Estimaciones multivariadas robustas para el COVID-19

Isabella Arango Restrepo
iarangor1@eafit.edu.co

Santiago Isaza Cadavid
sisazac@eafit.edu.co

Nicolás Rengifo Campo
nrengifoc@eafit.edu.co

Juan David Rengifo Castro
jdrengifoc@eafit.edu.co

2 de febrero de 2021

Resumen

El propósito de este trabajo es realizar un análisis estadístico multivariante a varios países afectados por el coronavirus. Utilizando variables como la capacidad médica, la mortalidad, el número de pruebas realizadas, las inversiones económicas entre otras. Con estas variables realizar modelos de regresión robusta y no paramétricos para buscar relaciones significativas entre ellas. De igual manera, análisis de componentes principales robusto y por último métodos de clasificación sobre los países para determinar si estaba preparado o no.

1. Introducción

Debido al alto grado de incertidumbre para evaluar la magnitud del Covid-19 en un país, se propone una estrategia estadística para generar indicadores, que permitan evaluar y representar la situación actual de esta problemática, a partir de las variables que presenten una menor incertidumbre, tales como la mortalidad, proporción de ocupación de la capacidad total del sistema de salud, entre otras. La rápida propagación del virus ha llevado a los gobiernos de todos los países, a enfrentarse a un reto sin precedentes, en el cual la mayoría de las variables más relevantes son inobservables. Esto ha llevado a la toma de decisiones sin fundamentos tangibles, no por negligencia, sino por la celeridad del este fenómeno y la inexistencia de métodos para evaluar la situación real. Teniendo en cuenta que estas decisiones tienen grandes efectos sobre sectores como el social, económico y salud, es pertinente y necesario, la construcción de herramientas que permitan evaluar de manera correcta la situación real, para así tomar decisiones que mitiguen los efectos en los sectores mencionados anteriormente.

Cabe resaltar que para realizar esta investigación se construyó una base de datos propia compuesta por múltiples fuentes reconocidas en el ambito internacional, tales como la CIA, la Unión Europea, World Meter, Our Wolrd in Data, IMF y el GHS Index. A partir de las fuentes mencionadas obtuvimos las siguientes variables: casos confirmados por 1000 habitantes, muertes confirmados por 100 habitantes, mortalidad confirmada, días desde el primer caso, médicos por 1000 habitantes, camas hospitalarias por 1000 habitantes, pruebas procesadas acumuladas, densidad poblacional, edad mediana, proporción urbana, PIB para el año 2019, el porcentaje del PIB invertido en salud y el índice GHS, el cual nos dice si un país esta bien preparado o no para una epidemia. Cabe resaltar, que todas estas variables, menos el PIB fueron construidas de manera que fueran comparables par a par para cada país, esto es, que sean proporcionales a alguna referencia. De esta manera, tenemos en su mayoría proporciones que evita el problema de múltiples unidades. Las variables mencionadas están ordenadas tal y como se mostraran en secciones posteriores.

2. Objetivos

- Proporcionar una mejor idea de la problemática por covid-19 para diversos países a partir de análisis estadísticos robustos y no paramétricos, sobre las variables de interés.

- Verificar que a partir de las variables de estudio sea posible realizar una clasificación veraz de la capacidad de un país para afrontar la crisis utilizando métodos de clasificación supervisada y no supervisada. La rápida propagación del nuevo virus que ataca a la humanidad ha llevado a los gobiernos de todos los países, a enfrentarse a un reto sin precedentes, en el cual la mayoría de las variables más relevantes son inobservables. Esto ha llevado a la toma de decisiones sin fundamentos tangibles, no por negligencia, sino por la celeridad del este fenómeno y la inexistencia de métodos para evaluar la situación real. Teniendo en cuenta que estas decisiones tienen grandes efectos sobre sectores como el social, económico y salud, es pertinente y necesario, la construcción de herramientas que permitan evaluar de manera correcta la situación real, para así tomar decisiones que mitiguen los efectos en los sectores mencionados anteriormente.

3. Modelos estadísticos

Los modelos tradicionales encontrados en [7] facilitan los primeros acercamientos para el análisis de datos multivariados, sin embargo, dada la naturaleza de los datos a analizar, una técnica tradicional puede no resultar eficiente dado que estos modelos imponen supuestos que no siempre son válidos y que pueden llevar a resultados de poca confiabilidad. Por su parte [9] y [3] ofrecen diferentes formas de hallar estructuras estadísticas como la matriz de covarianzas, a partir de coeficientes de correlación robusta como el de Spearman, además de reemplazar las desviaciones estándar involucradas por MAD, por sus siglas en inglés Median Absolute Deviation (Desviación Absoluta Media); con la finalidad de obtener robustez en los modelos, pues la presencia de datos atípicos es inevitable, por lo que se requieren modelos que toleren outliers sin influir en gran medida en los resultados finales.

3.1. Detección de outliers

Tomando lo propuesto en [2] se hace una detección de outliers o datos atípicos a partir de la distancia de Mahalanobis robusta, haciendo un ajuste en la matriz de covarianzas.

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n} \tag{1}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

$$\rho_k(i,j) = \frac{2 * (N_c - N_d)}{N(N-1)} \tag{3}$$

$$\rho_s(i,j) = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \tag{4}$$

$$Cov(X,Y) = \rho(X,Y)\sigma_X \sigma_Y, \tag{5}$$

Y la distancia de Mahalanobis entre un punto y su vector de medias.

$$D^2(x_i, \bar{x}) = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \tag{6}$$

Cambiando el coeficiente de correlación ρ de Pearson, por el de Spearman se obtienen resultados más robustos, así mismo, como cambiar el cálculo de las desviaciones estándares σ por MAD. Se clasifican como outliers aquellos registros que posean una distancia mayor en comparación que las demas; se puede tomar los k peores, o recortando por percentiles, o si se sabe la distribución de las distancias con un valor critico.

3.2. Regresión

3.2.1. Regresión Multivariada

Para las regresiones robustas se proponen variaciones, a partir de:

- Redefinición de la matriz de covarianza para hallar el vector de coeficientes que multiplican a las variables en el modelo de regresión lineal.
- Dado que la operación del cálculo de la inversa de una matriz puede ser erróneo computacionalmente usando la pseudoinversa de Moore-Penrose se propone una forma alternativa para hallar el vector de coeficientes que multiplican a las variables en el modelo de regresión lineal.
- Para análisis bivariados usaremos regresión no paramétrica de Nadaraya-Watson, que a diferencia del modelo lineal no hace ningún supuesto sobre el tipo de dependencia entre las variables.

$$\hat{\beta} = S_{XX}^{-1} S_{XY} \tag{7}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}^T \bar{X} \tag{8}$$

Así, nuestro modelo multivariado queda de la forma:

$$Y = \hat{\beta}^T X + \beta_0 \tag{9}$$

La forma estándar en la que se plantea el modelo de regresión lineal, esta dada de igual manera que en 7, sin embargo el vector de β se haya minimizando mínimos cuadrados y su forma esta dada por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{10}$$

3.2.2. Regresión Bivariada

Para la regresión bivariada lineal tradicional, se hereda la forma de calcular β como en 10, igual que el β_0 como en 8. Sin embargo, como se menciono anteriormente, dada la naturaleza de los datos, el modelo lineal estándar no es el método mas adecuado para realizar el análisis estadístico; en su contra parte se propone una regresión no paramétrica ponderada, Nadaraya-Watson, que se define a partir de: valor esperado $E(.)$, probabilidad condicional $P(A|B = b)$, ancho de banda h_n y un kernel para realizar la ponderación de los valores en el modelo de regresión, se utilizarán aquellos propuestos en [8]. La deducción del modelo se muestra a continuación:

$$\begin{aligned}
\hat{E}(y|x) &= \hat{g}(x) \\
&= \frac{\int y\hat{p}(x,y)dy}{\hat{p}(x)} \\
&= \frac{\frac{1}{nh_n}\sum_{i=1}^{n} y_i K(\frac{x_i - x}{h})}{\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \\
&= \sum_{i=1}^{n} y_i w_i(x)
\end{aligned} \tag{11}$$

$$\text{Donde } w_i(x) = \frac{K(\frac{x_i - x}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \text{ y } h = \hat{\sigma}_x n^{\frac{-1}{5}}$$

A continuación los kernel propuestos:

$$K(u) = \frac{1}{2} I_{[-1,1]}(u)$$

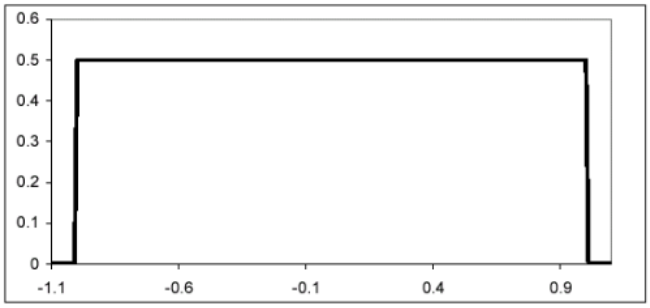
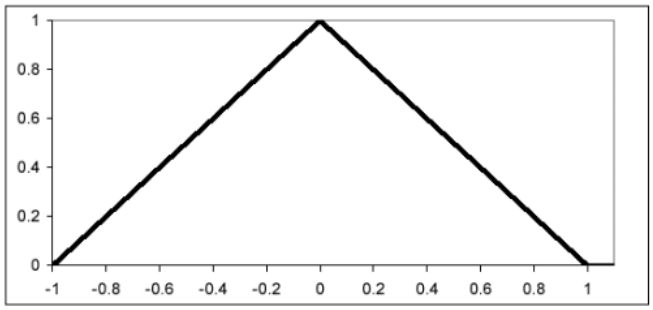


Figura 1: Kernel uniforme

$$K(u) = (1 - |u|)I_{[-1,1]}(u)$$



8u

Figura 2: Kernel triangular

$$K(u) = \frac{3}{4}(1 - u^2)I_{[-1,1]}(u)$$

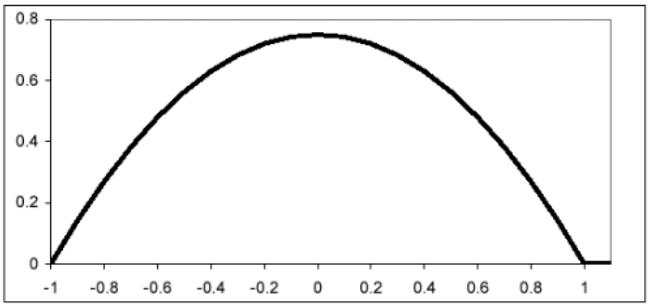


Figura 3: Kernel Epanechnikov

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

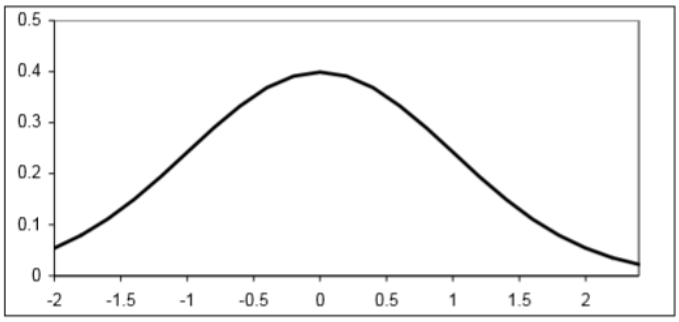


Figura 4: Kernel Gaussiano

3.3. Componentes principales

Para el análisis de los componentes principales partimos de la variación del cálculo de la matriz de covarianzas para hallar el primer componente principal, el cual usaremos como primera aproximación para determinar qué países están mejor o peor, teniendo en cuenta que los datos ya se han adimensionalizado con respecto a la población por país, para así poder realizar comparaciones de manera objetiva.
El primer componente se define como la combinación lineal de las variables que tienen máxima varianza.[7] Los valores del primer componente para cada observación o individuo se representa por un vector z_1, dado por $z_1 = Xa_1$. De igual manera su varianza es:

$$\frac{1}{n}z_1'z_1 = \frac{1}{n}a_1'X'Xa_1 = a_1'Sa1 \tag{12}$$

Donde S es la matriz de covarianza de los datos, en este caso, la matriz de covarianza robusta. Finalmente, si se desea maximizar esta varianza, se resuelve el problema mediante el multiplicador de Lagrange. Cuya solución es:

$$Sa_1 = \lambda a_1 \tag{13}$$

Lo que implica que a_1 es un vector propio de la matriz de covarianza robusta S y λ el valor propio correspondiente. Para determinar qué valor propio es la solución de 14 se multiplica por a_1', obteniendo la siguiente ecuación:

$$a_1'Sa_1 = \lambda a_1'a_1 = \lambda \tag{14}$$

A partir de 12 es lógico concluir que λ es la varianza de z_1, y a su vez, el mayor valor propio de la matriz S. [7]

3.4. Clasificaciones

Con la clasificación se pretende etiquetar a cada país como preparado o no para enfrentar la crisis, en base a un indicador encontrado en [4], en donde se clasifican de manera binaria los países. Se usan los métodos de clasificación provistos en [1] así como la regresión logística y máquinas de vector soporte para determinar si a partir de nuestro conjunto de datos se llega a esta clasificación.

3.4.1. Métodos de clasificación no supervisados

- **K-Medias**

K-medias es un algoritmo que consiste en dividir los datos en G grupos diferentes según sus características, siempre con el objetivo de minimizar la suma de las distancias entre los individuos y un dato escogido de forma aleatoria conocido como centroide. En pocas palabras, este método selecciona G datos de forma aleatoria y cada uno de ellos representa el centro de un grupo, posterior a esto, observa uno a uno los datos restantes, calcula la distancia a cada centroide y lo agrupa junto aquel que presente la menor distancia. Después de tener todos los datos agrupados, se recalcularán los centros a partir del promedio de los datos dentro del grupo analizado y nuevamente se procederá a clasificar cada individuo. [7] Este procedimiento se hará de forma iterativa hasta que la variabilidad intragrupo ya no pueda reducirse más, por lo cual resulta ser un problema de optimización considerando lo siguiente.

$$SCDG = \sum_{g=1}^{G} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \bar{x_{jg}})^2 \tag{15}$$

Por otro lado, es recomendable utilizar la siguiente regla empírica de Hartigan de forma iterativa con el objetivo de conocer cuál es el número de grupos necesarios, donde la condición de parada será cuando con este estadístico se obtenga un valor menor a 10 y en caso contrario incrementar un grupo más. [7]

$$F = \frac{SCDG(G) - SCDG(G+1)}{SCDG(G+1)/(n-G-1)} \tag{16}$$

- **K-Medoides**

K-Medoides es un algoritmo relacionado con k-medias, por lo cual su objetivo también es dividir el conjunto de datos en G grupos. En este nuevamente se escogen G centroides de forma aleatoria y se calcula la distancia de cada dato a estos, agrupándose con el centroide que se encuentre más cerca. Después de tener todos los datos agrupados, los nuevos centros se recalculan como el dato mediano de cada grupo y se realizará este procedimiento de forma iterativa hasta lograr que la variabilidad intragrupo no pueda reducirse más y se emplearán los mismos criterios explicados en k-medias.

3.4.2. Métodos de clasificación supervisados

Los métodos de clasificación supervisada son aquellos en los que es necesario dividir el conjunto de datos en dos grupos con el fin de entrenar el modelo y finalmente testearlo. Para esto se puede considerar el 70 % de los registros para entrenamiento y el 30 % restante para testeo, sin embargo, esta elección debe hacerse de forma aleatoria para evitar sesgos en una de las clases y para saber que estos conjuntos fueron seleccionados de forma correcta debe cumplirse que el accuracy o score de training sea mayor que el testing.

- **Discriminante Lineal**

El método de discriminante lineal tiene como objetivo encontrar una relación lineal entre un conjunto de variables categóricas y una variable respuesta de tipo binaria, además de tratar de definir una regla de decisión que permita predecir la clasificación de nuevos datos. Para este algoritmo es de gran importancia que las variables explicativas no se encuentren correlacionadas con las demás, pues podría producir sesgos hacia una clase. Por otro lado, esta estima la probabilidad de que un dato pertenezca a una de las clases posibles dependiendo de cual presente mayor probabilidad. En pocas palabras, este método para poder clasificar un nuevo dato calcula la distancia de este a la media de cada grupo y lo clasifica en el que presente la menor distancia.

- **Discriminante Cuadrático**

El método de discriminante cuadrático, también conocido como QDA por sus siglas en inglés (Quadratic Discriminat Analysis), es una variante del algoritmo de discriminante lineal, pero en este caso la regla de clasificación es la mínima distancia al cuadrado de un nuevo dato a las medias de cada uno de los grupos de clasificación. Sin embargo, también puede pensarse como una probabilidad de pertenencia a una de las dos clases como se mencionó en el método anterior y nuevamente la clasificación se basará en la máxima probabilidad [6].

- **K vecinos más cercanos**

Este método, normalmente conocido como KNN por sus siglas en inglés (k-nearest neighbors) es un método muy simple, pero con mucha precisión en la mayoría de los casos. Este se basa en la teoría de que las cosas similares siempre se encuentran cerca, es decir que clasifica los datos basados en una medida de similitud de todos los casos posibles, para esto utiliza una técnica de clasificación que considera los votos mayoritarios de los k vecinos alrededor de un punto y según las clases a los que estos pertenezcan el nuevo dato será clasificado. Sin embargo, una desventaja de este algoritmo es que es muy sensible a la métrica escogida para determinar la distancia de un punto a los demás, pues de esto dependerá quienes son los k vecinos más cercanos. Además,

una buena clasificación también dependerá del valor de k escogido inicialmente y al no existir una fórmula exacta para la obtención de este valor, muchas veces se trata de un caso de fallo y ensayo, pero se ha de tener en cuenta que k debe ser un número impar para evitar empates durante la votación.

- **Máquina de vector soporte**

El método máquina de vector soporte, normalmente conocido como SVM (Support Vector Machine) por sus siglas en inglés consiste en la búsqueda del hiperplano que permita dividir de la mejor manera al conjunto de datos en los posibles grupos de clasificación, para esto tiene en cuenta un conjunto de variables explicativas y una variable respuesta de tipo binaria. Para poder comprender de mejor forma el método, es importante conocer la definición de vector soporte e hiperplano, donde esta última puede ser pensado como una línea recta (en el caso bivariado) o un plano (caso multivariante) que permite dividir el conjunto de datos en dos grupos diferentes como es en el caso binario de tal forma que se obtenga el mayor margen posible entre los datos más cercanos a ambos lados del hiperplano y estos datos son los conocidos como los vectores soporte y después de establecer el hiperplano, es fácil reconocer las dos clases de clasificación. Sin embargo, en el caso en el que el conjunto de datos no es separable de una forma fácil, se introduce el concepto de kernel ya explicados anteriormente.

- **Regresión Logística**

La regresión logística es el método de clasificación más utilizado actualmente debido a que no requiere de muchos recursos computacionales. Este algoritmo permite estimar la relación existente entre un conjunto de variables explicativas y una variable respuesta de carácter binario, sin embargo, la relación se establece en términos de probabilidades de pertenencia a una clase u otra. Para facilitar la comprensión del método, vamos a dividir este en dos partes. La primera de ellas es considerarlo como un método lineal en el que es necesario estimar los coeficientes β dados por la función logística que puede expresarse de la siguiente manera [5]

$$\text{logit} \Rightarrow ln\left[\frac{p(x)}{1-p(x)}\right] = f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Figura 5: Función Logística

Sin embargo, para poder obtener una buena clasificación, es de gran importancia eliminar las posibles correlaciones entre las variables, pues en caso contrario, esto puede generar sesgos para una de las clases. Por esta razón, se eliminarán de manera progresiva aquellas variables que presenten el mayor p-valor hasta que todos estos sean menores a 0.05 y obtener así un modelo adecuado. En este momento, se comienza con la segunda fase del algoritmo, la cual consiste en la aplicación de la inversa de la función logística descrita por la siguiente fórmula

$$p(x) = \frac{e^{f(x)}}{1+e^{f(x)}} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} = \frac{1}{1+e^{-\beta_0-\sum_{i=1}^{k}\beta_i x_i}}$$

Figura 6: Inversa de la función logística

De esta forma se obtiene un comportamiento de una función sigmoide que varía entre cero y uno indicando una probabilidad, la cual será útil al momento de clasificar. Finalmente, para convertir las probabilidades en una predicción binaria al igual que la variable respuesta se considera un punto de corte, normalmente, de 0.5, y si las predicciones obtenidas son mayores a este valor pertenecerán a la clase que toma el valor de uno y en caso contrario a la clase de valor cero.

4. Resultados

4.1. Análisis Exploratorio

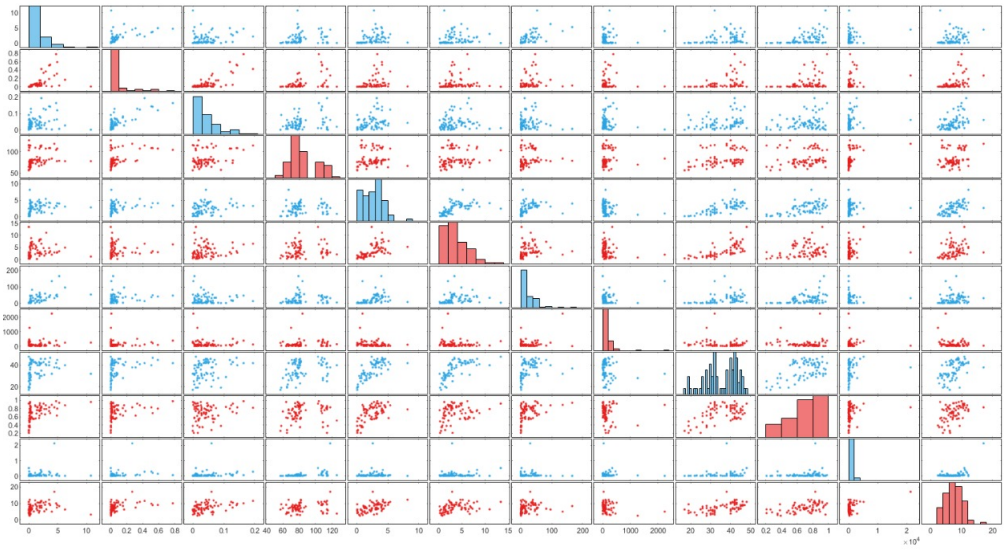


Figura 7: Plotmatrix de los datos

Se realizó el plot matrix para la base datos donde se logran evidenciar algunas relaciones lineales entre variables par a par y aunque no se muestra este comportamiento en todos los casos, es posible que en el caso multivariado las variables se relaciones de manera lineal. Además se puede observar para la cuarta variable existen dos poblaciones, recordemos que esta variable corresponde al número de días que llevaban los paises desde el primer caso confirmado hasta la fecha, esto tiene mucho sentido, pues tanto asía como europa llevan más tiempo que el resto del mundo. También se hace clara la presencia de múltiples datos atípicos, no solo el magnitud sino en comportamiento, por lo cual resulta racional emplear la distancia de Mahalanobis para eliminar estos datos atípicos en comportamiento que evitan la existencia de linealidad entre las variables par a par.

4.2. Detección de Outliers

La detección de datos atípicos se realizó para dos conjuntos de datos, el primero que incluye la variable categórica GHS index y el segundo que la excluye. Para medir las distancias entre cada observación (país) se implementó la distancia de Mahalanobis, pues esta presenta un buen comportamiento en el caso multivariado. Adicional a esto se realizó este mismo calculo de una manera robusta, el cual se logra al emplear la matriz de covarianza robusta previamente definida. Los resultados de los 14 países con mayor distancia se adjunto a continuación.

Con Categóricas			
Tradicional		Robusta	
Países	Distancia	Países	Distancia
Estados Unidos	67.05	Bélgica	30915.18
Qatar	54.87	Estados Unidos	20174.17
Baréin	54.23	España	14535.22
Islandia	34.23	Italia	11121.61
Japón	32.72	Reino Unido	10191.96
Bélgica	32.14	Francia	6446.09
Cuba	30.80	Qatar	6361.52
Zimbabwe	25.03	Suecia	4602.13
Bangladesh	22.82	Países Bajos	3595.08
Tailandia	21.21	Baréin	2447.95
Francia	19.73	Irlanda	2183.28
Nepal	19.37	Islandia	1384.18
España	18.04	Japón	1316.21
Bielorrusia	16.90	Bielorrusia	684.71

En la tabla de resultados que incluye las variables categóricas se puede observar que Estados Unidos es frecuente en ambos resultados. Esto en parte, debido a que EE.UU. a pesar de poseer el PIB más alto a nivel mundial, al igual que el mayor porcentaje de este invertido en salud, tiene una proporción considerable, del número de casos y muertes por 1000 habitantes. Esto gracias a las fútiles políticas de contingecia, efectuadas por el gobierno. Cabe resaltar que este dato atípico no corresponde a la magnitud de los casos y muertes, pues se espera que un país tenga muchos casos y muertos si no invierte en salud, lo que lo hace atípico, gracias a la distancia de Mahalanobis, es este comportamiento paradójico, alta inversión y alta proporción de la población infectada.

Sin Categóricas			
Tradicional		Robusta	
Países	Distancia	Países	Distancia
Estados Unidos	66.93	Bélgica	26192.66
Baréin	54.11	Estados Unidos	19710.98
Qatar	52.62	España	12194.02
Islandia	33.42	Italia	9587.83
Bélgica	31.03	Reino Unido	8884.00
Japón	30.54	Qatar	6250.10
Cuba	29.83	Francia	5645.99
Zimbabwe	22.59	Suecia	3783.00
Bangladesh	22.52	Países Bajos	3010.41
Francia	18.42	Baréin	2128.60
Nepal	17.59	Irlanda	1751.34
Tailandia	16.88	Japón	1174.01
Hungría	15.95	Islandia	1137.43
Bielorrusia	15.83	Bielorussia	652.58

Similar a la tabla anterior, Estados Unidos sigue siendo el dato más atípico de toda la base de datos. Posterior a la obtención de estos resultados, se analizaron diferentes alternativas para puntos de corte. El primero es retirar los 5 países más atípicos y el segundo es realizar un corte analítico donde se corta en el punto en el que los datos posteriores tienen distancias "similares". En secciones posteriores se retomarán estos dos tipos de corte y el tipo de distancia de Mahalanobis empleada.

4.3. Regresión

4.3.1. Regresión Multivariada

Para la regresión multivariada empleamos diferentes criterios a combinar, el primero es la base de datos empleada (con o sin GHS index), el segundo la distancia de Mahalanobis empleada para

realizar la detección de datos atípicos (normal o robusta), el tipo de recorte empleado (sin recorte, primeros cinco o analítico) y finalmente el tipo de regresión lineal aplicada (tradicional, robusta, Kendall o Spearman). Se realizaron todas los posibles combinaciones para las variables casos por 1000 habitantes y muertes por 1000 habitantes, pues resulta de interés comparar variables representativas de la situación del COVID-19 en un país, la primera con una alta incertidumbre y la segunda con poca. No obstante, solo se incluyeron en este artículo las combinaciones que otorgaban mejores resultados, a excepción de los métodos tradicionales, pues se pretende mostrar la mejoría de las variantes analizadas.

En la siguiente tabla se presentan los resultados para la base de datos sin la variable categórica y sin recortar datos atípicos, empleando la regresión lineal multivariada tradicional. La tabla muestra la variable explicada por las demás variables y el valor del R^2 ajustado respectivo, es decir, en que proporción se puede explicar la variable en cuestión respecto a las demás variables. Nótese que los resultados obtenidos no logran explicar las variables de una manera adecuada. pues en el mejor de los casos el R^2 ajustado máximo obtenido es de 0.767 para el número de muertes por 1000 habitantes, lo cual no alcanza la cota mínima deseada del 0.8.

Resulta interesante observar que las variables explicativas para esta variable corresponden al número de casos confirmados por 1000 habitantes y a la mortalidad confirmada (muertes sobre casos confirmados), por lo tanto, esta relación es trivial y se obtiene a partir de variables con un alto grado de incertidumbre. Estos resultados van encontra de nuestros objetivos, aunque queremos explicar las muertes, pretendemos hacerlo mediante variables robustas. Es por esto que para futuros trabajos, se sugiere no incluir la mortalidad confirmada y si los resultados son negativos buscar otras variables que puedan explicar el número de muertos confirmados por COVID-19, así como trabajar con el PIB per capita.

Sin recortar	
Variables	R2
Casos por 1000	0.558
Muertes por 1000	0.767
Mortalidad confirmada	0.666
Días de duración	0.251
Médicos por 1000	0.643
Camas por 1000	0.486
Densidad	0.186
Edad mediana	0.729
Proporción urbana	0.409
PIB 2019	0.388
PIB Sector Salud	0.552

Al analizar todas las posibles combinaciones mencionadas al comienzo de esta sección para casos confirmados por 1000 habitantes, obtenemos que las mejores aproximaciones se obtienes para los datos sin la variable categórica, recortando los datos de manera analítica al emplear la distancia de Mahalanobis robusta, usando la regresión lineal robusta. Nótese que las variables que explican los casos son las muertes confirmadas por 1000 habitantes, la mortalidad confirmada y las pruebas confirmadas acumuladas por 100 habitantes. Cabe resaltar, que el R^2 ajustado mejoró drásticamente, al pasar de 0.558 con la regresión lineal tradicional sin recortar a típicos, al 0.907 con la variación comentada anteriormente.

Analítico Robusto				
Casos por 1000	Índices	Variables	R2 Ajustado	F-Statistic
	2	Muertes por 1000	0.907	1.62E-30
	3	Mortalidad confirmada		
	7	Acumulado por 1000		

Tomando como variable explicada las muertes confirmadas por 1000 habitantes, obtenemos los

mejores resultados para la misma combinación mencionada anteriormente. En este caso obtenemos un mejor R^2 ajustado y las variables explicadas corresponden a los casos confirmados por 1000 habitantes y a la mortalidad confirmada.

Finalmente, hay que resaltar que las regresiones lineales multivariadas se realizaron con una cota del 0.05 para los p-values.

En la tablas que se muestra a continuación se evidencia el rendimiento superior de la regresión lineal multivariada robusta respecto a las demás variantes para nuestra base de datos.

Muertes por 1000	Índices	Variables	R2 Ajustado	F-Statistic
	1	Casos por 1000	0.91594	7.05E-33
	3	Mortalidad confirmado		

R2 Ajustado Muertes por 1000			
Paramétrico	Robusto	Kendall	Spearman
0.77605	0.91594	0.76919	0.78572

4.3.2. Regresión Bivariada

Como primer paso se realiza un análisis exploratorio de los datos, esto mediante un plotmatrix. Observando los resultados gráficos, las variables que presentan estructuras de interés en su relación son:

- Casos confirmados y número de muertes
- Personal médico y camas en UCI
- Personal Médico y la edad mediana de la población

Se muestran a continuación los resultados de realizar distintos tipos de regresiones, mediante el método lineal de mínimos cuadrados, mínimos cuadrados robustos, modelos con coeficientes hallados a partir de correlación tipo Spearman y Kendall, y el modelo de regresión no parametrica Nadaraya-Watson, que en general es el que mejor desmpeño presenta, dado que capta cualquier tipo de relacion en los datos, no necesariamente lineal como los otros métodos usados. Los coeficientes de determinación R^2 de los métodos se incluyen en las siguientes gráficas:

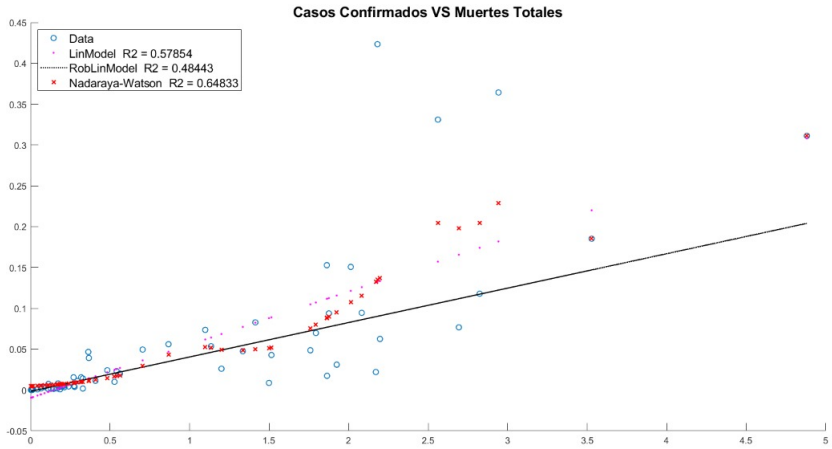


Figura 8: Gráfica de dispersión entre casos confirmados y muertes

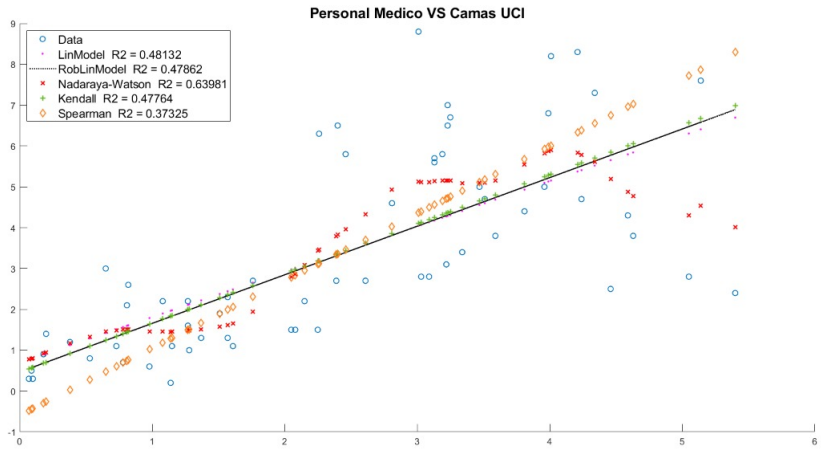


Figura 9: Gráfica de dispersión entre número de médicos y camas en UCI

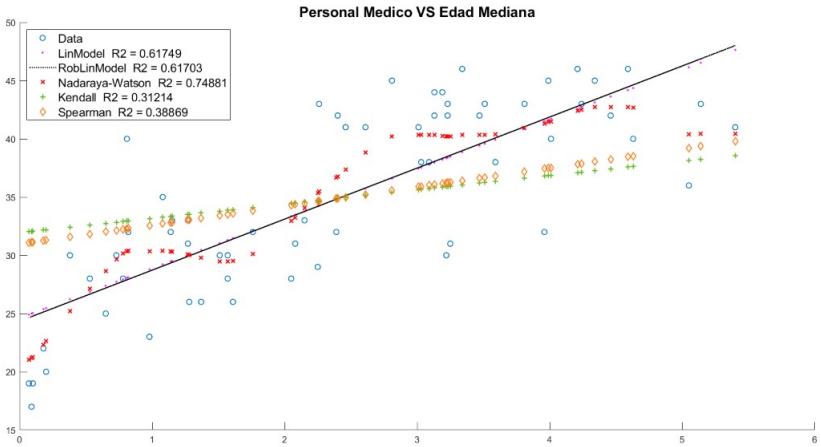


Figura 10: Gráfica de dispersión entre personal médico y la edad mediana de la población

Como se pudo evidenciar en las gráficas, los dados presentan un comportamiento altamente no lineal, por lo que el uso de la regresión Nadaraya-Watson se apropiado, pues independientemente de la forma de la nube de puntos, se ajusta al comportamiento con un buen ajuste.

4.4. Componentes Principales

En la siguiente tabla se presentan los valores de las direcciones de cada variable en el autovector asociado al mayor autovalor. Como se puede observar, la dirección más significativa es la de la variable Mortalidad, mientras que la menos importante para nuestro estudio es el PIB.

	Autovector
Casos	0.004600771
Muertes	-0.237847575
Mortalidad	0.971242752
Duración	0.000156075
Médicos	-0.00025803
Camas	0.000994715
Pruebas	0.000239016
Densidad	-1.32E-06
Edad	-0.000644169
Prop Urbana	-0.009622099
PIB	7.05E-07
Salud	-0.000862405

En esta tabla se encuentran los porcentajes de variabilidad explicada, por cada componente, es decir, la primera fila representa al primer componente, la segunda al segundo y así sucesivamente. Los resultados obtenidos muestran que el primer componente explica un poco más del 98 %. Siendo un resultado bastante positivo, pues si se trabaja con el primer componente como una única variable, se pasaría de trabajar con un problema de dimensión doce (pues la base de datos original tiene este número de variables) a solamente trabajar con una. Otro desarrollo que se tuvo en cuenta fue utilizar el primer componente principal como variable explicativa en los métodos de regresión. Esto es, quitar la variable muertes, calcular el primer componente principal (que igualmente explicaba la variabilidad por encima del 98 % y utilizar éste como variable explicativa para explicar el número de muertes. Los resultados obtenidos no fueron buenos, pues no se pudo establecer un modelo apropiado que explicara las muertes a partir del primer componente, es por esto que este desarrollo no se incluye en los resultados del estudio.

Variabilidad Explicada por componente
0.986465188
0.013363335
0.000126033
3.67E-05
7.26E-06
6.40E-07
4.54E-07
2.81E-07
1.09E-07
2.94E-09
1.72E-09
7.20E-11

4.5. Clasificación

Cada uno de los métodos acá presentados fueron realizados mediante validación cruzada con el propósito de obtener mejores resultados, excepto para el método de regresión logística pues en este caso se presentan los resultados de testeo, para el cual se consideró el 70 % para entrenar el modelo y 30 % para testearlo.

4.5.1. Kmedias

Estadísticos	Valor	
Score	0.8533	
Precision	1	0.8514
Recall	0.0833	1

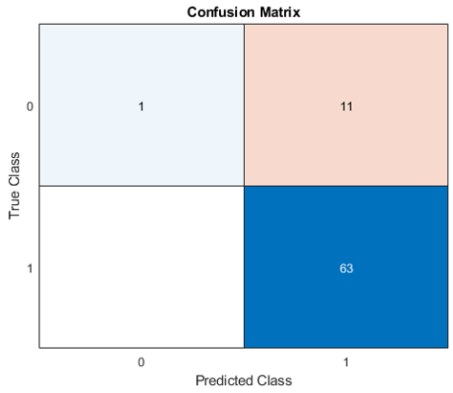


Figura 11: Matriz de confusión con Kmedias

4.5.2. Kmedoids

Estadísticos	Valor	
Score	0.1466	
Precision	0.1486	0
Recall	0.9147	0

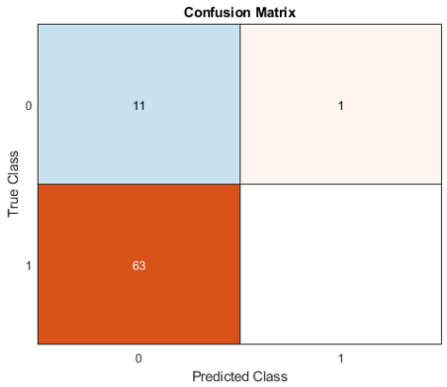


Figura 12: Matriz de confusión de Kmedoides

4.5.3. Discriminante lineal

Estadísticos	Valor	
Score	0.7723	
Precision	0.3684	0.9107
Recall	0.5883	0.8095

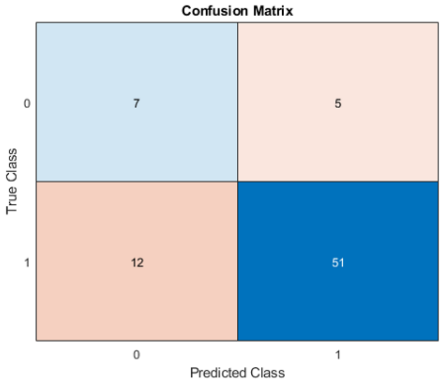


Figura 13: Matriz de confusión con Discriminante Lineal

Como puede evidenciarse, este método presenta un ajuste del 77 % aproximadamente, lo cual resulta ser un buen método a pesar de que comete errores durante la clasificación. Por ejemplo, si se observa el estadístico de precision, es decir si se predijo la cantidad que era de forma correcta, es notorio que este presenta un valor muy bajo para la primera clase, es decir la de los países que están bien preparados, pero muy alto para la segunda clase y lo mismo ocurre para el caso del recall, por lo que esto indica que la predicción no se hizo de forma correcta y la clase que presenta un mayor error de clasificación es la de los países bien preparados.

4.5.4. Discriminante cuadrático

Estadísticos	Valor	
Score	0.8667	
Precision	1	0.863
Recall	0.1667	1

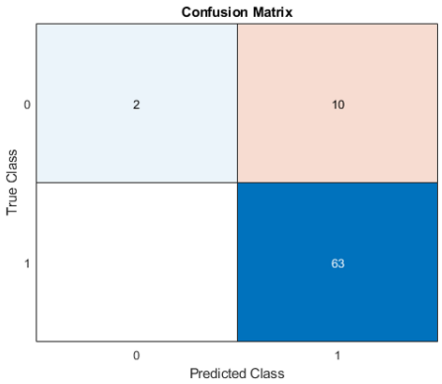


Figura 14: Matriz de confusión con Discriminante Cuadrático

Observando la matriz de confusión es fácil ver que para la clase de países bien preparados se predijo la cantidad que debía predecir y los clasificó de forma correcta, es decir 2 países y por eso se presenta una precisión de 1 para esta clase y un recall de 1 para la otra. Sin embargo, para la clase de países mal preparados también se predijo lo que era, pero cometió errores en la clasificación y por eso se ve tan reducido el recall de la primera clase y la precisión de la segunda. Además, para este método se logra obtener una muy buen ajuste del modelo, presentando así un score del 86.67 % aproximadamente.

4.6. KNN

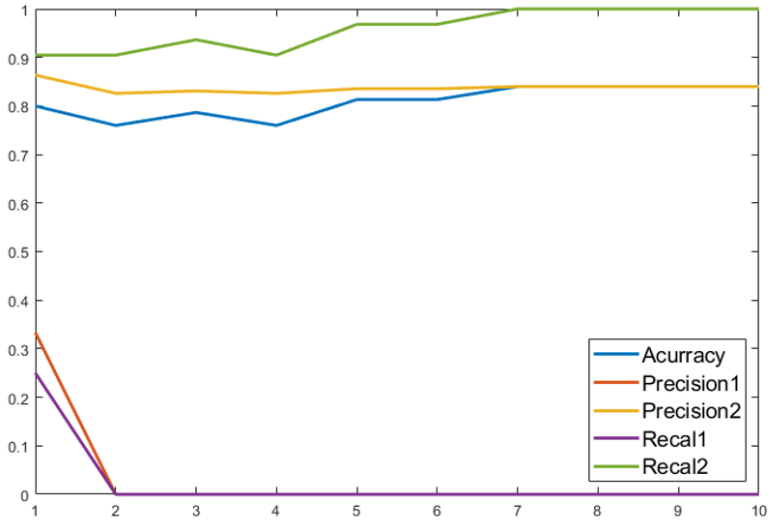


Figura 15: Métricas para diferente cantidad de vecinos

Para este método, se quiso observar el comportamiento de cada una de las métricas de clasificación consideradas con respecto al número de vecinos a considerar, el cual se varió entre uno y 10 vecinos. Luego, como puede observarse en la gráfica, al considerar solo un vecino es cuando se obtienen mejores valores para cada estadístico, aunque para la clase de no preparados se presenta un recall y una precisión demasiado baja. Sin embargo, en este caso no resulta tan grave clasificar un país preparado como no preparado, por lo tanto, si estas métricas se van a cero, lo cual ocurre si se consideran más de dos vecinos, no afecta tanto al modelo. Ahora, considerando esto anterior, puede verse que a partir de siete vecinos, el modelo parece estabilizarse presentando muy buenos resultados en precisión, recall y ajuste del modelo, por lo cual siete vecinos puede considerarse como un número adecuado.

4.6.1. VSM

Estadísticos	Valor	
Score	0.8533	
Precision	0.5455	0.9063
Recall	0.5	0.9206

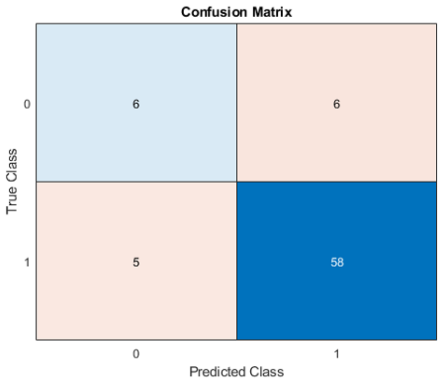


Figura 16: Matriz de confusión con VSM 3

Estadísticos	Valor	
Score	0.8	
Precision	0.3846	0.8871
Recall	0.4167	0.873

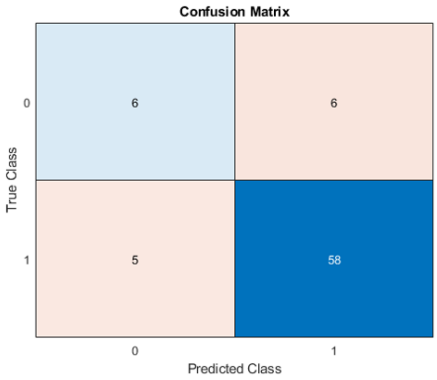


Figura 17: Matriz de confusión con VSM 5

En este caso también se aplicó validación cruzada, pero fue realizada por grupos, pues por individuo se incrementaba demasiado el tiempo computacional. En primer lugar se dividió el conjunto de datos en tres grupos diferentes y entre ellos se iban prediciendo, con los cuales se obtuvieron los primeros resultados presentados. Como puede observarse, para cada clase se predijo de más y además las clasifica de forma incorrecta, por esta razón las métricas de recall y precisión se ven afectadas. Sin embargo, se sigue presentando un score de 85.33 %, lo cual lo posiciona como un buen método al momento de clasificar este conjunto de datos.

Por otro lado, se varió el número de grupos de tres a cinco para ver esto como afectaba los resultados obtenidos anteriormente, pero como es de notarse todas las métricas redujeron en valor, lo cual indica que observaciones que estaba prediciendo y clasificando de forma correcta, ahora lo hace incorrectamente, por lo que el método resulta mejor con un número de grupos menor.

4.6.2. Discriminante Logístico

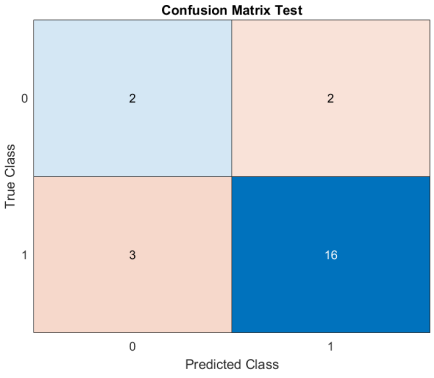


Figura 18: Matriz de confusión con discriminante logístico

Estadísticos	Valor	
Score	0.7826	
Precision	0.4000	0.8889
Recall	0.5000	0.8421

diFinalmente, en el caso de regresión logística para el conjunto de datos de testeo, puede observarse que nuevamente se presentan errores tanto en la predicción como en la clasificación, a

pesar de esto el modelo es capaz de predecir de forma correcta el 50 % de los países bien preparados y el 84 % de los no preparados. Sin embargo, si se compara su score con los métodos anteriores, puede concluirse que no es el mejor método de clasificación para los datos que se están considerando.

5. Conclusiones

- Aproximaciones robustas otorgan mejores resultados en el caso multivariado.
- Aproximaciones no parámetricas dan mejores resultados en el caso bivariado.
- Existen relaciones claras entre las variables consideradas.
- Las muertes se explican mejor respecto a las otras variables que los casos confirmados, además éstas al tener una menor incertidumbre permiten contemplar la posible construcción de un indicador robusto para diagnosticar la situación por COVID-19 en un país.
- Según el estudio, la variable menos significativa es el PIB.
- Es importante realizar un análisis exhaustivo de los datos, en lugar de usar herramientas estándar o preestablecidas.

Finalmente, como trabajo a futuro se plantea el desarrollo de la regresión multivariada no lineal (ej. Nadara-Watson). Otra ampliación a considerar es la búsqueda de relaciones robustas entre las muertes y otras variables. Y por último, la inclusión de datos faltantes, a partir de los modelos construidos.

Referencias

[1] Amparo Baíllo, A. G. *100 problemas resueltos de estadística multivariante (implementados en MATLAB)*. DELTA Publicaciones, 2008.

[2] Cabana, E., Lillo, R. E., and Laniado, H. Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators. *Statistical Papers* (2019), 1–27.

[3] Cabana, E., Lillo, R. E., and Laniado, H. Robust regression based on shrinkage with application to living environment deprivation. *Stochastic Environmental Research and Risk Assessment* (2020), 1–18.

[4] Index Project Team. `https://www.ghsindex.org/`, 2019.

[5] López Miranda, C. Modelo predictivo de riesgo de morosidad para créditos bancarios usando datos simulados.

[6] minitab. `https://bit.ly/3eoQ3qw`.

[7] Peña, D. *Análisis de datos multivariantes*. McGraw-Hill España, 2013.

[8] Rodríguez-Niño, N., et al. Un pronóstico no paramétrico de la inflación colombiana. *Borradores de Economía; No. 248* (2003).

[9] Wasserman, L. *All of nonparametric statistics*. Springer Science & Business Media, 2006.