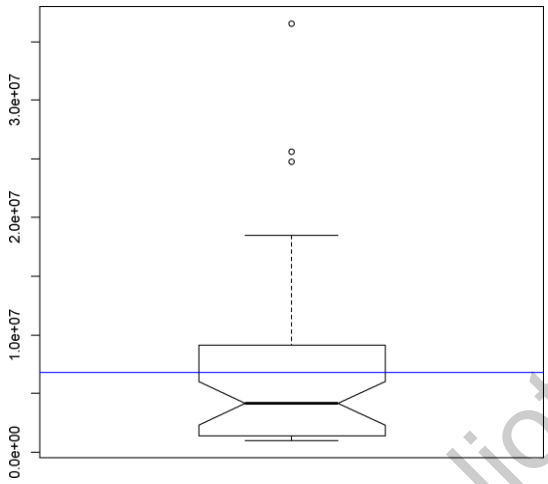

Figura 11: Diagrama de caja de la variable Stearic acid para la base de datos de mujeres.

el cual se detecta como la variable con mayor relevancia con un peso de $w_1 = 0.8904811$, sin embargo, se pueden observar algunos atípicos. Ahora se procederá a quitar estos valores atípicos mientras se registra el cambio de relevancia para la variable *Stearic acid* en la Tabla 7.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
1	0.8904811
Quitando objeto 28 con valor de 36, 564, 634	
3	0.8437433
Quitando el objeto 26 con valor de 25, 650, 045	
7	0.7979616
Quitando el objeto 25 con valor de 24, 792, 232	
10	0.7479771
Quitando el objeto 21 con valor de 18, 526, 645	
16	0.7138754

Tabla 7: Cambio de orden de relevancia y w_i de la variable Stearic acid.

Así vemos la variable Stearic acid había sido detectada como la más relevante con un w_i de 0.8904811 y quitando los valores atípicos fue identi cada con una relevancia de 16 y un w_i de 0.7138754.

Corticosterone

Su diagrama de caja individual se muestra en la Figura 12.

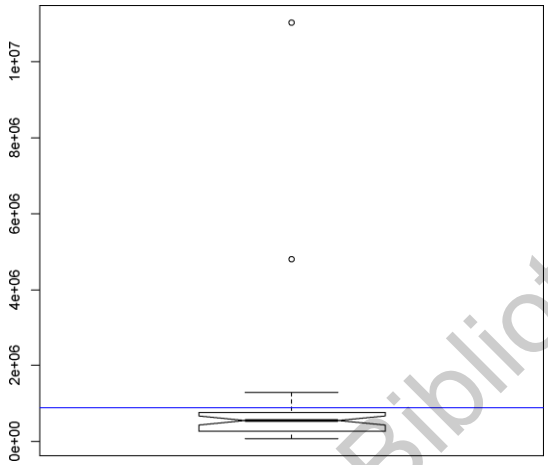

Figura 12: Diagrama de caja de la variable Corticosterone para la base de datos de mujeres..

el cual se detecta como la segunda variable con mayor relevancia con un peso de $w_2 = 0.8511226$. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *Corticosterone* en la Tabla 8.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
2	0.8511226
Quitando el objeto 29 con valor de 11, 033, 524	
10	0.7741171
Quitando el objeto 23 con valor de 4, 791, 405	
36	0.5001073

Tabla 8: Cambio de orden de relevancia y w_i de la variable Corticosterone.

Así vemos la variable Corticosterone originalmente se detectó como la segunda variable más relevante con un peso de 0.8511226, sin embargo, quitando los valores atípicos fue identi cada con una relevancia de 36 y un peso de 0.5001073.

Erucic Acid

Su diagrama de caja individual se muestra en la Figura 13.

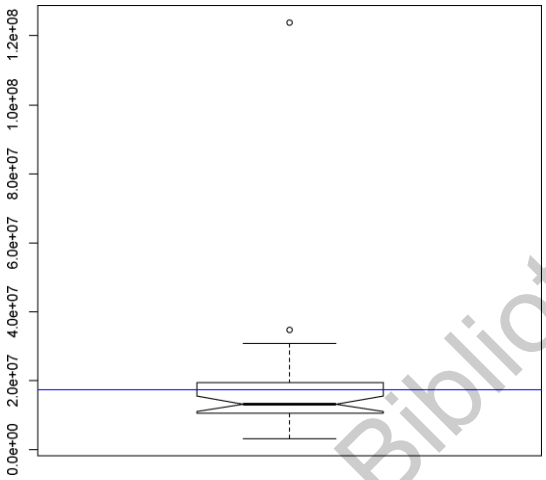

Figura 13: Diagrama de caja de la variable Erucic Acid para la base de datos de mujeres..

el cual se detecta como la sexta variable con mayor relevancia con un peso de 0.8043068. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *Erucic Acid* en la Tabla 9.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
6	0.8043068
Quitando el objeto 19 con valor de 123, 903, 672	
21	0.6410578
Quitando el objeto 24 con valor de 34, 807, 356	
30	0.5828769

Tabla 9: Cambio de orden de relevancia y w_i de la variable Erucic Acid.

Así vemos la variable Erucic acid, que había sido detectada como la sexta más relevante con un peso de 0.8043068, quitando los valores atípicos fue identi cada con una relevancia de 30 y un peso de 0.5828769.

PC 18:0 18:2

Su diagrama de caja individual se muestra en la Figura 14.

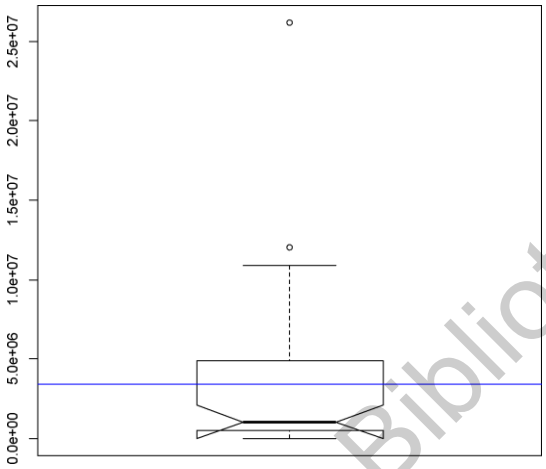

Figura 14: Diagrama de caja de la variable PC 18:0 18:2 para la base de datos de mujeres.

esta variable se detecta como la séptima variable con mayor relevancia con un peso de 0.8023021. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *PC* 18 : 0 18 : 2 en la Tabla 10.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
7	0.8023021
Quitando el objeto **31** con valor de **26, 194, 036**	
5	0.8205216
Quitando el objeto **38** con valor de **12, 010, 598**	
6	0.7967582
Quitando el objeto **39** con valor de **10, 909, 637**	
5	0.7774904
Quitando el objeto **41** con valor de **10, 866, 892**	
10	0.7754468
Quitando el objeto**32** con valor de **10, 697, 456**	
16	0.7637212
Quitando el objeto **29** con valor de **9, 459, 968**	
30	0.5828769

Tabla 10: Cambio de orden de relevancia y w_i de la variable PC 18:0 18:2.

Así vemos la variable PC 18:0 18:2 había sido detectada como la septima más relevante con un peso de **0.8023021** y quitando sus atípicos fue identi cada con una relevancia de 30 y con un peso de **0.5828769**, menor que el que tenía originalmente.

5.3. Metabolitos de hombre

A continuación, en la Tabla 11, se muestran las diez variables con mayores pesos y los valores respectivos

Relevancia	Variables	w_i
1	CER..16.1.22.1.	0.9855936
2	Gamma.Glutamyl.tryptophan	0.9407838
3	Gamma.Glutamyl.glutamate	0.9186337
4	TG..16.1.16.1.16.1.	0.8869246
5	TAG..16.0.16.0.16.1.	0.8695870
6	DG..18.1.18.1.	0.8653974
7	Gamma.Glutamyl.histidine	0.8374600
8	Lignoceric.acid	0.8310170
9	Corticosterone	0.8117083
10	Linoleoyl.carnitine	0.8081242

Tabla 11: Las 10 variables con mayor peso.

Para visualizar el comportamiento de estas variables se construyeron diagramas de caja de cada una, véase la Figura 15.

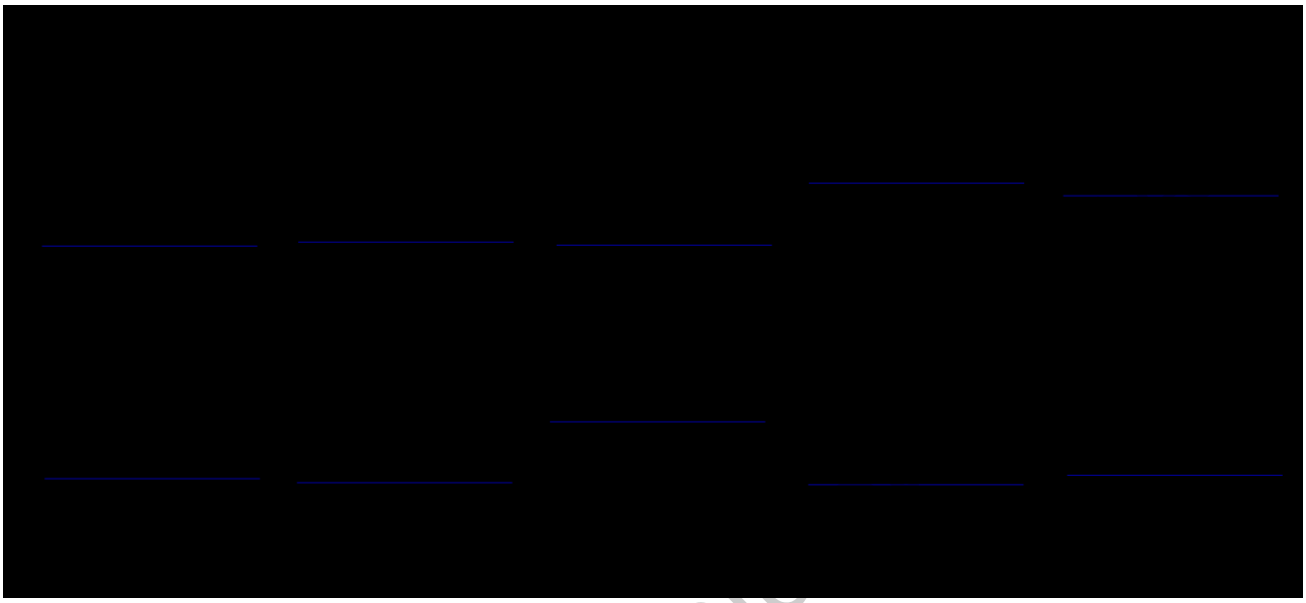

Figura 15: Diagramas de caja de las variables encontradas.

La línea azul en cada grá ca representa la media, esto es para observar cómo los datos atípicos tienden a crear una mayor separación entre la media y la mediana.
La Tabla 12 muestra un resumen de los valores más relevantes de cada variable.

Nombre de la variable	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo
CER..16.1.22.1	48922	172549	247326	1009876	324835	38417732
Gamma Glutamyl tryptophan	98401	266391	366380	782809	744553	12508780
Gamma Glutamyl glutamate	117833	183280	231429	365217	319127	4489268
TG..16.1.16.1.16.1.	10538630	15068212	21789732	21120761	25422706	35273008
TAG..16.0.16.0.16.1.	3147520	6205473	8966154	9891025	11342981	21644901
DG..18.1.18.1.	404633	29090306	37057264	38654846	39910406	281049199
Gamma.Glutamyl.histidine	66796	231236	565636	1009852	1477171	9291171
Lignoceric.acid	14630709	17129332	29038660	27367782	36230184	42686460
Corticosterone	15638	189535	349624	614603	763814	5789884
Linoleoyl.carnitine	139051	305206	559916	946793	1178808	5428702

Tabla 12: Resumen de medidas por variable.

Las variables resaltadas en la tabla anterior, *CER..*16.1.22.1, *ttamma.ttlutamyl.tryptophan*, *ttamma.ttlutamyl.glutamate* *TAtt..*16.0.16.0.16.1, *Dtt..*18.1.18.1., *ttamma.ttlutamyl.histidine*, *Corticosterone* y *Linoleoyl.carnitine*

contienen datos atípicos, esto se puede detectar por las grá cas de la Figura 15 y los valores en la Tabla 12.
Se quitaron los datos atípicos de cada variable y se repitió el algoritmo, para sabe si estos datos atípicos afectan la relevancia de los metabolitos de los hombres según el método ERGS.

CER..16.1.22.1

Su diagrama de caja original individual se muestra en la Figura 16.

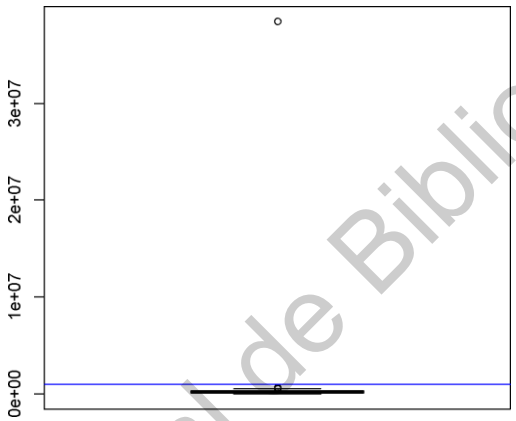

Figura 16: Diagrama de caja de la variable CER..16.1.22.1 para la base de datos de hombres.

esta variable se detecta como la variable con mayor relevancia con un peso de 0.9855936. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *CER*..16.1.22.1 en la Tabla 13.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
1	0.9855936
Quitando 6 atípicos mayores a 465, 000	
95	0.0000000

Tabla 13: Cambio de orden de relevancia y w_i de la variable CER..16.1.22.1.

Así vemos la variable CER.16.1.22.1 había sido detectada como la variable más relevante con un peso de 0.9855936 y quitando sus atípicos fue identi cada con una relevancia de 95 y con un peso de 0, mucho menor que el que tenía originalmente.

Gamma.Glutamyl.tryptophan

Su diagrama de caja original individual se muestra en la Figura 17.

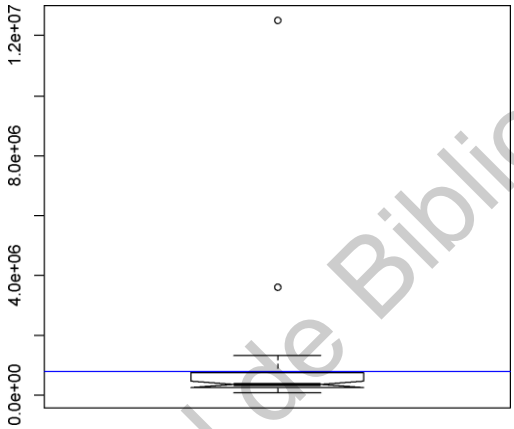

Figura 17: Diagrama de caja de la variable Gamma.Glutamyl.tryptophan para la base de datos de hombres.

esta variable se detecta como la segunda variable con mayor relevancia con un peso de 0.9407838. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *ttamma.ttlutamyl.tryptophan* en la Tabla 14.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
2	0.9407838

Tabla 14: Cambio de orden de relevancia y w_i de la variable Gamma.Glutamyl.tryptophan.

Se quitaron todos los valores mayores a 620, 000, así es como se eliminaron todos los valores atípicos. Una vez que no se tienen atípicos los cálculos del algoritmo fallas por la cantidad de NAs que se presentan.
Así vemos la variable Gamma.Glutamyl.tryptophan había sido detectada como la 2 más relevante con un peso de 0.9407838 y quitando sus atípicos el algoritmo falla para calcular su relevancia y su peso.

Gamma.Glutamyl.glutamate
Su diagrama de caja original individual se muestra en la Figura 18.

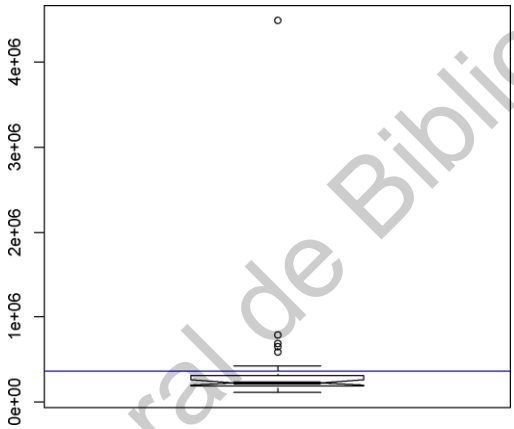

Figura 18: Diagrama de caja de la variable Gamma.Glutamyl.glutamate para la base de datos de hombres.

esta variable se detecta como la tercer variable con mayor relevancia con un peso de 0.9186337. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *ttamma.ttlutamyl.glutamate* en la Tabla 15.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
3	0.9186337

Tabla 15: Cambio de orden de relevancia y w_i de la variable Gamma.Glutamyl.glutamate.

Se quitaron todos los valores menores a 340, 000, así es como se eliminaron todos los valores atípicos. Una vez que no se tienen atípicos los cálculos del algoritmo fallas por la cantidad de NAs que se presentan.
Así vemos la variable Gamma.Glutamyl.glutamate había sido detectada como la 3 más relevante con un peso de 0.9186337 y quitando sus atípicos el algoritmo falla para calcular su relevancia y su peso.

TAG..16.0.16.0.16.1.

Su diagrama de caja original individual se muestra en la Figura 19.

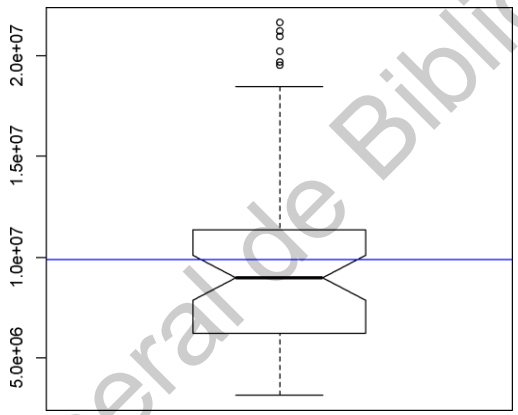

Figura 19: Diagrama de caja de la variable TAG..16.0.16.0.16.1. para la base de datos de hombres.

esta variable se detecta como la quinta variable con mayor relevancia con un peso de 0.8695870. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *TAtt*..16.0.16.0.16.1. en la Tabla 16.

Orden de relevancia	Peso asignado w_i
Con tod.os los objetos	
5	0.8695870
Quitando los valores menores a 16, 000, 000	
35	0.6550171

Tabla 16: Cambio de orden de relevancia y w_i de la variable TAG..16.0.16.0.16.1..

Así vemos la variable TAG..16.0.16.0.16.1. había sido detectada como la quinta variable más relevante con un peso de 0.8695870 y quitando sus atípicos fue identi cada con una relevancia de 35 y con un peso de 0.6550171, menor que el que tenía originalmente.

DG..18.1.18.1.

Su diagrama de caja original individual se muestra en la Figura 20.

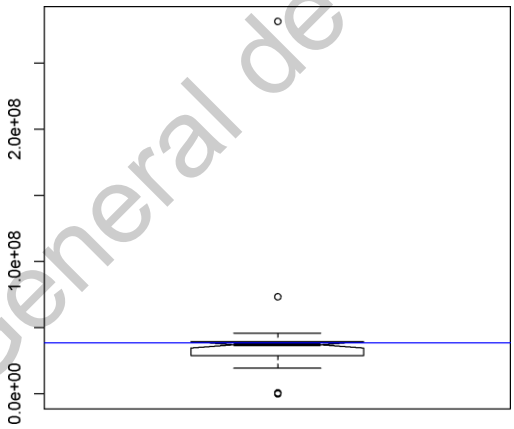

Figura 20: Diagrama de caja de la variable DG..18.1.18.1. para la base de datos de hombres.

esta variable se detecta como la sexta variable con mayor relevancia con un peso de 0.8653974. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *Dtt..*18.1.18.1. en la Tabla 17.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
6	0.8653974
Quitando los valores menores a 740000 y mayores que 60000000	
85	0.3312495

Tabla 17: Cambio de orden de relevancia y w_i de la variable DG. 18.1.18.1.

Así vemos la variable DG..18.1.18.1. había sido detectada como la 6 más relevante con un peso de 0.8653974 y quitando sus atípicos fue identi cada con una relevancia de 85 y con un peso de 0.3312495, menor que el que tenía originalmente.

Gamma.Glutamyl.histidine

Su diagrama de caja original individual se muestra en la Figura 21.

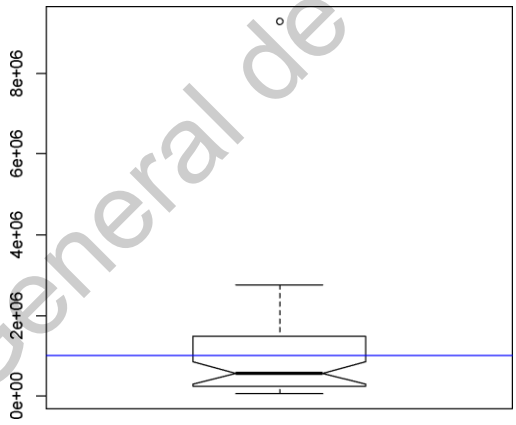

Figura 21: Diagrama de caja de la variable Gamma.Glutamyl.histidine para la base de datos de hombres.

esta variable se detecta como la séptima variable con mayor relevancia con un peso de 0.8374600. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *ttamma.ttlutamyl.histidine* en la Tabla 18.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
7	0.8374600
Quitando los valores mayores que 6, 000, 000	
27	0.6745086

Tabla 18: Cambio de orden de relevancia y w_i de la variable Gamma.Glutamyl.histidine.

Así vemos la variable Gamma.Glutamyl.histidine había sido detectada como la 7 más relevante con un peso de 0.8374600 y quitando un atípicos fue identi cada con una relevancia de 27 y con un peso de 0.6745086, menor que el que tenía originalmente.

Corticosterone

Su diagrama de caja original individual se muestra en la Figura 22.

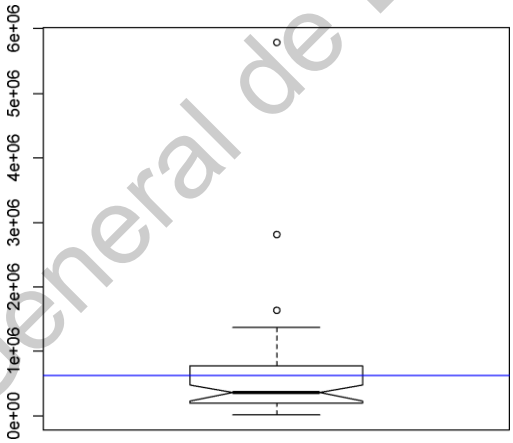

Figura 22: Diagrama de caja de la variable Corticosterone para la base de datos de hombres.

esta variable se detecta como la novena variable con mayor relevancia con un peso de 0.8117083. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *Corticosterone* en la Tabla 19.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
9	0.8117083
Quitando los valores mayores que 1, 300, 000	
52	0.4789272

Tabla 19: Cambio de orden de relevancia y w_i de la variable Corticosterone.

Así vemos la variable Corticosterone había sido detectada como la 9 más relevante con un peso de 0.8117083 y quitando sus atípicos fue identi cada con una relevancia de 52 y con un peso de 0.4789272, menor que el que tenía originalmente.

Linoleoyl.carnitine

Su diagrama de caja original individual se muestra en la Figura 23.

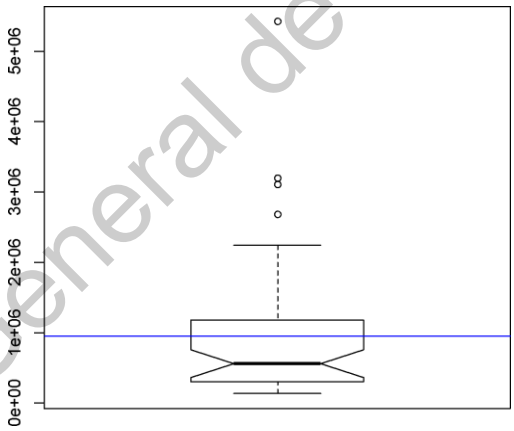

Figura 23: Diagrama de caja de la variable Linoleoyl.carnitine para la base de datos de hombres.

esta variable se detecta como la décima variable con mayor relevancia con un peso de 0.8081242. Ahora se procederá a quitar los valores atípicos mientras se registra el cambio de relevancia para la variable *Linoleoyl.carnitine* en la Tabla 20.

Orden de relevancia	Peso asignado w_i
Con todos los objetos	
10	0.8081242
Quitando los valores mayores que 2, 500, 000	
26	0.6815816

Tabla 20: Cambio de orden de relevancia y w_i de la variable Linoleoyl.carnitine.

Así vemos la variable Linoleoyl.carnitine había sido detectada como la 10 más relevante con un peso de 0.8081242 y quitando sus atípicos fue identi cada con una relevancia de 26 y con un peso de 0.6815816, menor que el que tenía originalmente.

5.4. Conclusión sobre la aplicación

Tanto en el caso de los metabolitos de mujeres como en el caso de los hombres, sucedió que los valores atípicos tuvieron mucha in uencia sobre los resultados del método ERGS en estas bases de datos. Primero se calcularon los pesos de las variables, después se detectaron valores atípicos y se eliminaron; se volvió a aplicar el método ERGS pero ninguna variable recuperó el valor de su peso. Esto indica que no se debe aplicar este método sin haber hecho una limpieza de atípicos a cualquier base de datos. El artículo de Chandra & Gupta [1] asevera que el algoritmo es robusto ante atípicos, estos ejemplos muestran que tal aseveración carece de fundamento.

6. Conclusión

Las conclusiones que se hacen en en el artículo de Chandra y Gupta [1] sobre este algoritmo son:

a. Se seleccionaron las variables de mayor peso con el propósito de clasi cación.

b. El método ERGS anula el efecto de los valores atípicos y grandes varianzas.

c. El algoritmo ERGS no requiere una estrategia de búsqueda computacionalmente extensiva ni criterios de evaluación a diferencia de otros algoritmos de selección de variables.

d. El algoritmo ERGS es rápido, sencillo de implementar y no requiere supuestos de distribución.

Sin embargo, algunas de estas conclusiones se hicieron de forma particular para el artículo de Chandra y Gupta ya que algunas fueron refutadas en este trabajo. Los siguientes puntos son las conclusiones de este trabajo de investigación sobre cada conclusión del artículo de Chandra y Gupta [1] respectivamente.

e. El algoritmo ERGS, a través de los rangos efectivos, selecciona un subconjunto de variables con el mayor peso, sin embargo en el artículo no se da una prueba como tal de la clasi cación después de la elección de estas variables, únicamente se usa el mismo ejemplo particular para mostrarlo. Para poder hacer esta a rmación se debería probar que la selección de variables para cualquier conjunto de datos da pie a una buena clasi cación.

f. Esta segunda conclusión no es una buena generalización, ya que en la sección 5, se aplicó el método sobre dos bases de datos en las cuales el hecho de que hubiera atípicos cambió el subconjunto de variables con los mayores pesos. Los atípicos sí tienen un efecto sobre el resultado de la aplicación del algoritmo ERGS.

g. Computacionalmente el algoritmo es sencillo de implementar y no requiere de la búsqueda de los criterios de evaluación ya que el ERGS recae sobre la construcción de los intervalos, los cuales se construyen con parámetros mencionados en la sección 3. Se deben de tomar algunas medidas previas al análisis porque los intervalos que se construyen a partir de μ_{ij} la cual es una medida de tendencia central que puede ser alterada por los valores atípicos de una variable.

h. Considerando el punto anterior, si los intervalos se construyen de forma simétrica usando la media para cada clase de cada variable entonces sí es relevante saber qué forma tiene la distribución antes de aplicar el algoritmo, esto con el propósito de saber cómo le va a afectar esta construcción de las amplitudes a los resultados del algoritmo ERGS. La media es una medida sensible a los atípicos y la varianza puede llegar a ser mayor por los atípicos; esto puede afectar en el cálculo de las amplitudes haciéndolas más gradese de lo que deberían o sesgadas hacia algún lado.

Las recomendacioes que se hacen para poder implementar el método son:

i. Normalizar: en bases de datos con donde p n es muy común que la distribución de éstos no sea normal y si este es el caso, las amplitudes pueden llegar a ser más grandes y el traslape de éstos pude llegar a ser mayor a lo qu debería; por esta razón se debe considerar la forma de los datos y, en caso de que no presenten una distribución simétrica, entoces se debe aplicar una transformación para simetrizar la distribución antes de aplicar el método ERGS.

j. Limpar atípicos: como se mencionó en el punto f., los atípicos afectan la robustéz del algoritmo. Con la ayuda de un análisis exploratorio, de debe hacer una limpieza a las bases de datos antes de aplicar el método ERGS.

k. Analizar cómo cambian los $w_i$cambiando el valor de γ: como se mencionó en la sección 3.3 se usa $\gamma = 1.732$ para garantizar que se incluyan el 66.7 % de los datos, sin embargo, no se presentan prueba sobre esta a rmación, por esta razón se sugiere justi car el valor de γ y analizar cómo este cambio afecta a los pesos.

Además de las conclusiones sobre el método ERGS y lo que implica; en lo personal este trabajo me hizo ver lo amplio que puede llegar a ser la estadística; a tal grado que puede llegar a in uir sobre la salud de las personas, habiendo estudiado en la Facultad de Ingeniería, nunca pensé que la estadística pudiera tener una responsabilidad social.

Por otro lado, me quedo con algo de preocupación al ver que investigaciones tan importantes como las citadas en esta tesis, tengan detalles teóricos que no se están considerando ni para el proceso de investigación ni se reporten como limitaciones o condiciones iniciales. Esto re eja lo indispensable que es el trabajo entre diferentes áreas y la comunicación de la información.

Considero que sí mejoré como estadística después de haber concluido este trabajo de investigación, . Mejoré la parte teórica en cuanto a lo que fue necesario para el desarrollo de este trabajo, sin embargo, lo más valioso que rescato es la importancia del trabajo interdiciplinario. Es de suma importancia que un estadístico haga un uso consiente de todas sus habilidades para poder resolver los problemas de áreas ajenas a la suya, esto lo llevo conmigo muy presentemente al ámbito laboral.

La formación que tenemos en la Licenciatura en Matemáticas Aplicadas sí contempla la redacción e investigación como asignaturas, sin embargo, me perece que hacer propio un trabajo de investigación a esta profundidad e importancia personal, potencializa estas habilidades de un nivel mucho mayor. Este trabajo me deja en claro el arduo trabajo que es la investigación, lo difícil que es el manejo de tiempos y la importancia de la correcta comunicación de ideas; no únicamente ser claro para uno mismo, sino para cualquier académico que lea. Aún queda mucho por mejorar pero agradezco a mi asesor y sinodales que tuvieron la paciencia y tiempo para ayudarme en estos ámbitos.

Agradecimientos

A mi abuelo Mario I. Villar Borja (D.E.P.), mi mamá Claudia Villar Hernández, mi hermano Pedro Barón Villar y mi esposo Allan Behnsen Romo.

Referencias

[1] Chandra, B., Gupta, M. (2011). An e cient statistical feature selection approach for classi cation of Gene expression data. Journal of Biomedical Informatics, 44(4), 529-535. DOI: 10.1016/j.jbi.2011.01.001

[2] Ayça Çakmak Pehlivanli. (2015). A novel feature selection scheme for high-dimensional data sets: four-Staged Feature Selection.Journal of Applied Statistics, DOI:10.1080/02664763.2015.1092112

[3] Xia, J.,Broadhurst, D.I., Wilson, M., Wishart, D.S. (2012). Translational biomarker discovery in clinical metabolomics: an introduction tutorial. Metabolimics, 9(2), 280-299. DOI:10.1007/s11306-012-0482-9

[4] Johnstone, I.M., Titterington, D.M. (2009). Statistical challenges of high-dimensional data. The Royal Society,376(2110), 4237-4253. DOI: 10.1098/rsta.2009.0159

[5] Wang, C., Gevertz, J.L. (2016). Finding causative genes from high-dimensional data: an appraisal of statistical and machine learning approaches. US National Library of Medicine National Institutes of Health, PubMed,15(4),321-47. DOI: 10.1515/sagmb-2015-0072

[6] Chandrashekar, G., Sahin, F. (2013). A survey on feature selection methods. Computers and Electrical Engineering,40(1), 16-28. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024

[7] Bylesjö, M. (2015). Extracting Meaningful Information from Metabonomic Data Using Multivariate Statistics. Mayer B. (eds) Bioinformatics for Omics Data. Methods in Molecular Biology (Methods and Protocols),1277, 137-146. DOI: 10.1007/978-1-4939-2377-9-11

[8] Miao, J., Niu, L. (2016). A Survey on Feature Selection.Procedia Computer Science,91, 919-926. DOI: 10.1016/j.procs.2016.07.111

[9] Dunkler D., Sánchez-Cabo F., Heinze G. (2011) Statistical Analysis Principles for Omics Data. Mayer B. (eds) Bioinformatics for Omics Data.Methods in Molecular Biology (Methods and Protocols), 719, 113-131. DOI: 10.1007/978-1-61779-027-0-5

[10] Guyon, I., Weston, J., Barnhill, S. et al. Machine Learning (2002) 46: 389. https://doi.org/10.1023/A:1012487302797

[11] Jacobson, R. (2013). 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?. EUA. IBM. Recuperado de: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/ en diciembre 2017.

[12] Wang. J, Zhou, S., et al. (2014). An Improved Feature Selection Based on E ective Range for Classi cation. The Scienti cWorld Journal. Recuperado el 11/08/2016 de: https://dx.doi.org/10.115/2014/972125

[13] James, G., Witten, D., Hastie, T. & Tibshirani, R. (2015). An Introduction to Statistical Learning (1). Springer, New York, NY. DOI: https://doi.org/10.1007/978-1-4614-7138-7

[14] The UN Secretary-General's Independent Expert Advisory Group. (2014). A World that Counts. Mobilising the Data Revolution for Sustainable Development. Recuperado desde: https://www.undatarevolution.org/report/

[15] Desjardins, J. (2017). What Happens in a Internet Minute in 2017?. Canadá. Visual Capitalist. Recuperado de: http://www.visualcapitalist.com/happens-internet-minute-2017/ en diciembre 2017.

[16] Taylor, C. (2017). Structured vs. Ustructured Data. EUA. Datamation. Recuperado de https://www.datamation.com/big-data/structured-vs-unstructured-data.html en diciembre 2017.

[17] Sánchez Turcios, R.A. (2015). t-Student. Usos y abusos. Revista Mexicana de Cardiología, Vol. 26 (1), p. 59-61.

[18] Metabolomics. (2018). En Oxford Dictionary. Recuperado de: https://en.oxforddictionaries.com/de nition/metabolomics

[19] Ramírez, Leticia. (2018). ¾Qué es Big Data?, retos y oportunidades. Notas del seminario. Seminario ¾Qué es...?. CIMAT. 24 agosto 2018. Recuperado de: https://www.youtube.com/watch?v=jyZU4Vl-c5g&feature=youtu.be

Anexos

Código para base de datos de colon

```
library(MASS)
library(boot)
library(parallel)
library(scatterplot3d)
library(plsgenomics)
data(Colon)
N<-dim(Colon$X)[1]
d<-dim(Colon$X)[2]
gamma<-1.732
c<-Colon$Y
C<-c("1","2")
L<-length(unique(c))
L==length(C)
#PROBABILIDADES
p<-matrix(0,1,L)
for(z in 1:L){
    y=0
    for(k in 1:N)
        if(c[k]==C[z])y<-y+1
    p[z]<-y
}
p<-p/N
#MU
mu<-matrix(0,d,L)
for(caract in 1:d){
    for(k in 1:L){
        s=0
        for(i in 1:N)
            if(c[i]==C[k])s<-s+Colon$X[i,caract]
    mu[caract,k]<-s/summary(as.factor(c))[k]
    }
}
```

```
#DESVIASIÓN ESTÁNDAR
sd<-matrix(0,d,L)
for(caract in 1:d){
    for(k in 1:length(C)){
        s=0
        for(i in 1:N)

           if(c[i]==C[k])s<-s+(Colon$X[i,caract]-mu[caract,k])^2
        sd[caract,k]<-sqrt(s/(summary(as.factor(c))[k]-1))
        }
}
#AMPLITUDES
Rmenos <- matrix(0,d,L)
Rmas <- matrix(0,d,L)
for(j in 1:L){
    for(i in 1:d){
        Rmenos[i,j]=mu[i,j]-(1-p[j])*gamma*sd[i,j]
        Rmas[i,j]=mu[i,j]+(1-p[j])*gamma*sd[i,j]
    }
}
#ORDEN
for(i in 1:d){
    Rmenos[i,]<-sort(Rmenos[i,])
    Rmas[i,]<-sort(Rmas[i,])
}
#FUNCIÓN PHI
phi<-function(i,j,k){
    if(Rmas[i,j]>Rmenos[i,k])rphi<-Rmas[i,j]-Rmenos[i,k] else rphi<-0
return(rphi)}
#ÁREA DE TRASLAPE
OA<-matrix(0,d,1)
t=1
for(i in 1:d){
```

```
    oa<-matrix(0,L,L)
    for(j in 1:(L-1)){
        for(k in (j+1):L){
            oa[j,k]<-phi(i,j,k) t=t+1
        }
    }
    OA[i,]<-sum(oa)
}
#COEFICIENTE DE ÁREA
AC<-matrix(0,d,1)
Max<-matrix(0,d,1)
min<-matrix(0,d,1)
for(i in 1:d){
    Max[i]<-max(Rmas[i,])
    min[i]<-min(Rmenos[i,])
    AC[i]<-OA[i]/(Max[i]-min[i])
}
#COEFICIENTE DE ÁREA NORMALIZADO
NAC<-matrix(0,d,1)
for(i in 1:d){
    NAC[i]<-AC[i]/max(AC)
}
#CALCULAR LOS PESOS
w<-matrix(0,d,1)
for(i in 1:d){
    w[i]<-1-NAC[i]
}
W<-sort(w,decreasing = TRUE,index.return=TRUE)
W$x[1,10]
Colon$gene.names[W$ix[1:10]]
```

Código para base de datos MLL

```
MLLread=read.csv("C:\\MLLdata.csv")
attach(MLLread)
```

```
dim(MLLread)
c<-as.character(name)
MLL<-MLLread[,-1]
dim(MLL)
N <- dim(MLL)[1]
d <- dim(MLL)[2]
gamma <- 1.732
C <- unique(c)
L <- length(unique(c))
#PROBABILIDADES
p<-matrix(0,1,L)
for(z in 1:L){

    y=0 for(k in 1:N)
    if(c[k]==C[z])y<-y+1 p[z]<-y

}
p<-p/N
#MU
muu<-matrix(0,d,L)
for(v in 1:d){

    muu[v,1]<-mean(MLL[1:24,v])
    muu[v,2]<-mean(MLL[25:44,v])
    muu[v,3]<-mean(MLL[45:N,v])

}
#DESVIASIÓN ESTÁNDAR
sdd<-matrix(0,d,L)
for(v in 1:d){

    sdd[v,1]<-sqrt(var(MLL[1:24,v]))
    sdd[v,2]<-sqrt(var(MLL[25:44,v]))
    sdd[v,3]<-sqrt(var(MLL[45:N,v]))

}
#AMPLITUDES
Rmenos <- matrix(0,d,L)
Rmas <- matrix(0,d,L)
gamma<-1.732
for(j in 1:L){
```

```
    for(i in 1:d){
        Rmenos[i,j]=muu[i,j]-(1-p[j])*gamma*sdd[i,j]
        Rmas[i,j]=muu[i,j]+(1-p[j])*gamma*sdd[i,j]
    }
}
#ORDEN
for(i in 1:d){
    Rmenos[i,]<-sort(Rmenos[i,])
    Rmas[i,]<-sort(Rmas[i,])
}
#FUNCIÓN PHI
phi<-function(i,j,k,Rmenos,Rmas){
    if(Rmas[i,j]>Rmenos[i,k])rphi<-Rmas[i,j]-Rmenos[i,k] else rphi<-0
return(rphi)}
#ÁREA DE TRASLAPE
OA<-matrix(0,d,1)
t=1
for(i in 1:d){
    oa<-matrix(0,L,L)
    for(j in 1:(L-1)){
        for(k in (j+1):L){
            oa[j,k]<-phi(i,j,k,Rmenos,Rmas) t=t+1
        }
    }
    OA[i,]<-sum(oa)
}
#COEFICIENTE DE ÁREA
AC<-matrix(0,d,1)
Max<-matrix(0,d,1)
min<-matrix(0,d,1)
for(i in 1:d){
    Max[i]<-max(Rmas[i,])
    min[i]<-min(Rmenos[i,])
    AC[i]<-OA[i]/(Max[i]-min[i])
```

```
}
#COEFICIENTE DE ÁREA NORMALIZADO }
NAC<-matrix(0,d,1)
for(i in 1:d){
    NAC[i]<-AC[i]/max(AC)
}
#CALCULAR LOS PESOS
w<-matrix(0,d,1)
for(i in 1:d){
    w[i]<-1-NAC[i]
}
W<-sort(w,decreasing = TRUE,index.return=TRUE)
W$x[1:10]
W$ix[1:10]
colnames(MLL)[W$ix[1:10]]
```

Código para la base de datos de metabolitos de mujeres

```
MMread=read.csv("C:\\MetabolitosMujeres.csv")
attach(MMread)
MMread<-MMread[,-1]
varef<-MMread[,1]
MM<-MMread[,-1]
N <- dim(MM)[1]
d <- dim(MM)[2]
gamma <- 1.732
L<-length(unique(varef))
#PROBABILIDADES
p<-rep(0,L)
for(z in 1:L){
    p[z]<-(summary(varef)/N)[z]
}
p
#MU
mu<-matrix(0,d,L)
for(v in 1:d){
```

```
    mu[v,1]<-mean(MM[1:10,v])
    mu[v,2]<-mean(MM[11:20,v])
    mu[v,3]<-mean(MM[21:28,v])
    mu[v,4]<-mean(MM[29:37,v])
    mu[v,5]<-mean(MM[38:N,v])
}
#DESVIASIÓN ESTÁNDAR
sdd<-matrix(0,d,L)
for(v in 1:d){
    sdd[v,1]<-sqrt(var(MM[1:10,v]))
    sdd[v,2]<-sqrt(var(MM[11:20,v]))
    sdd[v,3]<-sqrt(var(MM[21:28,v]))
    sdd[v,4]<-sqrt(var(MM[29:37,v]))
    sdd[v,5]<-sqrt(var(MM[38:N,v]))
}
#AMPLITUDES
Rmenos <- matrix(0,d,L)
Rmas <- matrix(0,d,L)
gamma<-1.732
for(j in 1:L){
    for(i in 1:d){
        Rmenos[i,j]=mu[i,j]-(1-p[j])*gamma*sdd[i,j]
        Rmas[i,j]=mu[i,j]+(1-p[j])*gamma*sdd[i,j]
    }
}
#ORDEN
for(i in 1:d){
    Rmenos[i,]<-sort(Rmenos[i,])
    Rmas[i,]<-sort(Rmas[i,])
}
#FUNCIÓN PHI
phi<-function(i,j,k,Rmenos,Rmas){
    if(Rmas[i,j]>Rmenos[i,k])
        rphi<-Rmas[i,j]-Rmenos[i,k] else rphi<-0 return(rphi)
```

```
}
#ÁREA DE TRASLAPE
OA<-matrix(0,d,1)
t=1
for(i in 1:d){

    oa<-matrix(0,L,L)
    for(j in 1:(L-1)){
        for(k in (j+1):L){
            oa[j,k]<-phi(i,j,k,Rmenos,Rmas)
            t=t+1
        }
    }
    OA[i,]<-sum(oa)

}
#COEFICIENTE DE ÁREA
AC<-matrix(0,d,1)
Max<-matrix(0,d,1)
min<-matrix(0,d,1)
for(i in 1:d){

    Max[i]<-max(Rmas[i,])
    min[i]<-min(Rmenos[i,])
    AC[i]<-OA[i]/(Max[i]-min[i])

}
#COEFICIENTE DE ÁREA NORMALIZADO
NAC<-matrix(0,d,1)
for(i in 1:d){

    NAC[i]<-AC[i]/max(AC)

}
#CALCULAR LOS PESOS
w<-matrix(0,d,1)
for(i in 1:d){

    w[i]<-1-NAC[i]

}
W<-sort(w,decreasing = TRUE,index.return=TRUE)
```

```
W$x[1:10]
W$ix[1:10]
colnames(MM)[W$ix[1:10]]
```

Código sobre la simulación

```
N<-62
d<-2000
M<-matrix(runif(N*d,0,10),N,d)
pos<-sample(d,3)
#pos<-sample(d,4) #para fB2
X1<-M[,pos[1]]
X2p<-M[,pos[2]]
X3p<-M[,pos[3]]
Z2<-2*X2p
Z3<-3*X3p
X2<-X1+0.35*Z2
X3<-X2+0.35*Z3
X4<-2/3*X1+2/3*X2+1/3*rnorm(1,0,1)
mu1<-mean(X1)
mu2<-mean(X2)
mu3<-mean(X3)
mu2p<-mean(X2p)
mu3p<-mean(X3p)
cA1<-88.4
cA2<-95
cA3<-4.9
cB1<-698
cB2<-6000
cB31<-10.5
cB32<-17
fA1<-as.numeric(2*X1+3*X2+4*X3+rnorm(1,0,1) > cA1)
fA2<-as.numeric(X1^2+X2^2+X3 > cA2)
fA3<-as.numeric(((X1-mu1)^2+(X2-mu2)+(X3-mu3)) > cA3)
fB1<-as.numeric(X1+X2+X3+X1*X2+X1*X3+X2*X3+X1*X2*X3 > cB1)
fB2<-as.numeric(X1*X2*X3*X4 > cB2)
fB3<-as.numeric((X1*X2 > cB31) & (X3 < cB32))
#FUNCIÓN
varef<-fA3
```

```
MM<-matrix(NaN,N,d)
cc=1
for(ill in 0:1)
    for(j in 1:N)
        if(varef[j] == ill){
           MM[cc,]<-M[j,]
           cc=cc+1
        }
#MM
varef<-sort(varef)
L<-length(unique(varef))
#PROBABILIDADES
p<-rep(0,L)
p[1]<-(length(varef)-sum(varef))/N
p[2]<-sum(varef)/N
#MU
mu<-matrix(0,d,L)
for(v in 1:d){
    mu[v,1]<-mean(MM[1:(length(varef)-sum(varef)),v])
    mu[v,2]<-mean(MM[(length(varef)-sum(varef)+1):N,v])
}
#DESVIASIÓN ESTÁNDAR
sd<-matrix(0,d,L)
for(v in 1:d){
    sd[v,1]<-sqrt(var(MM[1:(length(varef)-sum(varef)),v]))
    sd[v,2]<-sqrt(var(MM[(length(varef)-sum(varef)+1):N,v]))
}
#AMPLITUDES
Rmenos <- matrix(0,d,L)
Rmas <- matrix(0,d,L)
gamma<-1.732
for(j in 1:L){
    for(i in 1:d){
        Rmenos[i,j]=mu[i,j]-(1-p[j])*gamma*sd[i,j]
        Rmas[i,j]=mu[i,j]+(1-p[j])*gamma*sd[i,j]
```

```
    }
}
#ORDEN
for(i in 1:d){
    Rmenos[i,]<-sort(Rmenos[i,])
    Rmas[i,]<-sort(Rmas[i,])
}
#FUNCIÓN PHI
phi<-function(i,j,k,Rmenos,Rmas){
    if(Rmas[i,j]>Rmenos[i,k])rphi<-Rmas[i,j]-Rmenos[i,k] else rphi<-0
return(rphi)}
#ÁREA DE TRASLAPE
OA<-matrix(0,d,1)
t=1
for(i in 1:d){
    oa<-matrix(0,L,L)
    for(j in 1:(L-1)){
        for(k in (j+1):L){
            oa[j,k]<-phi(i,j,k,Rmenos,Rmas)
            t=t+1
        }
    }
    OA[i,]<-sum(oa)
}
#COEFICIENTE DE ÁREA
AC<-matrix(0,d,1)
Max<-matrix(0,d,1)
min<-matrix(0,d,1)
for(i in 1:d){
    Max[i]<-max(Rmas[i,])
    min[i]<-min(Rmenos[i,])
    AC[i]<-OA[i]/(Max[i]-min[i])
}
#COEFICIENTE DE ÁREA NORMALIZADO
```

```
NAC<-matrix(0,d,1)
for(i in 1:d){
    NAC[i]<-AC[i]/max(AC)
}
#CALCULAR LOS PESOS
w<-matrix(0,d,1)
for(i in 1:d){
    w[i]<-1-NAC[i]
}
W<-sort(w,decreasing = TRUE,index.return=TRUE)
W$x[1:10]
W$ix[1:30]
pos
t<-rep(0,length(pos))
for(h in 1:length(pos))
    for(r in 1:d)
        if(W$ix[r]==pos[h])
            t[h]<-r
t
```