

ESTUDIANDO LA FUNCIÓN DE PÉRDIDA PARA DATOS DISTRIBUCIONALES: APLICACIÓN A ÁRBOLES DE REGRESIÓN

CERVANTES, J. ¹

¹Universidad de Costa Rica

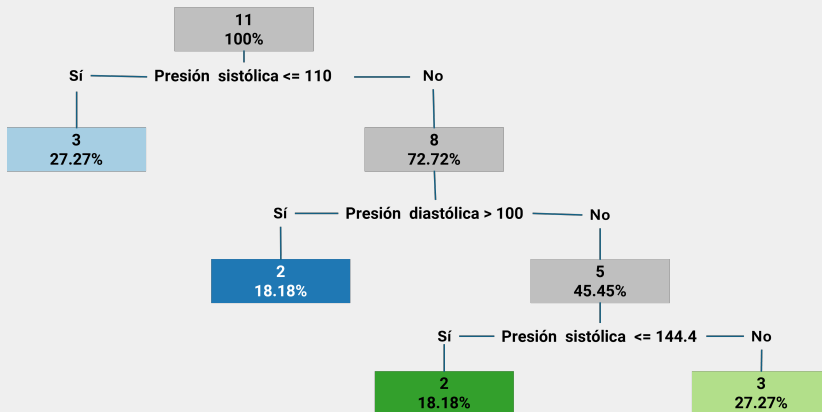
PRESENTADO POR:
JOSHUA CERVANTES

OCTUBRE, 2024

- 1 Introducción
- 2 Metodología
- 3 Aplicación
- 4 Conclusiones

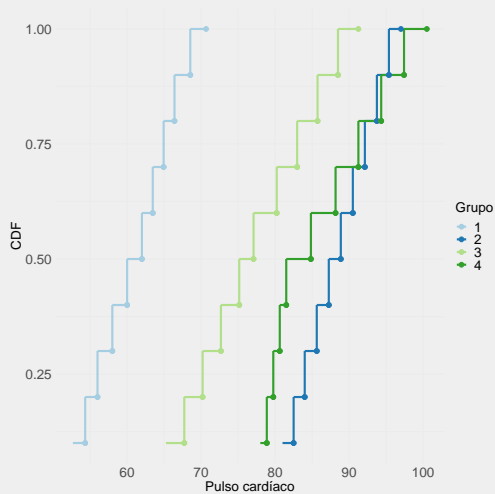
Se construyeron árboles “regresión” con datos tipos intervalo. A continuación se presenta un ejemplo de un árbol construido a partir los datos presentados en **Brito** y con el que se obtiene un $RMSE_M$ de 6.4779, un R^2 de 0.7193. y un Ω de 0.7739

Figura 1: Árbol para datos de cardiología



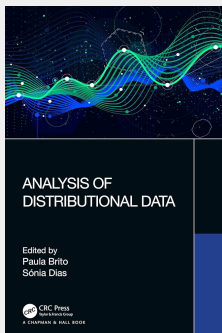
Fuente: Elaboración propia con datos de [1]

Figura 2: Ajuste probabilidad acumulada por grupo



Fuente: Elaboración propia con datos de [1]

Los objetos tipo histograma permiten comprender de mejor manera una amplia variedad de problemas, como se observó en el caso anterior, principalmente en el caso de regresión. El libro donde se ha encontrado un mayor material relacionado con regresión y pronosticación con este tipo de datos es en *Analysis of Distributional Data* (2024) de Paula Brito y Sonia Dias.



Trabajos estudiados:

- Se han encontrado trabajos los cuales tienen como objetivo trabajar datos distribucionales, realizar regresión sobre estos y realizar predicción. [1], [4], [1]
- Más allá del análisis simbólico de datos se encuentran trabajos en aprendizaje por refuerzo donde se emplea como variable de respuesta la distribución de los datos. [3], [7]
- Existen trabajos de árboles de decisión los cuales aprenden a partir de datos distribucionales como variable predictora. [5], [6]
- Hay trabajos donde el objetivo es aprender los parámetros que describen los de datos distribucionales como variable predictora. **Distributional regression forests for probabilistic precipitation forecasting in complex terrain** y **Investigating the Histogram Loss in Regression**.

- 1 Introducción
- 2 Metodología**
- 3 Aplicación
- 4 Conclusiones

Definition (Variable tipo histograma)

Sea Y una variable aleatoria numérica que puede tomar valores en un número finito de intervalos no solapados $\{l_j, u_j), j = 1, 2, \dots\}$ con $l_j \leq u_j$. Entonces si se agrega por algún concepto s_i la variable aleatoria tipo histograma de intervalos toma la forma de

$$Y(s_i) = Y_i = \{[l_{i1}, u_{i1}[, p_{i1}; \dots, [l_{im_i}, u_{im_i}[, p_{im_i}\}$$

Donde m_i es la cantidad de subintervalos del histograma para el individuo i (concepto s_i), p_{im_i} es el peso del subintervalo $[l_{im_i}, u_{im_i}[$ con $\sum_{j=1}^{m_i} p_{im_i} = 1$

Se asume distribución uniforme dentro de los intervalos.

Definition (Función distribución acumulada)

La función distribución acumulada es

$$F_i(x) = \begin{cases} 0 & x \leq l_{i1} \\ \frac{x-l_{i1}}{u_{i1}-l_{i1}}p_{i1} & l_{i1} \leq x < u_{i1} \\ p_{i1} + \frac{x-l_{i2}}{u_{i2}-l_{i2}}p_{i2} & l_{i2} \leq x < u_{i2} \\ \vdots & \\ 1 & x \geq u_{im_i} \end{cases}$$

Definition (Función cuantil)

La función distribución acumulada es

$$\psi_{Y_i}(t) = \begin{cases} l_{i1} + \frac{t}{w_{i1}}(u_{i1} - l_{i1}) & 0 \leq t < w_{i1} \\ l_{i2} + \frac{t}{w_{i2}}(u_{i2} - l_{i2}) & w_{i1} \leq t < w_{i1} + w_{i2} \\ \vdots & \\ l_{im_i} + \frac{t}{w_{im_i}}(u_{im_i} - l_{im_i}) & w_{im_i} \leq t \leq 1 \end{cases}$$

Con $w_{ij} = \sum_{k=1}^j p_{ik}$ si $j = 1, \dots, m_i$ y 0 en otro caso

Siguiendo a Irpino y Verde (2006) como lo cita [2] se presenta el algoritmo para construir los histogramas con la misma cantidad de intervalos, mismo peso y representados por sus histogramas.

Si se tiene el conjunto de histogramas Y_i con $i = 1, \dots, n$, sea W el conjunto de los pesos acumulados de las n distribuciones

$$W = \{w_{10}, \dots, w_{1m_1}, \dots, w_{nm_n}, \dots, w_{1m_1}\}$$

Se reorganizan los pesos sin repetición como

$$Z = \{w_0, \dots, w_j, \dots, w_m\}$$

Con $w_0 = 0$ y $w_1 = 1$ entonces se obtiene el histograma asociado a Y_i

$$H_i = \{[\psi_i(0), \psi(w_1)[, p_1; \dots; [\psi_i(w_{m-1}), \psi(w_m)[, p_m\}$$

Con $w_j = w_j - w_{j-1}$ y $j = 1, \dots, m$.

Nota: Se asume que todos los hitogramas anteriormente han sido construidos de esta forma por lo que se seguiran llamando Y_i , $\mathbf{X}_i = [X_{i1}, \dots, X_{ip}]$ va ser el vector de variables del individuo i y $\hat{Y}|\mathbf{X}_i$ es el valor predecido dado que se conoce \mathbf{X}_i .

[1] proponen un modelo llamado **Distribution and Symmetric Distribution Regression Model** el cual tiene la siguiente estructura:

$$\psi(t) = \psi_{\hat{Y}}(t) + e_i(t) = v + \sum_{j=1}^p a_j \psi_{X_j}(t) - b_j \sum_{j=1}^p b_j \psi_{X_j}(1-t) + e_i(t)$$

Donde ψ_{X_j} es la función cuantil de la variable X_j , $j \in \{1, \dots, p\}$, $t \in [0, 1]$ $a_j, b_j \geq 0$ y $v \in \mathbb{R}$. Los coeficientes son estimados minimizando

$$\sum_{i=1}^2 D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}_i}(t))$$

sujeto a $a_j, b_j \geq 0$ $j \in \{1, \dots, p\}$ y $v \in \mathbb{R}$.

Por otro lado [4] proponen su versión de la regresión lineal el cual es llamado **Two-component linear regression model** que es definido de la siguiente forma:

$$\psi_{Y_i}(t) = \beta_0 + \sum_{j=1}^p \beta_j \mu_{X_{ij}} + \sum_{j=1}^p \gamma_j \psi_{X_{ij}}^c(t) + e_i(t) = \psi_{\hat{Y}_i}(t) + e_i(t)$$

Donde se tiene que $\beta_j \in \mathbb{R}$, $j \in \{0, \dots, p\}$, $\psi_{X_{ij}}$ es la función cuantil del individuo i para la variable j , $\mu_{X_{ij}} = \int_0^1 \psi_{X_{ij}}(t) dt$ y $\psi_{X_{ij}}^c = \psi_{X_{ij}} - \mu_{X_{ij}}$.
Los coeficientes son estimados minimizando

$$\sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}_i}(t))$$

En este caso se seguirá la estructura propuesta por [5] **Breiman (1984)** para datos clásicos, adaptando la misma para poder trabajar con datos tipo histograma. Por lo que, se necesitará lo siguiente:

1. Una forma de poder medir el error de la estimación que se está realizando.
2. Un estimador.
3. Una forma de poder realizar separaciones y poder asignar los nuevos valores que sean observados, para datos tipo histograma.
4. Una regla que determine cuando un nodo es terminal.

En este caso el tercer punto puede tomarse un parámetro de penalización dada la complejidad del árbol tal y como lo realizan los autores de CART.

En este caso se quiere ver qué tanta similitud existe entre Y y \hat{Y} , de tal forma que se mida el error entre el valor real y la predicción. En el trabajo de predicción de series de tiempo de [2] se exploran distintas medidas de error entre el valor predicho y el observado. Los mismos concluyen que la mejor alternativa es utilizar una medidas de divergencia entre distribuciones.

Medida de divergencia	Definición
Kullback-Leibler	$D_{KL}(f, g) = \int_{\mathbb{R}} \log \frac{f(x)}{g(x)} f(x) dx$
Jeffrey	$D_J(f, g) = D_{KL}(f, g) + D_{KL}(g, f)$
Hellinger	$D_H(f, g) = \left[\int_{\mathbb{R}} (\sqrt{f(x)} - \sqrt{g(x)})^2 dx \right]^{1/2}$
Variación total	$D_{var}(f, g) = \int_{\mathbb{R}} f(x) - g(x) dx$
Wasserstein	$D_W(f, g) = \int_0^1 \psi_F(t) - \psi_G(t) dt$
Mallows	$D_M(f, g) = \left[\int_0^1 \psi_F(t) - \psi_G(t) ^2 dt \right]^{1/2}$
Kolgomorov	$D_K(f, g) = \max_{\mathbb{R}} F(x) - G(x) $

Cuadro 1: Medidas de divergencia más populares

La medida de divergencia de Mallows es la empleada, por lo siguientes aspectos:

- Esta definida en todo el soporte de las dos funciones de distribución.
- Los valores que puede tomar no están acotados por arriba, por lo que sí permite determinar qué tanta disimilitud existe entre las distribuciones, esto no sucede con Hellinger, variación total y Kolgomorov.
- Adicional tiene una interpretación intuitiva, la interpretación es como la de Distancia de Transportistas de Arena, siendo está interpretación la cantidad de esfuerzo requerido para pasar de una distribución a otra.
- Es una generalización de la distancia Euclídea.

Si se asume que se tienen dos histogramas con la misma cantidad de subintervalos y el mismo peso asociado en cada subintervalo, como se mostró anteriormente entonces se cumple lo siguiente:

Proposición

Si se asumen dos distribuciones empíricas Y_1 y Y_2 que pueden ser representadas por su funciones cuantil $\psi_{Y_1}(t)$ y $\psi_{Y_2}(t)$. Ambas escritas con la misma cantidad de m subintervalos y el mismo peso asociado en cada subintervalo. Entonces

$$D_M^2(\psi_{Y_1}(t), \psi_{Y_2}(t)) = \sum_{l=1}^m p_l \left[(c_{1l} - c_{2l})^2 + \frac{1}{3}(r_{1l} - r_{2l})^2 \right]$$

donde c_{1l} , c_{2l} y r_{1l} , r_{2l} con $l \in \{1, \dots, m\}$ son los centros y el rango del sub intervalo l de la distribución Y_1 y Y_2 respectivamente.

Demostración.

Si se considera la función cuantil $\psi_{Y_1}(t)$ y $\psi_{Y_2}(t)$ entonces se tiene

$$\begin{aligned} D_M^2(\psi_{Y_1}(t), \psi_{Y_2}(t)) &= \int_0^1 (\psi_{Y_1}(t) - \psi_{Y_2}(t))^2 dt \\ &= \sum_{l=1}^m \int_{w_{l-1}}^{w_l} \left[\left(c_{1l} + \left(\frac{2(t - w_l)}{w_l - w_{l-1}} - 1 \right) r_{1l} \right) - \left(c_{2l} + \left(\frac{2(t - w_l)}{w_l - w_{l-1}} - 1 \right) r_{2l} \right) \right]^2 dt \\ &= \sum_{l=1}^m \int_{w_{l-1}}^{w_l} \left[(c_{1l} - c_{2l}) + \left(\frac{2(t - w_l)}{w_l - w_{l-1}} - 1 \right) (r_{2l} - r_{1l}) \right]^2 dt \end{aligned}$$

Si se aplica el cambio de variable $v = \frac{t - w_l}{w_l - w_{l-1}}$ entonces se puede reescribir la integral como

$$\sum_{l=1}^m \int_0^1 p_l [(c_{1l} - c_{2l}) + (2v - 1)(r_{2l} - r_{1l})]^2 dv = \sum_{l=1}^m p_l [(c_{1l} - c_{2l})^2 + \frac{1}{3}(r_{2l} - r_{1l})^2]$$



Siguiendo lo propuesto por [5] y de forma análoga a lo visto en regresión lineal se debe encontrar un estimador \hat{Y} que minimice lo siguiente

$$\frac{1}{n} \sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}}(t))$$

En el caso de los datos clásicos el estimador que minimiza esto es el promedio. De una forma similar aquí el estimador es lo que llamarán [3] como promedio.

Definition (Promedio de Fréchet o Karcher)

Un promedio tipo Fréchet \bar{y} es un objeto x que lleva a la solución del siguiente problema de minimización

$$\bar{y} \sim \arg \min_x \sum_{i=1}^n w_i d^2(y_i, x)$$

Donde se tiene que d es una métrica. Por lo que si se asume que todos los individuos tienen el mismo peso lo que hay que hacer es encontrar \bar{Y} tal que

$$\bar{Y} \sim \arg \min_{\hat{Y}} \sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}}(t)) \quad (1)$$

Proposición

El valor mínimo para la ecuación (1) se encuentra para la función densidad que tiene asociada la función cuantil $\bar{\psi}(t)$ tal que

$$\bar{\psi}(t) = \frac{1}{n} \sum_{i=1}^n \psi_i(t), \quad \forall t \in [0, 1]$$

Demostración.

Se quiere solucionar lo siguiente

$$\arg \min_{\hat{Y}} \sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}}(t)) = \sum_{i=1}^n \int_0^1 (\psi_{Y_i}(t) - \psi_{\hat{Y}}(t))^2 dt$$

Si se toma \bar{Y} asociada a $\bar{\psi}$ y $t \in [0, 1]$ entonces

$$\frac{d}{d\bar{\psi}(t)} \left[\sum_{i=1}^n (\psi_i(t) - \bar{\psi}(t))^2 \right] = 0 \Rightarrow \bar{\psi}(t) = \frac{1}{n} \sum_{i=1}^n \psi_i(t)$$



Proposición

El valor mínimo para la ecuación (1) es el histograma que tiene asociados los centros y rangos

$$\bar{c}_j = \frac{1}{n} \sum_{i=1}^n c_{ij} \quad \bar{r}_j = \frac{1}{n} \sum_{i=1}^n r_{ij}, \quad j = 1, \dots, m$$

Demostración.

La prueba surge de reescribir la distancia de Mallows a partir de los centros y rangos y aplicar un argumento similar al anterior. Es decir encontrar los mínimos a partir de las derivadas. \square

A partir de lo anterior entonces se puede reescribir el histograma (baricentro) como

$$\bar{Y} = \{[\bar{c}_1 - \bar{r}_1, \bar{c}_1 + \bar{r}_1], p_1; \dots; [\bar{c}_m - \bar{r}_m, \bar{c}_m + \bar{r}_m], p_m\}$$

Entonces

$$\begin{aligned} \bar{l}_m &= \bar{c}_m - \bar{r}_m & \bar{u}_m &= \bar{c}_m + \bar{r}_m \\ &= \frac{1}{n} \left[\sum_{i=1}^m \left(\frac{l_i + u_i}{2} - \frac{u_i - l_i}{2} \right) \right] & &= \frac{1}{n} \left[\sum_{i=1}^m \left(\frac{l_i + u_i}{2} + \frac{u_i - l_i}{2} \right) \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^m l_i \right] & &= \frac{1}{n} \left[\sum_{i=1}^m u_i \right] \end{aligned}$$

Definition (Media simbólica empírica)

Se define la media empírica de acuerdo a lo propuesto por [4] como

$$\bar{Y}^* = \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^{m_i} c_{il} p_{il}$$

De lo anterior es fácil ver

$$\bar{Y}^* = \sum_{l=1}^m \bar{c}_l p_l$$

Esto muestra que la media empírica es el promedio de la media de Fréchet.

Dado que se tengan variables de tipo histograma, se debe idear una manera de hacer una pregunta binaria y establecer una regla de asignación. En este sentido se han encontrado dos alternativas la primera propuesta por [4] en Divisive Clustering of Histogram Data y también fue encontrada la opción de [5] y [6]. Se ha optado por la opción de Brito al considerarse más interpretable y de menor costo computacional, pero esto no significa que sea la mejor. En este caso las preguntas binarias están definidas como en el caso clásico

$$P_j := X_j \leq v_j, j = 1, \dots, p$$

Cada condición P_j lleva a una bipartición. En este caso un individuo i cumple la condición $P_j := X_j \leq v_j$ si solo si la mediana de X_{ij} es menor o igual que v_j .

Dado lo anterior entonces se tiene que los nodos pueden ser definidos reglas formadas por datos clásicos y datos de tipo histograma. Siguiendo a [5] se proceden a construir el árbol. El valor predicho en el nodo a es \bar{Y}^a , A es el árbol construido y \tilde{A} es el conjunto de los nodos terminales

$$R(A) = \frac{1}{n} \sum_{a \in \tilde{A}} \sum_{\mathbf{X}_i \in a} D_M^2(\psi_{Y_i}(t), \psi_{\bar{Y}^a}(t))$$

Si se toma

$$R(a) = \frac{1}{n} \sum_{\mathbf{X}_i \in a} D_M^2(\psi_{Y_i}(t), \psi_{\bar{Y}^a}(t))$$

Entonces

$$R(A) = \sum_{a \in \tilde{A}} R(a)$$

Definition

La mejor separación s^* de a es una separación que se encuentra en S , espacio de posibles separaciones, tal que decrece el error en $R(A)$. Más precisante para cualquier separación s del nodo a en a_L y a_R sea

$$\Delta R(s, a) = R(a) - R(a_L) - R(a_R)$$

La mejor separación s^* es aquella que

$$\Delta R(s^*, a) = \max_{s \in S} \Delta R(s, a)$$

En este caso para crecer los árboles es necesario tener un criterio de parada. En este caso se toma un valor mínimo de elementos que debería tener un nodo, y se cuenta con un error de complejidad

$$R_{\alpha}(A) = R(A) + \alpha|\tilde{A}|$$

El valor de α puede ser escogido por validación cruzada.

[1] propone una medida de ajuste que es la siguiente

Definition

Si se consideran el valor observado Y y un valor predicho \hat{Y} para una variable histograma con cuantiles $\psi_{Y_i}(t)$ y $\psi_{\hat{Y}|\mathbf{X}_i}(t)$. Si se considera la media simbólica empírica de **Billard** \bar{Y}^* entonces

$$\Omega = 1 - \frac{\sum_{i=1}^n D_M^2(\psi_{\hat{Y}|\mathbf{X}_i}(t), \psi_{Y_i}(t))}{\sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \bar{Y}^*)}$$

Este valor se encuentra entre 0 y 1 la prueba se puede revisar en [1]

Siguiendo la idea de [4], sin tomar en consideración la restricción a que el valor esté entre 0 y 1, entonces se define el siguiente estimador:

Definition (Error cuadrático relativo y pseudo R^2)

Si se considera Y_i los valores observados y $\hat{Y}|X_i$ los valores del estimador para una variable histograma, con cuantiles $\psi_{Y_i}(t)$ y $\psi_{\hat{Y}|X_i}(t)$. Y la media de Fréchet \bar{Y} entonces error cuadrático relativo es

$$RE = \frac{\sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}|X_i}(t))}{\sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\bar{Y}}(t))}$$

Y el pseudo R^2

$$R^2 = 1 - RE$$

En este caso tanto Ω como R^2 no tienen por qué ser mayores a 0, pero da una idea del ajuste del modelo. Un valor negativo indica que el error cometido es peor de que si se tomara la media de Fréchet o la media empírica.

- 1 Introducción
- 2 Metodología
- 3 Aplicación**
- 4 Conclusiones

A parte del modelo mostrado anteriormente propuesto [1] propone el siguiente modelo

$$\psi_{\hat{Y}|\mathbf{X}_i}(t) = \psi_{Constant}(t) + \sum_{j=1}^p a_j \psi_{X_{ij}}(t) - \sum_{j=1}^p b_j \psi_{X_{ij}}(1-t)$$

con $t \in [0, 1]$ y $a_j, b_j \geq 0$, y $j \in \{1, 2, \dots, p\}$ y

$$\psi_{Constant}(t) = \psi_c(t) + \psi_r(t)$$

$$\psi_c(t) = \begin{cases} c_v & 0 \leq t < w_1 \\ c_v + r_{v_1} + r_{v_2} & w_1 \leq t < w_2 \\ c_v + r_{v_1} + 2r_{v_2} + r_{v_3} & w_2 \leq t < w_3 \\ \vdots & \\ c_v + r_{v_1} + 2 \sum_{l=2}^{m-1} r_{v_l} + r_{v_m} & w_{m-1} \leq t < w_m \leq 1 \end{cases}$$

para cada subintervalo l

$$\psi_r(t) = \left(\frac{2(t - w_{l-1})}{w_l - w_{l-1}} - 1 \right) r_{v_l}, \quad w_{l-1} \leq t < w_l$$

con $r_{v_l} \geq 0$, $l \in \{1, 2, \dots, m\}$ y $c_v \in \mathbb{R}$

Cuadro 2: Conjunto de datos cardiológico, histograma

Frecuencia cardíaca	Presión sistólica	Presión diastólica
$\{[40; 60), 0.8; [60; 68], 0.2\}$	$\{[90; 95), 0.2; [95; 100], 0.8\}$	$\{[50; 60), 0.4; [60; 70], 0.6\}$
$\{[60; 70), 0.5; [70; 72], 0.5\}$	$\{[90; 110), 0.4; [110; 130], 0.6\}$	$\{[70; 80), 0.2; [80; 90], 0.8\}$
\vdots		
$\{[86; 89), 0.6, [89; 100], 0.4\}$	$\{[110; 135), 0.2; [135; 150], 0.8\}$	$\{[78; 88), 0.2; [88; 100], 0.8\}$

Fuente: Elaboración propia con datos de [1]

En el caso de [1] se obtiene los siguientes resultados

Cuadro 3: Parámetros y medidas para modelos DSD

Método	Parámetros					Medidas de ajuste		
	$v/\psi(t)$	a_1	b_1	a_2	b_2	Ω	R^{2**}	$RMSE$
DSD I	10.0366	0.0206	0.0551	0.8397	0	0.7452	0.6726	6.8484
DSD II	$\psi_{Constant}(t)$	0	0	0.77	0	0.7533	0.6830	6.7394

Fuente: Elaboración propia con datos de [1].

Nota: ** estimado a partir de valores presentados en paper original

En el caso del modelo propuesto aquí se tiene

Cuadro 4: Medidas para el modelos en tabla datos cardiológico

Método	Medidas de ajuste		
	Ω	R^{2**}	$RMSE$
DSD I	0.7352	0.6726	6.8484
DSD II	0.7533	0.6830	6.7394
Árbol, Min size = 2	0.7739	0.7193	6.4779

Fuente: Elaboración propia con datos de [1].

Figura 3: Conjunto de datos ozono, histograma

ID	Y Ozone.Conc. (ppb)		X ₁ Temperature (C)		X ₂ Solar Radiation (Watt/M ²)		X ₃ Wind Speed (m/Sec)	
	Bin	p	Bin	p	Bin	p	Bin	p
I1	[8.77 – 16.62)	0.01	[8.45 – 11.65)	0.01	[25.29 – 75.88)	0.01	[0.10 – 0.35)	0.01
	[16.62 – 17.54)	0.01	[11.65 – 13.06)	0.01	[75.88 – 108.27)	0.01	[0.35 – 0.41)	0.01

	[65.68 – 67.78)	0.01	[28.87 – 29.23)	0.01	[914.12 – 933.30)	0.01	[3.52 – 3.79)	0.01
	[67.78 – 89.60]	0.01	[29.23 – 30.18]	0.01	[933.30 – 942.00]	0.01	[3.79 – 4.48]	0.01
I2	Bin	p	Bin	p1	Bin	p	Bin	p
	[9.00 – 15.00)	0.01	[9.50 – 9.75)	0.01	[49.00 – 56.16)	0.01	[0.10 – 0.55)	0.01
	[15.00 – 17.00)	0.01	[9.75 – 10.38)	0.01	[56.16 – 71.50)	0.01	[0.55 – 0.80)	0.01

	[54.24 – 58.00)	0.01	[29.02 – 29.60)	0.01	[910.00 – 916.84)	0.01	[7.52 – 8.37)	0.01
I3	[58.00 – 63.00]	0.01	[29.60 – 30.70]	0.01	[916.84 – 944.00]	0.01	[8.37 – 9.60]	0.01
	Bin	p	Bin	p1	Bin	p	Bin	p
	[9.25 – 17.99)	0.01	[17.57 – 20.13)	0.01	[52.57 – 78.67)	0.01	[0.08 – 0.26)	0.01
	[17.99 – 20.31)	0.01	[20.13 – 20.63)	0.01	[78.67 – 105.48)	0.01	[0.26 – 0.38)	0.01

...	[62.38 – 64.11)	0.01	[36.10 – 36.42)	0.01	[979.18 – 990.02)	0.01	[3.77 – 4.07)	0.01
	[64.11 – 69.45]	0.01	[36.42 – 37.07]	0.01	[990.02 – 1020.00]	0.01	[4.07 – 4.81]	0.01
...

Fuente [4]

En el caso de [4] se otbeien los siguientes resultados

Cuadro 5: Parámetros y medidas para modelo de dos componentes

Método	Parámetros						
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
		Temp	Sol. Rad.	Wind Sp.	Temp	Sol. Rad.	Wind Sp.
Dos componentes	2.927	-0.346	0.070	0.395	0.915	0.018	1.887

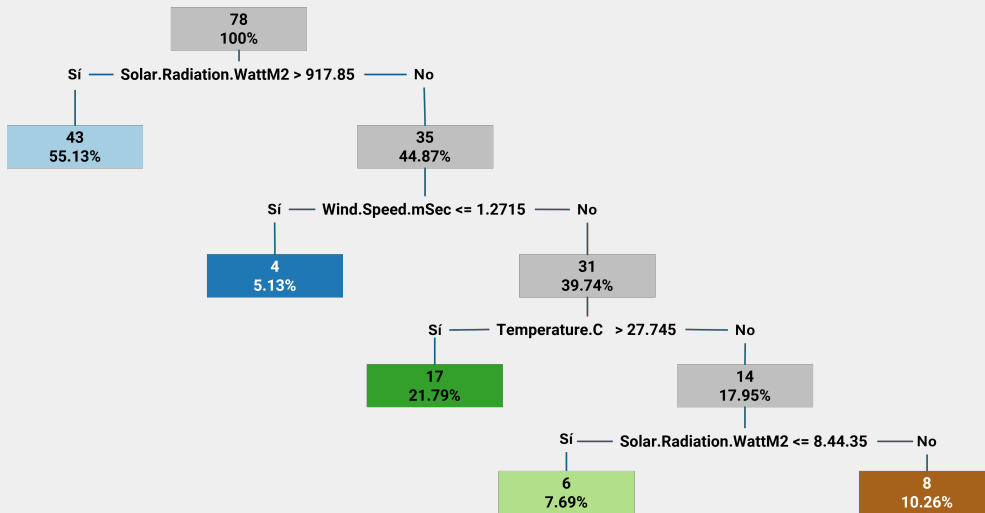
Fuente: Elaboración propia con datos de [4].

Cuadro 6: Medidas de los modelos en tabla de Ozono

Método	Medidas de ajuste		
	Ω	R^2	$RMSE$
Dos componentes	0.7420	0.4600	7.0000
Árbol, min size = 3	0.6286	0.2034	8.3179

Fuente: Elaboración propia con datos de [4].








Figura 4: Árbol para datos de ozono














Fuente: Elaboración propia con datos de [4].

- 1 Introducción
- 2 Metodología
- 3 Aplicación
- 4 Conclusiones**

- El método de árboles propuesto aquí muestra ser un buen estimador.
- Los árboles aquí presentados permiten interpretar de mejor manera los resultados obtenidos a partir del modelo.
- Se generaliza en cierta forma los árboles de regresión tradicionales.
- Se debe mejorar la implementación de estos árboles.
- Se pueden explorar otras alternativas para poder tomar la decisión de cómo realizar los cortes.
- Se debe buscar una mayor fundamentación estadísticas-probabilística.
- Estudiar a mayor profundidad aquellos trabajos de aprendizaje por refuerzo.

-  ARROYO, J. (2024). FORECASTING DISTRIBUTIONAL TIME SERIES. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 339-374)
-  ARROYO, J. & MATÉ, J. (2008). *MÉTODOS DE PREDICCIÓN PARA SERIES TEMPORALES DE INTERVALOS E HISTOGRAMAS*[TÉSIS PARA OBTENCIÓN DE DOCTORADO, UNIVERSIDAD PONTICIA COMILLAS]
-  BELLEMARE, M. G., DABNEY, W., & MUNOS, R. (2017, JULY). A DISTRIBUTIONAL PERSPECTIVE ON REINFORCEMENT LEARNING. IN *INTERNATIONAL CONFERENCE ON MACHINE LEARNING* (PP. 449-458). PMLR.
-  BILLARD, L. & DIDAY, E. (2006) *SYMBOLIC DATA ANALYSIS*. JOHN WILEY & SONS, LTD.
-  BREIMAN, L., FRIEDMAN, J., OLSHEN, R., & STONE, C. (1984). *CART: CLASSIFICATION AND REGRESSION TREES*.
-  BREINMAN, L. (2001) RANDOM FORESTS. *MACHINE LEARNING*, 45, (PP. 5-32).
-  DABNEY, W., ROWLAND, M., BELLEMARE, M., & MUNOS, R. (2018, APRIL). DISTRIBUTIONAL REINFORCEMENT LEARNING WITH QUANTILE REGRESSION. IN *PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE* (VOL. 32, No. 1).

-  DIAS, S., & BRITO, P. (2024). REGRESSION ANALYSIS WITH THE DISTRIBUTION AND SYMMETRIC DISTRIBUTION MODEL. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 295-317)
-  DIAS, S., & BRITO, P. (2024). FUNDAMENTAL CONCEPTS ABOUT DISTRIBUTIONAL DATA. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 3-35)
-  DIAS, S., & BRITO, P. (2024). DESCRIPTIVE STATISTICS BASED ON FREQUENCY DISTRIBUTION. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 37-56)
-  MARIE, C., & BRITO, P. (2024). DIVISIVE CLUSTERING OF HISTOGRAM DATA. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 128-137)
-  GURUNG, R. B., LINDGREN, T., & BOSTRÖM, H. (2016, MARCH). LEARNING DECISION TREES FROM HISTOGRAM DATA USING MULTIPLE SUBSETS OF BINS. IN *THE TWENTY-NINTH INTERNATIONAL FLAIRS CONFERENCE*.
-  GURUNG, R. B., LINDGREN, T., BOSTRÖM, H. (2016, MARCH). LEARNING DECISION TREES FROM HISTOGRAM DATA USING MULTIPLE SUBSETS OF BINS. IN *THE TWENTY-NINTH INTERNATIONAL FLAIRS CONFERENCE*.

-  HENNIG, C., & KUTLUKAYA, M. (2007). SOME THOUGHTS ABOUT THE DESIGN OF LOSS FUNCTIONS. *REVSTAT-STATISTICAL JOURNAL*, 5(1), (PP. 19-39).
-  IMANI, E., LUEDEMANN, K., SCHOLNICK-HUGHES, S., ELELIMY, E., & WHITE, M. (2024). INVESTIGATING THE HISTOGRAM LOSS IN REGRESSION. *ARXIV PREPRINT ARXIV:2402.13425*.
-  IRPINO, A., & ROSANNA, V. (2024). DESCRIPTIVE STATISTICS FOR NUMERIC DISTRIBUTIONAL DATA. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 57-79)
-  IRPINO, A., & ROSANNA, V. (2024). REGRESSION ANALYSIS OF DISTRIBUTIONAL DATA BASED ON A TWO-COMPONENT MODEL. EN *ANALYSIS OF DISTRIBUTIONAL DATA* (PP. 319-337)
-  SCHLOSSER, L., HOTHORN, T., STAUFFER, R., & ZEILEIS, A. (2019). DISTRIBUTIONAL REGRESSION FORESTS FOR PROBABILISTIC PRECIPITATION FORECASTING IN COMPLEX TERRAIN.