Regresión logística robusta para la clasificación de residuos sólidos

María Alejandra Vélez Clavijo, Valentina Moreno Ramírez, Alejandra Palacio Jaramillo, Juan José Wilches Rivas.
Asesor: Henry Laniado Rodas
Departamento de Ciencias Matemáticas, Escuela de Ciencias Exactas Universidad EAFIT
mavelezc1@eafit.edu.co, vmorenor@eafit.edu.co, apalacioj@eafit.edu.co, jjwilches@eafit.edu.co, hlaniado@eafit.edu.co

***Resumen*—La problemática ambiental que se ha vivido en las últimas décadas ha traído consecuencias graves para la vida en la tierra. En busca de contribuir a la protección y preservación del entorno, se pretende clasificar residuos sólidos en orgánicos e inorgánicos mediante la implementación de un modelo de regresión logística, y su versión robusta (es decir, insensible a datos anómalos). En este orden de ideas, se plantea un modelo, tanto clásico como robusto, el cual recibe un conjunto de imágenes que posteriormente se procesan para sacar características relevantes de sus bordes, los cuales, a partir de las regresiones logísticas antes mencionadas, permitirán clasificar el residuo en orgánico o inorgánico. Finalmente se evalúa la calidad de los modelos para analizar qué tan confiable es para la consecución del objetivo planteado**

***Palabras Clave*—Medio ambiente, Modelo matemático, Probabilidad, Residuos sólidos, Regresión logística, Regresión logística robusta.**

I. Introducción

En los últimos años, el planeta Tierra se ha visto envuelto en una gran cantidad de problemas que ponen en riesgo el bienestar de sus individuos. Uno de los más evidentes ha sido el calentamiento global, el cual ha afectado a las diversas formas de vida en el planeta, perjudicando la supervivencia de estos en el entorno. Con el objetivo de contribuir con el mejoramiento del medio ambiente, el ser humano se ha ido interesando por la innovación de tecnologías y el avance en la ciencia, procurando desarrollar una sociedad que cada día sea más consciente del cuidado ambiental, y las implicaciones que se tendrán a futuro al no darse su preservación.

Los residuos sólidos son de gran interés y preocupación para las autoridades, instituciones y comunidad, ya que en dichos residuos se identifica una verdadera amenaza contra la salud humana y el ambiente, específicamente en lo relacionado con el deterioro estético de los centros urbanos y del paisaje natural; en la proliferación de vectores transmisores de enfermedades y el efecto sobre la biodiversidad [1].

Con el propósito de aportar a la preservación del medio ambiente, este proyecto pretende facilitar procesos bases para el cambio ambiental, como lo es la correcta clasificación de los residuos sólidos. Para esto, se realizan dos modelos de clasificación con base en la regresión logística, en su versión clásica y en su versión robusta, los cuales van a permitir etiquetar los residuos como orgánicos o inorgánicos.

El modelo de clasificación va a ser entrenado y validado con imágenes de residuos, entre los cuales se encuentran cartones, latas, papel, entre otros; estos procesos van a permitir saber qué tan bien clasifica cada modelo, para determinar y comparar la calidad de cada uno y luego poder llevarlos a la práctica y darle uso en la vida diaria. Adicionalmente, se espera que a futuro el proyecto pueda ser implementado en sistemas de clasificación automática de residuos.

II. Materiales y métodos

Inicialmente, se tiene un conjunto de datos de 806 imágenes de residuos sólidos, de las cuales el 80 % van a ser destinadas al conjunto de entrenamiento y 20 % al conjunto de datos de validación.

II-A. Regresión logística binaria :

La regresión logística binaria es un método estadístico que predice el resultado de una variable categórica en función de una variable característica independiente. Dicha regresión permite clasificar conjuntos de observaciones en dos categorías dependiendo del valor que tome la variable categórica o predictora.

Definición. Sea ***y*** la variable dependiente, dicha variable es categórica y va a servir como variable predictora. Además, y es una variable binaria, la cual tendrá como posibles valores de sus etiquetas a 0 o 1, lo que se traduce a las etiquetas empleadas en el modelo de clasificación de residuos como orgánico o inorgánico.

Sean $x_1, x_2, \ldots, x_i$ las variables independientes que van a representar las características del borde del residuo. En general las x_i presentan las posibles condiciones que puedan incidir en la variable dependiente **y** .

Además, se tienen unos parámetros del modelo, los cuales son $w_1, w_2, \ldots, w_i$
La variable categórica **y**, se obtiene al hacer combinaciones lineales entre los parámetros del modelo y las características x_i . Para la construcción de dichas combinaciones lineales se van a tener dos vectores:

1. El vector que contiene las características del residuo

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix}$$

1. El vector transpuesto al vector de los parámetros

$$w^t = \begin{bmatrix} w_0 & w_1 & w_2 & \cdots \end{bmatrix}$$

Luego,

$$\begin{bmatrix} w_0 & w_1 & w_2 & \cdots \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix} = w^t x$$

$$w^t x = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

Para modelar la probabilidad de éxito con la regresión logística se necesita evaluar la logística estándar acumulada con la combinación lineal entre los parámetros del modelo y las variables características , de la siguiente manera:

$$F\left(w^t x\right) = \int_{-\infty}^{w^t x} \frac{e^u}{(1+e^u)^2} du$$
$$\int_{-\infty}^{w^t x} \frac{e^u}{(1+e^u)^2} du = 1 - \frac{1}{1+e^{w^t x}}$$

Luego, las probabilidades del modelo van a estar dadas por

$$p = \frac{e^{w^t x}}{1 + e^{w^t x}}$$

La regresión logística devuelve predicciones bien calibradas de forma predeterminada, ya que optimiza directamente la pérdida de registro [2]. Dicha calibración está dada por el regresor sigmoide, función que permitirá que la probabilidad que arroje el modelo esté entre 0 y 1.
La función sigmoide tiene la forma:

$$S(t) = \frac{1}{1 + e^{-t}}$$

Y por tanto la aplicación de la combinación lineal en la función sigmoide resultante es:

$$S\left(w^t x\right) = \frac{1}{1 + e^{-w^t x}}$$

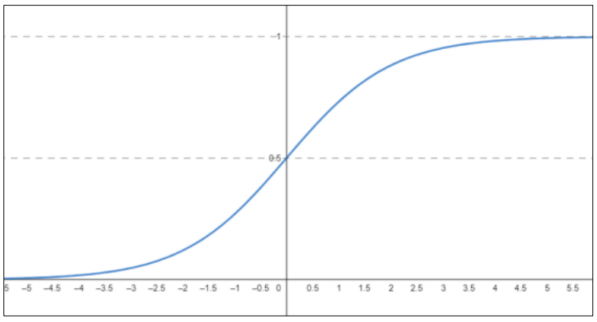

Fig. 1. Gráfica de la función sigmoide.

El objetivo del proyecto es clasificar los residuos en dos categorías, se va a tener entonces una primera expresión

$$P = (y = categoría|x)$$

la cual representa la probabilidad de que, **y** sea perteneciente a la categoría orgánico o inorgánico, dadas unas características **x** influyentes. Además, dichas probabilidades van a estar acompañadas de unos parámetros **w**.
De este modo, $P(y = inorgánico|x; w) = 1 - P(y = orgánico|x; w)$, lo que es lo mismo que $P(y = 0|x; w) = 1 - P(y = 1|x; w)$.
Dicha probabilidad en la función sigmoide quedaría:

$$S\left(w^t x\right) = P(y = 1 \mid x; w)$$

$$P(y = 1 \mid x; w) = \frac{1}{1+e^{-w^t x}}$$

Al calibrar el modelo, las clasificaciones estarán dadas por la siguiente función:

$$P(y = m \mid x; w) = \left\{ \begin{array}{l} 1 \text{ si } s\left(w^t x\right) \geq 0,5 \\ 0 \text{ si } s\left(w^t x\right) < 0,5 \end{array} \right\}$$

Ahora, se encontrará la expresión por la cual van a ser generados los parámetros a partir de la función que calibra el modelo.
Para estimar dichos parámetros **w** del modelo que representen de forma óptima a los datos, se debe emplear ya sea una función de pérdida o verosimilitud, con el objetivo de que dichos parámetros minimicen la pérdida y maximicen la verosimilitud. En este caso se maximizará la función de verosimilitud.
Sea R una muestra aleatoria de tamaño n tal que

$$R = \{(x_i, y_i)\,; i = 1, \ldots, n\}$$

donde $y_i = \{0, 1\}$ es el valor de **y** en el i-ésimo elemento de la muestra aleatoria y las x_i son las características independientes. [3]
Ahora, para maximizar la función de verosimilitud se va a tener en cuenta las condiciones iniciales del modelo, las cuales exigen que al ser una regresión logística binaria se obtengan solo dos posibles resultados: 0 o 1, los cuales son excluyentes entre sí, entonces se puede representar el espacio con el que se está trabajando, en una distribución de Bernoulli.
Asi, al tener las probabilidades $P\left(y_i = 0 \mid x_i; w\right) = 1 - S\left(w^t x_i\right)$ y $P\left(y_i = 1 \mid x_i; w\right) = S\left(w^t x\right)$, estas serán distribuidas con Benoulli de la siguiente manera:

$$P\left(y_i \mid x_i; w\right) = \left(S\left(w^t x_i\right)\right)^{y_i} \left(1 - S\left(w^t x_i\right)\right)^{1-y_i}$$

Luego, al definir p_i como la probabilidad de éxito (es decir,la probabilidad de que el residuo sea orgánico), esta va a estar dada por $p_i = S\left(w^t x_i\right)$ y se va a poder modelar la probabilidad de todas las observaciones independientes, asi:

$$P(Y \mid X; w) = \prod_{i=1}^{n} \left(p_i\right)^{y_i} \left(1 - p_i\right)^{1-y_i}$$

Dicha expresión resulta ser la función de verosimilitud L(w). La función de verosimilitud representa los datos de entrada, en este caso va a representar los datos que se tengan de los residuos, por lo cual se necesitan encontrar los parámetros ***w*** que permitan maximizarla.
Para maximizarla se sacan primero logaritmos a ambos lados de la expresión

$$\ln|L(w)| = \ln\left|\prod_{i=1}^{n}(p_i)^{y_i}(1-p_i)^{1-y_i}\right|$$

Por propiedades de los logaritmos,

$$\mathrm{l(w)} = \sum_{i=1}^{n} y_i \ln|p_i| + (1-y_i)\ln|1-p_i|$$

Ahora, para calcular el gradiente de dicha función a maximizar se utiliza la propiedad de suma de los gradientes,

$$\frac{\partial l(w)}{\partial w} = \sum_{i=1}^{n}\left(\frac{y_i}{p_i}\frac{\partial p_i}{\partial w} - \frac{(1-y_i)}{(1-p_i)}\frac{\partial p_i}{\partial w}\right)$$

Luego,

$$\frac{\partial p_i}{\partial w} = \frac{e^{-w^t x_i}}{\left(1+e^{-w^t x_i}\right)^2}$$

Recordando que $p_i = \frac{1}{1+e^{-w^t x_i}}$, entonces se puede escribir $\frac{\partial p_i}{\partial w}$ como $p_i(1-p_i)$, pues

$$p_i(1-p_i) = \left(\frac{1}{1+e^{-w^t x_i}}\right)\left(1-\frac{1}{1+e^{-w^t x_i}}\right)$$

$$p_i(1-p_i) = \left(\frac{1}{1+e^{-w^t x_i}}\right)\left(\frac{1+e^{-w^t x_i}-1}{1+e^{-w^t x_i}}\right)$$

$$p_i(1-p_i) = \left(\frac{1}{1+e^{-w^t x_i}}\right)\left(\frac{e^{-w^t x_i}}{1+e^{-w^t x_i}}\right)$$

$$p_i(1-p_i) = \left(\frac{e^{-w^t x_i}}{\left(1+e^{-w^t x_i}\right)^2}\right)$$

Luego, se multiplica dicha expresión por la característica x_i

$$\frac{\partial p_i}{\partial w} * \frac{x_i}{x_i} = p_i(1-p_i)$$

$$\frac{\partial p_i}{\partial w} * x_i = p_i(1-p_i) * x_i$$

Retomando la expresión $\frac{\partial l(w)}{\partial w}$ y reemplazando $\frac{\partial p_i}{\partial w}$ en la expresion, se obtiene:

$$\frac{\partial l(w)}{\partial w} = \sum_{i=1}^{n}\left(\frac{y_i}{p_i}p_i(1-p_i)x_i - \frac{(1-y_i)}{(1-p_i)}p_i(1-p_i)x_i\right)$$

$$\frac{\partial l(w)}{\partial w} = \sum_{i=1}^{n}\left(y_i(1-p_i)x_i - (1-y_i)p_i\right)$$

$$\frac{\partial l(w)}{\partial w} = \sum_{i=1}^{n}\left(y_i x_i - p_i y_i x_i\right) - \left(p_i x_i - y_i p_i x_i\right)$$

$$\frac{\partial l(w)}{\partial w} = \sum_{i=1}^{n}\left(y_i x_i - p_i y_i x_i - p_i x_i + y_i p_i x_i\right)$$

$$\frac{\partial l(w)}{\partial w} = \sum_{=1}^{n}\left(y_i x_i - p_i x_i\right)$$

$$\frac{\partial L(w)}{\partial w} = \sum_{i=1}^{n}\left(y_i - p_i\right)x_i$$

Para maximizar la función de verosimilitud se debe ir en sentido del gradiente, es decir, que el signo del gradiente es positivo. Así los parámetros quedarían de la forma,

$$w_{t+1} = w_t + \varepsilon\sum_{i=1}^{n}(y_i - p_i)x_i$$

Donde, w_{t+1} es la actualización de parmetros, w_t es del valor del parámetro anterior, ε representa una tasa de aprendizaje y $\sum_{i=1}^{n}(y_i - p_i)x_i$ es el gradiente.
De esta forma, se pueden obtener los parámetros **w** para el modelo.

II-B. Regresión logística binaria (versión robusta):

La regresión logística es un modelo sensible a los datos anómalos, los cuales son valores en el conjunto de datos que toman un valor completamente distinto o extremo al de la mayoría de estos; Como consecuencia, se pueden alterar los resultados y conclusiones del modelo y, por ello, resulta conveniente aplicar su versión robusta.
Cuando se habla de robustez en estadística, especialmente en el campo de estimación, se hace referencia a la poca sensibilidad de un modelo a la presencia de datos raros. Es decir, con base en el objetivo planteado, al aplicar robustez a la regresión logística, esta perderá sensibilidad frente a los datos raros que se tengan y, consecuentemente, arrojará resultados más precisos gracias a la reducción de dispersión en los datos.
Ahora bien, para volver robusta la regresión logística, se modifica la manera en la cual se obtienen los parámetros del modelo. De acuerdo con [**4**], el estimador de los w_i por mínimos cuadrados modificado está dado por:

$$w_0 = \bar{y} - \bar{x}tw_i$$
$$w_i = \sum_{xx}^{-1}\sum_{xy}$$

Tomando en cuenta lo anterior, es clave identificar que la expresión $\sum_{xx}^{-1}$ representa una matriz de covarianza, en este caso, de las características de los bordes de los residuos y $\sum_{xy}$ es la covarianza entre las características y la etiqueta de los residuos. Siendo así, para hacer robusta la estimación de estos parámetros, la matriz de covarianza juega un papel fundamental.
La matriz de covarianzas se estimará con coeficientes de correlación lineal como el de Kendall, Pearson y Spearman, teniendo en cuenta que la covarianza entre x e y se puede definir como:

1. $\mathrm{Cov}(x,y) = \rho_k S_x S_y$
2. $\mathrm{Cov}(x,y) = p_s S_x s_y$
3. $\mathrm{Cov}(x,y) = r s_x s_y$

Donde p_k es el coeficiente de correlación de Kendall, ρ_s es el coeficiente de correlación de Spearman, r es el coeficiente de correlación de Pearson, s_x es la desviación estándar de la caracteristica x y s_y es la desviación estándar de la característica y.

- *Coeficiente de correlación de Kendall:*

En estadística, el coeficiente de correlación de rango de Kendall, comúnmente conocido como coeficiente τ de

Kendall es un estadístico usado para medir la asociación ordinal entre dos cantidades medidas.
Este coeficiente se puede definir de la siguiente manera: sea $(x_1, y_1), \ldots, (x_n, y_n)$, una muestra aleatoria conjunta X e Y, cuando cualquier par de observaciones (x_i, y_i) y (x_j, y_j), donde i $<$ j se dice que son un par concordante si ambos $x_i > x_j$ y $y_i > y_j$ o el caso de que $x_i < x_j$ y $y_i < y_j$, de no ocurrir esto se dice que el par es discordante [4]. Aclarando lo anterior, el coeficiente de corrclación de Kendall se define como:

$$p_k = \frac{(\#\text{ de pares concordates}) \ -(\#\text{ de pares discordantes})}{\begin{pmatrix} n \\ 2 \end{pmatrix}}$$

- *Coeficiente de correlación de Spearman:*

El coeficiente de correlación de Spearman es una medida de asociación o interdependencia no paramétrica entre dos variables [5].
Para calcular este coeficiente, se sigue este procedimiento: Suponga que tiene n pares de datos con clasificaciones asociada $(u_1, u_2, \ldots, u_n)$ $(v_1, v_2, \ldots, v_n)$, donde los u_i son tomados en manera ascendente y los v_i serán la permutación de los u_i. El coeficiente de correlación del momento del producto de u_i y v_i se calcula por mínimos cuadiados como [5] :

$$S_s = \sum (u_i - v_i)^2$$

Luego,

$$_s = 1 - \frac{6S_s}{n(n^2-1)}$$

- *Coeficiente de correlación de Pearson:*

El coeficiente de correlación de Pearson es la razón entre la covarianza de dos variables y el producto de sus desviaciones estándar; considerándose una medida normalizada y mejorada de la covarianza. Al igual que con la covarianza, la medida solo puede reflejar una correlación lineal de variables e ignora muchos otros tipos de relación o correlación, además es sensible a la distribución de los datos [6].

Este coeficiente se define como:

$$r(x, y) = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

II-C. Procesamiento de imágenes

Como el modelo de clasificación de residuos trabaja con bases de datos de imágenes, se hizo necesario dar un enfoque a la parte de procesamiento de imágenes pretendiendo que el lector se relacione con el proceso que realiza la máquina.
Antes de proceder con la explicación del algoritmo usado para el análisis de los fotogramas, es necesario tener el concepto previo de una imagen digital, la cual es una representación bidimensional compuesta de un número finito de elementos, cada uno con una localidad y un valor particular. A estos elementos se les llama píxeles, siendo este el término utilizado para denotar la unidad mínima de medida de una imagen digital.
Ampliando la imagen en una zona cualquiera, se pueden apreciar estos valores, que se muestran en forma de matriz, correspondiéndose cada elemento de la matriz N_{ij} con las coordenadas en el plano $x = i, y = j$.
Asi pues, una imagen podra ser considerada como una funcion de dos dimensiones $f(x, y)$, donde x e y son las coordenadas de un plano que contiene todos los puntos de esta; y $f(x, y)$ es la amplitud en el punto (x, y), y se le llama intensidad o nivel de gris de la imagen en ese punto. El valor de esta intensidad está entre 0 y 255 . Fn el caso de que tanto las coordenadas x e y, como los valores de intensidad de la función f sean discretos y finitos, se habla de una imagen digital.

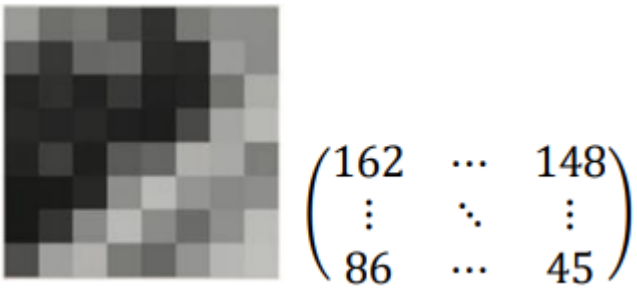

Fig. 2. Fragmento de una imagen con su respectiva matriz

Ahora bien, el procesamiento de imágenes digitales es el conjunto de técnicas que se aplican a las imágenes con el objetivo de mejorar la calidad o facilitar la búsqueda de información, en este modelo se pretende facilitar el reconocimiento de los archivos dados para su clasificación.
Clasificar las imágenes sin modificarlas resulta difícil tanto para la máquina como para el humano. Debido a esto se desarrollaron algoritmos que permiten facilitar esta tarea, entre ellos está el algoritmo Canny, el cual permite la detección de todos los bordes de la imagen usando métodos de detección de contornos mediante el empleo de máscaras de convolución y basado en la primera derivada, permitiendo de esta manera la reducción significativa de datos en una imagen, preservando las propiedades estructurales. [7]
El Algoritmo de detección de Canny se ejecuta en 5 etapas separadas:
1. Suavizado de la imagen: Debido a que la detección de bordes es un proceso susceptible al ruido en la imagen, el primer paso es eliminarlo o reducirlo lo más que se pueda, minimizando la cantidad de variaciones de intensidad entre píxeles vecinos, eliminando aquellos píxeles cuyo nivel de intensidad es muy diferente al de sus vecinos.
Este proceso se obtiene promediando los valores de intensidad de los píxeles en el entorno de vecindad con una máscara de convolución de media cero y desviación estándar . Sin embargo, se debe de tener cuidado de no realizar un suavizado excesivo, pues se podrían perder detalles de la imagen y provocar un pésimo resultado final.

A continuación, un ejemplo de este suavizado en una imagen:

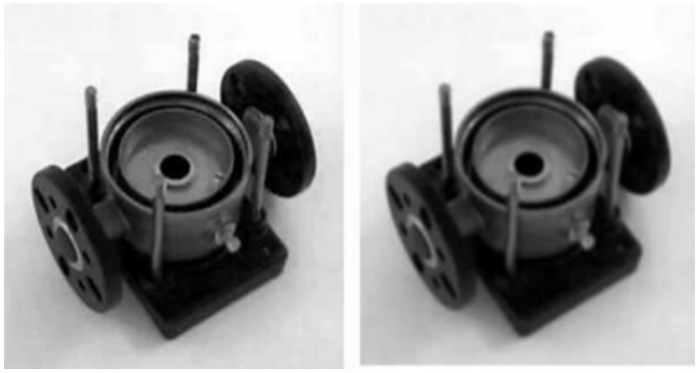

Fig. 3. Filtro de Gauss con una desviación estándar de $\sigma = 1{,}5$. [8]

2. Obtención de gradientes: Una vez que la imagen ha sido suavizada con el filtro Gaussiano, se calcula el gradiente de esta, el cual permite distinguir los bordes mediante su valor encontrado. Aquellos pixeles con un gradiente alto serán considerados bordes y para encontrarlos la imagen se filtra nuevamente, esta vez utilizando un Kernel Sobel, el cual es un operador diferencial discreto que calcula una aproximación al gradiente de la función de intensidad de una imagen $f(x, y)$.

Para cada punto de la imagen a procesar, el resultado del operador es tanto el vector gradiente correspondiente como la norma de este vector, dando la magnitud del mayor cambio posible, la dirección de este y el sentido desde oscuro a claro.
Matemáticamente el operador utiliza dos kernels de 3x3 elementos para aplicar convolución a la imagen original para calcular aproximaciones a las derivadas, un kernel para los cambios verticales y otro para las horizontales [9].
Sea A la matriz que contiene el valor de intensidsd de la imagen original, de esta manera se puede obtener el resultado del gradicnte horizontal G_x y el gradiente vertical G_y, calculándolos de la siguiente manera:

$$G_x = \begin{pmatrix} -1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ -2 & \cdots & 1 \end{pmatrix} * A$$

$$G_y = \begin{pmatrix} -1 & \cdots & 2 \\ \vdots & \ddots & \vdots \\ -2 & \cdots & 1 \end{pmatrix} * A$$

El borde de una imagen puede apuntar en diferentes direcciones, por lo que el algoritmo de Canny utiliza cuatro filtros para detectar horizontal, vertical y diagonal en los bordes de la imagen borrosa.
En cada punto de la imagen, los resultados de las aproximaciones de los gradientes horizontal y vertical pueden ser combinados para obtener la magnitud del gradiente mediante

$$G = \sqrt{G_x^2 + G_y^2}$$

Con esta información también se podrá calcular la dirección perpendicular a los bordes este usando la fórmula:

$$\theta = \tan^{-1} \frac{G_x}{G_y}$$

Fig. 4. Imagen suavizada vs Imagen con los valores de gradientes altos [8]

3. Supresión de falsos máximos: Esta técnica es utilizada para perfeccionar los bordes encontrados en el paso anterior. El proceso consiste en recibir la imagen resultante en la obtención de gradientes, en dicha imagen encontrar la orientación de los puntos de borde de la imagen, y tomar dos umbrales, el primero más pequeño que el segundo.
Para cada punto de la imagen se debe localizar el siguiente punto de borde no explorado que sea mayor al segundo umbral. A partir de dicho punto seguir las cadenas de máximos locales conectados en ambas direcciones perpendiculares a la normal del borde siempre que sean mayo res al primer umbral. Asíse marcan todos los puntos explorados y se almacena la lista de todos los puntos en el contorno conectado. Es así como en este paso se logra eliminar las uniones en forma de Y de los segmentos que confluyan en un punto.
Básicamente, se debe escanear la imagen para eliminar los píxeles que no formen parte de los bordes. Para esto se compara el valor de cada píxel con sus vecinos cercanos en la dirección del gradiente (perpendicular al borde). Si el valor del píxel es mayor que sus píxeles vecinos, entonces este es considerado un máximo local y el algoritmo lo acepta. De lo contrario, si el píxel resulta no ser un máximo local, entonces es suprimido. El resultado final será una imagen con bordes muy finos.

4. Umbral de histéresis: El procedimiento anterior logra determinar los píxeles que conforman los bordes con bastante precisión. Sin embargo, aún pueden quedar algunos píxeles provenientes del ruido o de variaciones en los colores de la imagen. En esta cuarta etapa se decide cuáles píxeles pertenecen realmente a bordes y cuáles no. Para ello, se deben fijar dos valores de umbral,$mínVal y máxVal$. Los píxeles con gradientes de intensidad mayores que $máxVal$ aceptados como pertenecientes a los bordes, mientras que los menores que $mínVal$ serán descartados. Los píxeles correspondientes a bordes con valores de gradientes que se encuentren entre estos dos umbrales son etiquetados como píxeles débiles. Estos últimos serán o no aceptados, dependiendo de su conectividad.

Si están conectados a píxeles “fuertes”, se consideran parte de los bordes; de lo contrario, también se descartan.

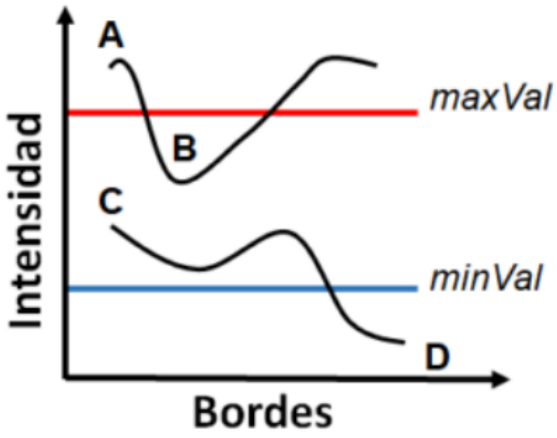

Fig. 5. Simplificación de umbral de histéresis

La fig.6 muestra el valor de la intensidad de los píxeles que conforman los bordes. En este caso, el píxel A será aceptado como parte del borde dado que su valor supera el umbral $máxVal$, mientras que el píxel D será descartado por tener un valor inferior a $mínVal$

Por otra parte, los píxeles B y C se consideran débiles por encontrarse entre los dos valores umbrales. Sin embargo, B será aceptado como parte de un borde, mientras que C no. La razón de esto es que B está conectado a A, que es un píxel fuerte, pero C sólo está conectado a píxeles débiles o descartados.

En relación con todo el proceso descrito anteriormente se puede decir que el algoritmo Canny, tiene como principal ventaja su gran adaptabilidad para poder ser aplicado a diversos tipos de imágenes, además de no disminuir su performance ante la presencia de ruido en la imagen original. Sin embargo, algunas de las desventajas que se puede identificar al implementar este algoritmo se encuentran en el suavizado, puesto que, si se aumenta el de la mascarilla, se logra reducir el ruido, pero esto lleva a difuminar los bordes y se pierde la calidad al momento de calcular la orientación. A pesar de esto al momento de procesar las imágenes del modelo se obtuvieron muy buenos resultados, así como se muestra en las siguientes imágenes:

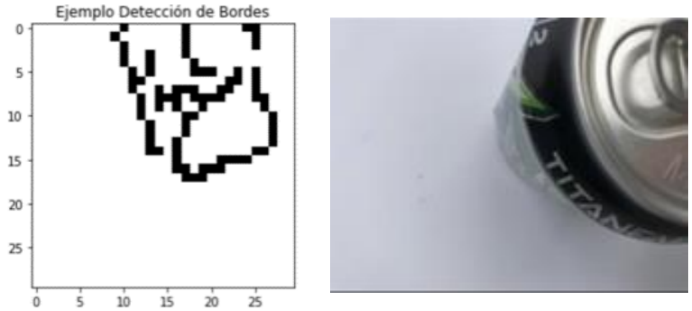

Fig. 6. Ejemplo 1 de detección de bordes

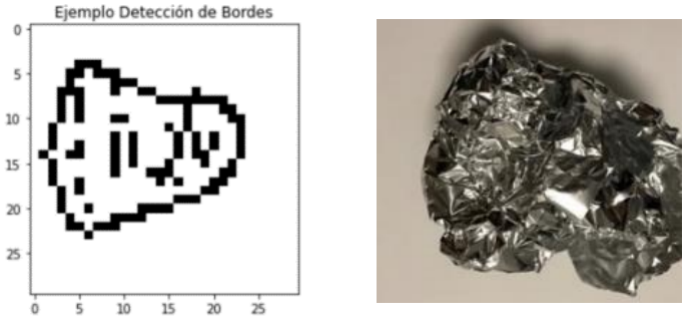

Fig. 7. Ejemplo 2 de detección de bordes.

II-D. Aplicación de los conceptos

Después de definir las funciones de preprocesamiento, es decir, la detección de bordes usando el algoritmo Canny, se aplica dicha función a las imágenes de entrenamiento para posteriormente entrenar el modelo de regresión logística binaria. Luego del entrenamiento, se realiza la predicción en muestras no observadas (test) con el modelo y, por último, se calculan las matrices de confusión, tanto de los datos de entrenamiento como de los datos del conjunto test y las métricas de evaluación para ambos para su posterior análisis.

III. RESULTADOS Y ANÁLISIS

Para poner a prueba y comparar los resultados que arrojan la regresión logística y la regresión logística robusta, se entrenan ambos modelos con el 80 % de los datos, es decir, con 641 imágenes y posteriormente se testea con el 20 % restante (162 imágenes). Además, los modelos dieron respuesta con base en 100 características de las imágenes. Los resultados obtenidos se muestran en sus respectivas matrices de confusión en las figuras 8,9, 10, y 11 y en la tabla I se evidencian sus métricas de evaluación para calificar la calidad del modelo

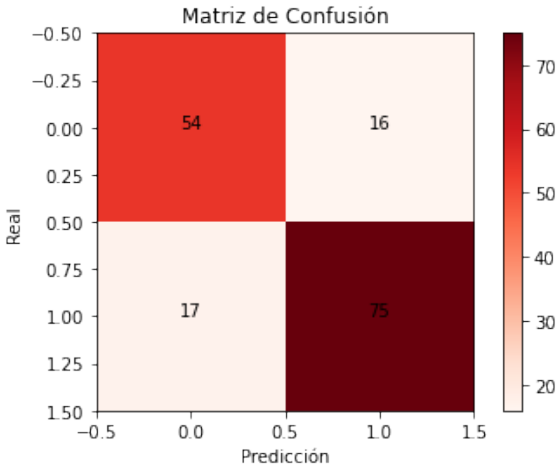

Fig. 8. Resultado obtenido a partir de la regresión logística clásica (máxima verosimilitud)

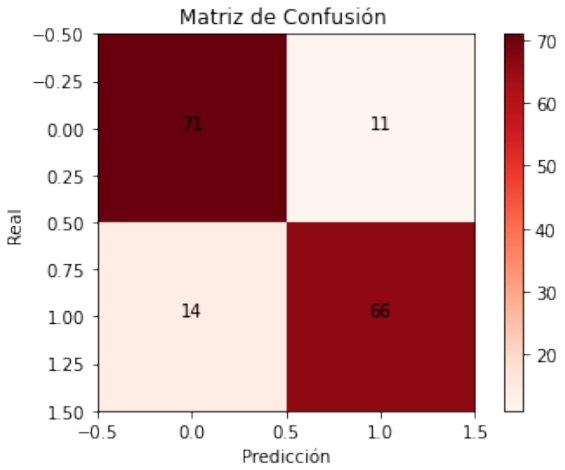

Fig. 9. Resultado obtenido de la regresión logística robusta realizada con el coeficiente de correlación de Spearman.

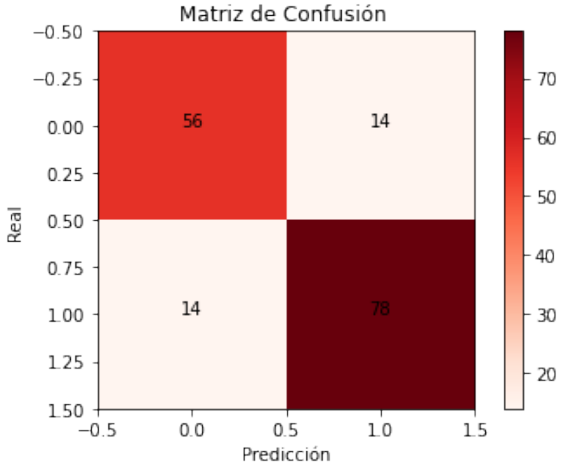

Fig. 10. Resultado obtenido de la regresión logística robusta realizada con el coeficiente de correlación de Kendall.

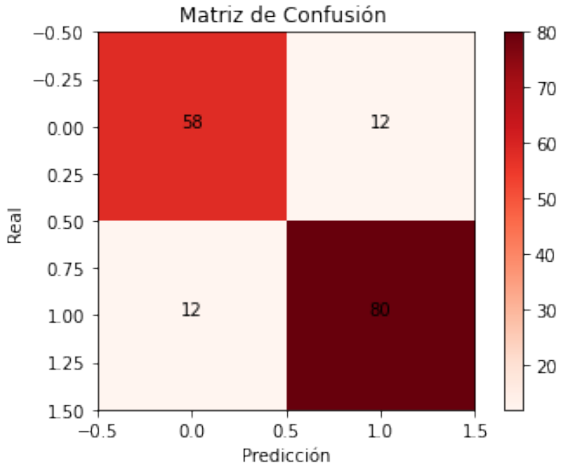

Fig. 11. Resultado obtenido de la regresión logística robusta realizada con el coeficiente de Pearson.

Cabe destacar que se realizaron varias predicciones y las matrices de confusión oscilaron alrededor de los valores presentados en las figuras anteriores.

TABLA I
MÉTRICAS DE EVALUACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA CON DIFERENTES TIPOS DE COEFICIENTE DE CORRELACIÓN

Métrica	Clásica	Kendall	Spearman	Pearson
Precisión	77.1 %	80 %	86.6 %	82.86 %
Sensibilidad	76.1 %	80 %	83.5 %	82.86 %
F1 Score	79.0 %	82.70 %	84.5 %	85.2 %

Las métricas utilizadas son: la precisión que es la relación entre las predicciones correctas y el número total de predicciones correctas previstas; la sensibilidad que se conoce como relación entre las predicciones positivas correctas y el número total de predicciones positivas; y el F1 score que se calcula a partir de la sensibilidad y la precision, es la métrica que mejor determina si el modelo clasifica los datos bien en sus clases.

De acuerdo con las métricas de evaluación del modelo de regresión logística con diferentes coeficientes de correlacion en la **TABLA I**, se puede observar que la regresión logística con Spearman es la mas precisa y sensible; con Pearson, se obtuvo una precisión y sensibilidad de 82.86 %, además el mayor F1 Score; con Kendall la precisión y sensibilidad oscilaron alrededor del 80 % y tuvo un F1 score menor a Spearman y Pearson, pero mayor a la regresion logística clásica. Respecto a la regresión logística clásica, se puede decir que es mas precisa que sensible, y acierta en un 77.1 % ´ en clasificar los residuos orgánicos cuando en realidad lo son.

Adicionalmente, para evaluar la robustez frente a datos raros de las regresiones en términos prácticos, se realiza una contaminación de los datos para evaluar los modelos con datos atípicos. Los resultados que se obtuvieron son los mostrados en las figuras 12, 13, 14 y 15 y sus respectivas métricas se encuentran en la **TABLA II**.

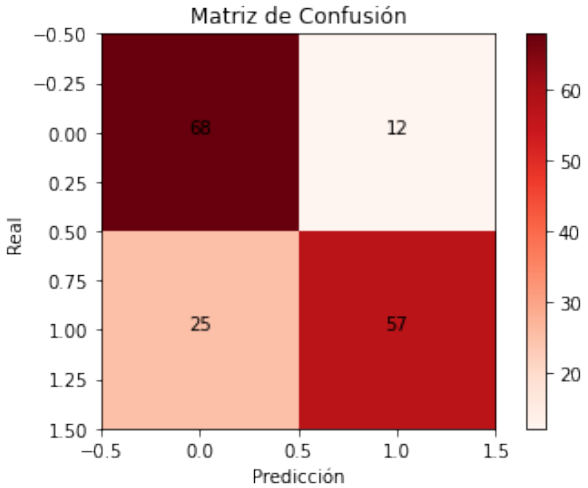

Fig. 12. Resultado obtenido a partir de la regresión logística clásica con los datos contaminados.

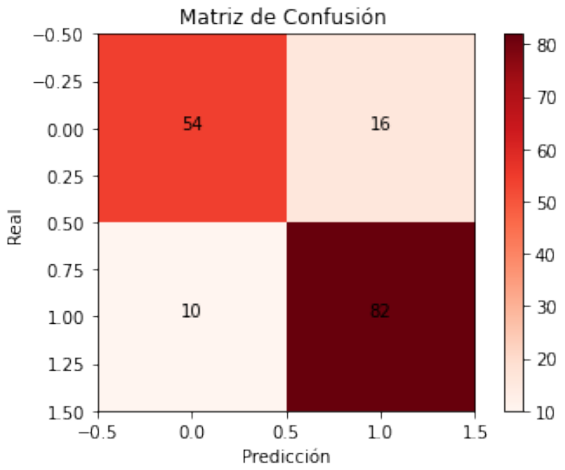

Fig. 13. Resultado obtenido a partir de la regresión logística robusta calculada con el coeficiente de correlación de Pearson y con los datos contaminados.

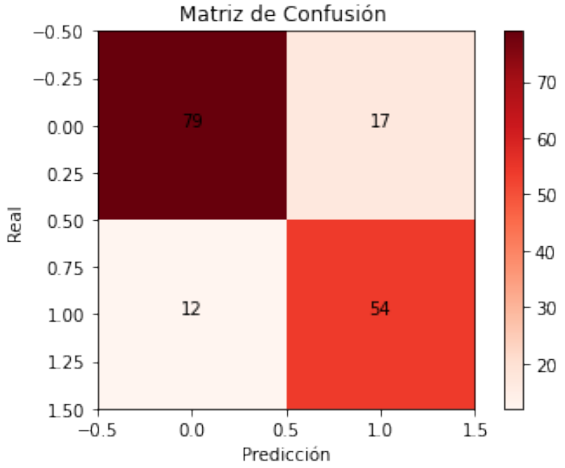

Fig. 14. Resultado obtenido a partir de la regresión logística robusta calculada con el coeficiente de correlación de Spearman y con los datoscontaminados.

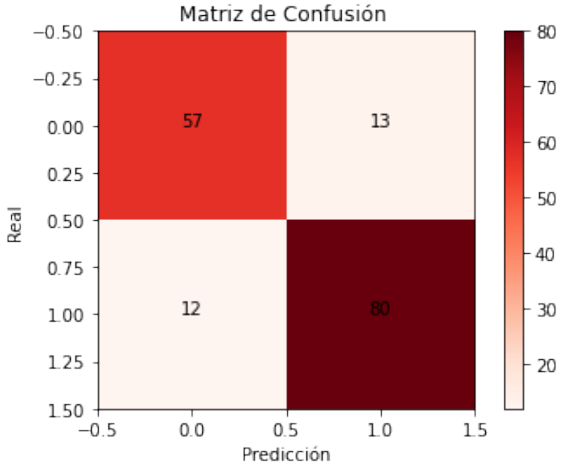

Fig. 15. Resultado obtenido a partir de la regresión logística robusta calculada con el coeficiente de correlación de Kendall y con los datoscontaminados.

TABLA II
MÉTRICAS DE EVALUACIÓN DE LAS REGRESIONES LOGÍSTICAS CON LOS DATOS CONTAMINADOS

Métrica	Clásica	Kendall	Spearman	Pearson
Precisión	85.00 %	81.43 %	82.29 %	0.77 %
Sensibilidad	73.121 %	82.61 %	86.81 %	0.84 %
F1 Score	77.20 %	84.50 %	82.09 %	84.00 %

Como se observa en la TABLA II, cuando se contaminan los datos, se tienen métricas F1 Score de las regresiones logísticas calculadas con los coeficientes de Kendal, Pearson y Spearman con valores de 84,5 %, 84 % y 82,09 %, respectivamente, mientras que la que obtuvo un porcentaje mas bajo fue la regresión logística clásica (77.2 %).

IV. CONCLUSIÓN

De acuerdo con los resultados obtenidos, se puede concluir que las regresiones logísticas calculadas con los coeficientes de Kendall, Pearson y Spearman son menos sensibles a datos atípicos respecto a la regresion logística clásica, pues los F1 score conseguidos luego de la contaminación de los datos oscilan alrededor de los que se calcularon para los tres casos en primer lugar, mientras que el F1 score para la regresión logística disminuyó respecto al anterior para este caso.
Es importante destacar que la regresion logística calculada con el coeficiente de Kendall es la mas robusta entre las 4 calculadas, debido a que tiene el F1 score más alto e incluso aumenta para los datos contaminados.
Finalmente, si bien la regresion logística clásica funciona correctamente, al ser más sensible a datos irregulares, se pueden obtener conclusiones erróneas o poco precisas, mientras que, usando su versión robusta, se puede conseguir mayor consistencia en los resultados.

REFERENCIAS

[1] C. Medina. “Manejo de residuos sólidos”. En: *Ciencia e Ingeniería Neogranadina* 8 (1999), págs. 135-144. DOI: https://doi.org/10.18359/rcin.1501.

[2] F. Pedregosa y col. “Scikit-learn: Machine Learning in Python”. En: *Journal of Machine Learning Research* 12 (2011), págs. 2825-2830.

[3] L. Flores. “Análisis estadístico de los factores de riesgo que influyen en la enfermedad Angina de Pecho”. En: *Tesis digitales UNMSM* (2002).

[4] Wikipedia. *Coeficiente de correlación de rango de Kendall — Wikipedia, La enciclopedia libre*. 2021.

[5] D. Plazas J. Vidal-Correa A. Tamayo. “Evaluation of Robust Covariance Estimation for Object Detection”. En: *Mathematical Engineering Universidad EAFIT* (2021).

[6] Wikipedia. *Coeficiente de correlación de Pearson — Wikipedia, La enciclopedia libre*. 2021.

[7] J. Valverde-Rebaza. “Detección de bordes mediante el algoritmo de Canny”. En: (2007).

[8] B. Hernández E. Aldair. “Graficación”. En: (2016). URL: https : / / beytiahernandez . wixsite . com / graficacionitc / algoritmode-canny..

[9] S. Patnaik Y.M. Yang. “Computing techniques in vision science”. En: *395 Springer* (2012).